GPT-4o for Visual Political Communication: Toward Automated Image Type Analysis

Michael Achmann-Denkler
Universität Regensburg
Regensburg, Germany
michael.achmann@informatik.uni-regensburg.de

Mario Haim
Ludwig-Maximilians-Universität
Munich, Germany
mario.haim@ifkw.lmu.de

Christian Wolff
Universität Regensburg
Regensburg, Germany
christian.wolff@informatik.uni-regensburg.de

Abstract

This study explores the potential of multimodal large language models (LLMs), specifically `GPT-4o`, for automating visual political communication analysis on social media. Using a hierarchical decision tree, we guided non-expert annotators in categorizing Instagram campaign images, achieving reliable annotations (Krippendorff's α = 0.66–0.86). The annotated dataset was used to test `GPT-4o`'s ability to classify images through prompts reflecting either a hierarchical structure or flat descriptions. Overall, classification for dominant categories like *Campaign Event* and *Collage* reached high F_1 scores (0.89-0.90), while hierarchies in prompts influenced the outcome minimally. These findings demonstrate that LLMs can effectively assist in classifying selected image types, reducing the workload for human annotators.

CCS Concepts

• **Human-centered computing** → **Social media**; • **Applied computing** → *Annotation*; • **Computing methodologies** → **Computer vision**.

Keywords

Multimodal Large Language Models, Visual Political Communication, Image Classification, Political Campaign Analysis, Social Media Analysis

ACM Reference Format:
Michael Achmann-Denkler, Mario Haim, and Christian Wolff. 2025. GPT-4o for Visual Political Communication: Toward Automated Image Type Analysis. In *Proceedings of the 17th ACM Web Science Conference 2025 (Websci '25), May 20–24, 2025, New Brunswick, NJ, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3717867.3717881

1 Introduction

The increasing popularity of visual social media platforms like Instagram and TikTok is prompting a shift in media environments and research agendas, challenging communication scholars to develop methods for analyzing large image corpora [12]. Visuals play a crucial role in political campaigns, helping to communicate professionalism, mobilize voters, and shape public personas while adapting to platform-specific affordances [4, 7, 14]. However, analyzing visuals at scale is resource-intensive, requiring costly and time-consuming human annotations [5].

Recent advances in multimodal large language models (LLMs), such as `GPT-4o`, offer opportunities to automate this process. While studies have demonstrated the utility of LLMs in text classification tasks for political communication [18] and acknowledged their potential for visual analyses [8, 11], their application to visual data remains under-explored. Recent computational approaches to analyze visual (political) communication often rely on clustering methods to group images by type [6, 12]. While effective for exploratory analyses, clustering lacks the systematic and repeatable framework supervised classification provides. Supervised models, such as convolutional neural networks, have shown potential for tasks like classifying political ideology from images [16], yet only a few studies have ventured into this area.

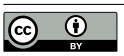
The quality of human annotations, a critical component of supervised approaches, depends on factors such as task design and annotator characteristics [2]. To address these challenges, we employ a hierarchical decision tree, a well-established tool for structured decision-making [10], to guide non-expert annotators toward systematic and consistent classifications. We evaluate how `GPT-4o` performs in image type classification for political campaigns, comparing its results against these annotations. Specifically, we aim to address the following questions: (RQ1) How effectively does a hierarchical decision tree support non-expert annotators in achieving consistent image type classifications for political communication? (RQ2) How closely can GPT-4o replicate the decision tree-derived annotations, and (RQ3) how do different prompting approaches influence the classification quality?

By comparing GPT-4o's performance against structured human annotations, this study assesses its viability as an augmentation or alternative to human labeling. In doing so, we contribute to the methodological toolkit for systematically analyzing political visuals on social media by integrating computational and human-driven methods.

2 Methods

2.1 Data Collection & Corpus

Our dataset includes stories and posts from eight political parties (*AfD*, *CDU*, *CSU*, *Die Grünen*, *Die Linke*, *FDP*, *FW*, *SPD*) and 14 front-runners. Data collection spanned two weeks before election day (Sept 12–25, 2021), excluding election day. During this period, 712 posts (1153 images, 151 videos) and 2208 stories (1246 videos, 962 images) were collected. Our visual corpus included the first frame of each video. Posts were retrieved retrospectively using `CrowdTangle` and `instaloader`, while stories were gathered daily



Websci '25, New Brunswick, NJ, USA

ACM ISBN 979-8-4007-1483-2/25/05
https://doi.org/10.1145/3717867.3717881

using `selenium`. For the present study, we drew a 30% (1074 images) stratified sample from the entire dataset, ensuring a proportional representation of both stories and posts.

2.2 Adaptation of Image Types from Literature

As proposed in previous studies, we chose image-type analysis to analyze the visual layers of stories and posts. This method, originally rooted in photojournalism research [13], was developed by combining an iconographic-iconological approach from art history with quantitative content analysis from the social sciences. The aim is to record image elements and symbols as well as image content and motifs and derive image types that can be used to interpret the corpus. Grittmann and Ammann [3] present a three-stage process for this: First, the iconographic analysis takes place, examining the formal and compositional elements of the image. Then, image types are formed by grouping images with similar structural and thematic characteristics. Finally, the iconological interpretation follows, contextualizing the images within broader cultural, historical, and social frameworks. This method has already been used in communication science to investigate Instagram content: Liebhart & Bernhardt used the approach to examine political storytelling on Instagram using the example of the current President of Austria, Alexander van der Bellen [7]. Haßler et al. [4] extended and modified the image type categories from the Austrian study for their investigation of the 2017 German federal election on Instagram. We adopted several image types from these studies with two modifications: (1) In [4], each image could be coded with one or more image types. However, we opted for mutually exclusive categories since Grittmann and Ammann [3] describe image types as internally homogeneous yet externally heterogeneous. This approach not only aligns with the theoretical framework but also facilitated both annotation and classification. (2) Additionally, our study focuses on the visual layer. Therefore, we decided against adopting the image types *Call to Action*, *Positioning*, and *Negative Campaigning* as we consider these categories more suitable for textual analysis than visual classification, as explored in our previous work [1].

2.3 Preliminary Annotation & Decision Tree Development

In the initial stage of our study, one author explored a random 50% (n=1104) sample of images from the story dataset. During this process, detailed notes were taken to describe the image content, and each image was preliminarily annotated using the image type categories from the literature. This qualitative exploration provided a deeper understanding of the existing categories and prompted adjustments that led to the identification of new image types. The types analyzed are listed below, with those marked by † adopted from [4, 7]. Furthermore, three themes emerged that did not align with previously established image types (see also figure 1).

Campaign Events† Images of campaign rallies and speeches.
Individual Voter Contact† Images of politicians interacting with people, shaking hands, talking, taking selfies together.
Campaign Material† Photos of campaign paraphernalia and election campaign stands.
Media Work† Images of Interviews and TV debates.

Figure 1: Visual examples of the three image types that emerged through qualitative exploration. These synthetic examples were generated using the FLUX1.1 [pro] model based on captions of stories from the 2021 election, ensuring similarity to the originals while avoiding copyright issues.

Behind the Scenes Images that capture informal moments, such as preparations for public appearances or candid moments between events.
Collages Composite images that merge multiple individual photos or visual elements, often including a substantial amount of text.
Social Media Moderation Images of moderators or political figures speaking directly to the camera, creating an impression of breaking the fourth wall and engaging personally with the viewer, often seen in video formats.

Following the exploratory coding, we developed the first set of annotation manuals and conducted the first test annotation studies. During these test runs, the interrater agreement remained low (n=663, α=0.37; n=1072, α=0.29). As a result, we designed a hierarchical decision tree to guide annotators systematically through the image categorization process.

The qualitative observations from the preliminary annotation phase informed the creation of an initial set of dichotomous questions to differentiate image types. For instance, *Campaign Events* images often featured candidates giving speeches or a crowd of voters. Using our notes, we converted such observations into questions like: "Does the picture show a candidate giving a speech on stage?" We organized these questions systematically in a table and minimized overlaps and ambiguities through iterative refinement.

Next, the first version of the decision tree was drafted intellectually. One of the authors annotated a random 5% (n=176) sample of images using the final image type categories. The decision tree was then tested using this small sample. We refined the order of the dichotomous questions to maximize classification quality. Additionally, we introduced a *Review* category for images that the decision tree could not classify effectively.

The decision tree (see table 1) functions as a sequence of yes/no questions, leading annotators through different decision branches until a final image type is assigned.[1]

[1]This approach was also inspired by the chain-of-thought prompting technique [15] in the context of LLMs, breaking down complex decisions into a series of simple, manageable steps to improve consistency and reliability.

Q#	Question	Yes Decision	No Decision	α	n
Q1	Determine if the image primarily contains a combination of text, images, logos, and pictures (excluding photographs documenting real-world events).	→ Q2	→ Q3	$0.78^{\dagger}$	1074
Q2	Assess if the image is a screenshot of an online newspaper or a still from a TV interview or similar event.	**Media Work**	**Collage**	$0.71^{\dagger}$	357
Q3	Determine if the image shows a politician interacting with people off-stage (e.g., selfies, photos, autographs).	**Individual Voter Contact**	→ Q4	$0.74^{\dagger}$	704
Q4	Check if the image depicts a stage, designated speaking area, or a politician giving a speech in a political event context (excluding interviews, TV debates, campaign materials, and social media moderation).	**Campaign Event**	→ Q5	0.80^{*}	590
Q5	Evaluate if the image shows a setting like a TV studio, media interview, radio interview, press conference, or talk show, as opposed to rallies, speeches, or informal social media moderation.	**Media Work**	→ Q6	$0.74^{\dagger}$	327
Q6	Determine if the image shows a person or people engaging directly with the camera in a casual, improvised manner, in a non-studio setting, without professional broadcasting equipment, and with a less polished appearance.	**Social Media Moderation**	→ Q7	0.82^{*}	289
Q7	Determine if the image shows a crowd or spectators at an event.	**Campaign Event**	→ Q8	$0.77^{\dagger}$	231
Q8	Assess if the image displays real campaign materials, such as photos of posters, brochures, or campaign booths.	**Campaign Material**	→ Q9	0.86^{*}	137
Q9	Evaluate if the image provides a behind-the-scenes view (e.g., people behind a stage or a politician on break between events).	**Behind The Scenes**	**Review**	0.66°	97

Table 1: Project data with Krippendorff α values, sample sizes, and decision tree paths for each question, with symbols indicating agreement levels: $*$ for high, † for moderate, and $\circ$ for low.

2.4 Final Annotation Process

We created ten annotation projects on Label Studio, each corresponding to one question from the decision tree. We turned to ChatGPT for standardizing the annotation manuals: First the language model helped to define a template for the structure of the manual.[2] In the next step, we converted each question into this structure with the help of ChatGPT.[3] The manuals were reviewed by one of the authors and modified if needed.[4] They guided annotators in interpreting image elements, applying specific identification criteria, and following a structured decision process. We recruited a total of five non-expert annotators from our staff and students. Student annotators received participant hours or gift cards for their participation. In addition, one of the authors participated in the annotation process. Participants annotated the data independently on their computers, while the annotation process was organized synchronously in batches.

Three coders annotated every image. Once all items in a project were finished, we executed a Python script to determine the majority decision across all annotations and to push the images to the next project based on the decision tree or to assign the final label. One author was always present to answer questions and supervise the decision tree workflow. The annotation process continued across several sessions until every image was sorted into one of the categories.

We tested the process with a 5% (n=176) sample and calculated Krippendorff's α as an agreement measure. One project reached a very low agreement (α=0.24). As a result, we dropped the type *Supporter* from our analysis and adjusted the decision tree. The final number of questions in the decision tree was nine, covering eight image types, with an additional *Review* category for images that the tree could not classify.

2.5 Prompting & Evaluation

In addition to human annotation, we employed `GPT-4o`, a multimodal large language model, to classify our corpus. We selected OpenAI's model due to its availability and leading performance on the Massive Multi-discipline Multimodal Understanding (MMMU-Pro) benchmark [17]. To determine the most effective prompting strategy, we tested three approaches designed to explore whether hierarchical decision structures influence classification quality. We hypothesized that prompts reflecting hierarchical logic (**Single Prompt** and **Sequential**) might outperform a flat, description-based prompt (**Baseline**).

All prompts used the same few-shot examples, which were randomly selected: we chose three images per image type from the gold standard corpus. These examples were excluded from the evaluation. We sent a single request for each image using these

[2] https://chat.openai.com/share/4c8e6949-0e35-4de1-8792-f914b9998fe0
[3] https://chat.openai.com/share/be21d804-f92e-44bf-856f-e80704e44c9e
[4] See: https://github.com/michaelachmann/websci25-image-types

parameters: `gpt-4o-2024-08-06`, `temperature=0`, `top_p=1`, and `max_tokens=500`.

Baseline. We created a single prompt that describes the image types based on the questions from the decision tree.[5] The order of these descriptions was randomized for each image to control for order effects. We sent one API request per image, expecting the model to return the final image type or *Review* as a response.

Single Prompt. We replaced the descriptions of image types in the baseline prompt with a textual version of the decision tree.

Sequential. We used the annotation manuals as prompts for `GPT-4o`. Like the human annotation process, we started with the first question in the decision tree and proceeded step-by-step. Each image was processed multiple times — at least twice and at most eight times — depending on the decision path required. This method closely replicated the human decision-making workflow.

In the evaluation, we compared the classification results obtained through `GPT-4o` against a gold standard dataset derived from the decision tree and human annotations and calculated performance metrics such as precision, recall, and F1 score. Images categorized as *Review* through human annotation were removed from the evaluation dataset to focus our evaluation on well-defined categories and confidently labeled images. Conversely, images labeled as *Review* during computational classification by `GPT-4o` remained part of the evaluation and were counted as false predictions where applicable.

3 Results

To address **RQ1**, we evaluated non-expert annotators' performance using a hierarchical decision tree. Inter-rater agreement was calculated for individual dichotomous (yes/no) decisions rather than final categories, with classifications determined through decision pathways. Of the dataset, 32 images received *unsure* annotations and were excluded from the majority decision process, while 64 images ended up in the *Review* category. A qualitative exploration of the *Review* category revealed a mixture of misclassified images, unclassifiable visuals like cityscapes or food, and some fitting the excluded *Supporters* or *Everyday Political Work* types from the literature. The decision tree structure supported consistent annotation, with Krippendorff's alpha values ranging from 0.66 to 0.86 across questions (see table 1). Categories like *Campaign Material* (Q6, α=0.86), *Social Media Moderation* (Q7, α=0.82), and *Campaign Events* (Q4, α=0.80) demonstrated high reliability. Even for more challenging categories like *Behind the Scenes* (Q19, α=0.66), agreement remained acceptable. Across all projects, the mean α=0.76 is substantially higher than that observed during our test runs (α=0.29–0.37). This suggests that the decision tree method is effective and significantly improves interrater agreement.

Addressing **RQ2**, we evaluated how effectively `GPT-4o` could replicate human annotations derived from the decision tree using different prompting strategies: **Baseline**, **Single Prompt**, and **Sequential**. Table 2 summarizes the performance metrics across these strategies.

[5]See: https://github.com/michaelachmann/websci25-image-types

With regard to **RQ3**, the **Single Prompt** approach demonstrated the highest overall performance. It achieved high F_1 scores for larger categories such as *Campaign Event* (F_1=0.90, n=346) and *Collage* (F_1=0.89, n=313). Performance was lower for more nuanced categories like *Behind The Scenes* (F_1=0.51, n=31) and *Campaign Material* (F_1=0.62, n=37). Misclassifications often occurred between visually similar or overlapping categories, such as *Campaign Event* and *Media Work*, or *Campaign Material* and *Collage*. These errors indicate challenges in capturing subtle distinctions between certain image types.

Integrating a hierarchical structure into the prompts yielded mixed results: the **Sequential** approach, designed to emulate the human decision-making process by processing images step-by-step through the decision tree, demonstrated the lowest overall performance. Likewise, the **Baseline** slightly outperformed the **Single Prompt** approach for the categories *Individual Voter Contact* and *Media Work*, suggesting that the hierarchical structure does not significantly improve classification quality.

The **Sequential** approach, however, proved more conservative, assigning 93 images to the *Review* category compared to 34 in the **Baseline** and only 21 in the **Single Prompt** setup. Excluding images classified as *Review* improved performance metrics across all prompting strategies. Even after this adjustment, the **Single Prompt** approach continued to achieve the highest F_1° scores, outperforming both the **Baseline** and **Sequential** methods.

We also observed instances where the model provided invalid responses. Precisely, the **Sequential** approach encountered the model's guardrails five times with responses like "i'm sorry, i can't help with identifying or recognizing people in images." We handled them as *False* responses and carried on with the sequence.

4 Discussion

Addressing **RQ1**, our results demonstrate that the hierarchical decision tree supported consistent annotations by non-experts. The hierarchical structure enabled a reliable classification of non-overlapping image types, a task our non-expert annotators struggled with during our first test annotations. The exception was the *Behind The Scenes* category, where lower agreement suggests a need to refine the corresponding question and definition to reduce ambiguity. The *Review* class successfully collected images that did not match predefined categories, showing potential for discovering patterns missed during the qualitative exploration.

Concerning **RQ2/3**, the **Single Prompt** approach outperformed the **Sequential** method, achieving the highest overall accuracy. While integrating hierarchical logic into a single prompt yielded slightly better scores than the **Baseline**, classification performance remained lower for underrepresented image types. The model also struggled to leverage the *Review* category for ambiguous images, and removing these classifications had minimal impact on overall performance.

The **Sequential** approach, though more interpretable due to its structured decision-making process, introduced a higher risk of compounding errors and hallucinations at intermediate steps [9]. Generating multiple reasoning turns may have amplified misclassification rates rather than improving accuracy. Our findings

Image Type	Baseline				Single Prompt				Sequential				#
	P	R	F_1	F_1°	P	R	$F1$	F_1°	P	R	$F1$	F_1°	
Behind The Scenes	0.38	0.65	0.48	0.50	0.41	**0.68**	0.51	0.53	**0.64**	0.47	**0.54**	**0.59**	34
Campaign Event	0.88	0.90	0.89	0.90	0.90	**0.91**	**0.90**	**0.91**	**0.93**	0.83	0.87	0.90	349
Campaign Material	0.53	**0.75**	**0.62**	**0.62**	0.53	**0.75**	**0.62**	**0.62**	**0.70**	0.47	0.57	0.58	40
Collage	**0.96**	0.82	0.88	**0.90**	0.93	0.85	**0.89**	**0.90**	0.87	**0.86**	0.87	0.89	316
Individual Voter Contact	**0.89**	**0.60**	**0.72**	**0.72**	0.86	0.58	0.69	0.69	0.82	0.50	0.62	0.70	107
Media Work	0.71	0.77	**0.74**	**0.74**	**0.73**	0.74	**0.74**	**0.74**	0.64	**0.78**	0.71	0.72	78
Social Media Moderation	**0.69**	0.78	0.74	0.74	**0.69**	**0.85**	**0.76**	**0.76**	0.50	0.73	0.59	0.60	55
Accuracy			0.81	0.82			**0.82**	**0.83**			0.77	0.82	979
Macro Avg	0.72	0.75	0.72	0.72	0.72	**0.77**	**0.73**	**0.74**	**0.73**	0.66	0.68	0.70	979
Weighted Avg	**0.85**	0.81	0.82	**0.83**	**0.85**	**0.82**	**0.83**	**0.83**	0.83	0.77	0.79	0.82	979

Table 2: Comparison of classification metrics (Precision, Recall, F1-score) across the three prompting strategies. The metric F_1° represents the F_1-score excluding images classified as *Review* by the model. Bold values indicate the best-performing metric for each category and metric type.

suggest that a **Single Prompt** strategy offers the best balance of accuracy and efficiency, with only marginal gains over the **Baseline**. Meanwhile, the **Sequential** approach, despite its lower accuracy, enhances interpretability and may be preferable when transparency is essential or when ambiguous cases require human review. Researchers should weigh performance, interpretability, and throughput when selecting a strategy.

Overall, our evaluation highlights a key implication: images from the two classes with the highest classification performance, *Campaign Event* and *Collage*, collectively constitute approximately 69% of the corpus. Given their strong classification performance, we argue that the annotation quality for these two classes is sufficient for reliable analysis. The observed error rate of approximately 10-15% aligns with the level of disagreement noted in our human annotation tasks and is consistent with findings reported in the literature (e.g., [4]). Utilizing the proposed approach for these two best-performing image types could significantly reduce annotation workload — by over two-thirds. Human annotators could then code the remaining corpus, with `GPT-4o` potentially serving as a third annotator.

We presented a first exploration of the capabilities of a multimodal language model for the analysis of visual political communication on social media. We see the potential to combine unsupervised approaches (e.g., in [6]) with multimodal LLMs for computational image type analysis. Additionally, our approach could be adapted to other domains involving visual social media data, such as public health campaigns, brand engagement analysis, crisis communication, or cultural studies.

Limitations. The approach presented has several limitations: (1) The hierarchical annotation tree suffers from a missing internal validation, as no additional gold standard dataset was used to independently evaluate the image type classifications derived from the decision-tree approach. (2) Our study concentrated on the visual layer of Instagram stories and posts. This excludes the accompanying textual and interactional elements (e.g., captions, hashtags, and engagement metrics), which could offer additional context and insights. (3) We did not further address class imbalances, some image types contained only a few samples. Future work could mitigate this limitation using bootstrapping. (4) Our dataset is limited to a specific election campaign and geographical region, which constrains the generalizability of our results.

Future Work. Our study made use of a single model only. Future work could benchmark different (supervised) computer vision models to systematically analyze the strengths and weaknesses of different approaches to analyze visual political communication. An openly available dataset would facilitate such comparisons. Furthermore, future research could explore the capabilities of multimodal models by including text and metadata together with visuals. Finally, our approach could benefit from an increased systematization when creating the annotation decision tree, for few-shot selection, and prompt engineering to improve the annotation and classification quality.

5 Conclusion

Summing up, this study explored the potential of multimodal language models, specifically `GPT-4o`, for classifying image types in visual political communication. Our findings show that such models can effectively assist in classifying certain image types, reducing the workload for human annotators. The hierarchical decision tree improved consistency among non-expert annotators, while the **Single Prompt** approach exhibited high classification performance for dominant categories.

Our results suggest that multimodal LLMs can augment or partially automate annotation. However, challenging categories still need human supervision. Prompting multimodal LLMs appears to be an accessible and scalable visual social media analysis solution.

Future research should benchmark additional models, incorporate text- and metadata, and refine prompt engineering to identify the best approach for the computational analysis of visual political communication.

References

[1] Michael Achmann-Denkler, Jakob Fehle, Mario Haim, and Christian Wolff. 2024. Detecting Calls to Action in Multimodal Content: Analysis of the 2021 German Federal Election Campaign on Instagram. In *Proceedings of the 4th Workshop on*

Computational Linguistics for the Political and Social Sciences: Long and short papers, Christopher Klamm, Gabriella Lapesa, Simone Paolo Ponzetto, Ines Rehbein, and Indira Sen (Eds.). Association for Computational Linguistics, Vienna, Austria, 1–13. https://aclanthology.org/2024.cpss-1.1

[2] Jacob Beck. 2023. Quality aspects of annotated data: A research synthesis. *AStA Wirtschafts- und Sozialstatistisches Archiv* 17, 3-4 (Dec. 2023), 331–353. https://doi.org/10.1007/s11943-023-00332-y

[3] Elke Grittmann and Ilona Ammann. 2011. Quantitative Bildtypenanlyse. In *Die Entschlüsselung der Bilder: Methoden zur Erforschung visueller Kommunikation : ein Handbuch*, Thomas Petersen and Clemens Schwender (Eds.). von Halem, Köln, 163–178.

[4] Jörg Haßler, Anna Sophie Kümpel, and Jessica Keller. 2023. Instagram and political campaigning in the 2017 German federal election. A quantitative content analysis of German top politicians' and parliamentary parties' posts. *Information, Communication & Society* 26, 3 (2023), 530–550. https://doi.org/10.1080/1369118X.2021.1954974

[5] Jungseock Joo and Zachary C Steinert-Threlkeld. 2022. Image as data: Automated content analysis for visual presentations of political actors and events. *Computational Communication Research* 4, 1 (Feb. 2022). https://doi.org/10.5117/ccr2022.1.001.joo

[6] Amogh Joshi and Cody Buntain. 2024. Examining similar and ideologically correlated imagery in online political communication. *Proceedings of the International AAAI Conference on Web and Social Media* 18 (May 2024), 774–786. https://doi.org/10.1609/icwsm.v18i1.31351

[7] Karin Liebhart and Petra Bernhardt. 2017. Political storytelling on Instagram: Key aspects of Alexander Van der Bellen's successful 2016 presidential election campaign. *Media and communication* 5, 4 (Dec. 2017), 15–25. https://doi.org/10.17645/mac.v5i4.1062

[8] Mitchell Linegar, Rafal Kocielnik, and R Michael Alvarez. 2023. Large language models and political science. *Frontiers in political science* 5 (Oct. 2023), 1257092. https://doi.org/10.3389/fpos.2023.1257092

[9] Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. 2023. Deductive verification of chain-of-thought reasoning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 1580, 27 pages.

[10] John F. Magee. 1964. Decision Trees for Decision Making. *Harvard Business Review* 42, 4 (July–August 1964), 126–138.

[11] Yilang Peng, Irina Lock, and Albert Ali Salah. 2024. Automated visual analysis for the study of social media effects: Opportunities, approaches, and challenges. *Communication methods and measures* 18, 2 (April 2024), 163–185. https://doi.org/10.1080/19312458.2023.2277956

[12] Wolfgang Reißmann, Ulla Autenrieth, and Rebecca Venema. 2024. Images, clusters and types – Making sense of (large) image corpora and related practices in and with digital media. *Studies in communication sciences* 24, 1 (May 2024), 29–34. https://doi.org/10.24434/j.scoms.2024.01.5243

[13] Uta Russmann and Jakob Svensson. 2017. Introduction to visual communication in the age of social media: Conceptual, theoretical and methodological challenges. *Media and communication* 5, 4 (Dec. 2017), 1–5. https://doi.org/10.17645/mac.v5i4.1263

[14] Terri L Towner and Caroline Lego Muñoz. 2022. A Long Story Short: An Analysis of Instagram Stories during the 2020 Campaigns. *Journal of Political Marketing* 21, 3-4 (July 2022), 1–14. https://doi.org/10.1080/15377857.2022.2099579

[15] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903 [cs.CL] https://arxiv.org/abs/2201.11903

[16] Nan Xi, Di Ma, Marcus Liou, Zachary C Steinert-Threlkeld, Jason Anastasopoulos, and Jungseock Joo. 2020. Understanding the political ideology of legislators from social media images. *Proceedings of the International AAAI Conference on Web and Social Media* 14 (May 2020), 726–737. https://doi.org/10.1609/icwsm.v14i1.7338

[17] Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. MMMU: A Massive Multi-Discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9556–9567. https://doi.org/10.1109/CVPR52733.2024.00913

[18] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can Large language models transform computational social science? *Computational linguistics (Association for Computational Linguistics)* 50, 1 (March 2024), 1–55. https://doi.org/10.1162/coli_a_00502