

# Cooking Up Politeness in Human–AI Information Seeking Dialogue

David Elswailer  
University of Regensburg  
Regensburg, Germany  
david@elsweiler.co.uk

Christine Elswailer  
University of Innsbruck  
Innsbruck, Austria  
Christine.Elsweiler@uibk.ac.at

Anna Ziegner  
University of Innsbruck  
Innsbruck, Austria  
Anna.Ziegner@student.uibk.ac.at

## Abstract

Politeness is a core dimension of human communication, yet its role in human–AI information seeking remains underexplored. We investigate how user politeness behaviour shapes conversational outcomes in a cooking-assistance setting. First, we annotated 30 dialogues, identifying four distinct user clusters ranging from *Hyperpolite* to *Hyperefficient*. We then scaled up to 18,000 simulated conversations across five politeness profiles (including impolite) and three open-weight models. Results show that politeness is not only cosmetic: it systematically affects response length, informational gain, and efficiency. Engagement-seeking prompts produced up to 90% longer replies and 38% more information nuggets than hyper-efficient prompts, but at markedly lower density. Impolite inputs yielded verbose but less efficient answers, with up to 48% fewer nuggets per watt-hour compared to polite input. These findings highlight politeness as both a fairness and sustainability issue: conversational styles can advantage or disadvantage users, and “polite” requests may carry hidden energy costs. We discuss implications for inclusive and resource-aware design of information agents.

## CCS Concepts

• **Information systems** → **Users and interactive retrieval**; • **Human-centered computing** → **Empirical studies in HCI**.

## Keywords

Politeness, Gen-AI, Task-based information seeking, mixed methods

### ACM Reference Format:

David Elswailer, Christine Elswailer, and Anna Ziegner. 2026. Cooking Up Politeness in Human–AI Information Seeking Dialogue. In *2026 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '26)*, March 22–26, 2026, Seattle, WA, USA. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3786304.3787870>

## 1 Introduction

Generative AI (GenAI) systems are increasingly embedded in everyday communication, from customer service and online learning to workplace collaboration and creative writing. As their use expands, they not only shape human communication [21, 42, 46] but are also shaped by the language practices of their users [48]. Optimised to produce coherent and broadly acceptable outputs [28], these systems risk contributing to a homogenisation of language and a

loss of linguistic diversity [24]. Since variation encodes cultural values, social identities, and interactional norms [1, 18], it should be accounted for in human–GenAI exchanges.

A central dimension of such variation is *politeness*, i.e. “how we maintain good relations and avoid interpersonal conflict through the use of different linguistic forms and strategies” [11]. Politeness norms differ across and within language communities [39, 40] and directly shape communication across different settings. In human–AI interactions, studies show that users value polite agents [23] but often adopt more direct styles themselves for efficiency [19, 30]. While human input varies widely, agent responses remain comparatively uniform [3], influencing perceptions of quality. Although prompt politeness has been shown to affect model output [49], we still know little about how politeness unfolds in task-oriented GenAI dialogues where efficiency and clarity are paramount.

At the same time, societal debates have framed politeness as inefficient, since it adds tokens that increase computational cost and energy use [12, 41]. Reducing interaction to machine-optimal forms, however, risks privileging technical constraints over human communicative needs and may disadvantage groups whose styles are underrepresented in training data. This makes politeness not only a question of usability but also of fairness and inclusion. Unlike well-studied demographic and content biases (e.g., gender [20], toxicity [34]), pragmatic variation such as politeness has rarely been studied in task-oriented GenAI contexts.

To address this gap, we focus on *task-based information-seeking dialogues*, where users pursue concrete goals but social aspects of communication remain relevant. Cooking assistance provides an exemplary testbed: it combines structured, stepwise information needs with opportunities for socially oriented behaviours such as thanking, hedging, or apologising. We take a two-step approach. First, we analyse Frummet et al.’s dataset of Wizard-of-Oz cooking conversations [14] to examine how users employ politeness and engagement strategies when interacting with a passive assistant. Second, we extend these insights through LLM–LLM simulation, systematically manipulating user politeness profiles across agent models. Our study addresses three central questions:

- (1) **Response length (RQ1):** Do politeness profiles elicit longer or shorter responses?
- (2) **Information transfer (RQ2):** Do politeness profiles affect the amount of information conveyed per step?
- (3) **Energy efficiency (RQ3):** Do politeness profiles influence the efficiency of information transfer, linking pragmatic variation to debates on sustainable and fair GenAI?

By combining empirical analysis of human data with controlled simulation, we show how politeness strategies impact efficiency and inclusivity in GenAI-mediated cooking tasks. Our findings



This work is licensed under a Creative Commons Attribution 4.0 International License. *CHIIR '26*, Seattle, WA, USA

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2414-5/26/03

<https://doi.org/10.1145/3786304.3787870>

provide both insight into pragmatic variation and practical guidance for designing conversational agents that balance efficiency with fairness.

## 2 Related Work

The related work is structured in two subsections. The first examines prior research on politeness within the field of pragmatics, while the second reviews studies on the linguistic dimensions of conversational information systems. Together, these discussions establish the foundation and rationale for our research agenda.

### 2.1 Variation in linguistic politeness

The systematic study of linguistic politeness in pragmatics, including the use of speech acts realising different politeness strategies, goes back to the seminal work of Lakoff, Brown & Levinson and Leech in the 1970s and 80s [6, 25, 26]. Since then, it has been demonstrated that politeness norms not only vary across different languages such as English and German [5, 16] but also within the same language [38, 40]. Intralingual variation is influenced by both demographic (e.g., region, age or gender) and situational factors (e.g., power, social distance or interactional setting) [4]. Age has, for instance, been shown to impact the level of directness in requests in German [22]. Regional differences are evident in the use of *please*, which is often avoided in American English, particularly when there is power imbalance, but is a routine politeness marker in British English [33]. The impact of communicative setting is evident, e.g., in digital communication, as different digital formats such as online chats, text messaging or email communication engender distinct politeness conventions [15, 31, 44]. These may even vary across languages or national varieties of languages, as the realisation of email requests by British and Australian undergraduate students to faculty highlights [31]. While the British students mark power differences through deference, Australian students treat their lecturers as social peers by emphasising closeness and common ground.

Politeness variation has important implications for human–GenAI interactions. If GenAI systems are more attuned to certain politeness norms than others, they may respond in ways that advantage some user groups while disadvantaging others, even within the same language community. Understanding politeness patterns is therefore critical to designing systems that can interpret and respond appropriately to diverse politeness styles.

### 2.2 Language in Information Interaction

The role of language in human–AI interaction has been studied from several complementary perspectives, with a focus on conversational style, system politeness, and user politeness.

Conversational style has been shown to influence the flow and perceived quality of interactions. Thomas et al. demonstrate that style can be systematically measured in spoken information-seeking dialogues, using features such as pronoun use, repetition, and speech rate [46]. Users were found to adapt their style toward the agent's, while style matching by the agent improved conversational smoothness and reduced the negative effects of errors [45]. More recent work confirms that LLMs likewise adapt syntactically to their interlocutor, aligning their output with user style [17]. The style features adopted by LLMs, e.g., complexity or friendliness,

moreover have an impact on user preferences, which are in turn influenced by user traits such as extroversion and agreeableness [7]. These findings underscore the role of linguistic alignment in human–agent dialogue, but they centre primarily on style rather than politeness as a pragmatic dimension.

System politeness has mostly been studied from the perspective of user perceptions [37] and system responses to input with varying (im)politeness levels [49]. Polite systems are generally perceived as more trustworthy, reliable, and acceptable [37]. LLMs have been shown to recognise (im)politeness, sometimes surpassing human-level performance [2, 29], and to adapt their replies to the politeness level of user input [36]. Politeness also influences system behaviour more subtly: across tasks and languages, polite prompts tend to elicit longer responses [49].

User politeness in human–AI interaction has received comparatively less attention. Existing studies suggest that users differentiate between humans and systems in their use of indirectness, with the use of mitigation strategies reserved for human interlocutors [30]. Perceptions of system intelligence have also been found to shape user behaviour, with participants being more likely to describe themselves as polite toward more advanced AI systems [50]. While some research reduces politeness to the presence of markers such as *please* and *thank you* [3, 50], more detailed research into the impact of gender highlights systematic variation, with women tending towards hedging and men towards direct request strategies in task-oriented ChatGPT interactions [47].

Variation in conversational style, system politeness, and user politeness all shape human–AI interaction, but existing research remains fragmented. Some studies reduce politeness to politeness markers or subjective impressions, and others examine isolated prompts rather than extended information-seeking conversations. Our limited understanding of more comprehensive interactional styles and politeness patterns restrict our ability to effectively and inclusively design GenAI systems. Systems that respond only to particular politeness patterns may advantage some users while disadvantaging others. A systematic investigation of politeness strategies in conversational search settings is therefore needed—an agenda our work aims to advance.

## 3 Analysing Frummet et al's data

In this section, we analyse data from Frummet et al. [14], who investigated user–assistant interaction in a cooking context using a *Wizard-of-Oz* experiment. In the study, participants followed step-by-step recipes via a chat interface and were randomly assigned to either an **active** or **passive** assistant condition. The active assistant proactively offered relevant background knowledge and assistance; the passive assistant responded only to direct user requests.

Participants were recruited via Prolific (native English speakers from the UK and US), and the recipe set consisted of six diverse and instructional meals. The original data, including all user utterances and system responses, were annotated and made publicly available.<sup>1</sup>

In the following subsections we outline our politeness annotation process, then present cluster visualisations, and qualitative descriptions of the clusters.

<sup>1</sup><https://github.com/AlexFrummet/cooking-with-conversations>

### 3.1 Annotation and Reliability

In our analyses, we focus on the 30 conversations in the **passive** condition, as this interaction style reflects how most current GenAI systems operate. In contrast, the active conversations more closely resembled human-human interaction, with the agent deliberately steering the exchange (e.g., prompting the user to consider specific steps in a cooking process, such that their curiosity for further information may be peaked).

We annotated utterances produced by both users and the agent in the passive condition. Each utterance was independently coded for the presence or absence of predefined politeness-related codes by a linguist. These codes were developed based on Leech's [27] classification scheme of politeness-sensitive speech acts, reflecting a goal-oriented approach to politeness, which is appropriate for our task-based cooking conversations with an assistant. Leech's scheme was slightly adapted and extended to capture all speech acts found in the interactions between the participants and the agent. As this study focusses on user politeness behaviour, the user utterances were annotated for the following categories:<sup>2</sup>

- **Requests** are speech acts in which a speaker asks the addressee to perform a future action that entails a cost to them. Since requests restrict the addressee's freedom of action, speakers may soften them, e.g., by presenting them as optional, to enhance their politeness. According to Leech [27], requests comprise a spectrum of directive speech acts, stretching from direct commands to mild suggestions. Since people phrase requests with different levels of directness and in more or less conventional ways, we coded them using five distinct strategy types.
  - (1) **Direct requests** such as commands directly ask for the requested action, typically without mitigation (e.g., *Show me how to chop the onions*).
  - (2) **Non-sentential requests** also directly state the requested action but are conveyed without using a complete sentence (e.g., *Next step*).
  - (3) **Conventionally indirect requests** give the addressee the option not to comply (e.g., *Can you show me...*).
  - (4) **Statement hints** do not directly mention the requested action, which must be inferred (e.g., *I'm ready to start*).
  - (5) **Information-eliciting questions**<sup>3</sup> encompass questions aiming at eliciting information from the interlocutor without asking whether they are able or willing to do this. By signalling the speaker's engagement and interest in continuing the interaction, they can increase politeness and help maintain harmony between interactants [27, 35] (e.g., *When do I add the stock?; Do you have a video for this step?*).
- **Politeness marker *please*** captures the use of *please* to mitigate the imposition of requests (e.g., *Next step, please*).
- **Apologies** intend to restore harmony between the speaker and addressee after an offence perceived as harmful to the speaker. In Frummet et al.'s data, there is only one apology by a participant, realised by *sorry*.

- **Absolving** codes the acknowledgement of apologies by the user by (a) **accepting** them (e.g., *That's ok*), (b) **denying** the reality of the offence (e.g., *No problem*), or (c) **ignoring** the offence due to its triviality (e.g., *Don't mention it*).
- **Thanking** is a face-enhancing expression of gratitude, for instance, for helpful information (e.g., *Thank you for the info*).
- **Compliment/praise/face-enhancing feedback** is an umbrella category comprising face-enhancing speech acts that positively evaluate the agent or its contributions (e.g., *Excellent; That's a great tip; This is really useful*).
- **Acknowledgements of responses**<sup>4</sup> include user reactions signalling receipt and understanding of the agent's response, thus showing their engagement (e.g., *I see; ok*). Acknowledgments including a positive evaluative comment (e.g., *That's helpful*) are additionally coded as face-enhancing feedback.
- **Declining** codes speech acts through which users turn down offers by the agent (e.g., *No thank you*).

In total 14 unique codes were applied to user utterances. Each utterance could be assigned multiple codes if more than one pragmatic function was present.

### 3.2 Validating the Annotation Process

To evaluate inter-rater reliability, a second linguist recoded a random sample of 20% of the data (6 conversations). We computed **Krippendorff's alpha** for each user code. Each utterance was coded in a binary fashion, either as an *application* of the code (present) or a *non-application* (absent), and reliability was calculated across all utterances. The mean  $\alpha = .769$ . All but one code  $\alpha > .60$ . These included *Thanking*, *Politeness marker please*, and *Compliment/praise*, which all reached perfect agreement ( $\alpha = 1.00$ ). The only code  $\alpha < .60$  was *Request - direct*, which showed lower agreement ( $\alpha = .39$ ). This was due to the low overall frequency of direct requests ( $N=6$ ) and a perceived overlap with non-sentential requests. One coder annotated *start* as *Request - direct* and the other as *Request - non-sentential*, an easy issue to resolve.

### 3.3 User Representation and Clustering

For each user, we computed a vector of the absolute frequencies of the selected politeness codes, aggregated across the annotated conversations. These frequencies were standardised (z-scores), producing a matrix, where each row represented a user and each column a code. We applied **k-means clustering** to group users based on their code usage profiles. The optimal number of clusters ( $k$ ) was determined using the **elbow method**, which suggested  $k = 3$  or 4. We opted for  $k = 4$  to better capture nuanced behavioural patterns, despite one cluster containing a single user, because this cluster represents a theoretically meaningful extreme of politeness behaviour that is rare in our small sample but plausibly more prevalent in the wider population.

### 3.4 Visualisation and Interpretation

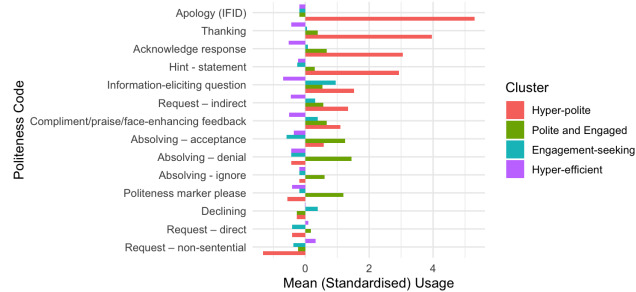
To interpret cluster characteristics, we calculated the **mean standardised usage** of each politeness code per cluster, visualising these means using bar plots to highlight which speech acts were

<sup>2</sup>Full details of the classification scheme can be found in the repository.

<sup>3</sup>Information-eliciting questions have been added to Leech's scheme.

<sup>4</sup>Acknowledgments of responses have been added to Leech's scheme.

more or less prominent in each cluster (Figure 1). This informed our characterisation of clusters based on the politeness and interactional strategies employed. Below, we summarise the main characteristics of each cluster providing examples from the data:



**Figure 1: Z-score standardised code frequency across clusters**

**Hyper-polite cluster.** This cluster is characterised by a very politeness-sensitive and engaged interactional style but it only comprises one participant. Thanking, face-enhancing feedback, acknowledging responses (e.g., *that is interesting; looks good; okay*), and information-eliciting questions (e.g., *do the basil leaves need to be organic?*) are frequently used. Less frequent but notable are indirect requests (e.g., *Can you show me what it would look like?*) and accepting agent apologies (e.g., *ok*). The participant avoids face-threatening direct and non-sentential requests, opting for more indirect statement hints to move along in the recipe instead (e.g., *I'm ready for the next step*) but does not use the politeness marker *please* at all. Notably, this is the only participant who apologises to the agent, in a repair of a preceding information-eliciting question (*what is the quantity of it sorry?*). These patterns suggest a very polite and engaged communication style catering to the addressee's positive face needs by showing an interest in the information they provide and offering face-enhancing feedback. This style is indicative of treating the agent more like a human interlocutor than a machine.

**Polite and engaged cluster.** This cluster is marked by a polite, face-enhancing, and engaged communication style that is employed with moderate frequency. It contains seven participants who fairly regularly use both face-threat mitigating strategies, such as indirect requests (e.g., *can you show me a picture*) and statement hints (e.g., *I think i'm ready to move to the next step*), and face-enhancing strategies, including acknowledging responses, thanking and compliments/face-enhancing feedback (e.g., *Awesome, thank you*), as well as absolving reactions to agent apologies. Notably, they use absolving strategies more often and with greater variation than other clusters (accepting (e.g., *that's okay*), ignoring the offence (e.g., *Doesn't matter*), denying the reality of the offence (e.g., *no worries*)). The politeness marker *please* is also more common than in any other cluster, often paired with indirect requests to further mitigate imposition (e.g., *can you tell me about the history of this recipe please?*) and with non-sentential requests to proceed (e.g., *Thanks, next step please*). Overall, the participants exhibit a polite, respectful and appreciative communication style, showing an interest in additional information.

**Engagement-seeking cluster.** The interactional style in this cluster is distinguished by a balance of politeness and efficiency. It includes six participants who frequently engage with the agent, particularly through varied information-eliciting questions (e.g. *where is the note about pecorino fiore sardo; Do I need to remove the zest?*), alongside a measured use of thanking, acknowledging responses, indirect requests, and face-enhancing feedback. These users tend to avoid statement hints, absolving responses to apologies, and direct requests, while using non-sentential requests to a limited degree. These rarely combine with the politeness marker *please*, which is not used in any other context. Their communication style reflects a preference for smooth, efficient interactions with high engagement and the occasional use of face-enhancing strategies.

**Hyper-efficient cluster.** This cluster exhibits low engagement and minimal use of politeness-related strategies such as thanking, acknowledging responses, face-enhancing feedback, or absolving reactions to apologies. It comprises 16 participants who use the fewest information-eliciting questions among all clusters and frequently rely on non-sentential requests (e.g., *next*) to move the process along, usually without mitigating devices like the politeness marker *please*. While they occasionally use indirect requests, they also sometimes resort to direct requests (e.g., *Remind me which types of cheese to use*). Their communication style is direct, task-oriented, and efficiency-driven, reflecting a tendency to treat the agent as a machine rather than a human conversational partner.

The politeness patterns manifested in the clusters represent a continuum from hyper-polite to hyper-efficient with two intermediate profiles reflecting distinct interactional goals and behaviour.

## 4 Simulating (Im)polite Conversations

To learn whether different (im)politeness strategies shape interactions with generative AI systems we performed a pre-registered experiment<sup>5</sup> using a controlled simulation in which two LLMs interacted: one acting as a *user* and the other as the *agent*. The following sub-sections outline the precise method employed, the hypotheses tested, the analysis plan, and the results.

### 4.1 Simulation Approach

**Recipe task domain.** We adopted the same six recipes used by Frummet et al. [14], all sourced from SeriousEats (<https://www.seriouseats.com>): *Parisian Gnocchi*, *Buttermilk-Brined Southern Fried Chicken*, *Duck à l'Orange*, *Savory Cheese Soufflé*, *Pesto alla Genovese*, and *Old-Fashioned Apple Pie*. Each recipe was decomposed into stepwise instructions (e.g., *score the duck skin and season with salt*). For each step we generated a set of neutrally phrased information needs spanning three categories: (1) task-related (e.g., *How deep should the scoring be?*), (2) science-related (e.g., *How does scoring influence fat rendering?*), and (3) history-related (e.g., *When did béchamel originate?*). These needs served as content “slots” that could be realised in different pragmatic styles.

**User profiles.** To operationalise variation in politeness, we defined multiple user profiles derived from our corpus analysis (i.e.,

<sup>5</sup>The pre-registration can be found here: <https://osf.io/mp3dt/overview>

*Hyperpolite, Polite and engaged, Engagement-seeking, and Hyper-efficient*). We additionally constructed a fifth profile, *Impolite*. Although we did not observe impolite behaviour in the naturalistic conversations collected by Frummet et al., there is evidence of users interacting impolitely with GenAI systems [47]. We added this profile to test how agents would react to impolite interactions in this context. The impoliteness features for this profile are based on a selective combination of Culpeper's impoliteness output strategies [8] and conventional impoliteness formulae [9, 10] that are relevant to the task-based cooking interaction between a user and an agent. The *Impolite* cluster is marked by face-threatening acts in the shape of imperatives and brusque commands and instances of withholding politeness through a lack of thanking, acknowledgments, or compliments. It moreover includes criticisms and disagreements (e.g., *This is irrelevant!*) as well as insults and taboo words (e.g., *You idiot, That is total crap!*) and is characterised by frequent abrupt moves to the next step without engagement. Each profile was characterised by probabilistic tendencies to use particular politeness strategies. These tendencies were described to the user-LLM in its system prompt, along with instructions to output turns in valid JSON containing: the utterance, the politeness codes expressed, the pragmatic intent, and whether the user wished to proceed to the next recipe step. This allowed us both to guide the LLM toward a target style and to log the realised behaviour.

*Agent behaviour.* The agent-LLM received the recipe step, the user's most recent utterance, and the conversation history. Its prompt instructed it to act as a knowledgeable cooking assistant, answering task, science, or history questions as they arose. In approximately fifteen percent of cases, particularly for science or history needs, the agent was prompted to *fail gracefully* by apologising for not having an answer. This enabled us to simulate how users in different clusters responded to agent apologies. Like the user, the agent's output was constrained to JSON, containing the text of its reply and a list of politeness codes expressed in the turn.

*Conversation flow.* Each recipe step was simulated as a short dialogue loop. The user-LLM produced an utterance guided by its politeness profile and the selected information need. The agent-LLM then responded with an answer, or, in some cases, with an apology. The loop continued until either the user signalled that they wished to proceed or a maximum of two questions had been asked about that step. This mirrors the conversations in Frummet et al.'s data. All turns were stored with their associated metadata, including a chronological event index.

*Data collection.* The simulation produced structured dialogue logs in tabular form. Each row corresponds to one utterance, annotated with: recipe step, need type, role (user vs. agent), text, pragmatic intent, politeness codes, and, where relevant, apologies. These logs allow us to (a) validate whether the generated user behaviours cluster according to the intended politeness profiles and (b) quantify the amount of information transferred and energy used under different conditions.

*Models.* We selected three open-weight instruction-tuned models in the 7–8B parameter class: *DeepSeek-8B*, *Llama-3.1-8B-Instruct*, and *Qwen-2.5-7B-Instruct*. This size range ensures comparability in computational demands while avoiding scale effects. The models

represent distinct training lineages and design aims: DeepSeek is a distilled model optimised for reasoning and instruction following [13], Llama-3.1 reflects Meta's broadly aligned and stylistically adaptive approach to multilingual assistant chat [32], and Qwen-2.5 incorporates extensive multilingual and cross-cultural training data with supervised fine-tuning and RLHF [43]. This diversity makes them well suited for investigating how user politeness strategies influence model responses.

Full code, including user- and agent-prompts, as well as the conversations themselves can be found in our repository.

*Validation.* To validate the approach, a set of 300 test conversations were simulated and 50 sampled to cover all 5 politeness profiles. Two linguists independently labelled the conversations based on the politeness profiles (clusters) employed by the user bot. The annotators achieved an accuracy 92% and 96%, respectively, which we accept as evidence for the simulations of the clusters behaving authentically.

*Sample size.* For the final experiment, we generated 18,000 conversations (5 politeness profiles  $\times$  6 recipes  $\times$  3 models  $\times$  200 conversations per cell). To determine this number we conducted Monte Carlo power simulations using conversation-level negative binomial models. Parameters calibrated from the 300 test conversations generated step-level data with a conversation random effect, then we aggregated to conversation-level for analysis. Power was the proportion of simulated datasets with a significant cluster  $\times$  model term ( $\alpha = .05$ ). Power simulations indicated that 200 conversations per cluster–recipe–model cell achieved 83% power (95% CI [77%, 90%]) to detect the planned interaction ( $\sim \pm 2\%$ ).

## 4.2 Hypotheses

Based on prior work in pragmatics and conversational style, as well as our initial simulations, we expect systematic differences across politeness profiles in how conversations unfold. Each hypothesis operationalises one research question (RQ1–RQ3).

*H1: Response length.* The length of agent responses (in tokens) will vary across politeness profiles. Polite or hyperpolite inputs may encourage more elaborated responses, whereas efficient and direct or impolite inputs may yield shorter, more task-focused answers.

*H2: Information transfer.* The number of new “information nuggets” per recipe step—defined as atomic, useful pieces of information not previously mentioned in the same utterance—will vary across politeness profiles.

*H3: Energy efficiency.* The efficiency of information transfer, operationalised as the ratio of information nuggets to energy consumption (Wh), will vary across politeness profiles.

## 4.3 Analysis Plan

The presented analyses follow the pre-registered analysis plan. Below we outline the dependent variables and how they are measured:

*Response length.* We measure the token length of agent responses in words.

*Information transfer.* To measure the amount of information conveyed, we adopted a nugget-counting approach similar to Frummet

et al., where an LLM is used to identify and count minimal factual units (“nuggets”). In our implementation, agent utterances were concatenated per step, which required handling duplicate information. We employed a few-shot prompting setup with the deepseek model to generate nugget counts.

We validated the nugget counter against a manually annotated test set ( $N = 158$  steps) with two coders, covering edge cases such as no information (e.g., apologies), duplicated content (rephrasings), mixed procedural and declarative knowledge, and high variability in information density. Model outputs correlated well with human judgments ( $r = .70$ ,  $MAE = 0.80$ ,  $bias = -0.22$ ) and showed excellent consistency across three runs ( $ICC(C, 3) = 0.994$ , 95% CI  $[0.992, 0.995]$ ); for comparison, inter-coder agreement was  $r = .73$ . This indicates that the metric is both valid and reliable, providing a stable basis to compare information density across conditions.

**Energy efficiency.** We compute an efficiency score (nuggets per Wh consumed).

## 5 Results

Following the pre-registered analysis plan, we analyse step-level outcomes with mixed-effects models that include *Cluster* (user politeness profile) and *AgentModel* as fixed factors (controlling for *Recipe*) and a random intercept for *Step* to account for the repeated measure (steps within a conversation). For H1–H3, we first run omnibus tests (e.g., ANOVA or analogue on the fitted mixed models) to determine whether *Cluster*, *AgentModel*, or their interaction explains outcome variance. If this is significant we conduct follow up pair-wise comparisons of estimated marginal means with Tukey adjustment. For accessibility, each hypothesis subsection is structured as follows: 1. Non-technical summary of the main take-away message, 2. Detailed statistical methods, 3. Detailed findings and commentary.<sup>6</sup>

### 5.1 H1: Do user politeness profiles change how long agents reply?

**Take-away.** Politeness profiles systematically change reply length: *Engagement-seeking* (ES) prompts elicit the longest answers; *Hyper-efficient* (HE) the shortest; *Polite and engaged* (PE) and *Hyperpolite* (HP) sit in between, with *Impolite* (IMP) typically shorter than PE.

**What we tested & why.** We modelled reply length at the step level using a linear mixed-effects model on  $\log(\text{words} + 1)$  with *Cluster*, *AgentModel*, and *Recipe* as fixed effects and a random intercept for *Step* (steps within a conversation are not independent). *Recipe* controls for content differences. Lengths are right-skewed; the log transform reduces the influence of extremes and makes effects interpretable as proportional (percentage) differences. For interpretability we report back-transformed estimated marginal means (EMMs) as *words per step*.

**Omnibus tests.** Mixed-effects ANOVA found reliable effects: *Cluster*:  $F_{4,16029} = 4450.29$ ,  $p < .001$ ; *AgentModel*:  $F_{2,16668} = 3981.49$ ,  $p < .001$ ; *Cluster*×*AgentModel*:  $F_{8,16026} = 47.16$ ,  $p < .001$ .

<sup>6</sup>For space reasons only key details are reported here. Full models and descriptives are available in our repository. [https://github.com/delsweil/CHIIR2026\\_politeness](https://github.com/delsweil/CHIIR2026_politeness)

Table 1 shows back-transformed EMMs for words (H1). Qwen produced on average shorter responses than the other models. However, across agent models consistent trends are apparent: ES has the longest replies and HE the shortest, with large within-model differences (see Table 2 for illustrative examples of how deltas vary across models and profiles).

### 5.2 H2: Do politeness profiles change how many nuggets agents convey?

**Take-away:** ES prompts yield the most nuggets per step, HE the fewest; PE generally exceeds IMP. However, *percentage increases in words are much larger than percentage increases in nuggets*. Some agents appear less open to influence from impoliteness on *information quantity*, even when their *reply length* still shifts.

**What we tested & why.** Nuggets are non-negative counts and were overdispersed, so we used a negative binomial mixed model (NB-GLMM) with *Cluster* and *AgentModel* as fixed effects, *Recipe* as a control, and a random intercept for *Conversation*.

**Omnibus tests.** Omnibus effects were tested via likelihood-ratio  $\chi^2$  tests comparing nested NB-GLMMs ( $\Delta\text{deviance} = 2 \Delta \log\text{-likelihood}$ ; df equals the number of fixed-effect parameters removed). Results indicated strong main effects of *Cluster* and *AgentModel*, and a reliable *Cluster*×*AgentModel* interaction: *Cluster*:  $\chi^2_{(4)} = 1895.38$ ,  $p < .001$ ; *AgentModel*:  $\chi^2_{(2)} = 1087.78$ ,  $p < .001$ ; *Cluster*×*AgentModel*:  $\chi^2_{(8)} = 239.22$ ,  $p < .001$ .

Table 1 reports EMMs for *nuggets per step* on the original count scale. The pattern is consistent across agent models: ES produces the most nuggets, HE the least, and PE exceeds IMP. Magnitudes vary by agent model, but the direction is stable. The overview in Table 2 quantifies these gaps, which can be up to 80.5% in words and 38.2% in nuggets.

**Pairwise exceptions (after Tukey).** Most cluster comparisons were significant after Tukey adjustment. The non-significant pairs were: llama: HP vs. IMP; qwen: PE vs. IMP. This may suggest that, for these models, *impoliteness did not reliably reduce the amount of information (nuggets per step)* relative to those polite styles. We interpret this as *reduced sensitivity to impoliteness in information quantity*, with three caveats: (i) this is *outcome-specific* (H1 shows length still changes, and H3 addresses efficiency); (ii) it is *pair-specific* (other IMP comparisons were significant); and (iii) non-significant  $\neq$  equivalence.

**Words vs. nuggets.** Our analyses in this subsection highlight that, across models, moving from HE to ES yields *much larger percentage increases in words* than in nuggets: roughly 60–90% more words for only about 20–40% more nuggets (see Table 2). Likewise, IMP often elicits *longer* replies than PE with only minimal nugget differences (approximately −4% to +8%). In short, *length grows faster than information*. This relates to H3, where we analyse *information density* and *energy efficiency* (nuggets per word; nuggets per Wh) in detail.

**Table 1: Back-transformed expected words per step (H1) and nuggets per step (H2), plus nuggets per 100 words (H3a) and nuggets per Wh (H3b), by Cluster  $\times$  Agent. 95% CIs shown as [LCL:UCL]. Estimates adjust for recipe and conversation clustering.**

| Agent    | Cluster            | Words [LCL:UCL]  | Nuggets [LCL:UCL] | Nuggets/100w [LCL:UCL] | Nuggets/Wh [LCL:UCL] |
|----------|--------------------|------------------|-------------------|------------------------|----------------------|
| deepseek | Hyperpolite        | 69.9 [69.1:70.7] | 2.03 [2.00:2.06]  | 1.61 [1.59:1.64]       | 1.654 [1.629:1.680]  |
| deepseek | Polite-Engaged     | 50.2 [49.5:50.9] | 1.87 [1.84:1.91]  | 1.92 [1.88:1.95]       | 1.979 [1.941:2.017]  |
| deepseek | Engagement-Seeking | 76.9 [76.0:77.8] | 2.16 [2.13:2.19]  | 1.61 [1.58:1.63]       | 1.598 [1.574:1.622]  |
| deepseek | Hyperefficient     | 42.6 [42.1:43.1] | 1.59 [1.56:1.62]  | 1.93 [1.90:1.96]       | 1.858 [1.826:1.891]  |
| deepseek | Impolite           | 66.2 [65.5:67.0] | 1.73 [1.70:1.76]  | 1.40 [1.37:1.42]       | 1.331 [1.309:1.353]  |
| llama    | Hyperpolite        | 67.9 [67.1:68.7] | 2.11 [2.07:2.14]  | 1.79 [1.76:1.81]       | 2.040 [2.010:2.071]  |
| llama    | Polite-Engaged     | 53.2 [52.5:54.0] | 2.03 [1.99:2.07]  | 2.15 [2.11:2.19]       | 2.527 [2.481:2.574]  |
| llama    | Engagement-Seeking | 69.8 [69.0:70.6] | 2.26 [2.23:2.29]  | 1.91 [1.88:1.94]       | 2.075 [2.045:2.105]  |
| llama    | Hyperefficient     | 42.8 [42.3:43.3] | 1.87 [1.84:1.90]  | 2.51 [2.47:2.55]       | 2.702 [2.659:2.746]  |
| llama    | Impolite           | 63.2 [62.4:63.9] | 2.11 [2.08:2.14]  | 1.94 [1.91:1.97]       | 2.005 [1.975:2.035]  |
| qwen     | Hyperpolite        | 53.0 [52.4:53.6] | 1.93 [1.90:1.96]  | 2.08 [2.05:2.11]       | 1.968 [1.937:1.999]  |
| qwen     | Polite-Engaged     | 38.3 [37.7:38.8] | 1.68 [1.65:1.72]  | 2.39 [2.34:2.44]       | 2.251 [2.206:2.297]  |
| qwen     | Engagement-Seeking | 57.5 [56.8:58.1] | 2.02 [1.98:2.05]  | 2.03 [2.00:2.06]       | 1.907 [1.878:1.936]  |
| qwen     | Hyperefficient     | 30.3 [30.0:30.7] | 1.46 [1.43:1.49]  | 2.45 [2.40:2.49]       | 2.144 [2.105:2.183]  |
| qwen     | Impolite           | 44.4 [43.9:44.9] | 1.63 [1.61:1.66]  | 1.98 [1.95:2.02]       | 1.642 [1.614:1.670]  |

**Table 2: Within-agent gaps across outcomes for two illustrative pairwise comparisons. Entries show  $\Delta$  and % (relative to the second cluster). H1 = words; H2 = nuggets; H3a = nuggets/100w; H3b = nuggets/Wh.**

| Agent    | ES-HE (W; N; N/100w; N/Wh)                                                       | PE-IMP (W; N; N/100w; N/Wh)                                                    |
|----------|----------------------------------------------------------------------------------|--------------------------------------------------------------------------------|
| deepseek | W: 34.3 (80.5)<br>N: 0.57 (35.9)<br>N/100w: -32.3 (-16.7)<br>N/Wh: -0.26 (-14.0) | W: -16.1 (-24.2)<br>N: 0.14 (8.2)<br>N/100w: 51.9 (37.1)<br>N/Wh: 0.65 (48.7)  |
| llama    | W: 27.0 (63.2)<br>N: 0.39 (20.7)<br>N/100w: -59.9 (-23.8)<br>N/Wh: -0.63 (-23.2) | W: -9.9 (-15.7)<br>N: -0.08 (-3.6)<br>N/100w: 20.6 (10.6)<br>N/Wh: 0.52 (26.0) |
| qwen     | W: 27.1 (89.4)<br>N: 0.56 (38.2)<br>N/100w: -41.8 (-17.1)<br>N/Wh: -0.24 (-11.1) | W: -6.2 (-13.9)<br>N: 0.05 (3.1)<br>N/100w: 40.3 (20.3)<br>N/Wh: 0.61 (37.1)   |

### 5.3 H3a: Do politeness profiles change information density?

**Takeaway.** Hyperefficient (HE) prompts yield the *densest* replies (more nuggets per 100 words), Engagement-seeking (ES) the *least dense*; Polite and engaged (PE) is consistently denser than Impolite (IMP).

**What we tested & why.** We analysed step-level *nugget counts* with a negative-binomial mixed model (log link). We included log(words) as an *offset*, which means effects are estimated on a *rate* scale (how many nuggets are produced *per word*) rather than on raw counts. Fixed effects were the *Cluster  $\times$  AgentModel* interaction and *Recipe* (as a control), and we added a random intercept for *Step* to account for repeated steps within conversations. Exponentiating coefficients gives *incidence-rate ratios*: values >1 indicate more nuggets per word than the reference level; values <1 indicate fewer.

Table 1 shows nuggets per 100 words by Cluster  $\times$  Agent. Across agents, HE has the highest density, while ES has the lowest. Within-agent deltas (Table 2) quantify the gaps. In short, ES tends to be

*word-inefficient*, while HE/PE concentrate information better.

**Omnibus Tests.** Likelihood-ratio  $\chi^2$  tests show significant effects for Cluster:  $\chi^2_{(4)} = 2097.36, p < .001$ ; AgentModel:  $\chi^2_{(2)} = 2633.61, p < .001$ ; Cluster $\times$ AgentModel:  $\chi^2_{(8)} = 335.62, p < .001$ .

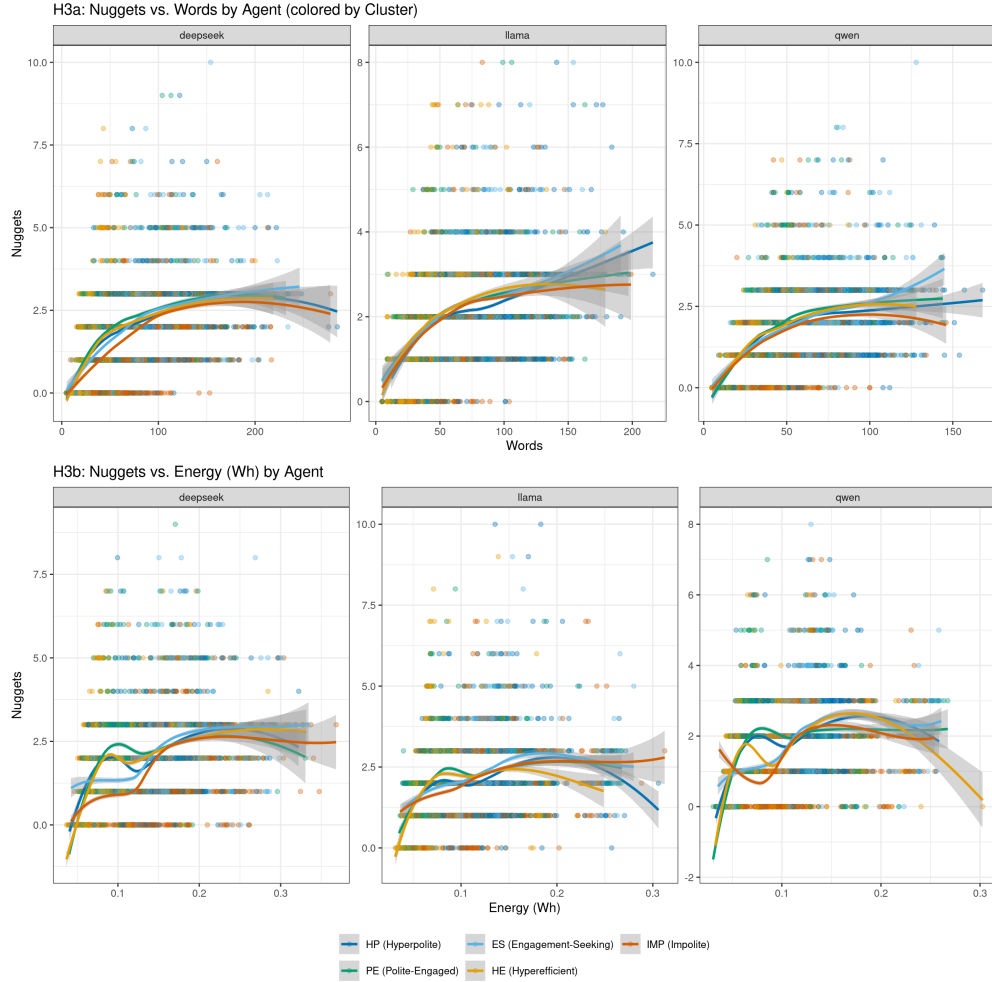
**Pairwise exceptions (Tukey-adjusted).** All but 6 density comparisons were significant. Non-significant pairs were: deepseek: HE vs. PE; HP vs. ES. llama: ES vs. IMP. qwen: ES vs. IMP; HE vs. PE; HP vs. ES. Unlike previous tests, there is no clear pattern here. Again we note the non-significant  $\neq$  equivalence.

*Note:* Our models fit means, however the LOESS curves in Fig. 2 show that nuggets rise steeply over the first tens of words and then flatten. The plateau occurs at slightly different points for different models, but after 50-100 words, additional text yields progressively smaller gains (often less than half a nugget), indicating diminishing returns in informational yield per extra word. Clusters differ mainly by the amount of information they provide at a given length, not by the presence of a different slope.

### 5.4 H3b: Do politeness profiles change energy efficiency?

**Take-away.** IMP is the least energy-efficient across agents. PE reliably beats IMP; HE and PE are the most efficient; ES is typically less energy-efficient than HE.

**What we tested & why.** We modelled the *nugget count* per step with a negative-binomial mixed model (log link) and included log(Wh) as an *offset*. The offset treats watt-hours as the exposure, so all effects are estimated on a *rate* scale (i.e., how many nuggets are produced *per unit of energy*). In other words, the model asks: holding energy usage constant, do different user clusters and agent models yield different *nuggets/Wh*? We used negative binomial because nugget counts are overdispersed (variance exceeds the mean).



**Figure 2: H3a/H3b efficiency diagnostics. Top: nuggets vs. words; bottom: nuggets vs. energy (Wh). Columns are agents; colours denote user clusters. LOESS smooths with 95% CIs over step-level points. X-axes are free per column to reflect very different scales. Patterns show early gains followed by plateaus (diminishing returns).**

Fixed effects were the  $Cluster \times AgentModel$  interaction (our focal factor) and (as in previous models) *Recipe* as a control for content differences. We added a random intercept for *Step*.

Table 1 reports nuggets/Wh across conditions, demonstrating small but significant absolute differences. Overall, *llama* is the most efficient model with the mean nuggets/Wh ranging from 2.005(IMP)–2.574(PE), compared to 1.331(IMP)–1.97(PE) for *deepseek*. Within-agent deltas (Table 2) highlight two stable patterns: ES vs HE: –14.0% (deepseek), –23.2% (llama), –11.1% (qwen); PE vs IMP: +48.7% (deepseek), +26.0% (llama), +37.1% (qwen). Thus, impoliteness is energy-inefficient; HE/PE deliver more nuggets per Wh.

**Omnibus Tests.** Likelihood-ratio  $\chi^2$  tests show significant effects for Cluster:  $\chi^2_{(4)} = 2705.87, p < .001$ ; AgentModel:  $\chi^2_{(2)} = 3345.16,$

$p < .001$ ; Cluster $\times$ AgentModel:  $\chi^2_{(8)} = 303.43, p < .001$ .

**Pairwise exceptions (Tukey-adjusted).** Again, most efficiency comparisons were significant, with only 6 non-significant pairs occurring, without any obvious trend.

*Note:* Fig. 2 shows a similar non-linearity for energy to that observed for information density in Section 5.3. We see modest increases in Wh initially buy noticeably more nuggets, but growth slows and can plateau at higher energy budgets. This suggests that extra computation tends to lead to verbosity, formatting, or hedging rather than new factual units once the main answers to questions relating to a step have been covered. The differences between clusters thus reflect how efficiently they turn computations into nuggets, not merely how much energy they use.

## 6 Discussion

Our two-staged study offers a first look at how politeness strategies shape interactions with generative AI in the context of task-oriented information seeking dialogues. Annotating 30 cooking dialogues with an extended version of Leech’s classification scheme for politeness-sensitive speech acts, we found users cluster along a spectrum from hyperpolite to hyperefficient, with intermediate styles reflecting distinct goals. Scaling up to 18k simulated dialogues, we show that politeness systematically influences reply length, informational yield, and energy use, suggesting that conversational style can have measurable effects on both content delivery and resource costs. Switching profiles can more than double the output length (+60–90%) and increase information nuggets (+20–40%), but at clear efficiency costs. Engagement-heavy styles give more but waste more, while impoliteness risks verbosity without value. The results also hint at diminishing returns: more words and watts eventually add little new information.

We begin by highlighting inclusion and environmental motivations for our work. Next, we revisit our results in light of these motivations and discuss their implications for the design of task-oriented information seeking, as well as directions for future work.

### 6.1 Inclusion & fairness

*Different styles, different rewards.* The different politeness strategies we observed resulted in consistent variation in outcomes. ES/HP users receive more informational units, but at higher costs in words and energy. HE users receive fewer units, but more concisely. While nugget counts do not equate to task utility, these systematic differences raise fairness concerns if certain user styles are consistently advantaged or disadvantaged. Politeness is culturally ingrained and varies according to demographic and situational factors. If systems do reward certain behaviours over others, this could amplify existing inequities.

*Accessibility & cognitive load.* Longer responses are not necessarily better. Even if they contain more information, users with time, attention, or bandwidth constraints may be penalised by verbosity.

*Sensitivity to impoliteness.* In our synthetic impolite profile, we observe clear disadvantages, though some models’ nugget counts are comparatively less sensitive to IMP prompts. We consider this as a positive inclusion signal: users should not have to ‘pay’ with politeness for access to information, particularly in high-stress or direct-speech contexts. At the same time, style-robustness and safety must co-exist: the aim is parity of informational help across everyday registers, not higher tolerance for harassment. Future work may wish to formalise a *Style-Response Parity* metric (e.g., nugget parity ratios and efficiency parity across clusters) and test interventions that improve parity without inflating costs or weakening guardrails.

### 6.2 Environmental & resource costs

*Polite polluters.* We found significant differences in the energy costs of different politeness strategies, with the more engaging strategies often being less efficient and more energy-hungry. While the effect sizes in absolute terms are small, these will scale, and will undoubtedly be amplified by larger models.

*Diminishing returns.* We observed consistent plateaus across models and politeness profiles, showing that additional words or energy provide ever smaller information gains. This suggests an opportunity to limited verbosity once the plateau begins, without reducing nuggets.

### 6.3 Design & Policy Implications

We identify a number of aspects for system designers and service providers to consider based on our findings:

*Parity across styles.* Systems should aim for a comparable information density between styles to avoid politeness becoming a hidden ‘tax’ for access to detail.

*Preference-aware generation.* Interfaces may allow users to set persistent preferences for response length or efficiency, ensuring that their chosen style, whether brief or elaborate, does not disadvantage them.

*Explanatory transparency.* Compact, human-readable rationales (e.g., “This reply is concise; click to expand”) may reduce pressure on users to adopt more engaging styles merely to obtain fuller answers.

*Guardrails for impoliteness.* If impolite prompts tend to inflate reply length without adding value, policy should limit verbosity by summarising content and offering structured follow-ups. This could prevent inefficiency spirals while maintaining access to useful information.

### 6.4 Limitations

Our study is limited to cooking conversations as an example of step-wise, task-oriented interaction. The behaviours we observed come from a small unrepresentative sample and the clusters we found were uneven in size. Generalisation, especially to other domains (e.g., learning, troubleshooting, health), remains to be validated. Moreover, information nuggets do not necessarily equate to utility. Future work should validate that nuggets correlate with task success, accuracy, and user satisfaction. Finally, for resource reasons, we only studied compact LLMs. Future work should aim to replicate with larger and multilingual models to test generalisability.

## 7 Conclusions and Future work

We studied how user politeness profiles shape GenAI behaviour in task-oriented cooking dialogues, combining empirical analysis of human interactions with large-scale LLM-LLM simulation. Our findings demonstrate that users’ conversational choices, alongside model characteristics, systematically affect response length, information transfer, and energy efficiency. This highlights important implications for system and policy design, suggesting that politeness can act as a hidden cost factor with consequences for both inclusion and sustainability. A deeper understanding of user motivations, including social needs, response preferences, and cross-cultural variation, is needed to guide the development of GenAI information seeking systems that are both inclusive and sustainable.

## References

- [1] Asif Agha. 2003. The social life of cultural value. *Language & communication* 23, 3-4 (2003), 231–273.
- [2] Marta Andersson and Dan McIntyre. 2025. Can ChatGPT recognize impoliteness? An exploratory study of the pragmatic awareness of a large language model. *Journal of Pragmatics* 239 (2025), 16–36.
- [3] Sabrina Barko-Sherif, David Elsewiler, and Morgan Harvey. 2020. Conversational agents for recipe recommendation. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. 73–82.
- [4] Anne Barron. 2021. *Synchronic and Diachronic Pragmatic Variability*. Cambridge University Press, 182–205.
- [5] Shoshana Blum-Kulka, Juliane House, and Gabriele Kasper (Eds.). 1989. *Cross-Cultural Pragmatics: Requests and Apologies*. Ablex Publishing Corporation, Norwood, NJ.
- [6] Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*. Vol. 4. Cambridge university press.
- [7] Rendi Chevi, Kentaro Inui, Thamar Solorio, and Alham Fikri Aji. 2025. How individual traits and language styles shape preferences in open-ended user-llm interaction: A preliminary study. *arXiv preprint arXiv:2504.17083* (2025).
- [8] Jonathan Culpeper. 1996. Towards an anatomy of impoliteness. *Journal of pragmatics* 25, 3 (1996), 349–367.
- [9] Jonathan Culpeper. 2011. *Impoliteness: Using language to cause offence*. Vol. 28. Cambridge University Press.
- [10] Jonathan Culpeper and Claire Hardaker. 2017. Impoliteness. In *The Palgrave handbook of linguistic (im) politeness*. Springer, 199–225.
- [11] Jonathan Culpeper, Michael Haugh, and Dániel Z. Kádár. 2017. Introduction. In *The Palgrave Handbook of Linguistic (Im)politeness*, Jonathan Culpeper, Michael Haugh, and Dániel Z. Kádár (Eds.). Palgrave Macmillan, London, 1–8.
- [12] Jules de Lavande. 2025. *How much energy does saying "thank you" to AI consume?* <https://huggingface.co/blog/jdelavande/thank-you-energy> Accessed: 2025-08-09.
- [13] DeepSeek-AI. 2024. DeepSeek-R1 Distill-Llama-8B. <https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-8B>
- [14] Alexander Frummet, Alessandro Spezzigiorin, David Elsewiler, Anton Leuski, and Jeff Dalton. 2024. Cooking with conversation: Enhancing user engagement and learning with a knowledge-enhancing assistant. *ACM Transactions on Information Systems* 42, 5 (2024), 1–29.
- [15] Sage L Graham and Claire Hardaker. 2017. (Im) politeness in digital communication. In *The Palgrave handbook of linguistic (im) politeness*. Springer, 785–814.
- [16] Juliane House and Dániel Z Kádár. 2021. *Cross-cultural pragmatics*. Cambridge University Press.
- [17] Florian Kandra, Vera Demberg, and Alexander Koller. 2025. LLMs syntactically adapt their language use to their conversational partner. *arXiv preprint arXiv:2503.07457* (2025).
- [18] Scott F Kiesling. 2013. Constructing identity. *The handbook of language variation and change* (2013), 448–468.
- [19] Krishna Kodur, Manizheh Zand, Matthew Tognotti, Sean Banerjee, Natasha Kholgade Banerjee, and Maria Kyrarini. 2025. Exploring the Dynamics of Human-Robot Interaction: Robot Error, Sentiment Analysis, and Politeness. *International Journal of Social Robotics* (2025), 1–14.
- [20] Hadas Koteck, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*. 12–24.
- [21] Tom Kouwenhoven, Max Peepkorn, Roy De Kleijn, and Tessa Verhoef. 2025. Shaping Shared Languages: Human and Large Language Models' Inductive Biases in Emergent Communication. *arXiv preprint arXiv:2503.04395* (2025).
- [22] Svenja Kranich, Hanna Bruns, and Elisabeth Hampel. 2021. Requests across varieties and cultures: Norms are changing (but not everywhere in the same way). *Anglistik: International Journal of English Studies* 32, 1 (2021), 91–114.
- [23] Shikhar Kumar, Eliran Itzhak, Yael Edan, Galit Nimrod, Vardit Sarne-Fleischmann, and Noam Tractinsky. 2022. Politeness in human-robot interaction: a multi-experiment study with non-humanoid robots. *International Journal of Social Robotics* 14, 8 (2022), 1805–1820.
- [24] Maria Kuteeva and Marta Andersson. 2024. Diversity and Standards in Writing for Publication in the Age of AI—Between a Rock and a Hard Place. *Applied Linguistics* 45, 3 (2024), 561–567.
- [25] Robin Lakoff. 1973. The logic of politeness: Or, minding your p's and q's. In *Proceedings from the annual meeting of the Chicago Linguistic Society*, Vol. 9. Chicago Linguistic Society, 292–305.
- [26] Geoffrey N. Leech. 1983. *Principles of pragmatics* (1. publ. ed.). Longman, London.
- [27] Geoffrey N Leech. 2014. *The pragmatics of politeness*. Oxford University Press.
- [28] Christoph Leiter, Ran Zhang, Yanran Chen, Jonas Belouadi, Daniil Larionov, Vivian Fresen, and Steffen Eger. 2024. Chatgpt: A meta-analysis after 2.5 months. *Machine Learning with Applications* 16 (2024), 100541.
- [29] Can Li, Bin Pang, Wenbo Wang, Lingshu Hu, Matthew Gordon, Detelina Marinova, Bitty Balducci, and Yi Shang. 2023. How well can language models understand politeness?. In *2023 IEEE Conference on Artificial Intelligence (CAI)*. IEEE, 230–231.
- [30] Eleonore Lumer, Clara Lachenmaier, Sina Zarriß, and Hendrik Buschmeier. 2023. Indirect politeness of disconfirming answers to humans and robots. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1808–1815.
- [31] Andrew John Morrison, Jack J Wilson, Bethan L Davies, and Michael Haugh. 2012. Getting stuff done: Comparing e-mail requests from students in higher education in Britain and Australia. *Journal of pragmatics* 44, 9 (2012), 1077–1098.
- [32] Meta AI. 2024. Llama-3.1-8B-Instruct. <https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>
- [33] M Lynne Murphy and Rachele De Felice. 2019. Routine politeness in American and British English requests: use and non-use of please. *Journal of Politeness Research* 15, 1 (2019), 77–100.
- [34] Nedjma Ousidhoum, Xinran Zhao, Tianqing Fang, Yangqiu Song, and Dit-Yan Yeung. 2021. Probing toxic content in large pre-trained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 4262–4274.
- [35] Christopher Peters, Catherine Pelachaud, Elisabetta Bevacqua, Maurizio Mancini, Isabella Poggi, and Università Roma Tre. 2005. Engagement capabilities for ecas. In *AAMAS'05 workshop Creating Bonds with ECAs*. 47.
- [36] Zhi Quan and Zhiwei Chen. 2025. Human-computer pragmatics trialled: some (im) polite interactions with ChatGPT 4.0 and the ensuing implications. *Interactive Learning Environments* 33, 2 (2025), 1020–1039.
- [37] Patrizia Ribino. 2023. The role of politeness in human-machine interactions: a systematic literature review and future perspectives. *Artificial Intelligence Review* 56, Suppl 1 (2023), 445–482.
- [38] Klaus Schneider and Anne Barron (Eds.). 2008. *Variational Pragmatics: A Focus on Regional Varieties in Pluricentric Languages*. Benjamins, Amsterdam and Philadelphia.
- [39] Klaus P Schneider and J César Félix-Brasdefer. 2022. Pragmatic variation across national varieties of pluricentric languages. *Pragmatics of Space* (2022), 637–78.
- [40] Klaus P Schneider and Maria Elena Placencia. 2017. (Im) politeness and regional variation. In *The Palgrave handbook of linguistic (im) politeness*. Springer, 539–570.
- [41] science.ORF.at. 2024. *Künstliche Intelligenz: Warum Höflichkeit unökologisch sein kann*. <https://science.orf.at/stories/3231413/> Accessed: 2025-08-09.
- [42] Cazalets Tanguy, Ruben Janssens, Tony Belpaeme, and Joni Dambre. 2025. Human Alignment: How Much Do We Adapt to LLMs?. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 603–613. doi:10.18653/v1/2025.acl-short.47
- [43] Team Qwen. 2024. Qwen2.5 Technical Report. *arXiv preprint* (2024). arXiv:2412.15115 <https://arxiv.org/abs/2412.15115>
- [44] Verena Thaler. 2012. *Sprachliche Höflichkeit in computervermittelter Kommunikation*. Vol. 70. Stauffenburg-Verl.
- [45] Paul Thomas, Mary Czerwinski, Daniel McDuff, Nick Craswell, and Gloria Mark. 2018. Style and alignment in information-seeking conversation. In *Proceedings of the 2018 conference on human information interaction & retrieval*. 42–51.
- [46] Paul Thomas, Daniel McDuff, Mary Czerwinski, and Nick Craswell. 2020. Expressions of style in information seeking conversation with an agent. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1171–1180.
- [47] Beste Tore. 2025. *Speaking with AI: Gender-based variation in politeness and communication strategies*. Master's thesis. VYTAUTAS MAGNUS UNIVERSITY.
- [48] Chenxi Wang, Yixuan Zhang, Lang Gao, Zixiang Xu, Zirui Song, Yanbo Wang, and Xiuying Chen. 2025. Under the Shadow of Babel: How Language Shapes Reasoning in LLMs. *arXiv preprint arXiv:2506.16151* (2025).
- [49] Ziqi Yin, Hao Wang, Kaito Horio, Daisuke Kawahara, and Satoshi Sekine. 2024. Should we respect LLMs? A cross-lingual study on the influence of prompt politeness on LLM performance. In *Proceedings of the Second Workshop on Social Influence in Conversations (SICoN 2024)*. 9–35.
- [50] Yicong Yuan, Mingyang Su, and Xiu Li. 2024. What Makes People Say Thanks to AI. In *International Conference on Human-Computer Interaction*. Springer, 131–149.