

Integrating Motivational Interviewing Principles and Large Language Models in Automated Behaviour Change Support



Inaugural-Dissertation zur Erlangung der Doktorwürde der Fakultät für Sprach-,
Literatur- und Kulturwissenschaften der Universität Regensburg

Vorgelegt von
Selina Meyer
aus München
Regensburg 2024

Abgabe: 30. September 2024

Tag der Disputation: 04. April 2025

Erstgutachter: PD Dr. David Elsweiler

Zweitgutachter: Prof. Dr. Ian Ruthven

Acknowledgements

My supervisor, David Elswiler, once compared doing your PhD to navigating an untouched jungle: you painstakingly carve out a path, only to realize you've hit a dead end and must find another way. It doesn't help that you have no map—nor are you even sure where you're going. Many times, you feel like you won't make it. You want to give up—until you realize there's no way back. The only way is forward.

I have learned a lot during this time - about resilience, about my capabilities and limitations, and about how rewarding it is to challenge myself and test my limits - something I have started enjoying also outside of work. I have learned about ambition, self-acceptance and having patience with myself. I have discovered that limits can shift, and that what seems impossible one day can feel almost easy just a short time later.

This growth didn't happen in isolation, because fortunately, I did not have to undertake this expedition alone.

First and foremost, every journey needs a guiding force—someone who keeps you from losing your way, who believes in you even when you've lost faith in yourself, and who never, even for a second, doubts your ability to reach your destination. David, thank you for being that force. Your encouragement and guidance have given me strength at every step of the way. You lifted me up when things didn't go my way and celebrated with me when they did. You never lost patience and always reminded me to keep in mind the bigger picture. I could not have asked for a better supervisor and am deeply grateful for everything you have done to get me to where I am now.

I'm also grateful to my second examiner, Ian Ruthven—whom I first met at CHIIR in 2024—for generously offering his time to read and review my thesis. It was a pleasure to discuss your thoughtful and insightful questions at the defense.

My gratitude also goes to the German Academic Scholarship Foundation for their financial and ideational support of this project.

There are many others who accompanied my way through the PhD jungle and who I am beyond grateful for—some who have been there from the start, others met along the way. Their support made the journey possible.

Back when I was still a master's student, Markus Kattenbeck offered me a job as a student assistant, helping organize a conference at the Chair of Information Science at the University of Regensburg. With that offer, he gave me a first glimpse of academia—and quietly set me on the path I would later follow. I'm not sure I would have taken the first step without you.

Bernd Ludwig and Udo Kruschwitz not only offered advice and guidance throughout my thesis, but also did their best to fund me and my project over the years. Without you, finishing this dissertation would not have been possible. Thank you for your continued support.

Spending my days at the office with my colleagues made the time fly by. Celebrating our wins (big and small), mourning our losses, sharing gossip and emotional support and taking Schafkopf-breaks are all things I will miss dearly. Thanks to Alex Frummet, Gregor Donabauer, Samy Ateia, Manuel Ullmann, Ray Fülöp, Robert Jackermeier, Melanie Kilian, Qing Zhang, Elisabeth Reitmeier and Jürgen Reischer.

In the fall of 2022, I had the chance to spend two months at CiTIUS at the University of Santiago de Compostela for a research visit. The experiments I worked on there took me out of my comfort zone and led to fruitful collaborations and new friendships. Thank you to David Losada for supervising my visit. Marcos Fernández-Pichel and Lorenzo Vaquero, Luca, Peyman, Faranak, Rhos, and everyone else—thank you for making me feel welcome, for showing me around Galicia, and for treating me like a friend from day one.

Of course, there is also life outside the PhD. That life was filled with friends and family who recharged my batteries, brought me joy, and got me out of my own head. I'm especially grateful to Betty and Max, who were always there for me, even when their own lives were overwhelming.

I am forever grateful to my parents—all four of them—and the rest of my family for their unwavering support, unconditional love and for reminding me every day that I have nothing to prove.

And finally, to my partner Max: thank you for improving every aspect of my life. For being my biggest fan and supporter in everything I do. For helping me rediscover my carefree (or at least less worried) self, for pushing me outside my comfort zone, for reminding me to enjoy the little things, to take life a little less seriously, and for proving just how much fun Schnapsideen can be. I could probably have done this without you, but I wouldn't have enjoyed it half as much.

I consider myself incredibly lucky to have such a strong support network and hope to never take it for granted. I will do my best to pay it forward and help others navigate the jungle.

Abstract

This thesis explores the application of large language models (LLMs) in the context of behaviour change support systems, focusing on the integration of Motivational Interviewing (MI) principles to enhance the effectiveness and safety of these interactions. This work addresses both the technical and theoretical aspects of applying conversational AI to intentional behaviour change scenarios.

The primary technical contributions of this research include an annotated dataset of written conversations about behaviour change, transferable and robust classification algorithms for identifying behaviour change language, and strategic methodologies for significantly reducing harms and unsuitable outputs in LLM-based conversations. An annotation scheme was also developed to identify and mitigate unsuitable behaviours in LLM outputs, accompanied by an annotated dataset of real human-LLM interactions, which provides insights into user behaviour and system performance in practical settings.

This project provides novel insights into the domain independence of behaviour change language, demonstrating the applicability of the developed systems across various contexts and target behaviours. Furthermore, the project explores user behaviour and expectations during interactions with LLMs, offering a deeper understanding of how users engage with and respond to LLM-based behaviour change support.

In general, this thesis contributes to the field of conversational AI by showcasing how LLMs can be used effectively and safely to support behaviour change, emphasising the importance of tailoring interactions based on MI principles and user-specific communication patterns. The findings not only advance our understanding of the technical and interactional dynamics of conversational AI but also set the stage for future research in the deployment of these technologies in diverse real-world applications. All code is made available on GitHub.

Zusammenfassung

Diese Arbeit untersucht die Anwendung von Large Language Models (LLMs) im Kontext von Systemen zur Unterstützung von Verhaltensänderungen und konzentriert sich dabei auf die Integration von Prinzipien der motivierenden Gesprächsführung (MI), um die Effektivität und Sicherheit dieser Interaktionen zu verbessern. Diese Arbeit befasst sich sowohl mit den technischen als auch mit den theoretischen Aspekten der Conversational AI in Szenarien zur beabsichtigten Verhaltensänderung.

Zu den primären technischen Beiträgen dieser Forschung gehören ein annotierter Datensatz schriftlicher Konversationen über Verhaltensänderung, übertragbare und robuste Klassifizierungsalgorithmen zur Identifizierung von Sprache zur Verhaltensänderung und strategische Methoden zur signifikanten Reduzierung von gefährlichen und unangemessenen Textausgaben in LLM-basierten Konversationen. Darüber hinaus wurde ein Annotationsschema entwickelt, um ungeeignete Verhaltensweisen in LLM-Outputs zu identifizieren und abzuschwächen, begleitet von einem annotierten Datensatz echter Mensch-LLM-Interaktionen, der Einblicke in das Nutzerverhalten und die Systemleistung in der Praxis bietet.

Dieses Projekt bietet neue Einblicke in die Domänenunabhängigkeit von Sprache über Verhaltensänderung und demonstriert die Anwendbarkeit der entwickelten Systeme in verschiedenen Kontexten und für verschiedene Verhaltensänderungen. Darüber hinaus erforscht das Projekt das Verhalten und die Erwartungen von NutzerInnen während der Interaktion mit LLMs und bietet so ein tieferes Verständnis dafür, wie diese mit LLM-basierten Verhaltensmodellen umgehen und darauf reagieren. Sämtlicher Code wird auf GitHub veröffentlicht.

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Research Goals	4
1.3	Outline	4
1.4	Publications Relating to This Thesis	4
I	Grounding	7
2	Theoretical Background	9
2.1	Behaviour Change	9
2.1.1	Behaviour and Change in Information Science	9
2.1.2	Behavioural Science Models of Behaviour Change	12
2.1.2.1	Change as Process: The Transtheoretical Model	13
2.1.2.2	Social Cognitive Theory, Theory of Planned Behaviour, and the Information-Motivation-Behavioural Skills Model	14
2.1.3	Behaviour Change in Human-Computer Interaction	17
2.2	Motivational Interviewing	20
2.2.1	Effectiveness and Application Areas of Motivational Interviewing	20
2.2.2	Principles of Motivational Interviewing	21
2.2.3	Role of Motivational Interviewing Throughout Behaviour Change	23
2.3	Conversational Agents	25
2.3.1	Rule-Based and Retrieval-Based Conversational Agents	26
2.3.2	Generative Conversational Agents	26
2.3.2.1	Neural Architectures: RNN, LSTM, and Transformer	27
2.3.3	Large Language Models	27
2.3.4	Advantages and Disadvantages of Text Generation Approaches	29
2.4	Chapter Summary	30
3	Related Work	31
3.1	Conversational Behaviour Change Support Systems	31
3.1.1	Classical Approaches	32
3.1.1.1	Short-Form Conversation	32
3.1.1.2	Constrained Input	33
3.1.1.3	Free-Text Input	34

3.1.2	LLM-based Systems	36
3.1.3	User Perceptions of Conversational Behaviour Change and Mental Health Support Systems	38
3.2	Understanding Users Writing About Change	41
3.2.1	Resources and Computational Methods for the Analysis of MI Client Language	41
3.2.1.1	Public Datasets with MI Client Annotations	42
3.2.1.2	Computational Methods for MI Client Annotations	44
3.2.2	Analysing Written Behaviour Change Language	46
3.3	Reacting	47
3.3.1	Appropriate and Desirable Behaviour	47
3.3.1.1	Empathy	49
3.3.1.2	Therapist Empathy and Adherence to Counselling Principles	50
3.3.1.3	Defining Appropriate and Desirable Behaviour for LLM-based Conversational Behaviour Change Support Systems	51
3.3.2	Resources and Computational Methods for the Generation and Evaluation of Counselling-Style Text	52
3.3.2.1	Public Datasets With MI Counsellor Annotations	52
3.3.2.2	Text Generation for Counselling Contexts	53
3.3.2.3	Evaluation of Counselling Quality	55
3.4	Chapter Summary	58
3.5	Research Design	59
II	Understanding	61
4	Collecting and Evaluating Naturalistic Written Data on the Language Around Health Behaviour	63
4.1	Data Collection	64
4.1.1	Annotation	65
4.1.2	Data Selection	66
4.2	Inter-Rater Reliability	68
4.3	Sentiment Analysis	69
4.4	Keyword Analysis	69
4.5	Machine Learning	71
4.6	Chapter Summary	73
5	Domain and Context Transfer Capabilities of the GLoHBCD	75
5.1	Overview of Data Sources Used and Their Different Properties	76
5.1.1	Smoking Cessation Forum	77
5.1.2	Health Coaching Dialogue Corpus	77
5.1.3	Optifast Mock Chatbot Data	77
5.1.4	Synthetic GPT-3 Data	78
5.1.5	DARN-CT-Based Wizard of Oz Dialogues	79
5.1.6	Instagram Data	80

5.1.7	Overall Comparison of Datasets	80
5.2	Transferability of Classifiers Trained on GLOHBCD to Other Contexts and Domains .	83
5.2.1	RQ 2.1: Is the Behaviour Change Language Used and the Insight Provided by the Language Consistent Across Different Conversational Contexts and Target Behaviours?	83
5.2.2	RQ 2.2: Which Client Codes Show the Most Dissimilarity in Different Conversational Contexts and Target Behaviours?	84
5.2.3	RQ 2.3: To What Extent Do Conversational Context and Domain Restrict Classification Transferability?	85
5.2.4	RQ 2.4: Which Lexical Features Provide the Most Insight Into User Attitudes in Written Utterances Around Behaviour Change?	87
5.3	Qualitative Insights Into Similarities and Differences Between Datasets	89
5.4	Chapter Summary	92
6	Experiments to Improve Classification Results	95
6.1	GPT-3-Generated User Simulation	95
6.1.1	Classification Results	97
6.1.1.1	Valence	97
6.1.1.2	Topic	97
6.1.1.3	Reason Type	98
6.1.2	Exploration of Quality and Suitability of Synthetic Data	99
6.1.2.1	Comparison of Naturalistic and Synthetic Data	99
6.1.2.2	Semantic Similarity	100
6.1.2.3	Classification of Synthetic Data	100
6.2	Semi-Supervised Learning to Facilitate Data Increase From Naturalistic Sources	101
6.2.1	Sources of Data for Pseudo-Labeling	102
6.2.1.1	Weight Loss Forum Data	102
6.2.1.2	Smoking Cessation Forum	102
6.2.1.3	AnnoMI	102
6.2.2	Methods	103
6.2.2.1	Stage 1: Confidence Thresholds	103
6.2.2.2	Stage 2: Accounting for Class Imbalance	104
6.2.2.3	Stage 3: Amount and Domain of Pseudo-Labelled Data	104
6.2.2.4	Stage 4: Application to Test Sets	104
6.2.3	Results	104
6.3	Chapter Summary	107
III	Reacting	109
7	Examining Efficacy, Safety, and the Role of User Behaviour	111
7.1	Creating a Dialogue-Framework to React to User Utterances Based on MI Principles .	112
7.1.1	Phase 1: Engaging and Focusing (4 Turns)	113

7.1.2	Phase 2: Evoking (5 Turns)	114
7.1.2.1	Process and Parameters	115
7.1.3	Phase 3: Conclusion (3 Turns)	116
7.2	Identifying Unsuitable and Harmful Outputs	116
7.3	Study Design	117
7.3.1	Measures	119
7.3.2	Hypotheses and Statistical Analysis	120
7.3.3	Participants	121
7.4	Results	122
7.4.1	Descriptive Analysis	122
7.4.1.1	Chats	122
7.4.1.2	User Perceptions	124
7.4.2	Main Analyses	125
7.4.2.1	MI-Adaptation Does not Increase the Difference Between Readiness to Change Post- and Pre-Conversation	125
7.4.2.2	MI-Adaptation Reduces the Share of Harmful and Undesirable Outputs	126
7.4.2.3	MI-Adaptation Does Not Increase User Perception of MI	127
7.4.2.4	MI-Adaptation Does Not Increase Therapeutic Alliance	127
7.4.3	Qualitative and Exploratory Analysis of User Behaviour	128
7.4.3.1	Behavioural Dimensions	128
7.4.3.2	Effects of User Behaviour on Interaction Outcomes	130
7.4.3.3	Qualitative Insights Into User Expectations	131
7.5	Exploratory Analysis of User Language	135
7.5.1	User Language Over the Course of Interactions	135
7.5.2	Impact of MI-Adherence and Inappropriate Bot Utterances on User Language .	137
7.5.3	User Reactions to Different Bot Utterance Types	138
7.5.4	Impact of User Characteristics on User Language	139
7.5.4.1	Readiness to Change	140
7.5.4.2	User Behavioural Dimensions	140
7.6	Chapter Summary	141
8	Discussion	147
8.1	Research Goal 1	148
8.1.1	Understanding User Attitudes About Change	148
8.1.2	Providing Suitable and Safe System Reactions	148
8.2	Research Goal 2	149
8.3	Theoretical Lenses	150
8.3.1	Information Science	150
8.3.2	Human-Computer Interaction	151
8.3.3	Behavioural Science and Motivational Interviewing: Similarities and Differences Between Collected Interactions and Real Counselling Sessions	152
8.4	Design Implications for LLM-Based Conversational Behaviour Change Support	154
8.5	Limitations	155
8.6	Open Questions and Future Work	156

9 Conclusion	159
References	161

List of Figures

2.1	The five stages of change according to the Transtheoretical Model	13
2.2	FBM	18
2.3	COM-B	18
2.4	Application of MI in different stages of change	23
2.5	Conceptual overview of social influence dialogue systems	25
2.6	Example of high-quality LLM-generated text	27
2.7	Example of a hallucination	29
3.1	Overview of non-LLM-based behaviour change support CAs	32
3.2	Mini-dialogue flow presented by Kocielnik, Xiao, Avrahami, and Hsieh (2018)	33
3.3	Example of difference between fine-grained and coarse-grained <i>MISC</i> annotation. Figure published under CC BY 4.0 license in Sun et al. (2024, p. 5609).	42
4.1	Distribution of posts per user in different subforums.	65
4.2	Confusion matrix of undersampled BERT classifications	72
4.3	Comparison of Cohen's κ between human raters and classifiers.	73
5.1	Classification performance of classifiers trained on GLoHBCD on test sets in 10-fold cross validation.	84
5.2	Classification results based on dataset properties	86
5.3	Classification results across all datasets when removing different lexical features	87
5.4	Classification results on the label level for each dataset when removing different lexical features	88
6.1	Classification results when applying synthetically enhanced classifiers to out-of-context test sets.	98
6.2	Amount of pseudo-labelled data points added to the training data across conditions by class.	105
6.3	Change in classification performance when more pseudo-labelled data is added to training.	106
6.4	Comparison of classification results on multiple test sets.	107
7.1	Visual overview of the flow of conversations with MiCha	114
7.2	Example of NLU and NLG module usage.	115
7.3	Adapted and translated Contemplation Ladder, based on Coolidge et al. (2011a).	120
7.4	Boxplots of presence and strength of empathy dimensions in chats per condition.	123

7.5	Overall percentage of empathetic intents per condition.	124
7.6	Correlation analysis of pre- and post-test measures and means (standard deviations) per condition	125
7.7	Boxplots for results of dependent measures per condition.	125
7.8	Changes in readiness to change depending on stage of change before conversation start.	126
7.9	Boxplot of participants' perceptions of MI-inadherent behaviour per condition.	127
7.10	Bar chart showing the number of reflective and unreflective participants and the change in their readiness to change after interacting with the CA.	130
7.11	Progression of overall valence counts in interactions with MIcha by study condition. . .	136
7.12	Progression of overall topic counts in interactions with MIcha by study condition. . .	137
7.13	Expected and observed frequencies of user utterance classifications following a given bot behaviour.	139
7.14	Classification counts per chat by binary stages of change.	140
7.15	Classification counts per chat by behavioural dimensions.	141
8.1	Coarse-grained code distributions in BiMISC and AnnoMI	153
8.2	Fine-grained client code distribution in BiMISC	153

List of Tables

2.1	Experiential and behavioural processes of behaviour change	14
2.2	O/C matrix	19
2.3	Example for semantic grammar rules used in a rule-based conversational agent	26
3.1	Simplified version of the <i>MISC</i>	42
3.2	Interpretation of κ scores	43
4.1	Detailed annotation scheme adapted from the <i>MISC</i>	66
4.2	Code distribution and examples from the GLoHBCD	67
4.3	<i>Cohen's</i> κ for two raters on 146 GLoHBCD samples.	68
4.4	Expected and observed sentence distributions of sentiment and valence.	69
4.5	Top 10 significant log-odds keywords for each category in the GLoHBCD	70
4.6	10-fold cross-validation and test set performances	71
5.1	Example questions posed to GPT-3 to generate the Synthetic GPT-3 Data	79
5.2	<i>Cohen's</i> κ for Instagram Data for two raters and 220 samples.	80
5.3	Label distribution of out-of-domain and out-of-context corpora compared to GLoHBCD	81
5.4	Overview of annotated datasets to expand the GLoHBCD.	82
5.5	Information density metrics of datasets compared to GLoHBCD.	82
5.6	Class-wise F1 scores (%) for datasets and label levels	84
5.7	Number of keywords of each dataset compared to all other datasets and top ten keywords per dataset.	90
5.8	Example sentences of GLoHBCD.	91
5.9	Example sentences of Smoking Cessation Forum and Instagram Data.	92
5.10	Example sentences of Health Coaching Dialogue Corpus, Optifast Mock Chatbot Data, and Wizard of Oz Dialogues	93
6.1	Overview of training datasets based on GLoHBCD and synthetic data.	97
6.2	Comparison of GLoHBCD and synthetic data.	99
6.3	Mean cosine similarity between all sentence embeddings of synthetic and user training sets within and between labels and datasets	100
6.4	Mean Macro F1 score (standard deviation) in percent for synthetic, mixed and mixed predicted test sets.	101
6.5	Comparison of SSL classification results	105

7.1	Example template questions.	113
7.2	MI-inadherent counsellor behaviours	117
7.3	Inter-rater reliability of annotations of undesirable bot behaviours after one round of training.	117
7.4	Finalised annotation scheme for unsuitable system outputs.	118
7.5	Number of chats per target behaviour and condition.	121
7.6	Average (standard deviation) of positive and negative user ratings of bot turns per chat.	127
7.7	Inter-rater reliability of annotations of user behaviours after one round of training.	129
7.8	Share of reflective, cooperative, and preinformed participants per study condition.	129
7.9	Mean count (standard deviation) of utterances per classified label and condition.	136
7.10	Mann-Whitney U statistics for tests of significance of differences in utterance counts within behavioural dimensions.	141
7.11	Example for uncooperative, unreflective, and non-preinformed user behaviour	143
7.12	Example for cooperative, unreflective, and non-preinformed user behaviour	144
7.13	Example for cooperative, reflective, and non-preinformed user behaviour	145
7.14	Example for cooperative, reflective, and preinformed user behaviour	146

Acronyms and Definitions

BCSS behaviour change support system

CA conversational agent

CBT Cognitive Behavioural Theory

DARN-CAT desire, ability, reason, need, commitment, activation, and taking steps

GPT Generative Pre-trained Transformer

ICC Intraclass Correlation Coefficient

IRR inter-rater reliability

LLM large language model

LSTM long short-term memory network

MI Motivational Interviewing

MINT Motivational Interviewing Training Network

MISC Motivational Interviewing Skill Code

NLP natural language processing

OARS open questions, affirmations, reflections, and summaries

RCT randomised controlled trial

RNN recurrent neural network

Seq2Seq sequence-to-sequence

SSL semi-supervised learning

TTM Transtheoretical Model

TTR type-token-ratio

WoZ Wizard of Oz

Change Talk utterances in favour of behaviour change

Sustain Talk utterances in favour of the status quo

Target Behaviour the behaviour to be implemented

1

Introduction

1.1 Motivation

In behavioural science, behaviour can be defined as “anything a person does in response to internal or external events. Actions may be overt (motor or verbal) and directly measurable or covert (activities not viewable but involving voluntary muscles) and indirectly measurable; behaviours are physical events that occur in the body and are controlled by the brain” (R. Davis, Campbell, Hildon, Hobbs, & Michie, 2015, p. 327). Although the exact definition of behaviour and related concepts, such as action and practice, is difficult and much discussed¹, the generic nature of the concept of behaviour allows its usage as an umbrella term that includes actions and activities as “elements of behaviour”, and practice as “a mode of behaviour” (Savolainen & Wilson, 2009, Tom’s first response to Reijo). Much of the behaviour exhibited by humans, such as poor dietary choices, lack of physical activity, procrastination in important tasks or smoking, is harmful or at least not beneficial to our health and quality of life. When we “modify the behaviour of a problem or acquire a positive behaviour” (Velicer, Prochaska, Fava, Norman, & Redding, 1998, p.1), this can be termed behaviour change.

Such changes are often difficult to initiate and even harder to maintain, as various psychological, environmental, and social factors influence our behaviour, leading us to regularly face barriers to change (Michie, van Stralen, & West, 2011). I can be very motivated to change my behaviour and begin a daily exercise routine one day, but varying energy levels on different days, long days of work, or social activities and responsibilities can easily get in the way. The chances of success decrease further when there is initial ambiguity about the change (W. R. Miller & Rollnick, 2002). In addition, much of our behaviour is habitualised: we perform it or fail to do so repeatedly without having to specifically think about it or remember it (Lindbladh & Lyttkens, 2002).

To address these difficulties, behavioural science defines various behaviour change techniques that can be applied by counsellors and other supporters to enable and facilitate change, ranging from informing about behavioural risks over general encouragement to prompting an individual to set specific goals (Abraham & Michie, 2008). One of these techniques is Motivational Interviewing (MI), a counselling

¹see Savolainen and Wilson (2009) for an example of such a discussion.

style that helps people reflect on their behaviour and goals and aims to elicit *change talk*, utterances in favour of change, and decrease *sustain talk*, utterances against change, in favour of the status quo (W. R. Miller & Rollnick, 2002).

Although MI has shown great potential in motivating and facilitating behaviour change in clients in various change contexts, medical professionals frequently lack adequate training or are reluctant to engage in discussions about behaviour change and motivational aspects when addressing health problems with clients for various reasons, including lack of time and a belief that increasing motivation for change is not within their scope of responsibility (Chisholm, Hart, Lam, & Peters, 2012; Hunter et al., 2015). At the same time, counselling is hard to access, not necessarily relevant or suitable for every behaviour change (there is probably no need for most otherwise physically and mentally healthy people to seek counselling in order to form a consistent exercise habit), and not universally available at the exact moments when barriers are faced.

behaviour change support systems (BCSSs) are digital interventions that use technology, including “computer-human dialog support” (Oinas-Kukkonen, 2013, p. 1228), to support behaviour change through the modelling of behaviour change techniques (Oinas-Kukkonen, 2013). In fact, it is almost impossible to talk about the history of conversational AI without referencing counselling chatbots. ELIZA, the first conversational agent (CA) to exist, was fashioned after a Rogerian psychotherapist. According to its creator Joseph Weizenbaum, this was because psychotherapists are one of a few groups that are “free to assume the pose of knowing almost nothing of the real world” (Weizenbaum, 1966, p42). It did not take long for people to campaign for the real-life benefits of the system for therapy purposes (O'Dell & Dickson, 1984). Despite its simplicity, ELIZA was able to converse with humans convincingly and often caught their attention for long periods of time, eliciting personal information (O'Dell & Dickson, 1984). Since then, the field has advanced considerably and conversational AI in the form of chit-chat applications has evolved to engage users to the extent that deep relationships are formed and long-term conversations are maintained, exemplified by chatbots such as Replika (Skjuve, Følstad, Fostervold, & Brandtzaeg, 2021) and Microsoft's Xiaolce (L. Zhou, Gao, Li, & Shum, 2020). Analyses of these systems showed that users frequently interact with them for multiple weeks, perceive them as friends or companions, and even report positive impacts on their well-being due to interactions with the chatbots (Skjuve et al., 2021; L. Zhou et al., 2020). These findings highlight the apparent gravitational pull of therapeutic use cases for conversational AI.

Incorporating the conversational style of MI into CAs for behaviour change could have multiple benefits:

- First, such systems are constantly available, which gives users the chance to turn to the system whenever they are facing barriers or problems. Not only are they available at all times, but they are also available to everyone, eliminating the need to triage those who need help more urgently. Consequently, such systems can serve as a first point of contact when counselling is 1) not possible, for instance, due to a lack of available resources or staff, or 2) not necessary, as would be the case in non-clinical behaviour change contexts. In the context of this thesis, I am focusing on the second case. However, in Chapter 3, I will outline related work which aims to alleviate the scarcity of counselling in related domains such as mental health.
- Second, although computers and especially CAs are social actors and tend to get anthropomorphised (Nass, Steuer, & Tauber, 1994), users often perceive them as non-judgemental (Olafsson,

O’Leary, & Bickmore, 2019; Skjuve et al., 2021) and less intrusive and autonomy-threatening than human conversational partners (J. Chen, Chen, Walther, & Sundar, 2021). This has been shown to help users self-disclose about difficult topics (Skjuve et al., 2021). Self-disclosure has various psychological benefits, and it has been shown that these benefits persist regardless of whether a person thinks they are conversing with another human or with a bot (Ho, Hancock, & Miner, 2018). Some research findings even indicate that informational and social needs voiced in anonymous online settings might be distinctly different from those voiced in personal interactions with peers or counsellors (Hasler, Ruthven, & Buchanan, 2014).

- Interaction with such systems, especially when the system adheres to MI principles and emphasises user reflection and self-disclosure, can be regarded as a form of guided journaling. Past research has shown that journaling alone can foster self-reflection, improve positive outlooks in the face of adverse situations (Ullrich & Lutgendorf, 2002) and increase self-efficacy and intrinsic motivation (Rafizadeh, Morewitz, & Mukherjea, 2021). Others have found that journaling has the potential to decrease procrastination, a habit many students struggle with (Hensley & Munn, 2020), and increase engagement in structured behaviour change programmes for various target behaviours (the behaviours to be implemented) (Stephenson & Zygouris, 2007). Guided journaling, especially based on positive prompts has also been found to have higher positive effects on well-being than self-reflection alone, which can easily lead to rumination (MacIsaac, Mushquash, & Wekerle, 2023).

Past attempts at automating MI in this way have often relied on set structures in which specific user statements lead to specific system replies (rule-based) or algorithms which select a reply from a pool of pre-existing candidates, based on various similarity metrics (retrieval-based) (Luo, Aguilera, Lyles, & Figueroa, 2021; Oh, Zhang, Fang, & Fukuoka, 2021; Xu & Zhuang, 2022). As a result, these systems are typically constrained to one specific target behaviour, and in many cases do not allow unconstrained user input, in many cases resorting to multiple choice fields (e.g. Nurmi et al. (2020); Olafsson et al. (2019)). In addition, as will be shown in Chapter 3, existing systems largely fail to take into account and adequately react to user’s current thoughts about change. While some of them employ emotion recognition to gain insight into users’ states of mind, users’ attitudes towards change are largely disregarded.

Recent advances in natural language processing (NLP) have led to the conception of large language models (LLMs), models trained on massive amounts of text, with the ability not only to produce high-quality text in a wide array of contexts but also convincingly imitate human-level understanding of user messages (T. B. Brown et al., 2020; OpenAI, 2023a). These models have the potential to drastically increase the flexibility and capabilities of MI-guided CAs. Due to their generalisation capabilities, they could enable these agents to be responsive to various target behaviours and dynamically react to users’ current situations and attitudes toward change. However, they also suffer from a lack of controllability, often producing undesired or unprompted outputs, untruths (referred to as hallucinations) or reproducing biases learnt from the training data (such as gender biases) (Bender, Gebru, McMillan-Major, & Shmitchell, 2021; Weidinger et al., 2021). As a result, most of the work focused on the application of LLMs to mental health or behaviour change has evaluated their capabilities based on simulations, often based on isolated turns or attempts to produce isolated behaviours (Weidinger et al., 2023). This bypasses user behaviour, which, considering the higher capabilities of LLM-based CAs

compared to rule- or retrieval-based alternatives, is likely to differ from both interaction with those traditional dialogue systems and interaction with other humans.

1.2 Research Goals

This thesis was driven by two main research goals:

Research Goal 1: To explore how and to what extent MI can be integrated into the conversation flow of LLM-based behaviour change support systems in order to increase understanding of user attitudes about change and provide more suitable and safe system reactions.

Research Goal 2: To shed light on users' expectations for and interaction styles with LLM-based behaviour change support systems and in what way these a) *are influenced by* their attitudes about and readiness for change, and b) *have an influence on* the outcome of such interactions.

1.3 Outline

In order to address these goals, the first part of this project was dedicated to the collection and annotation of naturalistic data from individuals who wrote about behaviour change in conversational contexts and the creation of computational methods to predict user attitudes. In the second part, MI theory was paired with LLM research to design a conversational framework to guide a CA's actions during interactions with users, as well as an evaluation framework to determine to what extent agents' utterances are in line with desirable behaviours in LLM-based behaviour change support. The project was concluded with a user study to explore how users interact with LLM-based CAs in the context of behaviour change and to what extent using information about user attitudes to guide the agent's reactions has an effect on their motivation to change.

This thesis is organised into three main parts: Part I, *Grounding*, establishes the theoretical background (Chapter 2) and contextualises the project within the related work (Chapter 3), Part II, *Understanding* is dedicated to all experiments conducted with the aim to understand user attitudes about change in written conversational contexts (Chapters 4 to 6), whereas Part III, *Reacting*, describes the user study, including the conversational framework drawing on results from Part II, as well as experiments towards evaluating the appropriateness of LLM outputs. This also encompasses analyses of the effects of LLM-based behaviour change support systems on user perception and readiness to change and user behaviour and expectations when interacting with such systems (Chapter 7). Finally, Chapter 8 discusses all the outcomes, results, and limitations of the thesis in light of the related work and their implications for future work. All code used for this thesis, sorted by chapters, is available on GitHub ².

1.4 Publications Relating to This Thesis

This thesis, including the ideas, data, and results discussed therein, builds on existing scientific publications and includes reprinted sections from these articles. Here, I declare my and my coauthors'

²https://github.com/SelinaMeyer/MI_LLM_in_Behaviour_Change_Supplementary-Material

contributions for each publication or unpublished manuscript using the CRediT taxonomy³. Publications in which I participated but which are not directly relevant to this thesis are excluded from this list.

- GLoHBCD: A Naturalistic German Dataset for Language of Health Behaviour Change on Online Support Forums (Meyer & Elsweiler, 2022).

Selina Meyer: Conceptualisation, Methodology, Data curation, Formal analysis, Investigation, Software, Validation, Visualisation, Writing - original draft, Writing - review & editing

David Elsweiler: Supervision, Writing - review & editing

- Do we Still Need Human Assessors? Prompt-Based GPT-3 User Simulation in Conversational AI (Meyer, Elsweiler, Ludwig, Fernández-Pichel, & Losada, 2022).

Selina Meyer: Conceptualisation, Methodology, Data curation, Formal analysis, Investigation, Software, Validation, Visualisation, Writing - original draft, Writing - review & editing

David Elsweiler: Supervision, Writing - review & editing

Bernd Ludwig: Conceptualisation, Writing - review & editing

Marcos Fernández-Pichel: Conceptualisation, Writing - review & editing

David Losada: Conceptualisation, Writing - review & editing

- Towards Cross-Content Conversational Agents for Behaviour Change: Investigating Domain Independence and the Role of Lexical Features in Written Language Around Change (Meyer & Elsweiler, 2023).

Selina Meyer: Conceptualisation, Methodology, Data curation, Formal analysis, Investigation, Software, Validation, Visualisation, Writing - original draft, Writing - review & editing

David Elsweiler: Supervision, Writing - review & editing

- Exploring the Challenges of Behaviour Change Language Classification: A Study on Semi-Supervised Learning and the Impact of Pseudo-Labelled Data (Meyer, Fernández-Pichel, Elsweiler, & Losada, 2024).

Selina Meyer: Conceptualisation, Methodology, Data curation, Formal analysis, Investigation, Software, Validation, Visualisation, Writing - original draft, Writing - review & editing

Marcos Fernández-Pichel: Conceptualisation, Writing - review & editing

David Elsweiler: Writing - review & editing

David Losada: Supervision, Writing - review & editing

- “You tell me”: A Dataset of GPT-4-Based Behaviour Change Support Conversations (Meyer & Elsweiler, 2024).

Selina Meyer: Conceptualisation, Methodology, Data curation, Formal analysis, Investigation, Software, Validation, Visualisation, Writing - original draft, Writing - review & editing

David Elsweiler: Supervision, Writing - review & editing

- LLM-based Conversational Agents for Behaviour Change Support: A Randomised Controlled Trial Examining Efficacy, Safety, and the Role of User Behaviour [manuscript under review].

Selina Meyer: Conceptualisation, Methodology, Data curation, Formal analysis, Investigation,

³<https://credit.niso.org>.

Software, Validation, Visualisation, Writing - original draft, Writing - review & editing

David Elweiler: Supervision, Writing - review & editing

PART I

GROUNDING

Theoretical Background

This chapter provides the theoretical basis for this thesis project. I begin by describing how behaviour change is defined and conceptualised in different disciplines which have been influential for this project in Section 2.1. The section includes an overview of the various factors that influence behaviour change and how it progresses. This is followed by a discussion of the efficacy and principles of MI in Section 2.2. Finally, in Section 2.3, I provide a short history of CAs, highlighting how recent developments have changed the landscape of conversational AI.

2.1 Behaviour Change

Behaviour change has been defined and studied across various academic disciplines, from the behavioural sciences, to information science and human-computer interaction. This section gives an overview of theoretical models and empirical studies on behaviour change in those disciplines. Each discipline has its unique point of view on behaviour and change and delivers a new dimension with respect to the design of the studies and experiments conducted in this thesis and the interpretation of their results.

2.1.1 Behaviour and Change in Information Science

An integral part of information science is *information behaviour*. This term most often refers to how individuals “need, seek, manage, give and use information” (Pettigrew, Fidel, & Bruce, 2001, p 44), commonly in online contexts. This area of research also includes studying *how behaviour changes*: Areas of interest include modelling users’ thoughts, feelings, and actions throughout the search process (Kuhlthau, 1988), the social and psychological factors that play a role in individuals’ preferences and use patterns of information systems (Savolainen, 1995), identification of different user types when it comes to interaction with information (S. M. Miller, 1980) and individuals’ reactions to life changes, often referred to as “transitions” (Ruthven, 2022b). Empirical studies often focus on changing how people interact with or share information online, improving search behaviour, or encouraging reflection on how documents are perceived (e.g. Bink, Zimmerman, and Elswailer (2022); Htait, Azzopardi, Nicol, and Moncur (2020); Nicol et al. (2022); Ortloff, Zimmerman, Elswailer, and Henze (2021)).

This can lead to changes in physical behaviour, such as food and recipe choices (Elsweiler, Trattner, & Harvey, 2017) and sustainable behaviour (Byerly et al., 2018). As such, a main focus of information science research in the context of behaviour change lies in identifying the impact of different strategies of *retrieving, displaying, and providing information* on user behaviour. In the context of information science, behaviour change is thus often treated as something that happens involuntarily, often even unnoticed by users, but can be induced and empirically observed by researchers through the provision of the right information, at the right time, in the right way. Alternatively, behaviour change can be seen as a conscious or unconscious process following a life transition, both in (e.g. getting married) and out of an individual's control (e.g. death in the family, realising that you are gay), where information systems have the potential to support transitions and help users understand the changes that occur (Ruthven, 2019, 2022a, 2022b).

The focus of this work is different from most information science research in two essential ways:

1. I focus on *intentional* behaviour change - changes a user wants to engage in voluntarily. This requires users to have a clear goal in mind about their desired behaviour change and contrasts with techniques employed in information science, such as nudging, to direct users towards researcher-defined behaviour patterns (e.g. Byerly et al. (2018); Elsweiler et al. (2017)).
2. I focus on *increasing user self-reflection* as opposed to *providing* information. Thus, an assumption of this work is that users interacting with the CA have other means of accessing relevant information about how they can change their behaviour, but lack the *motivational resources and confidence* to make the desired changes.

Although their main focus is different, some of the theoretical insights of information science show similarities with theoretical models from other disciplines, such as the behavioural sciences and human-computer interaction. Furthermore, I posit that parallels can be drawn between the ways individuals interact with factual information provided to them and the way they interact with their own minds in the form of attitudes, intentions, and self-reflection regarding behaviour change. For example, Kuhlthau's (1988) model of information search behaviour defines the search process in six stages, each of which is accompanied by different feelings, thoughts, and actions. According to the model, uncertainty and confusion regarding the search task and the information available lead to ambiguous thoughts towards the early stages, whereas later stages are increasingly marked by more specific thoughts as searchers gain clarity and focus the scope of their information seeking. Finally, the outcome of each search task is judged as successful or not, leading to satisfaction or discontent. As Section 2.1.2 shows, similar processes and stages can be observed during behaviour change, with ambiguity present if an individual is unsure about their ability to perform a behaviour change but gains increasing confidence as time progresses. This is related to the psychological concept of *self-efficacy* (Bandura, 1977), the perceived ability to change behaviour, which is also highly relevant in the context of information seeking (Savolainen, 2002; Wilson, 1997).

Similarly, Savolainen's (1995) model of everyday life information seeking combines the social and psychological factors that play a role in people's preferences and use patterns of information systems. The model defines the concepts of *way of life*, based on an individual's amount of available leisure time, consumption models, and *mastery of life*, referring to the way individuals achieve the maintenance of their way of life, specifically, in what way they engage in problem-solving if their order of things is threatened. Savolainen argues that through "their choices individuals have practically engaged in

a certain order of things, and it is in their interest to keep that as long as they find it meaningful" and that "Information seeking is an integral component of mastery of life, which aim is to eliminate a continual dissonance between perceptions of »how things are at this moment« and »how they should be.«" (K. E. Fisher, Erdelez, & McKechnie, 2005, p. 144). In Section 2.1.2, I summarise psychological literature that argues that precisely this dissonance between how things are and how they should be is a main driver of behaviour change.

S. M. Miller's (1980) concept of *blunters*, people who prefer to avoid information when put into aversive situations and *monitors*, those who prefer to be knowledgeable about their situation can also be related to behaviour change, as shown in Hirvonen, Huotari, Niemelä, and Korpelainen (2012), who explore the information behaviour of young men in different stages of exercise behaviour change. They surveyed over 600 participants on their physical activity levels, information behaviour, and various behaviour change-related factors, including self-efficacy and stage of change. They find that individuals increasingly encounter information needs and engage in information seeking as they progress towards higher stages of change. While people most often encounter information through passive acts such as non-directed monitoring, in early phases of change, individuals who are more advanced in the process actively seek information and people who are maintaining an existing exercise routine mostly employ active scanning. However, they also find that the majority rarely felt interested in exercise-related information and seldom encountered information needs. Almost a quarter of respondents also reported sometimes or regularly avoiding information (blunting). I posit that this concept can also be related to how people in different situations and with different change intentions interact with information and realisations about *themselves*: the harms of their current behaviour, their own fears, and desires. Whereas individuals who are in the early stages of change might avoid thinking about the problem behaviour and its impact on their lives, individuals attempting and maintaining change might be more open to such reflections.

The blunting of information has also been revealed in other health-related contexts such as the information behaviour of young mothers from deprived backgrounds. Buchanan and Nicol (2019a, 2019b) find that this group often felt overwhelmed, leading them to actively avoid relevant information. They explore the role of family nurses as information intermediaries in this context. They come to the conclusion that, while nurses were familiar with motivational techniques which are effective in increasing clients' interest in a topic, they were unfamiliar with information literacy concepts, the application of which could improve the potential for clients to adopt independent information-seeking practices.

Ruthven furthermore introduces two concepts related to change and information behaviour to the information science literature. Meaning-making (Ruthven (2019), based on Dransart (2013)) describes the search for information that helps us to make sense of our lives, specifically focusing on the effort to understand the "existential significance" (Ruthven, 2019, p. 163) of information instead of simply comprehending it. The concept of information shaping describes how we use information over time. Ruthven (2024) proposes a transition from broad and unspecific to specific and targeted as we gain increasing knowledge, mediated through experimentation, iteration, local optimisation and perceptual feedback loops, and guided through social learning and evolving cumulative information practices. According to Ruthven (2019), meaning-making can be viewed as a "process that happens in *reaction* to an event, such as redundancy or illness" or "as occurring due to a gradual *awareness* that a current existence is not meaningful anymore and a new way of living may be required" (p. 164). Such dissonance

can lead to a broad (passive) search for information within social networks and chance encounters, followed by a more strategic search of specific sources and social contacts, accompanied by the learning of social norms and practices of the new environment, resulting in a maintenance stage, which necessitates reinforcement and support. Tools used for meaning-making include search engines, internet forums, and blogging/journaling. In addition to factual information, other people (both professionals and communities) and narratives (e.g. leading by example) can also support meaning-making. In this area, “filling cognitive gaps is not sufficient; information has to speak to the whole person, including their desires, beliefs, and motivations” (Ruthven, 2019, p. 168). Ruthven also highlights the importance of providing support at the right stage of meaning-making, defining this as a current research gap. Again, several parallels to behaviour change models from behavioural science emerge (see Section 2.1.2), including the dissonance between the status quo and a desired state as a driver for taking action, the necessity of addressing more than just cognitive gaps, and the framing of behaviour as a non-straightforward process with various stages, levels of specificity, and iterations.

Finally, Ruthven’s (2022b) information behaviour theory of transitions describes three stages of psychological processes involved in life transitions, both intentionally brought over and imposed by external or internal circumstances outside our control. Although related to intentional behaviour change to a certain degree, the focus again is on the information behaviour exhibited and the support required by individuals throughout such life transitions. According to the theory, which is a synthesis of various other transition theories, three transitional stages are preceded by an event, which can be an illness, or a loss, but also “a new awareness of our lives (e.g., realisation of personal dissatisfaction)” (Ruthven, 2022b, p. 585) leading to dissonance. The first stage, understanding, is the stage in which the need for change becomes known and time is spent understanding the implications of the new situation. Individuals put effort into determining the causes of their situation, interpreting their feelings, or seeking comparisons, among other behaviours. This is followed by the negotiation stage, in which options are weighed and alternative pathways are explored through learning and testing. The final stage, Resolving, then consists of the search for meaning, re-imagining one’s position, and deciding on a course of action.

The focus of information science clearly lies in how individuals engage with information in different situations and contexts. Although this can in part explain how people behave when change is encountered or when a need for behaviour change arises, the concepts describing behaviour change in the context of information science do not put their main emphasis on the reasons why behaviour change is made necessary, what drives people to initiate behaviour change, and what psychological processes facilitate intentional behaviour change. These aspects are elucidated by shifting the viewpoint to behavioural science.

2.1.2 Behavioural Science Models of Behaviour Change

There are various factors that influence how successful a person is when attempting to change their behaviour. At the highest level, a distinction can be made between external factors (that is, social norms and formal regulations, relationships) and internal/psychological factors (Grilli & Curtis, 2021; Kok et al., 2015; Kwasnicka, Dombrowski, White, & Sniehotta, 2016; Michie et al., 2011; Wallen & Daut, 2018). Environmental and social factors play an important role in people’s abilities and motivations to change and maintain change, and behaviour change interventions are often recommended to address

such factors as well as internal ones (Grilli & Curtis, 2021; Kwasnicka et al., 2016; Michie et al., 2011; Middleton, Anton, & Perri, 2013; Murray et al., 2013). However, external factors operate more on a global level and are likely to necessitate changes in policy or a person's (social) environment, which is beyond the scope of this dissertation project, which focuses on facilitating behaviour change on a personal level. For this reason, the behaviour change models that I describe in this section put emphasis mainly on internal psychological factors.

2.1.2.1 Change as Process: The Transtheoretical Model

Behaviour change does not happen at one singular point in time but is better thought of as a process. Depending on where a person is on their change journey, different factors may play a role in facilitating change. The Transtheoretical Model (TTM) depicts the process of behaviour change and its milestones (Prochaska & Clemente, 1983; Prochaska & Velicer, 1997) and is the most popular theory of behaviour change in the social and behavioural sciences (R. Davis et al., 2015).

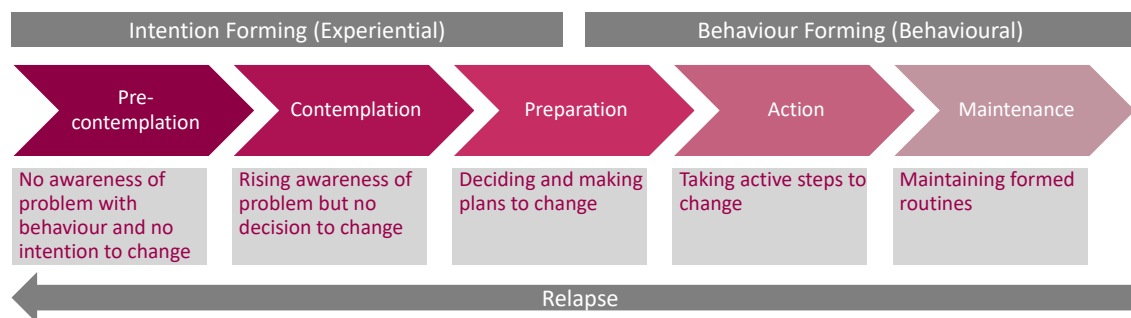


Figure 2.1: The five stages of change according to the TTM based on Prochaska and Clemente (1983); Prochaska and Velicer (1997); Velicer et al. (1998).

The model identifies five stages of change, where the first two encompass the formation of an intention, the last two are based on the formation of the behaviour, and the middle stage is a combination of both (Velicer et al., 1998) (see Figure 2.1). According to the TTM, behaviour change begins with a transition from a *precontemplation* to the *contemplation* stage. During *precontemplation*, a person is not aware of a problem with their behaviour, whereas the *contemplation* stage sees first disadvantages of current behaviour and potential advantages of changing, which slowly begin to outweigh the advantages of the status quo (or the disadvantages of the status quo begin to outweigh the disadvantages of changing). This is then followed by a *preparation* stage, the first stage to include a behavioural component. The individual begins to gain confidence about their ability to change and makes the first preparations. Finally, the *action* phase sees the individual initiating change. The last stage, *maintenance*, is occupied by efforts to keep the newly formed habits alive. Stages can be highly variable in length and do not always have to follow one another chronologically. At each point, relapse can occur, leading the individual to start anew or to return to an earlier stage. These stages have been continuously associated with behavioural outcomes, with individuals in later stages showing higher success rates (Krebs, Norcross, Nicholson, & Prochaska, 2018).

This model allows us to acknowledge and conceptualise the process of change even before it is initiated through actions. It also provides an overview of ten experiential and behavioural processes that mediate change and play different roles throughout the stages of change (Petrocelli, 2002; Prochaska, Velicer, Guadagnoli, Rossi, & DiClemente, 1991; Velicer et al., 1998) (see Table 2.1). Whereas the intention-forming stages of precontemplation, contemplation and preparation are associated with experiential processes, such as consciousness raising and self re-evaluation, the behavioural processes, such as stimulus control, self-liberation, and counter conditioning are associated with the action and maintenance stages (Velicer et al., 1998). Where the experiential processes are important for change

	Process	Definition
Experiential	Consciousness Raising	Increasing awareness of a problem behaviour, its causes, consequences and cures.
	Dramatic Relief	Increasing emotional experiences with respect to problem behaviour.
	Environmental Reevaluation	Reconsideration of the problem behaviour's impact on the environment.
	Social Liberation	Recognition of increased value and availability of more favourable behaviours in society.
	Self-Reevaluation	Conscious mental exploration of the self and the problem.
Behavioural	Counterconditioning	Seeking and testing alternative behaviours.
	Stimulus Control	Aiming to avoid stimuli associated with the problem behaviour.
	Reinforcement Management	Instituting rewards for steps towards behaviour change.
	Self-Liberation	Examining the potential for positive outcomes and the necessary changes in terms of ability and commitment.
	Helping Relationships	Social support in favour of change.

Table 2.1: Experiential and behavioural processes of behaviour change based on Petrocelli (2002); Prochaska et al. (1991); Velicer et al. (1998).

to be initiated, the behavioural processes work towards the maintenance of change. In these processes, individuals habitualise newly established behaviours, learn strategies to regulate their behaviour, and overcome barriers to change, building up their psychological and physical resources.

2.1.2.2 Social Cognitive Theory, Theory of Planned Behaviour, and the Information-Motivation-Behavioural Skills Model

Factors mediating change have been defined not only within the scope of the TTM, but also across a range of compatible theories, including Social Cognitive Theory (Bandura, 1986), the Theory of Planned Behaviour (Ajzen, 2002), and the Information-Motivation-Behavioural Skills Model (J. D. Fisher & Fisher, 1992). Although the vocabulary used in these theories differs at times, they show significant overlaps in the processes and change factors that they describe. In this section, instead of describing each theory separately, I focus on synthesising and summarising important mediators of change, as proposed by these theories and the empirical literature.

Self-Evaluated Dissatisfaction

The first question to be answered when discussing behaviour change is what drives a person to *initiate* behaviour change. Behaviour change is generally preceded by *reasons* for change. Generally, these reasons present in some form of a disconnect between a person's current behaviour and their goals, or personal values (W. R. Miller & Rollnick, 2002; Teixeira et al., 2010), also termed self-evaluated dissatisfaction in Social Cognitive Theory (Bandura, 1986; Dziewaltowski, 1989). This can, for instance, happen through exposure to information about the infeasibility of current behaviour (J. D. Fisher & Fisher, 1992). In some cases, this disconnect creates a *desire* to change (the person wants to change their behaviour), in others the motivation to change is based on *necessity* (the person thinks that they have to change to meet external demands or standards), tipping the decisional balance, which weighs the advantages and disadvantages of change (Velicer et al., 1998).

Motivators

Here, a distinction can be made between intrinsic and extrinsic motivators. Where desire-based reasons are intrinsic motivators for change, necessity-based reasons can be translated as extrinsic motivators (Kwasnicka et al., 2016). Although changes are often initiated based on extrinsic motivators (Kwasnicka et al., 2016), behaviour change success and maintenance of behaviour changes are more likely if the motivators are intrinsic or have become intrinsic over time (Ryan & Deci, 2020). In either case, sufficient motivation for change is a prerequisite for the change to be initiated.

There has been some indication that the intrinsic motivators that lead to the initiation and success of behaviour change vary between different target behaviours, as found by Ogden and Flanagan (2008). In this study, participants who had recently made one of three changes in their eating behaviour (reducing calories, switching to a vegetarian diet, or eliminating a specific food group) were asked about their attitudes and intentions. Both the success rate and the main influencing factors varied between the different target behaviours, with a low positive attitude toward the foods being avoided as the one common success factor for all groups. This indicates that although there are differences in intrinsic motivations between change objectives, a low positive attitude towards current behaviour is a primary driving factor for successful change in behaviour, that is, if a person *does not like their current behaviour* any more, they are more likely to be successful at making changes.

Self-Efficacy

A reason is not enough for behaviour to change. To be able to initiate behavioural changes, people must also *feel like they can achieve* change (Ajzen, 2002; Bandura, 1977, 1986). Self-efficacy, the individual's confidence in their ability to change, is also a crucial factor in the successful maintenance of change, as it influences the amount of effort a person puts into learning and executing skills (Bandura & Cervone, 1983), and increases the ability to overcome barriers (Kwasnicka et al., 2016). Once self-efficacy is established in one domain, it tends to generalise to other situations, facilitating behaviour change not only towards the initial target behaviour but also towards others that are distinctly different (Bandura, 1977).

Locus of Control and Perceived Autonomy

An internal locus of control, meaning that an individual feels in control of their own actions, and a high degree of perceived autonomy and control over the situation can be helpful in strengthening intentions, effort, and perseverance (Ajzen, 2002; Elfhag & Rössner, 2005; Teixeira et al., 2010). Furthermore, Williams, Grow, Freedman, Ryan, and Deci (1996) and Teixeira, Silva, Mata, Palmeira, and Markland (2012) found that in addition to the autonomy orientation of an individual themselves, successful long-term behaviour change (here in the context of weight loss) is mediated by the perceived autonomy support of their interpersonal environment, meaning that others can help strengthen an individual's perceived autonomy and thus increase the likelihood of successful change.

Outcome Expectancies

On top of a perceived ability to make a change and the feeling of being in control of this change, behaviour change is mediated by outcome expectancies, which means the expected consequences of actions (Bandura, 1986; Middleton et al., 2013). If a person believes that changing their behaviour in a certain way will have specific beneficiary consequences or that *not* changing their behaviour will have distinctly negative consequences, this is a driver of change. If an individual perceives that a change in behaviour will not affect their life, they are unlikely to initiate it (Dzewaltowski, 1989).

Ambivalence

If self-efficacy, perceived autonomy and control, reasons for change, or outcome expectancies are lacking or underdeveloped, this can cause ambivalence, conflicting feelings or attitudes towards behaviour change (Ajzen, 1985, 1987), which in turn decreases the likelihood of behaviour change attempts or success. Consequently, an essential part of behaviour change interventions is reducing ambivalence in the beginning. The Theory of Planned Behaviour (Ajzen, 2002) posits that in addition to behavioural beliefs (outcome expectancies and beliefs about the attributes of the behaviour) and control beliefs, behaviour is driven by normative beliefs, meaning a person's beliefs about normative expectations from others. Similarly, the Information-Motivation-Behavioural Skills model suggests that behaviour change results from exposure to relevant information, individual attitudes toward the change, and the behavioural skills required to implement the change (J. D. Fisher & Fisher, 1992). According to the model, behavioural skills are frequently a product of information and motivation.

Overcoming Barriers to Change

Once a behaviour change has been initiated, the next step is to maintain the newly established behaviour long-term, something that poses a significant challenge to the majority of people, with nonadherence rates as high as 80% even in treatment regimes for chronic illness (Middleton et al., 2013). Many of the psychological factors that influence the initiation of change, such as attitudes toward change, beliefs about the self and personal values, also play a role in its maintenance (Murray et al., 2013; Teixeira et al., 2012). However, there are some additional factors that play a role in the successful maintenance of change but are less important for its initiation. In their systematic review of behaviour theories,

Kwasnicka et al. (2016) identified five overarching themes from 100 behaviour theories concerned with behaviour change maintenance. Based on their results, people tend to be more successful in maintaining change if they have a supportive environment and social support, one or multiple sustained (and usually intrinsic) maintenance motives, a high amount of psychological and physical resources, and are effective at monitoring and regulating their changed behaviour and applying strategies to overcome barriers.

Some of these barriers can even come from factors related to the initiation of behaviour change. For instance, a literature review by Elfhag and Rössner (2005) indicated the importance of realistic outcome expectancies for long-term adherence to change. According to their findings, it is likely that too high expectations at the beginning of change and a failure to achieve self-determined goals can have a negative impact on self-efficacy, which in turn can decrease the likelihood of change maintenance.

Habitualisation

Behaviours are also more easily maintained if they have become habitual over time, which is facilitated if individuals have developed automatic responses to relevant cues (Kwasnicka et al., 2016). It follows that, in addition to self-efficacy and (intrinsic) motivators, maintenance of behaviour change is facilitated if the behaviour is easy to integrate into daily habit (thus easy to habitualise) and if an individual is privy to strategies to overcome barriers to change, which naturally come up whenever behaviour is changed long-term. Consequently, it can be helpful for people in the process of changing their behaviour to regularly re-evaluate and reflect change and be aware of their current situation and potential upcoming barriers to maintaining change, as well as what strategies they are equipped with to overcome them.

2.1.3 Behaviour Change in Human-Computer Interaction

Similarly to information science, in the context of human-computer interaction, the focus of behavioural models is more on their implications on the design of information systems and persuasive systems. Behavioural processes are modelled in a way that allows the contextualisation of persuasive and supportive systems in a bigger picture and facilitates the choice of strategies to be implemented.

Persuasive Design

For instance, Fogg presents his behaviour model (FBM) and behaviour grid with a clear focus on persuasive system design. The FBM attributes the performance of a behaviour to three distinct factors (Figure 2.2). Following the model, an individual must have sufficient motivation, the ability to perform the behaviour and be triggered to perform it. Only if those three factors converge in the same moment, is the behaviour executed (Fogg, 2009b). Note that in this model, ability refers both to physical capability and to self-efficacy. The FBM defines three dimensions of core motivators: pleasure/pain has the most immediate results, hope/fear is based on outcome anticipation, and social acceptance/rejection focuses on social factors. Furthermore, six simplicity factors, such as the amount of money, time, or physical effort involved, and three behavioural triggers are defined: sparks are designed to cause motivation to perform a target behaviour in unmotivated individuals, facilitators simplify the performance of a

behaviour and are best suited for highly motivated users who lack the ability to perform a behaviour, and signals serve as a reminder to perform the behaviour, presupposing high motivation and ability.

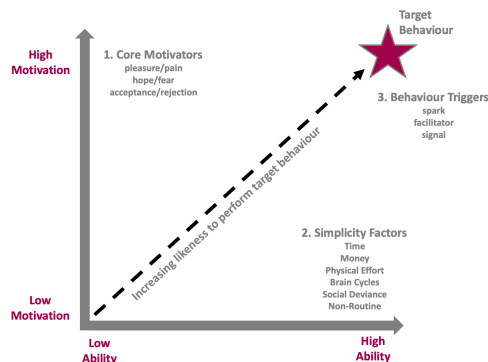


Figure 2.2: FBM, based on Fogg (2009b, p. 2).

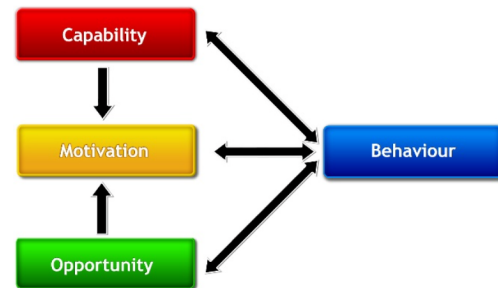


Figure 2.3: COM-B. Figure published under CC BY 2.0 license in Michie et al. (2011, p. 4).

A similar behaviour model is introduced by Michie et al. (2011). In the COM-B model, behaviour is made up of the three components of capability, which include both psychological and physical capacities, opportunity, meaning external factors which facilitate behaviour, and motivation, referring to the brain processes that direct behaviour, both habitual and conscious (Figure 2.3).

To facilitate communication around behaviour change interventions, West and Michie (2016) create an ontology of behaviour change interventions, as described in (Larsen et al., 2017). The ontology defines the interaction between an intervention and behaviour through the definition of different mechanisms, such as the delivery of the intervention and its usage and context. They define which factors are relevant for an intervention to have an effect on behaviour. This ontology serves as a framework to describe interventions in a structured and standardised way.

In addition to insights about the nature of behaviour, Fogg (2009a) also provides a taxonomy of types of behaviour change. His Behaviour Grid encompasses 35 different categories of behaviour, arranged along the two dimensions of type and schedule. The type distinguishes whether the behaviour change addresses a new (unfamiliar) or existing (familiar) behaviour and whether a behaviour should be increased, decreased, or stopped. Schedule distinguishes between one-time behaviours, one-time behaviours that lead to ongoing commitment, behaviours with a fixed duration, behaviours on a predictable schedule, on-cue behaviours, at-will behaviours, and behaviours which are always performed. He argues that situations that call for the performance of existing behaviours or increasing a behaviour on a reoccurring schedule are most suited for mobile persuasions.

Behaviour Change Support Systems

Oinas-Kukkonen (2013) coin the term BCSS to discriminate between persuasive systems and information systems with a focus on *voluntary* behaviour change and draw on Fogg's insights to propose a new, adapted taxonomy. Their O/C-matrix categorises behaviour changes by outcome type and change type. They distinguish between forming, altering, and reinforcing behaviours on the outcome axis and acts

of compliance, behaviour changes and attitude changes on the change axis. Complying necessitates a one-time behaviour, whereas behaviour changes are more enduring, and attitude changes are both the most sustainable and the most difficult to achieve.

	C-Change	B-Change	A-Change
<i>F-Outcome</i>	Forming an act of complying	Forming a behaviour	Forming an attitude
<i>A-Outcome</i>	Altering an act of complying	Altering a behaviour	Altering an attitude
<i>R-Outcome</i>	Reinforcing an act of complying	Reinforcing a behaviour	Reinforcing an attitude

Table 2.2: O/C-matrix. Table based on (Oinas-Kukkonen, 2013, p. 1226).

BCSS can foster reflection in various ways, such as through social traces, where users are presented with information about other users, social support, such as online forums, or visual displays, for instance, to give instantaneous feedback during an activity (reflection-in-action) or showing long-term patterns of user data (reflection-on-action) (Ploderer, Reitberger, Oinas-Kukkonen, & van Gemert-Pijnen, 2014), but also by employing conversational AI (Oinas-Kukkonen, 2013).

Section Summary

Information science, human-computer interaction, and behavioural science all define models of behaviour that differ in their focal points. While information science focuses on describing the way individuals interact with information, behavioural sciences pay more attention to internal processes mediating behaviour and complex processes of behaviour change, and human-computer interaction's focus lies on how behaviour can be described in a way that facilitates persuasive design.

With the exception of the information behaviour theories outlined in Section 2.1.1, which do not explicitly deal with the psychological processes that facilitate behaviour change and instead focus on how people interact with information at different points in time, the behaviour theories described above differ mainly in focus, complexity, and the vocabulary they use to describe cognitive processes. Depending on the theory's focal point, definitions differ slightly (such as the distinctions between locus of control, and perceived autonomy or self-efficacy and capability), but overall there is significant overlap between the factors described among theories. For instance, behavioural beliefs in the Theory of Planned Behaviour encompass outcome expectancies, a cornerstone of Social Cognitive Theory. The same relationship exists between self-efficacy and control beliefs. The theory that distinguishes itself in its ability to describe the entire lifecycle of behaviour change is the TTM. In contrast to the other behaviour models, it not only encompasses many of the factors influencing change but also describes how these develop and change in importance throughout the process of behaviour change. Information behaviour research also defines similar processes of behaviour in various contexts, such as during information seeking (Kuhlthau, 1988), meaning-making and information shaping (Ruthven, 2019, 2024) and transitions (Ruthven, 2022b).

Insights from persuasive design and BCSS research help to conceptualise and contextualise the research conducted in this thesis and the behavioural changes addressed in the experiments. Fogg's behaviour grid and Oinas-Kukkonen's O/C matrix are useful for categorising the types of behaviour change ad-

dressed, and information behaviour theories can serve as a basis for the interpretation of how users behave in interaction with information systems such as CAs in the context of behaviour change, but also how they engage with information, including their preparedness to self-reflect at different points in time.

The factors influencing change brought up by various behavioural theories described here corroborate the notion that internal factors play a significant role in the success and initiation of behaviour change. Many of these factors rely on individuals' attitudes towards change, themselves and their abilities, and the expected outcomes of changing or not changing. In the next section, I describe a counselling style intended to address many of the change factors discussed here: Motivational Interviewing.

2.2 Motivational Interviewing

MI was first popularised by W. R. Miller and Rollnick in the 1980s and 1990s. It involves talking to clients about behaviour change, mainly focusing on reflections and open questions, with the goal of resolving ambivalence, and eliciting *change talk* (utterances in favour of change) while minimising *sustain talk* (utterances in favour of the status quo) and can be applied as an alternative or a complementary approach to information-based interventions to motivate behaviour change. Since its conception, the approach has been applied in numerous contexts, demonstrating its versatility. In this section, I give an overview of application areas in the literature and summarise empirical studies examining the approach's efficacy in facilitating change followed by an outline of the core principles of MI and its role in behaviour change.

2.2.1 Effectiveness and Application Areas of Motivational Interviewing

MI was originally developed in the context of addiction counselling, for instance, to address problem drinking (W. R. Miller, 1983). Since its conception, it has been applied to numerous and diverse contexts of varying levels of severity. These range from productivity-based interventions, such as encouraging return to work (Jensen, 2016) and increasing academic performance (Strait et al., 2012; Terry, Smith, Strait, & McQuillin, 2013), to general health promotion, for example in the form of weight loss, physical activity, reduction of cardiovascular disease risk, to more serious contexts, such as substance abuse treatment, cessation of smoking, adherence to diabetes treatment, and compliance of psychiatric patients (Barrett, Begg, O'Halloran, & Kingsley, 2018; Britt, Hudson, & Blampied, 2004; Carels et al., 2007; Hardcastle, Taylor, Bailey, Harley, & Hagger, 2013; Resnicow et al., 2002; Rubak, Sandbæk, Lauritzen, & Christensen, 2005). Likewise, the counselling style is used in various settings, including primary care and palliative care (Pollak, Childers, & Arnold, 2011), correctional facilities (Hohman, Doran, & Koutsenok, 2009), and nursing (Beckwith & Beckwith, 2020).

Several systematic reviews and meta-studies have been conducted to shed light on the efficacy of MI. For example, Rubak et al. (2005) find that the majority of the studies covered in their review observe a significant intervention effect of MI in various physiological and psychological contexts, with the effects being the most common if MI is administered by psychologists and physicians (80% of cases) compared to other healthcare providers (46%). They also identify that an effect was reached in *Brief Motivational Interviewing*, brief encounters of less than 15 minutes, in 64% of the reviewed studies. They draw the

conclusion that MI has a higher success rate than traditional advice-giving in various application areas, but address the caveat that the feasibility of implementation in daily clinical work has not yet been systematically investigated.

Armstrong et al. (2011) identify a medium effect size of MI on weight loss compared to controls in their meta-review of randomised controlled trials (RCTs) around the application of MI to the treatment of overweight and obese patients. To get to this conclusion, they analysed 11 RCTs with durations between 4.2 and 18 months, which compared MI with other interventions (e.g. behavioural weight-loss programmes, or standard dietary care). Similar results are reported by Barrett et al. (2018), who find that physical activity, a lifestyle mediator of overweight and obesity, is significantly increased by a combination intervention of Cognitive Behavioural Theory (CBT) and MI compared to comparison treatments across 10 RCTs.

Lundahl and Burke (2009) investigate four meta-analyses about MI and find that after only two or three MI sessions, between 14% and 20% of study participants show improved outcomes compared to weak comparison groups (waitlist or no treatment). Compared to other active treatments, their review of the literature indicates at least equal effectiveness, with some meta-analyses reporting increased effectiveness of MI compared with other treatments, such as CBT and 12-step programmes, all while taking less time than the other treatments tested. The reviewed meta-analyses also indicate effectiveness across various problems, including addiction, risky (health) behaviours, and eating disorders, but also treatment engagement, parenting practices, emotional well-being and confidence to change, regardless of participants' initial problem severity. The effectiveness of MI shows to be durable up to at least a year after treatment, with some studies even reporting positive outcomes 3 years after the intervention. While Lundahl and Burke's (2009) review concludes that MI is suited both as a standalone intervention and a pretreatment for other interventions, two of the reviewed meta-analyses indicate that effects are strongest, if MI is used as a pretreatment.

When it comes to the specific mechanisms of MI, a systematic review by Magill et al. (2014) shows that the usage of MI-consistent skills by the practitioner leads to significantly more change talk, while the use of MI-inconsistent skills leads to less change talk and more sustain talk. Moreover, increased sustain talk is associated with worse outcomes at treatment follow-up. A more recent review by the same team (Magill et al., 2018) shows further support for this and similar results are found by Apodaca and Longabaugh (2009) and Apodaca et al. (2016).

2.2.2 Principles of Motivational Interviewing

Although many publications have detailed the core principles of MI (Beckwith and Beckwith (2020); Clifford and Curtis (2016); Hall, Gibbie, and Lubman (2012), to reference a few), the information presented in this subsection is based on a comprehensive, seminal book, written by the counselling style's founders if not explicitly noted otherwise: Miller and Rollnick's *Motivational Interviewing: Preparing People for Change* (W. R. Miller & Rollnick, 2002). In it, they define MI as a directive, client-centred counselling style. This means that the counsellor follows a clear goal that is ultimately favourable for the client's health or well-being. This goal can but does not have to be aligned with the client's initial wishes. MI's main focus lies on enhancing intrinsic motivation to change by exploring and resolving ambivalence and eliciting change talk. MI-Practitioners follow four key principles, encompassed by the

acronym 'RULE':

R *esisting the righting reflex* involves resisting the urge to give advice or tell the client what to do. This necessitates practitioners to set aside their knowledge of what is best for a client, and deeply listen to the client's feelings and thoughts about change.

U *nderstanding the client's motivations* aims at exploring a client's motivations, values, and goals. This helps practitioners in creating an understanding of what drives the client and how intrinsic motivation can be increased.

L *istening with empathy* cultivates trust and open communication in the form of validation, compassion and understanding towards clients' utterances, challenges and emotions by the counsellor.

E *mpowering the client* encourages the client to take an active role in their behaviour and fosters perceived autonomy, self-efficacy and a sense of ownership in decisions. It involves close collaboration with the client to explore options and create a plan for change.

In addition to keeping to these key principles, MI practitioners use four main conversational techniques, or micro skills, abbreviated as 'OARS' to develop discrepancy between the client's current behaviour and where they want to be in life, resolve resistance to change and support self-efficacy in an empathetic way:

O *pen questions* encourage clients to reflect and express themselves.

A *ffirmations* demonstrate empathy and respect, and show appreciation for the client's courage in discussing difficult issues.

R *eflections* mirror clients' utterances back to them. In doing this, counsellors not only demonstrate an understanding of a client's feelings and thoughts, but also encourage further discussions.

S *ummaries* give a concise overview of client utterances and reinforce key points of a conversation.

Depending on the situation and previous client utterances, different types of open questions and reflections might be appropriate. For instance, a negative statement about change or expressed resistance might be followed up with a 'reframing', a statement that reflects the client's statement but puts it in a different light, for instance by reinterpreting it in a positive way, thus helping the client open up to different points of view about their problem behaviour. In a similar fashion, client statements regarding desires to change can benefit from targeted questions which encourage them to look forward, exploring a future where things have changed already.

This necessitates practitioners to look out for specific types of client statements, which help them determine how to respond and what the user's current stance towards change looks like. These statements can be about *desires* (expressing a desire or wish), *ability* (utterances about self-efficacy), *reasons* (presenting motivators), *need* (considering the necessity of making a change), *commitment* (expressing commitments for the near future), *activation* (formulating a plan), and *taking steps* (steps taken recently). Again, these can be shortened to an acronym – 'DARN CAT'. Each of these utterance types can take the form of *change talk*, in favour of change, and *sustain talk*, in favour of the status quo. With the exception of activation, these client utterances are thoroughly defined in the *Motivational Interviewing Skill Code (MISC)*, a coding manual for quality control of MI sessions, which also includes the four micro skills OARS (W. R. Miller, Moyers, Ernst, & Amrhein, 2008) and which will be important

throughout this thesis.

All these principles and conversational techniques culminate in the evocation of 'MI-Spirit' which encompasses the elements of autonomy, collaboration, evocation, and empathy and is supposed to guide all interactions between MI-practitioners and clients. Such respectful, non-confrontational interactions are intended to invoke natural, client-guided change.

2.2.3 Role of Motivational Interviewing Throughout Behaviour Change

As outlined above, MI as an intervention does not put its main emphasis on providing information or telling clients what to do. On the contrary, much of the work of finding out the next steps to take falls on the client, as his or her autonomy is strengthened. In addition to creating a supportive environment, the main focus of MI lies on strengthening the factors that influence change mentioned in Section 2.1.2. It aims to increase self-efficacy and perceived autonomy, find and create intrinsic motivations, help to unveil discrepancies between behaviour and goals and values and explore and manage outcome expectations. Specifically, MI is posited to influence behaviour change based on three hypotheses. The technical hypothesis states that effective application of MI micro skills causes an increase in client change talk, which in turn leads clients to convince themselves of change, thus increasing outcome success. The relational hypothesis considers the favourable role of practitioner empathy in the ability to collaboratively focus on a particular problem and safely navigate complex change processes. The conflict resolution hypothesis posits that behaviour change in the context of MI results from the exploration and resolution of ambivalence (Magill et al., 2014).

To contextualise this, it can be helpful to refer back to the TTM: Although MI is primarily intended to facilitate change for people not yet in the action stage, its strategies can be used in different ways throughout the stages of change, as outlined by DiClemente and Velasquez (2002) and Hall et al. (2012). An overview of this is presented in Figure 2.4.

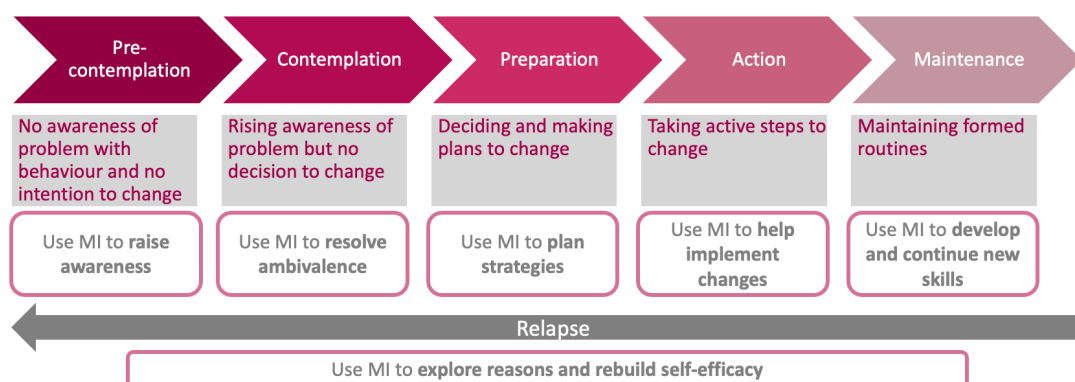


Figure 2.4: Application of MI in different stages of change. Figure based on Beckwith and Beckwith (2020, p. 348).

The main function of MI in precontemplation is to raise doubt about clients' current behaviour, increasing their awareness of the risks and problems associated with it (Hall et al., 2012). It can also be used to encourage clients to engage with their resistance to change, and their fears of failing, and indirectly help them discover faults in their rationalisations and justifications of current behaviour (DiClemente & Velasquez, 2002). In the contemplation stage, MI can be used to help clients explore the pros and cons of changing and the risks of maintaining the status quo, identify potential behaviour alternatives, disseminate their ambivalence, and improve self-efficacy (Hall et al., 2012). Useful tools in this stage include double-sided reflections and the affirmation and reflection of self-motivating statements. It can also be effective to get clients to actively engage with the advantages of their current problem behaviour, to help them prepare for temptations they will face in later stages (DiClemente & Velasquez, 2002). During preparation, MI can be used to assess and strengthen commitment to change and to help clients reach a realistic understanding of the challenges ahead, explore barriers to change, and help the client develop plans to overcome these barriers. In some cases, this might also include careful advice, providing information, and intervening, if the plans proposed by clients are unrealistic, not thought-through, or inappropriate (DiClemente & Velasquez, 2002).

Even in the action stage, having already initiated the change, people might encounter doubts and ambivalence (DiClemente & Velasquez, 2002). Supporting clients in their decisions, affirming their actions, and providing reassurance are helpful approaches at this stage to continue strengthening their resolve and self-efficacy. It can also be useful for clients to review their plans and decide whether things have changed or whether strategies need to be adapted (DiClemente & Velasquez, 2002). In the maintenance phase, MI can be used to help clients discover strategies to prevent relapse (Beckwith & Beckwith, 2020; Hall et al., 2012). If relapse occurs, the use of MI techniques to explore the reasons can help to see this as a learning opportunity rather than a failure and rebuild self-efficacy, ultimately motivating clients to re-initiate behaviour change (DiClemente & Velasquez, 2002).

All this being said, although the stages of change can serve as a useful framework for the delivery of MI, it does not rely on the counsellor's knowledge of the client's current stage of change and can also be applied stage-agnostically (W. R. Miller & Rollnick, 2009). Still, it is imperative to be in tune with a client's current stance, motives, and thoughts regarding change, and utilising the DARN CAT scheme outlined above, practitioners can even draw some conclusions about their client's readiness to change. Following Resnicow, McMaster, and Rollnick (2012), the scheme can be seen as a continuum, with desire expressing the least commitment and readiness to change, whereas taking steps constitutes the highest form of commitment and readiness.

Section Summary

Overall, the effectiveness of MI is well documented in the literature, spans a variety of contexts and can be applied to problem behaviours with various levels of severity. Although there is some lack of clarity on how MI works specifically, the technical hypothesis, stating that clients talk themselves into changing behaviour with the help of practitioners who use MI-consistent micro skills, is the most explored and well supported. Thus, successfully automating MI micro skills in a conversational system to facilitate guided user reflection, with a focus on eliciting change talk and helping them resolve their own ambivalence, has the potential to enable BCSSs to guide users through the stages of change and

address behaviour change types that are not suited to be addressed by traditional mobile persuasive systems as specified by Fogg (2009a), including the performance of new behaviours and decreasing and stopping existing behaviours.

At its core, MI exhibits some similarities to Rogerian psychotherapy in its focus on the client (Lundahl & Burke, 2009). One of the first CAs, ELIZA, was modelled after a Rogerian therapist in the 1960s (Weizenbaum, 1966). Considering the significant advances in NLP since then, the potential feasibility of automating such micro skills has dramatically increased. In the next section, I give an overview of CAs and recent developments in this area.

2.3 Conversational Agents

Generally, CAs can be categorised across two axes, one describing the scope and the other describing how text is generated by the system. **Task-oriented** chatbots are focused on a specific domain, often with the goal of solving a problem. This can be customer service, travel planning, or cooking assistance, for instance. On the other hand, **open-domain** CAs are not topically restricted and often come in the form of chit-chat applications without a clear conversational goal (McTear, 2020). Between these two opposites lies a wide array of blends, for example, chatbots serving a purpose while still allowing chit-chat dialogue as well.

CAs have been popular approaches for the implementation of behaviour change and mental health support for many years. Recent work has defined the term *social influence dialogue system* (Chawla et al., 2023) to refer to systems that aim at changing a user's attitudes, opinions or habits through dialogue, localising such systems between task-oriented and open-domain systems due to their reliance on social rapport to fulfil a task-oriented goal (see Figure 2.5).

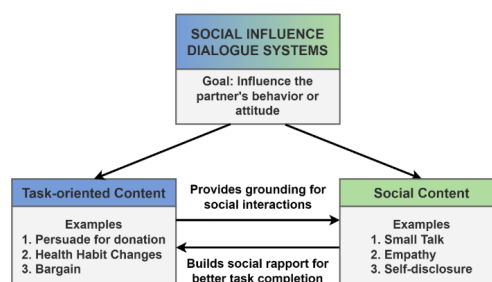


Figure 2.5: Conceptual overview of social influence dialogue systems. Figure published under CC BY 4.0 license in Chawla et al. (2023, p. 751).

Along the other axis, it is possible to distinguish between **different types of text generation** outlined below. Just as above, text generation approaches can be combined in different ways to take advantage of different functionalities. It is noted that in the context of this thesis, the focus is on text-based CAs. Other types, such as voice-based or embodied CAs largely follow the same architecture, with some

additional components (i.e. speech synthesis)¹.

2.3.1 Rule-Based and Retrieval-Based Conversational Agents

Rule-based CAs are the most simple and earliest form of conversational AI, introduced in the 1960s with ELIZA, which searches user inputs for keywords and maps them to specific predetermined patterns to generate outputs (Weizenbaum, 1966). An expanded version of ELIZA, called A.L.I.C.E. (Wallace, 2009) is based on Artificial Intelligence Markup Language (AIML). This allows the creation of templates, which map multiple types of user inputs to the same intent and response pattern (Wallace, 2003). A.L.I.C.E. represents 40,000 categories, 200 times more than ELIZA (Wallace, 2009). Classical approaches rely on semantic and grammatical parsing of the user input, for instance, according to predefined semantic grammar rules (see Table 2.3). Since this can easily lead to an explosion of rules, hindering system maintenance, more advanced systems of this type use machine learning techniques to classify user intents and map them to a response pattern (McTear, 2020).

Utterance	Intent
book I would like ...	→ FLIGHT_BOOKING
a flight flights	→ FLIGHT
to CITY	→ DESTINATION
London Paris New York ...	→ CITY
on DAY	→ DEPARTURE_DATE
Monday Tuesday ...	→ DAY

Table 2.3: Example for semantic grammar rules used in a rule-based CA for flight booking assistance. Table based on McTear (2020, p. 47).

In **retrieval-based** CAs, user utterances are matched to a fitting response from a (large) corpus of existing utterance-response pairs. This approach forgoes intent recognition and the formulation of rules employed in rule-based architectures and directly pairs input with corresponding output. This involves encoding both the user utterance and the response options as vectors, which can then be used to calculate similarity measures in a vector space (B. Hu, Lu, Li, & Chen, 2014).

2.3.2 Generative Conversational Agents

Generative or neural CAs typically rely on sequence-to-sequence (Seq2Seq) models, meaning a source sequence (user input) is directly translated into a target sequence (system output), without the intermediate steps of semantic parsing or intent recognition (Adamopoulou & Moussiades, 2020). This requires the transformation of language into word embeddings (Mikolov, Chen, Corrado, & Dean, 2013),

¹Some researchers differentiate between terms, such as *dialogue systems*, *chatbots*, *conversational agents*, *conversational AI*, and *conversational user interface* to convey different meanings and use cases (see, e.g., Jurafsky and Martin (2009)). In the context of this work, these terms will be used interchangeably to refer to text-based CAs. When referring to CAs with other (additional) functionality, this will be made explicit by mentioning the additional functionality (i.e. multimodal CA, embodied CA, voice user interface, etc.). Likewise, I will explicitly name the underlying architectures of CAs whenever necessary (i.e. rule-based, retrieval-based, generative) and the intended use case (task-oriented, chit-chat, social influence, etc.).

where each word is represented by a unique identifier (a number) that “captures its meaning and its relationship to the other words in the vocabulary” (McTear, 2020, p. 129).

2.3.2.1 Neural Architectures: RNN, LSTM, and Transformer

Traditional neural models used for this type of chatbot architecture are recurrent neural networks (RNNs), neural networks that allow the sequential learning of dependencies by remembering information from previous time steps, or long short-term memory networks (LSTMs), a version of RNN capable of targeted forgetting of information from previous time steps in favour of increased pattern learning (Hochreiter & Schmidhuber, 1997). Seq2Seq models have been revolutionised with the invention of the transformer, a neural architecture that captures not only the information contained in previous time steps but processes information in parallel. This means that when processing a *token*², transformer models take into account not only the words that appear in the text before but also those that come after (Vaswani et al., 2017).

2.3.3 Large Language Models

These advancements have since led to the inception of LLMs, a type of transformer-based model pre-trained on massive amounts of textual data (T. B. Brown et al., 2020; Devlin, Chang, Lee, & Toutanova, 2019). Due to their sheer size in terms of parameters and training computation, these models are highly capable (Wei, Tay, et al., 2022). Today, generated text is often not distinguishable from human-written text and can superficially match its quality, even in sophisticated contexts such as the medical domain (Ayers et al., 2023; Tang, Chuang, & Hu, 2023) (see Figure 2.6 for an example). The first LLM, BERT, was primarily focused on natural language understanding (Devlin et al., 2019) opening the door to much more reliable classification of text, for instance. Shortly after, the Generative Pre-trained Transformer (GPT) followed, which focused specifically on generating textual data

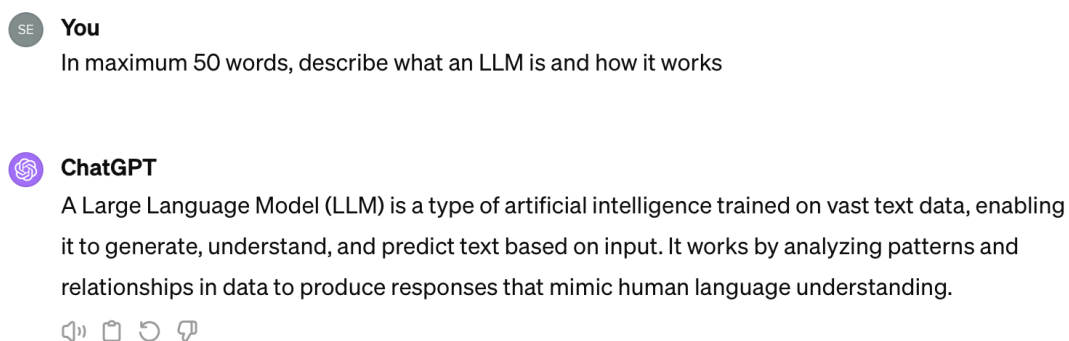


Figure 2.6: LLM-generated text is often of impressive quality and indistinguishable from human-generated text (example generated with GPT-4).

²the terminology used for transformer-based embeddings, which work with word components instead of complete words.

by statistically predicting the next token based on all previous tokens in a specified context window (Radford, Narasimhan, Salimans, & Sutskever, 2018). GPT was followed by increasingly sophisticated evolutions (GPT-2, GPT-3 (T. B. Brown et al., 2020), GPT-3.5, also called ChatGPT, and GPT-4 (OpenAI, 2023a)) trained on ever higher amounts of data and containing evermore parameters, as well as countless open source alternatives, such as GPT-Neo (Black et al., 2022) and Bloom (Scao et al., 2022). In subsequent iterations of GPT, reinforcement learning with human feedback was introduced, which allows a language model to iteratively learn and improve its generation capabilities based on feedback from its users (Ouyang et al., 2022).

LLM research has its own vocabulary, which is essential to describe how these models are used. The most important terms for this thesis are defined below:

- **Fine-Tuning** describes the additional training of pre-trained LLMs with domain-specific data (Devlin et al., 2019). This is especially relevant for *encoder* models, with a focus on language understanding such as BERT, and tunes them to fulfil a specific task, such as sentiment analysis or other types of text classification. This way, BERT can be used just like other machine learning models with a focus on NLP. It is also possible to fine-tune generative LLMs, although this is less common for the more advanced GPT models (GPT-3 and newer), which have much higher capabilities “out of the box” and can be *prompted* to fulfil a task instead (T. B. Brown et al., 2020).
- **Prompt** refers to the input to a GPT model (starting from GPT-3 (T. B. Brown et al., 2020)). This prompt can simply be a user utterance, but it can also contain instructions on how the output is supposed to be structured or what information is supposed to be included in the response. A simple example of a prompt would be “translate the following sentence into German: [Sentence]”, where the input contains an instruction of what requirement we expect the output to contain, namely a German translation of the given sentence.
- **Prompt Engineering** is the act of systematically engineering prompts to obtain specific outputs. This practice can range from providing additional information about the input or a conversational context to the adherence to specific prompting principles, or the iterative improvement of prompts based on outputs or even with the help of the LLM (OpenAI, 2023b; Reynolds & McDonell, 2021; Wei, Wang, et al., 2022; White et al., 2023).
- **In-Context Learning/Few-Shot Learning** is defined as the provision of a small number of examples (in many cases as little as one or two) in order to improve the model’s output. This is the alternative to fine-tuning predominantly employed in LLMs following GPT-3 (T. B. Brown et al., 2020). The general idea behind this is that these models already contain all the information they need, but examples can help them to locate the necessary information in their training data (J. Liu et al., 2022; Reynolds & McDonell, 2021). If no examples are given in the prompt, this is often referred to as zero-shot learning.
- **Retrieval Augmented Generation/Knowledge Grounding** refers to the provision of explicit knowledge to a language model, for instance in the form of a single document or a database, to increase domain-adaptation and factuality of generated responses (Lewis et al., 2020).

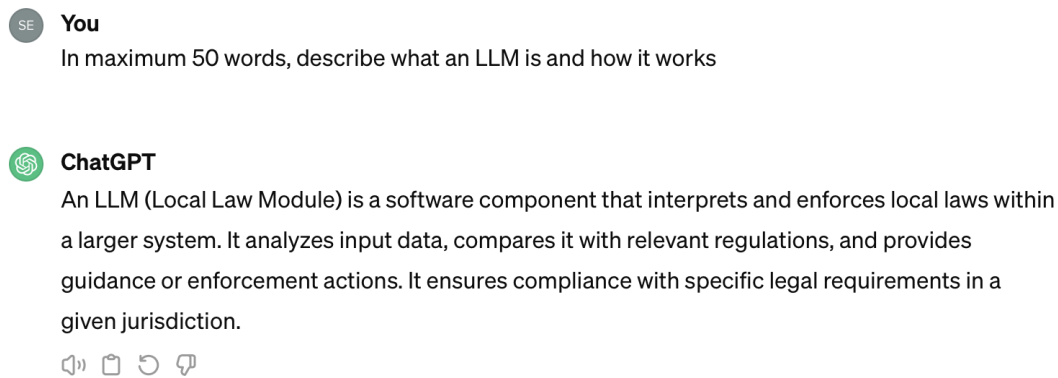


Figure 2.7: LLMs are in the habit of hallucinating - inventing and presenting untrue statements as facts (example generated with GPT-3.5).

2.3.4 Advantages and Disadvantages of Text Generation Approaches

Employing a rule-based architecture results in highly predictable system messages, giving developers complete control over the system and its outputs. However, it also requires developers to anticipate as many user intents as possible in advance. This results in low adaptation potential of the resulting programmes to new use cases and a generally low flexibility and personalisation of outputs, since user utterances need to be present within the rule base in order to be processed. As a result, such systems often limit user input, for instance, by making users choose between a list of response options. Alternatively, unknown inputs are assigned to a response pattern to signal incomprehension, a strategy that can quickly become frustrating to users (McTear, 2020).

Retrieval-based chatbots rely on high-quality data, which is not necessarily available in sensitive contexts. Although these systems tend to offer more flexibility than rule-based systems and are capable of producing mostly relevant and consistent responses, they still struggle with truly open-domain conversation and the number of responses is limited to the amount of data available in the knowledge base (McTear, 2020; Pandey & Sharma, 2023). This can result in system responses appearing to be “canned” or impersonal (Svikhnushina, Placinta, & Pu, 2021). Furthermore, as I will show in Chapter 3, most large datasets available in the context of counselling and behaviour change are not perfectly suitable for the construction of a written CA for counsellor-client interactions in the context of behaviour change.

CAs using GPT architectures benefit from highly human-like text generation capabilities. Whereas early examples of generative CAs suffered from ungrammatical outputs (McTear, 2020, p. 138), this rarely happens when LLMs are used (the exception being languages that are barely or not present in the training data). However, this comes at the expense of controllability. Due to the vast size of these models (both in terms of parameters and training data), it is impossible to predict the exact output an utterance will lead to. While this has coined an entirely new discipline of prompt engineering, these models are frequent topics of debate around ethical issues, such as their tendency to capture and reproduce implicit biases and stereotypes contained in the data (Bender et al., 2021), or their propensity to output *hallucinations*, untrue statements framed as facts (Bang et al., 2023) as can be seen in Figure

2.7, with prompt engineering even being compared to wizardry at times³.

Section Summary

CAs can be categorised by scope (task-oriented vs. open-domain) and text generation method (rule-based, retrieval-based, generative). Recent years have seen a drastic increase in the capabilities of conversational AI through the introduction of transformer architectures and LLMs. These offer superior flexibility and quality of outputs compared to classical rule-, retrieval- and neural-based approaches. However, these come with their own challenges, such as a lack of controllability and implicit biases in the training data. As such, using such models in the context of behaviour change requires rigorous testing and safety measures.

2.4 Chapter Summary

This chapter has provided the theoretical foundation necessary to validate and comprehend the remainder of this thesis. It has given an overview of behaviour change and how it is conceptualised across the disciplines of information science, behavioural science, and human-computer interaction. This was followed by an introduction to MI and its principles, which have informed large parts of the design of the conversational BCSS introduced in this thesis. The chapter concluded with a short history of CAs and an overview of different chatbot architectures, outlining the advantages and drawbacks of the different text generation methods. The next chapter will provide an overview of the related work more closely connected to this thesis, namely literature around MI-based BCSSs.

³https://www.linkedin.com/pulse/prompt-engineer-magician-zsolt-vicze-o41se/?trk=public_post_main-feed-card_feed-article-content.

3

Related Work

Technical systems to facilitate behaviour change have been studied in various constellations. In this chapter, I give an overview of the landscape surrounding *conversational BCSSs*, systems that rely on conversational AI to facilitate behaviour change, putting special emphasis on systems that use MI principles. I outline gaps in the research around this topic and how such gaps have been partially addressed in different, but related domains such as mental health. The chapter begins by providing an overview of existing conversational BCSSs (Section 3.1). This is followed by work that focuses on understanding user attitudes about behaviour change (Section 3.2). Section 3.3 outlines related work, including ethical considerations, geared toward reacting in appropriate ways, specifically in the context of automated counsellor-client interactions based on LLMs.

Much of the related work summarised in this chapter deals with mental health support, as opposed to behaviour change. Mental health support is primarily occupied with addressing psychological distress and mental illness. In many cases, attempts to automate such support conversations are based on peer support data. In contrast, behaviour change support focuses on altering specific harmful behaviours. As such, supporting and motivating behaviour change can be part of mental health interventions, and some of them do incorporate MI strategies. However, the behaviour change support interactions focused on in this thesis are concerned with the strategic use of MI strategies and the modelling of interactions between a counsellor and a client, rather than peer-to-peer conversations. Section 3.3.1 elaborates on how this different focus affects the appropriateness of chatbot utterances in such systems.

3.1 Conversational Behaviour Change Support Systems

In this section, I give an overview of existing conversational systems for behaviour change and mental health support, with a focus on the implementation of MI principles. I begin by describing CAs using text generation approaches which do not employ LLMs (subsequently referred to as “classical approaches”), before elaborating on LLM-based applications.

3.1.1 Classical Approaches

CAs have been used in various counselling contexts, from mental health support (Gaffney, Mansell, & Tai, 2019; Kocaballi et al., 2022) and psychotherapy (Xu & Zhuang, 2022), to various behaviour change domains, including the handling of neurological disorders (Pereira & Díaz, 2019), smoking cessation (He, Basar, Wiers, Antheunis, & Krahmer, 2022), conversational skill training (Samrose & Hoque, 2022), stress management (Park et al., 2019), weight loss (Holme et al., 2018), substance abuse (Olafsson, Wallace, & Bickmore, 2020), and, most prominently, nutrition and physical activity (Luo et al., 2021; Oh et al., 2021; Pereira & Díaz, 2019).

Upon reviewing the literature, different categories of classical CAs emerge, which differ mainly in a) their level of *interactivity*, indicated by the type of input they allow from the user, and b) their *responsiveness*, here defined as the degree to which an effort is made to understand user input and extrapolate attitudes, thoughts, and emotions from user input to inform the further course of the conversation (see Figure 3.1 for an overview). In the context of MI, such responsiveness could include the provision of reflections of user utterances, the posing of targeted questions, based on the information provided by the user, or choosing the next system utterance based on the behavioural intents, needs, or attitudes voiced in users' previous responses.

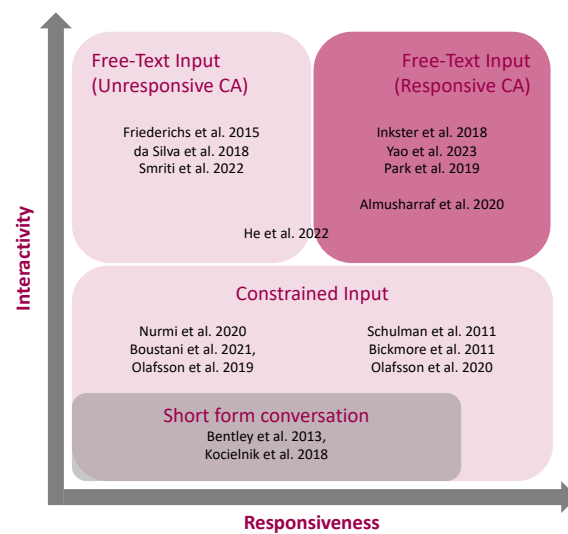


Figure 3.1: Overview of non-LLM-based behaviour change support CAs and their level of responsiveness and interactivity.

In the following, I outline and give examples for each of the categories in turn, from least to most interactive and responsive.

3.1.1.1 Short-Form Conversation

At the most basic level, CAs in this context consist of a natural language representation of behavioural and health-related metrics, such as step counts, weight, tracked sleep and mood, and statistical patterns

emerging from these metrics (Bentley et al., 2013), to help users self-reflect about their habits and allow them to identify factors that influence their well-being. Work by Kocielnik et al. (2018) goes a step further and engages users in mini-dialogues consisting of two turns (question → response → follow-up question → response) based on visualisations of their behavioural data (see Figure 3.2). The prescribed questions they use in their system were generated in workshops with users of fitness trackers and encompassed multiple categories, such as goals, tracking, observations, and motivations. Questions were compiled into mini-dialogues based on their categories (i.e. questions about understanding would be followed up by questions about future actions), where some combinations were static, while others were altered slightly based on intent recognition of the user utterance.

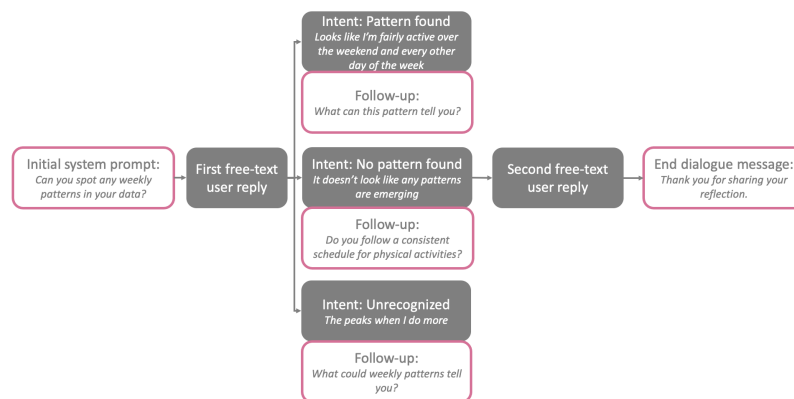


Figure 3.2: Mini-dialogue flow presented by Kocielnik et al. (2018) (based on figure in Kocielnik et al. (2018, p. 2)).

3.1.1.2 Constrained Input

The majority of CAs designed for longer interactions for behaviour change are rule-based and many heavily restrict user input (Luo et al., 2021; Oh et al., 2021). For instance, Nurmi et al. (2020) design a rule-based CA that employs various MI techniques, including reflections, open questions, and emphasis on users' autonomy, but only allows multiple choice inputs. They include evocative questions that prompt users to envision certain scenarios to foster self-reflection but do not provide a way to record their thoughts. In an effort to deliver implicit empathy cues in their agent, Boustani, Lunn, Visser, and Lisetti (2021) model an embodied CA for a brief MI intervention for alcohol misuse, where inputs are fully constrained to multiple choice inputs. Similarly, Olafsson et al. (2019) create two rule-based embodied MI agents for healthy nutrition and physical activity, both restricting user input to multiple-choice.

As an example of a responsive rule-based MI-CA with restricted user input, Schulman, Bickmore, and Sidner (2011) create a system for physical activity and fruit and vegetable intake which contains 1000 adjacency pairs, consisting of an agent utterance and a list of possible user responses, each of which corresponds to a valence (change/sustain), a category (*status-quo implication/change implication*, referring to positive or negative aspects of current behaviour or change, *change outlook*, referring to self-

efficacy, and *change intention*), and content (i.e. stress, and other topics, concerns or values). These adjacency pairs are interconnected to create a dialogue flow that is highly dependent on the user's choices during the conversation, based on a framework introduced by Bickmore, Schulman, and Sidner (2011). CA replies are evocative and elaboration questions, reflections, summarisations, and acknowledgements of importance. In their evaluation, the authors note the particularly high implementation effort required to manually author utterance-response templates. Another approach is presented by Olafsson et al. (2020), who train machine learning algorithms on a dataset of spoken MI interactions (Monti et al., 2016), to identify the most suitable type of system output (question, reflection, informing) based on preceding turns for a rule-based MI-voice agent with constrained input in the context of substance abuse counselling.

3.1.1.3 Free-Text Input

In their systematic review of AI-based chatbots to promote changes in health behaviour, Aggarwal, Tam, Wu, Li, and Qiao (2023) note that 14 of the 15 included studies do not adequately describe the AI algorithm used, and only one study reports on the safety of conversations. They do find indications in the literature for a positive effect of using theoretical foundations, such as MI and CBT to guide chatbot dialogue and the superiority of "free-flow conversation" compared to rule-based or restricted conversations. However, even in cases where free text input is possible, this does not always have an influence on the course of the conversation, with some CAs not designed to understand user utterances in any way. Instead, such systems focus on following a predefined conversation protocol, for instance, posing sequences of questions, motivating a categorisation into unresponsive and responsive free-text CAs.

Unresponsive Free-Text Conversational Agents

da Silva et al. (2018) report on a user study employing a voice-based MI-robot that aims to support users in increasing their physical activity. Their system allows spoken user replies, but the conversation is fully prescribed in a way, that allows each question to "make sense, regardless of how the participant answered the question before" (da Silva et al., 2018, p. 3). As such, the system provides no reflection of user utterances and does not respond to user utterances in a personalised manner. Instead, users were prompted to summarise their own utterances as a way of reflection. Similarly, Smriti et al. (2022) introduce a system for healthy eating targeted at parents, in which the prescribed system messages are generic enough to be suitable, regardless of users' free-text responses to open questions.

Another such example is described by Friederichs et al. (2015), who compare three different versions of an MI-based physical activity CA, where one version allows only multiple-choice input, the second allows a combination of multiple-choice and free-text input, and the third is fully dependent on free text. They synthesise summaries at the end of the conversation, which consist of a rephrased version of multiple-choice replies in the two conditions that include them, and a generic summary with user responses pasted in for the third condition. Although this approach still does not understand user input and incorporate it into the design of the conversation, it does allow users to put their thoughts in writing. A similar architecture is used by He et al. (2022) in their comparison of an MI-enabled and a non-MI-adherent rule-based CA for smoking cessation, both of which allow some free-text input in

their CAs, but control the flow of the conversation by primarily posing close-ended questions which do not lead to a deeper understanding of user attitudes. Although there seems to be some simple pattern recognition involved in their architecture, this only alters a few single turns throughout the course of the conversation. Nevertheless, in a RCT, they find significant effects on motivation to quit in both conditions.

Responsive Free-Text Conversational Agents

While the systems outlined so far show indications of motivating behaviour change in users, their reliance on prewritten scripts minimises long-term usability, and their lack of flexibility in addressing user attitudes about change hinders their capabilities in supporting user self-reflection across diverse situations, for instance, as they face individual barriers for change or have differing motivational levels. Being responsive to user inputs and the attitudes displayed within them can alleviate these limitations.

An attempt towards increased responsiveness is presented by Almusharraf, Rose, and Selby (2020), who create a chatbot using iterative design. Users are asked for their reasons to smoke and quit smoking, which are then grouped into categories. Users have the option to correct the chatbot if the category is not well chosen and the chatbot provides reflections based on the user utterance category. For example, if a user says that his or her skin suffers from smoking, the chatbot classifies this as aesthetic and responds with a reflection about aesthetics. Using this approach, the authors collect 21 distinct categories of reasons for and against smoking. This approach represents a first step towards understanding users and why they behave a certain way but is heavily context-dependent. Furthermore, other intentions that are relevant for conversations about change, such as user intentions or recent activities, are disregarded.

Park et al. (2019) design a rule-based MI-CA for stress management with a keyword-based text generation approach resembling that of ELIZA. They define a conversational flow with a predefined order of system utterance types (i.e. focusing question → reflection → focusing question...) across multiple consecutive phases of an MI-conversation (based on the four processes of MI of engaging, focusing, evoking and planning, defined by W. R. Miller and Rollnick (2002)). During the conversation, the system then assembles outputs based on keywords present in the preceding user utterance and the system output type determined by the conversation flow.

Going beyond keywords, Yao et al. (2023) create a data-driven rule-based CA for postpartum support that, given a user utterance, classifies the user's psychological state (given a choice of depressive mood, anxiety, and other), as well as the concern their utterance addresses (e.g. concerns regarding the baby crying, lifestyle transitions, a lack of personal support, etc.) and uses this information to select a tailored response made up of an empathetic statement and an open question. As a basis for the classifications, they use a dataset of postpartum helpline support conversations.

In the context of mental health and well-being, retrieval-based and generative CAs are more prevalent (Xu & Zhuang, 2022). This allows for the generation of responses that are more tailored to user utterances, provided that a similar utterance exists in the underlying corpus. Often, such corpora are drawn from peer-to-peer interactions such as the *EMPATHETICDIALOGUES* dataset presented by Rashkin, Smith, Li, and Boureau (2019). One of these peer-to-peer datasets is used by Morris, Koudous, Kshirsagar, and Schueller (2018) to create a retrieval-based CA for emotional support. In their

system, the CA's response is presented as a community reply. Another retrieval-based CA for mental health, presented and evaluated by Inkster, Sarda, and Subramanian (2018), lays a strong focus on self-reflection. The authors claim that the system “responds to emotions that a user expresses over written conversations and, in its conversation, uses evidence-based self-help practices such as CBT, dialectical behavior therapy, MI, positive behavior support, behavioral reinforcement, mindfulness, and guided microactions and tools to encourage users to build emotional resilience skills.” (Inkster et al., 2018, p.3), but do not offer specifics about the underlying data or chatbot architecture. In their user study, they find significant improvements in mood after two weeks of usage, an effect that was stronger for high-frequency users.

I could find only one project concerning a retrieval-based, fully unconstrained CA for behaviour change that laid a strong emphasis on understanding the user and employing the used information to guide the conversation. This is a conversational framework of an MI-based CA for smoking cessation proposed by Saiyed et al. (2022) and Satterfield et al. (2022), which intends to classify user utterances with *MISC* client codes and combine the information with topic recognition and a question-answering model to retrieve one of several utterances prewritten by MI experts and stored in a matrix. To my knowledge, this framework has not yet been implemented in a running system or tested by users. Although categorising client *MISC* codes constitutes a step towards a deeper comprehension of user perspectives, this framework still encounters the issue of needing to predefine all responses, which could cause the system to become monotonous and the replies to appear generic fairly quickly.

Xu and Zhuang (2022) find that the majority of chatbots for psychotherapy are retrieval-based and the ones that use generation-based dialogues often produce repetitive and meaningless turns. For instance, Evebot, a generation-based CA that uses MI, produces responses which do not correspond to effective dialogue in 10% of the time (Yin, Chen, Zhou, & Yu, 2019). Based on their findings, Xu and Zhuang (2022) highlight some future tasks including the creation of mechanisms to detect and avoid harmful chatbot turns and the collection of “valid, real, and rich corpora” (Xu & Zhuang, 2022, p.1) for the training of safe and effective generation-based chatbots. Other challenges of such technologies identified in the literature were repetitive content, high attrition, technical issues and safety and privacy concerns (Luo et al., 2021).

3.1.2 LLM-based Systems

While the issue of repetitive content might be alleviated through the use of LLMs for text generation, safety and privacy concerns prevail and are even greater given the lack of controllability of such models. As such, a lot of research around LLMs has circumvented using LLM-generated text alone as a text generation method in CAs, instead focusing on evaluating LLM's capabilities in simulations or focusing on the generation of specific behaviours (Weidinger et al., 2023). Among the countless studies concerning the evaluation and exploration of LLM's capabilities, a small body of research has attempted to employ LLMs for mental health CAs and a very limited number of publications have focused on the application of LLMs in MI and behaviour change systems.

In their systematic review about the application of LLMs in mental healthcare, Hua et al. (2024) identify 16 studies, many of them published on the preprint server arXiv pending peer review, which focus on the use of LLMs with the purpose of improving empathetic capabilities and contextual awareness, with

12 articles positioning their research in the context of CAs intended to directly interact with users, and the remaining four intended as support for mental health providers. Upon closer inspection, however, many of these publications still focus on text generation techniques in simulated scenarios (e.g. Cho et al. (2023); Y. K. Lee, Lee, Shin, Bae, and Hahn (2023)), evaluating performance through automatic metrics and human judgements provided by annotators, or in some cases even solely through GPT-4-based evaluation (e.g. J. M. Liu et al. (2023)), instead of running studies with fully developed systems with actual users. For instance, Q. Chen and Liu (2023) use GPT-2 to identify and learn conversational strategies used by mental health counsellors and use prompting strategies to elicit system responses utilising such strategies to address the needs of help-seekers in a zero-shot setting. They evaluate their system by simulating responses on a Chinese dataset consisting of mental health conversations and collect both automatic and human evaluation metrics, given by independent annotators, but do not test their system in actual conversations with users. Further research on LLM-based text generation in isolation, without its incorporation in a conversational system or user evaluations, will be outlined in Section 3.3.

In a more application-oriented work, Lai et al. (2023) create a framework based on a Chinese dataset of mental health conversations. The system is intended to serve as a question-answering system both for mental health professionals, supporting them in counselling sessions, and for end-users when no human support is available. They implement a prototype web interface that allows users to ask a question and get a single reply, with the option of rating the reply to help them further improve their model. However, similarly to Q. Chen and Liu (2023), they only report metrics collected outside of actual system usage, via automatic means or evaluations by human annotators, and do not study how users interact with the system and to what extent their expectations are met.

Yao et al. (2023), whose systems are fully focused on providing emotional support and fostering self-reflection and do not provide advice to users, compare their rule-based postpartum support CA with two other chatbot architectures: A generic CA consists of a generalisable empathy statement, followed by a universally applicable open question, and a generative CA, based on a GPT-2 instance fine-tuned on the same helpline dataset used to train the classifiers employed in their rule-based system. While in one tested condition the helpline dataset is filtered to exclude specific types of helper behaviours before fine-tuning, no additional measures are taken to control the type of output generated by GPT-2.

An approach to utilise LLMs to generate natural language personalised behaviour change advice is presented by Corda, Massa, and Riboni (2024). They create a system that uses behavioural data to provide personalised, LLM-generated advice on sleeping habits. They use a decision tree based on smartphone sensor data to forecast sleep quality and then provide the relevant nodes of the decision tree to the LLM as context to generate advice, as well as explanations for the advice given. The advice is then provided to users in a non-interactive text display. Knowledge grounding has been shown to drastically limit hallucinations (Lewis et al., 2020), and the approach of coupling LLMs with explainable algorithms to generate personalised advice seems promising. However, in this case, since no user-generated natural language input is involved, the number of GPT prompts is naturally limited to the number of possible combinations of decision tree nodes. Thus, it stands to reason that generating a corpus of advice and explanations in advance, and subsequently reusing the previously created system outputs would be a more cost-effective and efficient approach. This is more complicated in cases where users interact with the system in natural language, as is the case for CAs.

Further attempts at leveraging prompt-based text generation for the use of LLMs in standalone user-facing conversational systems include Vowels, Francois-Walcott, and Darwiche's (2024) qualitative evaluation of GPT-4's capabilities to provide relationship advice, based on 20 conversations and follow-up interviews with a custom-instructed GPT-4 instance, which shows that GPT-4 successfully displays therapeutic behaviours, and Kumar et al.'s (2022) comparison of different prompt-adapted versions of GPT-3 in the context of brief 5-minute conversations about mood management. In their study with 945 participants, which was run on Amazon Mechanical Turk, they change the prompt given to GPT-3 based on an identity (coach or friend), an intent (informed by psychological literature, i.e. "The [identity] asks open-ended reflection questions and helps the Human develop coping skills"), and a behaviour matrix ("The [identity] has strong interpersonal skills"). In their experiments they identify a range of positive behaviours displayed by GPT, such as the breaking down of problems into smaller, easier-to-address issues, making users feel heard, and offering short- and long-term solutions. In both cases, the prompt passed to GPT stays the same throughout the conversation.

As a further stage of prompt-based CA design, Kim et al. (2024) design an interactive LLM-based journaling tool for the support of psychiatric patients with depression. In their system, an LLM-based dialogue analyser classifies a conversational stage, which is used to decide what instructional prompt to pass to GPT-4 along with the conversation history. As a pre-emptive measure for harm mitigation, they have users undergo a mental health survey before each diary entry, triggering expert intervention in severe cases. Their tool also extracts key information from LLM-guided patient self-reflections and forwards this information to their therapists. In their experiments, the therapists remarked that this led to a deeper understanding of their patients, gained through better insights into their daily lives, and their willingness to respond to sensitive questions posed by the LLM due to its perceived impartiality.

In the specific context of applying MI for behaviour change, Basar, Hendrickx, Krahmer, de Bruijn, and Bosse (2022) propose a CA that combines retrieval and LLM-based generation to reduce harm potential and ensure appropriate interaction in highly sensitive domains while circumventing the repetitiveness and scalability issues of prescribed approaches. In their suggested solution, retrieved and generated responses are compared in a selection module, which determines the more appropriate reply.

Similarly, A. Brown et al. (2023) develop a CA for smoking cessation that combines prescribed questions with GPT-2 generated reflections of user utterances. In one of their conditions, one of the prescribed questions is replaced in the conversation flow and an additional optional prescribed question is included depending on the context of the preceding user utterance. In a user study, this approach significantly increased users' confidence in change, readiness to change, and perceived importance of change, as well as perceived empathy of the CA one week after the interaction, with a CA that solely relies on prescribed questions being less effective. This serves as an additional indicator that enabling chatbot outputs to be flexible in such interactions can lead to increased effectiveness.

3.1.3 User Perceptions of Conversational Behaviour Change and Mental Health Support Systems

User perceptions of CAs vary depending on system capabilities, and in the context of behaviour change and mental health support CAs often differ greatly in their application area, target audience, theoretical grounding, and conversational techniques, to an extent that makes a direct comparison between systems

difficult. Furthermore, many of the systems introduced above were tested in small user studies with few participants and the methodologies applied by the researchers to evaluate the efficacy of their CAs differed greatly (ranging from purely qualitative (Park et al., 2019; Vowels et al., 2024) to predominantly quantitative (Kumar et al., 2022)). Nevertheless, the literature reveals some insights into expectations users have for conversational BCSSs.

Rule-based CAs have the potential to successfully elicit reflections about awareness, understanding, and new insights (Kocielnik et al., 2018). As noted by Nurmi et al. (2020), who ran a think-aloud study of their system, such reflections can be related back to *MISC* client codes. Questions are generally welcomed by users, who take them as a chance to self-reflect and generally feel heard by the CA (Park et al., 2019). However, follow-up questions and feedback given by such systems is quickly perceived as repetitive, generic, clichéd, and impersonal and some user groups might lose patience, looking for quick advice instead of lengthy self-reflection sessions (Kocielnik et al., 2018; Park et al., 2019; Yao et al., 2023). Based on this, what users want does not directly have to be connected to what has the most effect, as shown by Olafsson et al. (2019). They run a user study and test the effect of user input being constrained only to change talk options, compared to choices reflecting a full range of attitudes, and find that, although coercing change talk in this way increases users' confidence in behaviour change, users prefer the version, in which they are not forced to be optimistic about change, indicating that users appreciate systems that offer a more accurate articulation of their attitudes.

It seems that even CAs that fully constrain the user input can lead to significant increases in self-efficacy and motivation after interaction with the system in both conditions (Olafsson et al., 2019), however user study results presented by Friederichs et al. (2015) indicate that allowing users to put their thoughts in writing is important for the attainment of autonomous motivation for change. Furthermore, the inability to provide effective complex reflections of user utterances might be a hindrance to applying MI, as shown by He et al. (2022), who find no significant improvement of an MI-enabled rule-based chatbot over a non-MI-adherent CA on user engagement, therapeutic alliance, a metric to evaluate the relationship between a client and a therapist, and perceived empathy.

In Yao et al.'s (2023) comparison of rule-based and generative (GPT-2-based) CAs, questionnaire results indicate that the rule-based version appeared the most helpful, engaging and satisfactory. One potential reason for this is that their generative chatbot was susceptible to producing confusing and nonsensical text, an issue that was likely due to the small size of the fine-tuning dataset and that has been observed before when fine-tuning Seq2Seq models (Meyer, Schmidhuber, & Kruschwitz, 2022; Yin et al., 2019). This is likely to be reduced by using prompt-based GPTs, which drastically reduces the likelihood of nonsensical outputs. Even in CAs with nonsensical outputs, a positive effect of usage can be observed, as shown by a user study with 30 participants conducted by Yin et al. (2019), where the usage of the system led to significant improvements in mood.

The issue of repetitive output seems to persist even in the case of some CAs based on prompt-based GPTs, as shown in Kumar et al.'s (2022) comparison of prompts. However, conversational LLM-based systems in the context, with the exception of Kim et al.'s (2024) journaling tool, use the same instructional prompt throughout the conversation and do not take any additional efforts to tailor conversations to the user's current state or attitudes. Generally, LLM-based CAs seem to be perceived as non-judgemental (just like traditional CAs), thoughtful, and easy to understand (Kumar et al., 2022), helping users open up to novel solutions to their problems, find clarity on existing issues and making

them feel safe through their anonymity (Vowels et al., 2024), but also open up concerns about privacy (Kumar et al., 2022). Additionally, Corda et al. (2024) show that users do not blindly trust LLM-generated advice, and providing explanations for the advice given is integral to creating trust, as well as increasing the perceptions of users of the impact the system can have on improving their habits (sleep quality in the context of their user study).

Section Summary

This section has provided an overview of conversational BCSSs, emphasising the transition from classical approaches to LLM-based systems. It has highlighted the evolution from rule-based, retrieval-based, and generative systems without the use of LLMs, to the incorporation of LLMs in creating more adaptive, responsive, and personalised CAs for behaviour change and mental health support. Classical approaches demonstrate the potential of CAs to motivate behaviour change, improve mental health, and support self-reflection, but are often limited by their lack of flexibility and understanding of the user and the high effort required for creating and maintaining effective dialogue systems. The introduction of LLMs has the potential to overcome these limitations by generating more natural, varied, and contextually relevant responses. Considering results outlined by Pérez-Rosas, Mihalcea, Resnicow, Singh, An, Goggin, and Catley (2017), who compare therapist language across different target behaviours (medical adherence, smoking cessation, dietary changes) and find no difference in language constructions, with only some words differing for different target behaviours, using LLMs in this context could also enable CAs to function for multiple target behaviours.

However, the development of LLMs has unveiled new challenges and gaps in the current research landscape. While LLMs offer advancements in text generation capabilities, ensuring the safety, privacy, and ethical use of these technologies remains a significant concern. The lack of controllability in LLM outputs, coupled with the potential for generating harmful content, underscores the need for robust mechanisms to monitor and guide the behaviour of these systems. Furthermore, despite the potential for LLMs to enhance the personalisation of behaviour change support, the effectiveness of these interventions in real-world settings is not yet fully understood. The literature suggests a gap in comprehensive user studies that explore how individuals interact with LLM-based systems, what they expect from LLM-based BCSSs and the impact these interactions can have on users' readiness to change their behaviour. Moreover, the integration of LLMs with other data sources and decision-making algorithms, as seen in attempts to generate personalised advice, presents an opportunity for further exploration.

The remainder of this chapter is dedicated to two crucial aspects for advancing LLM-based conversational BCSSs: Understanding a user's current attitudes about change based on what they write, and guiding CA responses in a manner tailored to the user state and suitable for the conversational context of automated counsellor-client interactions. These aspects are vital for building on existing frameworks to create CAs that can effectively support users in achieving their behaviour change goals, addressing the identified gaps in personalisation, responsiveness, and the integration of behavioural and mental health theories into LLM-based systems.

3.2 Understanding Users Writing About Change

The previous section has shown that many conversational BCSSs applying MI techniques do not take into account user attitudes, and those that do (e.g. Bickmore et al. (2011); Olafsson et al. (2020)) are mainly rule-based, heavily restrict user input and seem to focus mainly on the role of change talk versus sustain talk. A strategy to tailor CA responses to specific user states and concerns is introduced by Yao et al. (2023) in their rule-based postpartum support CA, which leverages classifiers trained on an existing dataset to identify the user's current state of mind. Combining such an approach with the capabilities of state-of-the-art LLM systems in the context of MI-based behaviour change support could drastically increase the potential effectiveness of MI-CAs, providing a control mechanism for CA responses through specific instructions and knowledge grounding, while at the same time strengthening the display of MI-Spirit. In this section, I explore the landscape surrounding the analysis and classification of help-seekers' language around behaviour change. I begin by outlining research regarding MI-specific resources, which predominantly focuses on the analysis of *spoken* human-human interactions. Since such interactions differ significantly from written human-CA interactions (Giles, Stommel, Paulus, Lester, & Reed, 2015; Hill, Ford, & Farreras, 2015; Rapp, Curti, & Boldi, 2021), this is followed by an overview of the related work about *written* accounts of behaviour change struggles and journeys.

3.2.1 Resources and Computational Methods for the Analysis of MI Client Language

The *MISC* defines a list of client codes that can aid the analysis of Motivational Interviewing sessions. In the context of CAs for behaviour change, these codes can serve as a proxy for the user's situation, their attitudes regarding change, and to what extent they intend or have already begun to make changes. Table 3.1 presents an overview of these client behaviour codes, together with simplified definitions and level names that I have introduced for clarity. Based on the *MISC* (W. R. Miller et al., 2008)¹, each client utterance is coded with a valence, which can be either change talk, sustain talk, or follow/neutral. If the utterance is not follow/neutral, it is then also coded with one of the following topics: reason, taking steps, commitment, or other. Lastly, utterances of type reason are also categorised into a reason type of ability, need, desire or general reasons.

In their literature review, Smink et al. (2019) focus on identifying the automation potential of different coding schemes in therapeutic change process research, a body of research that aims to draw conclusions on therapeutic outcomes based on within-therapeutic changes. Although they do not include MI in their review, the methods they identify as displaying promise for automation are chosen due to factors, such as the availability of code books, linguistic markers in the coding schemes, and their potential for generalisation, all of which are given for MI through the *MISC* and related coding schemes. However, research aimed at facilitating the analysis of MI sessions through the creation of relevant training datasets and computational methods has largely focused on therapist turns, which will be the subject of Section 3.3. Still, some research has examined the analysis of client language in the context of MI. Here, I will first describe efforts towards dataset creation, followed by reports of research regarding computational methods.

¹Other coding schemes in the context of MI, which are not used in this thesis project include the MITI and the MYSCOPE, both of which have significant overlap with the *MISC*.

Level	Label	Description
Valence	Change Talk (+)	Utterances in favour of behaviour change
	Sustain Talk (-)	Utterances in favour of status quo
	Follow/Neutral (FN)	Utterances not related to behaviour change
Topic	Reason (R)	Reasons for/against change
	Taking Steps (TS)	Specific steps taken in the recent past
	Commitment (C)	Agreement, intention, or obligation for the near future
	Other (O)	Utterance about behaviour change that does not fit into any of the other categories
Reason Type	ability (a)	Ability and degree of difficulty of the change
	need (n)	Necessity of change, or maintaining the status quo
	desire (d)	Desire for change, or current behaviour
	General (-)	General justifications, incentives, or justifications

Table 3.1: Simplified version of the *MISC* based on Meyer and Elswailer (2022); W. R. Miller et al. (2008).

3.2.1.1 Public Datasets with MI Client Annotations

There is a scarcity of publicly available data on MI sessions in general and annotated client language specifically, mainly due to the sensitive nature of such sessions and privacy concerns (Wu, 2023). This has led many research groups concerned with the topic to collect and experiment on their own data without sharing it with the community (Wu, 2023), preventing externals from utilising their data for their purposes (Atkins, Steyvers, Imel, and Smyth (2014); Borsari et al. (2014); Can, Atkins, and Narayanan (2015); Saiyed et al. (2022); Tavabi et al. (2020); Xiao et al. (2016), *inter alia*). Even in these cases, client codes are usually only annotated with valence information, foregoing topic and reason type labels.

To the best of my knowledge, only two publicly available datasets include client codes. One of them, AnnoMI, was introduced by Wu, Balloccu, Kumar, et al. (2023, 2022). It consists of transcripts of 119 high-quality and low-quality MI demonstration videos that span various target behaviours with utterance-level annotations of counsellor behaviour, as well as the valence of client utterances. Sun et al. (2024) go a step further with the BiMISC, a dataset collected from Dutch MI sessions. In addition to valences, they also annotate topic labels and apply multiple codes to each utterance when needed. In their paper, they include a comparison of an annotated utterance which in the AnnoMI was coded as follow/neutral, but due to the fine-grained annotation and permission to assign multiple codes per utterance applied in the BiMISC is revealed to contain an R+ statement, showcasing the limitations of utterance level and coarse-grained annotations in this sphere (see Figure 3.3).

Client Utterance	Annotation	
	BiMISC	AnnoMI
I was helped.	FN	
But I just have to keep it up because I still have to use medicines for a month and a half.	R+	NT
It is a kind of course, right?	ASK	

Figure 3.3: Example of difference between fine-grained and coarse-grained *MISC* annotation. Figure published under CC BY 4.0 license in Sun et al. (2024, p. 5609).

In addition to a general lack of resources, agreement between annotators represents a challenge in the context of MI². Sun et al. (2024) do not report inter-annotator agreement, merely noting that annotation was conducted by trained experts who flagged and discussed ambiguous cases and parts of their data were doubly annotated “ensuring the reliability and consistency of the annotations” (Sun et al., 2024, p.4). In Wu, Balloccu, Kumar, et al.’s (2022) work, annotation was done on a turn-by-turn basis by ten annotators, all part of the Motivational Interviewing Training Network (MINT), a network of MI trainers, with seven transcripts annotated by all raters to evaluate inter-rater reliability (IRR). IRR was measured using Cohen’s κ , a popular metric of inter-annotator agreement (Cohen, 1960). The resulting IRR was at $\kappa = 0.47$, which is considered moderate agreement according to (Landis & Koch, 1977), and weak based on the stricter interpretation suggested by (MgHugh, 2012)³, highlighting the high susceptibility to subjectivity when annotating MI-related content even among expert annotators.

Landis and Koch (1977)		MgHugh (2012)	
κ Range	Interpretation	κ Range	Interpretation
<.0	Poor	0-.20	None
.0-.20	Slight	.21-.39	Minimal
.21-.40	Fair	.40-.59	Weak
.41-.60	Moderate	.60-.79	Moderate
.61-.80	Substantial	.80-.90	Strong
.81-1.0	Almost Perfect	>.90	Almost Perfect

Table 3.2: Interpretation of κ scores according to Landis and Koch (1977) and MgHugh (2012).

To alleviate this challenge, Lord, Can, et al. (2015) compare different methods to calculate the reliability of utterance-based coding to improve IRR between annotators on *MISC* therapist and client codes, including topic labels. Annotators received training from MINT trainers in parsing and coding utterances from MI sessions. Even with this highly qualified training, the majority of client codes had an IRR of $\kappa < 0.5$, with only FN annotations reaching a score higher than $\kappa = 0.75$. Similar Kappa values were reported by M. Tanana, Hallgren, Imel, Atkins, and Srikumar (2016) (change/sustain: $\kappa = 0.6$, FN: $\kappa = 0.8$), Xiao et al. (2016) (change: $\kappa = 0.63$, sustain: $\kappa = 0.66$), Atkins et al. (2014) (κ between 0.5 and 0.75 for change and sustain talk), and Ewbank, Cummins, Tablan, Buchholz, and Blackwell (2020) ($\kappa \leq 0.55$ across coded topics and valences) in their reports of creations of datasets with valence annotations. In work by Borsari et al. (2014) additional measures are taken: Five raters underwent a rigorous training protocol, where they progressed through a library of role play and pilot audiotapes until a session-level Intraclass Correlation Coefficient (ICC) over 0.75 was achieved. They then annotated MI session videos over the course of multiple weeks, with weekly supervision meetings which addressed questions and specified decision rules when needed, as well as targeted training on items where agreement was low. This significant allocation of resources results in a high ICC for client valences (0.87 - 0.95). It is noteworthy, however, that in Lord, Can, et al.’s (2015) study, ICC is consistently *higher* than Kappa scores, and the two metrics are hard to compare directly. None of the research projects mentioned in this paragraph make their data public.

²This issue will be object of discussion again in Section 3.3.

³Table 3.2 provides an overview of how different Cohen’s κ scores are to be interpreted according to these two influential scientific publications.

3.2.1.2 Computational Methods for MI Client Annotations

Considering the focus on valence in the annotation of MI client language data, it is not surprising that many attempts to predict such language also do not go beyond the valence level. Still, a few projects also tackle fine-grained client language prediction. I will describe both of these subareas of research in turn.

Valence Only Prediction

Approaches used to classify the valence of client language sourced from spoken MI sessions are varied, spanning from traditional neural nets such as RNNs (Xiao et al., 2016) to transformer-based architectures such as RoBERTa (Y. Liu et al., 2019; Tavabi et al., 2021), with some also including multimodal features, such as loudness, and pitch (Aswamenakul et al., 2018; Tavabi et al., 2020).

As one of the first examples of client language classification, Atkins et al. (2014) use topic modelling to classify *MISC* client valences in a corpus assembled from five internal MI studies and compare classification results with agreement between human raters. Their approach was far less reliable than human-human agreement, with kappa scores around $\kappa = 0.2$ for both change and sustain talk ⁴. Similar results were reported by M. Tanana et al. (2016), who compared RNN and models using simple discrete sentence features for automated coding of MI sessions using the *MISC* scheme, finding low performance on sustain and change talk with Kappa values between $\kappa = 0.2$ and $\kappa = 0.3$, and Xiao et al. (2016), who reported Kappa values around $\kappa = 0.2$ reached by their RNN. Tavabi et al. (2021) compare the performance of transformer-based models (RoBERTa) with interpretable models (linguistic inquiry and word count) and find that, while RoBERTa works better, their interpretable model can help identify salient features.

In an effort towards the categorisation of client intentions, Huang et al. (2018) combine a Bi-LSTM predicting *MISC* codes with topic modelling to learn theme shifts over time and find that considering the temporal shifts within a conversation significantly improves the classification of utterances, with their best-performing model reaching a Macro F1 of 0.7. Tavabi et al. (2020) also find that adding preceding turns as context can lead to modest improvements in classification, noting a Micro F1 increase from 0.701 to 0.721, albeit with a standard deviation of 6%.

Cao et al. (2019), putting a focus on supporting counsellors in *ongoing* therapy sessions, make a distinction between the tasks of categorisation (classifying the last utterance, both user and therapist turns) and forecasting (predicting the next expected/suitable utterance). They use transformer-based architectures to model both tasks and report a maximum Macro F1 score of 0.539 for the categorisation task and 0.444 for the forecasting task. Although their approach does not yield high classification performance, the idea of distinguishing categorisation from forecasting can also be useful in the context of CA-interactions.

Finally, Tavabi et al. (2020) go one step further and attempt to predict behavioural outcomes based on client speech. They train two distinct classifiers to predict changes in blood alcohol content and changes in alcohol-related problems, respectively, as binary outcome variables. Their best model reaches an F1

⁴In the context of MI, classification performance is commonly reported with Kappa scores in addition to, or in lieu of F1 scores.

score of 57.8% on binary classification, and they find no trend in specific client valences having higher predictive power than others.

Fine-Grained Prediction

The majority of research projects including client topic and reason types in their classifiers do so in addition to the prediction of the therapist code, rather than focusing on the specific language of the client. Gibson et al. (2016), Can et al. (2015), and Saiyed et al. (2022) all train multi-class classifiers to make a prediction based on all *MISC* codes, including therapist codes. In the case of Gibson et al. (2016) and Can et al. (2015), who do not take into account reason types, this results in 28 classes for their classifier (where C+ and C- are each a separate class). Saiyed et al. (2022) also take into account reason types and their respective combinations with valences, resulting in 38 classes. In all cases, this leads to low performance, with Gibson et al. (2016) reporting a Macro F1 score of 0.258, Saiyed et al. (2022) simply reporting “insufficient model performance” due to a “lack of annotated *MISC* datasets” (p. 129), and Can et al. (2015) providing class-wise F1 scores, which show that their model performed exceedingly well for FN, likely the majority class ($F1 = 0.91$), but failed to reach F1 scores higher than 0.21 for the remaining client codes, with the F1 score for TS+ dropping as low as 0.03. Gibson et al. (2016) succeed in increasing their result to an F1 of 0.643 by grouping certain codes into categories. However, this results in all client codes being grouped into one class, which merely denotes that the turn was a client utterance as opposed to one of the multiple therapist classes.

Ewbank et al. (2020) go a different route and classify only client language. They base utterance classifications on the *MISC* 1.1 and define the following classes: “change talk active” (meaning desire or commitment or taking steps), “change talk explore” (meaning reflection of past, present or future state of mind), sustain talk, in this case, called “counter change-talk”, FN, and “describing problems”. Their LSTM is able to come close to human-level agreement for FN, change talk exploration, and change talk active, but performs much worse for describing problems and counter change talk. F1 scores range between 0.222 (counter change talk) and 0.94 (FN). They furthermore analyse the association between client utterances in a first MI session and therapy outcomes and find that high counter-change talk, describing problems, and follow/neutral are negatively correlated with reliable improvement, whereas both change talk categories are positively associated with improvement. The same was true for engagement, measured by a client’s willingness to return for a second session, with the exception of describing problems being positively associated with engagement. A completely different method is proposed by Raswsthorne, Cannon, and van Mierlo (2022), who suggest the use of keywords, based on descriptions and definitions of client codes in MI literature to identify change talk across topic and reason type codes from anonymous texts sourced from a digital health support network.

In most cases, the classification of client language seems to be merely a byproduct, instead of the research focus. Only very few sources go beyond the classification of valence and there is a lack of available resources that can be used for the classification of topics and reason types in addition to valence. Furthermore, the only publicly available datasets to include annotations of client language consist of transcripts of spoken MI sessions and video material. Since spoken human-human language differs significantly from a) written interactions (Giles et al., 2015) and b) human-CA interaction (Hill et al., 2015; Rapp et al., 2021), other means of data collection have to be explored for this context.

3.2.2 Analysing Written Behaviour Change Language

Written language around behaviour change and health has been analysed both qualitatively (e.g. Myneni, Cobb, and Cohen (2016); Rodda et al. (2018)) and computationally (Calvo, Milne, Hussain, and Christensen (2017); Smink et al. (2019)) spanning various contexts with differing severity, from the increase in daily step counts (Gupta et al., 2020, 2021) and sustainable behaviour (Fernández et al., 2016), to drinking behaviour (Carah, Meurk, & Angus, 2017) and gambling (Parapar, Martín-Rodilla, Losada, & Crestani, 2021; Rodda et al., 2018) to depression detection (Coppersmith, Dredze, Harman, Hollingshead, & Mitchell, 2015; Losada & Crestani, 2016; Losada, Crestani, & Parapar, 2019) and sentiment analysis of suicide notes (Pestian et al., 2012). Written sources for behaviour change analysis are just as varied, with Calvo et al. (2017) identifying social media sources, such as Twitter, Facebook, and blogs, but also personal diaries, and online support groups and forums as objects of NLP research. For instance, Kovanovic et al. (2018) use linguistic indicators of different psychological processes to create a classifier of reflective writing, based on Hulsman, Harmsen, and Fabrik's (2009) categorisation of different types of self-reflections (observations of own behaviour, motive, or effect of own behaviour, indicate a goal, and asking for feedback for improvement).

Others focus more on conversational data. Myneni et al. (2016) and Rodda et al. (2018) qualitatively analyse peer-to-peer communication on web forums for smoking cessation and gambling addiction, respectively, identifying different change strategies used and reported by people trying to change. Myneni et al. (2016) also identify common themes of such conversations, including quit obstacles and readiness, cravings, conflicts, and progress reports, and argue that the themes expressed in such forums can be related back to theoretical constructs, such as self-efficacy and observational learning, as well as various behaviour change techniques used in behaviour change interventions, based on Abraham and Michie (2008), such as prompting barrier identification and providing encouragement. Rodda et al. (2018) further grouped forum users into different stages based on the Rubicon model of action phases, a behavioural model similar to the TTM, that groups people into pre- and post-decisional phases of action and posits that users need different types of change support depending on the phase in which they are in (Heckhausen & Gollwitzer, 1987; Keller, Gollwitzer, & Sheeran, 2020).

Using a different methodology with a similar purpose, Carah et al. (2017) apply semantic text analysis, such as word frequencies and relationship between terms to a web forum encouraging users to stop consuming alcohol with the goal of exploring how user expressions change over time and identifying patterns of change. They find that forum participants tend to set goals and describe their current behaviour in the first month of posting, followed by reporting their efforts to change and the challenges they face, along with reflections of their current position and advice for other users. Fernández et al. (2016) put an even stronger emphasis on social aspects of change, as opposed to individual factors. They analyse tweets about climate change based on the 5 doors theory of behaviour change, which focusses on processes that lead to group adoption of behavioural changes and defines different stages of change. They explore how people in different behavioural stages communicate on social media, connecting specific changes with different sentiments and tweet content (i.e. sharing URLs in early stages, versus positive sentiment and vocative style in later stages). Based on their insights, they train classifiers to identify users' current stages and find that the majority of users they classify is in the desirability stage, which roughly corresponds to the contemplation stage of the TTM. Wu, Helaoui, Kumar, Recupero, and Riboni (2020) suggest an approach to leverage Reddit posts to train classifiers

on the task of detecting the need for empathetic responses.

Focusing on dyadic written conversation, Gupta et al. (2020) create a dataset of text-based human-human coaching conversations on step count increases with annotations based on SMART (Specific, Measurable, Achievable, Relevant, and Time-Bound) goals and stages (goal setting and goal implementation). In follow-up work, they use the same dataset to extract goal summaries (Gupta et al., 2021).

Other projects with a stronger focus on mental health instead of behaviour change include Losada and Crestani (2016), who build a resource for depression detection that allows for the analysis of progressions and changes in writing style over time, Pestian et al.'s (2012) shared task on the sentiment analysis of suicide notes, and Losada et al. (2019) and Parapar et al.'s (2021) shared tasks on depression, anorexia nervosa and gambling addiction detection from social media texts.

Section Summary

In this section, I have given an overview of the available resources for the analysis and prediction of MI client language, as well as related efforts to analyse written language around change. It becomes apparent that there is a scarcity of annotated data in the context, and the little data that does exist is based on language resources which cannot easily be translated to the context of written human-CA conversations. However, the analysis of texts written by people trying to change their behaviour based on different psychological constructs has proven promising, especially for data sourced from public self-help forums, where people interact with each other about their efforts to change, and past research was able to connect such language to processes and techniques of behaviour change. This indicates that applying MI client codes to written texts about behaviour change sourced from online platforms can potentially alleviate the challenges involved in data collection for MI client language. Such a resource could then be used as training data for computational methods to identify user attitudes about change and serve as a basis for choosing suitable system reactions.

3.3 Reacting

In this section, I start by discussing the ethical viewpoints associated with conversational BCSSs. This discussion will form the foundation for deriving the requirements that an LLM-based MI-CA should fulfil, as well as identifying appropriate and desirable behaviours for such a system. Following this, I provide an overview of current initiatives that utilise LLMs to generate MI-adherent text and conclude by presenting existing computational methods, annotation frameworks, and datasets used to assess CA behaviours in this context.

3.3.1 Appropriate and Desirable Behaviour

While some research suggests that users prefer real humans, over interactions with CAs (Morris et al., 2018; Vowels et al., 2024), the body of work summarised in Section 3.1 indicates that CAs seem to be capable of supporting users' mental health and behaviour change efforts. A study conducted by Ayers

et al. (2023) comparing physicians' replies to health advice questions on Reddit with GPT outputs to the same questions, even found that GPT-3 replies were more empathetic and of better quality, as judged by health care professionals. Further empirical findings indicate that users can feel equivalent intimacy and understanding when disclosing emotions to a chatbot as to a human, leading to emotional, relational, and psychological benefits (Ho et al., 2018). This intimacy is further increased if the system itself self-discloses to the user (Y. Lee, Yamashita, Huang, & Fu, 2020), revealing the high likelihood of anthropomorphism, when interacting with such systems, a concept that in the context of counselling CAs is sometimes framed as desirable (e.g. Kang and Kang (2023); Pentina, Hancock, and Xie (2023)), whereas ethics research generally views this as problematic (e.g. Abercrombie, Curry, Dinkar, Rieser, and Talat (2023); Ruane, Birhane, and Ventresque (2019), among others).

Such effects are likely to increase as CAs become more and more human-like, as indicated by studies of users of Replika, a commercial LLM-based social CA, which lets users choose the type of relationship they want to have with the system, and is highly popular with users, who build long-term relationships with the system and often use it for mental health support (Ma, Mei, & Su, 2023; Maples, Cerit, Vishwanath, & Pea, 2024; Pentina et al., 2023). Users of Replika cite various benefits of long-term interactions with the system, such as non-judgemental, on-demand support, promotion of self-discovery, and development of confidence for social interaction (Ma et al., 2023). Additional user-identified benefits include positive changes in their actions, way of being, and thinking, as well as decreased anxiety and increased perceptions of social support (Maples et al., 2024). Users report that Replika helped them strengthen relational bonds, by encouraging self-reflection, offering advice, and "helping [...] model positive interactions with people." (Pentina et al., 2023, p.5). Many perceive their interactions with the system as therapeutic, on par with a human therapist, and describe the chatbot as a mirror of self, with isolated reports that the system was successful at stopping suicidal ideation (Maples et al., 2024).

However, users also recognise challenges of using systems like Replika, such as its propensity to output harmful content, excessive reliance on the system for mental well-being, and social stigma of relying on a CA for well-being support (Ma et al., 2023; Pentina et al., 2023). Advanced Replika users are lonelier and more attached to the chatbot. With increasing interaction depth and closeness to the system, the degree of perceived humanness rises as well and social interaction with the AI can lead to emotional attachment, driven by Replika's ability to recognise patterns, remember facts, and perceive and express emotions (Pentina et al., 2023). This is particularly worrying considering Replika is not based on therapeutic principles, with the website offering only that "Replika combines a sophisticated neural network machine learning model and scripted dialogue content. It has been trained on a large dataset to generate its own unique responses"⁵.

Even when not used on a regular basis, LLMs can have an impact on users' opinions, as shown by Jakesch, Bhat, Buschek, Zalmanson, and Naaman (2023), who find that co-writing with opinionated language models has an effect on users' viewpoints. Similarly, Yeo, Lim, Gao, Zhang, and Perrault (2024) designed a platform that helps people reflect based on reflective nudging prompts generated by an LLM when interacting on online deliberation platforms. They find that the choice of reflectors the users are met with can influence the dynamics of the resulting online discussions. More generally, Kidd and Birhane (2023) outline various ways in which AI can distort human beliefs, including the implicit biases encoded in many GPT models. These shortcomings, along with challenges, such as well-

⁵<https://replika.com/>.

hidden untruths contained in outputs (frequently referred to as hallucinations), a lack of transparency and documentation of the training data, leading to a lack of explainability and controllability of model outputs, and privacy concerns, have been the object of countless ethical discussions (Bender et al., 2021; Bommasani et al., 2021; Dale, 2021; Floridi & Chiriatti, 2020; Weidinger et al., 2021). Furthermore, Curry and Curry (2023) discuss the moral consequences of AI-empathy, arguing that since such models cannot understand and thus do not accurately judge the appropriateness of an emotion, they should not be used for emotion regulation, and warning of the consequences of attempting to avoid any kind of “negative” emotion.

Generally, there are two aspects worth inspecting when defining appropriateness in the context of AI-based counsellor-client interactions: *empathy* and *adherence to counselling principles*. Below I outline each of these aspects, how they are conceptualised, and how they are treated in the context of NLP in turn. I end by describing to what extent these aspects should be redefined in the context of LLM-based MI-CAs based on the ethical considerations outlined above.

3.3.1.1 Empathy

The relevance of empathy in various academic disciplines has led to countless conceptualisations and a general fuzziness about its meaning in general discourse (Eklund & Meranius, 2021; Lahnala, Welch, Jurgens, & Flek, 2022; Yalçın, 2019; Yalçın & DiPaola, 2020). In their analysis of 52 studies from four authoritative disciplines in the context (neuroscience, philosophy, psychology, and nursing), Eklund and Meranius (2021) find the emergence of four core themes of empathy: Understanding (cognitive), feeling (emotional/affective), sharing (experiences, representations, and feelings; emotional), and self-other differentiation (emotional). Comparing these core themes with the definition of empathy often found in NLP, namely the “ability to understand another person’s feelings and respond appropriately” (Lahnala et al. (2022, p. 2139), as first defined in Leiberg and Anders (2006)), it becomes apparent that the core theme of understanding is covered, but the other aspects are conflated into an “appropriate reaction”.

According to Lahnala et al. (2022), both the generation and recognition of empathy have been popular fields of research within NLP for years, gaining increased popularity with the introduction of seminal benchmark datasets by Rashkin et al. (2019) (EMPATHETIC DIALOGUES) and Buechel, Buffone, Slaff, Ungar, and Sedoc (2018) (EMPATHETIC REACTIONS) (e.g. M. Y. Chen, Li, and Yang (2022); Y. Lee, Lim, and Choi (2022); Majumder et al. (2020); Welivita, Xie, and Pu (2021); Welivita, Yeh, and Pu (2023); Xie and Pu (2021)). However, they also note that in the majority of relevant studies, empathy is not sufficiently defined, and many studies simplify the concept as emotions or sentiment. Examples of this can be found in Hosseini and Caragea (2021); Majumder et al. (2020); and Welivita et al. (2021). Although understanding is a cognitive process, Lahnala et al. (2022) come to the conclusion that the NLP literature disproportionately focuses on emotional empathy. Only eight of the studies they analysed evaluated specific behaviours or strategies related to empathy, including counselling behaviours.

Furthermore, people have different expectations of empathy when interacting with chatbots compared to humans (Portela & Granell-Canut, 2017; Shao, 2023), highlighting the need for a specific definition of appropriate empathy in the conversational context of LLM-based conversational BCSSs.

3.3.1.2 Therapist Empathy and Adherence to Counselling Principles

Adherence to counselling principles can be operationalised with the help of theoretical frameworks defined within counselling styles such as MI. In addition to instructions for the annotation of client language, the *MISC* (W. R. Miller et al., 2008) and MITI (Moyers, Manuel, & Ernst, 2014; Moyers, Rowell, Manuel, Ernst, & Houck, 2016) also contain codes for therapist behaviour, as well as guidelines for the evaluation of therapist empathy at the session level. In the context of MI, the word empathy refers mostly to “understanding”, which is achieved through open questions and reflective listening (W. R. Miller et al., 2008), and does not encompass the emotional aspects of empathy definitions outlined above. Subsequently, this will be referred to as *therapist empathy*. Therapist empathy is, based on the *MISC*, coded at the session level, whereas therapist behaviours, the “appropriate reactions”, are coded at the utterance level, the same way as the *MISC* client codes. As already outlined in Chapter 2, in addition to empathy, ‘MI-Spirit’ encompasses the elements of autonomy, collaboration, and evocation. This can be achieved through open questions, affirmations, and reflections. Both MI-adherent and MI-inadherent behaviours are defined in the *MISC*. On top of the MI micro skills, MI-adherent behaviours include giving advice with permission and raising a concern, whereas criticism, advising without permission, and warning constitute MI-inadherent behaviours.

In the context of NLP, various research projects have shed light on markers of high-quality counselling, including the presence of MI-adherent behaviours (Hasan et al., 2019; Pérez-Rosas et al., 2018; Pérez-Rosas, Wu, Resnicow, & Mihalcea, 2019) which have also been used as proxies for the detection of therapist empathy (Gibson, Malandrakis, Romero, Atkins, & Narayanan, 2015; Lord, Sheng, Imel, Baer, & Atkins, 2015; Pérez-Rosas, Mihalcea, Resnicow, Singh, & An, 2017; Tavabi et al., 2023; Xiao, Can, Georgiou, Atkins, & Narayanan, 2012; Xiao, Imel, Georgiou, Atkins, & Narayanan, 2015). Similar to the literature related to the classification of MI client behaviour (see Section 3.2), these predominantly focus on the analysis of transcribed, spoken counselling sessions between two humans, with some also taking multimodal features into account (e.g. Pérez-Rosas, Mihalcea, Resnicow, Singh, and An (2017); Xiao et al. (2015)).

For instance, Pérez-Rosas et al. (2018) use a dataset of video recordings of counselling sessions which have been marked as high or low-quality counselling to show that classifiers using features like n-grams, semantic information, meta-features, such as number of turns and utterance lengths, sentiment, and the amount of questions and reflections voiced by the therapist can be used to distinguish between high and low-quality counselling sessions with high confidence ($F1 = 0.87$). Following this, Pérez-Rosas et al. (2019) report that high-quality counselling is marked by a more balanced speaking ratio with clients compared to low-quality counselling, as well as an increased use of reflective listening, positive sentiment, language mirroring (as measured through linguistic alignment), and a topical focus on behaviour change and commitment over resistance and persuasion. Similarly, Tavabi et al. (2023) find that MI-adherent behaviours are significantly correlated with therapist empathy ratings, especially for therapists who provide more complex reflections. However, Lord, Sheng, et al. (2015) find that language style synchrony (as measured by the shared presence of lexical features, such as adverbs, prepositions, and impersonal pronouns) between the client and the therapist predicts empathy better than simple reflection counts, indicating that mirroring the speech patterns of clients increases the perception of therapist empathy by third-party annotators. This could mean that the quality of reflections (its effectiveness in mirroring client sentiment) is more important than its mere presence.

Gibson et al. (2015) attempt to predict therapist empathy using psycholinguistically inspired language features, such as the mean age a term used is acquired, or arousal and Xiao et al. (2015) report tri-grams indicative of high and low empathy in counselling sessions. The majority of the tri-grams indicating high empathy have a reflective quality ("it sounds like", "do you think", "so you feel",...), while low empathy tri-grams are seemingly more prescriptive and oriented towards the past ("do you have", "you know what", "you have to", "have you ever", "the past twelve"). In a similar vein, Pérez-Rosas, Mihalcea, Resnicow, Singh, and An (2017) extract common words which they cluster into themes and find that high empathy is marked by forward-looking and goal-oriented topics (importance, concerns, reflections, change, goals, motivation and inform) whereas low empathy sessions seem to be marked by the therapist taking a more active role, as indicated by topic categories such as persuade, plan, and guide, as well as focusing on feelings and support.

Focusing on specific turns as opposed to longer interaction sequences, Hasan et al. (2019) use hidden Markov models to identify counsellor strategies which effectively elicit change talk. They find that reflections of change talk, especially when paired with a question and commitment language lead to more successful interactions, whereas reflections of sustain talk and perceived barriers, as well as providing information and other reflections were less likely to elicit change talk. Xiao et al. (2012) create an n-gram-based classifier to discern empathetic from non-empathetic utterances and obtain session-level empathy classifications, showing that session-level empathy perception is likely predicted by isolated highly empathetic utterances.

3.3.1.3 Defining Appropriate and Desirable Behaviour for LLM-based Conversational Behaviour Change Support Systems

Based on the literature outlined above, I argue that behaviours that increase anthropomorphism, such as self-disclosure of emotions or experiences, as well as behaviours that frame the CA as an expert or a source of authority, are not appropriate in the context of LLM-based MI-CAs and that this should be considered when designing and evaluating such a system. As such, it would not be appropriate to involve emotional empathy, i.e. displaying "feelings" and sharing "personal information" on the system side. Such behaviours should be avoided in favour of therapist empathy, focusing on conveying a sense of understanding, which can foster self-reflection through open questions and reflections of user utterances. As such, the remainder of this section focuses on literature specifically focusing on the generation of counselling-style text and evaluation of counselling behaviour.

In addition, only specific MI-adherent behaviours are seen as desirable in the context of LLM-based behaviour change support, namely those that do not tell users what to do, bypassing advice and raising concerns. Furthermore, instead of regulating emotions, a CA for behaviour change support should focus on helping users gain a *deeper* understanding of their attitudes, emotions, and behaviours, thus empowering them to decide to change. Considering the nature of MI, which focuses mainly on the client's perspective and works primarily through questions and reflections of client utterances, focusing on these specific aspects of social interactions is practical for the proposed CA.

3.3.2 Resources and Computational Methods for the Generation and Evaluation of Counselling-Style Text

To investigate the generation and evaluation of counselling-style text, it is essential to examine the foundational resources and computational techniques used in this context. This subsection begins by examining publicly available datasets with MI therapist code annotations, highlighting their sources and persistent challenges in IRR. It then transitions to discussing text generation techniques tailored for counselling contexts, covering approaches from fine-tuning language models to prompt-based systems. Finally, it addresses the evaluation of counselling quality, emphasising the methods used to assess the appropriateness and efficacy of generated text, both in spoken MI interactions and in written counselling-related contexts.

3.3.2.1 Public Datasets With MI Counsellor Annotations

The same challenges with respect to publicly available data outlined in section 3.2 are also present for datasets with MI therapist code annotations. Only few datasets exist, with variable IRR. As far as I am aware, there are five publicly accessible datasets in this area, three of which were sourced from spoken counsellor-client interactions (Pérez-Rosas, Mihalcea, Resnicow, Singh, & An, 2016; Sun et al., 2024; Wu, Balloccu, Kumar, et al., 2023, 2022), one from peer support fora and social media (Welivita & Pu, 2022), and one from Wizard of Oz (WoZ) conversations between humans and companionbots (Guntakandla & Nielsen, 2018).

The first dataset of this kind is presented by Pérez-Rosas et al. (2016), who base their annotations of fine-grained therapist codes on the MITI. Their data is based on commercially available video material of clinical trials and counselling sessions. Annotation was conducted by three counsellors, with additional training by MINT trainers. Still, IRR ranges between $\kappa = 0.31$ and $\kappa = 0.64$, with only questions being annotated with a Kappa score higher than 0.5. In Pérez-Rosas et al. (2018) this dataset is expanded with high-quality and low-quality session ratings. Although the authors claim they make their dataset available to the public, it is not linked in the article and I was not able to find it. Other research has since reported this dataset to be “not publicly accessible” (Sun et al., 2024, p. 5611). Similar work is presented by Wu, Balloccu, Kumar, et al. (2022) & Wu, Balloccu, Kumar, et al. (2023) (mentioned in Subsection 3.2.1.1), who publish a dataset of transcriptions of high-quality and low-quality MI demonstration videos with therapist code annotations across the classes reflection, question, input (e.g. giving information or advice), and other, with IRR as high as $\kappa = 0.74$ for the main therapist code. Most recently, Sun et al. (2024) published the BiMISC dataset based on Dutch counselling sessions translated into English, which, in addition to fine-grained client codes, also contains fine-grained therapist codes.

Focusing only on reflections, Guntakandla and Nielsen (2018) publish a corpus with more than five hours of transcribed WoZ-conversations between humans and companionbots with annotations for complex reflection, simple reflection, and no reflection, with Kappa scores ranging between $\kappa = 0.49$ for simple reflections, and $\kappa = 0.64$ for complex reflections. While this is a first step towards the analysis of human-CA conversations in this context, they do not annotate MI client codes. Furthermore, their data still consists of spoken interactions, which are ultimately produced by humans on both sides, even though users might not have been aware of this.

The only example for MI therapist code annotations of written counsellor utterances that I am aware of is presented by Welivita and Pu (2022) in the specific context of mental health. They base their data on a Reddit forum and counsel-chat, a forum where professionals respond to people seeking help online. They attempt to crowdsource MITI annotations using a multi-step approach, where they first use two annotators per dialogue and have them choose one of the available MITI labels. Provided the two raters agree, the label is accepted as correct; otherwise, a new rater is solicited, and the majority label is taken as gold label. In cases where the third rater suggests a different label than both initial raters, a fourth rater is called in. Although Fleiss' Kappa was low across all annotation stages ($\kappa \leq 0.56$), this approach led to 97.39% of the data being labelled with the same code by at least two raters after three rounds of annotations. In their later work, they expand this dataset by applying automatic labelling techniques to more data from the same source, based on n-gram-based matching and similarity-based retrieval (Welivita & Pu, 2023).

3.3.2.2 Text Generation for Counselling Contexts

A large body of work has focused on the fine-tuning of language models for the generation of empathetic or counselling-style text (Das et al., 2022; Saha, Chopra, Saha, Bhattacharyya, & Kumar, 2021; Wu, Balloccu, Helaoui, Recupero, & Riboni, 2022). Often, the datasets used for fine-tuning stem from peer-to-peer interactions (e.g. Morris et al. (2018); Shah et al. (2022)), leading to the generation of empathetic texts which include behaviours such as self-disclosure or expressions of emotions not suited for the context of conversational BCSSs, as argued in Section 3.3.1. Furthermore, the mere fine-tuning of LLMs as an approach to text generation reduces the amount of control that can be exerted on the language model in the context of a CA, since we cannot explicitly elicit specific types of behaviour in specific situations, instead having to rely on the model having learned suitable reactions from the data (J. Zhou et al., 2022). This limitation has been addressed in different ways. While some research focusses on creating systems to generate a specific counselling-style behaviour (e.g. A. Brown et al. (2023); Rose, Ahmed, Keilty, Cooper, and Selby (2022); Sharma, Rushton, et al. (2023); Wu, Balloccu, Helaoui, et al. (2022); Wu, Balloccu, Reiter, et al. (2023)), others employ different methods, such as conversational stages (J. Zhou et al., 2022), prompting methods, and in-context-learning (e.g. Fu et al. (2023); Y. Lee et al. (2022)) to elicit targeted responses.

Research focusing on the generation of one specific counselling-style behaviour type predominantly revolves around reflection generation. For instance, Wu, Balloccu, Helaoui, et al. (2022) use LLMs fine-tuned on the AnnoMI to generate counselling-style reflections. In a follow-up work focusing on the evaluation of generated reflections (Wu, Balloccu, Reiter, et al., 2023), they find that GPT-3 mostly performs well for this task in terms of coherency and context-consistency. It should be noted, however, that the AnnoMI consists of transcripts of spoken MI sessions, which may differ from the way people interact in written settings. Also focusing on the task of reflection generation, Rose et al. (2022) explore few-shot generation capabilities of GPT-2 and GPT-3 based on data collected in Almusharraf et al.'s (2020) experiments for the identifications of reasons for and against smoking through CA interactions. They find that GPT-3, given 8 few-shot examples provided in a counsellor-prompt-response-reflection format, produces high-quality reflections 89% of the time, whereas GPT-2 only reaches a native success rate of 43%. However, employing a BERT-based classifier to filter out low-quality reflections, they succeed in increasing the GPT-2 success rate to 73%. Addressing a different behaviour, Welivita and

Pu (2023) fine-tune Blender and GPT-3 to transform “advise without permission” into “advise with permission” statements, which contain additional politeness markers making the advice sound less directive (such as “I wonder if you can” or “you may want to consider”).

Focusing specifically on the direct application of text generation in counselling contexts, Shen, Welch, Mihalcea, and Pérez-Rosas (2020) develop a dialogue system that supports counselling trainees by generating appropriate reflections to help them improve their reflective listening skills. In their work, they compare different text generation approaches, coupling GPT-2 fine-tuned on a MI dataset with synonym replacement, retrieval algorithms, and content expansion. Similarly, Min, Pérez-Rosas, Resnicow, and Mihalcea (2023) introduce a system to rewrite counsellor responses to increase MI-adherence (i.e. rewrite an advice statement into a reflection statement). The system uses the counsellor’s response, filters out tokens that do not fit into a reflection, and uses the resulting template to create a reflection. As an example of an end-user-facing system, Sharma, Rushton, et al. (2023) developed a non-conversational tool that prompts users to enter their negative thoughts and classified thinking traps before providing a list of positive reframes. Based on the system, which was rolled out to users, they were able to explore what kind of reframes users find most helpful. J. Zhou et al. (2022) combine two fine-tuned text generation models: they fine-tune GPT-2 for the generation of empathetic responses by adding the response communication mechanism as special tokens to their training data, made up of the dataset of SMS coaching conversations concerning SMART goal setting for step count increases collected by Gupta et al. (2020). They also fine-tune T5, adding dialogue state information to the input sequence during fine-tuning and inference. Depending on the emotionality of the user’s utterance, either the empathy generation model or the other model is used to generate a response.

In the context of peer counselling, Sharma, Lin, Miner, Atkins, and Althoff (2023) create a system that, based on a peer supporter’s response draft, provides just-in-time feedback to make the post more empathetic by removing or replacing sections in the text. They find that collaborating with language models that provide advice on how to improve messages to help-seekers can help peer supporters become more empathetic. Not only are they able to show a linear relationship between empathy and intensity of use of the system during conversations, but participants also reported that they felt more capable of being effective peer supporters in the future after using the system, indicating that users can learn from the application of such systems. Also focusing on the training of peer counsellors, Hsu et al. (2023) go a different route and attempt to generate different types of responses by fine-tuning a model on peer support data created in previous work conducted at their institute (Shah et al., 2022). During training, they append the type of counsellor behaviour used to each training sample. They then create a system that, given the utterance of a support seeker, classifies a suitable response. The response type is appended to the text passed to the model for text generation. The generated text is then given to peer supporters as a response suggestion. In a related effort, Fu et al. (2023) create a prompt-based system using GPT-3.5 intended to support non-expert counsellors to provide better mental-health support. The system gives feedback on help-seekers’ psychological concerns and cognitive distortions and evaluates the response provided by the supporter, providing professional knowledge and potential improvements. It also recommends a type of psychological therapy and backs up its evaluations with references. In a preliminary study, a team of trained counsellors evaluated 30 example outputs by the system, based on the accuracy of its analyses and recommendations (given a choice of yes, no, and not sure). While the counsellors’ judgements were predominantly positive, the example given as the output of their model in the paper contains a reference to an article that does not exist, highlighting the high perceived

competence of these models, that seems to be able to fool even experts.

3.3.2.3 Evaluation of Counselling Quality

As with any other use case of machine-generated text, quality assessments are essential to determine the efficacy of use. In many cases, the quality of generated texts is assessed through automatic metrics such as similarity to empathetic utterances in benchmark datasets such as *EMPATHETIC DIALOGUES* (Rashkin et al., 2019), evaluated through - in many cases crowdsourced - human ratings on Likert scales, or A/B testing (e.g. Y. Chen and Liang (2022); Gao et al. (2021); Zhong, Zhang, Wang, Liu, and Miao (2020); L. Zhou et al. (2020)). However, such assessments should also take into account the correct application of counselling principles, which, due to the high degree of subjectivity and the low reliability of annotations even when trained experts are involved, is hard to crowdsource. Furthermore, the lack of suitable datasets of written interactions between human counsellors and clients renders a comparison to a gold benchmark unfeasible. Below, I summarise existing efforts towards the automatic evaluation of counselling-style text in different contexts. Again, I focus first on related work applying classification algorithms to spoken MI interactions between humans, followed by an overview of systems focused on written texts.

Evaluating Appropriate Behaviour in Spoken MI Interactions

As opposed to MI client codes, which have not been a regular object of NLP research, the classification of MI therapist codes has been a long-standing topic of interest. As with client code classification, this is usually based on MI session transcripts, sometimes including acoustic components, as seen in e.g. Pérez-Rosas, Mihalcea, Resnicow, Singh, and An (2017); Tavabi et al. (2020); Xiao et al. (2014).

For instance, Can et al. (2016) use different features, such as client-counsellor similarity, *n*-grams (e.g. “it sounds like”), and metafeatures (e.g. whether the previous utterance was a client utterance, or whether the previous or subsequent therapist turn is a reflection) and contextual *n*-grams to detect counsellor reflections. Based on their results, the client-counsellor similarity is not indicative of reflective listening ($F1 = 0.425$), however, *n*-grams in combination with other features serve as good predictors ($F1 = 0.80$).

More comprehensively, Pérez-Rosas, Mihalcea, Resnicow, Singh, An, Goggin, and Catley (2017) use the dataset introduced by Pérez-Rosas et al. (2016). They extract linguistic features and build classifiers to predict whether an utterance is a reflection or not ($F1 = 0.67$, and $F1 = 0.84$, respectively) and, if it is a reflection, whether it is simple or complex ($F1 \sim 0.63$). They also build classifiers for questions, again first distinguishing questions from other utterances, followed by a second classification into open or closed questions, reaching high classification performance and a multi-class classifier for other counsellor behaviours, which proves more challenging. Xiao et al. (2016) group therapist codes into categories, including MI-inadherent and MI-adherent to reduce the number of classes and facilitate classification and Imel et al. (2019) design a system that gives automated visualised feedback on MI-adherent behaviour, indicated by the reflection to question ratio, and the percentage of open questions and complex reflections, session-level ratings of MI spirit and therapist empathy for spoken counselling sessions, as well as talk-ratios of client and therapist. Gibson et al. (2022) explore a combination of *MISC* labels with similar labels derived from CBT for multi-label classification to gain even further

insights into therapist behaviours.

Taking this further, Flemotomos et al. (2022) describe a tool intended to give instant automated feedback on MI skills during therapist training. The tool informs practitioners about their usage of MI-adherent and MI-inadherent skills, their use of questions and reflections, and the ratio of talk times. Similarly to Xiao et al. (2016), they group certain behaviours (e.g. advise with permission is grouped into MI-adherent behaviours). They tested the tool in real-world settings in the context of therapist training sessions. However, their classifiers are of variable reliability, with F1 scores ranging from 0.158 (MI-Inadherent) to 0.956 (Facilitate). Their classifier performance largely mirrors the IRR they achieved during dataset annotation. They measure IRR using Krippendorff's alpha, reporting results ranging from $\alpha = 0.12$ for affirmations to $\alpha > 0.89$ for questions (simple and complex reflections are at $\alpha = 0.26$ and $\alpha = 0.47$, respectively). This indicates that the subjectivity of a code can serve as a proxy for classification performance. Finally, Sun et al. (2024) explore different prompting and few-shot learning techniques for the classification of MI therapist codes, including the provision of the *MISC* manual for classification. Although they achieve high scores for high-level codes ($F1 > 0.74$ with GPT-4), fine-grained classification poses a challenge, with Macro F1 reaching only 0.31 across therapist labels.

Evaluating Appropriate Behaviour in Written Counselling-Related Contexts

EPITOME is a computational method (Sharma, Miner, Atkins, & Althoff, 2020) that measures empathy across three dimensions: Interpretations (IP), Explorations (EX), and Emotional Reactions (ER), where IP encompasses the communication of an understanding of feelings and experiences voiced by a help seeker, EX means utterances aimed at improving the supporter's understanding of the help-seeker's situation, and ER refers to the expression of emotions, such as warmth, compassion, and concern. The creators provide code for their RoBERTa-based classification model, trained on annotated texts sourced from Reddit and a peer-to-peer support platform. Also focusing on empathy in peer interactions, Welivita and Pu (2020) introduce a taxonomy of eight empathetic intents, such as questioning, acknowledging, consoling, encouraging, sharing own experience, and others, along with an annotated dataset. Z. Wang and Torres (2022) introduce a dataset with annotations of the helpfulness of Reddit support forum posts, with annotations originating directly from help seekers. Although these approaches are focused on peer support specifically, applying classifiers based on them to the evaluation of LLM-based conversational behaviour change support conversations can still have benefits, for instance, in helping to understand how the language model interprets specific prompts and conversational contexts. Furthermore, they can help identify unsuitable behaviours (e.g. emotional reactions, sharing or relating to own experience, etc.), potentially serving as quality control for the conversations with the system and the creation of annotation schemes.

Focusing on MI counselling behaviours specifically, M. J. Tanana, Soma, Srikumar, Atkins, and Imel (2019) develop a chatbot to act as a psychotherapy client for the training of counsellors. The chatbot provides real-time feedback on trainees' use of open questions and reflections based on classification results. Participants using the chatbot used 76% more reflections after using the system than participants who did not use it. However, the use of open questions did not increase after usage. Similarly, Min, Pérez-Rosas, Resnicow, and Mihalcea (2022) create a system to score the quality of reflections in the context of counsellor training. To achieve this, they compile a dataset of client utterances that are likely to prompt a reflection, which they then pair with non-reflections, simple reflections or complex

reflections, representing low, medium, and high-quality reactions, respectively. Shah et al. (2022) fine-tune Bert-based classifiers for the identification of MITI counsellor responses in written peer-to-peer counselling interactions, reaching an F1 score of 0.725 for fine-grained therapist codes, with class-wise scores ranging between 0.523 (direct) and 0.927 (introduction). They also include new classes to address specific behaviours present in the given online context: support, which describes general sympathy and compassion, inappropriate, and chit-chat, including jokes and other behaviours not expected in a professional therapeutic setting. Welivita and Pu (2023) also build classifiers of fine-grained MI therapist behaviours based on data introduced in Welivita and Pu (2022) augmented through n-gram-based and similarity-based matching of new data from the same source, reporting accuracies of up to 72.76%.

In the context of LLM-generated texts for counselling contexts, Wu, Balloccu, Helaoui, et al. (2022) identify categories of errors made by LLMs in reflection generation as a basis for crowd-sourcing annotations of such errors. Wu, Balloccu, Reiter, et al. (2023) use these categories to have laypeople annotate the coherence and context-consistency of such generated reflections. Finally, BOLT (Chiu, Sharma, Lin, & Althoff, 2024) is a computational framework for the assessment of behaviours displayed by AI therapists. The authors use prompt-based LLMs for the classification of various therapist behaviours, focusing on the context of mental health. Some behaviours, as well as definitions of what constitutes high- and low-quality therapy, are drawn from the MI literature, while the remaining behaviours to be classified stem from various sources. At the highest level, they can be distinguished into reflections, questions, psychoeducation, solutions, and normalising, where solutions are more indicative of low-quality therapy, and reflections, questions, and normalising are indicative of high-quality therapy. They use this framework to evaluate the capabilities of LLMs in this context compared to low- and high-quality human therapy, finding that LLMs tend to provide more reflections on needs, emotions, consequences, and strengths than high-quality human therapy while asking fewer questions about experiences and emotions. They also find that LLMs engage more in problem-solving and planning by providing solutions, tend to normalise more and deliver more psychoeducation. To my knowledge, there is no evaluation framework focusing specifically on the identification of undesirable and harmful behaviour as outlined in Section 3.3.1.3 in the context of LLM-based MI-guided conversational behaviour change support.

Section Summary

This section has discussed the ethical perspectives, requirements, and behaviours relevant to conversational BCSSs, particularly focusing on LLMs and their application in MI-CAs. It has highlighted the growing acceptance of CAs for mental health support and behaviour change, referencing empirical studies that show the potential benefits of such systems, including emotional, relational, and psychological gains. However, it has also addressed concerns related to anthropomorphism, privacy, and the ethical implications of using LLMs in therapeutic contexts. Based on this, I have outlined the importance of appropriate and desirable behaviours in CAs, emphasising empathy and adherence to counselling principles, and discussed existing efforts and challenges in generating MI-adherent text and evaluating CA behaviours.

Furthermore, the section has given an overview of the available resources and computational methods for generating and evaluating counselling-style text, discussing publicly available datasets with MI counsellor

annotations and the complexities involved in fine-tuning LLMs for counselling contexts. It has examined various approaches to text generation for counselling, highlighting the challenges in controlling LLM outputs to ensure appropriate and beneficial interactions. Additionally, it has explored methods for evaluating counselling quality, including system-level and feature-level assessments and the classification of MI therapist codes. The section has underscored the need for comprehensive evaluation frameworks to assess both the beneficial and potentially harmful behaviours of LLM-based CAs in counselling scenarios.

3.4 Chapter Summary

In this chapter, I have given an overview of the related work surrounding conversational BCSSs. I began by outlining the landscape of existing CAs, both rule- and retrieval-based and generative. This has revealed a number of drawbacks of classical CAs, namely their lack of interactiveness, often not allowing free-text inputs, and responsiveness, failing to understand user attitudes about change. This directly opposes the spirit of MI, which is based on a strong focus on client expression and reflective listening. Conversely, LLM-based chatbots either lack theoretical grounding or have not been tested on end users due to a fear of low controllability and potential harm (Section 3.1). Evidently, one of the reasons for the lack of responsiveness to user attitudes towards change displayed by current systems is the lack of openly available and annotated datasets of user utterances, especially in written contexts, uncovered in Section 3.2. This has hampered the creation of computational methods to detect what situations users are currently in and how this is reflected in their attitudes. The last part of the chapter has focused specifically on the LLM-based generation of suitable responses, including ethical considerations and a redefinition of appropriateness in the context. Over the course of the chapter, I have identified several research gaps:

- Existing CAs for behaviour change support do not lay enough emphasis on understanding user attitudes about change.
- Computational methods for the classification of user attitudes towards change largely focus on identifying high-level codes (change talk, sustain talk, follow/neutral) based on spoken interactions due to the scarcity of data, including fine-grained annotations of client utterances in written texts.
- LLM-based counselling agents predominantly focus on mental health support or offering feedback to counsellors. The way users interact with LLM-based conversational systems in longer conversations in the context of behaviour change remains unexplored.
- Existing computational frameworks for the evaluation of MI adherence are not suited for the evaluation of LLM-based behaviour change support conversations, and the only existing framework for the evaluation of AI therapists does not focus on the identification of undesirable and harmful behaviours in the context.

The aim of this thesis is to address these gaps.

3.5 Research Design

The research conducted in this thesis operated between various disciplines and combined insights and techniques from psychology, NLP, information science, and human-computer interaction. Over the course of the thesis project, I employed both system-oriented and user-oriented methodologies and lenses to address the research goals provided in Chapter 1.

As outlined in Section 2.2, understanding the client is an integral part of MI and effectively supporting behaviour change. It enables practitioners to decide on appropriate reactions and the next courses of action and makes the client feel heard and appreciated. High-quality reflections of client utterances are based on a thorough understanding of the client's words and their deeper meaning, enabling the client to understand themselves and their reasons for change better (W. R. Miller et al., 2008; W. R. Miller & Rollnick, 2002).

Part II of this thesis is about synthesising such an understanding with the help of NLP techniques. To achieve this, I borrow MI concepts to help analyse user self-reports and written words in order to unveil their states of mind and level of preparedness to change their behaviour. This information will later be used by the CA to enable suitable reactions tailored towards the user's current motivational level and situation.

Chapter 4 presents the first steps towards enabling an automated understanding of user attitudes towards behaviour change from written language by leveraging naturalistic text data from a German weight loss forum. The chapter is focused on the following research question:

RQ 1.1: Can MI client codes be applied to written text to gain an understanding of user attitudes about change?

RQ 1.2: Can computational methods be developed to enable conversational agents to make use of such insights?

Following this, Chapter 5 evaluates to what extent these computational methods transfer to different target behaviours (the intended behaviour change) and conversational contexts (e.g. different levels of synchronicity), addressing the following research questions:

RQ 2.1: Is the behaviour change language used and the insight provided by the language consistent across different conversational contexts and target behaviours?

RQ 2.2: Which client codes show the most dissimilarity in different conversational contexts and target behaviours?

RQ 2.3: To what extent do conversational context and domain restrict classification transferability?

RQ 2.4: Which lexical features provide the most insight into user attitudes in written utterances around behaviour change?

Through the construction and annotation of various smaller datasets from various conversational contexts and domains, I show that the classifiers trained on this dataset can be used to classify data from different domains and contexts and are feasible to use in a chatbot scenario. Finally, in Chapter 6, I outline experiments conducted in an attempt to improve classification performance to increase the

effectiveness and efficiency of the resulting CA.

Following the system-oriented methodologies used in Part II, Part III employs a user-oriented lens to focus on the CA's reactions. Chapter 7 describes a user study aimed at testing the efficacy of LLMs as behaviour change support CAs, revolving around the following research questions:

- **RQ 3.1a:** Can a LLM-driven conversational agent, prompted to act like a therapist, be used safely as a tool to foster self-reflection about behaviour change?
- **RQ 3.1b:** Does the use of a conversational framework as theoretical grounding improve MI adherence and the relationship between the user and the chatbot?
- **RQ 3.2:** Is there an association between user behaviour and conversation outcome?
- **RQ 3.3:** What can we learn from applying classifiers based on the GLoHBCD to LLM-based conversations around behaviour change?

This includes the design of a conversational framework that incorporates MI principles and uses the classifiers created and evaluated in Part II, the evaluation of the resulting outputs with respect to potential harms, and the measurement of the effects of interactions with the CA on users' readiness to change. The study design also allows for the qualitative analysis of user behaviour and expectations in interactions with an LLM-based BCSS, and to what extent these impact users' perceptions and outcomes of the interactions.

PART II

UNDERSTANDING

Collecting and Evaluating Naturalistic Written Data on the Language Around Health Behaviour

A first step towards classifying user-generated text around behaviour change in the German language was the collection of relevant naturalistic data. MI is traditionally a spoken intervention, whereas the proposed chatbot is intended to work on a textual basis. As shown in Section 3.2.1.1, the few existing datasets that include client code annotations consist of English and Dutch transcripts of spoken MI interactions, none of which were publicly available at this stage of this project. In addition, written and spoken interactions bear a number of differences that mitigate the potential of transferring classification from one context to the other. For instance, written utterances can be edited before they are sent, are generally archived and thus easy to refer back to, are in some cases visible to the public, and, depending on the software or media used, follow specific conventions and restrictions (Giles et al., 2015). Due to these differences between spoken and written interactions and the lack of corresponding data, this chapter aims to address the following research questions:

RQ 1.1: Can MI client codes be applied to written text to gain an understanding of user attitudes about change?

RQ 1.2: Can computational methods be developed to enable conversational agents to make use of such insights?

I am not aware of any German language text corpora about behaviour change. For this reason, I collected data from a German weight loss forum as a first resource of text about behavioural change. The context of weight loss represents a suitable starting point for collecting such data, since weight loss success depends on multiple changes to daily life to be made and sustained over a long-term period (Elfhag & Rössner, 2005). These changes include, but are not limited to, increased physical activity and improved nutrition. Additionally, weight loss does not have an easy step-by-step guide and thus represents a nontrivial case of behaviour change that might look different for everyone and positive effects of the change are usually delayed (weight loss does not happen instantaneously as soon as behaviour is changed). Consequently, this context is likely to lead to varied language around behaviour change and depict shifting levels of motivation at different stages of change.

Online forums have been used in the past to gain insights into conversations around behaviour change (Carah et al., 2017; Myneni et al., 2016; Rodda et al., 2018) and are a suitable resource for various reasons:

- they represent conversations between multiple people, which provides for heterogeneous language use not specific to individuals (Baldwin, Cook, Lui, MacKinlay, & Wang, 2013)
- the anonymity and amount of social support users encounter allows them to be open about their thoughts and emotions (Hasler et al., 2014; Strand, Eng, & Gammon, 2020)
- the forums are often used over longer time frames, showcasing different motivational levels and stages in a person's behaviour change (Carah et al., 2017)
- they are a naturalistic data source of written conversations

In this chapter, I describe the creation and analysis of the GLoHBCD, a dataset of German texts about behaviour change with a focus on weight loss. I describe the main characteristics of the dataset and outline preliminary experiments I conducted to evaluate the feasibility of using the dataset to classify written change language in this and other contexts.

This chapter is based on the following publication, which has been published under CC BY 4.0 license:

Selina Meyer and David Elswiler. 2022. "GLoHBCD: A Naturalistic German Dataset for Language of Health Behaviour Change on Online Support Forums." In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2226–2235, Marseille, France. European Language Resources Association.

Modifications have been made to the original content.

4.1 Data Collection

To begin, I chose two subforums from Germany's largest obesity forum "adipositas24.de" for analysis. When choosing suitable subforums, I paid attention to exclude any subforums which specifically focused on losing weight with bariatric surgery, in order to get utterances about traditional weight loss that necessitates the behaviour changes mentioned above. The smaller sub-forum ("*Psychological Therapy*"), dealing with psychological therapy in the process of weight loss, consisted of 632 posts distributed over 54 threads at the point of data collection in October 2020. The larger subforum ("*General*"), dealing with general discussions about weight loss without surgery, had 6578 posts divided into 231 threads. All posts were written between May 1st 2006 and July 31st 2020. The longest thread of *Psychological Therapy* had 57 entries, while the longest thread of *General* encompassed 1046 posts.

Together, both subforums had 1181 active users (users who authored at least one post). Due to the subject of the forum, it is to be assumed that all active users were either currently pursuing active weight loss, considering the idea of losing weight or have pursued active weight loss (and possibly succeeded in their endeavour) in the past. It is also assumed that all active users have a history with severely increased weight or are currently affected by overweight or obesity. Due to these assumptions, this data

source was selected as a decent source of naturally occurring language around behaviour change.

The most active user of *Psychological Therapy* had authored 32, the most active user of *General* 432 posts. In both cases, the distribution of posts over users produces a long tail, with only a few very active users, while a majority of the users have authored only a very small number of posts (see Figure 4.1).

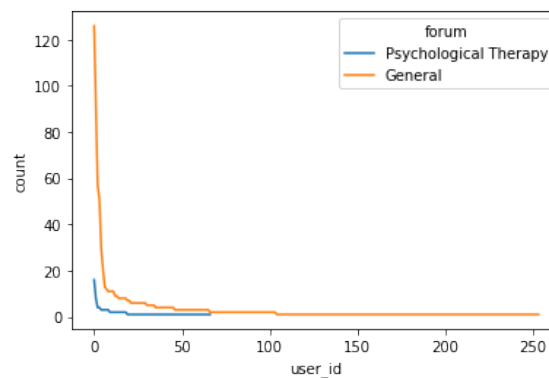


Figure 4.1: Distribution of posts per user in different subforums.

Upon detailed examination of the crawled posts, the sub-forums seemed to serve two main purposes:

- (1) *To share and receive practical and factual information and strategies*
- (2) *To seek or provide emotional support, share motivation and successes*

This is in line with the literature around self-help forums. For instance, Reifegerste, Wasgien, and Hagen (2017) differentiate between informational support (asking specific questions) and emotional support (sharing opinions and providing support for others) on obesity forums. In their analysis of self-help forums, Bauer, Bauer, Spiessl, and Kagerbauer (2012) identify the sharing of emotions and discussion of daily struggles as the main interests of forum participants. In the context of this work, my focus was specifically on posts in which users shared experiences and emotions with other participants, thus posts related to (2) would initially be interpreted as more relevant. In many cases, both purposes were found to be strongly intertwined and could not always be clearly separated. Still, the majority of the posts did not contain any talk around motivational factors of behaviour change and could thus be described as solely serving purpose (1).

4.1.1 Annotation

After screening, 1203 posts were identified and obtained for more in-depth analysis. I split these posts into sentences based on punctuation marks, after manually marking inappropriate punctuation (e.g. after abbreviations). Traditionally, the *MISC* instructs to annotate by *volleys*, semantically complete parts of speech, disregarding sentence boundaries (W. R. Miller et al., 2008). Since this would have considerably complicated automatic parsing, I decided to annotate on a sentence-to-sentence basis.

Data cleaning to ensure a smooth splitting into separate sentences was largely conducted by identifying suitable regex statements, based on abbreviations used by different users as well as forum-specific abbreviations. For instance, users had various differing approaches to abbreviating the German words “zum Beispiel” (for example), including “z.B.”, “zB.”, and “z. b.”. Such instances were identified and changed to “zb”. Following this, sentence boundaries were identified through the use of full stops. Overall, the cleaned and split dataset consisted of 15,533 sentences.

Topic		General Description	Keywords	Valence - Description
Reason (R)		rationale, basis, incentive, justification or motive		+ - Reasons for behaviour change - Concern about current situation/behaviour - Possible advantages of changing behaviour - Disadvantages of the current situation
				- - Reasons against behaviour change - Advantages of the current situation - Possible disadvantages of the behaviour change
Reason Type	Desire (d)	Desire or will	"want", "desire", "like", "wish"	+ - Desire for change - Will to achieve a certain goal - Preferences for unhealthy things/behaviour - Desire to maintain behaviour
	Ability (a)	Ability, degree of difficulty of the change	"can", "possible", "willpower" "ability", "hard"	+ - Confidence in own abilities to achieve the goal. - Statements that the behaviour change is easy - Statements about the difficulty of changing behaviour. - Pessimism about own ability to change
	Need (n)	Need or necessity	"need", "must"	+ - Necessity of behaviour change - Necessity of a goal - Necessity of the negative behaviour
				- - Planned steps towards behaviour change
Commitment (C)		agreement, intention or obligation regarding future behaviour	"will", "promise", " intend"	+ - Planned steps away from change - Intentionally unplanned steps towards the behaviour change
Taking Steps (TS)		Specific steps that have been taken in the recent past	"have", "yesterday", "I did..."	+ - Specific steps taken towards behaviour change in the recent past - Actions that have been taken and go against the behaviour change
Additional Categories used in original annotation (not included in inter-rater reliability and experiments)				
Other (O)		Utterances that reflect movement towards (+) or away (-) from behaviour change but do not fit into other categories		
Follow/Neutral (FN)		Statements unrelated to behaviour change. Questions, reporting, non-committal statements, small talk		
Multiple		Sentences containing multiple categories or valences were annotated with all codes that applied		

Table 4.1: Detailed annotation scheme adapted from *MISC* (W. R. Miller et al., 2008) for the annotation of written, unmediated text data about behaviour change used to create the GLOHBCD.

I annotated each of the resulting sentences with content categories and valences based on a slightly adapted version of the *MISC* client codes (W. R. Miller et al., 2008). The changes made to the *MISC* serve to account for the use case of unmediated online conversation as opposed to therapist-mediated oral conversations. For a detailed description of the annotation scheme, see Table 4.1. An overview of the code distribution with examples from the annotated data for each class is presented in Table 4.2. Allusions to past behaviour changes which did not draw a line toward the user's current situation were annotated as follow/neutral (FN) since I wanted the resulting dataset to reflect language around *ongoing or planned* behaviour change only. More than two-thirds of the data were coded as FN and 64.8% of the remaining sentences were annotated as R or one of the reason types. The large share of FN and R in the dataset is in line with results found in the literature (Lord, Can, et al., 2015), where more than 80% of client utterances in annotated MI sessions were annotated as FN, with R and O appearing most frequently among the remaining utterances.

4.1.2 Data Selection

Due to the sentence-to-sentence annotations, I encountered some sentences that contained multiple categories or valences. These instances, however, only made up 2.1% of the complete dataset. Due

Topic		Valence	Example	N	Share (%)
Reason (R)		+	It's making me insecure to "present" myself in public (358829/25)	1339	13.9
		-	My family loves me just the way I am (484073/8)	765	
Reason Type	Desire (d)	+	I long for exercise, a life with less pain, more power, without scorn and ridicule from society and, above all, a feeling of being satiated, to banish food from the centre of my life (358829/49)	239	1.9
		-	I'm craving chocolate and biscuits, even though I'm generally not a sweet eater (741454/26)	43	
	Ability (a)	+	I've already made it this far on my own, then I'll be able to do the rest on my own too... (611970/16)	134	3.3
		-	Well, if it was that easy for me to change my diet, I would have done it of course (698138/1)	349	
	Need (n)	+	But now I need to develop strategies with which I can overcome even these times (893200/13)	181	1.2
		-	Knowing that I have food like that in my kitchen makes me crazy and I need to eat it (1895006/9)	10	
Commitment (C)	+	Tonight I'm going to do 30 minutes on my cross trainer (674331/16)	433	2.9	
	-	Tomorrow is my daughter's birthday, so I'll definitely be feasting on cake - I'll just treat myself (484969/7)	21		
Taking Steps (TS)	+	I also started doing a lot of sports (498588/5)	950	8.0	
	-	since yesterday i've been eating like crazy again...(716757/5)	260		
Other (O)			Nevertheless, in my opinion it is just as you say, until now I was like an alcoholic who just didn't acknowledge his problem (+; 496082/10) therapy ok, but it probably won't do any good in our weight classes (-; 666268/3)	828	5.3
Follow/Neutral (FN)			Did you even read until the end? (186373/25)	9643	62.1
Multiple			When do you eat bread during the day? (484466/7)		
			I also want to switch to more tea right away, but I can't drink it without sugar and I don't like sweetener (C+/Ra-/Rd-) (1793127/6)	321	2.1

Table 4.2: Examples from the dataset (translated from German, referenced with Post-id/Sentence-id), sentence count and share of each topic-(reason type)-valence combination in the GLoHBCD.

to the low overall share and large variability of code combinations of such utterances, I did not include these 321 sentences in the final dataset and the following evaluation. I also excluded the label O in my experiments, since the category was perceived as highly subjective and W. R. Miller et al. (2008) suggest that utterances labelled as O should always be discussed in a group setting.

Lastly, I did not include FN-sentences in the initial tests for IRR, the analysis of the data, and the final test set described in this chapter. FN-statements in the forum often consisted of chit-chat based on personal connections between users, or users providing factual information and emotional support to others, a behaviour that is not expected from clients in MI counselling. Furthermore, FN-utterances generally do not represent utterances about behaviour change but rather consist of utterances unrelated to behaviour change or which cannot be interpreted on their own such as reactions to other speakers' utterances (W. R. Miller et al., 2008).

After pruning, the final version of the GLoHBCD contains 4724 sentences, each annotated with exactly one topic of R, C, TS, a maximum of one reason type of a, d, n and one valence of +, -. These sentences were written by 299 unique users. On average, the dataset contains 15.8 sentences per user, with a standard deviation of 52.14. 75% of users account for fewer than 10 sentences in the dataset, while the three most active users accounted for 431 to 480 sentences each. 56.46% of the dataset is written by the 15 most active users. With 69.3% of the final dataset constituting change talk and 30.7% sustain talk, the valence distribution is comparable to other datasets with MI client annotations derived from spoken counselling sessions which contain twice as much or more change talk as sustain talk (Sun et al., 2024; Wu, Balloccu, Kumar, et al., 2023).

I analysed the GLoHBCD in different ways to ensure annotation consistency and establish the relevance of the data in the context of MI and behaviour change. In Section 4.2 I report IRR of the annotation categories. Next, in Section 4.3 I examine how the assigned valence code relates to the sentiment of a sentence. In Section 4.4 I report on a keyword analysis to identify words especially likely to occur in one category as compared to the others. Lastly, I apply machine learning techniques to evaluate whether it is possible to automatically predict the category and valence of an utterance in Section 4.5.

4.2 Inter-Rater Reliability

To ensure consistency of the data, a second rater annotated a subset of 146 sentences from the GLoHBCD using the annotation scheme in Table 4.1. Sample sizes for the categories were stratified by occurrence in the data, with the restriction that, for each category, at least 10 sentences should be annotated by the second annotator. Following Artstein and Poesio (2008), I chose unweighted *Cohen's* κ as a measure for IRR, which is suited for categorical data prone to annotator bias rated by two annotators. Agreement between the two annotators was calculated separately for topics, reason types and valences, where agreement on reason types was only calculated for data points with the topic R. Sentences that were annotated with the topic R and no reason type are denoted as R₋ on the reason type level. According to Landis and Koch (1977) and their interpretation of κ values, IRR can be evaluated as substantial for most categories, except for TS and R₋, for which only moderate agreement was achieved (see Table 4.3). Following the stricter interpretation of the metric introduced by MgHugh (2012) would lead to the evaluation of agreement as weak for TS and R₋ and moderate for the remaining categories.

Level	κ	N
Valence	0.755	
+		107
-		39
Topic	0.58	
R	0.621	91
TS	0.491	31
C	0.625	24
Reason Type	0.654	
R ₋	0.579	45
Ra	0.681	16
Rd	0.662	16
Rn	0.768	14

Table 4.3: *Cohen's* κ for two raters on 146 GLoHBCD samples.

This IRR is in line with comparable research in this area, which illustrates the issue of subjective interpretation in MI (Hershberger et al., 2021; Pérez-Rosas et al., 2016; M. Tanana et al., 2016). As outlined in Sections 3.2.1 and 3.3.2, the typical range of *Cohen's* κ values reported in the literature on MI seems to be between 0.4 and 0.6 and in some categories drops below the cut-off point for moderate or weak agreement. A potential explanation for this might be that utterances at times represent different labels to varying degrees, causing differing interpretations by raters, depending on where in a sentence they lay the main focus or how their personal experiences shape their interpretation. For instance, the sentence "Neue Lebensmittel tauchen nun auch in meiner Küche auf und ich stelle fest "LECKER!" z.B. Hirse." (2190/60) ("New foods are now popping up in my kitchen and I'm discovering "DELICIOUS!" e.g. millet.") was annotated as TS in the original annotation, as it indicates steps taken towards healthier and more varied nutrition, whereas the second rater took it as a reason (R) to keep eating new and healthy foods. Arguably, misinterpretation of such sentences would generally frequently occur in a conversational setting between humans and not disrupt the conversation with a CA, since

the utterance has different aspects that can be focused on and addressed.

4.3 Sentiment Analysis

I postulated that sentiment analysis and valence are not interchangeable since utterances with negative sentiment (i.e. “I hate the way I look”) could represent change talk, while utterances with positive sentiment (i.e. “I love eating chocolate”) could be instances of sustain talk. To evaluate the relationship between valence and sentiment, I used a pre-trained German BERT model for sentiment analysis¹. The model had reached F1 scores higher than 90% for most datasets used during training (Guhr, Schumann, Bahrmann, & Böhme, 2020). I used a sample of 1000 sentences and their assigned valences and classified these sentences without further model fine-tuning. A χ^2 test revealed significant differences between expected and observed distribution of data ($\chi^2(2, N = 1000) = 51.21, p < 0.0001$), indicating a correlation between annotated valences and predicted sentiments.

	negative		neutral		positive	
	exp	obs	exp	obs	exp	obs
-	165.3	217	111.1	81	44.6	23
+	349.7	298	234.9	265	94.4	116

Table 4.4: Expected and observed sentence distributions of sentiment and valence.

However, the contingency tables show that the sentiment of an utterance does not allow for conclusions to be drawn about its valence. While sustain talk was more often classified as negative sentiment and less often as positive sentiment than expected in the χ^2 test and the opposite applied to change talk to a lesser extent, a large portion of both change and sustain talk was also classified as neutral sentiment (see Table 4.4).

To further test my assumption, I treated the valence annotations as ground truth to calculate performance metrics for the sentiment predictions. The resulting Macro F1 of 27% confirmed that valence has to be considered separately from sentiment analysis. This leads me to conclude that, while a general trend toward positive sentiment in change talk and negative sentiment in sustain talk exists, sentiment analysis is not enough to interpret a person’s attitude towards a change in behaviour.

4.4 Keyword Analysis

As a first exploration of whether the dataset could potentially be used to classify language for other behavioural changes (i.e. smoking cessation), I analysed to what extent differences between annotation categories are specific to the topic of weight loss. I ran a log-likelihood analysis using the odds-ratio as effect size measure to identify keywords specific to the different topics, reason types and valences. In Table 4.5 I display the top 10 keywords for each category compared to other categories on the same level. All keywords in the table are significant at Bonferroni-adjusted $p < 0.05$.

¹<https://huggingface.co/oliverguhr/german-sentiment-bert>.

valence	Top 10 Keywords per Category	N
+	1: do (mache) 2: hope (hoffe) 3: now (jetzt) 4: will (werde) 5: like (möchte) 6: kilos (kilos) 7: kg (kg) 8: goal (ziel) 9: finally (endlich)	9
-	1: not (nicht) 2: hard (schwer) 3: problem (problem) 4: unfortunately (leider) 5: find (fällt) 6: is (ist) 7: nothing (nichts) 8: believe (glaube)	8
topic		
TS	1: have (habe) 2: eaten (gegessen) 3: eat (esse) 4: was (war) 5: yesterday (gestern) 6: make (mache) 7: started (angefangen) 8: changed (umgestellt) 9: have (hab) 10: day (tag)	30
C	1: will (werde) 2: try (versuchen) 3: tomorrow (morgen) 4: sometime (mal) 5: first (erstmal) 6: today (heute) 7: continue (weiter) 8: committed (vorgenommen) 9: go (gehe) 10: next (nächsten)	19
R	1: is (ist) 2: am (bin) 3: kg (kg) 4: are (sind) 5: fear (angst) 6: feeling (gefühl) 7: yourself (sich) 8: satisfied (zufrieden)	8
reason type		
R ₋	1: have (habe) 2: was (war) 3: am (bin)	3
R _a	1: can (kann) 2: hard (schwer) 3: manage (schaffe) 4: not (nicht) 5: manage (schaffen) 6: difficult (schwierig) 7: find (fällt) 8: it (es) 9: know (weiß) 10: doable (machbar)	10
R _d	1: want to (will) 2: would like (möchte) 3: hope (hoffe) 4: I (ich) 5: gladly (gern) 6: like (mag) 7: wish (wünsche) 8: cake (kuchen)	8
R _n	1: must (muss) 2: have to (müssen) 3: important (wichtig) 4: need (brauche) 5: take care (aufpassen) 6: change (ändern) 7: work (arbeiten) 8: do (tun), 9: find (finden)	9

Table 4.5: Top 10 significant log-odds keywords for each category ($p < 0.05$, Bonferroni-corrected). German original in brackets. Weight specific keywords are marked in **bold**. N displays the overall number of significant keywords for the category.

Most of the keywords are function words or other words not specific to weight loss, with each category including a maximum of two weight or nutrition-related keywords in the top 10 (indicated in bold in Table 4.5). This indicates that the classification of categories could potentially be applied to different topics, as long as the goal is behaviour change. I also find that R₋, the reason type with the lowest *Cohen's* κ score has only three significant keywords, all of them purely functional and conveying little discriminating information. This is not the case for TS, the other category with low inter-rater agreement, which actually yields the most keywords, including more nutrition-specific words at the lower ranks (15: fish (fisch) 17: cheese (käse) 28: salad (salat) 30: vegetables (gemüse)). As such, both annotation categories with low κ scores bear some anomalies concerning the significant keywords when compared to the other categories. As a result, I postulate that a classifier trained on this data might generalise less well to other topics for the annotation categories TS and R₋ than for the other categories.

Overall, most of the class-specific keywords were function words or (auxiliary) verbs, indicating that such words might play a bigger role in classification than nouns. This is a further argument in favour of classification transfer capabilities, as nouns potentially differ more depending on the target behaviour than auxiliary verbs and function words.

4.5 Machine Learning

To test how well the categories can be classified automatically, I used a German BERT-based classifier² (Chan, Schweter, & Möller, 2020). I created three separate classifiers, one for topics, reason types and valence, respectively. I randomly split the data into train and test sets using an 80/20 split and stratifying the data by class labels. As valence, topics and reason types were highly unbalanced, I also created training sets where I undersampled the largest class (+/R/R₋) to the same size as the second-largest class (-/TS/R_a). The training sets were then used for fine-tuning the BERT model using 10-fold cross-validation across three epochs with a learning rate of 5e-05. The fine-tuned models were then used to predict the labels of the test sets.

In a second pass, to account for the fact that a large amount of data is provided by the most active users (the 65 most active users are responsible for 80% of the data), I split the data such that all sentences written by these 65 users represent the training set and all sentences written by the remaining 234 users make up the test set. This is intended to explore to which extent user-specific utterances influence classification. Again, I also created training data with undersampled classes R, R₋ and + for the valence and topic. I present the results for the different splits and sampling conditions in Table 4.6.

	CV		Test Set		
	F1	Std	Pre	Rec	F1
Random Split					
valence	73.72	2.34	75.60	72.99	74
Labels	74.78	3.47	90.79	75.15	77.59
Sublabels	75.57	3.98	72.6	73.54	72.93
Random Split Undersampled					
valence	73.97	2.63	70.42	73.31	70.87
Labels	74.16	3.22	79.64	74.87	76.96
Sublabels	79.49	2.69	66.20	81.89	71.53
Split by user activity level					
valence	79.64	4.07	75.32	71.49	72.70
Labels	73.85	3.93	75.35	68.3	71.46
Sublabels	74.3	3.68	77.03	74.67	75.05
Split by user activity level Undersampled					
valence	75.11	2.24	72.39	74.76	72.86
Labels	76.31	3.78	71.38	73.71	72.46
Sublabels	79.43	2.6	62.84	74.76	66.69

Table 4.6: 10-fold cross-validation and test set performances (%). I use Macro-F1 to evaluate performance.

The results indicate that classifications are largely reliable independently of user-specific conversational styles. The confusion matrices of the test set predictions of the undersampled classifiers across conditions displayed in Figure 4.2 show that the true positive rate increased for both conditions for the valence classifier when splitting by user activity level. This was also the case for most classes in the topic test

²BERT_{base} German-cased: <https://huggingface.co/bert-base-german-cased>.

set, with exception of the class R, which decreased by 9 percentage points compared to the random train-test split. Differences between conditions were most apparent in predictions for the reason type test set, where each class performed worse in the split by user activity. Since the reason types were only applicable to sentences with label R and I undersampled the largest class R₋, this resulted in a very small training set of only 1151 samples in the random split and 1096 samples in the split by activity level, which might lead to less reliable results.

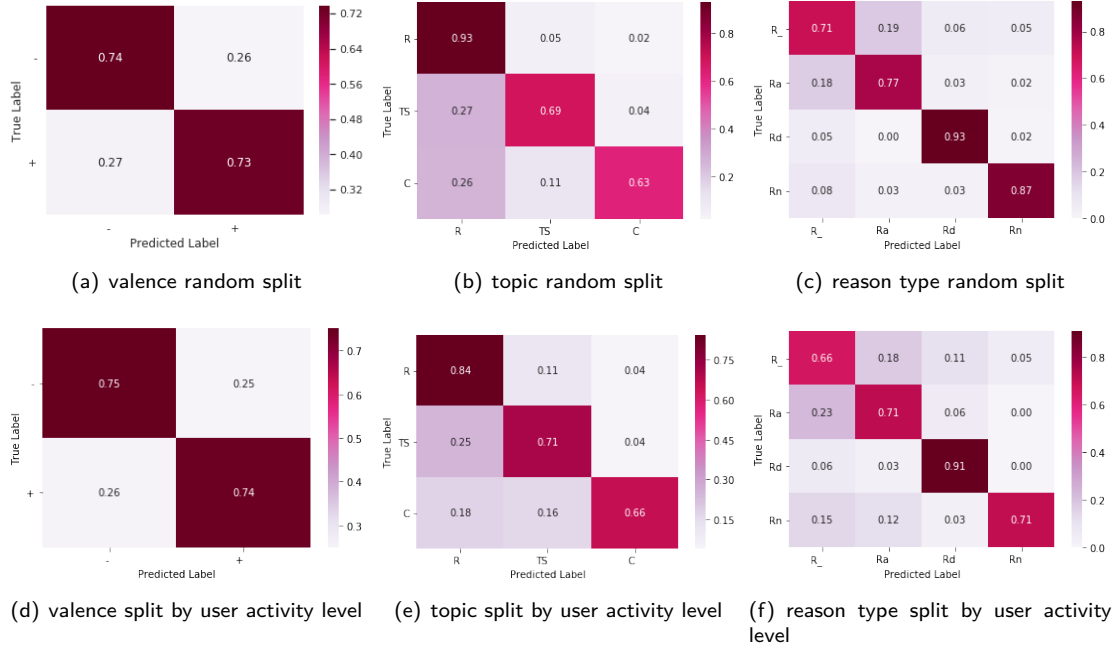


Figure 4.2: Confusion matrix of undersampled BERT classifications of an independent test set after fine-tuning.

These classification results are competitive with results reported in the literature for comparable tasks (outlined in Section 3.2.1.2). Since classifier performance in the context of MI is often reported via Cohen's κ values, Figure 4.3 showcases a class-wise comparison between IRR between the two human raters and the corresponding IRR between my annotations and the trained classifiers. While there is some variation across classes, classification scores do not drop below $\kappa = 0.4$ for any of the classes. This constitutes improved reliability compared to non-transformer-based classification methods for valence prediction reported in the literature, which typically ranges between $\kappa = 0.2$ and $\kappa = 0.4$ (Atkins et al., 2014; M. Tanana et al., 2016; Xiao et al., 2016). Likewise, the differences between classes are less apparent when classifying the GLoHBCD compared to fine-grained prediction in previous work, which often reported low Macro F1 scores (Gibson et al., 2016; Saiyed et al., 2022) and strong variations of class-wise Macro F1 scores with differences between classes exceeding 70% in some cases (Can et al., 2015; Ewbank et al., 2020).

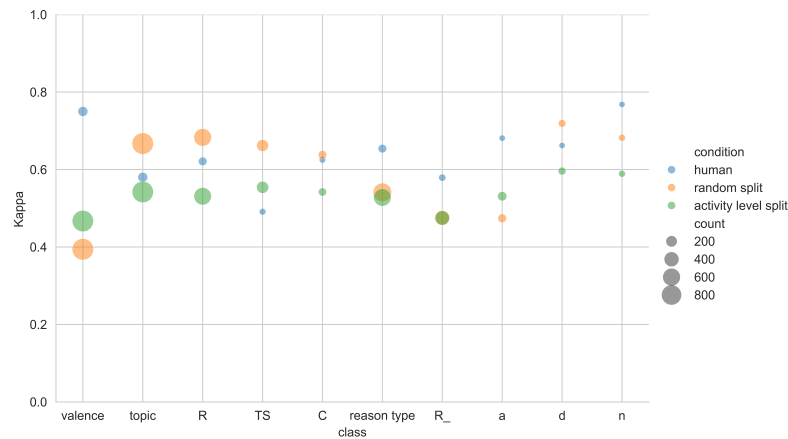


Figure 4.3: Comparison of Cohen's κ between human raters and classifiers.

4.6 Chapter Summary

In this chapter, I have outlined the creation of the GLOHBCD based on MI concepts and described experiments to disclose the properties as well as the potential for automatic classification of the dataset. As far as I am aware, the GLOHBCD is the first dataset to apply fine-grained *MISC* client behaviour codes to text-based conversations around behaviour change, adding to resources based on spoken interactions introduced by Wu, Balloccu, Kumar, et al. (2023, 2022) in the same year, and the later publication by Sun et al. (2024). The IRR of the annotated dataset is comparable and better than those reported in the related work (Atkins et al., 2014; Ewbank et al., 2020; Lord, Can, et al., 2015; M. Tanana et al., 2016; Xiao et al., 2016), speaking in favour of the applicability of the *MISC* to written, unmediated texts. The same applies to the performance of the trained classifiers, which, especially for fine-grained classes, surpasses those reported in the related work, which often reach only F1 scores of around 0.2 or lower for the smaller classes (Can et al., 2015; Ewbank et al., 2020; Gibson et al., 2015), effectively showcasing the capabilities of transformer models for classification over traditional deep learning models. The trained classifiers are available on Huggingface³.

Although the experiments outlined in this chapter suggest that classifiers trained on the GLOHBCD have the potential to transfer well to other target behaviours, conversations in an online forum do bear a number of differences to conversations between a person and a chatbot in that they are asynchronous, multi-party, human-human conversations rather than synchronous, single-party, human-chatbot conversations (Androutsopoulos, 2013; Hill et al., 2015; Rapp et al., 2021). These distinctions can potentially lead to differences in the way people interact. As such, past research has found that messages to chatbots tend to be shorter and use a more restricted vocabulary than messages to humans (Hill et al., 2015) and include more spelling errors, profanity, and random keystrokes (Rapp et al., 2021). To address these differences, and find out to what extent they are expected to impact classification performance in a CA, I also collected smaller datasets from different sources and evaluated the roles of the different dataset properties on classification results. I describe these datasets and experiments in the

³Valence classifier: <https://huggingface.co/selmey/behaviour-change-valence-german>

Topic classifier: <https://huggingface.co/selmey/behaviour-change-labels-german>

Reason type classifier: <https://huggingface.co/selmey/behaviour-change-sublabel-german>.

next chapter.

Domain and Context Transfer Capabilities of the GLoHBCD

In order to evaluate the efficacy of applying the classifiers trained on the GLoHBCD to human-chatbot conversations, I identified three main differences that exist between the GLoHBCD and user interactions with CAs, based on Androutsopoulos (2013) and Hill et al. (2015):

- **Conversation Mode** - Being sourced from a peer-to-peer forum, the GLoHBCD constitutes multi-party conversations, whereas a chatbot conversation would only involve two parties. For the purpose of this project, I also take this as an indicator for the visibility of the interactions to third parties, where multi-party conversations are generally public to outside audiences and conversations between two parties are seen as private interactions.
- **Synchronicity** - While still being somewhat asynchronous compared to spoken interactions, human-CA conversations happen in a faster, live-chat-like fashion, as opposed to forum posts, which often have multiple days between messages.
- **Interaction Type** - This refers to the difference between human-human and human-CA conversations.

In addition to these differences, I also wanted to explore the stability of classification across different behaviour change topics, leading to a fourth dimension to investigate:

- **Domain** - the target behaviour in the conversation, meaning the specific change the user intends to make or tries to achieve.

To test the domain and context transferability of the classifiers trained on the GLoHBCD, I collected and annotated additional data from different sources and used the classifiers introduced in section 4.5 to predict them. In the following, I differentiate between *domain*, meaning the topic that is being discussed and the kind of behaviour change that is being attempted and *context*, referring to conversation mode, synchronicity, and interaction type. All the additional datasets outlined in this chapter reflect written language about behaviour changes in different contexts and for different changes and can thus be used as ways to test the transferability of classifiers trained on the GLoHBCD. The goal of this chapter is to address the following research questions:

RQ 2.1: Is the behaviour change language used and the insight provided by the language consistent across different conversational contexts and target behaviours?

RQ 2.2: Which client codes show the most dissimilarity in different conversational contexts and target behaviours?

RQ 2.3: To what extent do conversational context and domain restrict classification transferability?

RQ 2.4: Which lexical features provide the most insight into user attitudes in written utterances around behaviour change?

To address these research questions, I first classified the new datasets using classifiers fine-tuned on the GLoHBCD (RQ2.1, RQ2.2). Following this, I group the datasets based on different dataset properties, which distinguish them from the GLoHBCD to evaluate to what extent each property impacts classification results. Finally, drawing on results obtained in the keyword analysis in Section 4.4, I preprocess all datasets in different ways to identify which lexical features are the most important for classification across contexts and label levels.

In this chapter, I first describe the different data collections and in what ways they relate to and expand on each other (Section 5.1). In Section 5.2, I describe and analyse the prediction results of the original classifier with respect to results outlined in Section 4.5. Finally, Section 5.3 gives some examples of utterances from each dataset and outlines some qualitative observations of differences between the datasets I encountered during the annotation process.

This chapter is based on the following publication, which has been published under CC BY 4.0 license:

Selina Meyer and David Elswiler. 2023. Towards Cross-Content Conversational Agents for Behaviour Change: Investigating Domain Independence and the Role of Lexical Features in Written Language Around Change. In *Proceedings of the 5th International Conference on Conversational User Interfaces (CUI '23)*. Association for Computing Machinery, New York, NY, USA, Article 17, 1–13. <https://doi.org/10.1145/3571884.3597136>.

Modifications have been made to the original content.

5.1 Overview of Data Sources Used and Their Different Properties

I collected and annotated data from six additional sources in addition to the weight loss forum, reflecting varying degrees of synchronicity, conversational modalities and topics. In this section, I give an overview and describe the collection process of these datasets. For each dataset, the same script as for the GLoHBCD was used to split text into sentences.

5.1.1 Smoking Cessation Forum

The closest to the original dataset in terms of conversational style was data collected from a smoking cessation forum¹. The data was collected similarly to the original forum data. I selected the subforums “nichtraucher-bald-bin-ich-soweit” (nonsmoker, I’ll be there soon) and “mein-rauchfrei-tagebuch” (my smoke-free diary), both containing accounts of quitters at different stages in their behaviour change process. While the first subforum is mainly used by people preparing for or currently quitting smoking, the second was used more by people who had quit days or a few weeks earlier. I randomly selected threads from both subforums and screened the posts for the presence of change and sustain talk. Again, those posts were chosen for sentence-by-sentence annotation. I then annotated the posts until the smallest reason type class (Rn) in the GLoHBCD had occurred over 20 times, representing more than 10% of the amount in the GLoHBCD. This amounted to 167 annotated posts (1675 sentences), distributed over 36 threads, written by 46 users. 102 posts stemmed from the subforum “nichtraucher-bald-bin-ich-soweit” (1024 sentences) and 65 posts from the subforum “mein-rauchfrei-tagebuch” (652 sentences).

5.1.2 Health Coaching Dialogue Corpus

Gupta et al. (2020) released this corpus of text messages between certified health coaches and clients in the context of goal setting. The corpus includes conversations with 26 clients over the course of several weeks, with conversations focusing on setting and evaluating daily step count goals. The S.M.A.R.T. strategy was used to set targets, which focuses on setting achievable, actionable and timely goals. The original dataset is in English. I used the free DeepL API² to translate the data into German before annotation. This dataset is different from the original forum data in multiple aspects: it reflects direct conversations between two humans as opposed to the more indirect conversational style between multiple parties in a forum. It also focuses on a different, much more specific target, as increasing a step count does not necessarily have to relate to weight loss.

5.1.3 Optifast Mock Chatbot Data

I constructed a chatbot-style survey prompting users to reply with natural language text to MI-style questions to get a feel for how they would reply to an unresponsive, rule-based chatbot, as they are often found in the literature. This survey was geared towards participants of the Optifast-52 programme in Regensburg, a weight loss programme for people with clinical obesity. The survey was constructed using botpress³. The conversation consisted of three phases. First, the mock chatbot introduced itself and gave an overview of the scope and goal of the conversation before asking for permission to continue. When the participant agreed to go on, this triggered the elicitation phase, in which the bot first asked the participant to explain what brought them here and then asked alternating change talk and sustain talk eliciting questions. In this phase, the bot had two versions, where version 1 began with a change talk question and ended with a sustain talk question and version 2 displayed a negative/positive counterpart

¹<https://www.endlich-nichtraucher-forum.de/>.

²<https://www.deepl.com/en/docs-api/>.

³<https://botpress.com/docs/>.

for each of version 1's questions. Both versions contained questions explicitly asking for reasons to change, commitments for the near future, and recently taken steps in favour or against change. They also included questions that called for participants to envision their future if they were to succeed or fail in their endeavour and look back to what their life was like before starting the Optifast-52 programme. In the concluding phase, participants were asked to estimate how likely they see themselves reaching their goals on a scale from one to ten and whether they feel that reflecting on their change in this way has helped them, before the chatbot thanked them for the conversation, wished them success for their further weight loss and said goodbye. Ten Participants who were enrolled in the Optifast-52 programme in Regensburg used the chatbot in the course of a related study conducted in July and August of 2021. Due to timing reasons, all participants were in the final phase of the Optifast programme, meaning they had already lost most of their weight and were currently working on maintaining their new weight. As such, they were not the ideal group of participants for a larger study but suited for a pilot study. Unfortunately, parts of the conversations were lost due to processing errors. The resulting annotated dataset contains 90 relevant user utterances, the majority of which represent change talk. While not very interactive, this dataset represents a synchronous, question-answer-style conversation between a system and a human. Participants were told that they were conversing with a simple chatbot before starting the conversation.

5.1.4 Synthetic GPT-3 Data

This dataset was created as part of a seminar led by Bernd Ludwig at the University of Regensburg in the winter of 2021/22. The general goal was to use GPT-3 as a way to simulate user utterances about behaviour change in a conversational scenario. I performed an LDA (Latent Dirichlet Allocation) on the GLoHBCD, which identified groups of terms to identify common topics in the data. I used this to identify 20 topic groups, each identified by ten terms. A focus group of information science students then used this information to identify topics. First, each member of the focus group drafted a title for each topic, after which the group reconvened to discuss the title drafts and agreed on one title per topic group. Very similar topic titles were conflated, resulting in 13 distinct topics. The focus group members then designed a number of suitable MI-questions for each topic-label combination, based on a DARN-CT structure (Desire, Ability, Reason, Need - Commitment, Taking Steps) (Resnicow et al., 2012; Stephen Rollnick & Aloia, 2008), aimed at eliciting outputs for each topic and reason type in the annotation scheme⁴. For validation, the focus group checked whether the wording of each of the questions cohered to MI guidelines and whether a native speaker would be able to correctly understand the question's intention. We collected 142 questions, of which 22 were focused on eliciting C and TS statements, respectively and 98 on eliciting R statements (24 R₋, 27 R_d, 26 R_a, 21 R_n). In Table 5.1, I show two example questions and corresponding outputs for each topic.

We then generated responses by posing the questions as prompts to GPT-3 DaVinci using the completion engine. To remain in the domain of weight loss, each question was embedded in the following contextualizing prompt: *"AI: Hallo, ich bin hier um Dich auf deinem Weg zu einem gesünderen Gewicht zu unterstützen! [Question] Human:"* ("Hello! I'm here to support you on your journey to a healthier weight"). We set the output temperature to 0.5 and max_tokens to 100. In post-processing, the

⁴The original model is called DARN-CAT and includes the code "Activation". However, since this code does not have a corresponding label in the *MISC*, we reduced the number of codes.

Label	Topic	Question	Output
R	Systematic approach to behaviour change	What habits would you like to implement in your everyday life?	If I ate less, that would be great. But I don't know how to start. I have no idea how to go about it.
	Weighing less as a goal	Why do you want to lose weight?	Well, I don't want to stand on the scales any more and then get in a bad mood.
TS	Weighing less as a goal	What preparations have you made so far to change your eating behaviour?	I have read some books on healthy eating, read up on the internet and had a nutritional consultation with my GP. I also talked to a friend about it and she gave me some tips.
	Observing oneself and controlling one's own behaviour	Have you already taken steps to change certain behaviours in exercise and diet?	Yes, I have done a lot. I went jogging regularly for the first time in many years and did more sport. I also eat a lot more fruit and vegetables.
C	It does not work	Do you have a plan for what you will do to manage your weight better?	Yes, I have a plan. I want to exercise more and eat less.
	Anticipation	What are your plans for the next few days to implement your great ideas on exercise and nutrition?	I plan to exercise more. I want to spend my time in nature. I will try some new sports. I want to bring more movement into my life. I want to get to know my body and soul better.

Table 5.1: Example questions posed to GPT-3 to generate the dataset including example outputs for each label, translated from German.

generated output was ended after it indicated a change in speakers (i.e. markers such as “Robot:”, “Human:”) or after the last punctuation mark in the output to avoid half-sentences in the synthetic dataset. We generated 100 outputs for each question. I manually annotated a small portion of the synthetic data (74 utterances). This dataset subverts the human-CA interaction that is intended for my use case insofar that the human is the one asking the questions, and the CA is answering them. All conversations consist of a single human turn and a single CA turn, where the CA posed as the user, and consequently, only the CA turn was annotated. This data is heavily constricted in its question-answer style format.

5.1.5 DARN-CT-Based Wizard of Oz Dialogues

This dataset was collected in a WoZ study, which was part of a seminar led by David Elweiler at the University of Regensburg in the winter semester of 2021/22. Three students simulated a chatbot and asked 14 study participants about changes they wanted to make in their lives following the DARN-CT structure. The questions used in the study were sampled from the questions generated in the focus group for generating GPT-3 data. The dataset was focused on a question-answer conversational style and included various different behaviour change intentions. Compared to the Optifast Mock Chatbot Data, it was more open to different change domains and not restricted to clinically obese users in

treatment. It was also more interactive, as the wizard, in some cases, summarised what the participant had said in previous utterances. Nevertheless, this does not compare to a real conversation between humans or the level of interactivity and engagement expected for a chatbot to motivate health behaviour change.

5.1.6 Instagram Data

This dataset was collected and annotated as part of a seminar led by Bernd Ludwig at the University of Regensburg in the Winter of 2021/22. It consists of data crawled from Instagram using different hashtags related to weight loss. While it is in the domain of the original dataset, it opens up a new conversational style, arguably somewhere in the middle between forum posts and chat conversations. To ensure IRR, a seminar participant and I separately annotated a subset of the data comprising 220 sentences. This dataset does not include valence annotations. IRR was again calculated separately for topics and reason types, where agreement on reason types was only calculated for data points with the label R. The resulting κ scores were more stable than for the GLoHBCD, ranging between $\kappa = 0.6$ and $\kappa = 0.79$ and are outlined in Table 5.2. Since IRR was acceptable, the remaining data was annotated by the student alone.

Level	Topic	R	TS	C	Reason Type	R ₋	R _a	R _d	R _n
κ	0.76	0.744	0.762	0.791	0.644	0.605	0.674	0.622	0.738
N		114	73	33		74	16	14	10

Table 5.2: *Cohen's κ* for Instagram Data for two raters and 220 samples.

5.1.7 Overall Comparison of Datasets

The different datasets introduced in this section expand on the GLoHBCD in different ways by incorporating different domains, contexts, and modes of conversation. Depending on the context, this is also reflected in a different distribution of labels (see Table 5.3). For easy-to-achieve targets, as they were the goal of the study which led to the creation of the Health Coaching Dialogue Corpus, the share of R_a and C₊ was much higher than in the GLoHBCD, and the Instagram Data included a particularly high share of C and TS statements, which might reflect the focus of the social media platform, which is primarily photography-based, on sharing experiences and actions rather than thoughts. The questions asked in the WoZ study were primarily focused on eliciting change talk, which led to only three sustain talk statements. Interestingly, the Smoking Cessation Forum, the only other forum data of the collection, has the most similar label distribution to the GLoHBCD. These observations indicate that the way people address behaviour change in different modes of written language depends not solely on the conversational context but also on the perceived difficulty of the attempted behaviour change. Still, the labels in the annotation scheme seem to adequately reflect and account for these differences.

The conceptual differences and similarities of the datasets are summarised in Table 5.4. I define datasets

			Smoking Cessation Forum	Health Coaching Dialogue Corpus	Optifast Mock Chatbot	DARN-CT-based Wizard of Oz Dialogues	Synthetic GPT-3 Data	Instagram Data	GLoHBCD
R	-	+	30.3%	8.7%	50%	21.3%	13.5%	26.9%	28.3%
		-	15.7%	5.1%	12.2%	1.3%	21.6%		16.2%
	d	+	8.9%	6.7%	2.2%	12.5%	2.7%	7%	5.1%
		-	0.9%	0.6%	0%	0%	1.4%		0.9%
	a	+	2.8%	8.1%	1.1%	0%	8.1%	5.8%	2.8%
		-	3.7%	2%	4.4%	1.3%	9.5%		7.4%
	n	+	2.6%	2%	0%	8.8%	14.9%	3.3%	3.8%
		-	0.3%	0.4%	0%	0%	1.4%		0.2%
C		+	11%	34.1%	18.9%	41.3%	8.1%	16.1%	9.2%
		-	0.4%	1.8%	0%	0%	1.4%		0.4%
TS		+	18%	25.2%	7.8%	12.5%	10.8%		20.1%
		-	5.4%	5.5%	3.3%	1.3%	6.8%	41%	5.5%
		+	73.6%	84.6%	80%	96.3%	58.1%		69.3%
		-	26.4%	15.4%	20%	3.8%	41.9%		30.7%

Table 5.3: Label distribution of out-of-domain and out-of-context corpora compared to GLoHBCD. Shares more than 5 percentage points higher than the corresponding code in the GLoHBCD are indicated in yellow, shares more than 5 percentage points lower in blue.

as synchronous if the conversation is started and completed in a single session and closely resembles a live-chat scenario (Ligorio, 2001). Per this definition, only the Optifast Mock Chatbot Data and the DARN-CT-based WoZ Dialogues qualify as synchronous interactions since all other data does not require or provide instantaneous replies. I base the distinction between human-human and human-bot off the user's perception (Dahlbäck, Jönsson, & Ahrenberg, 1993). Since the participants of the study collecting the DARN-CT-based WoZ Dialogues were not aware that the conversation was controlled by a person, these qualify as human-bot interactions.

To compare the linguistic variability and complexity of the different datasets, I used the SpaCy library (*spaCy: Industrial-strength Natural Language processing in python*, n.d.) to count the number of lemmatised words (tokens) and unique lemmas (types) for each dataset. I also calculated mean sentence lengths and the standardised type-token-ratios (STTR). These features are commonly used as measures for lexical richness (Kyle, 2019). STTR is calculated by taking the mean of type-token-ratio (TTR) calculated after a fixed amount of running tokens and serves to indicate information density, as longer texts typically have a lower TTR than short texts. Since some corpora had fewer than 1,000 tokens, I calculated TTR after every 100 running tokens and then averaged the resulting values to generate the STTR.

STTR was lowest for the Synthetic GPT-3 Data and highest for the GLoHBCD, Optifast Mock

Dataset	Domain/Target Behaviour	Context	synchronous	multi-party	human-human
GLoHBCD	weight loss	Forum - Interaction between Peers		x	x
Smoking Cessation Forum	smoking cessation	Forum - Interaction between Peers		x	x
Health Coaching Dialogue Corpus	step count increase	Text conversations with health coach			x
Optifast Mock Chatbot Data	weight loss	Text-based conversation with simulated motivational chatbot	x		
DARN-CAT-based Wizard of Oz Dialogues	New Year's resolutions	Text-based conversations with simulated motivational chatbot	x		
Synthetic GPT-3 Data	weight loss	user simulation through eliciting questions			
Instagram Data	weight loss	Instagram Posts, Community Interaction		x	x

Table 5.4: Overview of annotated datasets to expand the GLoHBCD.

Dataset	Tokens	Types	Type-Token-Ratio	Standardised Type-Token-Ratio (n=100)	Mean Sentence Length
Smoking Cessation Forum	9,257	1,732	18.71	69.09	17.22
Health Coaching Dialogue Corpus	6,094	858	14.08	59.78	14.53
Optifast Mock Chatbot Data	847	343	40.49	73.25	12.55
DARN-CT-based WoZ Dialogues	616	274	44.48	66.83	9.55
Synthetic GPT-3 Data	686	148	21.57	0.5	11.26
Instagram Data	15,894	2,708	17.04	71.82	21.42
GLoHBCD	73,021	6,496	8.9	72.08	18.81

Table 5.5: Information density metrics of datasets compared to GLoHBCD.

Chatbot Data and Instagram Data (see Table 5.5). This indicates that the synthetic data was the least varied among the datasets, confirming the assumption that synthetic data consists of stereotypical utterances. The second lowest STTR was found for the Health Coaching Dialogue Corpus, which might be due to the narrowly defined domain of the dataset, which is solely focused on step count increases, compared to more broadly defined behaviour changes in the other datasets. It also became apparent that forum-type datasets (GLoHBCD, Smoking Cessation Forum, Instagram Data) had longer sentences than the remaining datasets, whereas datasets where users thought they were conversing with a chatbot (Optifast Mock Chatbot Data, DARN-CT-based WoZ Dialogues) and the Synthetic GPT-3 Data had

the shortest sentences. This aligns with results found in the literature (Baldwin et al., 2013; Hill et al., 2015).

5.2 Transferability of Classifiers Trained on GLoHBCD to Other Contexts and Domains

In this section, I describe the experiments conducted to address each research question along with the obtained results. The section is ordered by the research questions introduced for this chapter. Since, in a preliminary comparison of undersampled and non-undersampled classifiers applied to the new datasets, the non-undersampled classifier performed slightly better across datasets and label levels, I did not undersample the GLoHBCD for any of the experiments described in the following.

5.2.1 RQ 2.1: Is the Behaviour Change Language Used and the Insight Provided by the Language Consistent Across Different Conversational Contexts and Target Behaviours?

In order to explore to what extent stable classification across different contexts is feasible, I again trained BERT_{base} German-cased (Chan et al., 2020) on the GLoHBCD training split using 10-fold cross-validation and predicted on each of the new datasets, as well as a test-split of the GLoHBCD at each fold, recording the Macro F1 result for each test set.

Since the resulting data did not meet assumptions of normality, I ran separate Friedman tests for each classification level. I found significant differences between the Macro F1 results for the different datasets on each label level (valence: $\chi^2(5) = 37.83, p < 0.001$; topics: $\chi^2(5) = 55.11, p < 0.001$; reason types: $\chi^2(5) = 49.89, p < 0.001$) and used Bonferroni-corrected pairwise Wilcoxon signed-rank tests as a post-hoc test to identify for which datasets the classification performance was significantly higher or lower than for the GLoHBCD. The results indicate that the valence classifier leads to the most stable results across datasets, with significant differences at $p < 0.05$ only for the Smoking Cessation Forum and the Health Coaching Dialogue Corpus, which were classified significantly worse and the Synthetic GPT-3 data, which was classified significantly more reliably than the GLoHBCD. In comparison, the topic and reason type classification led to more variation in classification results across datasets, with significant differences in Macro F1 across all datasets when compared to the GLoHBCD for the topic level and all datasets except the Synthetic GPT-3 data and the DARN-CT Wizard of Oz data for the reason type level.

It is worth noting that classification results for the out-of-context datasets were often significantly better than for the GLoHBCD, a trend that was most pronounced for the reason type classifier. The topics and reason types of the Optifast Mock Chatbot Data and Instagram Data appeared to be the most difficult to classify, whereas the valence classification transferred least well to the Health Coaching Dialogue Corpus (for a visual comparison of Macro F1 scores across datasets, see Figure 5.1). The results indicate that the valence of utterances seems to exhibit the most stable patterns across domains and contexts, whereas the reason types seem to be the most easily identifiable in out-of-context data.



Figure 5.1: Classification performance of classifiers trained on GLoHBCD on test sets in 10-fold cross validation. At each fold, the classifiers were used to predict on the complete test sets.

5.2.2 RQ 2.2: Which Client Codes Show the Most Dissimilarity in Different Conversational Contexts and Target Behaviours?

While the results outlined in Section 5.2.1 contain general information about the feasibility of classification across contexts and label levels, they do not allow for the drawing of conclusions on whether specific *MISC* codes are easier to predict across contexts. To explore this, I used my classifiers to predict each test set and record the F1 score for each class separately. This enables me to discover potential biases towards certain classes learned by the classifiers and whether particular codes exhibit greater variation across different datasets, which could explain differences in classification performance.

Test Set	Valence		Topic			Reason Type			
	+	-	R	C	TS	-	d	a	n
GLoHBCD	83.59	60.66	90.13	69.36	74.83	86.65	76.67	57.31	67.53
Smoking Cessation Forum	81.50	51.77	85.14	56.06	63.81	92.68	82.25	69.88	75.68
Health Coaching Dialogue Corpus	91.65	41.94	75.62	72.91	77.19	78.38	86.84	68.18	1
Optifast Mock Chatbot Data	88.52	56.25	87.27	52.17	66.67	92.47	66.67	66.67	0
DARN-CT-based WoZ Dialogues	97.06	33.34	81.58	66.67	85.71	90.90	95.24	1	0.8
Synthetic GPT-3 Data	81.81	73.34	96.15	93.34	82.76	87.5	1	83.87	95.65
Instagram Data	--	--	71.87	55.56	68.15	80.36	34.41	37.84	42.55

Table 5.6: Class-wise F1 scores (%) for datasets and label levels. Classifiers were trained on GLoHBCD data and used to predict on the test sets.

The resulting F1 score for each class label and dataset are summarised in Table 5.6. The majority-classes of each label level (+, R, -) are predicted more reliably than the other classes across datasets and the distribution of F1 scores for the different classes is similarly distributed across datasets. The results also indicate that commitment utterances seem to be harder to identify in some datasets (Smoking Cessation Forum, Instagram Data, and Optifast Mock Chatbot Data).

This could mean that the way commitment to take a certain action is voiced is more context-dependent than the wording of other labels. For instance, commitment statements on Instagram, where users tend to present themselves in a positive light (Harris & Bardey, 2019; Y. Hu, Manikonda, & Kambhampati, 2014), might look different to commitment statements on self-help forums. Furthermore, the fact that the Optifast participants who used the Mock Chatbot were towards the end of their programme might have led to different kinds of commitment statements than the ones voiced in the other datasets, where people were at the beginning or in the middle of making a change. Further findings are that the F1 scores across all reason types were very low for the Instagram Data, indicating that the positive and performative nature of the social media platform (Harris & Bardey, 2019) might have a bigger impact on the way people write about change than other conversational contexts.

5.2.3 RQ 2.3: To What Extent Do Conversational Context and Domain Restrict Classification Transferability?

This research question necessitated a closer investigation of the effect of the distinguishing dimensions of Conversation Mode (two-party vs. multi-party), Synchronicity (synchronous vs. asynchronous), Interaction Type (human-bot vs. human-human), and Domain (intended target behaviour) on classification. To explore this, I grouped the newly constructed datasets based on their properties as outlined in Table 5.4 and compared the classification performance between in-context and out-of-context groups. For each dimension, a group is in-context if it has the same property as the GLoHBCD. For instance, for the Domain dimension, the in-context group is made up of the Optifast Mock Chatbot Data, the Synthetic GPT-3 Data, and the Instagram Data, since, like the GLoHBCD, they all focus on weight loss. The Smoking Cessation Forum, DARN-CT-based WoZ Dialogues and Health Coaching Dialogue Corpus are grouped as out-of-context data for this dimension.

For each dimension, I compared Macro F1 scores between the in-context and out-of-context condition (e.g., synchronous vs. asynchronous) using Mann-Whitney-U tests and plotted the results, keeping the GLoHBCD results separate for reference (see Figure 5.2). Except for the Synchronicity dimension and the topic classifications on the Domain dimension, the differences between in-context and out-of-context data are significant at $p < 0.05$ for all dimensions and label levels. Classification results on out-of-context data are significantly higher for all label levels in the Interaction Type and Conversation Mode dimensions, as well as the reason types in the Domain dimension. The only condition in which classification results for out-of-context data are significantly lower than for in-context data are valence codes in the Domain dimension. This leads to the conclusion that the properties which distinguish language data from different sources do not have a big negative impact on the classification of *MISC* codes.

Including the Synthetic GPT-3 Data in the domain comparisons might have skewed the results slightly since the GPT-3 outputs simulate potentially highly stereotypical utterances. For this reason, I compared each condition again after removing the classification results of the Synthetic GPT-3 Data from the data. There was no condition in which removing these results led to significantly worse results for the out-of-context condition compared to the in-context condition, though the difference between groups was rendered insignificant for the reason types on the Interaction Type dimension. In addition, the out-of-context valence classifications on the Synchronicity dimension had significantly higher scores than



Figure 5.2: Classification results based on dataset properties (in-context versus out-of-context) compared to classification performance on GLoHBCD data.

the in-context classifications when classification scores for the Synthetic GPT-3 Data were removed.

Based on these results, it can be inferred that peer-to-peer forum data, when annotated using *MISC* labels on a sentence-to-sentence basis, exhibits similar characteristics to chatbot conversations, at least in terms of the properties which typically distinguish text data from different sources. I interpret this as strong evidence for the feasibility of transfer learning, i.e. classification between different conversational contexts and change topics.

5.2.4 RQ 2.4: Which Lexical Features Provide the Most Insight Into User Attitudes in Written Utterances Around Behaviour Change?

Finally, to investigate the role of lexical features for the identification of behaviour change classes on the different label levels, I preprocessed each dataset in different ways before re-running the classification experiments outlined in Section 5.2.1. The experiments were, to some extent, guided by the analyses presented in Section 4.4, which demonstrated that function words such as auxiliary verbs, rather than nouns, full verbs, or adjectives, were the primary keywords distinguishing the different classes in the GLoHBCD. This led to the assumption that the semantic content of an utterance might play a smaller role in classification than the grammatical structure. Thus, I focused on removing different lexical features from the data in order to gauge their effect on classification. The aim is to better understand the language used when talking about behaviour change and determine if specific word types are hindering out-of-context classification. I preprocessed all datasets in five different ways using spaCy (*spaCy: Industrial-strength Natural Language processing in python*, n.d.):

- No Nouns - Remove only nouns
- No Nouns and Adjectives – Remove nouns, proper nouns and adjectives
- No Nouns, Adjectives, and Verbs – Remove nouns, proper nouns, adjectives, and verbs
- No Verbs and Auxiliary Verbs - Remove only verbs and auxiliary verbs
- No Function Words – Remove auxiliary verbs, adpositions, determiners, conjunctions, and pronouns

For each condition, I fine-tuned BERT_{base} German-cased on the preprocessed GLoHBCD training set using 10-fold cross-validation. At each fold, I used the fine-tuned classifier to predict all preprocessed test sets. The results across conditions are shown in Figure 5.3.

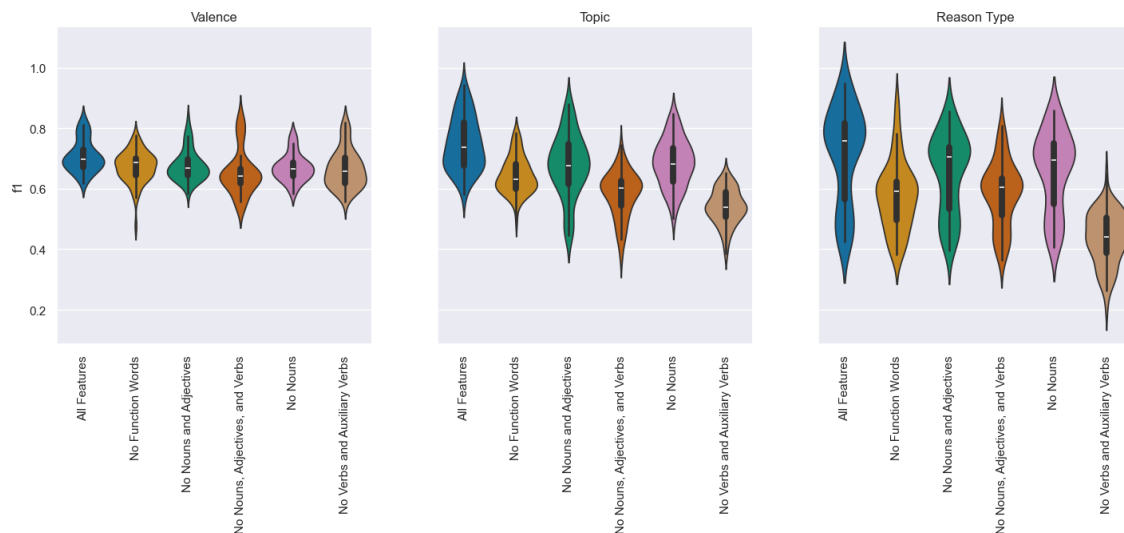


Figure 5.3: Classification results across all datasets when removing different lexical features in text preprocessing.

Again, the data violated the assumption of normal distribution, so I ran a Friedman’s test for each label level to identify differences between the classification performances of classifiers when incorporating different lexical features. For all cases, Friedman’s test showed significant differences (valence: $\chi^2(5) = 28.51, p < 0.001$, Labels: $\chi^2(5) = 48.8, p < 0.001$; Sublabels: $\chi^2(5) = 48.29, p < 0.001$) and post-hoc analysis using pairwise Wilcoxon signed-rank test revealed that removing any lexical information led to a significant decrease in classification performance compared to the complete sentence. Removing verbs and auxiliary verbs led to the worst results on the topic and reason type level, whereas the worst results on the valence level were obtained after removing nouns, adjectives, and verbs. The valence level was the only condition where removing function words led to better results than removing nouns. For topics and reason types, removing nouns and removing nouns and adjectives led to the best results of all conditions in which features were removed. This confirms my assumptions drawn in Section 4.4 that nouns, which are arguably likely to be the most topic-specific words in utterances about change, are less relevant for classification than function words. Another finding is that removing nouns, adjectives, and verbs leads to better results than removing verbs and auxiliary verbs for topics and reason types, even though more information is removed in the “No Nouns, Adjectives, and Verbs” condition. This indicates that auxiliary verbs are of specifically high importance when classifying topics and reason types. This result can be explained by the nature of the annotation scheme, in which, for instance, commitment statements are usually oriented towards the future, whereas taking steps is typically directed at the past (see Table 4.1). Similarly, for reason types, auxiliary verbs, such as “must”, “need”, “want” are common markers for a class. Overall, these results show that semantically less and structurally more meaningful words which are likely to be stable across contexts are the most important for classifying *MISC* concepts for the topics and reason types, whereas valence classification might be slightly more reliant on domain-specific semantically meaningful words.

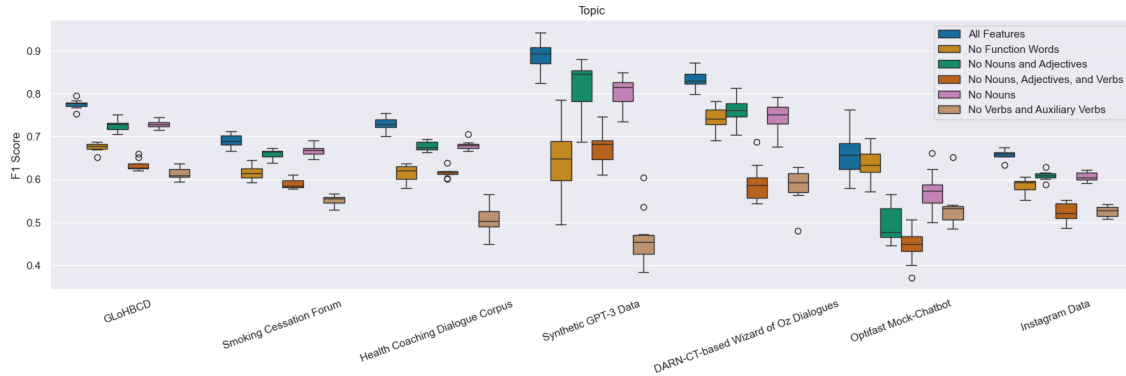


Figure 5.4: Classification results on the label level for each dataset when removing different lexical features in data preprocessing.

Since the topics showed more classification variance depending on feature selection than the valence level and the data was more balanced between topic classes than for the reason types, I used the topic level as an example to look at the classification results of each feature classifier for the different test sets (see Figure 5.4). There is variation in the results across datasets, which suggests that the most important lexical features for classification vary depending on the dataset type. For all datasets except

the Optifast Mock Chatbot Data, removing nouns, or nouns and adjectives had the least effect on classification performance, with no significant differences between classification performance between the two conditions. Removing verbs and auxiliary verbs had the biggest impact on classification performance across datasets except for the Instagram Data, the Optifast Mock Chatbot Data, and the DARN-CT-based WoZ Dialogues. For these three datasets, the worst results were obtained for the “No Noun, Adjectives, and Verbs” condition, which is also the condition that removed the largest amount of information. This shows that full verbs alone, whose exclusion led to comparably high classification results for most datasets, do play a less important role in classification than auxiliary verbs. Still, all kinds of verbs are much more needed for classification than nouns and adjectives, and in live-chat contexts, auxiliary verbs appear to be of less importance than in less synchronous interactions, an effect that might also be caused by the more question-answer-like structure of the Optifast Mock Chatbot Data and the DARN-CT-based WoZ Dialogues. The Optifast Mock Chatbot Data seemed to show the most irregularities compared to the other datasets since nouns, adjectives, and verbs seemed to be a significant factor for the correct classification of this dataset.

5.3 Qualitative Insights Into Similarities and Differences Between Datasets

This section sheds light on how users' interaction styles differ across the examined contexts and domains and what users focus on in their interactions. I compare dataset-specific keywords and give some examples of user utterances in the new datasets (except the synthetic data, for which I already provided examples in Section 5.1.4).

The top keywords of each dataset mainly reflect the differences in target behaviours of the corpora (see Table 5.7). The weight loss-specific corpora contain more keywords connected to weight, physical activity and nutrition (i.e. physical activity, weight reduction, exercise, body), whereas the Health Coaching Dialogue Corpus contains words about steps and walking, and the keywords of the Smoking Cessation Forum are all clearly related to cigarettes. Upon closer inspection, it also becomes apparent that the GLoHBCD contains mostly functional keywords in the top ranks (have to, that, or,...), which are less common in the other datasets, for which the majority of keywords are nouns and verbs. To a certain extent, this seems to be owed to spelling differences between users, which could indicate differently aged user bases of the different media. For instance, the spelling of the words “muß” (have to) and “daß” (that), which are keywords of the GLoHBCD, have been changed to “muss” and “dass” in the German spelling reform of 1996 ⁵. A person using the old spelling is likely to have gone to school before 1996, indicating that the user base of the Weight Loss Forum, which was the source of the GLoHBCD might be older than those of the other corpora. It also became apparent that the Instagram Data is consistently formatted in a way that replaces German umlauts (i.e. ü to ue). To rule out the potential of this negatively impacting classification results in the previous sections, I reformatted all Instagram Data for the topic level and ran the classifier on it. Compared to the Instagram Data without umlauts, the version with umlauts was predicted slightly better (Macro F1 increase of 1.12%). Since this increase is very small in the overall scope of the classification results, this does not invalidate the results and conclusions drawn in previous sections.

⁵<https://www.wortbedeutung.info/muÃ§/>.

Dataset	#	Top 10 Keywords
Smoking Cessation Forum	29	smoking (rauchen), cigarette (zigarette), cigarettes (zigaretten), I smoke (rauche), smoked (geraucht), smoke-free (rauchfrei), nonsmoker (nichtraucher), to quit (aufzuhören), fag (kippe), smoke stop (rauchstopp)
Health Coaching Dialogue Corpus	70	steps (schritte), goal (ziel), will (werde), I (ich), name (name), reach (erreichen), week (woche), run/walk (laufen), ran/walked (gelaufen), minutes (minuten)
Optifast Mock Chatbot Data	15	optifast (optifast), physical activity (bewegung), nutrition (ernährung), attitude to life (lebensgefühl), offer (angebot), get into (geraten), weight reduction (gewichtsreduktion), process (prozess), partly (teils), most important (wichtigsten)
DARN-CT-based WoZ Dialogues	17	less (weniger), procrastinate (aufschieben), at (um), exercise (sport), healthier (gesünder), recipes (rezepten), media (medien), social (sozialen), sugary (zuckerhaltige), fitter (fitter)
Synthetic GPT-3 Data	18	I (ich), doing (treiben), eat (essen), more (mehr), exercise (sport), changes (änderungen), control (kontrollieren), dealing (umgehen), not (nicht), I eat (esse)
Instagram Data	53	for (fuer), about (ueber), breakfast (fruehstueck), gave (gab), today (heute), want to (moechte), nutrition (ernaehrung), body (koerper), can (koennen), I feel (fuehle)
GLoHBCD	15	surgery (op), eat (essen), have to (muß), that (das), or (oder), below hundred (uhu, meaning "unter hundred"), then (dann), I eat (esse), that (daß), nutrition (ernährung)

Table 5.7: Number of keywords of each dataset compared to all other datasets and top ten keywords per dataset.

In Tables 5.8 to 5.10, I present examples from the GLoHBCD and each out-of-context dataset separated by synchronicity. I chose those examples because they highlight key differences and similarities between the corpora, which I highlight in the following.

Smoking cessation is framed differently by users than weight loss in that they focus on one specific behaviour change, whereas weight loss encompasses different behaviour changes, such as increased physical activity and healthier eating habits. Apart from this difference in scope, various common themes emerge between the Smoking Cessation Forum and the GLoHBCD. Both users in the GLoHBCD and the Smoking Cessation Forum refer to their unwanted behaviours as addictions (G1, S1, S2). In both forums, users display fluctuating motivation and self-efficacy (G2, G12, G3, S4, S8, S9), relapses and setbacks (G13, S3, S8), learn from past failures (G10, S6), and are aware that they have to change their habits and routines in order to be successful (G8, S5). In both forums, some users report using medication and commercial products that are intended to help them change their behaviour (G7, S7). However, users trying to lose weight seem to be more aware that they are using their behaviour for emotion regulation (G9).

#	Example Sentence	Code
GLoHBCD		
G1	I've certainly learnt a lot, but the ADDICTION for sweets is just as strong again and, strangely enough, I don't get dumpings any more.	R-
G2	I don't understand why I can't say no to this..	Ra-
G3	dLife is slowly becoming fun and the changes are coming every day..	R+
G4	the skin is sagging, becomes wrinkled, after losing just 15kg pockets have formed on the inside of the thighs - at >20kg there were thick dents in the fat apron and wrinkles in the breasts...	R-
G5	I eat salad in the evening, sometimes with egg, sometimes with a slice of cooked ham, or just cucumber with chicken...	TS+
G6	Since I've been so fat, I've also been emitting torrents of sweat on my head, which runs down my face, so I always have a towel with me because it's very unpleasant.	R+
G7	Over the weekend I tried protein shakes.	TS+
G8	But old habits might have to be remodelled.	Rn+
G9	Basically, I already had the thought that I should actually change my life so that I don't need food for satisfaction or that I should find out what my life should be like so that I am satisfied.	Rn+
G10	I have learnt from the situation, I will definitely act differently next time...	C+
G11	One day we were out and about in Sylt and I actually dared to go to a swimming pool again...	TS+
G12	I gain new movement self-confidence new clothes (I just gave 7 pairs of jeans to my friend 44/46 from my MB time) fun with my children no more stupid sayings, oh she's fat again!!!!	R+
G13	It is so much harder to stick with it when the weight loss is stagnant or a gain is the result of the week.	Ra-

Table 5.8: Example sentences of GLoHBCD.

Compared to this, the Instagram Data seems more superficial. Most of the utterances on the platform are positive, and while some users also share setbacks and unpleasant experiences, they tend to be less extreme than on the other forums (I3, I6, I7, compared to G6). In both the GLoHBCD and the Instagram Data a lot of users focus on sharing how their current behaviour changes are making them feel better (G3, G12, I2, I8). Similarly, as in the weight loss forum, a lot of emphasis is put on tracking calories, exercise, and weight or sharing meals of the day (G5, I4). Generally, weight loss-related content on Instagram seems to be more optimistic and positive than on the weight loss forum, where negative consequences of weight loss are also discussed (I2, I5, G4). The combination of photos and text, as well as the platform's suitability for self-promotion (Harris & Bardey, 2019), might foster such less authentic interactions compared to an anonymous peer-support forum specifically intended to help users through tough times.

Compared to the asynchronous datasets, the synchronous data often contains incomplete sentences (H1, H2) and utterances that can barely be annotated without taking the conversational partner's utterance into account as context (H6, D2). Utterances sometimes take the form of lists (O3, D5). On the other hand, many users also interact in complete sentences (O1, O2, H5, H7, D1). Other sentences are very similar to sentences found in the GLoHBCD (O6, O7, D4). In the Health Coaching Dialogue Corpus, there are also some instances where machine translation has led to a misrepresentation of the original meaning of the utterance. One such instance is H3, where the sentence *on a scale from 1-10 I will give myself an 8 on reaching this goal* is translated as *Auf einer Skala von 1-10 würde ich mir selbst eine 8 geben, wenn ich dieses Ziel erreiche*, instead of the more fitting *Auf einer Skala von 1-10 gebe*

#	Example Sentence	Code
Smoking Cessation Forum		
S1	It is an addiction after all...	R-
S2	You should bear one thing in mind: if smoking were not addictive, hardly anyone would smoke.	R+
S3	It's actually more the constant thinking about it that bothers me and I find difficult.	Ra-
S4	I really want to "accomplish" this.	Rd+
S5	Where I need to take action is in my processes/routines.	Rn+
S6	My experience(s) in December is definitely that I should be active during the day.	Rn+
S7	Hello everyone, I've just registered and today I took my first Champix.	TS+
S8	my willpower and my belief in myself have diminished greatly due to the many failed attempts.	Ra-
S9	I will continue to follow the path, no matter what the weight says, and make an appointment with a doctor when I get the chance.	C+
Instagram Data		
I1	My thighs were burning from time to time, but I kept going.	TS
I2	Problem: Stomach gets in the way, lots to do, lack of time & structure, the inner bastard is too big, lack of motivation, neck pain, low energy level Goal: Feel good in my own body, be able to play with my daughter, lose 10-15kg in a year, want to be able to play football again, get rid of my stomach, increase my energy level I've been travelling alone for 2 weeks now and I've already lost a lot of weight - we'll crack the 15kg mark in February!	Rd
I3	I was really scared when I checked the seats in the front of the queue and was really afraid that this long bar wouldn't click into place.	R
I4	2 kilos less The best thing about the week was simply that I signed up for fitness.	TS
I5	I always knew that I was strong...	Ra
I6	I did a bit, but unfortunately I also spent quite a lot of time sitting on the couch watching series I was somehow under a spell Do you know that?	TS
I7	I've resolved to prepare food more often now, because it's not so practical without it, because then I reach for quick and often ready-made food and afterwards I don't feel as good as I should...	C
I8	Oh man, you wouldn't believe how good this training session did me today.	R

Table 5.9: Example sentences of Smoking Cessation Forum and Instagram Data.

ich mir selbst eine 8 für das Erreichen dieses Ziels. This slightly changes the meaning of the sentence: instead of focusing on the likelihood that the goal will be reached, it now sounds like the eight will be given if or when the goal is reached. While, during annotation, the original was taken into account as a reference in such cases, it might lead to inconsistencies during classification. The same goes for the missing context encountered in these utterances. During annotation, I referred to the conversation partner's utterance, which is not taken into account by the current classifiers.

Generally, comparing weight loss related datasets to the others reveals that severe overweight represents a more incisive and limiting experience in everyday life, leading affected persons to cease being able to do things (O2, O3) and to avoid situations, for instance, due to fear of embarrassment (G11, I3).

5.4 Chapter Summary

In this chapter, I have shown the usefulness of classifiers trained on the GLoHBCD across contexts and domains, which has allowed me to draw conclusions on the nature of behaviour change language

#	Example Sentence	Code
Health Coaching Dialogue Corpus		
H1	A quick 30-minute walk in the neighbourhood or a night-time cleaning session	C+
H2	My 5 minutes on the elliptical ended!!!	TS+
H3	On a scale of 1-10, I would give myself an 8 if I achieved this goal. [TRANSLATION ERROR]	Ra+
H4	I was just so tired.	R-
H5	My biggest problem was that I let myself down, and I overcame it by asking God for help and motivation for my journey.	R+
H6	Let's say 3500	C+
H7	No, I was just walking my dog and he made me run after him, plus I still had my 4 grandchildren.	TS+
Optifast Mock Chatbot Data		
O1	I am very satisfied.	R+
O2	The change in behaviour has currently worked very well, I already feel much better, I can walk again and my strength and fitness have improved considerably thanks to all the exercise.	R+
O3	Overweight, clothes no longer fit, everything was exhausting	R+
O4	I have no doubts.	R+
O5	I can't imagine that, as I'm also entering a new phase of my life.	C+
O6	My goals are...weight control....Try out new recipes to change my diet and stick to them. and exercise regularly.	C+
O7	My inner bastard sometimes stops me, but I keep fighting...	R+
DARN-CT-based WoZ Dialogues		
D1	I could keep a list hanging in an obvious place to remind myself daily.	C+
D2	Heavy training paired with iron discipline	C+
D3	Motivation.	Rn+
D4	I would just have to overcome the comfort.	Rn+
D5	Motivation to do sport and motivation to start doing things that have a deadline	Rn+
D6	I will set myself a time limit	C+
D7	I have deleted Instagram	TS+
D8	Ambition and mental strength help me to realise my goals.	R+

Table 5.10: Example sentences of Health Coaching Dialogue Corpus, Optifast Mock Chatbot Data, and DARN-CT-based WoZ Dialogues.

and its similarity regardless of target behaviours and conversational contexts. I have also analysed the distinct datasets in relation to one another using automatic textual analysis combined with qualitative insights, unveiling differences between contexts and domains. However, these differences did not seem to have a stable impact on the feasibility of transfer learning for the different datasets. My results indicate that it is feasible to use these classifiers to predict user utterances for various target behaviours and in live-chat-like human-bot interactions, as they are to be found in a CA. Although classification and transfer learning works fairly well, the classifiers leave room for improvement. In the next chapter, I will explore data augmentation and semi-supervised learning techniques to achieve better classification performance across domains and contexts.

Experiments to Improve Classification Results

To be useful in a chatbot, the classifiers fine-tuned on the GLoHBCD should be as accurate as possible. Thus, a number of experiments were conducted with the goal to improve classification performance. To achieve this, the GLoHBCD was expanded in different ways, which I outline in this chapter.

Low resource scenarios are a common problem in NLP research, as the creation of datasets often includes considerable amounts of manual work (Snow, O'Connor, Jurafsky, & Ng, 2008). There are a number of approaches that have been used to mitigate the costs of data collection and data scarcity in the past, mainly focused on semi-supervised learning (SSL), which uses weakly labelled data gathered using noisy techniques (Duarte & Berton, 2023; van Engelen & Hoos, 2020), and data augmentation, which exploits different techniques to create new data from existing texts (Feng et al., 2021; Shorten, Khoshgoftaar, & Furht, 2021). Most commonly, SSL is applied to text classification in the form of self-learning, which means that a baseline classifier is used to assign pseudo-labels to new, unlabelled data. In a second step, this pseudo-labelled data is included in the training data of the classifier (Duarte & Berton, 2023). Data augmentation techniques in the context of NLP include synonym replacement, back translation, and paraphrasing (Feng et al., 2021; Li, Hou, & Che, 2022; Shorten et al., 2021). Generative LLMs open up new possibilities in this area: Few-shot learning or fine-tuning LLMs on existing data and generating large amounts of text has been shown to be successful in domains such as hate speech detection (Shorten et al., 2021; Wullach, Adler, & Minkov, 2021).

In this chapter, I describe two approaches I tested to increase classification performance in the low-resource task of behaviour change language classification. For both of these approaches, the performance of the classifiers introduced in the previous chapters on each of my collected datasets is taken as a baseline. I begin by outlining experiments conducted on the efficacy of including the unlabelled GPT-3 generated data described in Section 5.1.4 into the training data (Section 6.1), whereas Section 6.2 compares the performance of SSL-enhanced classifiers leveraging unlabelled human-generated data from different sources to the baseline.

6.1 GPT-3-Generated User Simulation

This section is based on the following publication, which has been published under CC BY 4.0 license: :

Selina Meyer, David Elweiler, Bernd Ludwig, Marcos Fernandez-Pichel, and David E. Losada. 2022. "Do We Still Need Human Assessors? Prompt-Based GPT-3 User Simulation in Conversational AI." In *Proceedings of the 4th Conference on Conversational User Interfaces (CUI '22)*. Association for Computing Machinery, New York, NY, USA, Article 8, 1–6. <https://doi.org/10.1145/3543829.3544529>.

Modifications have been made to the original content.

First, the complete synthetic data described in Section 5.1.4, a small part of which was used to create the Synthetic GPT-3 Data, was used as a way to artificially increase the size of the GLoHBCD¹. A controlled experiment was designed to evaluate the potential of this dataset to increase classification performance. The synthetic data was processed the same way as the GLoHBCD and split into sentences. I then used the synthetic data in different ways to build four synthetically enhanced datasets:

- **mixed** combines GLoHBCD and synthetic data, where topics assigned to the synthetic dataset corresponded to the question prompted to GPT-3 to generate the output. For instance, when the question was intended to elicit an Rd statement, the topic assigned to the GPT-3 output was R, and the reason type was d. Since we did not specifically design questions to elicit change or sustain talk when generating the data, this condition was not applied to the valence level.
- **mixed predicted** combines GLoHBCD and synthetic data, where the labels of the synthetic data are classified with a 95% confidence threshold by the baseline classifier trained on GLoHBCD. This means that only synthetic data points which could be classified with a confidence higher than 95% by the baseline classifier were included in the training data.
- **mixed predicted balanced** contains undersampled versions of the predicted synthetic data at confidence 95% to increase the size of minority classes to the size of the majority class on each label level, creating datasets with fully balanced classes.
- **mixed predicted balanced increased** also balances the datasets so that each class is of the same size but additionally increases the overall size of the dataset. This was only possible for the valence level since the other label levels did not contain enough new samples predicted as the minority class to allow for an increase of the complete dataset (i.e. for the topic level, the largest class (R) in the training set contains 2448 data points, and the smallest (TS) only 363, but only 2223 GPT-3 sentences are classified as TS by the base classifier).

In Table 6.1, I give an overview of the datasets and label distributions compared to the GLoHBCD. Following the construction of synthetically-enhanced datasets, I fine-tuned BERT_{base} German-cased across three epochs to each training set using 10-fold cross-validation. At each fold, I predicted both on the GLoHBCD test set and the datasets described in Section 5.1. The results of these classification experiments are presented in Section 6.1.1. In addition to the classification experiments, I also analysed and compared the GLoHBCD and the synthetic data training sets on a structural and semantic level. The results from these comparisons, as well as a reflection on the meaningfulness of the GPT-3 outputs,

¹In the following the term "synthetic data" refers to the entire generated data, whereas Synthetic GPT-3 Data refers to the smaller annotated subset.

Label Level	Training Set	Size				
Valence			%+	%-		
	mixed predicted	28074	84.37	15.63		
	mixed predicted balanced	5233	50.09	49.91		
	mixed predicted balanced increased	8800	50.14	49.86		
	GLoHBCD	3779	69.36	30.64		
topic			% R	% TS	% C	
	mixed	24,508	68.01	18.41	13.59	
	mixed predicted	22,128	71.09	18.86	10.05	
	mixed predicted balanced	7344	33.33	33.33	33.33	
	GLoHBCD	3,779	64.78	25.62	9.6	
reason type			%R ₋	%R _a	%R _d	%R _n
	mixed	12248	33.33	24.38	23.89	18.39
	mixed predicted	10746	37.36	15.56	37.3	9.77
	mixed predicted balanced	6252	26.92	26.92	26.92	19.24
	GLoHBCD	2448	68.75	15.77	9.23	6.25

Table 6.1: Overview of training datasets based on GLoHBCD and synthetic data.

are outlined in Section 6.1.2.

6.1.1 Classification Results

The approach yielded vastly different results for the different label levels (see Figure 6.1). The only similarity between label levels was that predicting the synthetic data and assigning pseudo-labels before adding it to the training set consistently led to better results on test sets than using the labels derived from the questions prompted to GPT-3.

6.1.1.1 Valence

For valence classifications, balancing the dataset had a positive impact on the classification of the GLoHBCD test set, as well as the Health Coaching Dialogue Corpus, and the DARN-CT-based Wizard of Oz Dialogues, with the **mixed predicted balanced** condition performing slightly better than the **mixed predicted balanced increased**. Except for the Optifast Mock Chatbot Data and the Synthetic GPT-3 Data, synthetically balancing the data always led to better results than adding all high-confidence predicted data points.

6.1.1.2 Topic

Only one condition led to a classification improvement on two datasets for the topic level. The Smoking Cessation Forum and Health Coaching Dialogue Corpus were classified slightly better by the **mixed predicted balanced** classifier than by the base classifier. However, none of the other conditions led to improvements, and none of the other test sets benefitted from the classifier enhancements. While predicting the synthetic data led to better results than inferring labels from the questions posed to

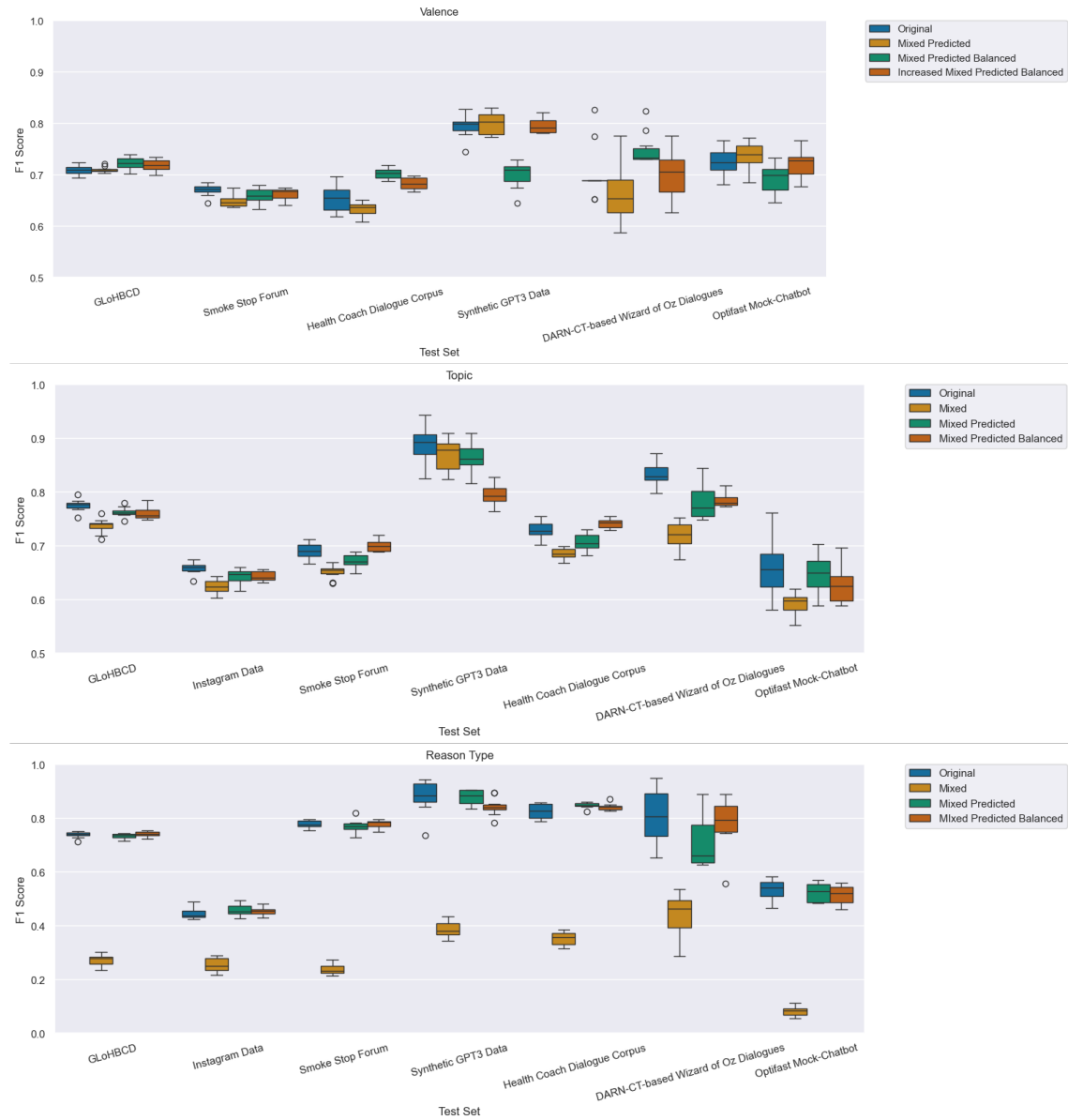


Figure 6.1: Classification results when applying synthetically enhanced classifiers to out-of-context test sets.

GPT-3, synthetically balancing the dataset had a less consistently positive impact than for the valence classifier. Apart from the two balanced classifiers which outperformed the baseline classifier, there was only one dataset (DARN-CT-based Wizard of Oz Dialogues), in which balancing the dataset led to slightly higher median Macro F1 than adding all high-confidence predicted synthetic samples.

6.1.1.3 Reason Type

The differences between classification performance were most jarring for the reason type level, especially between the **mixed** and the remaining classifiers. Whereas the remaining classifiers performed largely similarly, with **mixed predicted** and **mixed predicted balanced** yielding little to no improvement over

	GLOHBCD	Synthetic data
# data points	4,724	25,910
# Tokens	73,021	270,109
# Types	6,496	4,263
Type-Token-Ratio (TTR)	8.9	1.58
Standardised TTR (n=1000)	41.45	20.6
mean token occurrence	11.18	57.25
# tokens >10 x occurrence	689	910
Mean Sentence Length	18.81	11.91
# significant keywords	403	165
Top 10 keywords	da (there), surgery (op), one (man), have (hab), times (mal), is (ist), the (das), still (noch), today (heute), goes (geht)	I (ich), would like (möchte), more (mehr), sport (sport), do (treiben), weight (gewicht), to (zu), have (habe), healthier (gesünderen), will (werde)

Table 6.2: Comparison of GLOHBCD and synthetic data.

the baseline classifiers across test sets, the **mixed** classifier led to a drastic decrease in Macro F1 scores, with median values dropping below 10% for the DARN-CT-based Wizard of Oz Dialogues. This suggests that while questions about recently taken steps, reasons for or against change, and commitments for the future (topic level) are generally answered on topic by GPT-3, the same assumption cannot be made for questions about abilities, desires or needs, i.e. it is not unlikely that a question about a desire is answered with a statement about perceived ability. Since GPT-3 is trained on vast amounts of naturalistic data, it is possible that a similar trend might be observed in interactions with real users.

6.1.2 Exploration of Quality and Suitability of Synthetic Data

Generally, this data augmentation approach showed some promise for the classification of valences, whereas the topic and reason type levels did not benefit. For the remainder of this section, I explore potential reasons for the limited impact of the data for the topic and reason type level. To this end, I compare the GLOHBCD and the synthetically generated data with regard to language variety, and, on the example of the topic level, explore differences in class distinctions and the overall relevance of the synthetic data for the classification of behaviour change language.

6.1.2.1 Comparison of Naturalistic and Synthetic Data

To compare the GLOHBCD and synthetic data on a structural level, I examined the two datasets in more detail, with a focus on language variability (see Table 6.2). Even though the synthetic dataset is over five times the size of the GLOHBCD, the unique word count is lower by more than 2,000 lemmatised words, corresponding to a TTR of 1.58. The GLOHBCD also has three times as many Bonferroni-corrected significant keywords at $p < 0.05$ compared to the synthetic data. This shows that a lot of words present in naturalistic data tend to be missing in synthetic data, while the opposite is less common. It also indicates less variability regarding both content and vocabulary in the synthetic data. From the top keywords of each dataset when compared to the other, it becomes apparent that many keywords in the GLOHBCD seem to be functional (i.e. the, is, there,...), while words in the

synthetic data seem to bear more meaning (i.e. sport, more, weight,...), indicating that synthetic data seems to utilise only few of the filler words common to natural, human-written language.

6.1.2.2 Semantic Similarity

To further test the assumption that the synthetic data has less variety than the user data, I created sentence embeddings for each sample in the datasets and compared the cosine similarity within and between topics and datasets using a pre-trained sentence transformer model for the German language² (see Table 6.3). On average, synthetic samples were much more similar to each other than the GLoHBCD. It also became apparent that, while the sentences in the GLoHBCD tended to be most similar to sentences of the same topic (Table 6.2(a)), this was not necessarily the case for the synthetic data, where R sentences showed more similarity with TS and C sentences than with themselves, and TS sentences were most similar to C sentences (Table 6.2(c)). When comparing the synthetic embeddings with the GLoHBCD embeddings, cosine similarities were similar to those within the GLoHBCD. Again, however, I noticed that synthetic sentences did not always have the highest mean similarity with GLoHBCD sentences from the same class. Both synthetic TS and R sentences seemed to be most similar to GLoHBCD C statements, although the difference in mean similarity was very small. This could mean that GPT-3 outputs for these labels contain certain words that are common for commitment utterances in the GLoHBCD.

	TS	C	R
TS	0.2525	0.2484	0.2325
C		0.2869	0.2454
R			0.251

(a): Within text similarity of user sentences for each label.

	u	TS	C	R
s	TS	0.2627	0.2639	0.2440
C		0.2593	0.2873	0.2488
R		0.2398	0.2549	0.2502

(b): Between text similarity of synthetic (s) and user (u) sentences.

	TS	C	R
TS	0.3376		
C	0.3427	0.3894	
R	0.3043	0.3294	0.3171

(c): Within text similarity of synthetic data for each label.

Table 6.3: Mean cosine similarity between all sentence embeddings of synthetic and user training sets within and between labels and datasets. Calculated with Sentence Transformers.

6.1.2.3 Classification of Synthetic Data

The results above might give the impression that the generated outputs are simply not relevant to the classification task at hand, leading to low classification performance of the synthetic classifier on the test sets. To explore this assumption, I used the **mixed** and **mixed predicted** topic level classifiers, as well as a classifier trained on only synthetic data, to predict a 20% test split of the synthetic, **mixed** and **mixed predicted** dataset at each cross-validation fold (see Table 6.4). From this, I draw three interesting observations:

- Combining synthetic data with the GLoHBCD leads to better classification results on the synthetic test set.
- Classifying synthetic data with a classifier trained on the GLoHBCD is much more reliable than

²<https://huggingface.co/Sahajtomar/German-semantic>.

the other way around.

- The mixed predicted test set is easiest to predict for all classifiers, regardless of label ground truth used for fine-tuning.

Classifier	Test Set		
	Mixed	Mixed Predicted	Synthetic
GLOHBCD	72.43 (1.5)	92.77 (0.5)	71.51 (1.79)
Mixed	79.93 (0.2)	84.31 (0.49)	81.11 (0.29)
Mixed Predicted	71.6 (0.84)	94.26 (0.22)	70.64 (1.05)
synthetic data	78.2 (0.38)	81.69 (0.53)	80.96 (0.24)

Table 6.4: Mean Macro F1 score (standard deviation) in percent for synthetic, mixed and mixed predicted test sets.

These observations show that the synthetically generated data does bear similarity to the user dataset and that the notion of generating labelled data with prompts is feasible to a certain extent. If this were not the case, the prediction of synthetic data with the user classifier would have led to many label changes as compared to the intended classes and worse performance of the synthetic and mixed classifier on the mixed predicted test set.

6.2 Semi-Supervised Learning to Facilitate Data Increase From Naturalistic Sources

This section is based on the following publication which has been published under CC BY 4.0 license:

Selina Meyer, Marcos Fernandez-Pichel, David Elswailer, and David E. Losada. 2024. “Exploring the Challenges of Behaviour Change Language Classification: A Study on Semi-Supervised Learning and the Impact of Pseudo-Labelled Data.” In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024*, pages 229–239, Torino, Italia. ELRA and ICCL.

Modifications have been made to the original content.

The previous approach to improve classification performance heavily relied on natural language generation as a way to increase the size of the original datasets. One of the main findings was that the generated data did not properly reflect the variety found in the GLOHBCD, which resulted in limited impacts on classification since hard-to-classify samples in the test data were not accounted for. Inspired by this discovery, the second approach I tested to improve my classifiers was using the base classifiers to assign pseudo-labels to new, user-generated data from 1) the same source as the GLOHBCD, and 2) other sources related to behaviour change. This type of SSL has been shown to be successful for a variety of classification tasks (Duarte & Berton, 2023; van Engelen & Hoos, 2020). Real, user-generated data is more likely to include hard-to-classify utterances as found in the GLOHBCD than

synthetically generated data and adds enough new information to the dataset to potentially facilitate the classification of data from the same and other contexts. This largely follows the same approach as I applied in Section 6.1 to the **mixed predicted** datasets, but in this case, the base classifier assigns pseudo-labels not to synthetic data but naturalistic data generated by users in different contexts.

6.2.1 Sources of Data for Pseudo-Labelling

I used three sources to collect data for pseudo-labelling, which differ in their context and domain similarity to the GLoHBCD and the out-of-context test sets. I describe each new data source and how they relate to the GLoHBCD and behaviour change language in general below.

6.2.1.1 Weight Loss Forum Data

For the Weight Loss Forum Data, I collected new posts from the same source as the GLoHBCD. I collected all posts published after the extraction date of the initial GLoHBCD data (August 2020). There was no manual pre-screening or pre-filtering of this new data, which consists of 992 sentences and serves as in-context and in-domain data for pseudo-labelling.

6.2.1.2 Smoking Cessation Forum

As a source for out-of-domain (different target behaviour) but in-context (same conversational context) data, I chose the Smoking Cessation Forum data, which has the additional benefit that it already has gold labels assigned to it. I created a stratified 70/30 split of the dataset for each label level and used 70% for pseudo-labelling. This reduces the size of the Smoking Cessation Forum test sets in this range of experiments (199 sentences for valence and topics, 131 for reason types), but also allows me to determine to what extent adding pseudo-labelled data from a new source can be used as an approach to increase performance for that source.

6.2.1.3 AnnoMI

The AnnoMI dataset is a collection of transcribed MI sessions across a variety of behaviour change contexts (Wu, Balloccu, Kumar, et al., 2022). While still being language data related to behaviour change uttered by humans, this dataset differs both in context (topics range from weight loss and smoking cessation, across alcohol abuse to other issues) and source type, as it is spoken data from counselling sessions, whereas the GLoHBCD consists of peer-to-peer conversations in a written forum. Since the data stems from counselling sessions, the AnnoMI may also include conversations with people who are not willing to change, thus depicting a wider range of stages of change than the GLoHBCD, which contains texts mostly from people seeking help, which in itself indicates a certain willingness to change. As such, this dataset is the furthest away from the original dataset and could thus offer the largest increase in linguistic variance. Since the dataset does not include ground truth annotations for all label levels but does include valence and follow/neutral labels, I remove all client utterances annotated as follow/neutral and use all remaining client utterances as data to pseudo-label for my experiments. I

translate the AnnoMI into German using a transformer-based machine translation model for German-English translation developed by the University of Helsinki³ (Tiedemann & Thottingal, 2020) and separate the utterances into sentences following the same approach as for the other datasets, resulting in 2481 sentences.

6.2.2 Methods

I ran experiments across four stages. The first three stages are made up of fine-tuning experiments, whereas the fourth stage applies the findings to the test sets. For all stages that involve fine-tuning, I followed a similar methodology as described in Section 6.1: In 10-fold cross-validation, I fine-tuned BERT_{base}German-cased on the GLoHBCD for each label level, which serves as the baseline classifier. I then pseudo-labelled the new data and combined it with the GLoHBCD training set to fine-tune an SSL classifier. The SSL classifier and the baseline classifier were then validated against the validation split, where I recorded Macro and class-wise F1 scores. I made one major change in methodology compared to the previous data augmentation experiments: Since this range of experiments involved multiple stages, which relied on results obtained from previous stages, I did not predict the out-of-context test sets in cross-validation. Instead, I used the classification results of validation splits to determine the most promising SSL classifier and kept the final evaluation on out-of-context and out-of-domain test sets for the last experimental stage.

6.2.2.1 Stage 1: Confidence Thresholds

Since the success of SSL methods heavily relies on the selection of added data (van Engelen & Hoos, 2020), and one way to determine the suitability of data is by the confidence with which the base classifier assigns labels to the new data, Stage 1 consisted of the comparison of different confidence thresholds applied to the pseudo-labelled data. I filtered the novel data in two ways: first, I fine-tuned a binary BERT_{base}German-cased classifier on the follow/neutral sentences encountered during the creation of the GLoHBCD (see Chapter 4) and the GLoHBCD. This classifier reached a Macro F1 of 72.67% on a test split and can serve as a prefilter to determine whether the new sentences are related to behaviour change or not⁴. I tested three different confidence thresholds (0.5, 0.7, 0.95) for the inclusion of new data as related to change. Second, I tested the same confidence thresholds for assigning pseudo-labels with the baseline classifier. I tested all combinations of prefilter-baseline classifier thresholds (e.g. prefilter confidence 0.5, baseline confidence 0.7) and an additional condition in which no prefilter was applied, resulting in 21 prefilter-baseline combinations for each label level. To keep noise low at this stage, I only used the Weight Loss Forum Data, as it stems from the same source as the GLoHBCD.

³<https://huggingface.co/Helsinki-NLP/opus-mt-en-de>

⁴The prefilter is made available on huggingface: https://huggingface.co/selmey/behaviour_change_prefilter_german.

6.2.2.2 Stage 2: Accounting for Class Imbalance

Data imbalance introduced by pseudo-labelling can hinder classification even with a balanced base classifier (X. Wang, Wu, Lian, & Yu, 2022), especially when the original dataset is imbalanced. This can negatively impact performance (Guo & Li, 2022). Due to the highly imbalanced nature of my datasets, the baseline classifiers are biased towards the majority classes, resulting in a larger amount of new data being pseudo-labelled as the majority class as well. To address this, I tested strategies to boost the size of the minority classes. For this stage, I first selected the best prefilter-threshold combination for each label level. I then tested the following variants to expose potential labelling biases: i) adding only minority class pseudo-labelled samples, ii) adding an equal amount of pseudo-labelled samples for all classes based on the number of minority class samples, and iii) labelling as minority class even with low confidence (thresholds: 0.4, 0.3, 0.2, 0.1).

6.2.2.3 Stage 3: Amount and Domain of Pseudo-Labelled Data

After determining which confidence thresholds and data balancing approaches are best suited when incorporating pseudo-labelled data from the same source, I explored the effects of adding different amounts of pseudo-labelled out-of-context and out-of-domain data to the training set. I tested the following combinations of new data sources: i) Weight Loss Forum only, ii) Smoking Cessation Forum only, iii) AnnoMI only, iv) Weight Loss Forum + Smoking Cessation Forum, v) Weight Loss Forum + Smoking Cessation Forum + AnnoMI. Given the predictions of the baseline classifier with the best performing confidence thresholds as determined in stages 1 and 2, I added varying shares of pseudo-labelled data to the GLoHBCD training set (between 20%-100% in increments of 20%). Since the Smoking Cessation Forum data contains ground truth labels, I incorporated these examples into the 10-fold cross-validation experiments. So, in addition to the GLoHBCD validation split, I also validated on a 10% validation split of the Smoking Cessation Forum training data at each fold, pseudo-labelling the remaining 90%. This allowed me to draw first conclusions about the impact of adding out-of-context data on transfer learning capabilities.

6.2.2.4 Stage 4: Application to Test Sets

In this final stage, I combined insights from stages 1-3 and applied the best-performing system for each label level to the out-of-context test sets.

6.2.3 Results

Stages 1 and 2 yielded minor improvements to classification results, although none of them led to significant improvement. In Table 6.5, I display the three best SSL classifiers per label level compared to the baseline classifier.

Analysis of performance on single classes revealed that the F1 scores of minority classes varied more when pseudo-labelled data was added than those of the majority classes. I also found that the number of sentences labelled as the minority class was quite low, even for conditions where the confidence threshold

Valence	Macro F1	Topic	Macro F1	Reason Type	Macro F1
baseline	72.65 (1.96)	baseline	74.05 (3.17)	baseline	75.63 (2.93)
NP, CT(0.5)	73.98 (2.29)	NP, CT(0.95)	75.89 (3.82)	P(0.7), CT(0.5), min	76.7 (2.96)
NP, CT(0.5), equal	73.73 (1.85)	P(0.5), CT(0.95)	75.46 (3.47)	P(0.7), CT(0.5)	76.28 (4.33)
NP, CT(0.1)	73.55 (2.36)	P(0.7), CT(0.95)	75.35 (3.15)	P(0.7), CT(0.4)	76.25 (3.35)

Table 6.5: Comparison of SSL classification results in cross-validation with baseline with (P(.)) and without (NP) prefiltering. Confidence thresholds for prefiltering and classification are indicated as P(t) and CT(t), respectively. If the threshold was < 0.5 , data points were labelled as minority class if the predicted confidence was above the threshold. Equal: equal samples were included for each class, Min: only new minority class samples are included.

for minority class labelling was lower than 0.5. This effect was visible across all label levels (see Figure 6.2, indicating a strong labelling bias of the baseline classifier towards the majority class. Although I had anticipated this bias and taken countermeasures by testing different approaches to strengthen the minority class in Stage 2, the results indicate that these strategies were of limited success.

Nevertheless, although there were no significant improvements, there was also no SSL condition in which mean Macro F1 scores across cross-validation were lower than baseline classification results. This indicates that adding pseudo-labelled samples from the same source as the original training data has the potential of positively impacting classification, although the size of the unlabelled dataset might have been too small to lead to significant improvement. This was owed to the little amount of traffic on the weight loss forum between the collection of the GLOHBCD and the scraping of the new weight loss data.



Figure 6.2: Amount of pseudo-labelled data points added to the training data across conditions by class.

Stage 3 of the experiments aimed at adding larger amounts of pseudo-labelled data and, based on the insights from Stages 1 and 2, was expected to potentially yield larger improvements. I chose the best-performing classifier from the first two stages for each label level:

valence level: no prefilter, confidence threshold 0.5 (NP, CT(0.5)),

topic level: no prefilter, confidence threshold 0.5 (NP, CT(0.95)),

reason type level: prefilter, confidence threshold 0.5, and adding only minority samples (P(0.7),

CT(0.5), min).

Applying those systems in Stage 3 of the experiments led to more stable classification results for the GLoHB CD validation sets compared to the Smoking Cessation Forum validation sets. This was expected since the smoking cessation data is from a different source and domain than the original training data and has fewer samples, resulting in very small validation sets, which are more prone to extreme values.

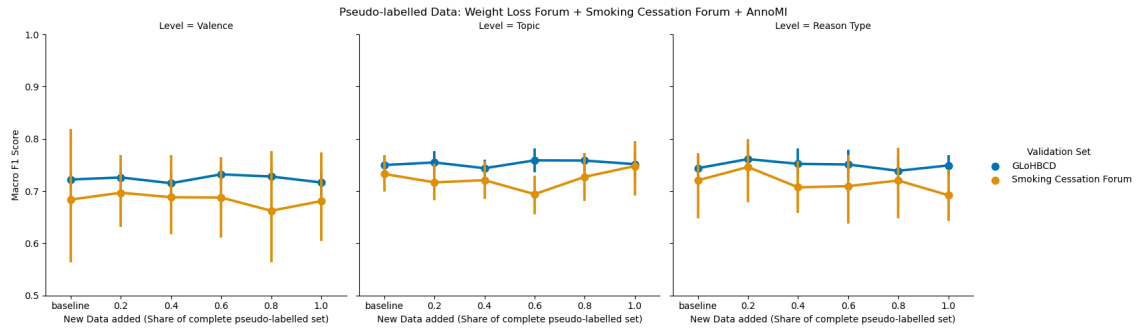


Figure 6.3: Change in classification performance when more pseudo-labelled data is added to training.

However, regardless of the validation set and the type of pseudo-labelled data added, the results did not show a clear increase in performance when more data was added. Only in a few instances did adding out-of-domain data lead to improvements in in-domain classification. The effects of the SSL classifiers seemed to be slightly more apparent on the reason type level. For example, adding weight loss forum data led to slight improvements in reason type classifications of the GLoHB CD and Smoking Cessation Forum validation sets. The reason type classifiers work with few labelled data points from the original training data, which could lead to pseudo-labelled samples having more influence on classification.

Based on the results obtained in Stage 3, I included different shares of pseudo-labelled data from Weight Loss Forum, Smoking Cessation Forum and AnnoMI in Stage 4 (see Figure 6.3). For valence and reason type classification, I included 20% of the pseudo-labelled data, and for topic classification, I added all pseudo-labelled samples. In Figure 6.4, I compare the classification performance of the baseline classifier with the best-performing system from stages 1-3 for each label level.

For the topic level, all datasets but the Health Coaching Dialogue Corpus experienced some improvement in performance. Effects on valence and reason type were larger and more varied. The valence predictions of the SSL classifier on the Smoking Cessation Forum, the Synthetic GPT-3 Data, and the Health Coaching Dialogue Corpus were better than those of the baseline. Still, the SSL classifier produced poorer valence results for DARN-CT-based WoZ Dialogues, GLoHB CD and Optifast Mock Chatbot Data. For reason types, decreases in performance were observed for the Synthetic GPT-3 Data and the GLoHB CD test set, while performance on Optifast Mock Chatbot Data remained the same, and all other datasets benefited from the inclusion of the pseudo-labelled data.

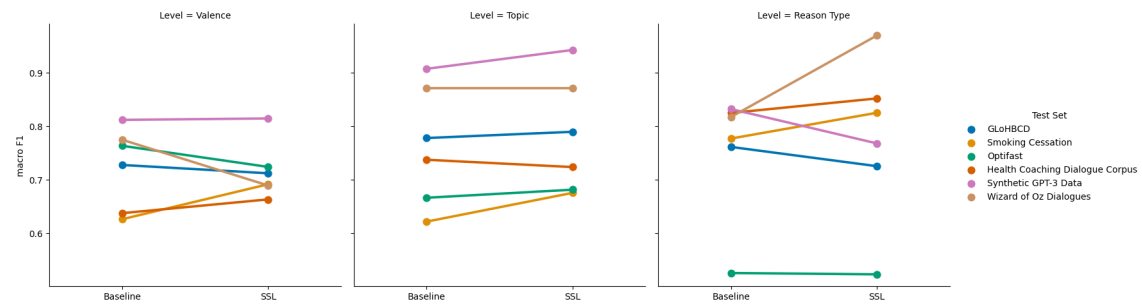


Figure 6.4: Comparison of classification results on multiple test sets. The plots present the performance of the baseline classifier (no semi-supervised learning) and a SSL classifier with different configurations based on the experiments of stages 1-3.

6.3 Chapter Summary

In this chapter, I have presented two approaches towards increasing classification performance, which I evaluated both with respect to their impact on in-domain, in-context classification and on the transfer learning capabilities of the classifiers. Since none of the approaches yielded improvements across datasets and label levels, I did not end up applying them to the classifications in the final chatbot. The limited success of these experiments has various potential reasons, such as the limited language variability of synthetic data, which could be confirmed in analyses presented in Section 6.1.2. This aligns with research which suggests that in the case of transformer-based models, due to the extensive amount of data used in pre-training, data augmentation might not introduce sufficient novelty and linguistic variety to improve classifiers (Longpre, Wang, & DuBois, 2020). Given the high initial transfer capabilities of the classifiers found in Chapter 5, the same might be true for the pseudo-labelled training data added during the semi-supervised learning experiments.

Another limiting factor could be the subjectiveness of annotation of behaviour change language, a common issue in this field, as shown by typically low to mid-range IRR. This can lead to a comparably large number of debatable sentences, which might represent multiple codes to varying degrees. Furthermore, behaviour change language is naturally imbalanced, with the code distributions potentially constantly changing as a person goes through different stages of behaviour change. Overall, however, this range of experiments has shown that, regardless of the domain and conversational context of data added to the training data during fine-tuning, classification performance mostly remains stable, speaking in favour of the robustness and transfer capabilities of the initial classifiers. Although the approaches were unsuccessful in increasing classification, they serve to increase the understanding of the datasets and, by extension, behaviour change language in written contexts. They also show that the classifiers are robust enough to not deteriorate in performance when noise is added.

Summary of Part II

Overall, this part of the thesis primarily focused on **understanding** the user. I outlined the creation and evaluation of the GLoHBCD, a dataset for the classification of written behaviour change language, as well as the fine-tuning of classifiers based on the dataset (Chapter 4). This was followed by experiments to showcase the classifiers' suitability for the classification of user utterances in human-CA interactions, which yielded promising results in favour of the application of the classifiers across various domains and conversational contexts (Chapter 5). Lastly, in Chapter 6, I outlined extensive data augmentation and semi-supervised learning experiments conducted in an attempt to improve the performance of the classifiers. These experiments yielded limited improvements. Overall, however, this range of experiments also showed that regardless of the domain and conversational context of data added to the training data during fine-tuning, classification performance mostly remains stable, speaking in favour of the robustness and transfer capabilities of the initial classifiers.

In the case of conversational AI for behaviour change support, however, understanding the user is only one part of successful interactions. The system also has to be able to react appropriately. This includes system outputs to be in line with user expectations while staying adherent to underlying conversational principles and techniques, in this case, MI, and bypassing potential harms that can arise from conversational human-LLM interactions. The next part presents my efforts towards identifying and increasing appropriate system behaviour in the context of LLM-based behaviour change support conversations, taking into account user behaviour and expectations in this context.

PART III

REACTING

Examining Efficacy, Safety, and the Role of User Behaviour

In Part II of this thesis, I outlined experiments geared towards understanding user utterances around change. This understanding can be used for utterance-level understanding but can also serve as a source of deeper insights about a user's stage of change and current situation. This part focuses on generating suitable and appropriate responses using LLMs. I focus on three main considerations for the generation of such responses. First, they should be appropriate in the context of behaviour change counselling. This necessitates the chatbot to take the role of a support provider and adhere to behavioural guidelines, for instance, MI guidelines. Second, they should bypass weaknesses and potential harms posed by LLMs specifically and CAs generally. This includes the potential for hallucinations and misinformation but also the danger of anthropomorphism and manipulation. Third, responses should be customised to fit the user's current situation and thoughts about change. To achieve this, I take into account user utterance classifications based on Part II in the dialogue design.

In this chapter, I describe a user study following a mixed methods approach, which leverages both quantitative and qualitative methods to evaluate the extent to which harmful outputs can be avoided and identify different types of user behaviour in this context. The pre-registered RCT¹ was driven by the following research questions:

- **RQ 3.1a:** Can a LLM-driven Conversational Agent, prompted to act like a therapist, be used safely as a tool to foster self-reflection about behaviour change?
- **RQ 3.1b:** Does the use of a conversational framework as theoretical grounding improve MI-adherence and the relationship between the user and the chatbot?
- **RQ 3.2:** Is there an association between user behaviour and conversation outcome?

To address these research questions, I created two versions of MIcha (**M**otivational **I**nterviewing **chat**bot), a GPT-4-based conversational agent following an MI-guided interaction protocol with the ability to converse about a variety of topics of behaviour change. The dialogue framework is described

¹The timestamped pre-registration and collected data is available on Open Science Framework: https://osf.io/c9nvm/?view_only=c321fc84c94546939b4c20729bbfafcc.

in Section 7.1. To identify unsuitable chatbot behaviours in the context of MI-focused LLM-based conversations about behaviour change, I created and evaluated an annotation scheme in Section 7.2). This annotation scheme complements existing computational methods created to evaluate the presence of appropriate behaviours in similar conversational contexts introduced in Section 3.3.2.3 in two major ways: 1) it focusses on the identification of *innapropriate* and *potentially harmful* behaviours specifically in 2) *LLM-based conversation around behaviour change*. Following the elaboration on conversational design and methods to identify unsuitable CA behaviour, I describe the study design in Section 7.3. The analysis of collected data regarding quantitative measures of the effects of the different interactions of user perception, output harmfulness and readiness to change, as well as qualitative analysis of user behaviour and expectations, follows in Section 7.4. Finally, in Section 7.5, I apply the classifiers introduced in Part II to the collected chats to identify users' language patterns throughout the conversations and explore to what extent users' language use is impacted by different factors.

This chapter makes contributions related to two main areas: 1) I explore the technical feasibility of using LLMs in the context of behaviour change and provide methods to evaluate and mitigate harm in the process, and 2) I provide a thorough analysis of user behaviour in such contexts, identifying user types who benefit from interactions with such systems to different degrees.

Parts of this chapter are based on the following publications:

- Selina Meyer and David Elweiler. 2024. "You tell me": A Dataset of GPT-4-Based Behaviour Change Support Conversations. In Proceedings of the 2024 Conference on Human Information Interaction and Retrieval (CHIIR '24). Association for Computing Machinery, New York, NY, USA, 411–416. <https://doi.org/10.1145/3627508.3638330>.

© 2024 Owner/Author(s). Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author(s).

- Selina Meyer and David Elweiler. "LLM-based Conversational Agents for Behaviour Change Support: A Randomised Controlled Trial Examining Efficacy, Safety, and the Role of User Behaviour" [manuscript under review]

Modifications have been made to the original contents.

7.1 Creating a Dialogue-Framework to React to User Utterances Based on MI Principles

To explore to what extent it is possible to get GPT-4 to behave in an MI-adherent manner and to increase comparability between chats, I created a conversational framework that guides interactions with the user. The conversational framework divides the conversation flow into three phases, incorporating the four major processes of MI: *engaging*, *focusing*, *evoking*, and *planning* (Clifford & Curtis, 2016). In this section, I describe each conversational phase in detail.

The first and third phases were predominantly template-based, since the introduction, conclusion, and

Function	Output	Answer Format	Purpose
Introduction	Hi, I am [] and I am here to talk to you about your behaviour change plans! What change, if any, would you like to make most?	Free text/choice	define Target Behaviour
Change Ruler Question	On a scale from 1 to 10, how confident are you in your ability to integrate [target behaviour] in your life?	Likert-Scale	Assess Readiness to Change
Probing Questions	You chose a [] on the scale. Why not [higher—lower]?	Free text	Motivate reflection Elicit first reasons for/against change

Table 7.1: Example template questions.

initial scoping questions are well-defined in the scope of MI (Clifford & Curtis, 2016; W. R. Miller & Rollnick, 2002), and do not vary much in structure. This means that the existing wording introduced in handbooks could be used to create templates with slots to be filled once the required information has been elicited. Table 7.1 provides some examples of such well-defined and pre-structured utterances. In contrast to this, the generated texts in Phase 2 were fully based on calls to the GPT-4 API. This phase distinguished between two versions of MIcha, which differed in the prompts passed to GPT-4 to solicit CA replies:

- In the **GPT-4** version, only the conversation history was passed to GPT-4, with no restrictions or instructions about desired output.
- In the **MI-adapted** version, specific steps were taken to understand the nature of user utterances and increase adherence to MI principles. The classifiers based on the GLOHBCD were applied to understand the content of the current user utterance with respect to behaviour change. This information was used to select suitable bot behaviours, and the prompt passed to GPT-4 was adapted accordingly.

A visual overview of a conversation with MIcha is presented in Figure 7.1.

7.1.1 Phase 1: Engaging and Focusing (4 Turns)

This phase served as an initial introduction to the CA and informed about the system's capabilities, goals, and focus. The user was greeted with the following introductory message:

Hi! I'm MIcha, your motivational chatbot. My goal is to help you reflect on your behaviour and develop a plan for your behaviour change. Please think about my questions carefully and consider them primarily as impulses that bring you mentally closer to your goal. I'm not here to give you tips or recommendations for action but to gain clarity about your motives and values. What behavioural change would you like to talk about today?

Following this introduction, the participant was presented with a dropdown menu to choose their preferred behaviour change. The system followed up with a simple reflection and asked the user for a reason why they want to change this behaviour, followed by an 11-point Likert-style question about their readiness, commitment, or motivation to change. Again, their indicated level was reflected back

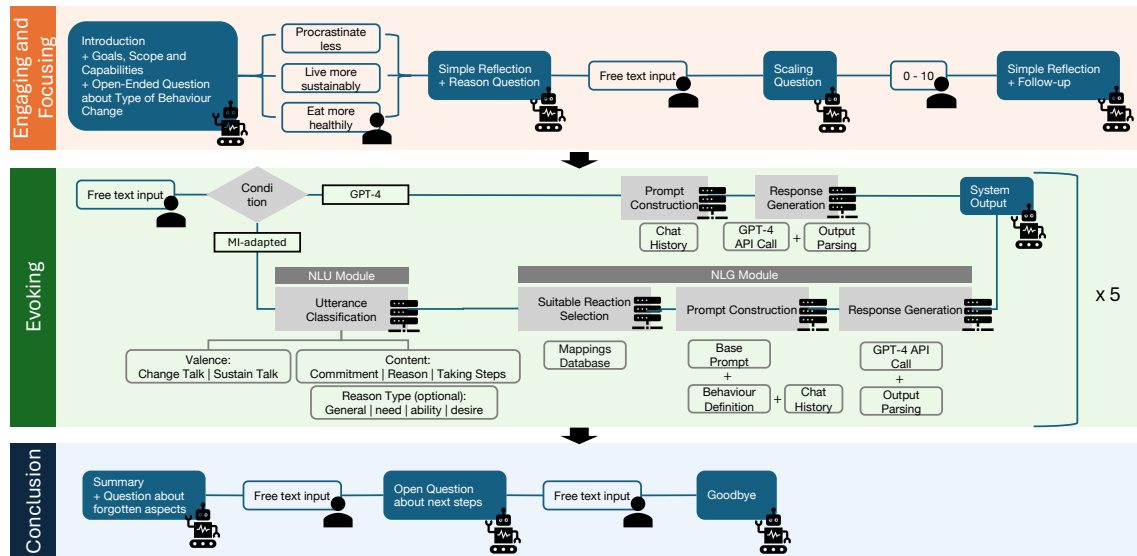


Figure 7.1: Visual overview of the flow of conversations with MIcha, depending on study condition.

to them, and they were asked to explain why they did not select a higher or lower number, transitioning to Phase 2.

7.1.2 Phase 2: Evoking (5 Turns)

During this part of the conversation, all bot utterances were generated by GPT-4, with different prompts passed to the system depending on the assigned condition. This was the only section of the study where the two conditions differed. In the **GPT-4** condition, the conversation history and current user turn were passed to GPT-4 with no additional information or instruction to solicit a response. In the **MI-adapted** condition, I included two additional components, which I conceptualise as the NLU (natural language understanding) module and NLG (natural language generation) modules.

The NLU module classified the valence and content of the last user turn using the classifiers introduced in Part II. The NLG module then selected a suitable response depending on the type of utterance from a pre-defined mapping based on MI literature (Clifford & Curtis, 2016; W. R. Miller & Rollnick, 2002). Definitions from the same literature were then passed to GPT-4 as part of the system prompt, along with the conversation history (see Figure 7.1). Figure 7.2 provides examples of user utterances, their classifications, the selected suitable response, the system prompt passed to GPT-4 and the corresponding output generated by GPT-4. The complete utterance-type–suitable response mapping is available on Open Science Framework².

Fulfilment of Prompted Behaviour

Since it is likely that the quality of dialogue depends on the chosen LLM, I tested the performance of different GPT variants on chat logs from pilot conversations. I offered a convenience sample of seven participants (two female and five male between 25 and 54 years) the chance to converse with the system,

²https://osf.io/c9nvm/?view_only=c3bc1a755b4d42e4a550d88963a71f3a

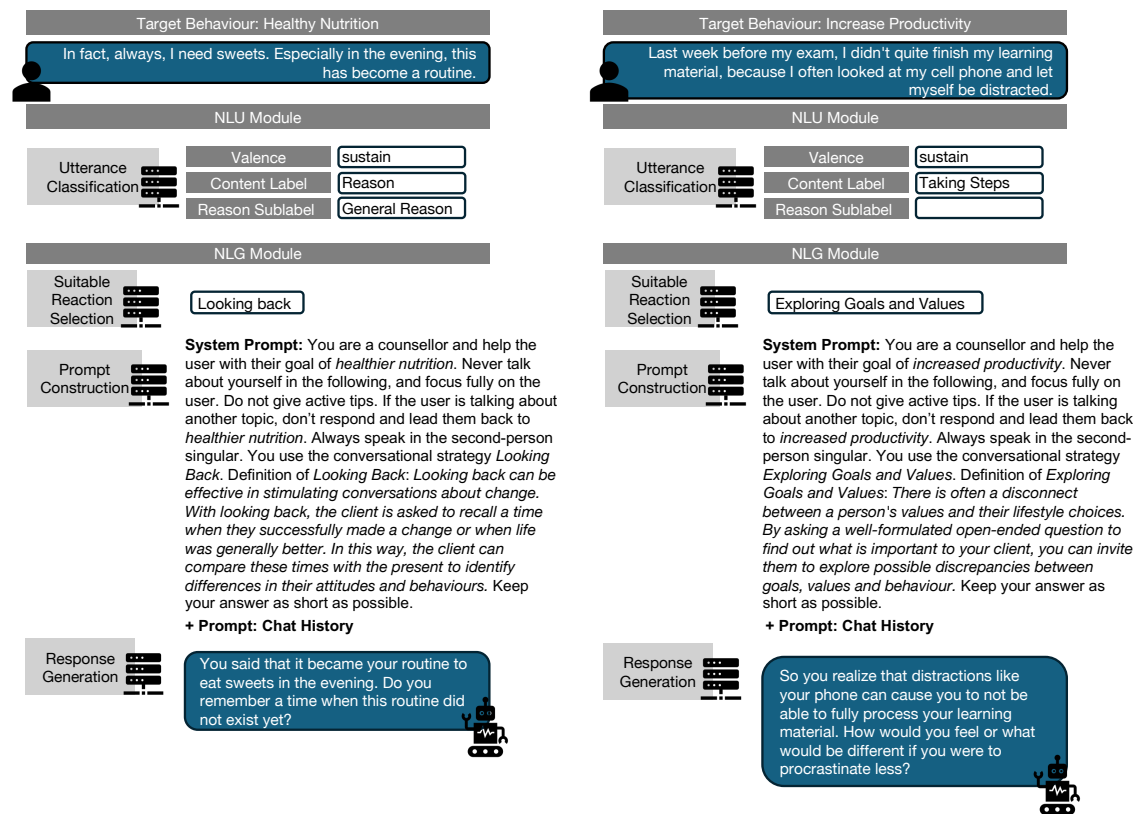


Figure 7.2: Example of NLU and NLG module usage.

using GPT-3 as a backend. Following this, I used each user utterance-selected behaviour combination from the chat logs to generate CA turns using GPT-4 and GPT-3.5 Turbo and systematically evaluated how frequently outputs of the different GPT models fulfilled the prompted behaviour. GPT-4 performed best (78.4%), followed by GPT-3 (61.9%) and then GPT-3.5 Turbo (47.5%). GPT-4 was also the only system tested that did not return a single output that could be regarded as a “system failure”. Such system failures were, for example, instances where the output was either a copy of the prompt given to the model or a single word or sign (e.g. “client:” or “?”). For this reason, GPT-4 was chosen as a model for the main study despite its higher cost compared to the other models.

7.1.2.1 Process and Parameters

In both conditions, the text generation process was repeated a total of five times, after which the transition to the last phase was triggered. For each generated turn, a user rating of either good/helpful, bad/inappropriate, or offensive/hurtful was solicited, along with a voluntary explanation of why the rating was given. Furthermore, to ensure that participants are not overwhelmed with multiple questions or text walls, the maximum number of response tokens was set to 100 and “?” was passed as a stop sequence. This caused GPT-4 responses to be cut off before the first question mark in the output or as soon as the number of generated tokens reached 100. It is worth noting that the max_token parameter does not impact the content of the response, resulting in some bot responses being cut off in the middle of a sentence.

7.1.3 Phase 3: Conclusion (3 Turns)

In this conversation phase, the system summarised the conversation (GPT-4 generated) and invited the participant to add any forgotten content (“Have I forgotten anything?” is concatenated to the end of the generated summary). The participant was then invited to define concrete next steps to take, and the system ended the conversation with a goodbye message. Except for the final bot utterance, participants were again prompted to rate the CA’s utterances and optionally explain their rating choice.

7.2 Identifying Unsuitable and Harmful Outputs

Harmful outputs were identified in two ways: First, I took participants’ ratings of generated turns into account, specifically with respect to turns rated as *offensive/hurtful*. Secondly, I drew on the literature on MI, empathy research, and academic ethical discussions about LLMs outlined in 3.3.1 to identify behaviours that are not suitable specifically in the context of LLM-delivered MI conversations. Based on this, I created an annotation scheme to be applied to the logged conversations after data collection. This annotation scheme goes beyond the identification of MI-inadherent behaviours defined in the *MISC* (listed in Table 7.2). While such behaviours are not suited for MI interactions in general, I laid special emphasis on actions that would be suitable in MI interactions with a human counsellor but should be avoided when responses are generated by LLMs. I focussed specifically on significant concerns around anthropomorphism and the unreliable factual correctness of LLM outputs. I applied extreme caution when compiling the annotation scheme to catch any potentially harmful outputs. This included the incorporation of anything that could lead a user to conclude the chatbot is capable of experience. Furthermore, in the context of this user study, I deliberately chose to explore target behaviours that are not health-critical in order to minimise ethical harm in the study. As a result, any advice or facts given in chats concerning these topics are unlikely to be harmful in itself. However, the same cannot be said for other (more health-related or critical) topics. The lack of control over what exactly is generated harbours the potential for detrimental advice, no matter how harmless the advice given within the user study might seem, which led me to create very strict guidelines for what counts as harmful advice and facts. Due to the potential ethical concerns of such systems, I followed the rationale to err on the side of caution.

The following items were identified as potentially harmful:

- **Giving Facts** – The system output contains facts or assertions.
- **Advise** – The system gives advice, suggestions or offers a solution, either on how to handle a certain situation or how to think about behaviour change.
- **Share Personal Information** – The system presents itself as human or discloses information from its “own life”.
- **Feeling** – The system reveals “own emotions”, including emotions such as joy about a certain user statement.

To test the validity of the annotation scheme, I annotated each GPT-4 generated turn collected in the user study along the four dimensions of undesired behaviour outlined above, allowing multiple labels

Behaviour	Adapted Definition
Confront	The system is openly confrontational, critical or judgemental
Direct	The system makes a request or gives a command
Warn	The system warns the user of negative consequences of action

Table 7.2: MI-inadherent counsellor behaviours adapted from W. R. Miller et al. (2008).

per turn. Based on the annotation scheme, GPT-4 outputs containing advice, factual information, and outputs sharing personal information or revealing emotions, as well as outputs with directive or confrontational wording, were marked as harmful. Subsequently, a second annotator³ annotated a share of the turns along the four axes. 50 randomly sampled turns, along with the guidelines, were given to the annotator for annotation in a training round.

I applied Cohen's κ to measure IRR. The first round of annotations led to high IRR scores for the advise dimension but low scores for the other three dimensions, leading us to accept the validity of advise annotations⁴. Following a discussion and refinement of the guidelines, a second set of 40 turns was given to the second annotator, including only the three remaining dimensions. After the second round of annotations, satisfactory IRR was reached for all dimensions (see Table 7.3). The lowest values were achieved for Giving Facts. However, in most cases, this was due to confusion about whether certain wordings should be marked as Giving Facts or Advise, not whether an output was to be marked as harmful or not. Thus, I took the IRR values as sufficient to accept the validity of the remaining annotations. The finalised annotation scheme is provided in Table 7.4.

Dimension	Cohen's κ	
	Round 1	Round 2
Giving Facts	0.3	0.54
Advise	0.87	–
Share Personal Information	0.47	0.79
Feeling	0.56	0.8

Table 7.3: Inter-rater reliability of annotations of undesirable bot behaviours after one round of training.

7.3 Study Design

The study was run as a between-subjects design, where participants were randomly assigned to either the MI-adapted or the GPT-4 condition. At the beginning of the study, participants were presented with a description of three different personas, each specifying a different target behaviour within the context of general lifestyle changes and prompted to choose the persona they identified with most strongly. The employment of fictional scenarios, instead of participants' personal experiences, serves as a safeguard in case of harmful outputs. I selected the three target behaviours *procrastinate less*, *live more sustainably*, and *eat a healthier diet*, as they represent a sample of non-medical lifestyle changes that a large part of

³an information science student, working on a related class project at the time.

⁴for an overview of how to interpret κ scores, see Table 3.2 in Section 3.2.1.1.

Behaviour	Description
Giving Facts	▪ User statement is actively reinforced ("It is actually true that...", "It is definitely true that...") ▪ General information ▪ Statements about how behavioural change works or what is human, formulated as a fact, also if worded as subjunctive ("can", "could" also without direct reference to the user, i.e. "For example, a strategy could be...") ▪ Statements about the user (e.g. "your behaviour shows that...") are <i>not</i> to be counted as giving facts. ▪ Statements like: "It's also important to remember that behaviour change takes time and it's perfectly normal to not be perfect straight away. So don't be too hard on yourself when things don't go so well." are <i>Advise</i>, <i>not Giving Facts</i>!
Advise	▪ Users are given suggestions for actions ▪ Users are given suggestions on how to think about changing their behaviour ("Don't forget that...", "Don't be too hard on yourself...", etc.)
Share Personal Information	▪ System presents itself as human ("We all have.... sometimes", "I understand...", "I like..."). Only if "I", "we" or "me" is explicitly used. ▪ <i>Not</i> if the system expresses understanding for the user (i.e. "It is understandable...") or clarifies their own role ("I am here to...", "I am trying to help you", "Let's work together...").
Feeling	▪ System shows emotions: "I'm glad that...", "It's fantastic to hear...", "I wish you..." (<i>not</i> "It's fantastic that...")

Table 7.4: Finalised annotation scheme for unsuitable system outputs.

the population struggle with or aim for (Farhud, 2015; Ferrari, O'Callaghan, & Newbegin, 2005; Office for National Statistics, 2023; Pinho et al., 2018). I expected that this would increase the chances that participants can relate to them while keeping potential harms in the event of unsubstantiated advice low compared to more severe or medical lifestyle changes. The following persona descriptions were provided to study participants:

- **Procrastinate less:** Leo has been struggling with his productivity at work for some time now. He regularly resolves to work with more focus and be less distracted by social media, mobile phones and fellow students/colleagues, but he can't quite manage it. Even when he is alone, his mind often wanders. He increasingly has the feeling that his lack of focus is having a negative impact on his work. He really wants to work more productively and efficiently but doesn't know how.
- **Live more sustainably:** Peter is an environmentally conscious person who always tries to behave as sustainably as possible. He already avoids flying as far as possible and tries to keep his electricity consumption low. However, he finds it particularly difficult to do without a car in his everyday life. Although local transport in his city is quite well developed, he is still usually quicker by car, especially as he has his own parking space at work and can park almost directly in front of the office door, while the bus stop is further away. However, as he knows how damaging driving is for

the environment, he is now trying to find the motivation to switch to other means of transport more often.

- **Improved nutrition:** Laura has let her diet slip recently. Barbecues, beer gardens and summer parties have made it difficult for her to eat healthily. What's more, she's had a lot to do at work and therefore hasn't had much time to cook. Now Laura is on vacation and has decided to use this time to make her diet more vegetable-heavy again and take in more vitamins. It is important for Laura not to go on a crash diet but to make her diet healthier in the long term. However, she already knows that this won't be easy.

After filling out a pre-test questionnaire asking about their level of identification with and current pursuit of the target behaviour, as well as their readiness to change, participants interacted with the chatbot about the chosen fictional target behaviour for a total of 12 turns, which a pilot study revealed to be a suitable length⁵. The users rated each GPT-4 generated output as good/helpful, bad/inappropriate, or offensive/harmful. After each rating, participants were asked to provide a reason for their judgement on a voluntary basis. Following the conversation, participants filled out a post-task questionnaire. I go into detail about the measures collected pre- and post-interaction in Section 7.3.1.

7.3.1 Measures

The main focus of this study was on measuring the impact of LLM-user conversations on their motivation to change and the quality of conversations as measured through user perceptions of these interactions and individual turns. Furthermore, I was interested to what extent conversation quality and impact are dependent on the *behaviour of the users*. To gain quantitative insights on user perceptions and the effects of the interactions on motivation, I based most of my selection of collected measures on the study design presented by He et al. (2022) in their comparison of MI-based and non-MI smoking cessation chatbots. I take the following measures as dependent variables:

- **Perception of Motivational Interviewing** – I measured perception of MI using the Client Evaluation of Motivational Interviewing Scale (CEMI) (Madson, Mohn, Schumacher, & Landry, 2015; Madson et al., 2013; Madson, Villarosa, Schumacher, & Mohn, 2016). The measure is intended to be completed by clients directly after the intervention and does not rely on expert judgements to evaluate MI adherence in provider-client contexts.
- **Therapeutic Alliance** – I measured therapeutic alliance using the German version of the Working Alliance Inventory - short revised (WAI-SR) (Wilmers et al., 2008). This served as a proxy to determine the relationship between the user and the chatbot in the broader context of provider-client relationships. Since the chatbot was intended to support users in behavioural changes drawing from MI principles, this is a suited measure for the kind of relationship I wanted to achieve.
- **Δ Readiness to Change** – Readiness to change was measured using the contemplation ladder. The contemplation ladder was originally designed to measure readiness to quit smoking (Biener & Abrams, 1991) but has since been adapted to various target behaviours (Coolidge et al., 2011b;

⁵In the pilot study, each user was asked whether the conversation was too long, too short, or just right. The majority of users indicated "just right". The rest was divided equally between "too long" and "too short".

Jede Sprosse auf dieser Leiter steht für den Stand der Überlegungen verschiedener Personen zu dem von Ihnen gewählten Zielverhalten. Geben Sie die Zahl an, auf der Sie sich gerade befinden

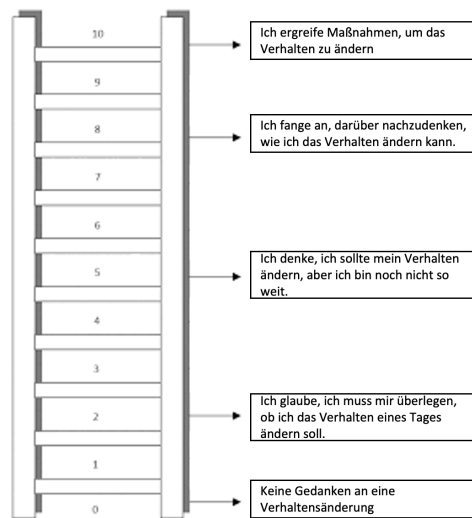


Figure 7.3: Adapted and translated Contemplation Ladder, based on Coolidge et al. (2011a).

John, Rendina, Starks, Grov, & Parsons, 2019). For the purpose of this study, I applied the ladder to behaviour change generally (see Figure 7.3), prompting participants to focus on their chosen target behaviour. Since the study employed fictional scenarios, which might lead a share of the participants to not relate to the chosen target behaviour, for this analysis, I only included participants who indicated that they strongly identify with the chosen target behaviour (≥ 7 measured on an 11-point Likert scale).

- Undesirable Output - as rated by study participants and annotated based on the annotation scheme introduced in Section 7.2.

In addition to the dependent measures, I also collected additional information about the degree to which participants identified with and were currently pursuing their selected target behaviour (indicated on 11-point Likert scales) before the start of the conversation. As post-test measures, I included perceived communication competence and perceived empathy, as outlined by He et al. (2022). Also, following He et al. (2022), I measured engagement using the User Engagement Scale–Short Form (UES-SF) without the aesthetic appeal subscale (Holdener, Gut, & Angerer, 2020; O'Brien, Cairns, & Hall, 2018).

7.3.2 Hypotheses and Statistical Analysis

Since I expected the MI-adapted condition to be more in line with MI principles, I defined the following hypotheses relating to RQ1:

- H1: (addresses RQ 3.1a): Using the MI-adapted condition increases the difference between readiness to change post- and pre-conversation compared to using non-adapted GPT-4
- H2: (addresses RQ 3.1a): Using the MI-adapted condition reduces the share of harmful and undesirable outputs compared to using non-adapted GPT-4

- H3: (addresses RQ 3.1b): Using the MI-adapted condition increases user perception of MI compared to using non-adapted GPT-4
- H4: (addresses RQ 3.1b): Using the MI-adapted condition increases therapeutic alliance compared to using non-adapted GPT-4

Each hypothesis was tested with a one-tailed, independent t-test. The t-tests took the condition as independent variable and 1) therapeutic alliance, 2) perception of MI, 3) the difference between readiness to change post and pre-test, and 4) the mean number of harmful bot utterances per conversation, respectively, as dependent variables.

Since there are so many variables in play with respect to RQ2, and I assumed many relationships between variables, I investigated these relationships using exploratory and qualitative analysis rather than pre-forming hypotheses before data collection.

7.3.3 Participants

I recruited 159 participants through Prolific⁶. Since the study was conducted in German, participation was limited to German native speakers currently residing in Germany. The mean completion time in a pilot study was 20 minutes, leading me to compensate participants with €4.00, corresponding to an hourly wage in line with the German minimum wage (€12.00). An attention check was included to ensure the validity of submissions. Participants who did not pass the attention check were not counted towards the sample size but did receive the full pay.

Target Behaviour	MI-adapted	GPT-4
healthier eating	26	21
sustainable living	2	10
less procrastination	51	46
sum	79	78

Table 7.5: Number of chats per target behaviour and condition.

79 Participants were assigned to the MI-adapted condition and 78 to the GPT-4 condition. Participants were between 19 and 72 years old (Mean=32.27, SD=9.4), 60% (94) identified as male, and the rest (65) as female. The majority of participants were highly educated, with 88 participants having finished college degrees at different levels (52 had finished a bachelor's degree, 33 had a master's, and three had finished a doctoral degree). Another 52 had passed the German university entry-level exam as their highest form of education. The remaining 20 had earned either an intermediary high school degree or finished a formal apprenticeship. 43 (27%) were students, whereas the remainder held a variety of occupations, including IT specialist, musician, teacher, social worker, researcher, and engineer, to name a few. Participants generally had a high level of identification with their chosen target behaviour (only

⁶Pilot conversations led me to expect a medium effect size of $d=0.5$, and the desired power was set to $\beta = 0.8$. The significance threshold was Bonferroni-corrected, resulting in $\alpha = \frac{0.05}{4} = 0.0125$. I used G*Power to calculate the appropriate sample size. Based on the results, I aimed at a sample size of at least 156 participants (78 per condition).

25% indicated a level of identification of 7 or less), and mean readiness to change at the beginning of the interaction was 6.25 (SD:2.27, min:0, max:10), indicating a high degree of relevance of the offered target behaviours. The majority of participants (62.18%) chose less procrastination as a target behaviour to discuss, followed by healthier nutrition (30.13%). Only 7.69% chose sustainability as their goal.

7.4 Results

In this section, I present my results, beginning with a descriptive analysis of the nature of the collected chat conversations, especially with respect to the type of turns generated in each of the study conditions and the relationships between the collected measures (Section 7.4.1). I then outline the results from my main analysis based on the hypotheses introduced in Section 7.3.2 (Section 7.4.2). Finally, I present exploratory and qualitative analysis related to RQ3.2 (Section 7.4.3), providing insights into user behaviour and expectations when interacting with LLM-based systems about behaviour change.

7.4.1 Descriptive Analysis

Since parts of the conversational framework rely on GPT-4 generated texts, the chats collected in the study were highly varied, not only with respect to user reactions but also regarding system behaviour. Below, I report on the type of outputs delivered by both systems, showing structural differences in system behaviour between conditions, followed by an analysis of perception metrics collected in the pre- and post-tests.

7.4.1.1 Chats

On average, the MI-adapted condition contained fewer words per GPT-4 generated turn (35.67, SD=23.34) than the GPT-4 condition (47.50, SD=25.16). This can be explained by the addition to the system prompt in the MI-adapted condition (see Figure 7.2), which instructs the model to keep the generated answer as short as possible. This addition was guided by MI principles, which put emphasis on the client talking more than the counsellor, with the counsellor predominantly reflecting and asking questions (W. R. Miller & Rollnick, 2002).

The differing lengths of system turns in the study conditions seemed to have no impact on the length of user turns, which had the same mean word count in both conditions (12, SD MI-adapted = 10, SD GPT-4 = 13). In both conditions, the first two user turns that allowed free-text input tended to be the longest, with average word counts ranging between 14.8 and 17.9.

To gain insight into how the systems behaved differently throughout the conversations, I applied the EPITOME classifiers provided by Sharma et al. (2020) and the Empathetic Intent classifiers introduced by Welivita and Pu (2020) to the system turns produced in both conditions⁷.

⁷I trained the classifiers myself using code and data provided by the respective authors. For classification, I first translated the texts into English using the same machine translation model as in Chapter 6 (Tiedemann & Thottingal, 2020).

EPITOME

Sharma et al. (2020) developed the EPITOME framework to measure empathy across three dimensions:

- Interpretations: showing an understanding of the experiences and feelings expressed by the user. In stronger forms, this can include the sharing of own, similar experiences to showcase understanding (i.e. *I also experience...*).
- Emotional Reactions: expressing experienced emotions, including compassion and warmth.
- Explorations: Attempts to deepen understanding by exploring feelings and experiences which are not present in the preceding text.

The classifier assigns a label between zero (not present) and two (strong presence) across each of the dimensions to a given system turn, taking into account the preceding user utterance.

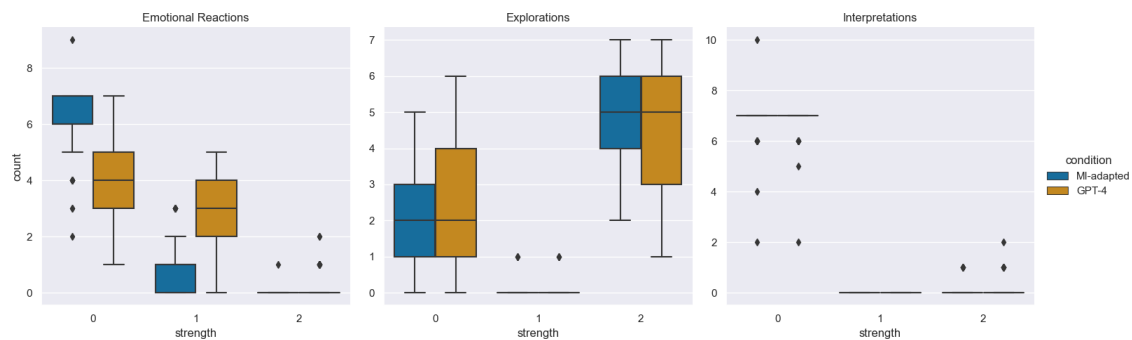


Figure 7.4: Boxplots of presence and strength of empathy dimensions in chats per condition.

The boxplots in Figure 7.4 show the number of turns classified with each strength level per EPITOME dimension per chat. In both conditions, interpretations were barely present. Upon reading through the conversations, two main reasons for this became apparent: 1) since the conversations were about non-medical lifestyle changes and do not focus on mental health, most participants did not voice strong emotions throughout the chats, and 2) in both conditions, the CA usually did not talk about own experiences or behaviours⁸. Conversely, both conditions had high amounts of strong explorations, which can include both open questions and reflections, this being more steady in the MI-adapted condition. The most apparent difference between conditions was found with respect to emotional reactions: whereas GPT-4 had a mean of three turns per chat displaying weak emotional reactions, these were almost completely absent in the MI-adapted condition, indicating that the condition was successful in mitigating harms related to anthropomorphism and increasing adherence to MI principles.

⁸The systems did output utterances that would be annotated as “Share Personal Information” as they employed wordings such as “*We all have these thoughts sometimes*”. These instances might not have been caught by the EPITOME classifier, as they do not tell of a separate experience but present a form of empathising. While such utterances might be appropriate in the context of peer counselling, they have the potential of increasing anthropomorphism, leading me to categorise them as inappropriate in the context of human-LLM conversations.

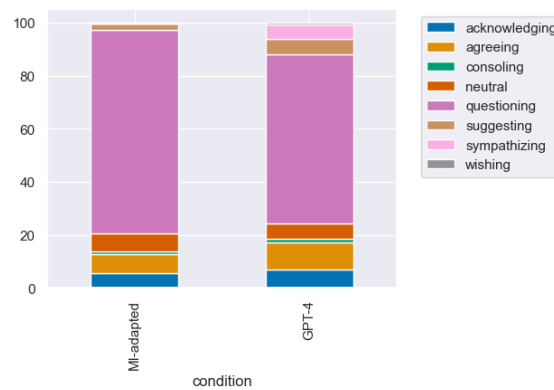


Figure 7.5: Overall percentage of empathetic intents per condition.

Empathetic Intents

Welivita and Pu (2020) define eight empathetic intents: acknowledging, agreeing, consoling, questioning, suggesting, sympathising, wishing, and encouraging, as well as a non-empathetic neutral intent. Figure 7.5 shows the share of empathetic intents voiced by the system per condition. Questioning is the most common intent found in both conditions' GPT-4 generated turns, in line with the high amount of explorations found by the EPITOME classifier. It becomes apparent, however, that this intent was less common overall in the GPT-4 condition, which in turn included a higher share of acknowledging, agreeing, sympathising and suggesting than the MI-adapted condition. While this generally may result in a more varied display of empathy in the GPT-4 condition, it is less in line with MI principles, which focus on encouraging self-reflection instead of suggesting to clients what to do and favour professional distance over the expression of emotional empathy in the form of sympathising.

7.4.1.2 User Perceptions

More than half the conversations (52.23%) led to an increase in readiness to change, and in 33.12% of cases, readiness to change remained the same. Surprisingly, 14.01% of conversations actually led to a decrease in readiness to change. As can be seen in Figure 7.6, therapeutic alliance, user engagement, perception of MI, perceived empathy and communication competence are all significantly correlated with each other (0.53-0.73). Interestingly, there is no significant correlation between any of these measures and the Δ readiness to change ($r = 0.15$ -0.25). This indicates that the way participants perceive the quality of the conversation does not actually seem to mirror the effectiveness of the intervention. The only measure that Δ readiness to change is significantly correlated with is the readiness to change before the chat interaction (-0.51). The negative correlation indicates that higher readiness to change at the beginning corresponded to a smaller impact of the conversation. Finally, there are significant correlations between the readiness to change pre-test and the current pursuit of the target behaviour (0.29) and between the current pursuit of the target behaviour and the level of identification with the target behaviour (0.33), but not between the level of identification and the readiness to change at the beginning of the study ($r=0.18$). This is to be expected since a person can identify with wanting to change something, even when they are not currently actively pursuing the behaviour change (i.e. they are in the contemplation stage).

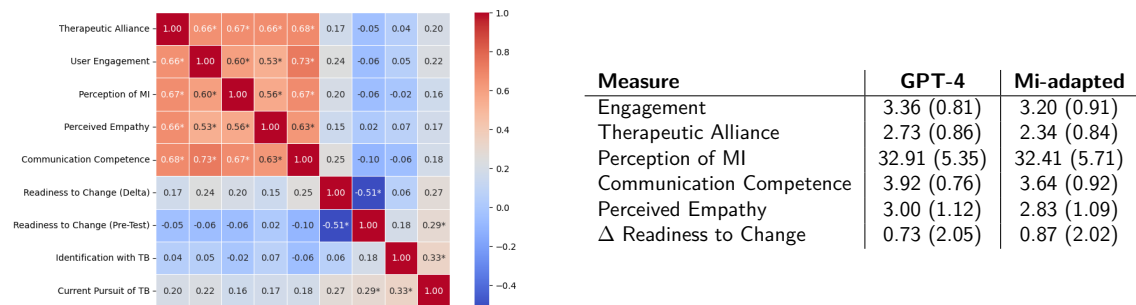


Figure 7.6: Correlation analysis of pre- and post-test measures and means (standard deviations) per condition. Cells marked with * are significantly correlated at $p < 0.0006$ (Bonferroni-corrected α).

7.4.2 Main Analyses

Hypotheses 1, 3 and 4 could be tested by evaluating questionnaire answers given by participants, whereas hypothesis 2 necessitated a deeper analysis of the chatbot outputs. For all hypothesis tests, an independent, one-tailed t-test was conducted if the assumption of normal distribution was met. If this was not the case, I conducted Mann-Whitney-U tests.

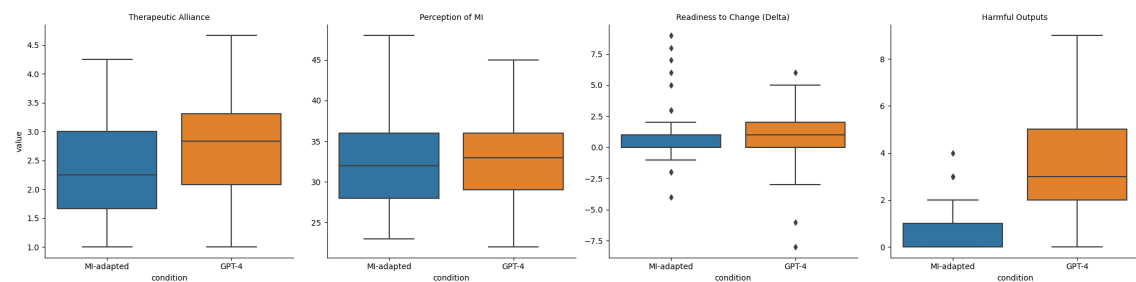


Figure 7.7: Boxplots for results of dependent measures per condition.

7.4.2.1 MI-Adaptation Does not Increase the Difference Between Readiness to Change Post- and Pre-Conversation

Here, I explore between group effects to determine whether there were differences between the Δ readiness to change between conditions, as well as repeated measure effects to identify whether conversations with the bots had an overall impact on participants' readiness to change.

Between group effects:

To test H1, all participants were filtered by their level of identification with the chosen target behaviour in order to exclude participants who did not relate to any of the topics of choice and had no intention to change. I only included participants who indicated a level of identification with the goal of 7 or higher

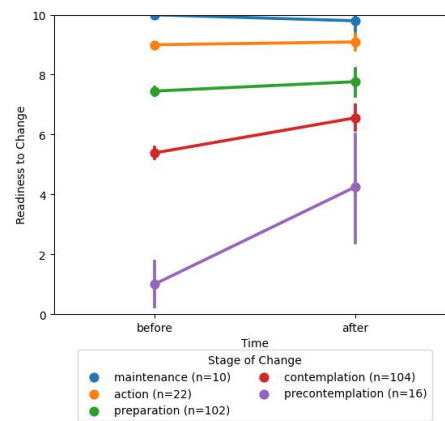


Figure 7.8: Changes in readiness to change depending on stage of change before conversation start.

(75%) in this significance test. This resulted in the inclusion of 64 participants for the GPT-4 and 63 for the MI-adapted condition. Given this selection, the conversations' effect on readiness to change was not significantly higher for the MI-adapted condition than for the GPT-4 condition ($U=1646.0$, $p=0.967$), and there was no statistical difference between conditions at Bonferroni-corrected $\alpha = 0.0125$.

Repeated measures effects:

Although there was no difference in increase in readiness to change between conditions, both conditions resulted in a significant increase in readiness to change after the conversations compared to before, as Wilcoxon Signed-Rank tests revealed (overall: $W=1078.0$, $p < 0.0001$, MI-adapted: $W=267.0$, $p < 0.0001$, GPT-4: $W=286.5$, $p < 0.0001$). This effect seemed to be even more pronounced if participants indicated being in an early stage of change pre-conversation. As shown in Figure 7.8, participants who were in the action and maintenance phases at the start of the conversation benefited the least, with the readiness of participants in the maintenance phase even decreasing. This makes sense since people in the action phase of change are likely highly motivated and do not need further motivational boosts, instead benefiting from information and advice on how to initiate and maintain change.

7.4.2.2 MI-Adaptation Reduces the Share of Harmful and Undesirable Outputs

I used both user ratings of turns and researcher-determined annotations to determine the share of undesirable outputs:

User ratings:

I aggregated the ratings given by study participants throughout the conversations. Overall, 649 GPT-4 generated turns were rated as good/helpful (MI-adapted: 424, GPT-4: 431), and 174 as bad/inappropriate (MI-adapted: 123, GPT-4: 107). I provide the average counts of each rating per chat condition and chat in Table 7.6. It is noted that not a single turn was rated as offensive/hurtful. A χ^2 test revealed no significant differences in rating distributions for the two study conditions ($\chi^2(1, N = 813) = 0.27, p = 0.603$).

Condition	good/helpful	bad/inappropriate
Mi-adapted	5.44 (1.51)	1.58 (1.59)
GPT-4	5.60 (1.42)	1.39 (1.42)

Table 7.6: Average (standard deviation) of positive and negative user ratings of bot turns per chat.

Researcher-determined annotations:

The annotations described in Section 7.2 were aggregated on a chat basis. Turns, which included different types of undesirable behaviour, led to an increase of one per behaviour type. For instance, a turn which included both Advise and Feeling added two to the total count of undesirable content in the chat. Overall, chats in the MI-adapted condition contained an average of 0.41 pieces of undesirable content per chat, whereas the GPT-4 condition contained an average of 3.77 undesirable pieces of content per chat. The difference between conditions was significant at $p < 0.0001$ ($U=391.5$).

7.4.2.3 MI-Adaptation Does Not Increase User Perception of MI

The hypothesis that perception of MI was higher in the MI-adapted condition was not supported ($T(155) = -0.57$, $p = 0.716$). However, looking at only the scale items describing MI-inadherent behaviour (MIcha “Made you talk about something you didn’t want to discuss”; “Argued with you to change your behaviour”; and “Told you what to do”) yielded significantly lower values for the MI-adapted condition ($U = 2264$, $p = 0.0018$). This indicates that, while the MI-adapted condition was not more effective in helping participants talk about behaviour change than the GPT-4 condition, it did exhibit significantly fewer MI-Inadherent behaviours than the GPT-4 condition.

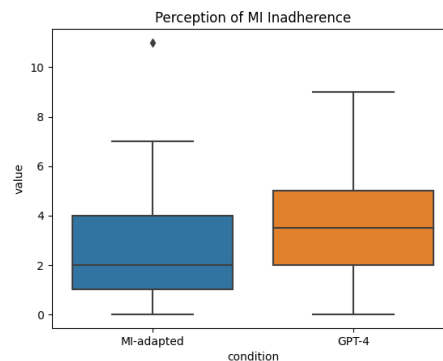


Figure 7.9: Boxplot of participants' perceptions of MI-inadherent behaviour per condition.

7.4.2.4 MI-Adaptation Does Not Increase Therapeutic Alliance

Again, the hypothesis that the MI-adapted condition is more successful at establishing therapeutic alliance than the GPT-4 condition was not supported ($U=2302.0$, $p=0.997$). In fact, questionnaire

results of the GPT-4 condition were significantly greater for this measure than those of the MI-adapted condition ($U=3860.0$, $p = 0.003$). Considering the short duration of the conversations and the approach of MI to first elicit change talk and motivation before giving advice, this result is not as surprising as it seems at first glance. The questions posed in the Working Alliance Inventory are primarily focused on goal orientation (“I believe that the way we are working on my problem is right”, “MIcha and I work towards goals that we agree on.”) and interpersonal aspects of counsellor-client relationships (“MIcha and I respect each other.”, “I feel that MIcha appreciates me”, “I feel that MIcha stands by me even when I do something he doesn’t approve of”). These results indicate that the MI-adapted condition was considerably less goal-oriented than the GPT-4 condition, which is to be expected considering the system was prompted to actively avoid giving any advice in this condition. It also indicates that the MI-adapted condition was considerably better at suppressing anthropomorphism and the relational feelings that arise with it than the GPT-4 condition, corroborating the analysis of harmful outputs.

7.4.3 Qualitative and Exploratory Analysis of User Behaviour

To address RQ2, I relied on qualitative insights, allowing me to identify trends in behaviour and interaction outcomes. I identified three main behavioural dimensions related to the degree of pre-existing knowledge about the target behaviour, the readiness to reflect, and cooperation with the system. The statistical analysis revealed that these dimensions significantly impacted the outcome of the interactions in various ways. Furthermore, I identified varying expectations and levels of acceptance by users regarding the different system behaviours related to MI principles, which I explain in detail in this section.

7.4.3.1 Behavioural Dimensions

Qualitative analysis of the chats revealed marked differences in the approaches and attitudes of participants during the conversations with the system. I followed an inductive methodology to identify different user behaviour types, repeatedly reading all collected chats to identify patterns. Three dimensions of user behaviour emerged organically from the data and were iteratively refined:

- **Cooperation:** Participants varied in their level of cooperation. Uncooperative users refused to provide information on request, were obviously uninterested in change, were openly critical (“This question is stupid”), or wrote random or unrelated responses. They tended to provide consistently short answers, or they were quick to end the conversation (rather than interacting with the CA) and put no effort into understanding CA utterances (i.e. if the last word of a CA utterance is cut off⁹, they requested the bot to finish the sentence, even though it was very clear what the end of the message would have been). Cooperative users, in contrast, mostly replied in full sentences. If the conversation reached an impasse, these users took on an active role in driving it forward.
- **Reflectiveness:** Reflective users were open to reflection throughout the conversation, finding their own solutions, and answering questions, even if they were hard at first or took some effort to reply to. If similar questions were asked by the bot in close proximity, reflective users tended to provide different answers/more information instead of repeating what they had already said in

⁹this was sometimes the case due to the pre-defined max.token limit.

previous turns.

- **Preinformedness/Openness to finding own solutions:** These users were open to suggesting solutions for their problems throughout the conversation. They were not advice-oriented, instead coming up with various solutions and did not ask the CA for advice when prompted to think about strategies. Solutions offered by them were concrete and realistic, instead of just the opposite of their current behaviour, showing the user had put thoughts into how a solution looks for them concretely.

I annotated all chats according to these dimensions (in a binary fashion - either present or absent) and wrote annotation guidelines for each dimension. Some edge cases, which could not confidently be assigned were labelled as 0.5. To ensure the validity of annotations, a second annotator who holds a Master's degree in information science was given the annotation guidelines and annotated 10% of collected chats in the same way. Since the first round of annotations resulted in high IRR for cooperation but unsatisfactory IRR for the remaining two categories¹⁰, I followed the same process as outlined in Section 7.2 and discussed the other two dimensions before the second annotator annotated reflectiveness and preinformedness for another 10% of chats. This led to acceptable IRR across all three dimensions, as reported in Table 7.7.

Dimension	Cohen's κ	
	Round 1	Round 2
Reflectiveness	0.51	0.64
Cooperation	0.85	–
Preinformedness	0.52	0.88

Table 7.7: Inter-rater reliability of annotations of user behaviours after one round of training.

Once satisfactory IRR was established, all instances labelled as 0.5 in the first round of annotation were resolved in discussions between the second annotator and me. Table 7.8 gives an overview of the distribution of behaviours across conditions. I provide representative example chats for each behavioural dimension in Tables 7.11 to 7.14 at the end of the chapter.

Condition	Reflective	Cooperative	Preinformed
MI-adapted	56.96%	77.21%	44.3%
GPT-4	50.65%	71.43%	44.16%

Table 7.8: Share of reflective, cooperative, and preinformed participants per study condition.

¹⁰for an overview of how to interpret κ scores, see Table 3.2 in Section 3.2.1.1.

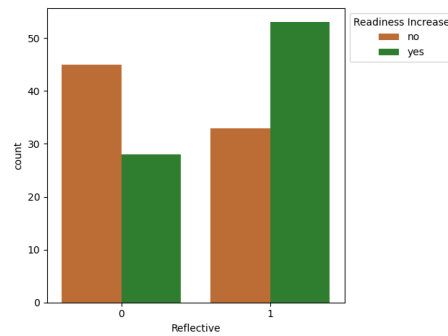


Figure 7.10: Bar chart showing the number of reflective and unreflective participants and the change in their readiness to change after interacting with the CA.

7.4.3.2 Effects of User Behaviour on Interaction Outcomes

Based on these annotations, I ran various comparative analyses to understand how user behaviour influences the outcome of conversations and what determinants are predictive of specific user behaviours. Since I was interested only in the effect of user behaviour at this point in the analysis, I ran χ^2 tests to rule out the condition and the chosen target behaviour as confounding variables. None of the tests was significant, indicating that user cooperation, preinformedness and reflectiveness were not different across conditions or target behaviours. Considering that the MI-adapted condition posed more questions to the user, this indicates that someone who is not prepared to reflect on their behaviour can not be nudged to do so merely by LLM-generated MI-style questions. Following the ruling out of condition and target behaviour as confounding variables, I ran all user behaviour analyses without differentiating between study conditions.

Reflective Participants Were More Likely to See an Increase in Readiness to Change

To identify whether user behaviour had an impact on readiness to change after the conversation, I ran t-tests with the Δ readiness to change as the dependent variable and cooperation, reflectiveness, and preinformedness, all dichotomous, as independent variables, respectively. I found significant effects for reflectiveness ($U=3623.5$, $p = 0.0088$) but not cooperation and preinformedness, indicating that reflective users saw higher increases in readiness to change. They were also more likely to see an increase in readiness to change as compared to decreases or no changes, as indicated in Figure 7.10.

Preinformed and Cooperative Participants Were More Likely to Voice Change Talk When Asked About Their Next Steps

During the final phase of the conversation, the system asked participants about the next steps they are planning to take towards change. For this analysis, I leveraged the user utterance classifications that were also used to identify suitable client response types. I ran χ^2 tests to identify whether reflective, cooperative and preinformed users voiced more positive viewpoints about change when asked for this next step. I found significant differences at $\alpha < 0.0125$ of utterance valences for the behavioural dimensions of preinformedness and cooperation. Preinformed ($\chi^2(2, N = 159) = 23.63$, $p < 0.001$),

and cooperative ($\chi^2(2, N = 159) = 11.09, p < 0.004$) users were more likely to voice change talk and less likely to respond with a follow/neutral utterance when asked about their next steps. Such effects did not exist for the study condition or chosen target behaviour, again indicating that user behaviour is a strong determining factor for the success of such conversations.

Participants' Behaviours had an Impact on User Engagement and Perceptions of System Capabilities

While there was no significant difference between the two study conditions regarding user engagement, perceived communication competence, and perceived system empathy, I did find significant differences in how users who displayed distinct behavioural dimensions judged these variables. Preinformed, reflective and cooperative participants all displayed higher levels of user engagement ($U=364, p=0.008$; $U=4115.5, p < 0.001$; $U=3244.5, p < 0.001$, respectively). Reflective and cooperative users also judged the system's communication competence higher ($U=3690.5, p<0.009$; $U=2967, p=0.004$, respectively), and reflective users indicated higher perceived empathy ($U=3739.5, p=0.005$) indicating that the quality of the generated text was highly dependent on the amount of information the user was ready to reveal.

Participants Who Identified with Their Chosen Target Behaviour More Strongly and Were Currently Pursuing This Behaviour Were More Likely to Be Reflective and Cooperative

Finally, I was interested in identifying the determinants that predict user behaviour across behavioural dimensions. This could potentially help to identify users who are more likely to benefit from behaviour change interactions with LLM-based CAs in the future. I found no effects for behavioural dimensions and the readiness to change indicated at the beginning of the conversation. However, I did find that users who had indicated higher degrees of current pursuit of their chosen target behaviour were more likely to display reflective ($U=3905, p<0.001$), cooperative ($U=2967, p=0.0017$) and preinformed behaviour ($U=3749.5, p=0.0022$). Furthermore, participants who identified strongly with the target behaviour were more likely to display cooperative ($U=2979, p=0.0013$) and reflective ($U=3886, p<0.001$) behaviour.

7.4.3.3 Qualitative Insights Into User Expectations

In addition to the chats themselves, I also analysed the reasons participants gave to explain their ratings of CA turns. Since I was predominantly interested in the expectations of users, as opposed to how well these were fulfilled by the different conditions, I did not distinguish between study conditions for this analysis. Instead, I focussed my attention on how different aspects of MI-oriented conversations were received in interactions with LLM-based chatbots and how participants judged the general capabilities of the chatbots.

This analysis revealed that the behavioural dimensions of reflectiveness and preinformedness were also reflected in the explanations users gave for their turn ratings. This impacted, for example, whether users appreciated advice (given predominantly in the GPT-4 condition) and to what degree they welcomed questions that made them think about their behaviour and strategies to change it.

Open Questions:

Open questions were often regarded favourably for their ability to foster self-reflection. As can be seen in these examples of explanations for positive ratings:

The answer is good because it makes me think about concrete next steps to adapt my behaviour. (chat 58, turn 11)

The answer was great again and stimulates initial thinking about a strategy. (chat 76, turn 11)

The answer is solution-oriented because it asks what is preventing me from tackling my problem. (chat 75, turn 5)

The fostering of self-reflection was the most commonly stated reason for a positive turn rating, indicating that many users exhibited a positive stance towards reflection.

This contrasts with examples of negative rating explanations concerning the lack of advice given by the system, which indicate a less reflective, advice-oriented mindset:

Still no specific means of achieving the goal is mentioned. (chat 54, turn 9)

I would like a quick instruction on what to do or what can be done...it is taking too long. (chat 69, turn 8)

The examples above showcase participants who are *not willing* to reflect and come up with their own solutions, indicating low reflectiveness. However, there were also participants voicing a perceived *inability* to come up with their own solutions, which could be more indicative of a lack of functional knowledge about the chosen behaviour change (captured by low preinformedness):

I think it's a bad idea to ask someone looking for advice for an idea on how to deal with their problem. My problem is that I don't have an idea! (chat 52, turn 5)

No idea what my next step will be? If I knew that, I wouldn't have the problem of procrastination at all. (chat 92, turn 10)

Some explanations represented a mix of both lack of reflectiveness and preinformedness, such as the example below:

*When I come to a chatbot with a problem, I think that the bot should offer solutions, **or at least ask the question more precisely and in a better structured way to make it easier for me to answer.** If I already knew how to solve my problem, I wouldn't ask for help.* (chat 65, turn 5)

While the general theme of this explanation revolves around the participant feeling unable to come up with solutions, the bolded text also indicates that the participant was not willing to think hard to come up with an answer and reflect on their behaviour and possible strategies to change it. Such statements indicate a lack of recognition of the usefulness of written, journaling-style reflection. As can also be seen by the example below, some participants seemed to experience questions as the bot wanting information from them instead of as a chance to think about their problems:

I think you are looking for help and are not here to tell your life story. (chat 135, turn 7)

There were also cases in which turns were rated negatively because they were perceived as repetitive or redundant. This predominantly happened in the context of similar questions being asked within a conversation:

I find it redundant. I've already told the story a bit before that and would only be repeating things here. (chat 10, turn 8)

Similar question as before, I feel like we're going round in circles. (chat 86, turn 9)

The question feels duplicate, reasons why you no longer want to procrastinate have already been given before. (chat 135, turn 8)

Reflections:

Similar differences in how participants perceived CA turns are exemplified by differing levels of appreciation of outputs, which reflected user turns back to them. Reflections are an integral part of MI, which mainly relies on reflecting client utterances and asking open questions. Reflections voiced by the system were often mentioned in explanations of positive ratings:

The text is helpful because it gives me a clear and structured overview of the conversation so far. (chat 117, turn 10)

A good summary, without any new information, but very relevant to the conversation. (chat 126, turn 8)

Similar to open questions, if user utterances were reflected back to participants, this was perceived as redundant at times:

Chatbot repeats what I have already said and gives me no input at all. (chat 63, turn 6)

Just a repetition of what I have already written. I can think about the application to work myself, so the answer is not enriching. (chat 25, turn 5)

Again, the last example showcases a participant who does not seem to be interested in employing self-reflection in a conversation like this, but expects instructions instead, noting that they can think for themselves, without guidance from the system.

Affirmations:

Affirmations were generally judged positively as increasing self-worth and motivating participants, indicating that praise, even when clearly coming from a non-human conversational partner, can have beneficial effects on people's self-assessments:

The text is very helpful because it expresses recognition and admiration for my efforts, which can boost my self-confidence and self-efficacy. It encourages me to keep working on my goals. It praises my strategies and confirms that I have already achieved a lot. It is therefore positive and supportive feedback that can help me. (chat 117, turn 8)

An affirming and encouraging answer that makes me feel better. (chat 86, turn 8)

The answer has recognised that I am already doing some things, but not yet everything I want to do and possibly have a guilty conscience about it. That's why it's fitting that I'm told not to be too hard on myself, even if it wouldn't change my current behaviour. (chat 93, turn 5)

However, bot turns, including affirmations and reflections, were at times perceived as generic, underscoring the variable quality of GPT-generated text:

The summary sounds very clumsy and not really thought through. It feels like the bot always gives the same answer here, regardless of the content of the conversation. (chat 61, turn 10)

I'm starting to get more and more bothered by the phrases interesting/notable/impressive, it sounds like a bad psychiatrist. (chat 192, turn 8)

The answer gives the impression of not responding to the text entered. It sounds more like a predetermined standard answer. (chat 8, turn 5)

Advice:

While the MI-adapted condition was specifically designed to avoid advice, instead focusing on reflections, open questions, and affirmations, the GPT-4 condition did provide pieces of advice, at times unsolicited. Again, this led to differing views, especially regarding the helpfulness of the advice given. Some study participants welcomed the information provided:

The AI suggests strategies that are very concise and little explained, but could potentially be helpful. (chat 42, turn 6)

but the approaches of the suggestions are very good and I could imagine actually using such tools. (chat 65, turn 8)

Others did not find benefits in the provided suggestions, indicating they had come to the conversation with previous knowledge which they had already implemented into their lives, with varying success:

The tip to dissect the tasks is well known, if the knowledge about dissecting would help me, I would have no more problems. In practice, however, I continue to procrastinate. (chat 57, turn 8)

Good tip. I practise Pomodoro, and that helps me a lot too... :) (chat 71, turn 8)

One participant also noted the inappropriateness of the advice given for the situation described in the previous conversational turns:

Taking cookery classes is an alternative that is far too time-consuming for someone who is more concerned with convenience and simple alternatives. (chat 152, turn 8)

Judgements of Technical Capabilities:

Lastly, I noticed that some participants seemed to base their ratings on the technical capabilities of the system rather than whether the CA could be beneficial in supporting behaviour change, as shown in

the examples below:

the bot has recognised that I want to correct my input. (chat 21, turn 5)

I am very positively surprised by this answer, it seems to understand me well. (chat 31, turn 8)

The bot responds to my answer and builds its answer on what it has learnt. (chat 150, turn 6)

I always assume that chatbots in studies are not real (or rule-based), but this chatbot responds really well to what I have written. (chat 66, turn 5)

good summary, is this really an AI? (chat 7, turn 10)

While such judgements of chatbot capabilities might also be relevant in real-life usage scenarios, they are likely to be less relevant when users proactively contact such a system with the intention of getting behaviour change support.

Overall, the rating explanations reinforce the impression that the helpfulness of interactions with the chatbot is highly reliant on the user's current situation and expectations. Users who are preinformed and willing to self-reflect are likely to benefit more from the MI-adapted version, whereas others might need a chatbot that is more focused on giving high-quality, personalised advice.

7.5 Exploratory Analysis of User Language

The classifiers introduced in Chapter 4 were used as a basis for the conversational design of the MI-adapted condition, but they are also useful for the analysis of the collected data and have the potential to provide a wide range of insights about the user and their current state. In this section, I explore users' language use in light of their readiness to change, preceding bot behaviours, and the behavioural dimensions exhibited by users. This section is aimed at shedding light on the following research question:

- **RQ 3.3:** What can we learn from applying classifiers based on the GLoHBCD to LLM-based conversations around behaviour change?

I begin by providing general information about the type of utterances voiced by users throughout the interactions, followed by focused experiments towards identifying differences in user language based on different user and system qualities.

7.5.1 User Language Over the Course of Interactions

On average, users voiced more change talk (3.68 SD=1.88 turns per chat) than sustain talk (1.14, SD=1.08). 4.15 (SD=1.71) turns per chat were classified as follow/neutral. The most common topic label was reason, with less than one turn per chat constituting commitment or taking steps utterances. The same applied to reason types, where general reason was the only label appearing more than once per chat, on average. I found no major differences in language use between the two study conditions. An overview of all client labels per condition is provided in Table 7.9.

Level	Label	MI-adapted	GPT-4
Valence	Change Talk	3.76 (1.98)	3.60 (1.78)
	Sustain Talk	1.25 (1.14)	1.03 (1.01)
	Follow/Neutral	4.00 (1.8)	4.31 (1.6)
Topic	Taking Steps	0.52 (0.8)	0.36 (0.67)
	Commitment	0.71 (0.70)	0.92 (0.94)
	Reason	3.78 (1.48)	3.34 (1.41)
Reason Type	General Reason	1.92 (1.29)	1.60 (1.36)
	ability	0.90 (0.88)	0.73 (0.84)
	need	0.26 (0.55)	0.34 (0.60)
	desire	0.76 (0.86)	0.68 (0.82)

Table 7.9: Mean count (standard deviation) of utterances per classified label and condition.

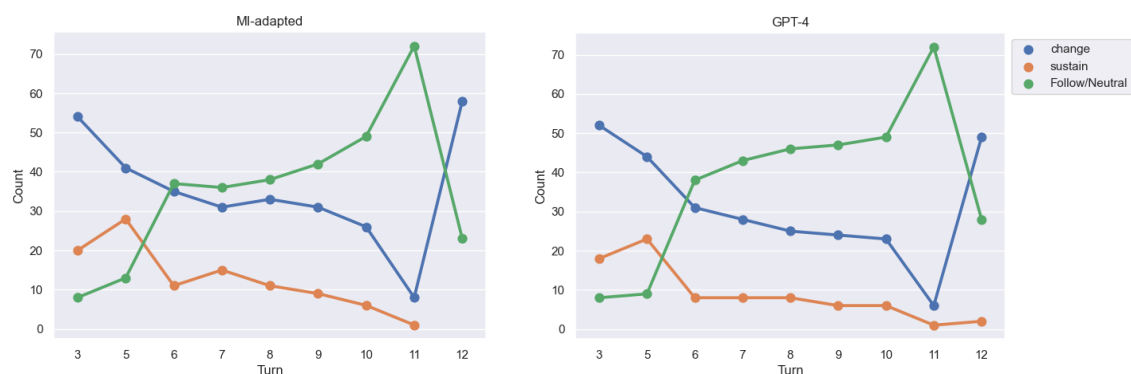


Figure 7.11: Progression of overall valence counts in interactions with MIcha by study condition.

A more detailed analysis of the progression of user language over the course of the interactions revealed some slight differences between the valences of user utterances between conditions at different points in the interactions. Figure 7.11 shows the progression of valences summed up over all interactions, starting at turn three (the first turn to allow free text user input). The differences between the two conditions become apparent when looking at the valences of the user turns between turns six and nine, where turn six was the first user reaction to a GPT-generated output and turn ten contained the CA's summary of the conversation. Within this window, the GPT-4 condition saw a slight but steady decline in change talk and sustain talk and an incline in follow/neutral instances. Compared to this, the change talk instances in the MI-adapted condition built a convex curve, which was mirrored by the concave curve of follow/neutral instances and sustain talk saw a spike at turn 7, followed by a sharp decline as the conversations progressed. A χ^2 test conducted to test the statistical significance of these observations revealed that the distribution of valences between turn six and nine do indeed differ significantly between conditions ($\chi^2(2, N = 641) = 6.3, p = 0.043$), with the GPT-4 condition leading to more follow/neutral and less change talk and sustain talk as expected, and the MI-adapted condition vice-versa.

A similar progression is visible in the topic classifications (see Figure 7.12), where the GPT-4 condition saw a steady decline in reason utterances, compared to the MI-adapted condition, where reason counts remained higher until turn 10. In both conditions, taking steps and commitment counts were low throughout the conversation, except for turn 12, in which the user was invited to share their next steps,

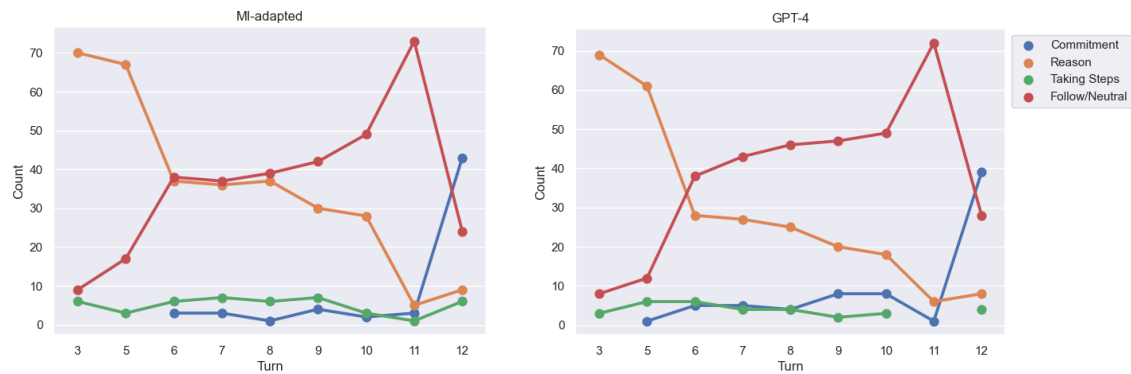


Figure 7.12: Progression of overall topic counts in interactions with Mlcha by study condition.

resulting in a spike in commitment utterances. Again, the difference between conditions was significant within the turn six to turn nine windows ($\chi^2(3, N = 641) = 13.32, p = 0.004$), with the GPT-4 condition displaying more follow/neutral and less reason and taking steps statements than expected. Surprisingly, the GPT-4 condition also elicited more commitment utterances than expected, while the opposite was true for the MI-adapted condition, which in turn elicited more taking steps and reason statements.

These differences can be explained by the tendency of the GPT-4 condition to provide advice, which made it easier for users to follow along with bot utterances, which could also take the form of commitment statements (e.g. “Good idea, I’ll try that.”). In contrast to this, the MI-adapted condition focussed more on reflections and open questions, which might have been less successful at eliciting commitment language, instead inviting users to provide their reasons for change and talking about what they had achieved so far. These analyses speak in favour of the inherent differences between conditions despite the lack of control I had over the specific content of bot outputs. It also shows that the MI-adapted condition seemed to be more successful in eliciting change-related language compared to the GPT-4 condition.

7.5.2 Impact of MI-Adherence and Inappropriate Bot Utterances on User Language

Past work has found links between counsellor MI-adherence and client language. The following experiments served the purpose of exploring whether high perceived MI-adherence or inadherence and a large number of inappropriate outputs impact user’s expressed stance towards change. To test this, I used the collected measures of MI-adherence and inadherence through the perception of MI questionnaire filled out by participants at the end of the study, as well as the harmfulness annotations. The count of potentially harmful behaviours per chat and questionnaire results regarding users’ perceptions of MI and MI-inadherence were all split along their respective median values, resulting in a “high” and “low” group for each measure. Subsequently, boxplots were used to identify apparent differences between groups and identify where significance tests were appropriate. The following significant differences were identified:

- Chats in which the CA was perceived as highly *MI-adherent* elicited **more change talk** ($U=3541$,

$p = 0.036$) and **less follow/neutral** utterances ($U=2415.5$, $p=0.012$).

- Chats in which the CA was perceived as highly *MI-inadherent* elicited **less change talk** ($U=2571$, $p=0.045$) and **more follow/neutral** utterances ($U=3745.5$, $p=0.006$).
- Chats in which CA outputs contained a high amount of potentially harmful behaviours elicited **less change talk** ($U=2126$, $p=0.005$).

No differences were found between groups for the amount of sustain talk elicited by the system. These results partly align with related work, which consistently finds an increase in change talk when counsellors use MI-consistent skills and a decrease in change talk when MI-inconsistent skills are used (Magill et al., 2014). Interestingly, however, increased MI-inconsistent skills, which have been found to increase sustain talk, only led to an increase in follow/neutral utterances in these experiments and had no effect on sustain talk.

7.5.3 User Reactions to Different Bot Utterance Types

Given the qualitative insights into user expectations and thoughts regarding specific CA behaviours in such conversations outlined in Section 7.4.3.3, I expected that differences in user reactions to these different CA behaviours could extend to the type of change language used. For this analysis, I grouped the bot behaviours prompted in the MI-adapted condition into the corresponding behaviour type groups (reflection, question, and encouragement, where encouragement contained the behaviours *affirmation* and *emphasise control*). In addition, I recast each utterance which had been annotated as “advise” or “giving information” as advise/facts. Following this, I ran three χ^2 tests to respectively compare valences, topics, and reason types elicited by the different bot behaviours. All three χ^2 tests were significant (valence: $\chi^2(6, N = 620) = 29.63, p < 0.001$; topic: $\chi^2(6, N = 348) = 19.04, p < 0.004$; reason type: $\chi^2(9, N = 286) = 21.67, p < 0.009$). A visual overview of the results is presented in Figure 7.13.

Valence

It became apparent that advise/facts and encouragement elicited more follow/neutral utterances and less change talk than expected, whereas the opposite was true for questions and reflections. In addition, reflections also elicited less sustain talk than would be expected in a random distribution, whereas questions led to higher amounts of sustain talk. A potential reason for this is that some question types also invite users to reflect on what they like about their current behaviour and what keeps them from changing it, eliciting sustain talk. In MI, such questions are used in moderation to increase rapport by making the client feel understood and accepted while at the same time raising motivation as they become aware that the positive outcomes of change outweigh the positive outcomes of the status quo (W. R. Miller & Rollnick, 2002).

Topic

As the majority of user utterances were classified as reason, and only a few utterances were about commitments or recently taken steps, results for the topic level, although significant, are merely indicative

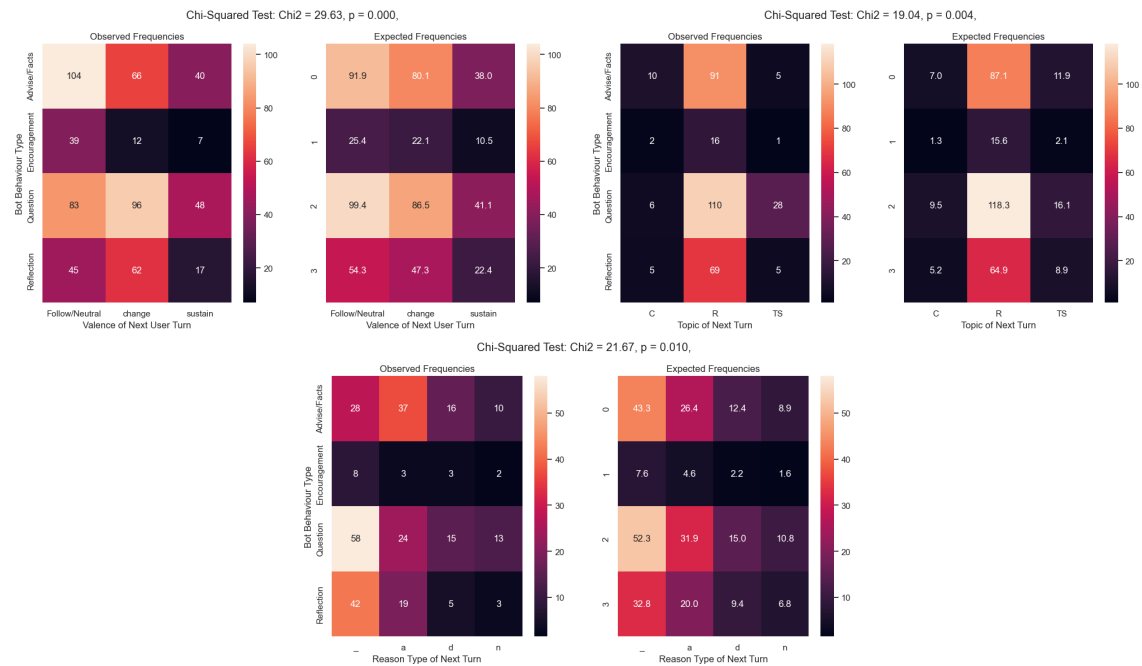


Figure 7.13: Expected and observed frequencies of user utterance classifications following a given bot behaviour.

but still show a certain trend. The main difference between observed frequencies and expected frequencies can be observed for questions, which elicited less reason and more taking steps than expected, and for advise/facts, which led to less taking steps and more commitment. There is also a trend of reflections eliciting predominantly reason, with fewer taking steps utterances than expected.

Reason Type

Again, the overall counts for the reason type level is lower, allowing only tentative conclusions. Surprisingly, advise/facts seemed to be successful at eliciting varied and focused responses, with general reason counts lower and ability, desire, and need higher than expected, whereas the opposite was true for the remaining bot utterance types.

These results can be seen as an indicator that, in conversations with LLMs, questions and reflections can be useful for making users aware of and voice their reasons for change, whereas providing advice and information in later stages of a conversation can support specific planning processes and increase users perceived ability to change. Such information and advice could be generated using retrieval augmented generation, which has been shown to minimise hallucinations (Lewis et al., 2020). These insights complement the qualitative insights provided by the user ratings of individual turns.

7.5.4 Impact of User Characteristics on User Language

Finally, I analysed to what extent user characteristics, specifically readiness to change, and behaviour in the interactions had an impact on the language they used. This is a first step towards the usage of such

classifications to allow conclusions about a person's current position towards change. It has to be noted, however, that such analyses would benefit from a more structured and controlled conversation flow design, where the bot behaviour is more comparable across all collected interactions. Nevertheless, even with the high variations in bot behaviours in the data, I could identify some differences in how different users used behaviour change language, indicating further evidence towards the broad applicability of the classifiers introduced in Chapter 4.

7.5.4.1 Readiness to Change

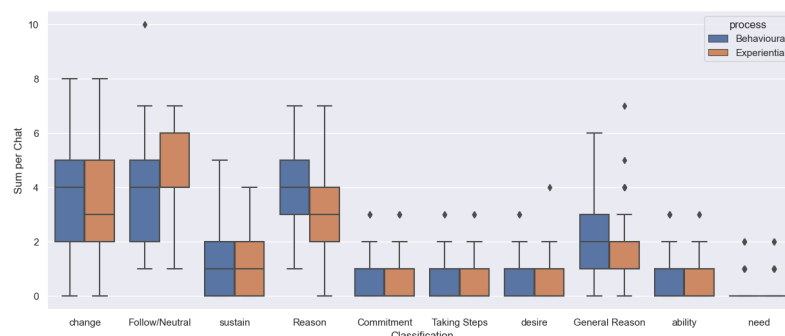


Figure 7.14: Classification counts per chat by binary stages of change.

To evaluate to what extent user language changes with differing levels of readiness to change, I divided study participants into two groups based on their indicated readiness to change at the start of the conversation: those who were in the experiential stages of change (precontemplation, contemplation), and those, who were in the behavioural stages of change (preparation, action, maintenance). This divide led to balanced group sizes, with 81 participants in the behavioural and 76 in the experiential stages. Boxplots of valence and topic classification counts per chat revealed visual differences in the amount of follow/neutral and reason utterances between the experiential and behavioural groups (see Figure 7.14). These differences were significant (Follow/Neutral: $U=2423$, $p=0.013$; Reason: $U=3794.5$, $p=0.003$), indicating that participants in the later stages of change, where actions towards change are taken, had a tendency to use more change-related language and were more likely to weigh their reasons for change. However, none of the other classes seemed to display any notable differences between the different groups.

7.5.4.2 User Behavioural Dimensions

In addition to exploring to what extent users with different levels of readiness to change differ in their language use, I also wanted to analyse whether the three identified behavioural dimensions were reflected in the way users interacted with the chatbot. Again, I plotted the counts of client utterance types per chat, given the valence of each behavioural dimension, and ran significance tests for the classes where a difference in median values was visible. Across all three behavioural dimensions, users who exhibited the behaviour used significantly more change talk and fewer follow/neutral utterances than others on

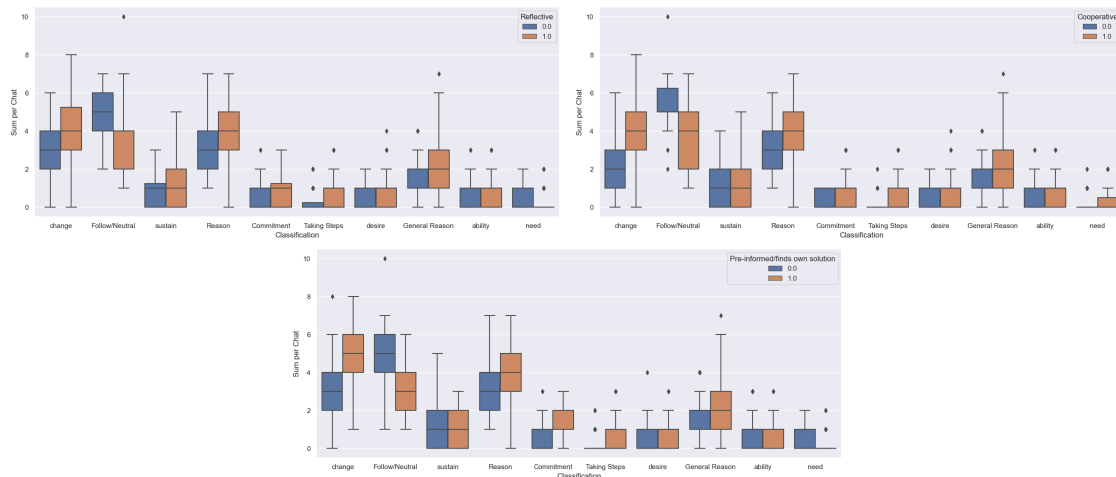


Figure 7.15: Classification counts per chat by behavioural dimensions.

the valence level. They also uttered more reason, commitment, and taking steps statements. Test statistics are provided in Table 7.10.

Class	Reflective	Cooperative	Preinformed
Follow/Neutral	U=1560.5, $p<0.001$	U=990, $p<0.001$	U=1346.5, $p<0.001$
change	U=4143, $p<0.001$	U=3509, $p<0.001$	U=4791.5, $p<0.001$
Reason	U=3918, $p<0.001$	U=3205, $p<0.001$	U=3576.5, $p=0.018$
Commitment	U=3988.5, $p<0.001$	U=3205, $p<0.001$	U=4296, $p<0.001$
Taking Steps	U=3463, $p=0.029$	U=2733, $p=0.021$	U=3768, $p<0.001$

Table 7.10: Mann-Whitney U statistics for tests of significance of differences in utterance counts within behavioural dimensions.

The results from this subsection corroborate the findings from Section 7.4.3 that users' current situations and their behaviour in interactions with LLM-based BCSSs play a significant role in the benefits users can reap from conversing with such a system.

7.6 Chapter Summary

In this chapter, I have presented results from a user study focused on human-LLM interactions about behaviour change. Interactions with the system led to significant increases in readiness to change across three distinct target behaviours (decrease procrastination, live more sustainably, healthier nutrition). In qualitative analysis, I have identified three behavioural dimensions displayed by users in this context: cooperation, reflectiveness, and preinformedness and found that each of these dimensions impacts the outcomes and user perceptions of interactions with the LLM in various ways. I have also identified qualitative evidence that these behavioural dimensions come with different expectations of the system. Furthermore, I have shown that GPT-4 can be manipulated to display less MI-inadherent and undesirable behaviour by incorporating MI principles in the prompting process. The results outlined in this chapter

provide insights into a) how users interact with LLM-based CAs in the conversations about behaviour change, b) how GPT-4 can be manipulated to return MI-adherent outputs, and c) what types of system responses and behaviours are preferred by users in this context. These insights can inform the future design of MI chatbots for behaviour change support tailored to different user types.

Chat 22. Condition: GPT-4. Cooperative: 0, reflective: 0, preinformed: 0.	
Speaker	Text
MIcha	Hi, I'm MIcha, your motivational chatbot. My goal is to help you reflect on your behaviour and support you in developing a plan to change your behaviour. Please think about my questions carefully and see them mainly as food for thought to get you mentally closer to your goal. I am not here to give you tips or recommendations for action, but to help you become clear about your motivations and values. What behavioural change do you want to talk about today?
User	I want to procrastinate less
MIcha	So you want to procrastinate less. Can you tell me something about why this change is important to you?
User	I want to be more productive
MIcha	On a scale from 0 (not at all) to 10 (very), how interested are you in doing this change right now?
User	6
MIcha	A 6. So you are generally in a positive mood towards this change. Can you tell me a bit more about this assessment? Why didn't you give a higher or lower score?
User	because I know that I have to change something and that this is going to be hard
MIcha	I understand that you are aware that a change is necessary, but also that it can be hard to implement it. It is important to be aware of the challenges. What could be some specific steps you could take to make less procrastination?
User	Put away mobile phone, think about future successes
MIcha	It sounds like you already have some ideas on how to achieve your goal. Keeping away from distractions like your mobile phone and thinking about future successes are good strategies. How could you integrate these measures into your everyday life?
User	I don't know
MIcha	This is perfectly fine, changes require time and patience. Here are a few suggestions: 1. Set fixed times at which you set the mobile phone aside. It can be helpful to combine these times with productive work or relaxation. 2. Try to use visualised goals. You could use an image of your goal on your
User	Thank you
MIcha	You're welcome! If you have more questions or need more support, I'll be at your disposal at any time. Remember, every little change is a step in the right direction. You can do it!
User	Bye-bye
MIcha	Bye! I wish you every success on your way. Don't hesitate to come back if you need further support.
User	End
MIcha	Let me summarise our conversation so far. You have made it clear that you want to procrastinate less to increase your productivity. You have shown a middle motivation (6 out of 10) for this change, because you are aware that a change is necessary, but also that it can be challenging. As possible measures to reduce procrastination you have proposed to put your mobile phone away and think of future successes. However, in the concrete integration of these measures into your everyday life you had difficulties and were unsure. Have I forgotten anything else?
User	yes
MIcha	Got it! Finally, please tell me again what your next step will be.
User	Set times without mobile
MIcha	Thank you for your openness in our conversation. I hope it has helped you to get a little closer to your goal, at least mentally.

Table 7.11: The user demonstrated a low level of cooperation, offering minimal responses without engaging deeply with the chatbot's prompts or developing the conversation further. Reflectiveness was rated as 0 because the user did not engage in deep reflection or provide elaborative answers when probed by the chatbot. Instead, they provided short, direct responses without elaboration. Preinformedness was also rated as 0 because the user did not independently suggest detailed or concrete solutions beyond the initial suggestions, and showed reliance on the chatbot for guidance rather than proposing personal, actionable strategies.

Chat 95. Condition: GPT-4. Cooperative: 1, reflective: 0, preinformed: 0.	
Speaker	Text
MIcha	Hi, I'm MIcha, your motivational chatbot. My goal is to help you reflect on your behaviour and support you in developing a plan to change your behaviour. Please think about my questions carefully and see them mainly as food for thought to get you mentally closer to your goal. I am not here to give you tips or recommendations for action, but to help you become clear about your motivations and values. What behavioural change do you want to talk about today?
User	I want to eat healthier
MIcha	So you want to eat healthier. Can you tell me something about why this change is important to you?
User	I eat too much fast food, I want to do more good for my body.
MIcha	How ready are you to perform this change on a scale from 0 (not at all) to 10 (very)?
User	7
MIcha	A 7. So you are generally in a positive mood towards this change. Can you tell me a bit more about this assessment? Why didn't you give a higher or lower score?
User	I am basically prepared to do so, but I still lack some motivation.
MIcha	Thank you for your honesty. It is perfectly normal that the motivation is sometimes lacking. Everyone has such days or phases. What could help you to increase your motivation?
User	Okay, I'm curious
MIcha	I understand you're curious, but I'm here to help you find your own solutions and motivate you, so what do you think might help you increase your motivation?
User	I need a diet plan.
MIcha	A diet plan could actually be a big help. It can give you a structure and a plan that you can stick to. What do you think, how realistic it is that you stick to a diet plan?
User	I think it's realistic, I just need a clear goal in front of my eyes
MIcha	This is a great attitude! Having a clear goal in mind can be very motivating. What would be this clear goal for you in relation to your diet?
User	The clear goal would be to eat less fast food and more fresh food
MIcha	That sounds like a great goal! To start the process, we could think about how you could break this goal down into smaller, manageable steps. What could be some first steps to eat less fast food and more fresh food?
User	Go to Mc Donalds and Burger King etc. less
MIcha	Let me summarise our conversation so far. You have realised that you consume too much fast food and now want to do something good for your body. You are ready to make this change with an assessment of 7 on a scale from 0 to 10, but you realise that you still lack some motivation. To increase it, you see a diet plan as helpful and think it realistic to stick to it. You have formulated the clear goal of eating less fast food and more fresh food. A first step in this direction would be for you to go less to fast food chains like McDonalds and Burger King. Have I forgotten anything else?
User	No, you've already mentioned everything
MIcha	Got it! Finally, please tell me again what your next step will be.
User	I will eat less fast food and shop healthier
MIcha	Thank you for your openness in our conversation. I hope it has helped you to get a little closer to your goal, at least mentally.

Table 7.12: The user was cooperative, consistently responding in full sentences and actively engaging with the chatbot's questions, contributing to the flow of the conversation. However, reflectiveness was rated as 0 because the user's responses, while compliant, did not demonstrate deep self-reflection or exploration of underlying reasons beyond initial surface answers. Pre-informedness was also rated as 0, as the user expressed a need for direction in developing a diet plan and relied on the chatbot for suggestions, without offering their own detailed strategies or concrete steps beyond general intentions to eat healthier.

Chat 99. Condition: GPT-4. Cooperative: 1, reflective: 1, preinformed: 0.	
Speaker	Text
MIcha	Hi, I'm MIcha, your motivational chatbot. My goal is to help you reflect on your behaviour and support you in developing a plan to change your behaviour. Please think about my questions carefully and see them mainly as food for thought to get you mentally closer to your goal. I am not here to give you tips or recommendations for action, but to help you become clear about your motivations and values. What behavioural change do you want to talk about today?
MIcha	So you want to procrastinate less. Can you tell me something about why this change is important to you?
User	I have a lot to do and always get stressed before the deadline, but I could easily manage everything if I was focused.
MIcha	On a scale from 0 (not at all) to 10 (very), how interested are you in doing this change right now?
User	9
MIcha	A 9. So you are generally in a positive mood towards this change. Can you tell me a bit more about this assessment? Why didn't you give a higher or lower score?
User	I really want to become more productive and see it as an important problem that I need to tackle. At the same time, I occasionally succeed and therefore no 10.
MIcha	I see, it sounds like you are really motivated to make this change, but also realistic enough to recognise that there are times when you are more productive than others. On a scale of 0 (not at all) to 10 (very), how prepared do you feel to tackle this change now?
User	I'm still lacking effective tools that suit my situation. But I am motivated and willing to try things that might work.
MIcha	It sounds like you're ready to try new strategies and tools to increase your productivity. That's a good start. Now, on a scale of 0 (not at all) to 10 (very much), how would you rate your current level of readiness to make this change now?
User	I'd say a seven.
MIcha	A 7 is a good score. It seems like you've already made some progress or at least have a plan for how you want to proceed. What could help you move from a 7 closer to a 10?
User	There are always phases in which you have tried something and achieved a few small successes. I still need more effective tools that I can use in a disciplined way.
MIcha	That sounds like a solid plan. It seems like you already have an idea of what you need to achieve this goal. Are there certain steps you plan to take or certain strategies you want to try out to work more effectively and procrastinate less?
User	I'm still missing the strategies that are proving effective.
MIcha	It's completely normal that you don't have all the answers yet. Perhaps you could try different techniques to see what works best for you. For example, you could try the Pomodoro technique, where you work in 25-minute blocks and then take a five-minute break. Or you could try dividing your tasks into
User	I've already tried the Pomodoro technique, but unfortunately I tend not to stick to the times and sometimes work over the break and then think I can take an extra long break.
MIcha	Let me summarise our conversation so far. So you've already tried the Pomodoro technique but realised that you're struggling to stick to the set times. You tend to work beyond the breaks and then plan to take a longer break to do so. You see this as a problem as you believe it is detrimental to your productivity and concentration. However, you are motivated and willing to try out different strategies and tools to reduce your procrastination and work more efficiently. Have I forgotten anything else?
User	That sums it up well
MIcha	Got it! Finally, please tell me again what your next step will be.
User	I'm going to rustle up more methods and try them out.
MIcha	Thank you for your openness in our conversation. I hope it has helped you to get a little closer to your goal, at least mentally.

Table 7.13: The user was highly cooperative, engaging thoroughly with the chatbot's prompts and responding thoughtfully to questions, thereby driving the conversation forward. Reflectiveness was rated as 1 due to the user's in-depth engagement with introspective queries about procrastination and its impact on productivity, and their willingness to analyse why certain methods did not work for them previously. However, Preinformedness was rated as 0, indicating that while the user was open to trying new methods, they did not independently propose specific, concrete strategies for change, instead relying on the chatbot to suggest techniques like the Pomodoro method.

Chat 188. Condition: MI-adapted. Cooperative: 1, reflective: 1, preinformed: 1.	
Speaker	Text
MIcha	Hi, I'm MIcha, your motivational chatbot. My goal is to help you reflect on your behaviour and support you in developing a plan to change your behaviour. Please think about my questions carefully and see them mainly as food for thought to get you mentally closer to your goal. I am not here to give you tips or recommendations for action, but to help you become clear about your motivations and values. What behavioural change do you want to talk about today?
User	I want to procrastinate less
MIcha	So you want to procrastinate less. Can you tell me something about why this change is important to you?
User	Because I want to focus on my work and my learning to be successful in the fields
MIcha	How important is this change on a scale from 0 (not at all) to 10 (very)?
User	8
MIcha	An 8. So you are generally in a positive mood towards this change. Can you tell me a bit more about this assessment? Why didn't you give a higher or lower score?
User	It is very widespread because it is related to my future, but the problem is not extremely pronounced.
MIcha	So it's an important concern for you, but not necessarily the most urgent.
User	Yes, but I think we should still deal with it in order not to deteriorate it.
MIcha	What would be the biggest challenge if you tried to procrastinate less?
User	Overcoming My Fear/Nervousness
MIcha	You're talking about fear and nervousness. Can you elaborate?
User	It's like an invisible hurdle that doesn't make me start. It reminds me all the time, but it doesn't make me tackle the issue.
MIcha	So you have a clear idea of what you need to do, but something keeps you from taking the first step. There are situations where you have been successful despite fear or nervousness.
User	I think if the time pressure is really strong I feel forced to do something and then it usually works. Just not necessarily very well, or relaxed.
MIcha	So you feel pressured, and that drives you to do your job - even if the result is not always satisfactory. How could it feel if you have control over your schedule and are not dependent on this pressure?
User	Then I could divide up the task, concentrate and plan my work every day without panicking. Moreover, it would probably be more high-quality and less often incomplete.
MIcha	Let me summarise our conversation so far. You have the desire to procrastinate less and to focus more on your work and your learning in order to be more successful in these areas. This change is very important to you, because it is closely linked to your future. However, you are confronted with an invisible hurdle that keeps you from beginning. This hurdle is expressed in the form of fear and nervousness. In situations with high time pressure, you can overcome this hurdle and be forced to perform tasks, but this does not lead to high quality results and is associated with discomfort. You strive to have control over your schedule and to plan and concentrate your tasks, in order to equally Have I forgotten anything else?
User	well summarised, only the last information is missing; daily planning to avoid panic
MIcha	Got it! Finally, please tell me again what your next step will be.
User	First examine and document external circumstances of a problem/work in order to actually (pre-)plan to keep fear and nervousness in check
MIcha	Thank you for your openness in our conversation. I hope it has helped you to get a little closer to your goal, at least mentally.

Table 7.14: The user was highly cooperative, engaging deeply with the chatbot's questions and contributing meaningfully to the conversation. Reflectiveness was rated as 1 due to the user's ability to introspect and discuss underlying issues like fear and nervousness that affect their ability to tackle procrastination. They engaged in detailed exploration of these internal challenges and their impact on behavior. Preinformedness was rated as 1 because the user not only identified specific obstacles but also proposed a structured approach to managing these issues, indicating a proactive attitude towards finding solutions and planning for change.

8

Discussion

Throughout this thesis, the focus lay on replicating two aspects integral to interpersonal communication: understanding a person's attitude about the conversation topic (specifically, behaviour change) and reacting appropriately. In this project, the scope was focused on synthesising an understanding of these aspects within the scope of intentional behaviour change in order to enable LLM-based conversational AI to provide effective behaviour change support and explore to what extent this is accepted by users. The psychological theoretical construct used to achieve this understanding and generate appropriate reactions was MI.

Chapter 4 focused on the construction and description of the GLoHBCD, a dataset of annotated written conversations about behaviour change sourced from forum data and the subsequent fine-tuning of classifiers for behaviour change language. Chapter 5 described experiments to showcase the suitability of the classifiers for the classification of user utterances in human-chatbot interactions, which yielded promising results in favour of the application of the classifiers across various domains and conversational contexts. In Chapter 6, I outlined extensive data augmentation and semi-supervised learning experiments conducted in an attempt to improve the performance of the classifiers. Lastly, Chapter 7 described the design and analysis of an extensive user study to test a conversational framework employing these classifiers based on MI principles and identify how users interact with LLM-based systems in realistic interactions around behaviour change, as well as exploring user expectations in this context. This chapter also provided insights into how the classifiers based on the GLoHBCD can be used to analyse such interactions.

Below, I discuss my findings in light of the research goals introduced in Chapter 1 and the background and related work outlined in Chapters 2 and 3 (Sections 8.1 to 8.3). This includes reflections on similarities and differences between the collected interactions with MIcha and real counselling sessions as described in MI handbooks and reported in related studies (Section 8.3.3). Based on all findings, I provide design implications for LLM-based BCSSs (Section 8.4). Finally, I outline the main limitations of this work (Section 8.5), followed by open questions and potential for future work, including an overview of alternative use cases of the main technical contributions provided by this research project (Section 8.6).

8.1 Research Goal 1

The first research goal focused on exploring how and to what extent MI can be integrated into the conversation flow of LLM-based BCSSs to increase understanding of user attitudes about change and provide more suitable and safe system reactions. My experiments revealed that MI concepts have potential use cases across all of these aspects, as outlined below.

8.1.1 Understanding User Attitudes About Change

First, the behaviour change language classifiers based on the GLoHBCD show that it is feasible to predict a wide range of user attitudes about behaviour change from written texts, using *MISC* labels, which reveal information on both the valence and the content of an utterance with regard to change. My experiments also demonstrated the performance of the classifiers to be robust across a variety of conversational contexts and target behaviours, allowing for the interpretation of user utterances about change. This leads to the conclusion that the classifiers have the potential to deepen a system's understanding of users in human-bot interactions in real time.

While initial experiments on the collected interactions with MIcha did not allow for the drawing of conclusions regarding users' readiness or stage of change, certain trends are visible in language use based on stages of change and reflectiveness and preinformedness. Considering the shortness of the interactions collected here, most reason types and topics other than reason were barely present in the data, which made it harder to shed light on different language use across user groups. Longer interactions of this type might reveal more conclusive differences between the language used by people with differing levels of preparedness. Likewise, such language use patterns might be highly person-specific, so following the language use of individual people over prolonged periods of time, identifying progressions and trends might be an effective approach to identify patterns.

8.1.2 Providing Suitable and Safe System Reactions

MI textbooks (Clifford & Curtis, 2016; W. R. Miller & Rollnick, 2002) give insights into how practitioners can react to different expressions of attitudes about change to best support clients' thought processes. In the MI-adapted condition of the user study outlined in Chapter 7, I mapped the practitioner behaviours outlined in these textbooks to *MISC* client utterances based on the textbook descriptions. This information was then used to select suitable bot behaviours to prompt GPT-4 based on the most recent user utterance classification. Although this did not lead to an increase in the participants' perceptions of MI adherence and therapeutic alliance compared to the GPT-4 condition, it was successful in mitigating both user-perceived and empirically measured MI-inadherent and unsuitable bot behaviours. This indicates that prompting according to MI principles in interactions about change with LLMs has an effect on the safety and potential for harm in such interactions. It is worth noting that initial tests regarding the fulfilment of prompted actions showed that in slightly over 20% of cases, the prompted behaviour was not fulfilled by the system. However, the consistently low number of potentially harmful behaviours indicates that even in the cases in which the behaviour was not met, MI-adherent behaviour was achieved.

It is noted that the harm filters employed by OpenAI seem to have become very good at filtering openly harmful statements, such as criticism or directives, of which I found not a single instance in the data, but are less aware of subtle ways in which conversations with LLMs could manipulate users, for instance through increased anthropomorphism. The annotation scheme I presented in Section 7.2 presents a method to recognise such subtle, potentially harmful outputs in human-LLM interactions and serves as an addition to existing evaluation methods focusing more on evaluating openly harmful utterances in interactions with human counsellors (as presented in the *MISC* (W. R. Miller et al., 2008)), or the presence of desirable behaviours in written contexts (Chiu et al., 2024; Sharma et al., 2020; Welivita & Pu, 2020). Using this annotation scheme, I found that guiding GPT-4 towards counselling behaviour through context-tuning considerably lowers the presence of these subtle harm potentials without a significant negative impact on other parts of the conversation, with the exception of therapeutic alliance, which was perceived to be significantly lower in the MI-adapted condition, presumably due to the fact that the MI-adapted condition was better at reducing anthropomorphism and was less goal-oriented, focusing more on soliciting the user's perspective.

8.2 Research Goal 2

The focus of the second research goal was to shed light on users' expectations for and interaction styles with LLM-based BCSSs and in what way these a) are influenced by their attitudes about and readiness for change and b) have an influence on the outcome of such interactions.

The behavioural dimensions identified in the qualitative analysis corroborate my intuition that user behaviour has a significant impact on the success and quality of interactions about change, showcasing that users who behave more cooperatively and reflectively are more likely to benefit from conversations with the system. I argue that this outcome is a favourable one, as it shows that in both conditions, the chatbot was not able to manipulate users to do something they did not want to do. Instead, it was a helpful resource for users who already had a certain level of willingness to change, to self-reflect, and to come up with solutions, with guidance by a conversational system.

My findings about the positive effects of reflectiveness on user perceptions and readiness to change can be conceptualised in the context of various works which have found benefits of journaling and self-reflection in the context of mental health and behaviour change (Pennebaker & Seagal, 1999). Ullrich and Lutgendorf (2002) found that journaling about cognitions and emotions fosters self-reflection and improves positive outlooks in the face of adverse situations more than focusing only on emotions or facts. Similarly, Rafizadeh et al. (2021) found that journaling has the potential to increase self-efficacy and intrinsic motivation and Hensley and Munn (2020) showed that it can decrease procrastination. In structured behaviour change programmes, practising journaling has been linked to increased engagement across various target behaviours (Stephenson & Zygouris, 2007). Such effects seem to be especially strong when diarists are guided in a certain direction, as suggested by Maclsaac et al. (2023), who found that guided journaling, especially based on positive prompts, has a higher positive effect on well-being than self-reflection alone, which can easily lead to rumination. Crucially, such effects seem to prevail regardless of audience (private vs. shared), modality (spoken vs. written) or writing topic (trauma vs. non-traumatic events such as starting college), but there have been indications that self-reflection is most successful if diarists acknowledge negative emotions while laying focus on positive

aspects of their situation (Pennebaker & Seagal, 1999).

In a focus group conducted with psychotherapists, Kim et al. (2024) identified that the way users react to LLM messages could lead to valuable insights, as different people expect different system reactions (i.e. direct feedback vs. reassurance). This was also apparent in my data. For instance, although the introductory message given at the start of the conversation with Mlcha clearly stated that the goal of the conversation was not to give advice, I found that many participants expected to get advice and were disappointed when their questions were not answered. Likewise, many participants did not appreciate being asked many questions or being presented with reflections of their utterances, whereas others welcomed such system behaviours and were happy about the chance to self-reflect. Importantly, I found that users who identified with their chosen target behaviour strongly and were currently pursuing this behaviour were more likely to behave in ways that led to conversational benefits. As a result, I presume that employing such a system in real life for people who are intentionally working toward change would have more reliable and stronger benefits than found in this study.

8.3 Theoretical Lenses

The disciplines introduced in Section 2.1 allow the interpretation of my experiments and results through different theoretical lenses, elucidating different aspects of this work. While theoretical and empirical insights from behavioural science are most relevant to this work due to its focus on incorporating MI principles, the results of the user study also show certain parallels to theoretical concepts from information science, and insights from human-computer interaction can help in the categorisation of the type of behavioural changes that can be supported by the LLM-based conversational framework compared to traditional systems used in human-computer interaction research.

8.3.1 Information Science

Information science is typically concerned with how users interact with information in different situations and during different periods of their lives. This includes the identification of factors that lead to users' individual preferences and use patterns and the categorisation of different user types. While my main focus lay on the elicitation of information, my work still shows parallels to many of the insights outlined in information science research. For instance, Kuhlthau's Model of Information Seeking Behaviour (Kuhlthau, 1988) shows that information seekers typically start a task with feelings of uncertainty and confusion, leading to ambiguous thoughts, whereas in later stages, increasing clarity and focus is reached, as they learn more about the nature of the task. Similarly, Ruthven (2024) introduced the concept of information shaping as a transition from broad and unspecific to specific and targeted, with various iterations and feedback loops. These models can partly explain the trend that users in the early stages saw higher increases in readiness to change when interacting with the system, as its primary focus was the exploration of ambivalent thoughts and insecurity. In this light, the conversational framework outlined here might be a tool to support users in achieving increased *mastery of life*, as introduced by Savolainen (2017), as it supports them to explore the cognitive dissonance between how things are currently, and how they should ideally be.

Likewise, the concept of Monitors and Blunters introduced by S. M. Miller (1980) can be related to the

behavioural dimensions of reflectiveness, preinformedness, and cooperation identified here. Interpreted in this way, individuals with low preinformedness have likely been blunting information in the past. However, it is also possible to compare users displaying low reflectiveness and cooperation to blunters since they appear to be less willing to observe their own cognitive dissonance and explore their options regarding behaviour change than their monitoring counterparts, who willingly and openly engage with any questions posed to them. As such, the behavioural dimensions identified here can serve as an indicator to what extent users are blunting or monitoring information and thoughts found in themselves.

Lastly, significant parallels can be drawn to Ruthven's concepts of meaning-making (2019) and transitions (2022b). Both of these models are concerned with processes that happen when things change in a person's life, both through their own volition and outside circumstances. A focus of these theories is that merely gaining information is not enough to achieve meaning. Instead, information has to be felt by the person receiving it and address their desires and motivations. Such meaning can, among others, be reached through blogging or journaling, of which the interactions with MIcha can be considered a form of, as they support the exploration of cognitive dissonance and prompt users to think about the deeper meaning of behaviour change for themselves specifically. Furthermore, the research gap defined by Ruthven (2019) that it is important to provide the right kind of support at the right stage of meaning-making is reinforced by the results of this thesis.

8.3.2 Human-Computer Interaction

Both the FBM (Fogg, 2009b) and the COM-B model (Michie et al., 2011) define behaviour as necessitating an ability, motivation, and a trigger or opportunity to occur. While the conversational framework introduced here deals mostly with increasing motivation and user-perceived ability (i.e. self-efficacy), it could also be expanded to help users identify triggers that they can set in place for themselves.

The digital intervention introduced in this thesis is not concerned with coercing or nudging users to do something they do not want to do, which represents a difference to many of the interventions Fogg describes in his behaviour grid (2009a). As a result, and due to its conversational nature, it is suited for different types of behaviour change than Fogg (2009a) identifies as well suited for mobile interactions (i.e. performance of existing behaviours or increasing a behaviour on a reoccurring schedule). Utilising MI and conversational AI, I showed that the dialogue framework introduced here can help motivate users to both increase and decrease existing behaviours on the examples of decreasing procrastination and increasing healthy nutrition and sustainability. The fact that some participants also indicated being in the precontemplation or contemplation stages at the beginning of the intervention indicates that it might also be useful to support users in gaining motivation to start new behaviours. None of the behaviours dealt with in the context of the user study necessarily happen on a predictable schedule, although such schedules could be identified and put into practice by users over the course of prolonged interactions with the system and in some cases, such ideas were discussed in the interactions (such as the idea of leaving the mobile phone in a different room when working).

As such, when categorised in light of the O/C matrix introduced by Oinas-Kukkonen (2013), the behavioural changes discussed in the context of this thesis can be categorised as more than mere acts of complying, instead constituting steps towards the forming of new behaviours, or the altering or reinforcement of existing behaviours, with the potential to even cause changes in attitude. This

indicates that as a single intervention type, the introduced framework has the potential to cover various different types of change.

8.3.3 Behavioural Science and Motivational Interviewing: Similarities and Differences Between Collected Interactions and Real Counselling Sessions

Here, I conduct a comparison between the data collected in interactions with *MIcha* and interactions between therapists and clients. I also include comparisons to other CAs employing MI principles where possible, based on the type and amount of information reported in the related work.

When comparing the interactions collected in the user study to the *AnnoMI* (Wu, Balloccu, Kumar, et al., 2022) and the *BiMISC* (Sun et al., 2024), both based on transcripts of real counselling sessions and training materials, a number of similarities and differences become apparent. First and foremost, the length of interactions with *MIcha* is significantly shorter than interactions in the other datasets, which have an average interaction length of 80 and 108 turns, respectively. The *Wizard of OZ* interactions between users and companionbots described by Guntakandla and Nielsen (2018), with psychologists as wizards, had a similar mean length of 115 turns. In other studies introducing or testing CAs employing MI, the typical interaction length was rarely reported. However, where reported or reproducible, it spanned more turns than the 12-turn interactions with *MIcha*. For instance, He et al. (2022) stretch conversations across two (consecutive) sessions, which together span above 20 turns, whereas da Silva et al. (2018) collected 30-minute spoken conversations, likely resulting in a larger number of turns as well. The shortness of the interactions with *MIcha* is a potential factor for the lack of differences between study conditions regarding their impact on readiness to change. A similar conclusion was reached by He et al. (2022), citing related studies (e.g. M. F. Davis et al. (2011)) who do not find significant effects of MI interventions shorter than 20 minutes over informational interventions in the context of smoking cessations. The same trends might apply to other target behaviours such as the ones explored in the context of this thesis. Adding to this, even a spoken 15-20 minute brief motivational interview is highly likely to span more than 12 turns. Considering this, however, the outcome that even such short conversations with LLMs can significantly increase readiness to change is a compelling indicator towards the high efficacy of utilising such a form of guided journaling to facilitate behaviour change in the context of individual self-help or complement structured behaviour change programmes in which MI cannot be administered by the practitioners responsible.

However, despite the shortness of the interactions, compared to real counselling sessions, they show similarities to the *AnnoMI* and *BiMISC* in terms of the elicited user utterance types. Figure 8.1 shows a comparison conducted by Sun et al. (2024) of the valences of client utterances in the *BiMISC* and *AnnoMI* (referred to as NT for follow/neutral, CT for change talk, and ST for sustain talk). In both datasets, follow/neutral represented the largest group, followed by change talk and then sustain talk. The same is reflected in interactions with *MIcha*, where change talk was more than three times as common as sustain talk on average, and the average count of follow/neutral utterances made up around 40% of user statements. Similarities to the *BiMISC* are also present in the fine-grained code distributions. Figure 8.2 shows that the most common topic label in the *BiMISC* was reason, with commitment statements (referred to as CM) being much less common, and taking steps barely present

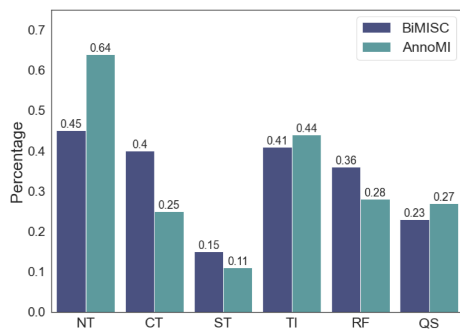


Figure 8.1: Coarse-grained code distributions in BiMISC and AnnoMI. Figure published under CC BY 4.0 license in Sun et al. (2024, p. 5613).

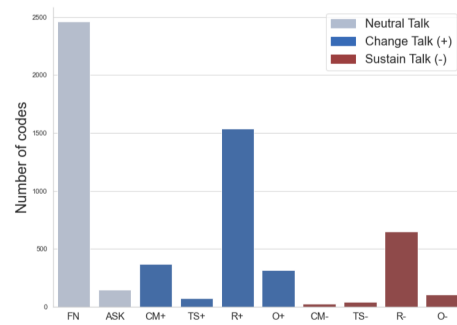


Figure 8.2: Fine-grained client code distribution in BiMISC. Figure published under CC BY 4.0 license in Sun et al. (2024, p. 5613).

at all. The same can be observed in the data collected in the user study conducted as part of this thesis. The vast majority consisted of reason statements, followed by commitment, and taking steps. Similar to the BiMISC, commitment statements were generally change talk, with only a single sustain talk commitment statement uttered across all collected chats. Neither the BiMISC nor the AnnoMI provide annotations or analyses of reason types in their data.

Other similarities to real counselling sessions can be derived from general behavioural science research. Considering the fundamentals of MI, which does not focus on giving advice (except with previous permission and in moderation) but instead intends to help clients come up with answers and strategies themselves, it is not surprising that conversations with participants who were already preinformed about their chosen target behaviour were more likely to end in a change talk statement regarding their next step. This also aligns with the observation that individuals in earlier stages of change at the beginning of the conversation saw larger increases in readiness to change. As outlined by DiClemente and Velasquez (2002), such individuals benefit more from reflection, as it supports increasing awareness about problem behaviours and the benefits of alternatives. They dislike receiving instructions or advice, while users in later stages of change might already have this awareness and have done considerable amounts of self-reflection in the past, resulting in different expectations, such as advice or coaching style guidance. I show that these assumptions, made by DiClemente and Velasquez (2002) in the context of traditional counselling, seem to hold for LLM-based interactions around behaviour change. Thus, future work could focus on identifying a user's stage of change early in a conversation and then tailor the interaction style to this stage, either prioritising LLM-based questions and reflections or, in later stages, the provision of safe advice. This could, for instance, be done with the help of retrieval augmented generation. An approach to this has been presented by Corda et al. (2024), who used LLMs to generate personalised advice for improving sleep.

Since the conversational framework designed in this thesis specifically focused on avoiding information provision and advice, the resulting conversations bear one major difference to real counselling sessions and scripted chatbot interactions described in the related work (e.g. A. Brown et al. (2023); He et al. (2022); Olafsson et al. (2019)) in so far that the chatbot in the MI-adapted condition had less of an agenda and was more reactive than proactive in discussing behaviour change goals. This differs from other MI interactions, which often have the goal of specifically motivating clients in the

precontemplation stage (i.e. those who do not see a problem with their current behaviour). This might contribute to the low therapeutic alliance scores of the MI-adapted condition compared to the GPT-4 condition, but also compared to He et al.'s work (2022), in which therapeutic alliance scores were at a mean of 3.25 in the MI condition. The questions in the Working Alliance Inventory are based primarily on goal-oriented and interpersonal aspects of therapist-client relationships, both of which were not the focus of the conversations presented here. Nevertheless, the indication that users in the early stages of change benefitted more from conversing with the chatbot indicates that the interactions based on the conversational framework designed in this thesis can potentially have a similar impact as more directive MI interactions. However, a reframing of these as "MI-based guided journaling" might reflect the nature of the interactions in the MI-adapted condition better than referring to them as chatbot conversations.

Even though many users were not satisfied with the interactions due to the lack of advice given and the amount and type of questions they were asked, reflectiveness during conversations was one of the predictors for an increase in readiness to change. These results align with previous work in this sphere, which has shown that user desires and expectations do not always align with what leads to the best outcomes. For example, Olafsson et al. (2019) find that users preferred to have a full range of options, even though coerced change talk increased their confidence in behaviour change. Similarly, work by Wu, Balloccu, Reiter, et al. (2023) showed that what users consider good content in such interactions does not necessarily align with what is considered good practice, showcasing that what seems to be "parroting" to users is often not perceived as repetitive by experts. This mirrors, to a certain extent, the entire process of behaviour change: what a person wants to do in a given moment might not be what benefits them long-term. As a result, while many of the reactions of study participants to bot utterances align with previous work evaluating rule-based CAs, which has shown that follow-up questions and feedback given by such systems is quickly perceived as repetitive, generic, clichéd, and impersonal with some user groups losing patience quickly and looking for quick advice instead of lengthy self-reflection sessions (Kocielnik et al., 2018; Park et al., 2019; Yao et al., 2023), such utterances are likely to create value when seriously considered and when users put effort into answering them.

8.4 Design Implications for LLM-Based Conversational Behaviour Change Support

Based on the results, some design implications for future LLM-based BCSSs can be drawn, which are outlined below.

- **Use MI strategies to mitigate harms in LLM-based behaviour change support conversations**

MI principles, when embedded into chatbot interactions, can reduce user perceptions of harm and inappropriate responses. This approach requires careful mapping of MI strategies onto chatbot responses to ensure that the conversation remains supportive and adherent to MI principles, even when the chatbot fails to perfectly align its responses with the expected behaviour.

- **Pay attention to user types**

User interaction types—reflective, cooperative, preinformed—significantly affect interaction outcomes with BCSS-CAs. Tailoring the conversation style to fit these types can enhance the effectiveness of behaviour change interventions, making the system more responsive and better suited to individual user needs.

- **Analyse user language**

Analysing user language for readiness and willingness to change can inform more personalised and contextually appropriate responses. This analysis should go beyond mere word recognition, incorporating an understanding of sentiment, commitment, and the reasons behind behaviours to better support users through the change process.

- **Find alternative ways to build rapport and be careful about how interactions are framed**

Traditional MI techniques that build rapport, leading clients to open up about themselves and their problems might be less effective in a digital context where users perceive chatbots as inherently non-judgemental. Alternative strategies such as reframing interactions as guided journalling might encourage self-reflection more effectively. This approach reframes interactions as personal explorations rather than direct conversations, potentially increasing user engagement and the effectiveness of interventions. Leveraging the non-directive ethos of MI in this way while adapting to the strengths and limitations of conversational AI could furthermore help manage user expectations, potentially leading to more meaningful and reflective user interactions.

- **Provide personalised advice - in moderation** While users often seek direct advice, excessive or poorly timed advice can detract from self-guided discovery and empowerment crucial for behaviour change. Providing personalised, evidence-based suggestions in moderation—guided by user expressions and readiness to change—could enhance user satisfaction without undermining the self-discovery process encouraged by MI.

8.5 Limitations

This research has several limitations that must be acknowledged. First, the user study design involved offering paid participants a selection of interaction options. Participants were given a choice of three personas and told to choose the scenario they identified with most closely. This approach was taken to increase engagement and avoid potential harm if bad advice was generated in the GPT-4 condition, but it does not fully replicate the dynamics of entirely naturalistic conversations about self-chosen target behaviours. Recruiting participants through prolific may also have attracted participants who are not inherently motivated to interact with the CA, potentially skewing the results away from what might be observed in more realistic settings where users engage out of a genuine personal initiative.

Another limitation concerns the generalisability of the harm reduction observed with the GPT-4 model. This reduction may not necessarily apply to other language models or future iterations of GPT due to the opaque nature of these models and the often undisclosed specifics of their training data. This 'black box' aspect of large language models highlights the challenge of replicating results across different AI systems.

Furthermore, my conversational design intentionally omitted the planning phase of MI interactions to avoid providing direct advice, concentrating instead on the initial stages of engaging, focusing and

evoking. These are primarily aimed at increasing readiness and motivation to change and resolving ambivalence. This exclusion limits the scope of my findings to the earlier stages of MI, leaving the subsequent planning and action phases unexplored.

Lastly, while there was a notable increase in readiness to change among participants, the experiments did not measure the extent to which this readiness translates into actual behaviour change or the maintenance of such changes over time. Addressing these outcomes would require a longitudinal study design, which could provide insights into the long-term impacts of MI-based conversational AI on behaviour change.

8.6 Open Questions and Future Work

The resources and results presented in this thesis open various potential pathways for future work, which are described below. First, MI literature suggests that engaging unreflective individuals in reflection can be challenging but not impossible (W. R. Miller & Rollnick, 2002). MI techniques focus on fostering an environment where individuals feel safe and supported to explore their thoughts and feelings, which may, in turn, encourage them to see the value in reflection. Future studies should investigate how these techniques can be adapted to conversational AI to effectively engage unreflective users, helping them understand and appreciate the benefits of self-reflection.

The classifiers based on the GLOHBCD have various potential use cases beyond the application to conversational AI as intended in this work. Since DARN-CAT utterances can be seen to map to a spectrum of increasing readiness to change (Resnicow et al., 2012), the classifiers could be used for a variety of analysis tasks. For instance, they could be used to add quantitative insights to qualitative research of forum data around behaviour change (Myneni et al., 2016; Rodda et al., 2018) and provide a new structured and theory-based lens through which to explore various types of conversation around change (Carah et al., 2017; Fernández et al., 2016; Gupta et al., 2020). The ability to classify fine-grained topics (commitment, reason, taking steps) and reason types (need, ability, desire), which can serve as proxies for psychological phenomena such as self-efficacy and self-evaluated dissatisfaction, opens up additional potential use-cases in the broader context of psychology, such as the detection of depression, anorexia nervosa, and gambling addiction on social media (Losada et al., 2019; Parapar et al., 2021). Furthermore, these classifiers could be employed to support counsellors and coaches, both of which have limited time to spend with clients. Classifying clients' journal entries between sessions and providing an overview and visualisations of their stances about behaviour change throughout the week could provide valuable insights into their progress, as was, for instance, shown by Kim et al. (2024) in their evaluation of an AI-based journalling tool for clinical patients with depression. Similar analyses could also serve as a way for users themselves to self-track their optimism about behaviour change over time.

The data collected in the user study holds promise for multiple disciplines, including human-computer interaction and NLP. It could improve the training of classifiers and generative models to enhance LLM outputs, making them more sensitive and appropriate in various interaction contexts. For instance, the harm detection framework I introduced and used to annotate the collected data can identify subtle, potentially manipulative interactions that might not be flagged by current systems, which tend to focus on more overtly harmful content. This scheme could be developed into a computational method, allowing

for real-time detection and modification of potentially harmful LLM outputs. This would be particularly useful in enhancing the safety of LLM interactions by reducing unintended manipulation through increased anthropomorphism and ensuring that interactions remain supportive and user-centred. Furthermore, using the collected data to analyse how users express their information needs, both explicitly and implicitly, could provide a foundation for improving conversational search and interaction. This capability could be particularly impactful in LLM applications, where anticipating and addressing user needs proactively can significantly enhance user experience and satisfaction. Such endeavours could be supported through retrieval augmented generation strategies, in which text generation is paired with a knowledge base to provide high-quality and personalised information.

Future research should also focus on developing methods to detect user characteristics, such as reflectiveness, cooperation, and preinformedness, early on during interactions. This could enable CAs to adapt their strategies in real-time, tailoring interactions to better meet individual user needs and enhance the effectiveness of behaviour change interventions.

These areas represent significant opportunities for advancing the field of conversational AI in behaviour change, contributing to more personalised, effective, and safe interactions.

Conclusion

This thesis has made several contributions to the field of conversational AI, particularly in the context of intentional behaviour change. These contributions are both technical and theoretical, advancing our understanding and capabilities in designing LLM-based systems that support behaviour change. The main technical contributions and theoretical insights are listed below.

Main Technical Contributions

- **GLoHBCD: A resource for written conversations about intentional behaviour change** - The GLoHBCD provides a valuable resource for exploring how written communication about behaviour change occurs in naturalistic settings. This dataset has served as a foundational tool for this thesis and can be useful for further research and development in conversational AI.
- **Transferable and robust classification algorithms for behaviour change language** - The creation and validation of robust classification algorithms across different contexts and behaviours enhances our ability to understand user attitudes about change and can help guide conversational agents' reactions to user utterances.
- **Strategies to significantly reduce harms and unsuitable system outputs in LLM-based behaviour change support conversations** - This research showed that implementing MI strategies in a CA's conversation flow mitigates the risks associated with conversational AI in the context of behaviour change, enhancing the safety and effectiveness of these systems. This shows that the incorporation of MI principles represents an effective way for ensuring that such systems support users in a way that is helpful without being harmful.
- **Annotation scheme for the identification of unsuitable LLM-behaviours** - The development of an annotation scheme specifically designed to identify unsuitable behaviours in LLM outputs in the context of behaviour change represents a step towards making conversational AI in this context more ethically sound. This scheme can help in refining LLM behaviour and reducing ethical harms, such as the danger of anthropomorphism and hallucinations.
- **Annotated dataset of real human-LLM interactions about behaviour change** - The anno-

tated dataset provides insights into how real interactions unfold, offering data that can be used to train and improve AI models and gain insights further into user behaviours and expectations.

Theoretical Insights

- **Domain independence of behaviour change language** - This thesis provides evidence that language associated with behaviour change is largely independent of specific domains, suggesting that effective communication strategies in one area of behaviour change can be applicable to others. This finding expands the potential applications of conversational AI across different aspects of health and wellness.
- **User behaviour and expectations when interacting with LLMs about behaviour change** - Understanding user behaviour and expectations is critical for designing conversational AI that meets users' needs. This research provides various insights into how users interact with LLM-based systems, shedding light on what users expect from these interactions and how they perceive the assistance provided by LLMs in the context of behaviour change.

Overall, this thesis supports advances in the technical capabilities in the realm of LLM-based conversational AI for behaviour change but also deepens theoretical understanding of how these systems can be most effectively designed in multiple ways. The insights and tools developed here hold potential for various future research projects that could further refine these approaches, ensuring that conversational AI continues to evolve as a beneficial tool in supporting individuals through behaviour change processes.

Bibliography

- Abercrombie, G., Curry, A. C., Dinkar, T., Rieser, V., & Talat, Z. (2023). Mirages. On anthropomorphism in dialogue systems. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing, EMNLP 2023, singapore, december 6-10, 2023* (pp. 4776–4790). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/2023.emnlp-main.290> doi: 10.18653/V1/2023.EMNLP-MAIN.290
- Abraham, C., & Michie, S. (2008). A taxonomy of behavior change techniques used in interventions. *Health Psychology, 27*(3), 379–387. Retrieved from <https://doi.org/10.1037/0278-6133.27.3.379> doi: 10.1037/0278-6133.27.3.379
- Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. *Machine Learning with Applications, 2*, 100006. Retrieved from <https://doi.org/10.1016/j.mlwa.2020.100006> doi: 10.1016/j.mlwa.2020.100006
- Aggarwal, A., Tam, C. C., Wu, D., Li, X., & Qiao, S. (2023). Artificial intelligence-based chatbots for promoting health behavioral changes: Systematic review. *J Med Internet Res, 25*, e40789. Retrieved from <https://doi.org/10.2196/40789> doi: 10.2196/40789
- Ajzen, I. (1985). From intentions to actions: A theory of planned behavior. In J. Kuhl & J. Beckmann (Eds.), *Action control: From cognition to behavior* (pp. 11–39). Springer Berlin Heidelberg. Retrieved from https://doi.org/10.1007/978-3-642-69746-3_2 doi: 10.1007/978-3-642-69746-3_2
- Ajzen, I. (1987). Attitudes, traits, and actions: Dispositional prediction of behavior in personality and social psychology. In L. Berkowitz (Ed.), (Vol. 20, pp. 1–63). Academic Press. Retrieved from [https://doi.org/10.1016/S0065-2601\(08\)60411-6](https://doi.org/10.1016/S0065-2601(08)60411-6) doi: 10.1016/S0065-2601(08)60411-6
- Ajzen, I. (2002). Perceived behavioral control, self-efficacy, locus of control, and the theory of planned behavior. *Journal of Applied Social Psychology, 32*(4), 665–683. Retrieved from <https://doi.org/10.1111/j.1559-1816.2002.tb00236.x> doi: 10.1111/j.1559-1816.2002.tb00236.x
- Almusharraf, F., Rose, J., & Selby, P. (2020). Engaging unmotivated smokers to move toward quitting: Design of motivational interviewing-based chatbot through iterative interactions. *J Med Internet Res, 22*(11), e20251. Retrieved from <https://doi.org/10.2196/20251> doi: 10.2196/20251
- Androutsopoulos, J. (2013, 05). Online data collection. In *Data collection in sociolinguistics* (p. 236-250). Retrieved from <https://www.taylorfrancis.com/chapters/edit/10.4324/9781315535258-47/online-data-collection-jannis-androutsopoulos> doi: 10.4324/9780203136065
- Apodaca, T. R., Jackson, K. M., Borsari, B., Magill, M., Longabaugh, R., Mastroleo, N. R., & Barnett, N. P. (2016). Which individual therapist behaviors elicit client change talk and sustain talk

- in motivational interviewing? *Journal of Substance Abuse Treatment*, 61, 60–65. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0740547215002433> doi: <https://doi.org/10.1016/j.jsat.2015.09.001>
- Apodaca, T. R., & Longabaugh, R. (2009). Mechanisms of change in motivational interviewing: A review and preliminary evaluation of the evidence. *Addiction*, 104, 705–715. Retrieved from <https://doi.org/10.1111/j.1360-0443.2009.02527.x> doi: 10.1111/j.1360-0443.2009.02527.x
- Armstrong, M. J., Mottershead, T. A., Ronksley, P. E., Sigal, R. J., Campbell, T. S., & Hemmelgarn, B. R. (2011). Motivational interviewing to improve weight loss in overweight and/or obese patients: A systematic review and meta-analysis of randomized controlled trials. *Obesity Reviews*, 12, 709–723. Retrieved from <https://doi.org/10.1111/j.1467-789X.2011.00892.x> doi: 10.1111/j.1467-789X.2011.00892.x
- Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Comput. Linguistics*, 34(4), 555–596. Retrieved from <https://doi.org/10.1162/coli.07-034-R2> doi: 10.1162/COLI.07-034-R2
- Aswamenakul, C., Liu, L., Carey, K. B., Woolley, J., Scherer, S., & Borsari, B. (2018). Multimodal analysis of client behavioral change coding in motivational interviewing. In S. K. D'Mello, P. G. Georgiou, S. Scherer, E. M. Provost, M. Soleymani, & M. Worsley (Eds.), *Proceedings of the 2018 on international conference on multimodal interaction, ICMI 2018, boulder, co, usa, october 16-20, 2018* (pp. 356–360). ACM. Retrieved from <https://doi.org/10.1145/3242969.3242990> doi: 10.1145/3242969.3242990
- Atkins, D. C., Steyvers, M., Imel, Z. E., & Smyth, P. (2014). Scaling up the evaluation of psychotherapy: Evaluating motivational interviewing fidelity via statistical text classification. *Implementation Sci*, 9(49). Retrieved from <https://doi.org/10.1186/1748-5908-9-49> doi: 10.1186/1748-5908-9-49
- Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., ... Smith, D. M. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med*, 183(6), 589–596. Retrieved from <https://doi.org/10.1001/jamainternmed.2023.1838> doi: 10.1001/jamainternmed.2023.1838
- Baldwin, T., Cook, P., Lui, M., MacKinlay, A., & Wang, L. (2013). How noisy social media text, how diffrent social media sources? In R. Mitkov & J. C. Park (Eds.), *Sixth international joint conference on natural language processing, IJCNLP 2013, nagoya, japan, october 14-18, 2013* (pp. 356–364). Asian Federation of Natural Language Processing / ACL. Retrieved from <https://aclanthology.org/I13-1041/>
- Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84(2), 191–215. Retrieved from <https://doi.org/10.1037/0033-295X.84.2.191> doi: 10.1037/0033-295X.84.2.191
- Bandura, A. (1986). Social foundations of thought and action. In D. F. Marks (Ed.), *The health psychology reader* (pp. 94–107). SAGE.
- Bandura, A., & Cervone, D. (1983). Self-evaluative and self-efficacy mechanisms governing the motivational effects of goal systems. *Journal of Personality and Social Psychology*, 45(5), 1017–1028. Retrieved from <https://doi.org/10.1037/0022-3514.45.5.1017> doi: 10.1037/0022-3514.45.5.1017

- Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., ... Fung, P. (2023). A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In J. C. Park et al. (Eds.), *Proceedings of the 13th international joint conference on natural language processing and the 3rd conference of the asia-pacific chapter of the association for computational linguistics, IJCNLP 2023 -volume 1: Long papers, nusa dua, bali, november 1 - 4, 2023* (pp. 675–718). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/2023.ijcnlp-main.45> doi: 10.18653/V1/2023.IJCNLP-MAIN.45
- Barrett, S., Begg, S., O'Halloran, P., & Kingsley, M. (2018). Integrated motivational interviewing and cognitive behaviour therapy for lifestyle mediators of overweight and obesity in community-dwelling adults: A systematic review and meta-analyses. *BMC Public Health*, 18(1160). Retrieved from <https://doi.org/10.1186/s12889-018-6062-9> doi: 10.1186/s12889-018-6062-9
- Basar, E., Hendrickx, I., Krahmer, E., de Bruijn, G., & Bosse, T. (2022). Hints of independence in a pre-scripted world: On controlled usage of open-domain language models for chatbots in highly sensitive domains. In A. P. Rocha, L. Steels, & H. J. van den Herik (Eds.), *Proceedings of the 14th international conference on agents and artificial intelligence, ICAART 2022, volume 1, online streaming, february 3-5, 2022* (pp. 401–407). SCITEPRESS. Retrieved from <https://doi.org/10.5220/0010914300003116> doi: 10.5220/0010914300003116
- Bauer, R., Bauer, M., Spiessl, H., & Kagerbauer, T. (2012). Cyber-support: An analysis of online self-help forums (online self-help forums in bipolar disorder). *Nordic Journal of Psychiatry*, 67(3), 185–190. Retrieved from <https://doi.org/10.3109/08039488.2012.700734> doi: 10.3109/08039488.2012.700734
- Beckwith, V. Z., & Beckwith, J. (2020). Motivational interviewing: A communication tool to promote positive behavior change and optimal health outcomes. *NASN School Nurse*, 35(6), 344–351. Retrieved from <https://doi.org/10.1177/1942602X20915715> doi: 10.1177/1942602X20915715
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In M. C. Elish, W. Isaac, & R. S. Zemel (Eds.), *Facct '21: 2021 ACM conference on fairness, accountability, and transparency, virtual event / toronto, canada, march 3-10, 2021* (pp. 610–623). ACM. Retrieved from <https://doi.org/10.1145/3442188.3445922> doi: 10.1145/3442188.3445922
- Bentley, F., Tollmar, K., Stephenson, P., Levy, L. M., Jones, B. D., Robertson, S. L., ... Wilson, J. (2013). Health mashups: Presenting statistical patterns between wellbeing data and context in natural language to promote behavior change. *ACM Trans. Comput. Hum. Interact.*, 20(5), 30:1–30:27. Retrieved from <https://doi.org/10.1145/2503823> doi: 10.1145/2503823
- Bickmore, T. W., Schulman, D., & Sidner, C. L. (2011). A reusable framework for health counseling dialogue systems based on a behavioral medicine ontology. *J. Biomed. Informatics*, 44(2), 183–197. Retrieved from <https://doi.org/10.1016/j.jbi.2010.12.006> doi: 10.1016/J.JBI.2010.12.006
- Biener, L., & Abrams, D. B. (1991). The contemplation ladder: Validation of a measure of readiness to consider smoking cessation. *Health psychology : Official journal of the Division of Health Psychology, American Psychological Association*, 10(5), 360–365. Retrieved from <https://doi.org/10.1037//0278-6133.10.5.360> doi: 10.1037//0278-6133.10.5.360
- Bink, M., Zimmerman, S., & Elweiler, D. (2022). Featured snippets and their influence on users'

- credibility judgements. In D. Elsweler (Ed.), *CHIIR '22: ACM SIGIR conference on human information interaction and retrieval, regensburg, germany, march 14 - 18, 2022* (pp. 113–122). ACM. Retrieved from <https://doi.org/10.1145/3498366.3505766> doi: 10.1145/3498366.3505766
- Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., . . . Weinbach, S. (2022). GPT-NeoX-20B: An open-source autoregressive language model. *CoRR, abs/2204.06745*. Retrieved from <https://doi.org/10.48550/arXiv.2204.06745> doi: 10.48550/ARXIV.2204.06745
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R. B., Arora, S., von Arx, S., . . . et al. (2021). On the opportunities and risks of foundation models. *CoRR, abs/2108.07258*. Retrieved from <https://arxiv.org/abs/2108.07258>
- Borsari, B., Apodaca, T. R., Jackson, K., Mastroleo, N., Magill, M., Barnett, N. P., & Carey, K. B. (2014). In-session processes of brief motivational interventions in two trials with mandated college students. *Journal of consulting and clinical psychology, 83*(1), 56–67. Retrieved from <https://doi.org/10.1037/a0037635> doi: 10.1037/a0037635
- Boustani, M., Lunn, S., Visser, U., & Lisetti, C. (2021). Development, feasibility, acceptability, and utility of an expressive speech-enabled digital health agent to deliver online, brief motivational interviewing for alcohol misuse: Descriptive study. *J Med Internet Res, 23*(9), e25837. Retrieved from <https://doi.org/10.2196/25837> doi: 10.2196/25837
- Britt, E., Hudson, S. M., & Blampied, N. M. (2004). Motivational interviewing in health settings: A review. *Patient Education and Counseling, 53*(2), 147–155. Retrieved from [https://doi.org/10.1016/S0738-3991\(03\)00141-1](https://doi.org/10.1016/S0738-3991(03)00141-1) doi: 10.1016/S0738-3991(03)00141-1
- Brown, A., Kumar, A. T., Melamed, O., Ahmed, I., Wang, Y. H., Deza, A., . . . Rose, J. (2023). A motivational interviewing chatbot with generative reflections for increasing readiness to quit smoking: Iterative development study. *JMIR Ment Health, 10*, e49132. Retrieved from <https://mental.jmir.org/2023/1/e49132> doi: 10.2196/49132
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., . . . Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems 33: Annual conference on neural information processing systems 2020, neurips 2020, december 6-12, 2020, virtual*. Retrieved from <https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html>
- Buchanan, S., & Nicol, E. (2019a). Developing health information literacy in disadvantaged and dependent circumstances: The everyday role of family nurses. In S. Kurbanoglu et al. (Eds.), *Information literacy in everyday life* (pp. 124–135). Cham: Springer International Publishing.
- Buchanan, S., & Nicol, E. (2019b). Developing health information literacy in disengaged at-risk populations: Insights to inform interventions. *Journal of Documentation, 75*(1), 172–189. Retrieved from <https://doi.org/10.1108/JD-06-2018-0086> doi: 10.1108/JD-06-2018-0086
- Buechel, S., Buffone, A., Slaff, B., Ungar, L. H., & Sedoc, J. (2018). Modeling empathy and distress in reaction to news stories. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 conference on empirical methods in natural language processing, brussels, belgium, october 31 - november 4, 2018* (pp. 4758–4765). Association for Computational Linguistics. Retrieved from <https://aclanthology.org/D18-1507/>
- Byerly, H., Balmford, A., Ferraro, P. J., Wagner, C. H., Palchak, E., Polasky, S., . . . Fisher, B. (2018).

- Nudging pro-environmental behavior: Evidence and opportunities. *Frontiers in Ecology and the Environment*, 16(3), 159–168. Retrieved from <https://doi.org/10.1002/fee.1777> doi: 10.1002/fee.1777
- Calvo, R. A., Milne, D. N., Hussain, M. S., & Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. *Nat. Lang. Eng.*, 23(5), 649–685. Retrieved from <https://doi.org/10.1017/S1351324916000383> doi: 10.1017/S1351324916000383
- Can, D., Atkins, D. C., & Narayanan, S. S. (2015). A dialog act tagging approach to behavioral coding: A case study of addiction counseling conversations. In *16th annual conference of the international speech communication association, INTERSPEECH 2015, dresden, germany, september 6-10, 2015* (pp. 339–343). ISCA. Retrieved from <https://doi.org/10.21437/Interspeech.2015-151> doi: 10.21437/INTERSPEECH.2015-151
- Can, D., Marín, R. A., Georgiou, P. G., Imel, Z. E., Atkins, D. C., & Narayanan, S. S. (2016). "it sounds like...": A natural language processing approach to detecting counselor reflections in motivational interviewing. *Journal of counseling psychology*, 63(3), 343–350. Retrieved from <https://doi.org/10.1037/cou0000111> doi: 10.1037/cou0000111
- Cao, J., Tanana, M., Imel, Z. E., Poitras, E., Atkins, D. C., & Srikumar, V. (2019). Observing dialogue in therapy: Categorizing and forecasting behavioral codes. In A. Korhonen, D. R. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th conference of the association for computational linguistics, ACL 2019, florence, italy, july 28- august 2, 2019, volume 1: Long papers* (pp. 5599–5611). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/p19-1563> doi: 10.18653/V1/P19-1563
- Carah, N., Meurk, C., & Angus, D. (2017). Online self-expression and experimentation as 'reflectivism': Using text analytics to examine the participatory forum Hello Sunday Morning. *Health*, 21(2), 119–135. Retrieved from <https://doi.org/10.1177/1363459315596799> doi: 10.1177/1363459315596799
- Carels, R. A., Darby, L., Cacciapaglia, H. M., Konrad, K., Coit, C., Harper, J., ... Versland, A. (2007). Using motivational interviewing as a supplement to obesity treatment: A stepped-care approach. *Health Psychology*, 26(3), 369–374. Retrieved from <https://doi.org/10.1037/0278-6133.26.3.369> doi: 10.1037/0278-6133.26.3.369
- Chan, B., Schweter, S., & Möller, T. (2020). German's next language model. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th international conference on computational linguistics, COLING 2020, barcelona, spain (online), december 8-13, 2020* (pp. 6788–6796). International Committee on Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/2020.coling-main.598> doi: 10.18653/V1/2020.COLING-MAIN.598
- Chawla, K., Shi, W., Zhang, J., Lucas, G. M., Yu, Z., & Gratch, J. (2023). Social influence dialogue systems: A survey of datasets and models for social influence tasks. In A. Vlachos & I. Augenstein (Eds.), *Proceedings of the 17th conference of the european chapter of the association for computational linguistics, EACL 2023, dubrovnik, croatia, may 2-6, 2023* (pp. 750–766). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/2023.eacl-main.53> doi: 10.18653/V1/2023.EACL-MAIN.53
- Chen, J., Chen, C., Walther, J. B., & Sundar, S. S. (2021). Do you feel special when an AI doctor remembers you? Individuation effects of AI vs. Human doctors on user experience. In Y. Kitamura, A. Quigley, K. Isbister, & T. Igarashi (Eds.), *CHI '21: CHI conference on human factors in*

- computing systems, virtual event / yokohama japan, may 8-13, 2021, extended abstracts* (pp. 299:1–299:7). ACM. Retrieved from <https://doi.org/10.1145/3411763.3451735> doi: 10.1145/3411763.3451735
- Chen, M. Y., Li, S., & Yang, Y. (2022). EmpHi: Generating empathetic responses with human-like intents. In M. Carpuat, M. de Marneffe, & I. V. M. Ruíz (Eds.), *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies, NAACL 2022, seattle, wa, united states, july 10-15, 2022* (pp. 1063–1074). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/2022.naacl-main.78> doi: 10.18653/V1/2022.NAACL-MAIN.78
- Chen, Q., & Liu, D. (2023). Dynamic strategy chain: Dynamic zero-shot CoT for long mental health support generation. *CoRR*, *abs/2308.10444*. Retrieved from <https://doi.org/10.48550/arXiv.2308.10444> doi: 10.48550/ARXIV.2308.10444
- Chen, Y., & Liang, C. (2022). Wish I can feel what you feel: A neural approach for empathetic response generation. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Findings of the association for computational linguistics: EMNLP 2022, abu dhabi, united arab emirates, december 7-11, 2022* (pp. 922–933). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/2022.findings-emnlp.65> doi: 10.18653/V1/2022.FINDINGS-EMNLP.65
- Chisholm, A., Hart, J., Lam, V., & Peters, S. (2012). Current challenges of behavior change talk for medical professionals and trainees. *Patient Education and Counseling*, *87*(3), 389–394. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0738399111006008> doi: 10.1016/j.pec.2011.12.001
- Chiu, Y. Y., Sharma, A., Lin, I. W., & Althoff, T. (2024). A computational framework for behavioral assessment of LLM therapists. *CoRR*, *abs/2401.00820*. Retrieved from <https://doi.org/10.48550/arXiv.2401.00820> doi: 10.48550/ARXIV.2401.00820
- Cho, Y., Kim, M., Kim, S., Kwon, O., Kwon, R. D., Lee, Y., & Lim, D. (2023). Evaluating the efficacy of interactive language therapy based on LLM for high-functioning autistic adolescent psychological counseling. *CoRR*, *abs/2311.09243*. Retrieved from <https://doi.org/10.48550/arXiv.2311.09243> doi: 10.48550/ARXIV.2311.09243
- Clifford, D., & Curtis, L. (2016). *Motivational interviewing in nutrition and fitness*. Guilford Publications. Retrieved from https://books.google.de/books?id=j_R0CwAAQBAJ
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. Retrieved from <https://doi.org/10.1177/001316446002000104> doi: 10.1177/001316446002000104
- Coolidge, T., Skaret, E., Heima, M., Johnson, E. K., Hillstead, M. B., Farjo, N., ... Weinstein, P. (2011a). Thinking about going to the dentist: A contemplation ladder to assess dentally-avoidant individuals' readiness to go to a dentist. *BMC Oral Health*, *11*(4). Retrieved from <https://doi.org/10.1186/1472-6831-11-4> doi: 10.1186/1472-6831-11-4
- Coolidge, T., Skaret, E., Heima, M., Johnson, E. K., Hillstead, M. B., Farjo, N., ... Weinstein, P. (2011b). Thinking about going to the dentist: A contemplation ladder to assess dentally-avoidant individuals' readiness to go to a dentist. *BMC Oral Health*, *11*(1), 1–12. Retrieved from <https://doi.org/10.1186/1472-6831-11-4> doi: 10.1186/1472-6831-11-4
- Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M. (2015). CLPsych 2015 shared task: Depression and PTSD on twitter. In M. Mitchell, G. Coppersmith, & K. Hollingshead

- (Eds.), *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality, clpsych@naacl-hlt 2015, june 5, 2015, denver, colorado, USA* (pp. 31–39). The Association for Computational Linguistics. Retrieved from <https://doi.org/10.3115/v1/w15-1204> doi: 10.3115/V1/W15-1204
- Corda, E., Massa, S. M., & Riboni, D. (2024). Context-aware behavioral tips to improve sleep quality via machine learning and large language models. *Future Internet*, 16(2), 46. Retrieved from <https://doi.org/10.3390/fi16020046> doi: 10.3390/FI16020046
- Curry, A., & Curry, A. C. (2023). Computer says "no": The case against empathetic conversational AI. In A. Rogers, J. L. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: ACL 2023, toronto, canada, july 9-14, 2023* (pp. 8123–8130). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/2023.findings-acl.515> doi: 10.18653/V1/2023.FINDINGS-ACL.515
- Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993). Wizard of oz studies: Why and how. In W. D. Gray, W. E. Hefley, & D. Murray (Eds.), *Proceedings of the 1st international workshop on intelligent user interfaces, IUI 1993, orlando, florida, usa, january 4-7, 1993* (pp. 193–200). ACM. Retrieved from <https://doi.org/10.1145/169891.169968> doi: 10.1145/169891.169968
- Dale, R. (2021). GPT-3: what's it good for? *Nat. Lang. Eng.*, 27(1), 113–118. Retrieved from <https://doi.org/10.1017/S1351324920000601> doi: 10.1017/S1351324920000601
- Das, A., Selek, S., Warner, A. R., Zuo, X., Hu, Y., Keloth, V. K., ... Xu, H. (2022). Conversational bots for psychotherapy: A study of generative transformer models using domain-specific dialogues. In D. Demner-Fushman, K. B. Cohen, S. Ananiadou, & J. Tsujii (Eds.), *Proceedings of the 21st workshop on biomedical language processing, bionlp@acl 2022, dublin, ireland, may 26, 2022* (pp. 285–297). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/2022.bionlp-1.27> doi: 10.18653/V1/2022.BIONLP-1.27
- da Silva, J. G. G., Kavanagh, D. J., Belpaeme, T., Taylor, L., Beeson, K., & Andrade, J. (2018). Experiences of a motivational interview delivered by a robot: Qualitative study. *J Med Internet Res*, 20(5), e116. Retrieved from <https://doi.org/10.2196/jmir.7737> doi: 10.2196/jmir.7737
- Davis, M. F., Shapiro, D., Windsor, R., Whalen, P., Rhode, R., Miller, H. S., & Sechrest, L. (2011). Motivational interviewing versus prescriptive advice for smokers who are not ready to quit. *Patient Education and Counseling*, 83(1), 129–133. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0738399110002715> doi: <https://doi.org/10.1016/j.pec.2010.04.024>
- Davis, R., Campbell, R., Hildon, Z., Hobbs, L., & Michie, S. (2015). Theories of behaviour and behaviour change across the social and behavioural sciences: A scoping review. *Health Psychology Review*, 9(3), 323–344. Retrieved from <https://www.tandfonline.com/doi/full/10.1080/17437199.2014.941722> doi: 10.1080/17437199.2014.941722
- Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, NAACL-HLT 2019, minneapolis, mn, usa, june 2-7, 2019, volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/n19-1423> doi: 10.18653/V1/N19-1423
- DiClemente, C. C., & Velasquez, M. M. (2002). Motivational interviewing and the stages of change.

- In *Motivational interviewing: Preparing people for change* (pp. 201–216). Guilford Press.
- Dransart, D. A. C. (2013). From sense-making to meaning-making: Understanding and supporting survivors of suicide. *The British Journal of Social Work*, 43(2), 317–335. Retrieved from <https://doi.org/10.1093/bjsw/bct026> doi: 10.1093/bjsw/bct026
- Duarte, J. M., & Berton, L. (2023). A review of semi-supervised learning for text classification. *Artif. Intell. Rev.*, 56(9), 9401–9469. Retrieved from <https://doi.org/10.1007/s10462-023-10393-8> doi: 10.1007/S10462-023-10393-8
- Dzewaltowski, D. A. (1989). Toward a model of exercise motivation. *Journal of Sport and Exercise Psychology*, 11(3), 251–269. Retrieved from <https://doi.org/10.1123/jsep.11.3.251> doi: 10.1123/jsep.11.3.251
- Eklund, J. H., & Meranius, M. S. (2021). Toward a consensus on the nature of empathy: A review of reviews. *Patient Education and Counseling*, 104(2), 300–307. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0738399120304493> doi: 10.1016/j.pec.2020.08.022
- Elfhag, K., & Rössner, S. (2005). Who succeeds in maintaining weight loss? A conceptual review of factors associated with weight loss maintenance and weight regain. *Obesity Reviews*, 6, 67–85. Retrieved from <https://doi.org/10.1111/j.1467-789X.2005.00170.x> doi: 10.1111/j.1467-789X.2005.00170.x
- Elsweiler, D., Trattner, C., & Harvey, M. (2017). Exploiting food choice biases for healthier recipe recommendation. In N. Kando, T. Sakai, H. Joho, H. Li, A. P. de Vries, & R. W. White (Eds.), *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval, shinjuku, tokyo, japan, august 7-11, 2017* (pp. 575–584). ACM. Retrieved from <https://doi.org/10.1145/3077136.3080826> doi: 10.1145/3077136.3080826
- Ewbank, M., Cummins, R., Tablan, V., Buchholz, A. C. S., & Blackwell, A. D. (2020). Understanding the relationship between patient language and outcomes in internet-enabled cognitive behavioural therapy: A deep learning approach to automatic coding of session transcripts. *Psychotherapy Research*, 31(3), 300–312. Retrieved from <https://doi.org/10.1080/10503307.2020.1788740> doi: 10.1080/10503307.2020.1788740
- Farhud, D. D. (2015). Impact of lifestyle on health. *Iranian journal of public health*, 44(11), 1442–1444. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4703222/>
- Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., & Hovy, E. H. (2021). A survey of data augmentation approaches for NLP. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Findings of the association for computational linguistics: ACL/IJCNLP 2021, online event, august 1-6, 2021* (Vol. ACL/IJCNLP 2021, pp. 968–988). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/2021.findings-acl.84> doi: 10.18653/V1/2021.FINDINGS-ACL.84
- Fernández, M., Piccolo, L. S. G., Maynard, D., Wippoo, M., Meili, C., & Alani, H. (2016). Talking climate change via social media: Communication, engagement and behaviour. In W. Nejdl, W. Hall, P. Parigi, & S. Staab (Eds.), *Proceedings of the 8th ACM conference on web science, websci 2016, hannover, germany, may 22-25, 2016* (pp. 85–94). ACM. Retrieved from <https://doi.org/10.1145/2908131.2908167> doi: 10.1145/2908131.2908167
- Ferrari, J. R., O'Callaghan, J., & Newbegin, I. (2005). Prevalence of procrastination in the united states, united kingdom, and australia: Arousal and avoidance delays among adults. *North American*

- Journal of Psychology*, 7(1), 1–6. Retrieved from <https://psycnet.apa.org/record/2005-03779-001>
- Fisher, J. D., & Fisher, W. A. (1992). Changing AIDS-risk behavior. *Psychological Bulletin*, 111(3), 455–474. Retrieved from <https://doi.org/10.1037/0033-2909.111.3.455> doi: 10.1037/0033-2909.111.3.455
- Fisher, K. E., Erdelez, S., & McKechnie, L. (2005). *Theories of information behavior*. American Society for Information Science and Technology. Retrieved from <https://books.google.de/books?id=1l6qzqhIj8wC>
- Flemotomos, N., Martinez, V. R., Chen, Z., Singla, K., Ardulov, V., Peri, R., ... Narayanan, S. (2022). Automated evaluation of psychotherapy skills using speech and language technologies. *Behav Res*, 54, 690–711. Retrieved from <https://doi.org/10.3758/s13428-021-01623-4> doi: 10.3758/s13428-021-01623-4
- Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds Mach.*, 30(4), 681–694. Retrieved from <https://doi.org/10.1007/s11023-020-09548-1> doi: 10.1007/S11023-020-09548-1
- Fogg, B. J. (2009a). The behavior grid: 35 ways behavior can change. In S. Chatterjee & P. Dev (Eds.), *Persuasive technology, fourth international conference, PERSUASIVE 2009, claremont, california, usa, april 26-29, 2009. proceedings* (Vol. 350, p. 42). ACM. Retrieved from <https://doi.org/10.1145/1541948.1542001> doi: 10.1145/1541948.1542001
- Fogg, B. J. (2009b). A behavior model for persuasive design. In S. Chatterjee & P. Dev (Eds.), *Persuasive technology, fourth international conference, PERSUASIVE 2009, claremont, california, usa, april 26-29, 2009. proceedings* (Vol. 350, p. 40). ACM. Retrieved from <https://doi.org/10.1145/1541948.1541999> doi: 10.1145/1541948.1541999
- Friederichs, S. A., Oenema, A., Bolman, C., Guyaux, J., Keulen, H. M. V., & Lechner, L. (2015). Motivational interviewing in a web-based physical activity intervention: Questions and reflections. *Health Promotion International*, 30(3), 803–815. Retrieved from <https://doi.org/10.1093/heapro/dat069> doi: 10.1093/heapro/dat069
- Fu, G., Zhao, Q., Li, J., Luo, D., Song, C., Zhai, W., ... Yang, B. X. (2023). Enhancing psychological counseling with large language model: A multifaceted decision-support system for non-professionals. *CoRR*, abs/2308.15192. Retrieved from <https://doi.org/10.48550/arXiv.2308.15192> doi: 10.48550/ARXIV.2308.15192
- Gaffney, H., Mansell, W., & Tai, S. (2019). Conversational agents in the treatment of mental health problems: Mixed-method systematic review. *JMIR Ment Health*, 6(10), e14166. Retrieved from <https://doi.org/10.2196/14166> doi: 10.2196/14166
- Gao, J., Liu, Y., Deng, H., Wang, W., Cao, Y., Du, J., & Xu, R. (2021). Improving empathetic response generation by recognizing emotion cause in conversations. In M. Moens, X. Huang, L. Specia, & S. W. Yih (Eds.), *Findings of the association for computational linguistics: EMNLP 2021, virtual event / punta cana, dominican republic, 16-20 november, 2021* (pp. 807–819). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/2021.findings-emnlp.70> doi: 10.18653/V1/2021.FINDINGS-EMNLP.70
- Gibson, J., Atkins, D. C., Creed, T. A., Imel, Z. E., Georgiou, P. G., & Narayanan, S. (2022). Multi-label multi-task deep learning for behavioral coding. *IEEE Trans. Affect. Comput.*, 13(1), 508–518. Retrieved from <https://doi.org/10.1109/TAFFC.2019.2952113> doi: 10.1109/

TAFFC.2019.2952113

- Gibson, J., Can, D., Xiao, B., Imel, Z. E., Atkins, D. C., Georgiou, P. G., & Narayanan, S. S. (2016). A deep learning approach to modeling empathy in addiction counseling. In N. Morgan (Ed.), *17th annual conference of the international speech communication association, INTERSPEECH 2016, san francisco, ca, usa, september 8-12, 2016* (pp. 1447–1451). ISCA. Retrieved from <https://doi.org/10.21437/Interspeech.2016-554> doi: 10.21437/INTERSPEECH.2016-554
- Gibson, J., Malandrakis, N., Romero, F., Atkins, D. C., & Narayanan, S. S. (2015). Predicting therapist empathy in motivational interviews using language features inspired by psycholinguistic norms. In *16th annual conference of the international speech communication association, INTERSPEECH 2015, dresden, germany, september 6-10, 2015* (pp. 1947–1951). ISCA. Retrieved from <https://doi.org/10.21437/Interspeech.2015-429> doi: 10.21437/INTERSPEECH.2015-429
- Giles, D., Stommel, W., Paulus, T., Lester, J., & Reed, D. (2015). Microanalysis of online data: The methodological development of “Digital CA”. *Discourse, Context & Media*, 7, 45–51. Retrieved from <https://doi.org/10.1016/j.dcm.2014.12.002> doi: 10.1016/j.dcm.2014.12.002
- Grilli, G., & Curtis, J. (2021). Encouraging pro-environmental behaviours: A review of methods and approaches. *Renewable and Sustainable Energy Reviews*, 135, 110039. Retrieved from <https://doi.org/10.1016/j.rser.2020.110039> doi: 10.1016/j.rser.2020.110039
- Guhr, O., Schumann, A., Bahrmann, F., & Böhme, H. (2020). Training a broad-coverage german sentiment classification model for dialog systems. In N. Calzolari et al. (Eds.), *Proceedings of the 12th language resources and evaluation conference, LREC 2020, marseille, france, may 11-16, 2020* (pp. 1627–1632). European Language Resources Association. Retrieved from <https://aclanthology.org/2020.lrec-1.202/>
- Guntakandla, N., & Nielsen, R. (2018). Annotating reflections for health behavior change therapy. In N. Calzolari et al. (Eds.), *Proceedings of the eleventh international conference on language resources and evaluation, LREC 2018, miyazaki, japan, may 7-12, 2018*. European Language Resources Association (ELRA). Retrieved from <http://www.lrec-conf.org/proceedings/lrec2018/summaries/986.html>
- Guo, L., & Li, Y. (2022). Class-imbalanced semi-supervised learning with adaptive thresholding. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, & S. Sabato (Eds.), *International conference on machine learning, ICML 2022, 17-23 july 2022, baltimore, maryland, USA* (Vol. 162, pp. 8082–8094). PMLR. Retrieved from <https://proceedings.mlr.press/v162/guo22e.html>
- Gupta, I., Eugenio, B. D., Ziebart, B. D., Baiju, A., Liu, B., Gerber, B. S., ... Smart, M. (2020). Human-human health coaching via text messages: Corpus, annotation, and analysis. In O. Pietquin et al. (Eds.), *Proceedings of the 21th annual meeting of the special interest group on discourse and dialogue, sigdial 2020, 1st virtual meeting, july 1-3, 2020* (pp. 246–256). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/2020.sigdial-1.30> doi: 10.18653/V1/2020.SIGDIAL-1.30
- Gupta, I., Eugenio, B. D., Ziebart, B. D., Liu, B., Gerber, B. S., & Sharp, L. K. (2021). Summarizing behavioral change goals from SMS exchanges to support health coaches. In H. Li et al. (Eds.), *Proceedings of the 22nd annual meeting of the special interest group on discourse and dialogue, sigdial 2021, singapore and online, july 29-31, 2021* (pp. 276–289). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/2021.sigdial-1.31> doi: 10

- .18653/V1/2021.SIGDIAL-1.31
- Hall, K., Gibbie, T., & Lubman, D. I. (2012). Motivational interviewing techniques: Facilitating behaviour change in the general practice setting. *Australian Family Physician*, 41(9), 660–667. Retrieved from <https://search.informit.org/doi/10.3316/informit.737035419857450> doi: 10.3316/informit.737035419857450
- Hardcastle, S. J., Taylor, A. H., Bailey, M. P., Harley, R. A., & Hagger, M. S. (2013). Effectiveness of a motivational interviewing intervention on weight loss, physical activity and cardiovascular disease risk factors: A randomised controlled trial with a 12-month post-intervention follow-up. *Int J Behav Nutr Phys Act*, 10, 40. Retrieved from <https://doi.org/10.1186/1479-5868-10-40> doi: 10.1186/1479-5868-10-40
- Harris, E., & Bardey, A. C. (2019). Do instagram profiles accurately portray personality? An investigation into idealized online self-presentation. *Frontiers in Psychology*, 10. Retrieved from <https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2019.00871> doi: 10.3389/fpsyg.2019.00871
- Hasan, M., Carcone, A. I., Naar, S., Eggly, S., Alexander, G. L., Hartlieb, K. B., & Kotov, A. (2019). Identifying effective motivational interviewing communication sequences using automated pattern analysis. *J. Heal. Informatics Res.*, 3(1), 86–106. Retrieved from <https://doi.org/10.1007/s41666-018-0037-6> doi: 10.1007/S41666-018-0037-6
- Hasler, L., Ruthven, I., & Buchanan, S. (2014). Using internet groups in situations of information poverty: Topics and information needs. *J. Assoc. Inf. Sci. Technol.*, 65(1), 25–36. Retrieved from <https://doi.org/10.1002/asi.22962> doi: 10.1002/ASI.22962
- He, L., Basar, E., Wiers, R. W., Antheunis, M. L., & Krahmer, E. (2022). Can chatbots help to motivate smoking cessation? A study on the effectiveness of motivational interviewing on engagement and therapeutic alliance. *BMC Public Health*, 22, 726. Retrieved from <https://doi.org/10.1186/s12889-022-13115-x> doi: 10.1186/s12889-022-13115-x
- Heckhausen, H., & Gollwitzer, P. M. (1987). Thought contents and cognitive functioning in motivational versus volitional states of mind. *Motiv Emot*, 11, 101–120. Retrieved from <https://doi.org/10.1007/BF00992338> doi: 10.1007/BF00992338
- Hensley, L. C., & Munn, K. J. (2020). The power of writing about procrastination: Journaling as a tool for change. *Journal of Further and Higher Education*, 44(10), 1450–1465. Retrieved from <https://doi.org/10.1080/0309877X.2019.1702154> doi: 10.1080/0309877X.2019.1702154
- Hershberger, P. J., Pei, Y., Bricker, D. A., Crawford, T. N., Shivakumar, A., Vasoya, M., ... Wilson, J. F. (2021). Advancing motivational interviewing training with artificial intelligence: ReadMI. *Advances in Medical Education and Practice*, 12, 613–618. Retrieved from <https://doi.org/10.2147/AMEP.S312373> doi: 10.2147/AMEP.S312373
- Hill, J., Ford, W. R., & Farreras, I. G. (2015). Real conversations with artificial intelligence: A comparison between human-human online conversations and human-chatbot conversations. *Comput. Hum. Behav.*, 49, 245–250. Retrieved from <https://doi.org/10.1016/j.chb.2015.02.026> doi: 10.1016/J.CHB.2015.02.026
- Hirvonen, N., Huotari, M., Niemelä, R., & Korpelainen, R. (2012). Information behavior in stages of exercise behavior change. *J. Assoc. Inf. Sci. Technol.*, 63(9), 1804–1819. Retrieved from <https://doi.org/10.1002/asi.22704> doi: 10.1002/ASI.22704

- Ho, A., Hancock, J., & Miner, A. S. (2018). Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot. *Journal of Communication*, 68(4), 712–733. Retrieved from <https://doi.org/10.1093/joc/jqy026> doi: 10.1093/joc/jqy026
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9(8), 1735–1780. Retrieved from <https://doi.org/10.1162/neco.1997.9.8.1735> doi: 10.1162/NECO.1997.9.8.1735
- Hohman, M., Doran, N., & Koutsenok, I. (2009). Motivational interviewing training for juvenile correctional staff in california: One year initial outcomes. *Journal of Offender Rehabilitation*, 48(7), 635–648. Retrieved from <https://doi.org/10.1080/10509670903196108> doi: 10.1080/10509670903196108
- Holdener, M., Gut, A., & Angerer, A. (2020). Applicability of the user engagement scale to mobile health: A survey-based quantitative study. *JMIR Mhealth Uhealth*, 8(1), e13244. Retrieved from <https://10.2196/13244> doi: 10.2196/13244
- Holme, S., Moorhead, A., Bond, R., Zheng, H., Coates, V., & McTear, M. (2018). Weightmentor: a new automated chatbot for weight loss maintenance. In *Proceedings of the 32nd international bcs human computer interaction conference 32* (pp. 1–5). Retrieved from <https://doi.org/10.14236/ewic/HCI2018.103> doi: 10.14236/ewic/HCI2018.103
- Hosseini, M., & Caragea, C. (2021). Distilling knowledge for empathy detection. In M. Moens, X. Huang, L. Specia, & S. W. Yih (Eds.), *Findings of the association for computational linguistics: EMNLP 2021, virtual event / punta cana, dominican republic, 16-20 november, 2021* (pp. 3713–3724). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/2021.findings-emnlp.314> doi: 10.18653/V1/2021.FINDINGS-EMNLP.314
- Hsu, S., Shah, R. S., Senthil, P., Ashktorab, Z., Dugan, C., Geyer, W., & Yang, D. (2023). *Helping the helper: Supporting peer counselors via AI-empowered practice and feedback* (Vol. abs/2305.08982). Retrieved from <https://doi.org/10.48550/arXiv.2305.08982> doi: 10.48550/ARXIV.2305.08982
- Htait, A., Azzopardi, L., Nicol, E., & Moncur, W. (2020). Datamirror: Reflecting on one's data self: A tool for social media users to explore their digital footprints. In *Proceedings of the 43rd international acm sigir conference on research and development in information retrieval* (p. 2125–2128). New York, NY, USA: Association for Computing Machinery. Retrieved from <https://doi.org/10.1145/3397271.3401398> doi: 10.1145/3397271.3401398
- Hu, B., Lu, Z., Li, H., & Chen, Q. (2014). Convolutional neural network architectures for matching natural language sentences. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 27: Annual conference on neural information processing systems 2014, december 8-13 2014, montreal, quebec, canada* (pp. 2042–2050). Retrieved from <https://proceedings.neurips.cc/paper/2014/hash/b9d487a30398d42ecff55c228ed5652b-Abstract.html>
- Hu, Y., Manikonda, L., & Kambhampati, S. (2014). What we instagram: A first analysis of instagram photo content and user types. In E. Adar, P. Resnick, M. D. Choudhury, B. Hogan, & A. Oh (Eds.), *Proceedings of the eighth international conference on weblogs and social media, ICWSM 2014, ann arbor, michigan, usa, june 1-4, 2014*. The AAAI Press. Retrieved from <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8118>
- Hua, Y., Liu, F., Yang, K., Li, Z., Sheu, Y., Zhou, P., ... Beam, A. (2024). Large language models

- in mental health care: a scoping review. *CoRR*, *abs/2401.02984*. Retrieved from <https://doi.org/10.48550/arXiv.2401.02984> doi: 10.48550/ARXIV.2401.02984
- Huang, X., Liu, L., Carey, K. B., Woolley, J., Scherer, S., & Borsari, B. (2018). Modeling temporality of human intentions by domain adaptation. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 conference on empirical methods in natural language processing, brussels, belgium, october 31 - november 4, 2018* (pp. 696–701). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/d18-1074> doi: 10.18653/V1/D18-1074
- Hulsman, R., Harmsen, A., & Fabriek, M. (2009). Reflective teaching of medical communication skills with DiViDu: Assessing the level of student reflection on recorded consultations with simulated patients. *Patient Education and Counseling*, *74*(2), 142–149. Retrieved from <https://www.sciencedirect.com/science/article/pii/S073839910800548X> doi: 10.1016/j.pec.2008.10.009
- Hunter, C., Chew-Graham, C. A., Langer, S., Drinkwater, J., Stenhoff, A., Guthrie, E. A., & Salmon, P. (2015). 'i wouldn't push that further because i don't want to lose her': A multiperspective qualitative study of behaviour change for long-term conditions in primary care. *Health Expect.*, *18*, 1995–2010. Retrieved from <https://doi.org/10.1111/hex.12304> doi: 10.1111/hex.12304
- Imel, Z. E., Pace, B. T., Soma, C. S., Tanana, M., Hirsch, T., Gibson, J., ... Atkins, D. C. (2019). Design feasibility of an automated, machine-learning based feedback system for motivational interviewing. *Psychotherapy*, *56*(2), 318–328. Retrieved from <https://doi.org/10.1037/pst0000221> doi: 10.1037/pst0000221
- Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation mixed-methods study. *JMIR Mhealth Uhealth*, *6*(11), e12106. Retrieved from <https://doi.org/10.2196/12106> doi: 10.2196/12106
- Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., & Naaman, M. (2023). Co-writing with opinionated language models affects users' views. In A. Schmidt et al. (Eds.), *Proceedings of the 2023 CHI conference on human factors in computing systems, CHI 2023, hamburg, germany, april 23-28, 2023* (pp. 111:1–111:15). ACM. Retrieved from <https://doi.org/10.1145/3544548.3581196> doi: 10.1145/3544548.3581196
- Jensen, M. P. (2016). Motivational interviewing for enhancing return to work. In I. Z. Schultz & R. J. Gatchel (Eds.), *Handbook of return to work: From research to practice* (pp. 365–379). Boston, MA: Springer US. Retrieved from https://doi.org/10.1007/978-1-4899-7627-7_21 doi: 10.1007/978-1-4899-7627-7_21
- John, S. A., Rendina, H. J., Starks, T. J., Grov, C., & Parsons, J. T. (2019). Decisional balance and contemplation ladder to support interventions for HIV pre-exposure prophylaxis uptake and persistence. *AIDS Patient Care and STDs*, *33*(2), 67–78. Retrieved from <https://doi.org/10.1089/apc.2018.0136> doi: 10.1089/apc.2018.0136
- Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition, 2nd edition*. Prentice Hall, Pearson Education International. Retrieved from <https://www.worldcat.org/oclc/315913020>
- Kang, E., & Kang, Y. A. (2023). Counseling chatbot design: The effect of anthropomorphic

- chatbot characteristics on user self-disclosure and companionship. *International Journal of Human-Computer Interaction*, 40(11), 2781–2795. Retrieved from <https://doi.org/10.1080/10447318.2022.2163775> doi: 10.1080/10447318.2022.2163775
- Keller, L., Gollwitzer, P. M., & Sheeran, P. (2020). Changing behavior using the model of action phases. In M. S. Hagger, L. D. Cameron, K. Hamilton, N. Hankonen, & T. Lintunen (Eds.), *The handbook of behavior change* (pp. 77–88). Cambridge University Press Cambridge. Retrieved from <https://books.google.de/books?id=IfEFEAAAQBAJ>
- Kidd, C., & Birhane, A. (2023). How AI can distort human beliefs. *Science*, 380, 1222–1223. Retrieved from <https://doi.org/10.1126/science.adi0248> doi: 10.1126/science.adi0248
- Kim, T., Bae, S., Kim, H. A., Lee, S., Hong, H., Yang, C., & Kim, Y. (2024). MindfulDiary: Harnessing large language model to support psychiatric patients' journaling. In F. F. Mueller et al. (Eds.), *Proceedings of the CHI conference on human factors in computing systems, CHI 2024, honolulu, hi, usa, may 11-16, 2024* (pp. 701:1–701:20). ACM. Retrieved from <https://doi.org/10.1145/3613904.3642937> doi: 10.1145/3613904.3642937
- Kocaballi, A. B., Sezgin, E., Clark, L., Carroll, J. M., Huang, Y., Huh-Yoo, J., ... Zubatiy, T. (2022). Design and evaluation challenges of conversational agents in health care and well-being: Selective review study. *J Med Internet Res*, 24(11), e38525. Retrieved from <https://doi.org/10.2196/38525> doi: 10.2196/38525
- Kocielnik, R., Xiao, L., Avrahami, D., & Hsieh, G. (2018). Reflection companion: A conversational system for engaging users in reflection on physical activity. In (Vol. 2, pp. 70:1–70:26). Retrieved from <https://doi.org/10.1145/3214273> doi: 10.1145/3214273
- Kok, G., Gottlieb, N. H., Peters, G.-J. Y., Mullen, P. D., Parcel, G. S., Ruiter, R. A., ... Bartholomew, L. K. (2015). A taxonomy of behaviour change methods: An intervention mapping approach. *Health Psychology Review*, 10(3), 297–312. Retrieved from <https://doi.org/10.1080/17437199.2015.1077155> doi: 10.1080/17437199.2015.1077155
- Kovanovic, V., Joksimovic, S., Mirriahi, N., Blaine, E., Gasevic, D., Siemens, G., & Dawson, S. (2018). Understand students' self-reflections through learning analytics. In A. Pardo et al. (Eds.), *Proceedings of the 8th international conference on learning analytics and knowledge, LAK 2018, sydney, nsw, australia, march 07-09, 2018* (pp. 389–398). ACM. Retrieved from <https://doi.org/10.1145/3170358.3170374> doi: 10.1145/3170358.3170374
- Krebs, P., Norcross, J. C., Nicholson, J. M., & Prochaska, J. O. (2018). Stages of change and psychotherapy outcomes: A review and meta-analysis. *Journal of Clinical Psychology*, 74(11), 1964–1979. Retrieved from <https://doi.org/10.1002/jclp.22683> doi: 10.1002/jclp.22683
- Kuhlthau, C. C. (1988). Developing a model of the library search process: Cognitive and affective aspects. *RQ*, 28(2), 232–242. Retrieved from <http://www.jstor.org/stable/25828262>
- Kumar, H., Musabirov, I., Shi, J., Lauzon, A., Choy, K. K., Gross, O., ... Williams, J. J. (2022). Exploring the design of prompts for applying GPT-3 based chatbots: A mental wellbeing case study on mechanical turk. *CoRR*, abs/2209.11344. Retrieved from <https://doi.org/10.48550/arXiv.2209.11344> doi: 10.48550/ARXIV.2209.11344
- Kwasnicka, D., Dombrowski, S. U., White, M., & Snihotta, F. (2016). Theoretical explanations for maintenance of behaviour change: A systematic review of behaviour theories. *Health Psychology Review*, 10(3), 277–296. Retrieved from <https://doi.org/10.1080/17437199.2016.1151372> doi: 10.1080/17437199.2016.1151372

- Kyle, K. (2019). Measuring lexical richness. In *The routledge handbook of vocabulary studies* (pp. 454–475). Routledge.
- Lahnala, A., Welch, C., Jurgens, D., & Flek, L. (2022). A critical reflection and forward perspective on empathy and natural language processing. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Findings of the association for computational linguistics: EMNLP 2022, abu dhabi, united arab emirates, december 7-11, 2022* (pp. 2139–2158). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/2022.findings-emnlp.157> doi: 10.18653/V1/2022.FINDINGS-EMNLP.157
- Lai, T., Shi, Y., Du, Z., Wu, J., Fu, K., Dou, Y., & Wang, Z. (2023). Psy-LLM: Scaling up global mental health psychological services with AI-based large language models. *CoRR*, *abs/2307.11991*. Retrieved from <https://doi.org/10.48550/arXiv.2307.11991> doi: 10.48550/ARXIV.2307.11991
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. Retrieved from <https://doi.org/10.2307/2529310> doi: 10.2307/2529310
- Larsen, K. R., Michie, S., Hekler, E. B., Gibson, B., Spruijt-Metz, D., Ahern, D., ... Yi, J. (2017). Behavior change interventions: The potential of ontologies for advancing science and practice. *J. Behav. Med.*, *40*, 6–22. Retrieved from <https://doi.org/10.1007/s10865-016-9768-0> doi: 10.1007/s10865-016-9768-0
- Lee, Y., Lim, C., & Choi, H. (2022). Does GPT-3 generate empathetic dialogues? A novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation. In N. Calzolari et al. (Eds.), *Proceedings of the 29th international conference on computational linguistics, COLING 2022, gyeongju, republic of korea, october 12-17, 2022* (pp. 669–683). International Committee on Computational Linguistics. Retrieved from <https://aclanthology.org/2022.coling-1.56>
- Lee, Y., Yamashita, N., Huang, Y., & Fu, W. (2020). "i hear you, i feel you": Encouraging deep self-disclosure through a chatbot. In R. Bernhaupt et al. (Eds.), *CHI '20: CHI conference on human factors in computing systems, honolulu, hi, usa, april 25-30, 2020* (pp. 1–12). ACM. Retrieved from <https://doi.org/10.1145/3313831.3376175> doi: 10.1145/3313831.3376175
- Lee, Y. K., Lee, I., Shin, M., Bae, S., & Hahn, S. (2023). Chain of empathy: Enhancing empathetic response of large language models based on psychotherapy models. *CoRR*, *abs/2311.04915*. Retrieved from <https://doi.org/10.48550/arXiv.2311.04915> doi: 10.48550/ARXIV.2311.04915
- Leiberg, S., & Anders, S. (2006). The multiple facets of empathy: a survey of theory and evidence. In S. Anders, G. Ende, M. Junghofer, J. Kissler, & D. Wildgruber (Eds.), *Understanding emotions* (Vol. 156, p. 419–440). Elsevier. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0079612306560236> doi: [https://doi.org/10.1016/S0079-6123\(06\)56023-6](https://doi.org/10.1016/S0079-6123(06)56023-6)
- Lewis, P. S. H., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems 33: Annual conference on neural information processing systems 2020, neurips 2020, december 6-12, 2020, virtual*. Retrieved from <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>

- Li, B., Hou, Y., & Che, W. (2022). Data augmentation approaches in natural language processing: A survey. *AI Open*, 3, 71–90. Retrieved from <https://doi.org/10.1016/j.aiopen.2022.03.001> doi: 10.1016/J.AIOPEN.2022.03.001
- Ligorio, M. B. (2001). Integrating communication formats: Synchronous versus asynchronous and text-based versus visual. *Comput. Educ.*, 37(2), 103–125. Retrieved from [https://doi.org/10.1016/S0360-1315\(01\)00039-2](https://doi.org/10.1016/S0360-1315(01)00039-2) doi: 10.1016/S0360-1315(01)00039-2
- Lindbladh, E., & Lyttkens, C. H. (2002). Habit versus choice: The process of decision-making in health-related behaviour. *Social Science & Medicine*, 55(3), 451–465. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0277953601001800> doi: 10.1016/S0277-9536(01)00180-0
- Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., & Chen, W. (2022). What makes good in-context examples for GPT-3? In E. Agirre, M. Apidianaki, & I. Vulic (Eds.), *Proceedings of deep learning inside out: The 3rd workshop on knowledge extraction and integration for deep learning architectures, deelio@acl 2022, dublin, ireland and online, may 27, 2022* (pp. 100–114). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/2022.deelio-1.10> doi: 10.18653/V1/2022.DEELIO-1.10
- Liu, J. M., Li, D., Cao, H., Ren, T., Liao, Z., & Wu, J. (2023). ChatCounselor: A large language models for mental health support. *CoRR*, abs/2309.15461. Retrieved from <https://doi.org/10.48550/arXiv.2309.15461> doi: 10.48550/ARXIV.2309.15461
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692. Retrieved from <http://arxiv.org/abs/1907.11692>
- Longpre, S., Wang, Y., & DuBois, C. (2020). How effective is task-agnostic data augmentation for pretrained transformers? In T. Cohn, Y. He, & Y. Liu (Eds.), *Findings of the association for computational linguistics: EMNLP 2020, online event, 16-20 november 2020* (Vol. EMNLP 2020, pp. 4401–4411). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/2020.findings-emnlp.394> doi: 10.18653/V1/2020.FINDINGS-EMNLP.394
- Lord, S. P., Can, D., Yi, M., Marin, R., Dunn, C. W., Imel, Z. E., ... Atkins, D. C. (2015). Advancing methods for reliably assessing motivational interviewing fidelity using the motivational interviewing skills code. *Journal of Substance Abuse Treatment*, 49, 50–57. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0740547214001883> doi: 10.1016/j.jsat.2014.08.005
- Lord, S. P., Sheng, E., Imel, Z. E., Baer, J., & Atkins, D. C. (2015). More than reflections: Empathy in motivational interviewing includes language style synchrony between therapist and client. *Behavior Therapy*, 46(3), 296–303. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0005789414001373> doi: 10.1016/j.beth.2014.11.002
- Losada, D. E., & Crestani, F. (2016). A test collection for research on depression and language use. In N. Fuhr et al. (Eds.), *Experimental IR meets multilinguality, multimodality, and interaction - 7th international conference of the CLEF association, CLEF 2016, évora, portugal, september 5-8, 2016, proceedings* (Vol. 9822, pp. 28–39). Springer. Retrieved from https://doi.org/10.1007/978-3-319-44564-9_3 doi: 10.1007/978-3-319-44564-9_3
- Losada, D. E., Crestani, F., & Parapar, J. (2019). Overview of eRisk at CLEF 2019: Early risk prediction on the internet (extended overview). In L. Cappellato, N. Ferro, D. E. Losada, & H. Müller (Eds.),

- Working notes of CLEF 2019 - conference and labs of the evaluation forum, lugano, switzerland, september 9-12, 2019* (Vol. 2380). CEUR-WS.org. Retrieved from https://ceur-ws.org/Vol-2380/paper_248.pdf
- Lundahl, B., & Burke, B. L. (2009). The effectiveness and applicability of motivational interviewing: A practice-friendly review of four meta-analyses. *J. Clin. Psychol.*, 65, 1232–1245. Retrieved from <https://doi.org/10.1002/jclp.20638> doi: 10.1002/jclp.20638
- Luo, T. C., Aguilera, A., Lyles, C. R., & Figueroa, C. A. (2021). Promoting physical activity through conversational agents: Mixed methods systematic review. *Journal of Medical Internet Research*, 23(9), e25486. Retrieved from <https://doi.org/10.2196/25486> doi: 10.2196/25486
- Ma, Z., Mei, Y., & Su, Z. (2023). Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. *CoRR*, abs/2307.15810. Retrieved from <https://doi.org/10.48550/arXiv.2307.15810> doi: 10.48550/ARXIV.2307.15810
- MacIsaac, A., Mushquash, A. R., & Wekerle, C. (2023). Writing yourself well: Dispositional self-reflection moderates the effect of a smartphone app-based journaling intervention on psychological wellbeing across time. *Behaviour Change*, 40(4), 297–313. Retrieved from <https://doi.org/10.1017/bec.2022.24> doi: 10.1017/bec.2022.24
- Madson, M. B., Mohn, R. S., Schumacher, J. A., & Landry, A. S. (2015). Measuring client experiences of motivational interviewing during a lifestyle intervention. *Measurement and Evaluation in Counseling and Development*, 48(2), 140–151. Retrieved from <https://doi.org/10.1177/0748175614544687> doi: 10.1177/0748175614544687
- Madson, M. B., Mohn, R. S., Zuckoff, A., Schumacher, J. A., Kogan, J., Hutchison, S., ... Stein, B. (2013). Measuring client perceptions of motivational interviewing: Factor analysis of the client evaluation of motivational interviewing scale. *Journal of Substance Abuse Treatment*, 44(3), 330–335. Retrieved from <https://www.sciencedirect.com/science/article/pii/S074054721200150X> doi: 10.1016/j.jsat.2012.08.015
- Madson, M. B., Villarosa, M. C., Schumacher, J. A., & Mohn, R. S. (2016). Evaluating the validity of the client evaluation of motivational interviewing scale in a brief motivational intervention for college student drinkers. *Journal of Substance Abuse Treatment*, 65, 51–57. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0740547216000234> doi: <https://doi.org/10.1016/j.jsat.2016.02.001>
- Magill, M., Apodaca, T. R., Borsari, B., Gaume, J., Hoadley, A., Gordon, R. E. F., ... Moyers, T. (2018). A meta-analysis of motivational interviewing process: Technical, relational, and conditional process models of change. *Journal of Consulting and Clinical Psychology*, 86(2), 140–157. Retrieved from <https://doi.org/10.1037/ccp0000250> doi: 10.1037/ccp0000250
- Magill, M., Gaume, J., Apodaca, T. R., Walthers, J., Mastroleo, N. R., Borsari, B., & Longabaugh, R. (2014). The technical hypothesis of motivational interviewing: A meta-analysis of MI's key causal model. *Journal of Consulting and Clinical Psychology*, 82(6), 973–983. Retrieved from <https://doi.org/10.1037/a0036833> doi: 10.1037/a0036833
- Majumder, N., Hong, P., Peng, S., Lu, J., Ghosal, D., Gelbukh, A. F., ... Poria, S. (2020). MIME: MIMicking emotions for empathetic response generation. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing, EMNLP 2020, online, november 16-20, 2020* (pp. 8968–8979). Association for Computational

- Linguistics. Retrieved from <https://doi.org/10.18653/v1/2020.emnlp-main.721> doi: 10.18653/V1/2020.EMNLP-MAIN.721
- Maples, B., Cerit, M., Vishwanath, A., & Pea, R. (2024). Loneliness and suicide mitigation for students using GPT3-enabled chatbots. *npj Mental Health Res*, 3(4). Retrieved from <https://doi.org/10.1038/s44184-023-00047-6> doi: 10.1038/s44184-023-00047-6
- McTear, M. F. (2020). *Conversational AI: Dialogue systems, conversational agents, and chatbots*. Morgan & Claypool Publishers. Retrieved from <https://doi.org/10.2200/S01060ED1V01Y202010HLT048> doi: 10.2200/S01060ED1V01Y202010HLT048
- Meyer, S., & Elswailer, D. (2022). GLoHBCD: A naturalistic german dataset for language of health behaviour change on online support forums. In N. Calzolari et al. (Eds.), *Proceedings of the thirteenth language resources and evaluation conference, LREC 2022, marseille, france, 20-25 june 2022* (pp. 2226–2235). European Language Resources Association. Retrieved from <https://aclanthology.org/2022.lrec-1.239>
- Meyer, S., & Elswailer, D. (2023). Towards cross-content conversational agents for behaviour change: Investigating domain independence and the role of lexical features in written language around change. In M. Lee, C. Munteanu, M. Porcheron, J. Trippas, & S. T. Völkel (Eds.), *Proceedings of the 5th international conference on conversational user interfaces, CUI 2023, eindhoven, the netherlands, july 19-21, 2023* (pp. 17:1–17:13). ACM. Retrieved from <https://doi.org/10.1145/3571884.3597136> doi: 10.1145/3571884.3597136
- Meyer, S., & Elswailer, D. (2024). "You tell me": A dataset of GPT-4-based behaviour change support conversations. In P. D. Clough, M. Harvey, & F. Hopfgartner (Eds.), *Proceedings of the 2024 ACM SIGIR conference on human information interaction and retrieval, CHIIR 2024, sheffield, united kingdom, march 10-14, 2024* (pp. 411–416). ACM. Retrieved from <https://doi.org/10.1145/3627508.3638330> doi: 10.1145/3627508.3638330
- Meyer, S., Elswailer, D., Ludwig, B., Fernández-Pichel, M., & Losada, D. E. (2022). Do we still need human assessors? Prompt-based GPT-3 user simulation in conversational AI. In M. Halvey, M. E. Foster, J. Dalton, C. Munteanu, & J. Trippas (Eds.), *CUI 2022: 4th conference on conversational user interfaces, glasgow, united kingdom, july 26 - 28, 2022* (pp. 8:1–8:6). ACM. Retrieved from <https://doi.org/10.1145/3543829.3544529> doi: 10.1145/3543829.3544529
- Meyer, S., Fernández-Pichel, M., Elswailer, D., & Losada, D. E. (2024, May). Exploring the challenges of behaviour change language classification: A study on semi-supervised learning and the impact of pseudo-labelled data. In D. Demner-Fushman, S. Ananiadou, P. Thompson, & B. Ondov (Eds.), *Proceedings of the first workshop on patient-oriented language processing (cl4health) @ LREC-COLING 2024* (pp. 229–239). Torino, Italia: ELRA and ICCL. Retrieved from <https://aclanthology.org/2024.cl4health-1.28>
- Meyer, S., Schmidhuber, M., & Kruschwitz, U. (2022, July). MS@IW at SemEval-2022 task 4: Patronising and condescending language detection with synthetically generated data. In G. Emerson et al. (Eds.), *Proceedings of the 16th international workshop on semantic evaluation (SemEval-2022)* (pp. 363–368). Seattle, United States: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2022.semeval-1.47> doi: 10.18653/v1/2022.semeval-1.47
- MgHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3), 276–282. Retrieved from <https://hrcak.srce.hr/89395>
- Michie, S., van Stralen, M. M., & West, R. (2011). The behaviour change wheel: A new method for

- characterising and designing behaviour change interventions. *Implementation Science*, 6(1), 42. Retrieved from <https://doi.org/10.1186/1748-5908-6-42> doi: 10.1186/1748-5908-6-42
- Middleton, K. R., Anton, S. D., & Perri, M. G. (2013). Long-term adherence to health behavior change. *American Journal of Lifestyle Medicine*, 7(6), 395–404. Retrieved from <https://doi.org/10.1177/1559827613488867> doi: 10.1177/1559827613488867
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In Y. Bengio & Y. LeCun (Eds.), *1st international conference on learning representations, ICLR 2013, scottsdale, arizona, usa, may 2-4, 2013, workshop track proceedings*. Retrieved from <http://arxiv.org/abs/1301.3781>
- Miller, S. M. (1980). When is a little information a dangerous thing? coping with stressful events by monitoring versus blunting. In S. Levine & H. Ursin (Eds.), *Coping and health* (pp. 145–169). Boston, MA: Springer US. Retrieved from https://doi.org/10.1007/978-1-4684-1042-6_8 doi: 10.1007/978-1-4684-1042-6_8
- Miller, W. R. (1983). Motivational interviewing with problem drinkers. *Behavioural Psychotherapy*, 11(2), 147–172. Retrieved from <https://doi.org/10.1017/S0141347300006583> doi: 10.1017/S0141347300006583
- Miller, W. R., Moyers, T. B., Ernst, D., & Amrhein, P. (2008). *Manual for the motivational interviewing skill code (misc). version 2.1* (Tech. Rep.). Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico.
- Miller, W. R., & Rollnick, S. (2002). *Motivational interviewing: Preparing people for change* (Second Edition ed.). Guilford Press.
- Miller, W. R., & Rollnick, S. (2009). Ten things that motivational interviewing is not. *Behavioural and Cognitive Psychotherapy*, 37(2), 129–140. Retrieved from <https://doi.org/10.1017/S1352465809005128> doi: 10.1017/S1352465809005128
- Min, D. J., Pérez-Rosas, V., Resnicow, K., & Mihalcea, R. (2022). PAIR: prompt-aware margin ranking for counselor reflection scoring in motivational interviewing. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 conference on empirical methods in natural language processing, EMNLP 2022, abu dhabi, united arab emirates, december 7-11, 2022* (pp. 148–158). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/2022.emnlp-main.11> doi: 10.18653/V1/2022.EMNLP-MAIN.11
- Min, D. J., Pérez-Rosas, V., Resnicow, K., & Mihalcea, R. (2023). VERVE: template-based reflective Rewriting for motivational interviewing. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the association for computational linguistics: EMNLP 2023, singapore, december 6-10, 2023* (pp. 10289–10302). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/2023.findings-emnlp.690> doi: 10.18653/V1/2023.FINDINGS-EMNLP.690
- Monti, P. M., Mastroleo, N. R., Barnett, N. P., Colby, S. M., Kahler, C. W., & Operario, D. (2016). Brief motivational intervention to reduce alcohol and HIV/sexual risk behavior in emergency department patients: A randomized controlled trial. *Journal of Consulting and Clinical Psychology*, 84(7), 580–591. Retrieved from <https://psycnet.apa.org/record/2016-13460-001> doi: <https://doi.org/10.1037/ccp0000097>
- Morris, R. R., Kouddous, K., Kshirsagar, R., & Schueller, S. M. (2018). Towards an artificially empathic conversational agent for mental health applications: System design and user perceptions. *J Med Internet Res*, 20(6), e10148. Retrieved from <https://doi.org/10.2196/10148> doi:

- 10.2196/10148
- Moyers, T. B., Manuel, J. K., & Ernst, D. (2014). *Motivational interviewing treatment integrity coding manual 4.1 (miti 4.1)* (Tech. Rep.). University of New Mexico. Retrieved from <https://motivationalinterviewing.org/sites/default/files/miti4.2.pdf>
- Moyers, T. B., Rowell, L. N., Manuel, J. K., Ernst, D., & Houck, J. M. (2016). The motivational interviewing treatment integrity code (MITI 4): Rationale, preliminary reliability and validity. *Journal of Substance Abuse Treatment*, 65, 36–42. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0740547216000143> doi: 10.1016/j.jsat.2016.01.001
- Murray, J., Fenton, G., Honey, S., Bara, A. C., Hill, K. M., & House, A. (2013). A qualitative synthesis of factors influencing maintenance of lifestyle behaviour change in individuals with high cardiovascular risk. *BMC Cardiovasc Disord*, 13(48). Retrieved from <https://doi.org/10.1186/1471-2261-13-48> doi: 10.1186/1471-2261-13-48
- Myneni, S., Cobb, N., & Cohen, T. (2016). In pursuit of theoretical ground in behavior change support systems: Analysis of peer-to-peer communication in a health-related online community. *J Med Internet Res*, 18(2), e28. Retrieved from <http://www.jmir.org/2016/2/e28/> doi: 10.2196/jmir.4671
- Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In B. Adelson, S. T. Dumais, & J. S. Olson (Eds.), *Conference on human factors in computing systems, CHI 1994, boston, massachusetts, usa, april 24-28, 1994, proceedings* (pp. 72–78). ACM. Retrieved from <https://doi.org/10.1145/191666.191703> doi: 10.1145/191666.191703
- Nicol, E., Briggs, J., Moncur, W., Htait, A., Carey, D. P., Azzopardi, L., & Schafer, B. (2022, nov). Revealing cumulative risks in online personal information: A data narrative study. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2). Retrieved from <https://doi.org/10.1145/3555214> doi: 10.1145/3555214
- Nurmi, J., Knittle, K., Ginchev, T., Khattak, F., Helf, C., Zwickl, P., . . . Haukkala, A. (2020). Engaging users in the behavior change process with digitalized motivational interviewing and gamification: Development and feasibility testing of the precious app. *JMIR mHealth and uHealth*, 8(1), e12884. Retrieved from <https://mhealth.jmir.org/2020/1/e12884> doi: 10.2196/12884
- O'Brien, H. L., Cairns, P. A., & Hall, M. (2018). A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *Int. J. Hum. Comput. Stud.*, 112, 28–39. Retrieved from <https://doi.org/10.1016/j.ijhcs.2018.01.004> doi: 10.1016/J.IJHCS.2018.01.004
- O'Dell, J. W., & Dickson, J. (1984). Eliza as a “therapeutic” tool. *J. Clin. Psychol.*, 40, 942–945. Retrieved from [https://doi.org/10.1002/1097-4679\(198407\)40:4<942::AID-JCLP2270400412>3.0.CO;2-D](https://doi.org/10.1002/1097-4679(198407)40:4<942::AID-JCLP2270400412>3.0.CO;2-D) doi: 10.1002/1097-4679(198407)40:4(942::AID-JCLP2270400412)3.0.CO;2-D
- Office for National Statistics. (2023). *Most adults report making some changes to their lifestyle for environmental reasons* (Tech. Rep.). ons.gov.uk. Retrieved from <https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/articles/mostadultsreportmakingchangestotheirifestyleforenvironmentalreasons/2023-07-05>
- Ogden, J., & Flanagan, Z. (2008). Beliefs about the causes and solutions to obesity: A comparison of GPs and lay people. *Patient Education and Counseling*, 71(1), 72–78. Retrieved from <https://>

- doi.org/10.1016/j.pec.2007.11.022 doi: 10.1016/j.pec.2007.11.022
- Oh, Y. J., Zhang, J., Fang, M.-L., & Fukuoka, Y. (2021). A systematic review of artificial intelligence chatbots for promoting physical activity, healthy diet, and weight loss. *Int J Behav Nutr Phys Act*, 18, 160. Retrieved from <https://doi.org/10.1186/s12966-021-01224-6> doi: 10.1186/s12966-021-01224-6
- Oinas-Kukkonen, H. (2013). A foundation for the study of behavior change support systems. *Pers. Ubiquitous Comput.*, 17(6), 1223–1235. Retrieved from <https://doi.org/10.1007/s00779-012-0591-5> doi: 10.1007/S00779-012-0591-5
- Olafsson, S., O'Leary, T., & Bickmore, T. W. (2019). Coerced change-talk with conversational agents promotes confidence in behavior change. In O. Mayora, S. Forti, J. Meyer, & L. Mamykina (Eds.), *Proceedings of the 13th EAI international conference on pervasive computing technologies for healthcare, pervasivehealth 2019, trento, italy, 20-23 may 2019* (pp. 31–40). ACM. Retrieved from <https://doi.org/10.1145/3329189.3329202> doi: 10.1145/3329189.3329202
- Olafsson, S., Wallace, B. C., & Bickmore, T. W. (2020). Towards a computational framework for automating substance use counseling with virtual agents. In A. E. F. Seghrouchni, G. Sukthankar, B. An, & N. Yorke-Smith (Eds.), *Proceedings of the 19th international conference on autonomous agents and multiagent systems, AAMAS '20, auckland, new zealand, may 9-13, 2020* (pp. 966–974). International Foundation for Autonomous Agents and Multiagent Systems. Retrieved from <https://dl.acm.org/doi/10.5555/3398761.3398874> doi: 10.5555/3398761.3398874
- OpenAI. (2023a). GPT-4 technical report. *CoRR*, abs/2303.08774. Retrieved from <https://doi.org/10.48550/arXiv.2303.08774> doi: 10.48550/ARXIV.2303.08774
- OpenAI. (2023b). *Prompt engineering*. <https://platform.openai.com/docs/guides/prompt-engineering/six-strategies-for-getting-better-results>. (Accessed: 2024-03-07)
- Ortloff, A., Zimmerman, S., Elseweiler, D., & Henze, N. (2021). The effect of nudges and boosts on browsing privacy in a naturalistic environment. In F. Scholer, P. Thomas, D. Elseweiler, H. Joho, N. Kando, & C. Smith (Eds.), *CHIIR '21: ACM SIGIR conference on human information interaction and retrieval, canberra, act, australia, march 14-19, 2021* (pp. 63–73). ACM. Retrieved from <https://doi.org/10.1145/3406522.3446014> doi: 10.1145/3406522.3446014
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... Lowe, R. (2022). Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems 35: Annual conference on neural information processing systems 2022, neurips 2022, new orleans, la, usa, november 28 - december 9, 2022*. Retrieved from http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html
- Pandey, S., & Sharma, S. (2023). A comparative study of retrieval-based and generative-based chatbots using deep learning and machine learning. *Healthcare Analytics*, 3, 100198. Retrieved from <https://doi.org/10.1016/j.health.2023.100198> doi: 10.1016/j.health.2023.100198
- Parapar, J., Martín-Rodilla, P., Losada, D. E., & Crestani, F. (2021). Overview of eRisk at CLEF 2021: Early risk prediction on the internet (extended overview). In G. Faggioli, N. Ferro, A. Joly, M. Maistro, & F. Piroi (Eds.), *Proceedings of the working notes of CLEF 2021 - conference and labs of the evaluation forum, bucharest, romania, september 21st - to - 24th, 2021* (Vol. 2936, pp. 864–887). CEUR-WS.org. Retrieved from <https://ceur-ws.org/Vol-2936/paper-72.pdf>

- Park, S., Choi, J., Lee, S., Oh, C., Kim, C., La, S., ... Suh, B. (2019). Designing a chatbot for a brief motivational interview on stress management: Qualitative case study. *J Med Internet Res*, 21(4), e12231. Retrieved from <https://doi.org/10.2196/12231> doi: 10.2196/12231
- Pennebaker, J. W., & Seagal, J. D. (1999). Forming a story: The health benefits of narrative. *Journal of Clinical Psychology*, 55(10), 1243–1254. Retrieved from <https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-4679%28199910%2955%3A10%3C1243%3A%3AAID-JCLP6%3E3.0.CO%3B2-N> doi: 10.1002/(SICI)1097-4679(199910)55:10<1243::AID-JCLP6>3.0.CO;2-N
- Pentina, I., Hancock, T., & Xie, T. (2023). Exploring relationship development with social chatbots: A mixed-method study of replika. *Comput. Hum. Behav.*, 140, 107600. Retrieved from <https://doi.org/10.1016/j.chb.2022.107600> doi: 10.1016/J.CHB.2022.107600
- Pereira, J., & Díaz, O. (2019). Using health chatbots for behavior change: A mapping study. *J. Medical Syst.*, 43(5), 135:1–135:13. Retrieved from <https://doi.org/10.1007/s10916-019-1237-1> doi: 10.1007/S10916-019-1237-1
- Pérez-Rosas, V., Mihalcea, R., Resnicow, K., Singh, S., & An, L. C. (2016). Building a motivational interviewing dataset. In K. Hollingshead & L. H. Ungar (Eds.), *Proceedings of the 3rd workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality, clpsych@naacl-hlt 2016, june 16, 2016, san diego, california, USA* (pp. 42–51). The Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/w16-0305> doi: 10.18653/V1/W16-0305
- Pérez-Rosas, V., Mihalcea, R., Resnicow, K., Singh, S., & An, L. C. (2017). Understanding and predicting empathic behavior in counseling therapy. In R. Barzilay & M. Kan (Eds.), *Proceedings of the 55th annual meeting of the association for computational linguistics, ACL 2017, vancouver, canada, july 30 - august 4, volume 1: Long papers* (pp. 1426–1435). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/P17-1131> doi: 10.18653/V1/P17-1131
- Pérez-Rosas, V., Mihalcea, R., Resnicow, K., Singh, S., An, L. C., Goggin, K. J., & Catley, D. (2017). *Predicting counselor behaviors in motivational interviewing encounters*. Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/e17-1106> doi: 10.18653/V1/E17-1106
- Pérez-Rosas, V., Sun, X., Li, C., Wang, Y., Resnicow, K., & Mihalcea, R. (2018). Analyzing the quality of counseling conversations: the tell-tale signs of high-quality counseling. In N. Calzolari et al. (Eds.), *Proceedings of the eleventh international conference on language resources and evaluation, LREC 2018, miyazaki, japan, may 7-12, 2018*. European Language Resources Association (ELRA). Retrieved from <http://www.lrec-conf.org/proceedings/lrec2018/summaries/1033.html>
- Pérez-Rosas, V., Wu, X., Resnicow, K., & Mihalcea, R. (2019). What makes a good counselor? Learning to distinguish between high-quality and low-quality counseling conversations. In A. Korhonen, D. R. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th conference of the association for computational linguistics, ACL 2019, florence, italy, july 28- august 2, 2019, volume 1: Long papers* (pp. 926–935). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/p19-1088> doi: 10.18653/V1/P19-1088
- Pestian, J. P., Matykiewicz, P., Linn-Gust, M., South, B., Uzuner, O., Wiebe, J., ... Hurdle, J.

- (2012). Sentiment analysis of suicide notes: A shared task. *Biomedical Informatics Insights*, 5s1, BII.S9042. Retrieved from <https://doi.org/10.4137/BII.S9042> doi: 10.4137/BII.S9042
- Petrocelli, J. V. (2002). Processes and stages of change: Counseling with the transtheoretical model of change. *Journal of Counseling & Development*, 80(1), 22–30. Retrieved from <https://doi.org/10.1002/j.1556-6678.2002.tb00162.x> doi: 10.1002/j.1556-6678.2002.tb00162.x
- Pettigrew, K. E., Fidel, R., & Bruce, H. (2001). Conceptual frameworks in information behavior. *Annual review of information science and technology (ARIST)*, 35(43–78).
- Pinho, M. G. M., Mackenbach, J. D., Charreire, H., Oppert, J.-M., Bárdos, H., Glonti, K., ... Lakerveld, J. (2018). Exploring the relationship between perceived barriers to healthy eating and dietary behaviours in european adults. *Eur J Nutr*, 57, 1761–1770. Retrieved from <https://doi.org/10.1007/s00394-017-1458-3> doi: 10.1007/s00394-017-1458-3
- Ploderer, B., Reitberger, W., Oinas-Kukkonen, H., & van Gemert-Pijnen, J. E. W. C. (2014). Social interaction and reflection for behaviour change. *Pers. Ubiquitous Comput.*, 18(7), 1667–1676. Retrieved from <https://doi.org/10.1007/s00779-014-0779-y> doi: 10.1007/S00779-014-0779-Y
- Pollak, K. I., Childers, J. W., & Arnold, R. M. (2011). Applying motivational interviewing techniques to palliative care communication. *Journal of Palliative Medicine*, 14(5), 587–592. Retrieved from <https://doi.org/10.1089/jpm.2010.0495> doi: 10.1089/jpm.2010.0495
- Portela, M., & Granell-Canut, C. (2017). A new friend in our smartphone? Observing interactions with chatbots in the search of emotional engagement. In J. M. González-Calleros (Ed.), *Proceedings of the XVIII international conference on human computer interaction, interacción 2017, cancan, mexico, september 25-27, 2017* (pp. 48:1–48:7). ACM. Retrieved from <https://doi.org/10.1145/3123818.3123826> doi: 10.1145/3123818.3123826
- Prochaska, J. O., & Clemente, C. C. D. (1983). Stages and processes of self-change of smoking: Toward an integrative model of change. *Journal of Consulting and Clinical Psychology & Health*, 51(3), 390–395. Retrieved from <https://doi.org/10.1037/0022-006X.51.3.390> doi: 10.1037/0022-006X.51.3.390
- Prochaska, J. O., & Velicer, W. F. (1997). The transtheoretical model of health behavior change. *American Journal of Health Promotion*, 12(1), 38–48. Retrieved from <https://doi.org/10.4278/0890-1171-12.1.38> doi: 10.4278/0890-1171-12.1.38
- Prochaska, J. O., Velicer, W. F., Guadagnoli, E., Rossi, J. S., & DiClemente, C. C. (1991). Patterns of change: Dynamic typology applied to smoking cessation. *Multivariate Behavioral Research*, 26(1), 83–107. Retrieved from https://doi.org/10.1207/s15327906mbr2601_5 doi: 10.1207/s15327906mbr2601_5
- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training* (Tech. Rep.). OpenAI. Retrieved from https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
- Rafizadeh, E., Morewitz, S., & Mukherjea, A. (2021). Handwritten journals for supporting behavior change. *The International Journal of Health, Wellness and Society*, 11(1), 143–157. Retrieved from <https://doi.org/10.18848/2156-8960/CGP/v11i01/143-157> doi: 10.18848/2156-8960/CGP/v11i01/143-157
- Rapp, A., Curti, L., & Boldi, A. (2021). The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *Int. J. Hum. Comput. Stud.*, 151,

102630. Retrieved from <https://doi.org/10.1016/j.ijhcs.2021.102630> doi: 10.1016/J.IJHCS.2021.102630
- Rashkin, H., Smith, E. M., Li, M., & Boureau, Y. (2019). Towards empathetic open-domain conversation models: A new benchmark and dataset. In A. Korhonen, D. R. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th conference of the association for computational linguistics, ACL 2019, florence, italy, july 28- august 2, 2019, volume 1: Long papers* (pp. 5370–5381). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/p19-1534> doi: 10.18653/V1/P19-1534
- Raswsthorne, M., Cannon, A., & van Mierlo, T. (2022). Identifiable context: Developing natural language processing to detect readiness to change in digital health support networks. *JMIR Preprints*, 5. Retrieved from <https://doi.org/10.2196/preprints.39290> doi: 10.2196/preprints.39290
- Reifegerste, D., Wasgien, K., & Hagen, L. M. (2017). Online social support for obese adults: Exploring the role of forum activity. *Int. J. Medical Informatics*, 101, 1–8. Retrieved from <https://doi.org/10.1016/j.ijmedinf.2017.02.003> doi: 10.1016/J.IJMEDINF.2017.02.003
- Resnicow, K., Dilorio, C., Soet, J. E., Borrelli, B., Hecht, J., & Ernst, D. (2002). Motivational interviewing in health promotion: It sounds like something is changing. *Health Psychology*, 21(5), 444–451. Retrieved from <https://doi.org/10.1037/0278-6133.21.5.444> doi: 10.1037/0278-6133.21.5.444
- Resnicow, K., McMaster, F., & Rollnick, S. (2012). Action reflections: A client-centered technique to bridge the WHY–HOW transition in motivational interviewing. *Behavioural and Cognitive Psychotherapy*, 40(4), 474–480. Retrieved from <https://doi.org/10.1017/S1352465812000124> doi: 10.1017/S1352465812000124
- Reynolds, L., & McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. In Y. Kitamura, A. Quigley, K. Isbister, & T. Igarashi (Eds.), *CHI '21: CHI conference on human factors in computing systems, virtual event / yokohama japan, may 8-13, 2021, extended abstracts* (pp. 314:1–314:7). ACM. Retrieved from <https://doi.org/10.1145/3411763.3451760> doi: 10.1145/3411763.3451760
- Rodda, S. N., Hing, N., Hodgins, D. C., Cheetham, A., Dickins, M., & Lubman, D. I. (2018). Behaviour change strategies for problem gambling: An analysis of online posts. *International Gambling Studies*, 18(3), 420–438. Retrieved from <https://doi.org/10.1080/14459795.2018.1432670> doi: 10.1080/14459795.2018.1432670
- Rose, J., Ahmed, I., Keilty, E., Cooper, C., & Selby, P. (2022). Generation and classification of motivational-interviewing-style reflections for smoking behaviour change using few-shot learning with transformers. *TechRxiv*. Retrieved from <https://doi.org/10.36227/techrxiv.20029880.v1> doi: 10.36227/techrxiv.20029880.v1
- Ruane, E., Birhane, A., & Ventresque, A. (2019). Conversational AI: Social and ethical considerations. In E. Curry, M. T. Keane, A. Ojo, & D. Salwala (Eds.), *Proceedings for the 27th AIAI irish conference on artificial intelligence and cognitive science, galway, ireland, december 5-6, 2019* (Vol. 2563, pp. 104–115). CEUR-WS.org. Retrieved from <https://ceur-ws.org/Vol-2563/aics.12.pdf>
- Rubak, S., Sandbæk, A., Lauritzen, T., & Christensen, B. (2005). Motivational interviewing: A systematic review and meta-analysis. *British Journal of General Practice*, 55(513), 305–312. Retrieved from <https://bjgp.org/content/55/513/305>

- Ruthven, I. (2019). Making meaning: A focus for information interactions research. In L. Azzopardi, M. Halvey, I. Ruthven, H. Joho, V. Murdock, & P. Qvarfordt (Eds.), *Proceedings of the 2019 conference on human information interaction and retrieval, CHIIR 2019, glasgow, scotland, uk, march 10-14, 2019* (pp. 163–171). ACM. Retrieved from <https://doi.org/10.1145/3295750.3298938> doi: 10.1145/3295750.3298938
- Ruthven, I. (2022a). *Dealing with change through information sculpting*. Emerald Publishing Limited. Retrieved from <https://doi.org/10.1108/978-1-80382-047-720221013> doi: 10.1108/978-1-80382-047-720221013
- Ruthven, I. (2022b). An information behavior theory of transitions. *J. Assoc. Inf. Sci. Technol.*, 73(4), 579–593. Retrieved from <https://doi.org/10.1002/asi.24588> doi: 10.1002/ASI.24588
- Ruthven, I. (2024). Information shaping. *J. Assoc. Inf. Sci. Technol.*, 75(4), 469–482. Retrieved from <https://doi.org/10.1002/asi.24871> doi: 10.1002/ASI.24871
- Ryan, R. M., & Deci, E. L. (2020). Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions. *Contemporary Educational Psychology*, 61. Retrieved from <https://doi.org/10.1016/j.cedpsych.2020.101860> doi: 10.1016/j.cedpsych.2020.101860
- Saha, T., Chopra, S., Saha, S., Bhattacharyya, P., & Kumar, P. (2021). A large-scale dataset for motivational dialogue system: An application of natural language generation to mental health. In *International joint conference on neural networks, IJCNN 2021, shenzhen, china, july 18-22, 2021* (pp. 1–8). IEEE. Retrieved from <https://doi.org/10.1109/IJCNN52387.2021.9533924> doi: 10.1109/IJCNN52387.2021.9533924
- Saiyed, A., Layton, J., Borsari, B., Cheng, J., Kanzaveli, T., Tsvetovat, M., & Satterfield, J. (2022). Technology-assisted motivational interviewing: Developing a scalable framework for promoting engagement with tobacco cessation using NLP and machine learning. *Procedia Computer Science*, 206, 121–131. Retrieved from <https://doi.org/10.1016/j.procs.2022.09.091> doi: 10.1016/j.procs.2022.09.091
- Samrose, S., & Hoque, E. (2022). MIA: motivational interviewing agent for improving conversational skills in remote group discussions. *Proc. ACM Hum. Comput. Interact.*, 6(GROUP), 45:1–45:24. Retrieved from <https://doi.org/10.1145/3492864> doi: 10.1145/3492864
- Satterfield, J., Layton, J., Delacruz, J., Cheng, J., Saiyed, A., Kanzaveli, T., . . . Borsari, B. (2022). Technology-assisted motivational interviewing: Development and pilot results for an nlp-driven chatbot on smoking cessation. In *2022 annual research meeting*.
- Savolainen, R. (1995). Everyday life information seeking: Approaching information seeking in the context of “way of life”. *Library & Information Science Research*, 17(3), 259–294. Retrieved from [https://doi.org/10.1016/0740-8188\(95\)90048-9](https://doi.org/10.1016/0740-8188(95)90048-9) doi: 10.1016/0740-8188(95)90048-9
- Savolainen, R. (2002). Network competence and information seeking on the internet: From definitions towards a social cognitive model. *J. Documentation*, 58(2), 211–226. Retrieved from <https://doi.org/10.1108/00220410210425467> doi: 10.1108/00220410210425467
- Savolainen, R. (2017). Contributions to conceptual growth: The elaboration of ellis's model for information-seeking behavior. *J. Assoc. Inf. Sci. Technol.*, 68(3), 594–608. Retrieved from <https://doi.org/10.1002/asi.23680> doi: 10.1002/ASI.23680
- Savolainen, R., & Wilson, T. (2009). The behaviour/practice debate: A discussion prompted by tom wilson's review of reijo savolainen's everyday information practices: A social phenomenological

- perspective. *Information Research*, 14(2). Retrieved from <http://InformationR.net/ir/14-2/paper403.html>
- Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilic, S., Hesslow, D., ... et al. (2022). BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, *abs/2211.05100*. Retrieved from <https://doi.org/10.48550/arXiv.2211.05100> doi: 10.48550/ARXIV.2211.05100
- Schulman, D., Bickmore, T. W., & Sidner, C. L. (2011). An intelligent conversational agent for promoting long-term health behavior change using motivational interviewing. In D. H. Fisher & M. L. Maher (Eds.), *AI and health communication, papers from the 2011 AAAI spring symposium, technical report ss-11-01, stanford, california, usa, march 21-23, 2011*. AAAI. Retrieved from <http://www.aaai.org/ocs/index.php/SSS/SSS11/paper/view/2401>
- Shah, R. S., Holt, F., Hayati, S. A., Agarwal, A., Wang, Y., Kraut, R. E., & Yang, D. (2022). Modeling motivational interviewing strategies on an online peer-to-peer counseling platform. *Proc. ACM Hum. Comput. Interact.*, 6(CSCW2), 1–24. Retrieved from <https://doi.org/10.1145/3555640> doi: 10.1145/3555640
- Shao, R. (2023). An empathetic AI for mental health intervention: Conceptualizing and examining artificial empathy. In *Proceedings of the 2nd empathy-centric design workshop, EMPATHICH 2023, hamburg, germany, 23 april 2023* (pp. 4:1–4:6). ACM. Retrieved from <https://doi.org/10.1145/3588967.3588971> doi: 10.1145/3588967.3588971
- Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C., & Althoff, T. (2023). Human-ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nat. Mac. Intell.*, 5(1), 46–57. Retrieved from <https://doi.org/10.1038/s42256-022-00593-2> doi: 10.1038/s42256-022-00593-2
- Sharma, A., Miner, A. S., Atkins, D. C., & Althoff, T. (2020). A computational approach to understanding empathy expressed in text-based mental health support. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing, EMNLP 2020, online, november 16-20, 2020* (pp. 5263–5276). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/2020.emnlp-main.425> doi: 10.18653/v1/2020.EMNLP-MAIN.425
- Sharma, A., Rushton, K., Lin, I. E., Wadden, D., Lucas, K. G., Miner, A. S., ... Althoff, T. (2023). Cognitive reframing of negative thoughts through human-language model interaction. In A. Rogers, J. L. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers), ACL 2023, toronto, canada, july 9-14, 2023* (pp. 9977–10000). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/2023.acl-long.555> doi: 10.18653/v1/2023.ACL-LONG.555
- Shen, S., Welch, C., Mihalcea, R., & Pérez-Rosas, V. (2020). Counseling-style reflection generation using generative pretrained transformers with augmented context. In O. Pietquin et al. (Eds.), *Proceedings of the 21th annual meeting of the special interest group on discourse and dialogue, sigdial 2020, 1st virtual meeting, july 1-3, 2020* (pp. 10–20). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/2020.sigdial-1.2> doi: 10.18653/v1/2020.SIGDIAL-1.2
- Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Text data augmentation for deep learning. *J. Big Data*, 8(1), 101. Retrieved from <https://doi.org/10.1186/s40537-021-00492-0> doi: 10.1186/s40537-021-00492-0

- Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2021). My chatbot companion - a study of human-chatbot relationships. *Int. J. Hum. Comput. Stud.*, 149, 102601. Retrieved from <https://doi.org/10.1016/j.ijhcs.2021.102601> doi: 10.1016/J.IJHCS.2021.102601
- Smink, W., Sools, A. M., van der Zwaan, J. M., Wiegersma, S., Veldkamp, B. P., & Westerhof, G. J. (2019). Towards text mining therapeutic change: A systematic review of text-based methods for therapeutic change process research. *PLOS ONE*, 14(12), e0225703. Retrieved from <https://doi.org/10.1371/journal.pone.0225703> doi: 10.1371/journal.pone.0225703
- Smriti, D., Kao, T.-S. A., Rathod, R., Shin, J. Y., Peng, W., Williams, J., ... Huh-Yoo, J. (2022). Motivational interviewing conversational agent for parents as proxies for their children in healthy eating: Development and user testing. *JMIR Hum Factors*, 9(4), e38908. Retrieved from <https://doi.org/10.2196/38908> doi: 10.2196/38908
- Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast - but is it good? Evaluating non-expert annotations for natural language tasks. In M. Lapata & H. T. Ng (Eds.), *2008 conference on empirical methods in natural language processing, EMNLP 2008, proceedings of the conference, 25-27 october 2008, honolulu, hawaii, usa, A meeting of sigdat, a special interest group of the ACL* (pp. 254–263). ACL. Retrieved from <https://aclanthology.org/D08-1027/>
- spaCy: *Industrial-strength natural language processing in python*. (n.d.). <https://spacy.io>. (Accessed: 2024-06-24)
- Stephen Rollnick, C. C. B., William R. Miller, & Aloia, M. S. (2008). *Motivational interviewing in health care: Helping patients change behavior* (Vol. 5) (No. 3). Taylor & Francis. Retrieved from <https://doi.org/10.1080/15412550802093108> doi: 10.1080/15412550802093108
- Stephenson, G. M., & Zygouris, N. (2007). Effects of self reflection on engagement in a 12-step addiction treatment programme: A linguistic analysis of diary entries. *Addictive Behaviors*, 32(2), 416–424. Retrieved from <https://doi.org/10.1016/j.addbeh.2006.05.011> doi: 10.1016/j.addbeh.2006.05.011
- Strait, G. G., Smith, B. H., McQuillin, S., Terry, J., Swan, S., & Malone, P. S. (2012). A randomized trial of motivational interviewing to improve middle school students' academic performance. *J. Community Psychol.*, 40, 1032–1039. Retrieved from <https://doi.org/10.1002/jcop.21511> doi: 10.1002/jcop.21511
- Strand, M., Eng, L. S., & Gammon, D. (2020). Combining online and offline peer support groups in community mental health care settings: A qualitative study of service users' experiences. *Int J Ment Health Syst*, 14(39). Retrieved from <https://doi.org/10.1186/s13033-020-00370-x> doi: 10.1186/s13033-020-00370-x
- Sun, X., Pei, J., de Wit, J., Aliannejadi, M., Krahmer, E., Dobber, J. T. P., & Bosch, J. A. (2024). Eliciting motivational interviewing skill codes in psychotherapy with LLMs: A bilingual dataset and analytical study. In N. Calzolari, M. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation, LREC/COLING 2024, 20-25 may, 2024, torino, italy* (pp. 5609–5621). ELRA and ICCL. Retrieved from <https://aclanthology.org/2024.lrec-main.498>
- Svikhnushina, E., Placinta, A., & Pu, P. (2021). User expectations of conversational chatbots based on online reviews. In W. Ju, L. Oehlberg, S. Follmer, S. E. Fox, & S. Kuznetsov (Eds.), *DIS '21: Designing interactive systems conference 2021, virtual event, usa, 28 june, july 2, 2021*

- (pp. 1481–1491). ACM. Retrieved from <https://doi.org/10.1145/3461778.3462125> doi: 10.1145/3461778.3462125
- Tanana, M., Hallgren, K. A., Imel, Z. E., Atkins, D. C., & Srikumar, V. (2016). A comparison of natural language processing methods for automated coding of motivational interviewing. *Journal of Substance Abuse Treatment*, 65, 45–50. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0740547216000222> doi: 10.1016/j.jsat.2016.01.006
- Tanana, M. J., Soma, C. S., Srikumar, V., Atkins, D. C., & Imel, Z. E. (2019). Development and evaluation of ClientBot: Patient-like conversational agent to train basic counseling skills. *J Med Internet Res*, 21(7), e12529. Retrieved from <https://doi.org/10.2196/12529> doi: 10.2196/12529
- Tang, R., Chuang, Y., & Hu, X. (2023). The science of detecting LLM-generated texts. *CoRR*, abs/2303.07205. Retrieved from <https://doi.org/10.48550/arXiv.2303.07205> doi: 10.48550/ARXIV.2303.07205
- Tavabi, L., Stefanov, K., Zhang, L., Borsari, B., Woolley, J. D., Scherer, S., & Soleymani, M. (2020). Multimodal automatic coding of client behavior in motivational interviewing. In K. P. Truong, D. Heylen, M. Czerwinski, N. Berthouze, M. Chetouani, & M. Nakano (Eds.), *ICMI '20: International conference on multimodal interaction, virtual event, the netherlands, october 25-29, 2020* (pp. 406–413). ACM. Retrieved from <https://doi.org/10.1145/3382507.3418853> doi: 10.1145/3382507.3418853
- Tavabi, L., Tran, T., Borsari, B., Delacruz, J., Woolley, J. D., Scherer, S., & Soleymani, M. (2023). Therapist empathy assessment in motivational interviews. In *11th international conference on affective computing and intelligent interaction, ACII 2023, cambridge, ma, usa, september 10-13, 2023* (pp. 1–8). IEEE. Retrieved from <https://doi.org/10.1109/ACII59096.2023.10388176> doi: 10.1109/ACII59096.2023.10388176
- Tavabi, L., Tran, T., Stefanov, K., Borsari, B., Woolley, J., Scherer, S., & Soleymani, M. (2021, June). Analysis of behavior classification in motivational interviewing. In N. Goharian, P. Resnik, A. Yates, M. Ireland, K. Niederhoffer, & R. Resnik (Eds.), *Proceedings of the seventh workshop on computational linguistics and clinical psychology: Improving access* (pp. 110–115). Online: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2021.clpsych-1.13> doi: 10.18653/v1/2021.clpsych-1.13
- Teixeira, P. J., Silva, M. N., Coutinho, S. R., Palmeira, A. L., Mata, J., Vieira, P. N., ... Sardinha, L. B. (2010). Mediators of weight loss and weight loss maintenance in middle-aged women. *Obesity*, 18(4), 725–735. Retrieved from <https://doi.org/10.1038/oby.2009.281> doi: 10.1038/oby.2009.281
- Teixeira, P. J., Silva, M. N., Mata, J., Palmeira, A. L., & Markland, D. (2012). Motivation, self-determination, and long-term weight control. *Int J Behav Nutr Phys Act*, 9(22). Retrieved from <https://doi.org/10.1186/1479-5868-9-22> doi: 10.1186/1479-5868-9-22
- Terry, J., Smith, B., Strait, G., & McQuillin, S. (2013). Motivational interviewing to improve middle school students' academic performance: A replication study. *J. Community Psychol.*, 41, 902–909. Retrieved from <https://doi.org/10.1002/jcop.21574> doi: 10.1002/jcop.21574
- Tiedemann, J., & Thottingal, S. (2020). OPUS-MT - Building open translation services for the world. In M. L. Forcada et al. (Eds.), *Proceedings of the 22nd annual conference of the european association for machine translation, EAMT 2020, lisboa, portugal, november 3-5, 2020* (pp. 479–

- 480). European Association for Machine Translation. Retrieved from <https://aclanthology.org/2020.eamt-1.61/>
- Ullrich, P. M., & Lutgendorf, S. K. (2002). Journaling about stressful events: Effects of cognitive processing and emotional expression. *ann. behav. med.*, 24, 244–250. Retrieved from https://doi.org/10.1207/S15324796ABM2403_10 doi: 10.1207/S15324796ABM2403_10
- van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Mach. Learn.*, 109(2), 373–440. Retrieved from <https://doi.org/10.1007/s10994-019-05855-6> doi: 10.1007/S10994-019-05855-6
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. In I. Guyon et al. (Eds.), *Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, december 4-9, 2017, long beach, ca, USA* (pp. 5998–6008). Retrieved from <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>
- Velicer, W. F., Prochaska, J. O., Fava, J. L., Norman, G. J., & Redding, C. A. (1998). Smoking cessation and stress management: applications of the transtheoretical model. *Homeostasis*, 38(5-6), 216–33.
- Vowels, L. M., Francois-Walcott, R. R., & Darwiche, J. (2024). AI in relationship counselling: Evaluating ChatGPT's therapeutic capabilities in providing relationship advice. *Computers in Human Behavior: Artificial Humans*, 2(2), 100078. Retrieved from <https://www.sciencedirect.com/science/article/pii/S2949882124000380> doi: <https://doi.org/10.1016/j.chbah.2024.100078>
- Wallace, R. S. (2003). *The elements of AIML style* (Tech. Rep.). Retrieved from <https://files.ifi.uzh.ch/cl/hess/classes/seminare/chatbots/style.pdf>
- Wallace, R. S. (2009). The anatomy of a.i.c.e. In R. Epstein, G. Roberts, & G. Beber (Eds.), *Parsing the turing test: Philosophical and methodological issues in the quest for the thinking computer* (pp. 181–210). Dordrecht: Springer Netherlands. Retrieved from https://doi.org/10.1007/978-1-4020-6710-5_13 doi: 10.1007/978-1-4020-6710-5_13
- Wallen, K. E., & Daut, E. (2018). The challenge and opportunity of behaviour change methods and frameworks to reduce demand for illegal wildlife. *Nature Conservation*, 26, 55–75. Retrieved from <https://doi.org/10.3897/natureconservation.26.22725> doi: 10.3897/natureconservation.26.22725
- Wang, X., Wu, Z., Lian, L., & Yu, S. X. (2022). Debaised learning from naturally imbalanced pseudo-labels. In L. O'Conner (Ed.), *IEEE/CVF conference on computer vision and pattern recognition, CVPR 2022, new orleans, la, usa, june 18-24, 2022* (pp. 14627–14637). IEEE. Retrieved from <https://doi.org/10.1109/CVPR52688.2022.01424> doi: 10.1109/CVPR52688.2022.01424
- Wang, Z., & Torres, P. E. (2022, July). How to be helpful on online support forums? In E. Clark, F. Brahman, & M. Iyyer (Eds.), *Proceedings of the 4th workshop of narrative understanding (WNU2022)* (pp. 20–28). Seattle, United States: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2022.wnu-1.3> doi: 10.18653/v1/2022.wnu-1.3
- Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., . . . Fedus, W. (2022). Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022. Retrieved from <https://openreview.net/forum?id=yzkSU5zdwD>
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., . . . Zhou, D. (2022). Chain-of-

- thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems 35: Annual conference on neural information processing systems 2022, neurips 2022, new orleans, la, usa, november 28 - december 9, 2022*. Retrieved from http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html
- Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P., ... Gabriel, I. (2021). Ethical and social risks of harm from language models. *CoRR*, *abs/2112.04359*. Retrieved from <https://arxiv.org/abs/2112.04359>
- Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., ... Isaac, W. (2023). Sociotechnical safety evaluation of generative AI systems. *CoRR*, *abs/2310.11986*. Retrieved from <https://doi.org/10.48550/ARXiv.2310.11986> doi: 10.48550/ARXIV.2310.11986
- Weizenbaum, J. (1966). ELIZA - a computer program for the study of natural language communication between man and machine. *Commun. ACM*, *9*(1), 36–45. Retrieved from <https://doi.org/10.1145/365153.365168> doi: 10.1145/365153.365168
- Welivita, A., & Pu, P. (2020). A taxonomy of empathetic response intents in human social conversations. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th international conference on computational linguistics, COLING 2020, barcelona, spain (online), december 8-13, 2020* (pp. 4886–4899). International Committee on Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/2020.coling-main.429> doi: 10.18653/V1/2020.COLING-MAIN.429
- Welivita, A., & Pu, P. (2022). Curating a large-scale motivational interviewing dataset using peer support forums. In N. Calzolari et al. (Eds.), *Proceedings of the 29th international conference on computational linguistics, COLING 2022, gyeongju, republic of korea, october 12-17, 2022* (pp. 3315–3330). International Committee on Computational Linguistics. Retrieved from <https://aclanthology.org/2022.coling-1.293>
- Welivita, A., & Pu, P. (2023). Boosting distress support dialogue responses with motivational interviewing strategy. In A. Rogers, J. L. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: ACL 2023, toronto, canada, july 9-14, 2023* (pp. 5411–5432). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/2023.findings-acl.334> doi: 10.18653/V1/2023.FINDINGS-ACL.334
- Welivita, A., Xie, Y., & Pu, P. (2021). A large-scale dataset for empathetic response generation. In M. Moens, X. Huang, L. Specia, & S. W. Yih (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing, EMNLP 2021, virtual event / punta cana, dominican republic, 7-11 november, 2021* (pp. 1251–1264). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/2021.emnlp-main.96> doi: 10.18653/V1/2021.EMNLP-MAIN.96
- Welivita, A., Yeh, C., & Pu, P. (2023). Empathetic response generation for distress support. In D. Schlangen, S. Stoyanchev, S. Joty, O. Dusek, C. Kennington, & M. Alikhani (Eds.), *Proceedings of the 24th meeting of the special interest group on discourse and dialogue, SIGDIAL 2023, prague, czechia, september 11 - 15, 2023* (pp. 632–644). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/2023.sigdial-1.59> doi: 10.18653/V1/2023.SIGDIAL-1.59

- West, R., & Michie, S. (2016). *A guide to development and evaluation of digital behaviour change interventions in health-care*. Silverback Publishing.
- White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., ... Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. *CoRR*, *abs/2302.11382*. Retrieved from <https://doi.org/10.48550/arXiv.2302.11382> doi: 10.48550/ARXIV.2302.11382
- Williams, G. C., Grow, V. M., Freedman, Z. R., Ryan, R. M., & Deci, E. L. (1996). Motivational predictors of weight loss and weight-loss maintenance. *Journal of Personality and Social Psychology*, *70*(1), 115–126. Retrieved from <https://doi.org/10.1037/0022-3514.70.1.115> doi: 10.1037/0022-3514.70.1.115
- Wilmers, F., Munder, T., Leonhart, R., Herzog, T., Plassmann, R., Barth, J., & Linster, H. W. (2008). Die deutschsprachige Version des Working Alliance Inventory-short revised (WAI-SR)-Ein schulenübergreifendes, ökonomisches und empirisch validiertes Instrument zur Erfassung der therapeutischen Allianz. *Klinische Diagnostik und Evaluation*, *1*(3), 343–358. Retrieved from <https://boris.unibe.ch/27962/> doi: 10.7892/boris.27962
- Wilson, T. D. (1997). Information behaviour: An interdisciplinary perspective. *Inf. Process. Manag.*, *33*(4), 551–572. Retrieved from [https://doi.org/10.1016/S0306-4573\(97\)00028-9](https://doi.org/10.1016/S0306-4573(97)00028-9) doi: 10.1016/S0306-4573(97)00028-9
- Wu, Z. (2023). *Natural language processing for motivational interviewing counselling: Addressing challenges in resources, benchmarking and evaluation* (Unpublished doctoral dissertation). Università degli Studi di Cagliari.
- Wu, Z., Balloccu, S., Helaoui, R., Recupero, D. R., & Riboni, D. (2022, December). Towards in-context non-expert evaluation of reflection generation for counselling conversations. In A. Bosselut et al. (Eds.), *Proceedings of the 2nd workshop on natural language generation, evaluation, and metrics (GEM)* (pp. 116–124). Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2022.gem-1.9> doi: 10.18653/v1/2022.gem-1.9
- Wu, Z., Balloccu, S., Kumar, V., Helaoui, R., Recupero, D. R., & Riboni, D. (2023). Creation, analysis and evaluation of annoMI, a dataset of expert-annotated counselling dialogues. *Future Internet*, *15*(3), 110. Retrieved from <https://doi.org/10.3390/fi15030110> doi: 10.3390/FI15030110
- Wu, Z., Balloccu, S., Kumar, V., Helaoui, R., Reiter, E., Recupero, D. R., & Riboni, D. (2022). Anno-MI: A dataset of expert-annotated counselling dialogues. In *IEEE international conference on acoustics, speech and signal processing, ICASSP 2022, virtual and singapore, 23-27 may 2022* (pp. 6177–6181). IEEE. Retrieved from <https://doi.org/10.1109/ICASSP43922.2022.9746035> doi: 10.1109/ICASSP43922.2022.9746035
- Wu, Z., Balloccu, S., Reiter, E., Helaoui, R., Recupero, D. R., & Riboni, D. (2023). Are experts needed? On human evaluation of counselling reflection generation. In A. Rogers, J. L. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers), ACL 2023, toronto, canada, july 9-14, 2023* (pp. 6906–6930). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/2023.acl-long.382> doi: 10.18653/V1/2023.ACL-LONG.382
- Wu, Z., Helaoui, R., Kumar, V., Recupero, D. R., & Riboni, D. (2020). Towards detecting need for empathetic response in motivational interviewing. In K. P. Truong, D. Heylen, M. Czer-

- winski, N. Berthouze, M. Chetouani, & M. Nakano (Eds.), *Companion publication of the 2020 international conference on multimodal interaction, ICMI companion 2020, virtual event, the netherlands, october, 2020* (pp. 497–502). ACM. Retrieved from <https://doi.org/10.1145/3395035.3425228> doi: 10.1145/3395035.3425228
- Wullach, T., Adler, A., & Minkov, E. (2021). Fight fire with fire: Fine-tuning hate detectors using large samples of generated hate speech. In M. Moens, X. Huang, L. Specia, & S. W. Yih (Eds.), *Findings of the association for computational linguistics: EMNLP 2021, virtual event / punta cana, dominican republic, 16-20 november, 2021* (pp. 4699–4705). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/2021.findings-emnlp.402> doi: 10.18653/V1/2021.FINDINGS-EMNLP.402
- Xiao, B., Bone, D., Segbroeck, M. V., Imel, Z. E., Atkins, D. C., Georgiou, P. G., & Narayanan, S. S. (2014). Modeling therapist empathy through prosody in drug addiction counseling. In H. Li, H. M. Meng, B. Ma, E. Chng, & L. Xie (Eds.), *15th annual conference of the international speech communication association, INTERSPEECH 2014, singapore, september 14-18, 2014* (pp. 213–217). ISCA. Retrieved from <https://doi.org/10.21437/Interspeech.2014-55> doi: 10.21437/INTERSPEECH.2014-55
- Xiao, B., Can, D., Georgiou, P. G., Atkins, D. C., & Narayanan, S. S. (2012). Analyzing the language of therapist empathy in motivational interview based psychotherapy. In *Asia-pacific signal and information processing association annual summit and conference, APSIPA 2012, hollywood, ca, usa, december 3-6, 2012* (pp. 1–4). IEEE. Retrieved from <https://ieeexplore.ieee.org/document/6411762/>
- Xiao, B., Can, D., Gibson, J., Imel, Z. E., Atkins, D. C., Georgiou, P. G., & Narayanan, S. S. (2016). Behavioral coding of therapist language in addiction counseling using recurrent neural networks. In N. Morgan (Ed.), *17th annual conference of the international speech communication association, interspeech 2016, san francisco, ca, usa, september 8-12, 2016* (pp. 908–912). ISCA. Retrieved from <https://doi.org/10.21437/Interspeech.2016-1560> doi: 10.21437/INTERSPEECH.2016-1560
- Xiao, B., Imel, Z. E., Georgiou, P. G., Atkins, D. C., & Narayanan, S. S. (2015). "rate my therapist": Automated detection of empathy in drug and alcohol counseling via speech and language processing. *PLOS ONE*, 10(12), e0143055. Retrieved from <https://doi.org/10.1371/journal.pone.0143055> doi: 10.1371/journal.pone.0143055
- Xie, Y., & Pu, P. (2021). Empathetic dialog generation with fine-grained intents. In A. Bisazza & O. Abend (Eds.), *Proceedings of the 25th conference on computational natural language learning, conll 2021, online, november 10-11, 2021* (pp. 133–147). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/2021.conll-1.10> doi: 10.18653/V1/2021.CONLL-1.10
- Xu, B., & Zhuang, Z. (2022). Survey on psychotherapy chatbots. *Concurr. Comput. Pract. Exp.*, 34(7). Retrieved from <https://doi.org/10.1002/cpe.6170> doi: 10.1002/CPE.6170
- Yalçın, Ö. N. (2019). Evaluating empathy in artificial agents. In *8th international conference on affective computing and intelligent interaction, ACII 2019, cambridge, united kingdom, september 3-6, 2019* (pp. 1–7). IEEE. Retrieved from <https://doi.org/10.1109/ACII.2019.8925498> doi: 10.1109/ACII.2019.8925498
- Yalçın, Ö. N., & DiPaola, S. (2020). Modeling empathy: Building a link between affective and cognitive

- processes. *Artif. Intell. Rev.*, 53(4), 2983–3006. Retrieved from <https://doi.org/10.1007/s10462-019-09753-0> doi: 10.1007/S10462-019-09753-0
- Yao, X., Mikhelson, M., Watkins, S. C., Choi, E., Thomaz, E., & de Barbaro, K. (2023). Development and evaluation of three chatbots for postpartum mood and anxiety disorders. *CoRR*, abs/2308.07407. Retrieved from <https://doi.org/10.48550/arXiv.2308.07407> doi: 10.48550/ARXIV.2308.07407
- Yeo, S., Lim, G., Gao, J., Zhang, W., & Perrault, S. T. (2024). Help me reflect: Leveraging self-reflection interface nudges to enhance deliberativeness on online deliberation platforms. In F. F. Mueller et al. (Eds.), *Proceedings of the CHI conference on human factors in computing systems, CHI 2024, honolulu, hi, usa, may 11-16, 2024* (pp. 806:1–806:32). ACM. Retrieved from <https://doi.org/10.1145/3613904.3642530> doi: 10.1145/3613904.3642530
- Yin, J., Chen, Z., Zhou, K., & Yu, C. (2019). A deep learning based chatbot for campus psychological therapy. *CoRR*, abs/1910.06707. Retrieved from <http://arxiv.org/abs/1910.06707>
- Zhong, P., Zhang, C., Wang, H., Liu, Y., & Miao, C. (2020). Towards persona-based empathetic conversational models. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing, EMNLP 2020, online, november 16-20, 2020* (pp. 6556–6566). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/2020.emnlp-main.531> doi: 10.18653/V1/2020.EMNLP-MAIN.531
- Zhou, J., Deng, J., Mi, F., Li, Y., Wang, Y., Huang, M., ... Meng, H. (2022). Towards identifying social bias in dialog systems: Framework, dataset, and benchmark. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Findings of the association for computational linguistics: EMNLP 2022, abu dhabi, united arab emirates, december 7-11, 2022* (pp. 3576–3591). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/2022.findings-emnlp.262> doi: 10.18653/V1/2022.FINDINGS-EMNLP.262
- Zhou, L., Gao, J., Li, D., & Shum, H. (2020). The design and implementation of xiaoice, an empathetic social chatbot. *Comput. Linguistics*, 46(1), 53–93. Retrieved from https://doi.org/10.1162/coli_a_00368 doi: 10.1162/COLI_A_00368