NLPSI 2025

The First Workshop on Integrating NLP and Psychology to Study Social Interactions (NLPSI) @ICWSM '25

Proceedings of the Workshop

June 23, 2025

This issue of the ICWSM Workshop Proceedings contains papers published at the workshops held on June 23rd, 2025 at the 19th International AAAI Conference on Web and Social Media (ICWSM 2025).
Copenhagen, Denmark

Workshop Chairs
Luca Rossi (University of Copenhagen)
Shruti Padke (Drexel University)
Akhil Arora (Aarhus University)

Published on: 2025-06-05

Preface

This volume presents the proceedings of the First Workshop on Integrating NLP and Psychology to Study Social Interactions (NLPSI 2025), held in conjunction with the 19th International AAAI Conference on Web and Social Media (ICWSM 2025) in Copenhagen, Denmark.

The workshop was motivated by the observation that recent advances in Natural Language Processing offer new opportunities to study psychological phenomena in language at a scale that was previously out of reach. NLP allows us to move beyond controlled, small-scale experiments and conduct large-scale studies that engage with key theories and research questions in psychology, while psychological research designs and theories in turn offer valuable grounding for NLP work on communication and social interaction.

The workshop received 25 submissions, of which we accepted 16: 5 long papers, 3 short papers, and 8 extended abstracts. Among the extended abstracts, 6 were non-archival and are therefore not included in the proceedings. The accepted contributions covered a broad range of topics, including multi-agent LLM simulations of social interaction, negotiation, and persuasion; computational modeling of psychological constructs such as motives, basic psychological needs, and well-being; the use of generative agents to study behavior change and belief revision; pragmatic and narrative-based approaches to interpersonal communication; and the development of lexical resources for social media analysis.

The workshop featured an invited talk by Bennett Kleinberg (Tilburg University & University College London), titled "Human and Machine Behaviour – Bridging Psychology and Computational Methods." Bennett Kleinberg is an Associate Professor in Behavioural Data Science at Tilburg University (Department of Methodology and Statistics) and University College London (Department of Security and Crime Science).

We are grateful to the ICWSM 2025 workshop chairs for enabling these proceedings to be included in the ACL Anthology. We would also like to warmly thank the members of our Program Committee; your time investment and thorough reviews played a key role in shaping the quality of this workshop. Finally, we thank the authors for their contributions and the participants for making this first edition of NLPSI a success.

Aswathy Velutharambath, Sofie Labat, Neele Falk, Flor Miriam Plaza-del-Arco, Roman Klinger and Véronique Hoste
Program Chairs

Organizing Committee

Organizers

Aswathy Velutharambath, University of Bamberg
Sofie Labat, Ghent University
Neele Falk, University of Stuttgart
Flor Miriam Plaza-del-Arco, Bocconi University
Roman Klinger, University of Bamberg
Véronique Hoste, Ghent University

Program Committee

Reviewers

Liesbeth Allein

Christopher Bagdon, Elisa Bassignana

Alessandra Teresa Cignarella, Aidan Combs

Luna De Bruyne, Orphee De Clercq, Loic Delanghe, Esra Dönmez

Neele Falk

Lynn Greschner

Karina H Halevy

Els Lefever

Aaron Maladry, Yarik Menchaca Resendiz

Sean Papay, Nadine Probol

Jocelyn J Shen, Pranaydeep Singh, Marco Antonio Stranisci

Tarun Tater, Enrica Troiano

Cynthia Van Hee, Eva Maria Vecchi

Amelie Wuehrl

Federico Zimmerman

Keynote Talk

Human and Machine Behaviour – Bridging Psychology and Computational Methods

Bennett Kleinberg
Tilburg University & University College London

Bio: Bennett Kleinberg is an Associate Professor in Behavioural Data Science at Tilburg University (Department of Methodology and Statistics) and University College London (Department of Security and Crime Science). His research explores the interplay and advancement of computational methods and psychological research to study human behaviour. He previously worked at the UCL Dawes Centre for Future Crime and earned his PhD from the Department of Psychology at the University of Amsterdam. His lab addresses two core questions: how computational methods can enhance our understanding of human behaviour, and how psychological research methods can inform our understanding of computational model behaviour, with a methodological focus on textual data and psychological processes, including deception detection, human resilience, and machine behaviour.

Table of Contents

Simulating Persuasive Dialogues on Meat Reduction with Generative Agents

Georg Ahnert[1*†], **Elena Wurth**[1*], **Markus Strohmaier**[1,2,3], **Jutta Mata**[1,4]

[1]University of Mannheim
[2]GESIS - Leibniz Institute for the Social Sciences, Cologne
[3]Complexity Science Hub, Vienna
[4]Max Planck Institute for Human Development, Berlin

Abstract

Meat reduction benefits human and planetary health, but social norms keep meat central in shared meals. To date, the development of *communication strategies that promote meat reduction while minimizing social costs* has required the costly involvement of human participants at each stage of the process. We present work in progress on simulating multi-round dialogues on meat reduction between *Generative Agents* based on large language models (LLMs). We measure our main outcome using established psychological questionnaires based on the *Theory of Planned Behavior* and additionally investigate *Social Costs*. We find evidence that our preliminary simulations produce outcomes that are (i) consistent with theoretical expectations; and (ii) valid when compared to data from previous studies with human participants. Generative agent-based models are a promising tool for identifying novel communication strategies on meat reduction—tailored to highly specific participant groups—to then be tested in subsequent studies with human participants.

1 Introduction

Research Objectives Reducing meat consumption offers substantial benefits for human and planetary health, yet entrenched social norms continue to place meat at the center of shared meals (Godfray et al. 2018). Vegans, vegetarians, and flexitarians have the potential to challenge these norms and drive important social change (Judge et al. 2024). However, in mixed-diet social settings, they often self-silence to avoid social costs (e.g., Bolderdijk and Cornelissen 2022; Romo and Donovan-Kicken 2012). To date, the development of communication strategies that promote meat reduction while minimizing social costs has required human participants at each stage of the process (e.g., Carfora et al. 2019; Pabian et al. 2020). This is not only very costly—both financially for the researcher and in terms of participant effort—but also severely restricts the range of communication strategies and participant diversity that can be explored. Generative Agent simulations based on large language models (LLMs) have the potential to overcome these limitations (Bail 2024).

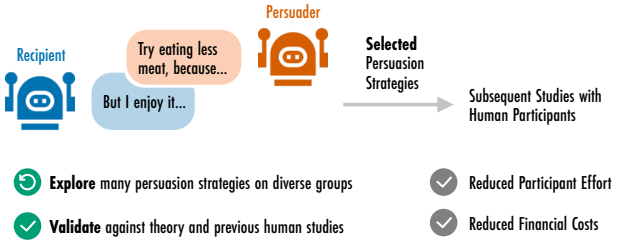


Figure 1: **Simulation of Persuasive Dialogues.** We use generative agents to develop and select persuasion strategies for meat reduction with minimal social costs, to be tested in subsequent studies with human participants.

Approach We present work in progress on simulating multi-round dialogues on meat reduction between *generative agents* based on LLMs, as shown in Figure 1. Related work on generative agent simulations already shows promising results (Park et al. 2022; Törnberg et al. 2023). Still, it remains an open question how well such simulations can inform subsequent studies with human participants, especially in our context of dialogues on meat reduction. This paper therefore addresses the following **Research Question**:

To what extent can generative agent simulations reliably reproduce persuasive human dialogues about meat reduction? To answer this, we validate the behavior of our agents both internally and against data from prior experiments with human participants. This validation is a necessary step before using these simulations to inform the design of future studies involving human subjects.

Results Initial experiments with the `Llama 3` family of models indicate that generative agents can effectively simulate persuasive communication, producing reliable and valid response patterns across key psychological constructs. However, issues such as uniformity in some constructs and the potential influence of instruction-tuning on strategies will require further investigation.

Generative agent-based models can be a promising tool for identifying novel communication strategies for meat reduction, tailored to highly specific participant groups, to then be tested in subsequent studies with human participants.

*These authors contributed equally.
†corresponding author: ahnert [at] uni-mannheim [dot] de

Proceedings of the First Workshop on Integrating NLP and Psychology to Study Social Interactions (NLPSI) @ICWSM '25, pages 1–10

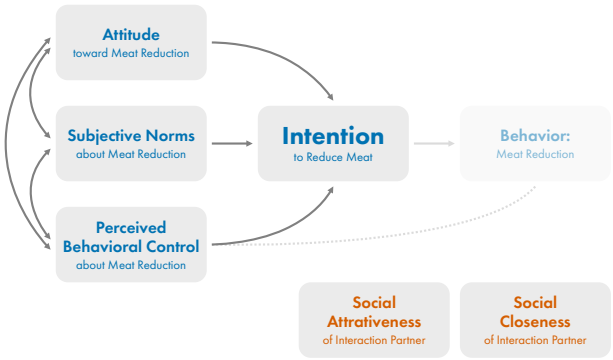


Figure 2: **Constructs Measured for the Recipient, Comprising Theory of Planned Behavior (TPB) and Social Costs.** Behavior itself (i.e., meat reduction) is unobservable to us in our simulation, but previous research has established that self-reported **Intention to Reduce Meat**—our main outcome—adequately predicts actual reduction (Berndsen and Van Der Pligt 2005; Saba and Di Natale 1998).

2 Background

Studies on Meat Reduction The Theory of Planned Behavior (TPB; Ajzen 1991) explains behavior through three core constructs: *Attitudes*, *Subjective Norms*, and *Perceived Behavioral Control*. These shape behavioral *Intentions*, which in turn predict actual behavior. On average, intentions account for 28% of variance in health behavior (Sheeran et al. 2005). In meat reduction, these variables reflect recipient traits as follows: attitudes map personal beliefs about meat, norms relate to perceived expectations of others, and control captures confidence in one's ability to change eating habits. From theory and previous research, all three constructs reliably predict the intention to eat less meat, which in turn predicts actual meat consumption (Berndsen and Van Der Pligt 2005; Saba and Di Natale 1998).

Social Simulations with Generative Agents A rapidly growing body of research investigates the feasibility of using large language models (LLMs) to model survey responses (e.g., Argyle et al. 2023; Bisbee et al. 2023; Ahnert et al. 2025), or the outcomes of experimental studies (e.g., Hewitt et al. 2024; Aher, Arriaga, and Kalai 2024). LLMs promise to facilitate novel exploratory studies of human behavior (Bail 2024), but recent research indicates that LLMs might not always accurately mimic human study participants (Tjuatja et al. 2024). It remains an open question how well socio-demographically prompted LLMs can represent human study participants (Sen et al. 2025)—in particular, LLMs were shown to "flatten" the reported attitudes of identity groups (Wang, Morgenstern, and Dickerson 2024).

Generative Agents use LLMs to simulate individuals and their interactions. Previous research applied generative agents, for example, to simulate communication on social media platforms (Park et al. 2022; Törnberg et al. 2023). Persuasive dialogues between generative agents have been found to favor communication strategies similar to humans (Vaccaro et al. 2025), but also to converge to the inherent biases of the underlying LLMs (Taubenfeld et al. 2024). We apply a similar generative agent setup to a novel context: persuasion for meat reduction with minimal social costs. This allows us to draw from the well-established literature in health psychology to statistically validate our generative agent-based model and to assess its potential for informing subsequent studies with human participants. We test a variety of open-weight LLMs, since previous research found that model size is an important predictor of negotiation success (Davidson et al. 2024).

3 Approach

Psychological Constructs Based on two current meta-analyses (Harguess, Crespo, and Hong 2020; Kwasny, Dobernig, and Riefler 2022), we identify several personal characteristics relevant to meat consumption and its reduction, including demographics, values, and personality traits, as shown in Table 1. Based on these central characteristics and to test the potential of our method, we develop two contrasting personas. According to the two meta-analyses mentioned above, these personas should represent opposite ends of the expected susceptibility to meat reduction appeals: a progressive, open-minded, younger woman with high education and income living in an urban setting (*Easy to Persuade Recipient*), and a conservative, older male with limited education and low income residing in a rural area (*Hard to Persuade Recipient*).

We define the Theory of Planned Behavior (TPB) as our theoretical framework and the corresponding variables as our outcomes for persuasion, as shown in Figure 2. Consequently, we use validated instruments to assess these constructs reliably. As part of our simulation, we extract survey responses from the agent to be persuaded (*Recipient*), with **Intention to Reduce Meat** as our main outcome. The questionnaire measures our 6 central constructs: *Attitudes*, *Subjective Norms*, and *Behavioral Control* of the Recipient; the Recipient's *Intention to Reduce Meat Consumption*; as well as the *Social Closeness* and the *Social Attractiveness*

Persona 1: Easy to Persuade Recipient	**Persona 2: Hard to Persuade Recipient**
demographics	**demographics**
female, younger, living in the city, high income	male, older, living on the countryside, low income
values	**values**
self-transcendence, openness to change, encouraging empathy towards animals	self-enhancement, conservation, discouraging empathy towards animals
personality traits	**personality traits**
open, conscientious	conservative, careless

Table 1: **Contrasting Recipient Personas.** For our initial experiments, we focus on two contrasting personas that provide the greatest potential for identifying existing effects, as they represent opposite ends of the expected susceptibility to meat reduction arguments.

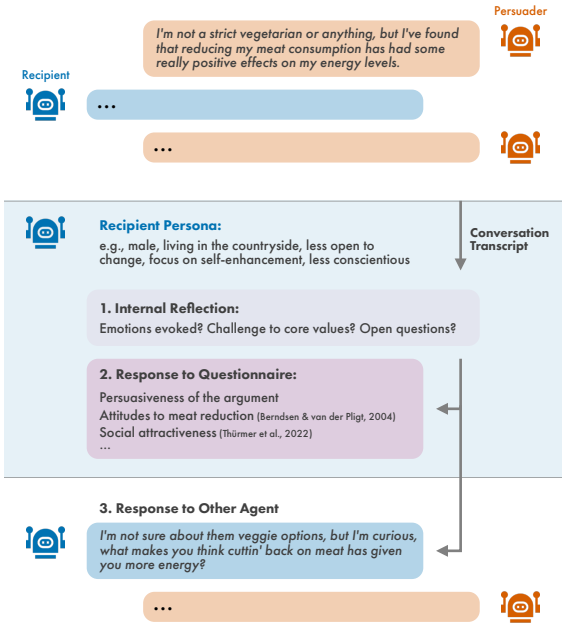


Figure 3: **Simulation Setup.** We simulate dialogues in which a Persuader agent aims to convince a Recipient agent to reduce their meat consumption. In each round of the simulation, agents first (1) perform internal reflection, and (2) answer a questionnaire, before they (3) generate a response to the other agent.

of the Persuader, as perceived by the Recipient. The first 4 constructs comprise the aspects of the TPB that we can measure in our simulation, while the latter 2 constructs measure Social Costs—see Appendix A for the full questionnaire. We expect individuals with favorable attitudes, strong normative support, and high perceived control to be more receptive to meat reduction messages.

Simulation Setup We simulate meat reduction dialogues between two generative agents (*Persuader* and *Recipient*) as shown in Figure 3. Every conversation is started by the Persuader and consists of 5 *rounds*, in which both agents produce 1 message each. To produce a new message, an agent goes through the following steps: it **(1) performs internal reflection**; Recipient agents **(2) answer a questionnaire**; and finally, an agent **(3) generates a response** to the other agent. The Persuader has explicit instructions to persuade the other agent to reduce meat, while the Recipient is instructed to adopt one of the personas shown in Table 1. In our initial simulations, we do not prompt the Persuader to follow a specific persuasion strategy. We explain the simulation process in the following and publish all respective prompts as part of our code under MIT license.[1]

To perform **internal reflection**, all agents receive a transcript of the conversation so far and are instructed to reflect on *which emotions were evoked*, *which core values were challenged or supported*, and *which questions or uncertainties there are*. Only the Recipient agent, then, fills out the **questionnaire** described above. Finally, each agent **generates a response** to the other agent. This response takes into account only the agent's persona, the transcript of the conversation so far, and the agent's internal reflection. To ensure that the questionnaire does not affect the simulation, we remove it from the LLM's context.

For our initial analyzes, we simulate 200 dialogues for both contrasting Recipient *personas* that are shown in Table 1, i.e., 400 conversations in total, with 10 messages exchanged in each dialogue. We extract 10 independent questionnaire answers from the Recipient persona to improve robustness. We opt for *open-weight* LLMs to facilitate replication (Barrie, Palmer, and Spirling 2024), and run our simulation with 3 distinct LLMs to investigate the impact of model size: `Llama 3.3 70B`, `Llama 3.1 8B`, and `Llama 3.2 3B`. For the questionnaire, we employ *structured outputs* tailored to the respective answer options, to ensure that we obtain valid responses even from the smallest model. We run all simulations with Llama's default temperature of 0.6 and use `vllm` with *automatic prefix caching* to improve model throughput—the combined runtime for all simulations on two NVIDIA H100 in parallel was $\approx$ 70 hours.

Validation To assess whether generative agent simulations can meaningfully replicate persuasive human dialogues about meat reduction, we implement a two-step validation strategy. First, we perform an **internal validation** by evaluating whether the agents' responses exhibit psychometric properties comparable to those observed in human data. Using established psychological constructs and measurement instruments, we assess internal consistency, convergent and discriminant validity, and distributional plausibility. Second, we conduct an **external validation** by comparing the simulated results against empirical data from prior human studies with similar persuasion tasks (see Table 2). This dual approach enables us to examine both the internal coherence and the empirical realism of the simulated dia-

Construct	Original Study	*n*
Attitude & Intention to Reduce Meat	Pabian et al. (2020)	47
Subjective Norms & Behavioral Control	Wyker and Davison (2010)	204
Social Attractiveness	Thürmer, Stadler, and McCrea (2022)	260
Social Closeness	Monin, Sawyer, and Marquez (2008)	70
Threat of Freedom	Dillard and Shen (2005)	196
Meat Attachment	Graça, Calheiros, and Oliveira (2015)	318

Table 2: **Studies with Human Participants Used for External Validation.** We compare published means against the survey responses from our generative agent simulation.

[1] https://github.com/dess-mannheim/MeatlessAgents

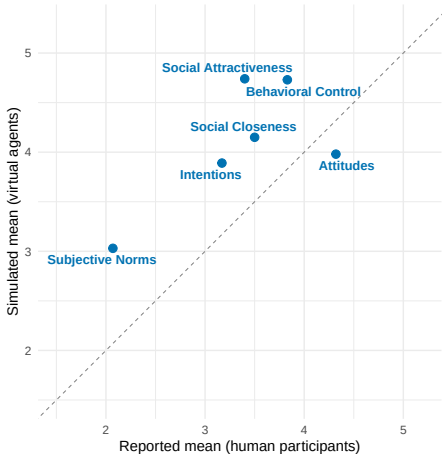


Figure 4: **Simulated Survey Responses Are on Average Close to Responses from Human Participants.** Mean values of our six central constructs based on human-reported data and simulated agent data. Each point represents one construct. The dashed diagonal line indicates perfect agreement between simulated and human means. All response scales range from 1 to 7.

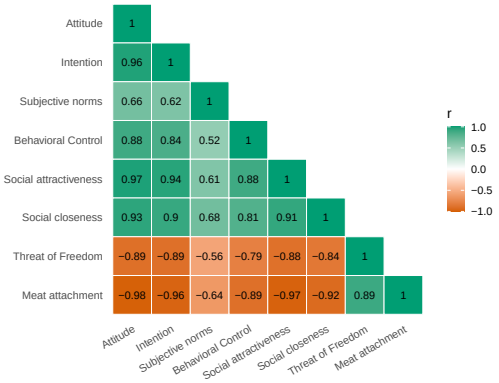


Figure 5: **Strong Pearson Correlations** between *Attitude*, *Intention*, *Behavioral Control*, *Social Attractiveness* and *Social Closeness* support **Convergent Validity**. Weaker correlations with theoretically more distinct constructs such as *Subjective Norms* support **Discriminant Validity**. *Threat of Freedom* and *Meat Attachment* were included to help assess if constructs form a distinct and coherent cluster.

logues. Crucially, this validation serves as a necessary foundation: only if our simulation proves **sufficiently reliable and valid** can it be used to **inform the design of future experiments**—with real human participants—under comparable conditions.

4 Results

To evaluate the quality of our simulation, we conducted several reliability and validity analyses. All results shown below refer to `Llama 3.3 70B`. Respective results for the smaller Llama models are provided in Appendix B. Note that with smaller Llama models, reliability and validity of the measures also decrease, but to a still acceptable level.

Descriptive Statistics & Distributions Appendix Figure 7 visualizes the distributions of the measures across all conversations and the two personas using boxplots. *Attitude* and *Intention* showed the highest median values and the highest variability in the sample. These wide distributions likely reflect the contrasting tendencies of the two personas, resulting in a broader spread of responses. *Behavioral Control*, *Social Attractiveness*, and *Social Closeness* revealed similarly high median values with more narrow spreads. This suggests more uniform perceptions across personas in terms of perceived control and social perception of the persuader. These constructs may have been modeled less distinctly between personas or are inherently perceived as less polarized. *Subjective Norms* exhibited both the lowest median and least variability, indicating that both personas shared similarly low perceptions of social expectations to reduce meat consumption. This uniformity could reflect limitations in the way social dynamics were depicted within the simulation framework.

To assess the internal consistency of the measured constructs, we calculated Cronbach's Alpha coefficients for each scale. The results indicated excellent reliability for most constructs (e.g. *Meat Attachment*, *Attitude*, *Intention*, *Social Attractiveness*, and *Social Closeness* $\alpha s \geq .98$). In contrast, *Subjective Norms* ($\alpha = .76$) and *Behavioral Control* ($\alpha = .70$) showed comparatively lower but still acceptable internal consistency. A detailed overview of the reliability coefficients, including comparisons with the values obtained from human samples, is provided in Appendix Table 3.

Comparing Simulated and Human Data To assess the similarity between simulated and human data, we compared the mean values of key constructs with those reported in previous studies using the same measures with human participants. As we did not have access to the original datasets, the comparisons were based on published mean values. Table 2 shows the original studies belonging to the respective constructs with their respective number of human participants. A scatterplot of the simulated versus reported means is shown in Figure 4. Most constructs produced higher mean values in the simulation compared to human data, with the exception of *Attitudes*, where the simulated mean was slightly lower. The largest discrepancy was observed for *Subjective Norms*, which were generally low in both cases, but especially in the human samples. In general, the points cluster closely around the identity line, indicating a high degree of similarity between the simulated and human-generated data, particularly in the relative positioning of the constructs. This supports the validity of the simulation in capturing psychologically plausible response patterns. However, these findings should be interpreted as preliminary evidence.

Construct Validity To further assess the construct validity of the simulated responses, we examined the intercorrelations between all measured variables, as visualized in Figure 5. The results reveal strong positive correlations between constructs that are theoretically linked, such as *Attitude*, *Intention*, *Social Attractiveness*, and *Social Closeness* (e.g., $r = .96$ between Attitude and Intention). This supports convergent validity. Correlations with less strongly related constructs, such as relations with *Subjective Norms*, are notably lower (e.g., $r = .52$ between Behavioral Control and Subjective Norms), supporting discriminant validity. These two constructs were deliberately included as control variables to assess whether the central constructs form a conceptually distinct cluster. However, some correlations between central variables are exceptionally high (approaching or exceeding $r = .95$), which is uncommon in empirical data and may reflect the polarized nature of the simulated personas. This could lead to overly homogeneous response patterns, limiting generalizability, and possibly inflating convergence artificially. Future simulations could benefit from a broader and more nuanced range of personas to capture more natural variance.

Persuasion Effectiveness Figure 6 shows the reported intention to reduce meat consumption—our main outcome—and each point represents the mean response of a simulated Recipient across the 3 respective questionnaire items and 10 repeated survey responses. As expected, the Hard to Persuade Recipient indicates a much smaller intention for meat reduction than the Easy to Persuade Participant. For `Llama 3.3 70B`, we also observe much more variance in the responses from the Hard to Persuade Recipient. Surprisingly, during the first rounds of the conversation, the simulated survey responses for the Hard to Persuade Recipient show a decreasing intention to reduce meat consumption. This could be interpreted as a backlash to the persuasion attempt, a pattern that is also observed with human participants (Shen, Sheer, and Li 2015). After approximately 3 conversation rounds, we observe an increasing mean intention for meat reduction, across both participants and all LLMs. This trend continues in round 5 for the Hard to Persuade Recipient and `Llama 3.3 70B`, which we take as a reason to extend future simulation to more rounds.

5 Conclusion

Our preliminary findings demonstrate that generative agents can engage in plausible discussions about meat reduction. However, limitations such as uniformity in subjective norm responses require further investigation. As next steps, we will extend the simulation to more rounds and run targeted simulations in which we instruct the Persuader to employ a specific persuasion strategy. Additionally, we will investigate more diverse personas.

Encouraging more people to reduce meat consumption remains a vital goal for both individual well-being and planetary health. Generative agent-based simulations offer great potential for large-scale exploration of persuasion strategies for meat reduction, which will then be tested in subsequent studies with human participants.

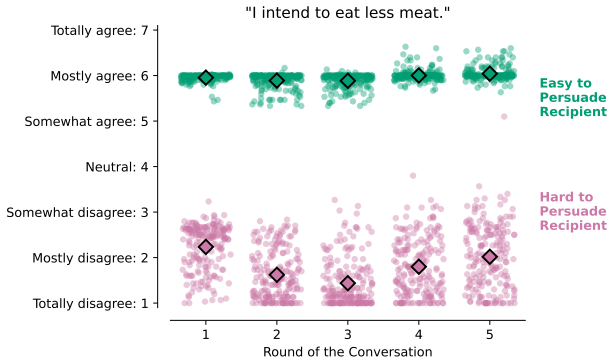


Figure 6: **Recipients' Intention to Reduce Meat.** Combined survey responses after each conversation round, where ◊ indicates the mean over all conversations per round. Following our expectations, the Easy to Persuade Recipient responds with a generally high intention to reduce meat. Over the first rounds of the conversation, the Hard to Persuade Recipient shows an initially decreasing intention to reduce meat, similar to human participants (Shen, Sheer, and Li 2015).

Ethical Considerations

From an ethical standpoint, our findings underscore both the potential and risks associated with using large language models for persuasive communication. While our study focuses on promoting reduced meat consumption—a goal aligned with individual and planetary health—the underlying persuasive strategies and technical setup are generalizable to a wider range of topics.

This generalizability is methodologically promising, but it also raises important ethical considerations. The same techniques used to encourage socially beneficial behavior change could potentially be repurposed to mislead or exert undue influence. However, effective persuasion still depends on a person's openness to change—people are unlikely to be swayed to adopt behaviors they fundamentally reject, especially those they view as ethically wrong. Moreover, we would expect persuasive dynamics in humans to be more complex and often less effective, given that conversational agents are inherently designed to be accommodating, assistive, and agreeable—traits that may amplify the success of persuasive strategies in artificial contexts compared to real human interactions. Nonetheless, without clear guidelines and safeguards, there remains a real risk of these technologies being used to promote misinformation, ideological agendas, or exploitative commercial practices.

We anticipate that applications of persuasive AI will continue to expand rapidly. Therefore, we believe it is essential to publish and critically examine what LLMs are already capable of in this space. Our goal is not only to inform future research, but to contribute to a proactive and transparent discourse about the ethical governance of persuasive AI technologies.

References

Aher, G.; Arriaga, R. I.; and Kalai, A. T. 2024. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies.

Ahnert, G.; Pellert, M.; Garcia, D.; and Strohmaier, M. 2025. Extracting Affect Aggregates from Longitudinal Social Media Data with Temporal Adapters for Large Language Models. ArXiv:2409.17990 [cs].

Ajzen, I. 1991. The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2): 179–211.

Argyle, L. P.; Busby, E. C.; Fulda, N.; Gubler, J. R.; Rytting, C.; and Wingate, D. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3): 337–351.

Bail, C. A. 2024. Can Generative AI improve social science? *Proceedings of the National Academy of Sciences*, 121(21): e2314021121.

Barrie, C.; Palmer, A.; and Spirling, A. 2024. Replication for Language Models.

Berndsen, M.; and Van Der Pligt, J. 2005. Risks of meat: the relative impact of cognitive, affective and moral concerns. *Appetite*, 44(2): 195–205.

Bisbee, J.; Clinton, J.; Dorff, C.; Kenkel, B.; and Larson, J. 2023. Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. preprint, SocArXiv.

Bolderdijk, J. W.; and Cornelissen, G. 2022. "How do you know someone's vegan?" They won't always tell you. An empirical test of the do-gooder's dilemma. *Appetite*, 168: 105719.

Carfora, V.; Catellani, P.; Caso, D.; and Conner, M. 2019. How to reduce red and processed meat consumption by daily text messages targeting environment or health benefits. *Journal of Environmental Psychology*, 65: 101319.

Davidson, T. R.; Veselovsky, V.; Josifoski, M.; Peyrard, M.; Bosselut, A.; Kosinski, M.; and West, R. 2024. Evaluating Language Model Agency through Negotiations. ArXiv:2401.04536 [cs].

Dillard, J. P.; and Shen, L. 2005. On the Nature of Reactance and its Role in Persuasive Health Communication. *Communication Monographs*, 72(2): 144–168.

Godfray, H. C. J.; Aveyard, P.; Garnett, T.; Hall, J. W.; Key, T. J.; Lorimer, J.; Pierrehumbert, R. T.; Scarborough, P.; Springmann, M.; and Jebb, S. A. 2018. Meat consumption, health, and the environment. *Science*, 361(6399): eaam5324.

Graça, J.; Calheiros, M. M.; and Oliveira, A. 2015. Attached to meat? (Un)Willingness and intentions to adopt a more plant-based diet. *Appetite*, 95: 113–125.

Harguess, J. M.; Crespo, N. C.; and Hong, M. Y. 2020. Strategies to reduce meat consumption: A systematic literature review of experimental studies. *Appetite*, 144: 104478.

Hewitt, L.; Ashokkumar, A.; Ghezae, I.; and Willer, R. 2024. Predicting Results of Social Science Experiments Using Large Language Models.

Judge, M.; Bouman, T.; Steg, L.; and Bolderdijk, J. W. 2024. Accelerating social tipping points in sustainable behaviors: Insights from a dynamic model of moralized social change. *One Earth*, 7(5): 759–770.

Kwasny, T.; Dobernig, K.; and Riefler, P. 2022. Towards reduced meat consumption: A systematic literature review of intervention effectiveness, 2001–2019. *Appetite*, 168: 105739.

Monin, B.; Sawyer, P. J.; and Marquez, M. J. 2008. The rejection of moral rebels: Resenting those who do the right thing. *Journal of Personality and Social Psychology*, 95(1): 76–93.

Pabian, S.; Hudders, L.; Poels, K.; Stoffelen, F.; and De Backer, C. J. S. 2020. Ninety Minutes to Reduce One's Intention to Eat Meat: A Preliminary Experimental Investigation on the Effect of Watching the Cowspiracy Documentary on Intention to Reduce Meat Consumption. *Frontiers in Communication*, 5: 69.

Park, J. S.; Popowski, L.; Cai, C.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2022. Social Simulacra: Creating Populated Prototypes for Social Computing Systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, 1–18. Bend OR USA: ACM. ISBN 978-1-4503-9320-1.

Romo, L. K.; and Donovan-Kicken, E. 2012. "Actually, I Don't Eat Meat": A Multiple-Goals Perspective of Communication About Vegetarianism. *Communication Studies*, 63(4): 405–420.

Saba, A.; and Di Natale, R. 1998. A study on the mediating role of intention in the impact of habit and attitude on meat consumption. *Food Quality and Preference*, 10(1): 69–77.

Sen, I.; Lutz, M.; Rogers, E.; Garcia, D.; and Strohmaier, M. 2025. Missing the Margins: A Systematic Literature Review on the Demographic Representativeness of LLMs.

Sheeran, P.; Milne, S.; Webb, T. L.; and Gollwitzer, P. M. 2005. Implementation Intentions and Health Behaviour. In Conner, M., ed., *Predicting Health Behaviour: Research and Practice with Social Cognition Models*, 276–323. New York: Open Univ. Pr. ISBN 978-0-335-21177-7.

Shen, F.; Sheer, V. C.; and Li, R. 2015. Impact of Narratives on Persuasion in Health Communication: A Meta-Analysis. *Journal of Advertising*, 44(2): 105–113.

Taubenfeld, A.; Dover, Y.; Reichart, R.; and Goldstein, A. 2024. Systematic Biases in LLM Simulations of Debates. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 251–267. Miami, Florida, USA: Association for Computational Linguistics.

Thürmer, J. L.; Stadler, J.; and McCrea, S. M. 2022. Intergroup Sensitivity and Promoting Sustainable Consumption: Meat Eaters Reject Vegans' Call for a Plant-Based Diet. *Sustainability*, 14(3): 1741.

Tjuatja, L.; Chen, V.; Wu, S. T.; Talwalkar, A.; and Neubig, G. 2024. Do LLMs exhibit human-like response biases? A case study in survey design. ArXiv:2311.04076 [cs].

Törnberg, P.; Valeeva, D.; Uitermark, J.; and Bail, C. 2023. Simulating Social Media Using Large Language

Models to Evaluate Alternative News Feed Algorithms. ArXiv:2310.05984 [cs].

Vaccaro, M.; Caoson, M.; Ju, H.; Aral, S.; and Curhan, J. R. 2025. Advancing AI Negotiations: New Theory and Evidence from a Large-Scale Autonomous Negotiations Competition. Version Number: 1.

Wang, A.; Morgenstern, J.; and Dickerson, J. P. 2024. Large language models cannot replace human participants because they cannot portray identity groups. ArXiv:2402.01908 [cs].

Wyker, B. A.; and Davison, K. K. 2010. Behavioral Change Theories Can Inform the Prediction of Young Adults' Adoption of a Plant-based Diet. *Journal of Nutrition Education and Behavior*, 42(3): 168–177.

A Questionnaires

In the following section, the questionnaires used in this study to assess all measured outcomes are described. An overview of the six key constructs is provided in Figure 2.

To assess the *persuasiveness* of the presented arguments, participants were asked to rate the item *"How persuasive did you find this argument?"* on a 7-point Likert scale.

Behavioral change was measured with the item *"To what extent did this argument make you consider changing your behavior?"*, also using a 7-point Likert scale.

Attitudes toward meat reduction were assessed using a four-item semantic differential scale based on Berndsen and Van Der Pligt (2005), rated on a 7-point scale:

- Pleasant – Unpleasant
- Useful – Useless
- Favorable – Unfavorable
- Good – Bad

Intentions to reduce meat consumption were captured with a three-item scale adapted from Graça, Calheiros, and Oliveira (2015), using a 7-point Likert scale:

- I intend to eat less white meat.
- I intend to eat less red meat.
- I intend to eat less processed meat.

Subjective norms were measured with four items adapted from Berndsen and Van Der Pligt (2005), each rated on a 7-point scale:

- How much do you feel your friends want you to reduce your meat consumption in the next year?
- How much do you feel your family want you to reduce your meat consumption in the next year?
- How much do you feel health experts want you to reduce your meat consumption in the next year?
- How much do you feel your colleagues want you to reduce your meat consumption in the next year?

Perceived behavioral control was assessed via two items based on Wyker and Davison (2010), using a 7-point Likert scale:

- How much personal control do you feel you have about reducing your meat consumption in the next year?
- To what extent do you see yourself as being capable of reducing your meat consumption in the next year?

Attachment to meat was measured using the 20-item Meat Attachment Questionnaire (MAQ) developed by Graça, Calheiros, and Oliveira (2015), on a 5-point Likert scale. Items include:

- To eat meat is one of the good pleasures in life.
- Meat is irreplaceable in my diet.
- According to our position in the food chain, we have the right to eat meat.
- I feel bad when I think of eating meat.
- I love meals with meat.
- To eat meat is disrespectful towards life and the environment.
- To eat meat is an unquestionable right of every person.
- Meat consumption is crucial to my balance.
- A full meal is a meal with meat.
- I'm a big fan of meat.
- If I couldn't eat meat I would feel weak.
- If I was forced to stop eating meat I would feel sad.
- By eating meat I'm reminded of the death and suffering of animals.
- Eating meat is a natural and undisputable practice.
- I don't picture myself without eating meat regularly.
- Meat sickens me.
- I would feel fine with a meatless diet.
- Meat consumption is a natural act of one's affirmation as a human being.
- A good steak is without comparison.
- Meat reminds me of diseases.

The *perceived threat to freedom* was assessed using four items from Dillard and Shen (2005), rated on a 7-point Likert scale:

- The person tried to make a decision about my own diet for me.
- The person tried to influence me and my diet.
- The person threatened my freedom to decide about my own diet.
- The person tried to pressure me regarding my diet.

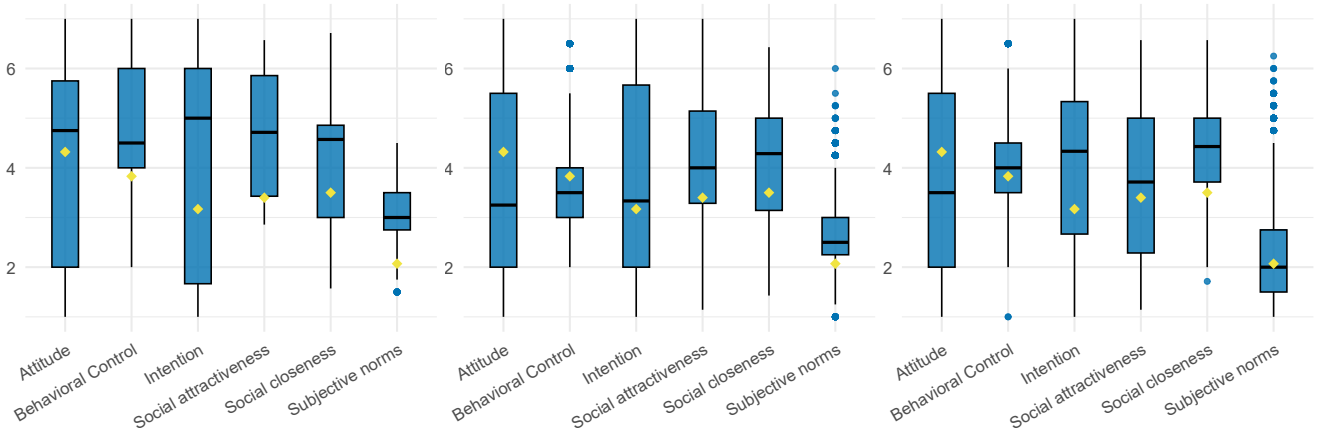
Social attractiveness was measured with seven items adapted from Thürmer, Stadler, and McCrea (2022), using a 7-point Likert scale:

- To what extent do you think your interaction partner is intelligent?
- ...trustworthy?
- ...friendly?
- ...open?
- ...likeable?
- ...respectable?
- ...interesting?

Social closeness was measured via seven items based on Monin, Sawyer, and Marquez (2008), rated on a 7-point Likert scale:

- I would like to get to know the other person better.
- I would like to have lunch with the other person sometime.
- I would like to work together with the other person on a task.
- I would like to have a conversation with the other person about food.
- I felt emotionally close to the other person.
- I wanted to have a conversation with the other person.
- I felt like I made friends with the other person.

B Additional Results

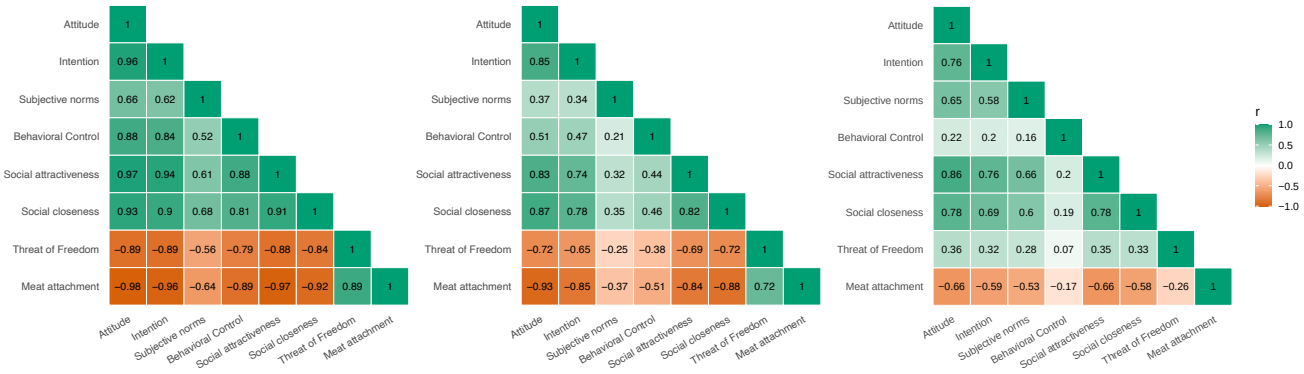


(a) **Llama3.3 70B** (b) **Llama 3.1 8B** (c) **Llama 3.2 3B**

Figure 7: **Smaller `Llama` models show less consistent construct response distributions.** The distribution of model-generated responses is shown across `Llama models (70B, 8B, 3B)` for key psychological constructs. Yellow diamonds indicate the mean from human samples for each construct. Compared to the 70B model, smaller models exhibit reduced variability and more outliers in several constructs (e.g., *Behavioral Control*, *Subjective Norms*), suggesting a loss of distributional alignment with human responses as model size decreases.

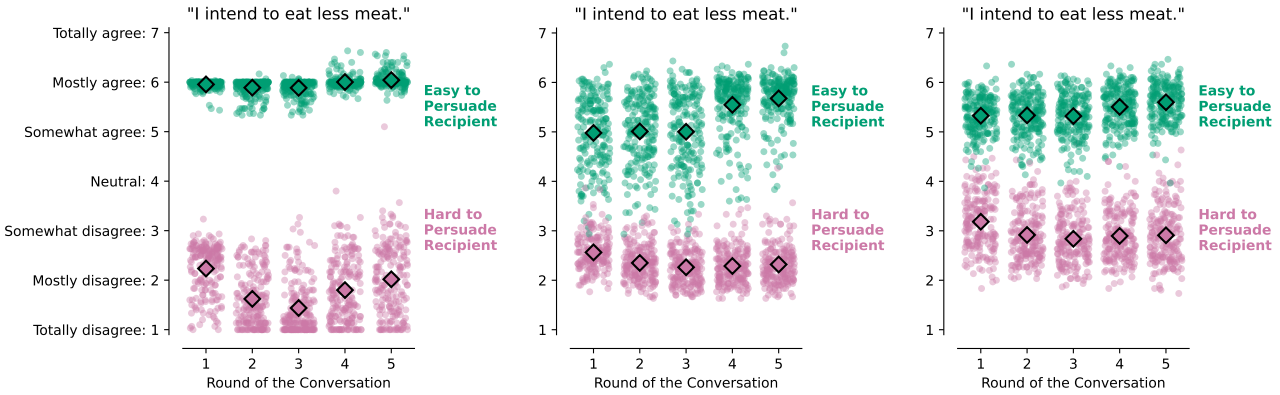
Scale	#items	Llama 3.3 70B	Llama 3.1 8B	Llama 3.2 3B	Human Samples	Original Study
Theory of Planned Behavior						
Attitude	4	.99	.97	.92	.86	Pabian et al. (2020)
Intention	3	.99	.93	.88	.85	Pabian et al. (2020)
Subjective Norms	2	.76	.59	.78	.75	Wyker and Davison (2010)
Behavioral Control	2	.70	.66	.68	.70	Wyker and Davison (2010)
Social Costs						
Social Attractiveness	7	.98	.93	.96	.91	Thürmer, Stadler, and McCrea (2022)
Social Closeness	7	.98	.94	.82	.91	Monin, Sawyer, and Marquez (2008)
Additional Constructs						
Threat of Freedom	4	.93	.87	.60	.84	Dillard and Shen (2005)
Meat Attachment	20	.99	.98	.71	.92	Graça, Calheiros, and Oliveira (2015)

Table 3: **Reliability of psychological constructs decreases with smaller Llama models.** Cronbach's Alpha values for each scale are compared across `Llama models (3.3 70B, 3.1 8B, 3.2 3B)` and human participant data to assess internal consistency. While larger models consistently show high reliability ($\alpha \geq .70$), several scales fall below the acceptable threshold in smaller models (shown in red)—most notably *Behavioral Control*, and *Threat of Freedom*. This suggests a degradation in the coherent representation of multi-item constructs as model size decreases.



(a) **Llama 3.3 70B** (b) **Llama 3.1 8B** (c) **Llama 3.2 3B**

Figure 8: **Correlation strength weakens with smaller Llama models, while overall construct patterns remain stable.** To assess construct validity, we examine Pearson correlations among psychological constructs across three Llama models of varying sizes. While general correlation patterns—including the two added constructs *Threat of Freedom* and *Meat Attachment*—remain largely stable, the strength of correlations systematically decreases as model size reduces. This suggests a loss of representational consistency in smaller models.



(a) **Llama 3.3 70B** (b) **Llama 3.1 8B** (c) **Llama 3.2 3B**

Figure 9: **Recipients' Response to: "I Intend to Eat Less Meat"**, obtained from simulations run with `Llama 3.3 70B`, `Llama 3.1 8B` and `Llama 3.2 3B`. Combined survey responses after each conversation round, where ◊ indicates the mean over all conversations per round. Compared to `Llama 3.3 70B`, we observe much greater variance in survey responses for simulations performed with smaller models. The intention of the Hard to Persuade Recipient stays more stable over time for the smaller models, while the intention of the Easy to Persuade Recipient shows a stronger increase in round 4, especially for `Llama 3.1 8B`.

Lexpansion: Evaluating Dictionary Based Lexicon Expansion for Social Media Analysis

Mohamed Bahgat[1], **Steven R Wilson**[2], **Walid Magdy**[1]
[1]University of Edinburgh
[2]University of Michigan-Flint
m.bahgat@ed.ac.uk, steverw@umich.edu, wmagdy@inf.ed.ac.uk

Abstract

Lexicons are indispensable tools for textual analysis through labelled term associations. Despite their utility, lexicons are static and require manual effort to curate and maintain. Regular updates are essential to stay relevant amid semantic shifts and neologisms during language evolution. In this work, we explore the potential of supervised learning for expanding lexicons using dictionaries. We study the effect of using dictionaries with varying properties such as noise, size, labels, structure and curation method. Definitions are used as input features to a transformer model (BERT) that assigns categories to terms. We analyse the expansions using varying English dictionaries and lexicons for estimated accuracy, coverage and labelling consistency and apply the expanded versions to a downstream task. Our analyses show dictionary based expansion is a robust approach. We release our expanded lexicons, code, and pretrained models.

1 Introduction

Lexicons have been broadly used for text analysis in domains such as politics (Elbagir and Yang 2019), social attitudes (Gao and Huang 2017), mental health (Zanwar et al. 2023) and more. Even in the era of LLMs, lexicons are useful to interpret results, build lightweight models or add emphasis on relevant features (Li et al. 2020).

However, lexicons are static, yet language evolves over time. Various societal and demographic factors were shown to affect language change (Labov 2011; Milroy and Milroy 2017). These changes are amplified by social media both syntactically (Maity et al. 2016) and semantically (McGillivray et al. 2022). Only a limited number of lexicons are updated to cope with language changes, like LIWC, (Pennebaker, Francis, and Booth 2001; Pennebaker et al. 2007, 2015; Boyd et al. 2022) which has been revised at 7-year intervals. An alternative is using resources such as dictionaries to expand lexicons (Bahgat, Wilson, and Magdy 2022). Dictionaries are updated frequently and capture language changes (Nguyen, McGillivray, and Yasseri 2018). For example, in March 2023, Oxford English Dictionary (OED) added 700 new terms like "deepfake" and "groomzilla" while revising 1,400[1]. Other dictionaries are even more sensitive. In Urban Dictionary (UD), a crowdsourced dictionary, "selfie" was first defined in September 2009 versus August 2013 for OED (Nguyen, McGillivray, and Yasseri 2018). UD contains a variety of slang terms and is useful in relevant applications (Wilson et al. 2020a).

Downstream tasks that analyse content from social media would benefit largely from such increased coverage especially for terms that commonly surface on these forums.

In this work, we present "Lexpansion", a framework for adding new terms to lexicons by obtaining candidates from dictionaries and analysing results. We specifically address categorical lexicons which comprise lists of terms and their corresponding categories. Dictionaries used in our work are lists of words and corresponding senses. We describe senses as definitions composed of meaning and any examples provided for a given term. Dictionary definitions that correspond to lexicon terms are used to train a model to classify other unseen dictionary terms into one of the categories. Models are evaluated on held-out lexicon terms and their corresponding dictionary definitions.

Our work focuses on understanding the effect lexicon and dictionary properties have on dictionary-based lexicon expansion. We propose a framework to (1) train a model once then extract candidates from batches of data without the need to retrain the model, (2) compare results from different dictionaries for lexicon expansion, (3) check category assignments for the same and across different dictionaries for multiple definitions for the same term, and (4) study the impact of using expanded lexicons on downstream tasks. Our method targets lowering lexicon update costs and enables adaptation to language shifts with affordable, accessible models. We discuss English resources, but given our work does not rely on language specific features, it applies to other languages. Our code, models and expanded lexicons are available publicly[2].

2 Related Work

Lexicons are used in sociolinguistics and computational social science including analysing political speeches (Liu and Lei 2018), elections discussions on social media (Tumasjan et al. 2010), hate speech (ElSherief et al. 2018), social hierarchies and cohorts (Kacewicz et al. 2014), mental health

[1]https://www.oed.com/information/updates/march-2023/

[2]https://github.com/mabahgat/lexpansion

Proceedings of the First Workshop on Integrating NLP and Psychology to Study Social Interactions (NLPSI) @ICWSM '25, pages 11–27

disorders (Coppersmith, Dredze, and Harman 2014), and emotions versus values (Bahgat, Wilson, and Magdy 2020). Lexicons are used as additional features with machine learning models (Biggiogera et al. 2021; Cutler et al. 2021), augmented during training neural layers for domain specific information (Lei, Yang, and Yang 2018; Li et al. 2020) or emphasize features from a target task (Baziotis et al. 2018).

2.1 Early Automated Lexicon Creation and Expansion

Neologisms often emerge in response to cultural, technological, or social changes (Čilić and Plauc 2021), which poses challenges to lexicons' relevance, including within non-English languages(Bies et al. 2014; Cassidy et al. 2022). Generating or expanding lexicons automatically provides a solution. For example, for morphological lexicons, Oliver, Castellón, and Màrquez (2003) and Oliver and Tadic (2004) expanded two morphologically rich languages, Russian and Croatian. Kaufmann and Pfister (2010) attempted to predict morphosyntactic features for out of vocabulary using a statistical model and applied it to German. Watkinson and Manandhar (1999), Thomforde and Steedman (2011) and Wang, Kwiatkowski, and Zettlemoyer (2014) expanded automatically Combinatory Categorial Grammar lexicons. There were also attempts to expand WordNet in French Sagot and Fišer (2012). Similarly, Seppälä, Barque, and Nasr (2012) used rules to generate a semantic lexicon for French adjectives while Cai and Yates (2013) proposed expanding the lexicon used for semantic parsing through pattern matching.

As for categorical lexicons, a common approach is adding synonyms of current entries from resources like WordNet (Miller 1998) as in the case of Badaro et al. (2018) which expanded *DepecheMood* (Staiano and Guerini 2014) creating *EmoWordNet* containing 1.8 times more terms. Similarly, Shaikh et al. (2016) added hyponyms to get 5 times more terms. WordNet is manually curated with high quality terms and relations but suffers from being stale[3] unlike dictionaries that we use in this work as source of candidates that gets updated with new terms.

2.2 Dictionaries

Other work used dictionaries for expansion. Bahgat, Wilson, and Magdy (2022) used Urban Dictionary (UD) to expand LIWC2015 categories. The terms were classified by using their corresponding definitions. However, that work is limited to a single dictionary and does not investigate the impact of lexicon and dictionary properties. We obtained better results using Wiktionary with LIWC2015 and only slightly lower results using OPTED, a much smaller dictionary. Similarly, Wu, Morstatter, and Liu (2018) used UD to build *Slang Sentiment Dictionary* (*SlangSD*) lexicon. The approach leverages sentiment entries from existing lexicons as seed terms and assigns sentiment strength to new candidate terms based on their co-occurrence with seed terms in UD social media posts. However, this method focuses solely on sentiment and is susceptible to errors due to co-occurrence reliance. Wiktionary (WK) was also used for creating and expanding lexicons. Bajčetić and Declerck (2022) added English pronunciation information from WK to WordNet to disambiguate terms with the same spelling but different meanings. Sajous et al. (2010) extended the relations in WK. The work computed graphs derived from Bengali WK including translations, synonyms, and glosses to identify synonym relations via random walks through these graphs. Those relations were then manually reviewed. Di Natale and Garcia (2024) identified new terms by leveraging shared senses derived from aligning terms in language-to-language dictionaries. That is, when a term from a source language is mapped to a term from a destination language, all terms in the source language that correspond to the destination language term are considered as candidates. But language-to-language dictionaries are not expected to have the same coverage and fast evolution as dictionaries defining terms in the same language.

[3]Latest update was 2006 as mentioned in https://wordnet.princeton.edu/download/current-version

2.3 Embeddings and Tailored Features

More approaches rely on word embeddings where terms are represented by vectors. Vectors for semantically similar terms have smaller differences (Mikolov et al. 2013). That was exploited by Empath (Fast, Chen, and Bernstein 2016) to select candidate terms that are neighbouring the existing terms in LIWC lexicon in the embeddings space for manual review. The space was built using fictional text with more emotional content which is more relevant to LIWC, a psychometric lexicon. Similarly, Wilson, Shen, and Mihalcea (2018) created a lexicon by starting with seed terms for personal values and then adding candidate terms neighbouring the seed terms in the embeddings space. Again, the terms were reviewed manually to increase quality.

In other work, tailored features were used to classify terms. (Avancini et al. 2006) expanded lexicons for new domains by computing *tf/idf* feature vectors for terms from domain corpora that were fed into *ADABoost.MH*KR classifier to label them. The resulting lexicon had higher coverage than a test target domain corpus. In another case, Staiano and Guerini (2014) construct a document-by-emotion and word-by-document matrices to estimate the labels for each word to expand *DepecheMood* to create *DepecheMood++*. Those matrices were estimated from web articles where users labelled each with 8 emotions. However, these methods require additional specialized corpora to compute features for new terms rather than only relying on dictionary definitions.

Compared to using dictionaries where we can build a model once and then classify new terms, adding new terms in the above approaches would require rebuilding the embeddings or retraining the classifiers to maintain accuracy.

2.4 Transformers and Large Language Models

Lexpansion assigns class labels to word sequences similar to, for example, Word Sense Disambiguation (WSD) (Bevilacqua et al. 2021). Huang et al. (2019) concatenated context and gloss and fine tuned BERT (Devlin et al. 2019) to predict if the gloss is correct while Yap, Koh, and Chng (2020) predicted the gloss directly by using glosses from WordNet and manually annotated corpora, but new glosses

or words will not be found in WordNet. Also, manually annotated corpora will rarely be updated relative to dictionaries. In another case, ChatGPT was used to annotate terms automatically (Marcondes et al. 2024). These annotations are then reviewed manually. LLMs do a "fair" job in those experiments but still require manual review.

3 Lexpansion

Lexpansion leverages dictionary term definitions as inputs to predict a lexicon category using a classifier. An experiment expanding a lexicon can train multiple models, each with a different dictionary, allowing us to compare and analyse the results. We formulate lexicon expansion as a multi-class classification problem where a term is assigned a lexicon category based on the term's corresponding definition.

3.1 An Expansion Run

A lexpansion *run* attempts to expand a lexicon. A run can be viewed as a pipeline which takes a lexicon to expand, one or more dictionaries as source of candidates, exclusion lists used to filter out entries in dictionaries, and a configuration file for model parameters. The configuration file includes lexicon categories to be expanded, dictionaries used, training and testing set sizes, which exclusion lists to use, threshold to filter out dictionary entries if applicable, model parameters and model labelling confidence threshold applied when generating the final expanded lexicons. The pipeline steps are processing dictionary entries, creating the training and testing sets, training the model, running the evaluation on the test set, generating a report and optionally running on the full dictionary to generate an expanded version of the lexicon. If the run involved multiple dictionaries, the evaluation report would provide a comparison. Also optionally, output for multiple models trained on different dictionaries can be used in majority voting when assigning labels to terms.

3.2 Lexicons

We expand three lexicons that vary in size and complexity. Appendix A lists example entries.

Values (Wilson, Shen, and Mihalcea 2018) is a lexicon that captures personal values. The authors created an initial hierarchy with seed terms. More candidate terms are proposed by selecting neighbouring terms to seed terms from a word embeddings space. The resulting list of terms are expected to be noisy, so crowd-sourced human annotators verified the terms and their assigned categories. The categories we selected for expansion are "life", "parents", "truth", "religion", "social", "feeling-good", "children", "animals", "learning", "order" and "accepting-others". These are top level categories that were represented enough in the largest of our selected dictionaries. We applied such constraint to obtain enough training and testing examples.

LIWC2015 (Pennebaker et al. 2015) is an iteration of Linguistic Inquiry and Word Count (LIWC) which measures various psychometric properties in input text. LIWC is a hierarchical lexicon that links terms to categories that represent different properties: Linguistic and Grammar Dimensions; physiological, cognitive, perceptual, and biological processes; drives; time orientation; relativity; personal concerns; and informal language. Under each, there can be subcategories. We selected a subset of the top level categories. These were: "affect", "social", "cogproc" (Cognitive Process), "percept" (Perceptual Process), "bio" (Biological Process), "drives", "relativ" (Relativity), "pconcern" (Personal Concerns), and "informal". Function words category with subcategories of personal pronouns, adverbs, numbers and others was left out as new terms can be identified by other means. Note that LIWC2015 uses wildcards indicated by "*" which only appears at the end of a term. For example, "enjoy*" matches "enjoy" and "enjoyable". Wildcards expand coverage but introduce noise so we opted not to use them. LIWC2015 only matched unigrams.

Lexicon	Values	LIWC2015	LIWC22
Count	1,664	6,547	12,400

Table 1: Number of all terms in each lexicon.

LIWC22 (Boyd et al. 2022) is the next iteration after LIWC2015. LIWC22 has a revised set of categories. Some categories were added, such as *Culture*, *Personal Concerns* was renamed as *Life Style*, and *Relativity* was deleted altogether. LIWC22 has also increased coverage using nearly twice as many entries compared to LIWC2015 and more flexible wild cards appearing at any place in terms. LIWC22 also includes n-grams as "civil unrest" or "not in the mood". We expand a subset of LIWC22 top categories: "Culture", "Conversation", "Cognition", "Drives", "Social", "Physical", "Affect", "Lifestyle" and "Perception". Similar to LIWC2015 case, we excluded other categories that can be detected by other means. Table 1 compares lexicon term counts.

3.3 Dictionaries

Dictionaries contain a list of terms and their definitions as well as other information. A term belonging to a dictionary can have multiple definitions. A definition can be composed of multiple parts, but in most cases it will have a meaning and example. Parts are concatenated together to form a definition. We used three dictionaries that we detail below.

Urban Dictionary (*UD*) is a crowd sourced online dictionary. Authors contribute by adding either new *terms* that did not exist or new *definitions* to already existing terms. Definitions are composed of a meaning and usage example(s). Some definitions are faithful such as "The feeling when your heart overflows with the joy of living and being able to just be free and have fun." for the term "happy." However, other definitions are noisy; some are opinionated, like "happy : something that no Gen Z'er experiences" or humorous, like "what UD wants you to buy all the time" as a definition of "mug"[4]. UD users are allowed to up-vote or down-vote a definition. Votes help definitions bubble up, but some highly voted definitions are still considered noise. A common case

[4]Referring to mug advertisements on UD website.

Dict.	Terms	Def.	DPT	TPD
UD	1,974,242	3,534,963	1.79	54
WK	1,065,301	1,376,430	1.29	17.4
OPTED	111,616	176,045	1.58	11.2

Table 2: For each dictionary: number of unique terms, number of definitions, Definitions per term (DPT) and Tokens per definition (TPD).

Dictionary	Values	LIWC2015	LIWC22
UD	608	3,725	81,627
WK	964	5,055	194,478
OPTED	670	3,174	15,463

Table 3: Counts of terms matched from each dictionary.

is proper nouns which usually describe a person known to the author with votes based on popularity rather than soundness. UD also has a significant amount of offensive content (Nguyen, McGillivray, and Yasseri 2018) and contains slang or informal definitions and terms (Wu, Morstatter, and Liu 2018) that are commonly used on social media. Therefore, it is sensitive to new terms which surface on social media (Wilson et al. 2020b) or news (Bahgat, Wilson, and Magdy 2022). UD has a significant number of terms that do not exist in formal common dictionaries.

Wiktionary (*WK*) is another crowd-sourced online dictionary. In contrast to UD, WK is highly moderated so definitions have higher quality and less noise. Definitions are also more detailed, providing some or all of definitions, examples, pronunciation, etymologies, translations and more. For our task, we use meanings and examples only. WK has more content compared to formal dictionaries, but fewer terms and definitions to UD, as shown Table 2.

The Online Plain Text English Dictionary (*OPTED*)[5], is a formal English dictionary. Compared to other formal dictionaries it smaller and not updated. Nevertheless, OPTED is a good representation of classical static dictionaries with formal content, less slang, and minimal noise representing the other side of spectrum compared to UD.

Table 2 compares the three dictionaries. Although OPTED is the smallest, OPTED lists terms not found in the other dictionaries as "testamentize" and "mollities" which are defined as "To make a will" and "With a narrow mouth, as the shell of certain gastropods." respectively. UD has the highest number of terms and definitions, but there are terms that exist in OPTED and WK but not in UD such as "innoxious" which is defined as "Free from crime; pure; innocent.". Also, Table 3 shows lexicon matches for each dictionary.

Dictionaries are processed to filter out noisy entries. Names and stop words were also excluded. More lists can be used to filter out other terms such as offensive or biased ones although this would depend on the application as in some cases such terms would be needed.

[5]OPTED is based on "The Project Gutenberg text of Webster's Unabridged Dictionary" which contains the 1913 US Webster's Unabridged Dictionary. Can be found at https://www.mso.anu.edu.au/~ralph/OPTED/

3.4 Selecting Train and Test Samples

Our superset of samples includes all dictionary definitions for terms matched by a lexicon along with the lexicon category for each term. There can be multiple definitions for the same term and each is considered as a separate training sample. That is, dictionary definitions composed of the meaning (and optionally, examples) are used as input to the model while the lexicon categories are the target labels. In case we can have a metric for dictionary definition quality, low quality definitions are excluded before selecting any samples for either the training or testing sets.

The samples are divided into training and testing sets, ensuring that all samples linked to terms assigned to the test set remain exclusively within it, i.e., if a term in the test set has multiple definitions, none of them appear in the training set. The terms for the test set are selected at random and its distribution of labels matches the distribution of labels in the overall available samples. For terms with multiple definitions, if there is a quality metric then the best one is selected otherwise a definition is selected at random.

Next, terms are selected for the training set. The remaining terms and their corresponding definitions are included in the training set. In UD case, only the top 10 definitions in terms of the difference between the up votes and down votes for each term are included in the training set. Definitions with more down votes than up votes are also excluded even if those are in the top 10. This strategy allows us to exclude noisy definitions that are expected in UD.

An experiment for expanding a lexicon can use multiple dictionaries to allow comparing the results from each dictionary. In that case, a single set of terms are selected where all these terms exist in all the used dictionaries. The test set used to evaluate each model built with different dictionary uses the definitions from that dictionary for the selected terms. The training set is curated the same way as a single dictionary.

3.5 Model

We use pretrained BERT (Devlin et al. 2019) "bert-base-uncased" with a linear layer. All models were fine tuned with $5e^{-5}$ learning rate and 5 epochs. Our implementation allows swapping to other model architectures. We chose BERT for its availability, popularity given the relatively good performance, and ability to train on less expensive hardware. However, our overall framework can be used with any supervised text classification model.

3.6 Tuning Precision

The precision of labels assigned by the model can be tuned by studying Precision and Recall against the value of model confidence assigning these labels. Picking a higher confidence threshold will filter out terms assigned labels with confidence scores below that threshold, resulting in fewer terms added to the lexicon, but with improved precision. The threshold can be selected depending on the desired outcome.

3.7 Using Majority Vote

Precision can be further improved by combining label assignments to a term from different dictionary models for

the same lexicon. Two methods were considered: majority agreement and unanimous model agreement, where a new term will not be added unless multiple dictionary models agree on the same class for a given term. This would result in fewer added terms but may improve precision.

4 Experimental Setup

4.1 Baselines

To assess the proposed method, we ran two baselines to answer different questions. One is whether a simple model achieves comparable performance and the other is whether large language models (LLMs) perform better on expansion task without any fine tuning. We chose those baselines as opposed to previous work as previous work either expanded lexicons with a limited categories or does not address evolving language that the dictionaries allow us to solve.

Our simple baseline was **Max Category Count**, following Bahgat, Wilson, and Magdy (2022). Synonyms and relevant words in the definition are expected to have matching labels. A term is assigned a label equal to the most frequent label matched in its definition. To avoid bias, the candidate term is removed from the definition. For example, the word "pie" is defined in WK as "a type of pastry that consists of an outer crust and a filling. the family had steak and kidney pie for dinner and cherry pie for dessert."[6]. LIWC2015 "pie" labels as "bio". The definition included the words "kidney", "dinner" and "desert" which are all "bio" terms while "social" label was matched once for "family" and "relativ" label was matched once for "outer".

The second baseline uses **ChatGPT** which yielded impressive results on a range of tasks (Bang et al. 2023; Laskar et al. 2023). We sought to explore whether it could serve as a stand-in replacement for our classification approach. ChatGPT was prompted to label terms with LIWC categories given their definition. We experimented with a range of prompts to find the best style which was *ChatGPT 4 with few shot learning, with lexicon name*. Details are in Appendix B.

4.2 Train and Test Sets Splits

LIWC2015 and LIWC22 test set size were 1,000 samples while Values had 200 given its smaller term count. Training set counts are listed in Table 4. From a dictionary standpoint, WK had the most matches with existing lexicon terms, while OPTED had the fewest. From a lexicon standpoint, LIWC22 had significantly more coverage than the others. We opted not to use wildcard matching for LIWC2015 as we saw empirically that it led to a significant number of noisy terms being matched. Values had the fewest matches.

For UD, to build the training set, we used the top 10 definitions. We use up-votes and down-votes added by UD users as a proxy to define quality. While we believe this is a good approximation, it is worth mentioning that votes can reflect popularity like in opinionated definitions especially when dealing with politicians or public figures.

[6]https://en.wiktionary.org/wiki/pie

	UD	WK	OPTED
Values	2,198	3,158	1,535
LIWC2015	17,508	25,179	11,765
LIWC22	182,691	263,018	31,691

Table 4: Training samples count for each pair.

Method	Precision	Recall	F-Score
Max Count	0.34	0.30	0.27
ChatGPT	0.43	0.41	0.38
Lexpansion	**0.59**	**0.57**	**0.57**

Table 5: Comparing Lexpansion to baselines for LIWC2015.

5 Results and Discussion

This section discusses our results. All results were computed using "scikit-learn" (Pedregosa et al. 2011) as multi-class classification and are macro-averaged given class imbalance. Results include all outcomes without filtering using model confidence unless mentioned otherwise.

5.1 Results Compared to Baselines

We compare *Lexpansion* against *Max Count* and *ChatGPT with few-shots* to expand LIWC2015 with UD. The results of held out test set are in Table 5. Compared to a simplistic *Max Count* our method performs much better making the cost of fine-tuning worthwhile. Also, while our purpose-built fine-tuned model is much smaller, it performed better than the none-specialized ChatGPT.

5.2 Multi-Dictionary Experiments

Table 6 shows the results for lexicon-dictionary pairs on held out test set. UD models performed better for Values and LIWC22 while lagging behind WK models for LIWC2015. Although the model built using OPTED was trained on much smaller data (Table 3), the model was not far behind compared to models trained using WK, which had the most data. UD was mostly better in the cases of Values and LIWC22, slightly lagging behind WK for LIWC2015.

5.3 Tuning Results and Using Voting

Table 6 shows test set results running our models. Precision increases beyond 0.8 by increasing the confidence

Lexicon	Dict.	Precision	Recall	F-Score
Values	OPTED	0.67	0.67	0.66
	WK	0.68	0.70	0.68
	UD	**0.69**	**0.75**	**0.71**
LIWC2015	OPTED	0.54	0.54	0.53
	WK	**0.61**	**0.59**	**0.59**
	UD	0.59	0.57	0.57
LIWC22	OPTED	0.67	0.65	0.66
	WK	0.76	0.65	0.68
	UD	**0.77**	**0.77**	**0.77**

Table 6: Results for lexicon-dictionary pairs.

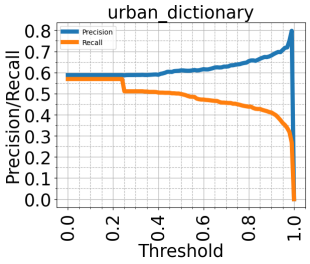


Figure 1: Precision-recall curve for LIWC2015 & UD

Lexicon	Vote	Count	Prec.	Recall	F-Score
Values	2/3	95	0.72	0.69	0.70
	3/3	81	0.89	0.86	0.87
LIWC2015	2/3	246	0.66	0.70	0.67
	3/3	482	0.89	0.95	0.91
LIWC22	2/3	519	0.75	0.84	0.78
	3/3	310	0.95	0.98	0.96

Table 7: Test set results using multiple dictionary models. "Vote" represents agreement in labels while "Count" how many samples are included while computing the metrics.

threshold to 0.98. Figure 1 shows this trade off for expanding LIWC2015 using UD. Figures for the other lexicon-dictionary pairs are in Appendix C.

Another approach, Table 7 shows counts of added terms when two out of three or all three models trained using different dictionaries agree on a label. Note that figures listed in the table were generated without applying a confidence threshold. While precision is higher in these cases, the number of candidate terms is significantly less.

5.4 Expansion Counts

We apply our models to all entries in each dictionary to generate a list of candidate terms for expanding existing lexicons. Table 8 shows the counts for each lexicon-dictionary pair that were assigned labels with confidence more than 0.98. The table also shows the intersection between labels assigned across different dictionaries when only considering the label with the highest confidence.

	Values	LIWC2015	LIWC22
OPTED	9,685	29,543	37,779
WK	163,477	266,021	216,599
UD	354,477	569,009	765,208
Words $UD \cap WK$	15,629	26,008	16,922
Labels $UD = WK$	9,943	13,824	8,622
Words in all	1,723	4,813	4,112
Labels in all	793	2,129	1,398

Table 8: Terms added with model confidence > 0.98. "Words" show common terms assigned *any label*. "Labels" show common terms assigned *same label*.

5.5 Agreement within Dictionaries

We examine variations in labels assigned to terms having two or more definitions within the same dictionary. Different definitions might still convey similar meanings for a term, which may generate the same label. One example is "semantics" shown in Table 9 where both definitions are related to linguistics. LIWC2015 labels corresponding to each definition were the same "Cognitive Process". Another example is "wheedle" from OPTED dictionary where both definitions convey similar ideas and were assigned the same LIWC22 label "Social." "launder" was also labelled with "life" from Values label with one definition being related to monetary transactions and the other with sanitation as both fall under the parent category "life" through two different subcategories; "wealth" and "health." In other cases, definitions can convey different meanings, for which models should assign different labels. For example, one definition of "abrade" illustrated a person's exhaustion, while another described how a physical surface would feel. The corresponding predicted LIWC2015 labels were "affect" and "perception" respectively. Another example is "fam" which relates to family in one definition but to an online store in another. A final example is "gale," with two definitions; one related to wind and was assigned LIWC22 "Perception" while the other was related to *laughter* and was assigned "Affect". More examples are in Appendix D. Currently, when generating the final expanded lexicon, only one label is assigned to each term. That label would be the one generated by the model with the highest confidence between the different definitions that correspond to a given term.

5.6 Agreement Across Different Dictionaries

Next, we compare labels assigned by our models for the same term in different dictionaries for the same lexicon. Definitions in different dictionaries can have varying or similar meanings. Table 12 shows definitions for "analytics". Those definitions conveyed the same meaning thus generating the same LIWC22 "Cognition" label for all dictionaries. In the case of "acai berry", OPTED did not have an entry while WK had a proper definition, while UD had an opinionated one. The UD definition was still relevant to food and health and was assigned the same label "Physical" (a "food" category in LIWC22) as WK case. For "cheesy", the models for the three dictionaries assigned the same label "Physical" as those definitions were food related. But there were also definitions in WK and UD that were conveying a sentiment towards a person. In both cases, the term was labelled "Affect" which matches the definitions. A final example is "assailant". The definitions from each dictionary provide close meanings but with different variations. The model for each dictionary assigned different labels. Note labels assigned by OPTED and WK models match LIWC22 lexicon label, while UD's does not. Table 10 shows examples of terms that were either matched or different across dictionaries.

We study terms with similar definitions and how consistent their labels are. Considering only terms with two or more definitions, UD had the highest number of definitions per term. However, when calculating the average number of

Term	Lexicon	Dict.	Label	Definition
semantics	LIWC2015	WK	cogproc	a branch of linguistics studying the meaning of words. semantics is a foundation of lexicography.
			cogproc	the individual meanings of words, as opposed to the overall meaning of a passage. the semantics of the terms used are debatable. the semantics of a single preposition is a dissertation in itself.
wheedle	LIWC22	OPTED	Social	To entice by soft words; to cajole; to flatter; to coax.
			Social	To grain, or get away, by flattery.
launder	Values	UD	life	wash scrub clean tidy intermediary To conceal the source of money as by channeling it through an intermediary. Look at us. We're such nerds we're looking up money laundering in the dictionary.
			life	laundrette laundromat clothes An error prone means of removing stains and odor from clothing, often causing them to trade colors, tear, shrink, or to disappear entirely. I accidentally laundered my whites with a red shop rag and now I have pink underwear and socks- I wonder if the missing sock is now pink or white.
abrade	LIWC2015	WK	affect	to wear down or exhaust, as a person; irritate.
			percept	to cause the surface to become more rough.
fam	Values	UD	social	derived from the word "family" . referring to people that are extremely close; as if a family member. what's crackin fam?
			life	Acronym for the Furcadia Alt Market (altmarket.net) Did you see the alts listed on FAM?
gale	LIWC22	WK	Percept	a very strong wind, more than a breeze, less than a storm; number 7 through to 9 winds on the 12-step beaufort scale. it's blowing a gale outside ...
			Affect	an outburst, especially of laughter. a gale of laughter the slightest hint of smugness would have had the nation leaning over our shoulders to blow out the birthday candles with a gale of reproach and disapproval.

Table 9: Candidate terms and their corresponding labels. Note that some definitions are truncated for space.

Lexicon	Term	Top label when using		
		UD	WK	OPTED
Values	enact	life	life	life
	crisis	life	social	children
	nonmated	-	social	-
	clingy	social	social	-
LIWC2015	hiberdating	p. concern	-	-
	leap	relativ	relativ	relativ
	venture	drives	cogproc	relativ
	gaga	affect	affect	-
LIWC22	yonder	Perception	Perception	Perception
	frickle	Drives	Physical	Affect
	mousy	-	Perception	-
	botted	Culture	Culture	-

Table 10: Examples of terms added

unique labels assigned to terms with two or more definitions, the difference between UD and the two dictionaries was small, indicating that definitions in UD often have similar meanings. These values are shown in Table 11.

6 Applying Expansion on Downstream Task

In this section we apply the resulting expanded lexicons on "Cohort Analysis" (Bahgat, Wilson, and Magdy 2020) as a downstream task and analyse the differences in outcomes when compared to running the same using the original lexicons. We aim to study the effect of utilising the expanded lexicon on a non-trivial analysis task.

Cohort Analysis studies how different cohorts affected by mental health challenges relate different concepts. For example, authors might relate "Family" concept to "Home" if

	OPTED	WK	UD
Values	3.42 vs 2.00	3.34 vs 2.10	5.79 vs 2.24
LIWC2015	3.12 vs 2.16	3.17 vs 2.19	6.79 vs 2.51
LIWC22	3.10 vs 2.16	3.36 vs 2.18	8.48 vs 2.47

Table 11: For terms with two or more definitions, average definitions/term versus average unique labels assigned.

they feel family is the most relevant to home while in other cases "Wealth" might be more relevant to "Home" if home is considered as the shelter or source of income. A cohort is represented by the authors of a subreddit. The subreddits included in that study are */r/depression*, */r/SuicideWatch*, */r/mentalhealth*, */r/BPD*, */r/ptsd*, */r/bipolar2*, */r/rapecouncling*, */r/socialanexity* and */r/StopSelfHarm*. The statistics for each subreddit are listed in Table 13. Consequently, the number of tokens from each subreddit corpus vary between tens of millions to a few millions. Word embeddings modelling the language used by the affected cohort is built using all posts from that subreddit. Those word embeddings were built using fasttext (Joulin et al. 2016) to generate skipgram word embeddings for each subreddit. The vector embeddings size was set to 100. A concept encoded by a lexicon category is represented in the embeddings space by the centroid of all terms included in that category. Relevant concepts are expected to appear close to each other. For a given concept all other concepts are ranked starting with the closest. To avoid ranking concepts that are generally close rather than specifically for a given cohort, ranks are compared to their counterparts in a neutral subreddit */r/IAMA* and reranked according to the difference between how close two concepts

	Dict.	Definition	Label
analytics	OPTED	The science of analysis.	Cognition
	WK	the principles governing any of various forms of analysis.	Cognition
	UD	Adjective:A person who is able to analyze and deduce.Noun:To analyze. Adjective:don't worry, he's a analytic.	Cognition
acai berry	OPTED	-	-
	WK	a fruit that grows on the brazilian wild palmberry tree, euterpe olearacea, used for nutritive support as an antioxidant. it is similar in size to a grape.	Physical
	UD	berry scam pyramid scheme health fat Acai is a berry native to South America that is pretty healthy, but hasn't been scientifically proven to be any healthier than many other types of berries ...	Physical
cheesy	OPTED	Having the nature, qualities, taste, form, consistency, or appearance of cheese.	Physical
	WK	of or relating to cheese. this sandwich is full of cheesy goodness.	Physical
	WK	overdramatic, excessively emotional or clichéd, trite, contrived. a cheesy song; a cheesy movie another night, when the local entertainers had gone home, gould went into the empty lounge to play piano with a cheesy string of colored lights overhead and bongo drums at his side.	Affect
	UD	cheese cheddar mozzarella gouda brie an adjective to describe cheese cheddar is cheesy	Physical
	UD	cheezy sentimental dramatic superficial emotional Sentimental and/or dramatic, yet superficial and unconvincing. The word cheesy is often defined simply as "sentimental", but there is a key distinction between the two. A situation is cheesy when the characters, their motives, and/or their emotions ...	Affect
assailant	OPTED	One who, or that which, assails, attacks, or assaults; an assailer.	Drives
	WK	a hostile critic or opponent. [...] the assailants of the quill have their honour as much at heart as the assailants of the sword.	Social*
	UD	assail ambush waylay attack rob mug A person who attacks someone on a road or a masked attacker. He was attacked by an unknown assailant.	Cognition*

Table 12: LIWC22 models assignments. Labels with "*" got confidence below 0.98 while Dashes "-" got below 0.6.

Subreddit	Submissions	Vocab.	Tokens
depression	766,971	66,661	125,883,951
IAmA	402,415	60,455	22,208,754
SuicideWatch	296,647	44,552	49,142,594
mentalhealth	106,750	31,853	16,177,175
BPD	89,471	26,886	13,686,784
socialanexity	74,802	22,041	9,674,364
ptsd	21,545	17,042	4,194,729
rapecouncling	14,907	14,488	4,401,250
bipolar2	10,519	9,436	1,424,287
StopSelfHarm	7,965	7,011	1,059,160
Average	*179,199*	*30,043*	*24,785,305*

Table 13: Counts for each subreddit. The data used matches Bahgat, Wilson, and Magdy (2020).

are in a specific subreddit and that neutral subreddit. This is called "relative rank" which is used in comparison moving forward. This task employs two lexicons, LIWC2015 and values, and uses most of the their categories. The content was curated from Reddit, a social media website that allows users to share anonymously. The users use significant slang which is expected to contain new terms not picked up by original lexicons. Thus, we picked up this task to evaluate the downstream performance of the expanded lexicons as (1) it uses more than one lexicon with a wide set of categories and (2) the content curated is expected to contain a lot of slang missing from the original lexicons.

6.1 Coverage

We first compare the coverage of the expanded versus the original lexicons for each subreddit. Coverage is hypothesized to be important as it allows the inclusion of more properties and signals from text leading to a more accurate analysis. For example, if terms relevant to a concept are not matched by the lexicon, it would be under-represented or completely missed in the analysis. The expanded versions of the lexicons include terms added from the three dictionaries mentioned earlier. Table 14 shows the percentage increase for both vocabulary and tokens. For the smaller Values lexicon, the coverage increased significantly for both vocabulary and tokens. While the original LIWC2015 contained more common vocabulary which is apparent given the lower coverage increase in tokens compared to vocabulary, there are missing slang terms which were more commonly used after the lexicon was issued or less frequently used terms given the high cost of including those such as "kabbalah" or the slang "g-d" as in the following excerpt from */r/SuicideWatch* "i've been studying ***kabbalah*** for ages, i'm 18, all i want is to meet ***g-d***.". Table 15 shows examples of newly matched terms where some can be considered slang such as "meds", "mekill", and "crack".

6.2 Impact

While coverage is higher, but the effect of newly matched terms has to be verified. As such, we investigate the changes in findings when using the expanded lexicons. We empirically to chose to filter out labels that were assigned with a confidence less than 0.8. We study how many concepts moved closer or further from others and if those are warranted given the content in the corresponding subreddits. We classify changes into *None*, *minor*, *moderate* and *significant* correspond to zero change, 1 to 4 inclusive, 5 to 9 inclusive and 10 or above respectively. *Minor* changes can be ignored as they can occur because of slight changes while *moderate*

	Values %		LIWC2015 %	
Subreddit	**Voc.**	**Tokens**	**Voc.**	**Tokens**
bipolar2	+46.52	+86.57	+36.21	+8.72
bipolar	+35.39	+86.77	+31.97	+7.96
BPD	+36.62	+87.6	+32.10	+7.18
CPTSD	+35.25	+86.29	+31.99	+6.99
depression	+23.64	+87.56	+23.36	+5.81
mentalhealth	+30.32	+86.84	+28.16	+6.87
ptsd	+43.98	+86.77	+36.78	+7.13
socialanxiety	+40.76	+87.49	+33.92	+6.07
StopSelfHarm	+55.78	+87.99	+35.91	+5.99
SuicideWatch	+27.75	+87.68	+26.38	+6.11
IAMA	+31.06	+84.82	+29.55	+10.22
Average	*+37.01*	*+86.94*	*+31.48*	*+7.19*

Table 14: Coverage increase in percentages for each cohort. Tokens are all words appearing in text, while vocabulary is the unique set of words.

Term	**Count**	**LIWC2015 Label**
meds	17,674	Biological Process (bio)
online	16,053	Personal Concerns (pconcern)
overdose	7,200	Biological Process (bio)
wasted	6,486	Affect
mekill	4,424	Personal Concerns (pconcern)
noose	3,503	Personal Concerns (pconcern)
disability	3,253	Personal Concerns (pconcern)
crack	1,064	Affect
meth	917	Biological Process (bio)
goodbyes	869	Social
consent	724	Cognitive Process (cogproc)
worthlessness	717	Affect
oblivion	717	Affect

Table 15: Examples of terms matched by the expanded version of LIWC2015 but not the original version along with their counts in "SuicideWatch" and labels.

might be less impactful to the analysis.

Table 16 shows the distribution of changes. An example of newly highlighted findings is the increased relevance of the "Children" concept relative to "Death" among SuicideWatch authors. This shift suggests a stronger connection between the two topics. Inspection of relevant posts reveals themes of troubled childhoods, exposure to or committing abuse or grief over child loss are linked to early suicidal thoughts. Also, "Affect" which indicates emotions gained relevance compared to "Cognitive Process" which lost relevance, emphasizing that the authors process the concept of "death" more emotionally rather than rationally. Other changes are obvious where categories like "life", "feeling-good", "accepting-others" and "social" for both LIWC-2015 and Values had a significant drop in their relevance to "death" for the same subreddit. Those changes show that the extra coverage indeed rendered relevant findings that better aid the analysis task.

Subreddit	**Significant**	**Moderate**	**Minor**	**None**
bipolar2	1,250	1,084	3,723	915
bipolar	1,269	864	3,746	1,093
BPD	1,294	815	3,813	1,050
CPTSD	1,214	821	3,918	1,019
depression	1,216	784	3,779	1,193
mentalhealth	1,182	818	3,840	1,132
ptsd	1,250	828	3,861	1,033
socialanxiety	1,283	863	3,780	1,046
StopSelfHarm	1,334	976	3,874	788
SuicideWatch	1,179	758	3,792	1,1161

Table 16: Changes in Relative Rank for different subreddits.

7 Conclusion and Future Work

In our work, we presented "Lexpansion" to expand lexicons and analyse with new terms from regularly-updated dictionaries. Dictionary term definitions that exist in the original lexicon are used to train models that label new terms with target lexicon categories. Using Lexpansion, we expanded three lexicons using three dictionaries which vary in properties like formality, sensitivity to new terms, term count, and number of definitions per term. Our analysis of the expanded lexicons shows the effectiveness of Lexpansion. Dictionary-based expansion was also demonstrated to be more effective than other baselines including simple voting and ChatGPT-based while applying the expanded lexicons to a downstream task rendered new findings.

For future work, we aim to identify metrics to quantify parameters like sensitivity to language shifts, ambiguity, and coverage as well as applying our method to multiple languages. We also plan to expand other types of lexicons such as WordNet using similarity of dictionary definitions. Moreover, while currently a single label is assigned to each term, we plan to study assigning multiple labels for terms that their definitions represent multiple senses.

References

Avancini, H.; Lavelli, A.; Sebastiani, F.; and Zanoli, R. 2006. Automatic expansion of domain-specific lexicons by term categorization. *ACM Transactions on Speech and Language Processing (TSLP)*, 3(1): 1–30.

Badaro, G.; Jundi, H.; Hajj, H.; and El-Hajj, W. 2018. EmoWordNet: Automatic expansion of emotion lexicon using English WordNet. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, 86–93.

Bahgat, M.; Wilson, S.; and Magdy, W. 2022. LIWC-UD: Classifying Online Slang Terms into LIWC Categories. In *14th ACM Web Science Conference 2022*, 422–432.

Bahgat, M.; Wilson, S. R.; and Magdy, W. 2020. Towards Using Word Vector Embeddings Space for Better Cohort Analysis. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Bajčetić, L.; and Declerck, T. 2022. Using Wiktionary to create specialized lexical resources and datasets. In *Proceedings of the thirteenth language resources and evaluation conference*, 3457–3460.

Bang, Y.; Cahyawijaya, S.; Lee, N.; Dai, W.; Su, D.; Wilie, B.; Lovenia, H.; Ji, Z.; Yu, T.; Chung, W.; et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Baziotis, C.; Nikolaos, A.; Chronopoulou, A.; Kolovou, A.; Paraskevopoulos, G.; Ellinas, N.; Narayanan, S.; and Potamianos, A. 2018. NTUA-SLP at SemEval-2018 Task 1: Predicting Affective Content in Tweets with Deep Attentive RNNs and Transfer Learning. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, 245–255. New Orleans, Louisiana: Association for Computational Linguistics.

Bevilacqua, M.; Pasini, T.; Raganato, A.; and Navigli, R. 2021. Recent trends in word sense disambiguation: A survey. In *International Joint Conference on Artificial Intelligence*, 4330–4338. International Joint Conference on Artificial Intelligence, Inc.

Bies, A.; Song, Z.; Maamouri, M.; Grimes, S.; Lee, H.; Wright, J.; Strassel, S.; Habash, N.; Eskander, R.; and Rambow, O. 2014. Transliteration of arabizi into arabic orthography: Developing a parallel annotated arabizi-arabic script sms/chat corpus. In *Proceedings of the EMNLP 2014 workshop on Arabic natural language processing (ANLP)*, 93–103.

Biggiogera, J.; Boateng, G.; Hilpert, P.; Vowels, M.; Bodenmann, G.; Neysari, M.; Nussbeck, F.; and Kowatsch, T. 2021. BERT meets LIWC: Exploring State-of-the-Art Language Models for Predicting Communication Behavior in Couples' Conflict Interactions. In *Companion Publication of the 2021 International Conference on Multimodal Interaction*, 385–389.

Boyd, R. L.; Ashokkumar, A.; Seraj, S.; and Pennebaker, J. W. 2022. The Development and Psychometric Properties of LIWC-22.

Cai, Q.; and Yates, A. 2013. Large-scale semantic parsing via schema matching and lexicon extension. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 423–433.

Cassidy, L.; Lynn, T.; Barry, J.; and Foster, J. 2022. TwittIrish: a universal dependencies treebank of Tweets in modern Irish. In *60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics (ACL).

Čilić, I. Š.; and Plauc, J. I. 2021. Today's usage of neologisms in social media communication. *Društvene i humanističke studije*, 6(1 (14)): 115–140.

Coppersmith, G.; Dredze, M.; and Harman, C. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, 51–60.

Cutler, A. D.; Carden, S. W.; Dorough, H. L.; and Holtzman, N. S. 2021. Inferring Grandiose Narcissism From Text: LIWC Versus Machine Learning. *Journal of Language and Social Psychology*, 40(2): 260–276.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Di Natale, A.; and Garcia, D. 2024. LEXpander: applying colexification networks to automated lexicon expansion. *Behavior Research Methods*, 56(2): 952–967.

Elbagir, S.; and Yang, J. 2019. Twitter sentiment analysis using natural language toolkit and VADER sentiment. In *Proceedings of the international multiconference of engineers and computer scientists*, volume 122, 16.

ElSherief, M.; Kulkarni, V.; Nguyen, D.; Wang, W. Y.; and Belding, E. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Fast, E.; Chen, B.; and Bernstein, M. S. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4647–4657. ACM.

Gao, L.; and Huang, R. 2017. Detecting online hate speech using context aware models. *arXiv preprint arXiv:1710.07395*.

Huang, L.; Sun, C.; Qiu, X.; and Huang, X.-J. 2019. GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3509–3514.

Joulin, A.; Grave, E.; Bojanowski, P.; Douze, M.; Jégou, H.; and Mikolov, T. 2016. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Kacewicz, E.; Pennebaker, J. W.; Davis, M.; Jeon, M.; and Graesser, A. C. 2014. Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, 33(2): 125–143.

Kaufmann, T.; and Pfister, B. 2010. Semi-automatic extension of morphological lexica. In *Proceedings of the International Multiconference on Computer Science and Information Technology*, 403–409. IEEE.

Labov, W. 2011. *Principles of linguistic change, volume 3: Cognitive and cultural factors*, volume 3. John Wiley & Sons.

Laskar, M. T. R.; Bari, M. S.; Rahman, M.; Bhuiyan, M. A. H.; Joty, S.; and Huang, J. X. 2023. A Systematic Study and Comprehensive Evaluation of ChatGPT on Benchmark Datasets. *arXiv preprint arXiv:2305.18486*.

Lei, Z.; Yang, Y.; and Yang, M. 2018. Sentiment lexicon enhanced attention-based LSTM for sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Li, W.; Zhu, L.; Shi, Y.; Guo, K.; and Cambria, E. 2020. User reviews: Sentiment analysis using lexicon integrated two-channel CNN–LSTM family models. *Applied Soft Computing*, 94: 106435.

Liu, D.; and Lei, L. 2018. The appeal to political sentiment: An analysis of Donald Trump's and Hillary Clinton's speech themes and discourse strategies in the 2016 US presidential election. *Discourse, Context & Media*, 25: 143–152.

Maity, S. K.; Ghuku, B.; Upmanyu, A.; and Mukherjee, A. 2016. Out of vocabulary words decrease, running texts prevail and hashtags coalesce: Twitter as an evolving sociolinguistic system. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, 1681–1690. IEEE.

Marcondes, F. S.; Gala, A. d. C.; Rodrigues, M.; Almeida, J. J.; and Novais, P. 2024. Lexicon Annotation with LLM: A Proof of Concept with ChatGPT. In *International Conference on Hybrid Artificial Intelligence Systems*, 190–200. Springer.

McGillivray, B.; Alahapperuma, M.; Cook, J.; Di Bonaventura, C.; Penuela, A. M.; Tyson, G.; and Wilson, S. 2022. Leveraging time-dependent lexical features for offensive language detection. In *Proceedings of the 1st Workshop of Ever Evolving NLP, EMNLP 2022*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Miller, G. 1998. *WordNet: An electronic lexical database*. MIT press.

Milroy, J.; and Milroy, L. 2017. *Varieties and Variation*, chapter 3, 45–64. John Wiley & Sons, Ltd. ISBN 9781405166256.

Mishra, S.; Khashabi, D.; Baral, C.; Choi, Y.; and Hajishirzi, H. 2022. Reframing Instructional Prompts to GPTk's Language. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Findings of the Association for Computational Linguistics: ACL 2022*, 589–612. Dublin, Ireland: Association for Computational Linguistics.

Nguyen, D.; McGillivray, B.; and Yasseri, T. 2018. Emo, love and god: making sense of Urban Dictionary, a crowd-sourced online dictionary. *Royal Society open science*, 5(5): 172320.

Oliver, A.; Castellón, I.; and Màrquez, L. 2003. Use of internet for augmenting coverage in a lexical acquisition system from raw corpora: application to russian. In *Proceedings of the Workshop IESL 2003. Internationa Workshop on Information Extraction for Slavonic and other Central and Eastern European Languages.*

Oliver, A.; and Tadic, M. 2004. Enlarging the Croatian Morphological Lexicon by Automatic Lexical Acquisition from Raw Corpora. In *LREC*.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Pennebaker, J. W.; Boyd, R. L.; Jordan, K.; and Blackburn, K. 2015. The development and psychometric properties of LIWC2015. Technical report.

Pennebaker, J. W.; Chung, C. K.; Ireland, M.; Gonzales, A.; and R., B. 2007. The development and psychometric properties of LIWC2007. 1–22.

Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001): 2001.

Sagot, B.; and Fišer, D. 2012. Automatic extension of WOLF. In *GWC2012-6th International Global Wordnet Conference*.

Sajous, F.; Navarro, E.; Gaume, B.; Prévot, L.; and Chudy, Y. 2010. Semi-automatic endogenous enrichment of collaboratively constructed lexical resources: Piggybacking onto wiktionary. In *International Conference on Natural Language Processing*, 332–344. Springer.

Seppälä, S.; Barque, L.; and Nasr, A. 2012. Extracting a semantic lexicon of french adjectives from a large lexicographic dictionary. In *Joint Conference on Lexical and Computational Semantics*.

Shaikh, S.; Cho, K.; Strzalkowski, T.; Feldman, L.; Lien, J.; Liu, T.; and Broadwell, G. A. 2016. ANEW+: Automatic expansion and validation of affective norms of words lexicons in multiple languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 1127–1132.

Staiano, J.; and Guerini, M. 2014. Depeche Mood: a Lexicon for Emotion Analysis from Crowd Annotated News. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 427–433.

Thomforde, E.; and Steedman, M. 2011. Semi-supervised CCG lexicon extension. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 1246–1256.

Tumasjan, A.; Sprenger, T.; Sandner, P.; and Welpe, I. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the international AAAI conference on web and social media*, volume 4, 178–185.

Wang, A.; Kwiatkowski, T.; and Zettlemoyer, L. 2014. Morpho-syntactic lexical generalization for CCG semantic parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1284–1295.

Watkinson, S.; and Manandhar, S. 1999. Unsupervised lexical learning with categorial grammars. In *Unsupervised Learning in Natural Language Processing*.

White, J.; Fu, Q.; Hays, S.; Sandborn, M.; Olea, C.; Gilbert, H.; Elnashar, A.; Spencer-Smith, J.; and Schmidt, D. C. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Wilson, S. R.; Magdy, W.; McGillivray, B.; Garimella, K.; and Tyson, G. 2020a. Urban dictionary embeddings for slang NLP applications. ACL.

Wilson, S. R.; Magdy, W.; McGillivray, B.; and Tyson, G. 2020b. Analyzing temporal relationships between trending terms on twitter and urban dictionary activity. In *12th ACM Conference on Web Science*, 155–163.

Wilson, S. R.; Shen, Y.; and Mihalcea, R. 2018. Building and Validating Hierarchical Lexicons with a Case Study on Personal Values. In *International Conference on Social Informatics*, 455–470. Springer.

Wu, L.; Morstatter, F.; and Liu, H. 2018. SlangSD: building, expanding and using a sentiment dictionary of slang words for short-text sentiment classification. *Language Resources and Evaluation*, 52(3): 839–852.

Yap, B. P.; Koh, A.; and Chng, E. S. 2020. Adapting BERT for Word Sense Disambiguation with Gloss Selection Objective and Example Sentences. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 41–46.

Zanwar, S.; Wiechmann, D.; Qiao, Y.; and Kerz, E. 2023. SMHD-GER: A Large-Scale Benchmark Dataset for Automatic Mental Health Detection from Social Media in German. In *Findings of the Association for Computational Linguistics: EACL 2023*, 1496–1496.

A Lexicons Category Examples

Tables 17, 18 and 19 shows example terms from selected categories. Note that in the case of LIWC2015 and LIWC22 lexicons, there are wild cards used and denoted using "*" that would allow matching any characters inserted at the corresponding position. In LIWC2015 cases, those wild cards can only appear at the end, while in LIWC22 wild cards can appear in the middle as well.

Category	Examples		
Life	liv	reside	lifetimes
Parents	daddy	mothers	grandparents
Truth	loyal	faithfully	reality
Religion	prayers	obedience	worships
Social	charity	respectful	relatives
Feeling-good	fervor	joyful	relish
Children	teenager	sons	teenage
Animals	zebras	zoos	iguanas
Learning	lectured	school	university
Order	easily	reassurance	simplicity
Accepting-others	permissive	condone	patient

Table 17: Examples from selected categories in Values.

Category	Examples		
Affect	defense	grr*	badly
Social	grandpa*	cc	flirtatious
Cognitive Process	guessing	wonder	perceiv*
Perceptual Process	cooling	yell	grips
Biological Process	spit	popcorn	wellbeing
Drives	dependent	tentativ*	lacking
Relativity	stuck	ate	emailed
Personal Concerns	ambitions	mecca	apartment*
Informal	4ev*	whoo	haha*

Table 18: Examples from selected categories in LIWC2015 lexicon. The asterisk (*) symbol denotes the presence of a wildcard for that term.

Category	Examples		
Culture	email*	european	veto
Conversation	btw	ttyl	yeah
Cognition	know	but	afaik
Drives	own	approval	regime
Physical	gym*	fish	arm
Lifestyle	game*	hiking	work
Perception	look* for	dizzled	out

Table 19: Examples from selected categories in LIWC22. The "*" symbol denotes a wildcard.

B Labelling Candidate Terms with ChatGPT

A challenge though when working with LLMs is to find the prompt that would result in the best outcome possible. Asking for the same information using different questions can result in different answers some of which are less accurate or do not provide the required information. For example, adding more content to the question text can cause ambiguity. When enquiring about the value of GDP in the US through ChatGPT through the question "How is the economy doing in the US in terms of GDP value?" would return the answer "I don't have real-time data, so I can't provide the current GDP value for the U.S. It's a dynamic figure that changes regularly. To get the latest information, you might want to check reliable economic sources or government reports for the most accurate and up-to-date GDP data." which doesn't really provide the value. But when a shorter more to the point question is used "What is the GDP of US?", the answer provided "I don't have real-time data, but as of my last update in 2022, the GDP of the United States was around $21 trillion. For the most accurate and up-to-date information, you might want to check the latest reports or official sources." gives the value requested.

To overcome similar issues with querying LLMs, there is work targeting how best to query LLMs for outcomes that correctly answers the questions in hand. One of that work described *prompt patterns* that are analogous to software patterns (White et al. 2023). The work proposed 15 different patterns that addressed a range of cases from question refinement, generating list of facts or infinite generation. Each pattern would have a name, intent and motivation and would provide a structure, examples and how the pattern's use affects and improves the output. Another approach is to *reframe* (or rephrase) a given prompt in a way that would give more desirable results (Mishra et al. 2022). Reframing a prompt would include five key items: (a) avoiding abstract concepts and replacing those low-level patterns that describe the desired output, (b) avoiding descriptive paragraphs and use itemised lists instead while avoiding negations and replacing them with affirmative statements, (c) transforming a larger task into smaller and simpler set of tasks, (d) adding details that would constraint the output, and (e) being specific about the instructions without adding extra details.

While these items mainly target conversational scenarios, we have applied some within our experiments even though our queries were in question-and-answer format[7]. The con-

[7]Using the *completion* API for ChatGPT.

tent of the prompts were stripped to the minimum number of words as long as it remains correctly constructed fluent sentences. Also, the expansion task in itself was reframed into a word labeling task with a single term being presented to the model per query.

There are still different choices to be made on how to format a prompt. For one, there was a choice of whether the low-level detail of which lexicon the target labels stated in the prompt belong to. We hypothesised that if the model has information about the lexicon it would benefit the accuracy of the answer. So, ChatGPT 3.5 was queried about the three lexicons in our experiments. The answers showed that it the model only had information about LIWC2015 but not LIWC22 nor Values. Note that ChatGPT 3.5 training data contains content up to September 2021. Only the most recent ChatGPT 4 model; *gpt-4-1106-preview* had training data with content up to April 2023 (*gpt-4* training data had only up to September 2021). It was able to identify LIWC22 correctly, but still not Values lexicon. Thus, and to optimize cost, when experimenting for baseline performance and whether to include the lexicon name or not, only LIWC2015 mention was added to the prompts and we only tested that baseline on expanding LIWC2015.

Another choice was how to provide the definition within the prompt. The meaning and example could appear separate in the prompt with each labeled with what it was. The other was to provide the whole definition as a context for the word. A final choice that was to consider to test asking the model to generate only the label within the answer. This adds emphasis on the label and simplifies parsing of the answer as the model sometimes generate more details in the answer that can be irrelevant for our task. Examples for both prompt styles separating meaning and example and adding the whole definition as context are shown in Table 20.

In addition to how the prompt is generated, we also experimented with zero versus few shot learning. To generate prompts for few-shot learning, examples were prepended to the best zero-shot prompt based on the above choices. We tested different number of example terms per selected label. Each label was paired with a single term as opposed to stating different terms in bulk that belong to a corresponding label. This makes the prompt clearer to the model and avoid any word sequence pattern that can be confusing. Examples for the few-shot cases are show in Table 20.

To query ChatGPT API, we used *text-davinci-003*[8] and *gpt-4-1106-preview*[9]. The first is the most capable model for ChatGPT 3.5 and is designed as a completion model. That is, it returns an answer for the provided query without assuming that this is a part of chat conversation. The second model is the most recent one at the time of writing. We have used *gpt-4-1106-preview* is a snapshot of *gpt-4* where *gpt-4* goes through continuous model upgrades. The snapshot should be available for an extended duration which can be used for reproducing the results. Although it is designed for conversation, but according to API documentation, the model can also be used for completion requests. Given cost considerations, we started off by testing different prompt styles on a sample of the data against ChatGPT 3.5. The styles that showed promise were then tested against the full test set. Furthermore, the highest scoring prompt style was then used to test against ChatGPT 4.

Table 21 shows results from experimenting the different choices of prompts. The zero shot learning styles used was "Definitions as Context" from Table 20, while few shot learning used was "Few Shot Learning with Definition as Context". When the lexicon name was not mentioned in the prompt, the name was omitted from the sentence preceding listing the different labels. The prompt style of *ChatGPT 4 with few shot learning, with lexicon name* was then used in further comparisons.

ChatGPT API is a paid service with *davinci* model being relatively expensive. To save cost we have run the baseline against the task of expanding LIWC2015 using UD. We assume that this task will be representative of the other two if the difference between our method and the baselines is significant. Also, when we queried about LIWC2015, ChatGPT responded with information about it. But when queried about LIWC22, could not define it as its data is only up to September 2021. ChatGPT did not have information as well about Values lexicon.

C Precision/Recall Curves

Figure 2 shows the precision and recall curves against model confidence for labels assigned by models trained using entries from OPTED, WK and UD against Values, LIWC2015 and LIWC22 lexicons. To obtain higher precision for newly added terms, new candidate terms can be filtered out based on selecting a threshold to reject labels assigned with model confidence less than that threshold. A side effect to the higher threshold is lower recall; that is, the less number of new candidate terms will be available for use. It is left for the users of Lexpansion method to decide a threshold based on the downstream tasks they aim to use the expanded lexicons for by either picking a lower threshold that would increase the number of candidate terms but add more noise or increasing the threshold to get more quality terms but with significantly less number of new terms.

D Labelling Consistency with Multiple Definitions

Definitions in UD are commonly repeated or even redundant conveying similar meanings. For example, the term "emo" had 315 definitions labelled by our model with high confidence using UD model while for WK model the same term had only 3. Out of those 315 definitions, 136 contained the word "music". Note that without filtering out lower confidence labels, the counts would be $1,382$ and 10 for UD and WK respectively out of which 684 had the word "music" in UD's case. Another example is the word "duh" with 47 definitions in UD and only 2 in WK. One final example is the term "carcolepsy" found only in UD. That term had 5 definitions or a total of 13 when including ones labelled with less

[8] https://platform.openai.com/docs/models/gpt-3-5

[9] https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo

Style: *Split Definition*

For a word with the following meaning:
"a word commonly used to describe an emotional state in which the person feels a sense of having no hope; usually during a deep depression."
in this example
"As I lay awake, alone in my bed, I cannot help but become overwhelmed by this feeling of blah."
which label should be assigned to the word blah **from the following** *"LIWC2015"* **labels:** *affect, social, cogproc, percept, bio, drives, relativ, pconcern, informal***?**
Respond with label only

Style: *Definition as Context*

In the following context:
"a word commonly used to describe an emotional state in which the person feels a sense of having no hope; usually during a deep depression. As I lay awake, alone in my bed, I cannot help but become overwhelmed by this feeling of blah."
which label should be assigned to the word blah **from the following** *"LIWC2015"* **labels:** *affect, social, cogproc, percept, bio, drives, relativ, pconcern, informal***?**
Respond with label only

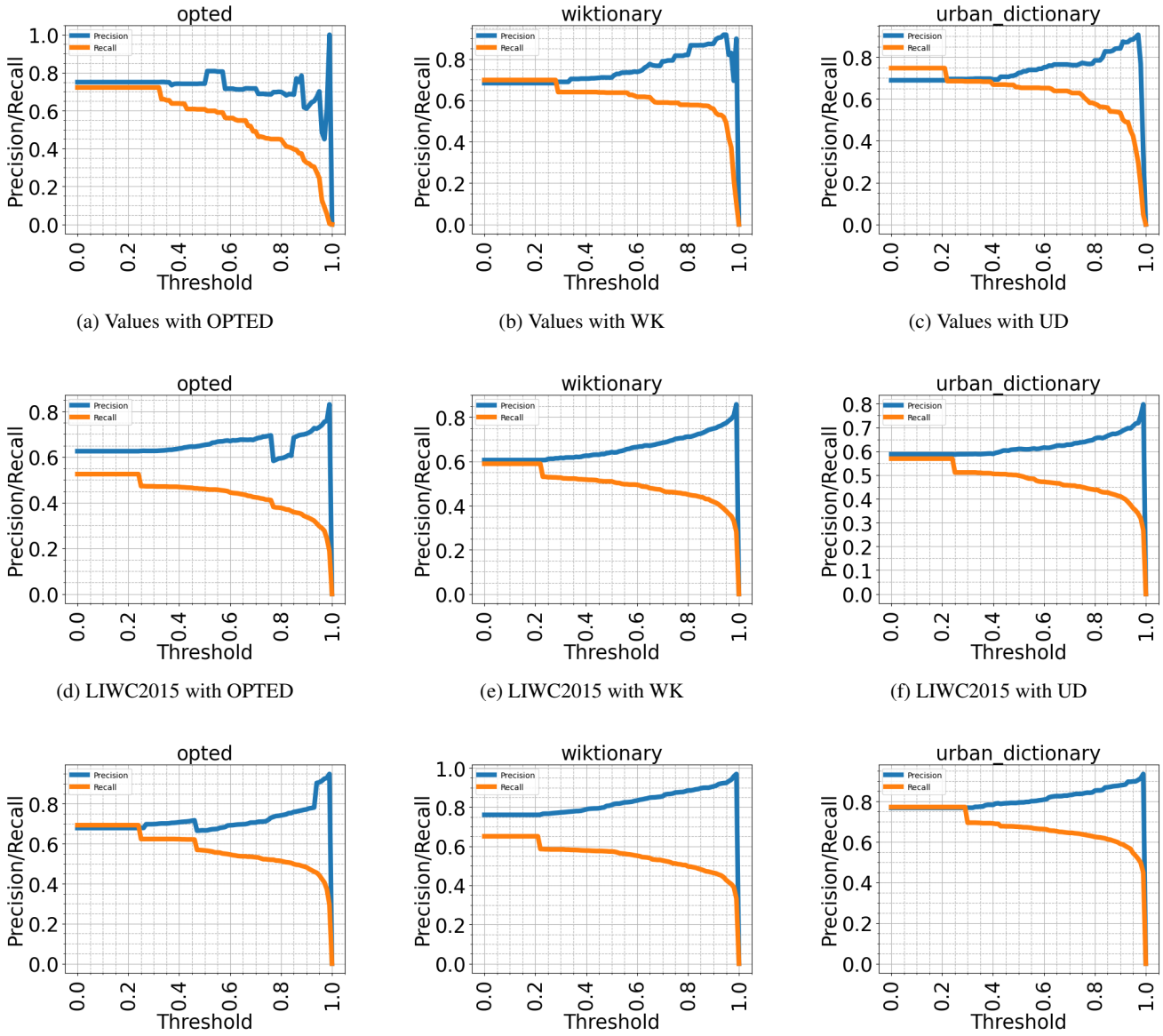
Style: *Few Shot Learning with Split Definition*

Using the following LIWC2015 labels for the following words as example:
"fake" **labeled as** *"affect"*
...
"principal" **labeled as** *"drives"*
...
"mwah" **labeled as** *"informal"*
For a word with the following meaning:
"a word commonly used to describe an emotional state in which the person feels a sense of having no hope; usually during a deep depression."
in this example
"As I lay awake, alone in my bed, I cannot help but become overwhelmed by this feeling of blah."
which label should be assigned to the word blah **from the following** *"LIWC2015"* **labels:** *affect, social, cogproc, percept, bio, drives, relativ, pconcern, informal***?**
Respond with label only

Style: *Few Shot Learning with Definition as Context*

Using the following LIWC2015 labels for the following words as example:
"fake" **labeled as** *"affect"*
...
"principal" **labeled as** *"drives"*
...
"mwah" **labeled as** *"informal"*
For a word with the following meaning:
"a word commonly used to describe an emotional state in which the person feels a sense of having no hope; usually during a deep depression."
in this example
"As I lay awake, alone in my bed, I cannot help but become overwhelmed by this feeling of blah."
which label should be assigned to the word blah **from the following** *"LIWC2015"* **labels:** *affect, social, cogproc, percept, bio, drives, relativ, pconcern, informal***?**
Respond with label only

Table 20: Examples of different prompt style we experimented with. **Bold text** indicates a template used with every prompt. *Italic text* indicates lexicon dependant part of the template. The rest is the content of the definition.



(a) Values with OPTED (b) Values with WK (c) Values with UD

(d) LIWC2015 with OPTED (e) LIWC2015 with WK (f) LIWC2015 with UD

(g) LIWC22 with OPTED (h) LIWC22 with WK (i) LIWC22 with UD

Figure 2: Precision and Recall curves against model confidence for different lexicon-dictionary pairs.

Prompt Style	Precision	Recall	F-Score
ChatGPT 3.5 with zero shot learning, no lexicon name	0.24	0.27	0.20
ChatGPT 3.5 with zero shot learning, lexicon name	0.25	0.27	0.20
ChatGPT 3.5 with few shot learning, lexicon name	0.40	0.35	0.30
ChatGPT 4 with few shot learning, lexicon name	**0.43**	**0.41**	**0.38**

Table 21: Macro-average scores on the full test set for different prompts styles. Note that for all cases we prompt the model to return the label only.

	OPTED	WK	UD
Values	7.90	2.63	1.43
LIWC2015	2.51	2.73	1.28
LIWC22	3.49	7.37	4.16

Table 22: Terms with two or more definitions, the ratio of consistent labels to differing ones.

than 0.98 confidence. All definitions were related to sleeping in a car. The above are shown in Table 23.

To quantize how consistent or redundant definitions are within the same dictionary, we compute the ratio of terms with two or more definitions assigned the same label for all definitions and the number of terms assigned more than one. The ratios are shown in Table 22.

E Limitations

The proposed work only addressed categorical lexicons which are composed of term-labels pairs. This does not cover other types and structures of lexicons such as WordNet. Nevertheless, dictionaries can be beneficial in expanding such lexicons (using similarity for example) but it is out of this work's scope. Another challenge dealing with categories with very limited number of entries. In that case, new candidate terms for these categories are also limited to none which does not allow us to analyse or compare the dictionary impact on expanding these terms. Our work also used the publicly and freely available dictionaries but not other dictionaries that would require subscription. While these were convenient for the current research purposes, the use of other dictionaries can provide more insight on evaluating Lexpansion. The presented Lexpansion and analysis only addressed labelling terms with a single category although the expanded lexicons do not have such a constraint. Our work also does not address possible biases that might exist in crowd sourced dictionaries such as in the case of UD which contains significant offensive stereotypes. We are also unable to generate new terms for lexicon categories that were never matched by the lexicon when annotating the dictionaries to generate the training and testing set. Finally, while our method can be applicable to different languages, this work was only addressed English.

Term	Lexicon	Dictionary	Label
emo	***LIWC2015***	***UD***	
	music bands depressed sometimes fun style unique 1) An emotional person. They are not depressed all the time and some are acually very happy at times. They do smile, they don't sit in a corner crying all day. Some are actually quite popular and laugh and joke around lots ...		affect
	a) short for the term emotional...in a musical sense b) music derived in the 80's... with such bands as Rites of Spring, Texas is the Reason, and more. c) can be used to describe a person who listens to emo, can relate to most of it and then cry because they can relate to it and not just because its emo. emo music is not punk. emo music usually contains lyrics which have a desperate side to them ...		affect
	music selling out corporate rock whoredom bland armchair rebels Bland, shallow corporatist pastiche of Punk that embodies non of the values or raw honest emotional integrity of Punk but continually dances on its grave non the less.. Strangely popular. I just saw Emo kings My Chemical Romance advertise Guitar Hero on XBox Live Marketplace ...		affect
	Emo is genre of hardcore punk music started in the mid eighties by a band called Rites Of Spring. Their lead singer had been a big fan of the band Minor Threat, and after their breakup, had decided to start his own band. It was similar in the musical sense, however the lyrics were different ...		pconcern
	LIWC2015	***WK***	
	a young person who is considered to be over-emotional or stereotypically emo.		affect
	depressed. criticism drapes a black velvet cape across the puddle that interrupts the path to change, to be emo about it.		affect
	associated with youth subcultures embodying emotional sensitivity. the one thing everyone agrees on is that they've never encountered a band that claimed to be emo. trevor looks kind of emo, rail thin, dark hair, guyliner, wears black all the time.		affect
duh	***LIWC22***	***UD***	
	Duh means "No sh*t sherlock" and/or "Thank you captain obvious" Him: I'm 18 Me: DUH!		Conversation
	An expression that someone says when another person says something obvious or dumb. Person 1: Hey look I can see myself in the mirror! Person 2: duh.		Conversation
	replacement for the sarcastic retort, "No Kidding!" "Hey, the Sun is bright" "Duh!"		Conversation
	LIWC22	***WK***	
	Disdainful indication that something is obvious. it's hot in the desert. - well, duh!		Conversation
	Indication of mock stupidity. duhhh, I'm jasmine, I can't even tie my shoe laces right!		Conversation
carcolepsy	***LIWC22***	***UD***	
	sleep attack sleep creep sleep hole sleep monster sleepathon. a condition affecting buddies on a trip who fall asleep as soon as the car starts moving, providing no company or driving help. Joe slept the whole way here, I think he suffers from carcolepsy.		Physical
	narcolepsy sleep roadtrip dws drive Pronounced: Kar-ko-lep-see The inability to stay awake and alert when in a car, or any other thing that moves, such as trains, planes, and busses. The act of passing out while in a car regardless of passenger or driver status. Roadtrip? Count me in, but I can't drive. I have carcolepsy and we'd all die.		Physical
	sleep carcolepsy tired road trip nubman a condition characterized by brief attacks of deep sleep brought on by simply being in a car Dude, your car is so comfortable it gives me carcolepsy		Physical

Table 23: Example terms with multiple definitions across several dictionaries, and the lexicon categories that were predicted for these definitions.

Automatically Coding Implicit Motives in Picture Story Exercises: The Automated Motive Coder

Max Brede[1], **Felix Schönbrodt**[2], **Birk Hagemeyer**[3], **Veronika Lerche**[1]

[1]Christian-Albrechts-Universität zu Kiel
[2]Ludwig-Maximilians-Universität München
[3]Friedrich-Schiller-Universität Jena

Abstract

The Picture Story Exercise (PSE) is a projective measure in personality psychology where individuals create narratives based on ambiguous images. Traditionally, the coding of these narratives has been labor-intensive. We introduce the Automated Motive Coder (AMC), which employs recent advances in natural language processing and machine learning to automate the coding of PSE narratives. Trained on an extensive dataset, the AMC demonstrates accuracy comparable to expert coders for both original and translated texts. The model offers support for multiple languages that were absent in prior methods while improving in accuracy and speed. To illustrate its effectiveness, we tested and successfully replicated the established psychological effect of gender difference in the affiliation motive. The AMC can be utilized through established machine learning tools, offering a pragmatic and reliable method for coding across several languages. This tool provides an option to reduce the workload involved in PSE coding, promoting efficiency and consistency in motive assessment.

HuggingFace —
https://huggingface.co/automatedMotiveCoder/setfit

Introduction

A pivotal assumption of motivational psychology is that individuals differ in their propensity to seek out specific classes of affectively charged goal states (McClelland et al. 1953; Schultheiss and Köllner 2021). For instance, individuals differ in their aspirations of goals pertaining to the most commonly investigated motivational needs for achievement (nAch; a concern for excellence and success), power (nPow; a concern for social impact and prestige), and affiliation-intimacy (nAff; a concern for positive social interactions and relationships). Such dispositions are termed implicit motives because individuals are not aware of their motivational inclinations. Consequently, implicit motives cannot be inferred from self-reports, but have to be assessed in an indirect way. Beginning with nAch in the 1950s, Picture Story Exercises (PSEs) have become the established method of implicit motive measurement (McClelland et al. 1953; Pang 2010).

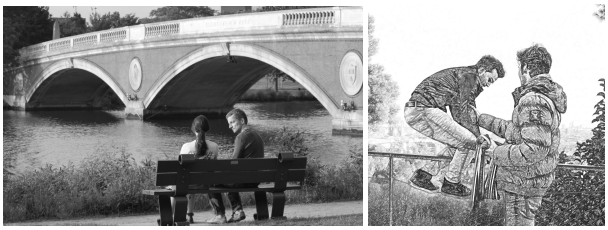


Figure 1: Examples of images used as stimuli in Picture Story Exercises (PSEs). The images are sourced from a collection of candidate PSE images published in Schönbrodt et al. (2021). The left image is titled "newpic14" by Oliver Schultheiss (CC-BY 4.0); the right image is titled "burglars" (CC0).

PSEs are research variants of the Thematic Apperception Test (TAT; Morgan and Muray 1935). Participants are instructed to invent imaginative stories in response to picture cues showing ambiguous, often social, scenes. Typically, a sequence of four to eight pictures that pull for imagery related to the targeted motive domains are presented for 10 – 15 seconds, and participants are given 4 or 5 minutes time per picture to write down a story that describes the depicted scene (Pang 2010). See Figure 1 for example images. The stories are then coded for the appearance of motive-related contents as defined in empirically validated coding systems that were developed for the assessment of the respective motives (for an overview of classical motive coding systems, see Smith 1992).

The process of PSE coding is effortful and requires substantial amounts of time, labor, and expertise. Coders need to undergo training in order to apply the coding rules accurately, which according to Pang (2010) should include at least 12 hours of practice per coder. For a trained and experienced coder, the coding of a single PSE-story takes about 2 – 5 min, which amounts to a coding time of 12 – 30 min. per participant for a typical six-picture PSE. In addition, it is recommended that PSE-stories are coded by at least two independent coders (Pang 2010), which doubles the effort. A second potential problem pertains to the objectivity of PSE coding by human coders. Coders are required to achieve high agreement with expert codings during coder training

Proceedings of the First Workshop on Integrating NLP and Psychology to Study Social Interactions (NLPSI) @ICWSM '25, pages 28–38

(reported requirements are usually a category agreement of 85%, see Winter 1994) and, in case of multiple independent coders, with each other. Nonetheless, the objectivity of codings might still be impaired by coder drift over time (Pang 2010) or differences in coding practices between different labs.

Previous Work

Because of the limited practical efficiency and potential problems of objectivity that come with human coders, motive researchers have been striving to develop automatized, computer-based alternatives for decades (e.g., Green et al. 1967; Smith 1968). The approaches for automated motive coding reach from classic natural language processing methods like bag of word approaches (e.g., Schultheiss 2013) to embedding model based RNN-approaches (e.g., Pang and Ring 2020). Most recently, Nilsson et al. (2024) published transformer-based automated motive coding approaches based on RoBERTa large (Liu et al. 2019) and GBERT (Chan, Schweter, and Möller 2020). Crucially, all of these earlier efforts were fundamentally limited to unilingual processing. Although the two models presented by Nilsson et al. (2024) showed performance comparable to that of expert coders, their reliance on unilingual base models represents a significant constraint.
Tunstall et al. (2022) published a contrastive learning paradigm (described below) that leverages transformer-based embedding models implementing recent advancements. This approach additionally allows for basing classifier architectures on Sentence Transformer (Reimers and Gurevych 2019) models, that are trained to be capable of embedding texts in a variety of languages. We aim to improve on the models trained by Nilsson et al. (2024) by utilizing the training paradigm described by Tunstall et al. (2022) to (1) improve inference speed and accuracy and (2) allow our model to classify texts written in more than one language.

Methods

Training and holdout material

The main dataset is a combined data base of German PSE stories from 27 studies, which have been coded sentence-wise with the Manual for Scoring Motive Imagery in Running Text (Winter 1994) by several trained coders (see Schönbrodt et al. 2021 for details of the data base). Each sentence is labelled with one of the four motive classes nAch, nAff, nPow or "no motive" (null). See Table 1 for examples of coded sentences. In rare cases (see below), multiple classes were assigned. For the computation of a motive score on person level, first all motive labels are summed across all stories that a person wrote. As such a score is typically correlated with the length of stories (longer stories with more sentences have more potential to accumulate motive labels), these sum scores are subsequently corrected for word count (see below).

According to Winter (1994), one category per sentence can be coded for each of the three motives. Thus, motive coding basically takes place at the level of single sentences. However, there are deviations from this principle. Most notably for this context is the so-called second-sentence rule. This rule states that the same motive class cannot be assigned to two consecutive sentences. Since this rule leads to an inflation in null-codings, this rule has been ignored by a large portion of the stories.

The studies used different sets of pictures, including the standard set of six pictures suggested by Schultheiss and Pang (2007), some pictures from the classical Thematic Apperception Test (Morgan and Muray 1935), and some new pictures. Upon publication, the full PSE data base was split into a public set and a holdout set, with the latter not being publicly released. We used the public set for training, and the non-public holdout set as an additional test set.

For both training and holdout set, we removed all sentences with 5 or fewer characters (503 total sentences). We then removed stories with fewer than 30 words (Smith, Feld, and Franz 1992). As the unit of analysis in Winter's coding system is the sentence, we furthermore removed stories which had fewer than two sentences. This step removed 1,018 out of 26,376 stories. Furthermore, we removed sentences that were potentially affected by the second-sentence coding rule (5.8%).

For the training set, we also removed sentences that had more than one motive category (0.8%) as the classification algorithm is trained on single motive codings only. In the holdout set, we did not remove the mixed categories, because we wanted to keep the real stories for testing. These selections resulted in a training set of 131,110 sentences in 20,463 stories, written by 3,637 persons, and a holdout set of 33,546 sentences in 4,892 stories, written by 900 persons. Final class frequencies were: nAff (training: 14.63%, holdout: 16.89%), nAch (training: 9.65%, holdout: 9.28%), nPow (training: 13.92%, holdout: 11.69%), null (training: 61.80%, holdout: 58.16%), and mixed categories in the holdout set: 3.98%.

The holdout set was additionally translated to English using DeepL to allow a heuristic test of the multilingual performance of the final model. This was also done to allow for a fair comparison with Nilsson et al. (2024), who followed this approach on the same base-dataset to train a classifier for English texts. Neither the holdout set nor its translation were used in any way during the training of the model.

Training Procedure

The training process was split into two sequential stages, closely adhering to the methodology described by Tunstall et al. (2022). This approach initiates with the contrastive fine-tuning of a Sentence Transformer model (stage one), subsequently transitioning to the training of a classification model leveraging the refined embeddings (stage two).

Stage one involved a contrastive fine-tuning procedure applied to selected Sentence Transformer models. The candidate models were on the one hand a selection of well-tested models published by the authors of the Sentence Transformers library (Reimers and Gurevych 2020) and on the other hand models that performed well on the MTEB benchmark suite for multilingual classification (Muennighoff et al.

Text	Coded Motive
Frederike carries out her experiment routinely and secures her 1st place again.	nAch
The two are sitting in a restaurant and look happy .	nAff
Michael knew that his worst enemy, the well-known mafia boss Luigi, had boarded this ship, his last chance to wipe him out once and for all.	nPow
Next he will land roughly on the ground.	null

Table 1: Translated example sentences from the dataset published by Schönbrodt et al. (2021). The *Coded Motive* column indicates the motive assigned by the expert coders.

2023). For the whole list of candidate models, see Appendix 1.

This fine-tuning-procedure aimed at enhancing the model's understanding and encoding of complex sentence structures specific to motivational imagery within the PSE narratives. The models were trained to generate closely packed embeddings for sentences with similar motive codes while distancing embeddings for sentences of differing motives. This was done by using pairs of sentences from the training set to maximize the cosine similarity between the embeddings of sentences with the same motive class and minimize it for sentences with different motive classes. This paradigm is meant to greatly reduce the amount of necessary labeled data, since it effectively uses each sentence $n - 1$ times instead of once, where n is the number of sentences in the set.

In this first stage, we trained on randomly sampled subsets of varying sizes from the training dataset. A hyperparameter defined which stratifying variables to use in sampling, ranging in value from one to six. A value of one indicated that only the 'motive class' was used as a stratifying variable, while a value of two included both 'motive class' and 'picture ID'. Higher values incorporated additional variables in this order: participant gender (3), participant age group (4), the study ID (5), and the coding lab (6). Thus, the motive class was always part of the stratification, and the other five variables could be optionally added. Stratifying for the target (in our case, the motive class) has been shown to improve cross-validation results (Kohavi 1995). The other possible stratifying variables were chosen and ordered based on their plausible potential to influence the sentence structure and content related to motivational imagery within the PSE narratives. The size of the subsets sampled from the training set ranged from 1 to 5 sentences for each combination of stratifying variables. This number of sentences was also determined by a hyperparameter. If there were fewer sentences available for a given combination than required, the sentences of this combination were oversampled to fulfill the target size. If a combination was not present in the training set, it was ignored.

The fine-tuning phase utilized the Tree-structured Parzen Estimator (TPE) algorithm as implemented in Optuna (Watanabe 2023) for hyperparameter optimization. The TPE algorithm was employed to search through a predefined space of possible hyperparameters, with the objective of finding the combination that yielded the best performance on a validation set[1]. The validation set was randomly sampled from the prepared training set according to the same stratification procedure as described above. It was made sure that the validation and training sets did not overlap. In cases where all available sentences for a given combination of stratifying variables were included in the training set, the combination was omitted from the validation set. For the complete set of hyperparameters tuned, see Appendix 1. The training was implemented using the HuggingFace Sentence Transformers library (Reimers and Gurevych 2019) and set to save the best model according to the performance on the validation set. The F1-score of the cosine with an automatically found threshold was used as the metric for evaluation and as the hyperparameter optimization target. The best model trained using this procedure was then employed to embed the complete training set as well as the holdout set and its English translation.

Stage two used the embedded training set to train multilabel-classification models. These models predicted for each sentence, whether it contained imagery relevant to nAch, nAff, or nPow, or if it contained no motive-relevant imagery at all. All trained classification heads consisted of a OneVsRestClassifier as implemented in scikit-learn (Pedregosa et al. 2011) with one of a series of possible classifier models as its estimator. The classifier models, along with their associated settings, were determined using again the TPE algorithm, as implemented in the Optuna library. Notably, we employed a hyperparameter to determine the motive class weighting strategy during training. This strategy was selected to allow the model to adjust its emphasis on different motive classes and compensate for the significant motive class imbalance within the dataset. By adjusting the weight each motive class was given during training, the model could better handle imbalance and improve prediction accuracy across all motive classes. For the complete list of hyperparameters and their ranges as well as further details on the configuration used during training, please refer to Appendix 2.

Ten-fold cross-validation (CV) was employed to train and evaluate the models. The separation into folds was conducted at the person level, ensuring that no sentences written by the same individual appeared in both the training and

[1]In this context, *validation set* refers to the dataset used to calculate hyperparameter optimization targets. Thus, the term does not refer to the psychological quality criterion of validity.

evaluation set of a fold. The folds were stratified according to the uncorrected motive score of a person, their age group, gender and the lab that coded the sentences.

Performance evaluation

Two evaluation performance metrics - one at the sentence level and the other at the person level - were each averaged over all 10 folds, and both were used as optimization targets for hyperparameter tuning: As a metric on sentence level, the average F1-score was chosen. This score was calculated as the weighted average of the four F1 scores, which were determined based on the true and false predictions for each of the four classes (nAch, nAff, nPow, Null) compared to the others on a per-class basis. For the evaluation at the person level, we first calculated the so-called corrected motive scores, i.e., the aggregated value of a person's implicit motive strength based on all stories they wrote. Next, for each motive class separately, we computed the correlation between the predicted and true corrected motive scores (nAch, nAff, nPow). Finally, we averaged the three motive correlation coefficients. The Null-class was left out in this step, since it proved to be inflated for nearly all trained models. We employed the correction-procedure described in Schönbrodt et al. (2021) to calculate the corrected motive scores and applied this procedure both for the motive class probabilities predicted by our model and the motive classes coded by the expert coders (i.e., the true motive classes). This procedure aims to correct for the positive correlation between motive sum scores and story length (Pang 2010) by conducting the following steps:

1. Summation of predicted motive probabilities (model-prediction) and summation of expert-coded motive codings separately for each motive.
2. Robust regression for each implicit motive of the sum scores from step 1 divided by 1,000 on word count per subject (In our case, the lmrob-implementation of the extended method suggested in Koller and Stahel (2011) from the robustbase-R-package (Maechler et al. 2024) was used.)
3. Extraction of residuals as corrected motive scores per implicit motive

We used the Optuna implementation to determine all Pareto-optimal runs. Of those, the final classifier settings were selected based on the mean of the F1-score and the mean correlation, both scaled so that the maximum value reached in the Pareto-optimal runs was equal to one. The settings with the highest mean of these two scaled metrics were chosen as the best performing model and a final classifier was trained using these settings on the full training set.

While optimizing the model for agreement with expert coders - more specifically in terms of the correlation between predicted and coded motive scores - it is important to establish what constitutes a satisfactory level of agreement. Guidelines suggest that a minimum of 85% category agreement with precoded examples during training is required, as outlined by Winter (1994). To contextualize this benchmark using Pearson correlation metrics, Schultheiss (2013) provides relevant data, reporting inter-coder correlations of .79, .74, and .86 for nPow, nAch, and nAff, respectively. These correlations were achieved after the coders were trained in a joint lab with common practices and (implicit and explicit) coding norms. Similarly, Schultheiss, Liening, and Schad (2008) reported inter-coder correlations of .86, .70, and .81 for the same motive domains and the same coder-training procedure. Taken together, these findings suggest that among human coders, Pearson correlations of .70 or higher are typically expected, with values rarely exceeding .85.

Since these correlations are achieved under quite optimal conditions, they may not be indicative of agreement after coder drift has occurred. Phenomena like coder drift lead to declining agreement, as does the presence of coders from multiple labs (who never coded jointly or were calibrated in any way) in the data set. Realistically, expectations for inter-coder correlations are likely lower, especially if the coders have not been trained in the same laboratory. In sum, a model that reaches the values typically reported for human coders demonstrates good performance.

Pang and Schultheiss (2005) and Schultheiss and Pang (2007) additionally suggest one-way random effects intra-class correlations (ICCs) for the evaluation of model performance. However, we argue that this measure is not entirely valid for the motive scores generated by our model. Our motive scores are calculated based on the class-probabilities of the motives while the codings generated by human coders are binary ratings per sentence (i.e., motive class present or not). Since the type I ICC is highly sensitive to mean level differences between coders, this measure might skew the results. Transforming the probabilistic output into a binary rating by setting a threshold at .5 would be an option to transform the probabilities to be usable in the ICC-calculation. But Nilsson et al. (2024) and previous unpublished trials at training an automated coding solution we ran showed that using the summed probabilities resulted in more reliable estimates for the motive scores than the transformed binary ratings.

As an additional measure of the validity of the models' predictions, we computed Standardized Root Mean Square Residuals ($SRMR$). This measure indicates the deviation of two correlation matrices, in our case the intercorrelations of the scores for the three implicit motives. Ideally, automated coding reconstructs the motive intercorrelations from manual coding, where lower $SRMR$ values are better.

Lastly, to gauge the criterion validity of the automated codings, we examined whether there are differences in affiliation motive strength between female and male participants. Meta-analytic findings show that after word count correction, women have higher motive scores in the implicit affiliation motive than men. Drescher and Schultheiss (2016) report an average Cohen's $d = 0.45; 95\% - CI = [0.37, 0.53]$ for this effect. Accordingly, we computed the standardized mean difference (Cohen's d) between women's and men's affiliation scores obtained using the AMC as the gender effect performance measure. Effects should be comparable to those found in the human codings and to the literature.

To benchmark our best model against the approach published by Nilsson et al. (2024), we employed both their mod-

els for English and German texts on our holdout set and its English translation. Performance was evaluated using the just discussed measures, providing a direct comparison of our model's efficacy against the classifiers trained by Nilsson et al. (2024), both their approach based on the GBert-embeddings and the RoBERTa-embedding based approach.

Results

Remember that the primary training target for all models was the sentence-level prediction of motive class. Notably, the best embedding model (i.e., after stage 1 of the Tunstall et al. 2022 approach) reached a cosine-based F1-score of .61 on the randomly sampled validation set. Thus, even without a classification model, the performance was already quite impressive. This embedding model was based on the multilingual-e5-large-variant of the architecture presented in Wang et al. (2024) with the contrastive loss function implemented in the Sentence Transformer library (Reimers and Gurevych 2019). All settings can be seen in Appendix 1.

The final combination of embedding-model (stage 1) and classifier (stage 2) performing best in the hyperparameter optimizations achieved an average F1-score of .77 on the train folds and an average F1-score of .77 on the validation folds. This result was reached with an Elastic Net classifier with stochastic gradient descent learning as implemented in scikit-learn. For alle other settings, see Appendix 2. Importantly, the model also performed well on the holdout set (F1: .76) with only slight decreases in performance for the translated holdout set (F1: .73).

On the person level, our best model reached correlations between model- and human-coded motive scores of on average $r = .71$ for nAch, $r = .83$ for nAff and $r = .76$ for nPow over all ten validation folds. The findings for the holdout set and its translation are in a comparable range (see Table 2).

As reported in the Methods section, we used, next to the aforementioned correlations, two additional metrics as indicators for model performance on the person level.

First, we calculated the $SRMR$ for all datasets to test whether our automated coding approach resulted in a deviation in the intercorrelations between the model-coded and the human-coded implicit motive scores. As example we show the intercorrelations for the holdout set in Table 3. The table shows only minor differences in the intercorrelations for all motive scores.

These matrices were used to calculate the $SRMR$ as the square root of the mean of the squared differences of the matrices. The $SRMR$s were the highest on the translated to English holdout set ($SRMR = .0626$) with $SRMR$s for Train, Validation and German holdout set of .0159, .0490, and .0430, respectively.

In addition, we calculated Cohen's d as a measure of the gender difference in the motive score for affiliation. The results for the gender differences found in the training and holdout sets are shown in Table 4. Women showed a higher implicit affiliation motive than men in all datasets, ranging from $d = 0.39$ to 0.57 with the largest effects in the holdout set ($d = 0.53$) and its translation ($d = 0.57$). Interestingly, all effect sizes, except for the average of the validation folds, are larger for the automatically coded motive scores than for the ones coded by humans.

Comparison to previous approaches

We compared the performance of our model on the holdout set to that of the GBERT and RoBERTa models published by Nilsson et al. (2024). Since the GBERT model is exclusively trained for German texts and the RoBERTa model for English texts, we applied the GBERT model only to the original German holdout set and the RoBERTa model to the translated-to-English holdout set. All model predictions for the models published by Nilsson et al. (2024) were calculated using the R-script presented in their paper. To simplify comparison, we only calculated motive scores for sentences present in both the English and the German set coded by both models published by Nilsson et al. (2024), which is a subset of the data presented above.[2]

The results on both holdout sets are presented in Table 5. Our model surpasses the performance reached by the GBERT model on all motive scores for the German holdout set. This effect is less clear for the holdout set that was translated to English. Here, the RoBERTa model reached slightly lower correlations than our AMC for all motive scores except for nPow, where no substantial difference can be seen.

Our model additionally generates closer intercorrelations as measured in the $SRMR$, both on the German holdout set ($SRMR = .04$ for our model vs $SRMR = .11$ for the GBERT approach) and the translated-to-English holdout set ($SRMR = .06$ for our model vs $SRMR = .08$ for the RoBERTa approach).

The gender effect measured as Cohen's d of the difference in the affiliation motive score between women and men is slightly higher for the models published by Nilsson et al. (2024), both on the German holdout set ($d = 0.53$, 95% CI [0.39, 0.67] for our model vs $d = 0.65$, 95% CI [0.50, 0.79] for the GBERT approach) and the translated-to-English holdout set ($d = 0.57$, 95% CI [0.43, 0.71] for our model vs $d = 0.64$, 95% CI [0.49, 0.78] for the RoBERTa approach).

To test for differences in inference time, we repeatedly ran our model and both models published by Nilsson et al. (2024) on a virtual headless ubuntu 22.04 LTS private cloud computing instance with 126 GB of RAM and 64 assigned virtual cores of an Intel Xeon Processor (Icelake). The instance was additional fitted with two Nvidia A40 graphics cards. All classifications were run for 25 times for the first 1,000 sentences in the holdout/translated holdout set, both on GPU and CPU.

Our model took an average of 1.7 seconds ($SD = 0.96, min = 1.5, max = 6.4$) for the German and an average of 1.41 seconds ($SD = 0.03, min = 1.3, max = 1.5$) of inference time for the English holdout set on GPU.

The models by Nilsson et al. (2024) were far slower, with runtimes of an average of 304 seconds ($SD = 2.97, min =$

[2]The reason for this partial selection is a combination of issues with the DeepL-translation and some data cleaning implicit in the text package, Nilsson et al. (2024) used to ship their models.

	N	r			
		nAch	nAff	nPow	Null
Train	3298.3*	.71 [.69, .73]†	.83 [.82, .84]†	.77 [.75, .78]†	.93 [.93, .94]†
Validation	363.2*	.70 [.65, .75]†	.81 [.78, .85]†	.78 [.74, .82]†	.92 [.90, .94]†
Holdout Set	900.0	.74 [.71, .77]	.79 [.77, .82]	.74 [.70, .76]	.93 [.92, .94]
Translated Holdout Set	900.0	.74 [.71, .77]	.79 [.76, .81]	.73 [.70, .76]	.93 [.93, .94]

Note:

95%-Confidence Intervals in brackets.

* Averaged over all folds. Note that due to the stratification on person level, not all folds consisted of the exact same amount of samples. This was due to some combinations of stratifying variables being present in few persons, leading to their absence in some of the training and/or validation folds.

† The correlations and CIs reported for the validation and train folds are those of the runs with the largest CI out of all ten folds.

Table 2: Correlations between human- and model-coded motive scores on person level. Both human- as well as model-coded scores were transformed to consider the number of written words in the PSEs as described in the Methods section

	Human codings		Model predictions	
Motive	nAff	nPow	nAff	nPow
nAch	.03	-.08	.08	-.01
nAff		-.34		-.31

Table 3: Intercorrelations of the 3 implicit motive scores, based on human codings and model-predicted probabilities for the holdout set.

	d	
	Human codings	Model predictions
Train*	0.35	0.41
Validation*	0.40	0.39
Holdout Set	0.44 [0.30, 0.58]†	0.53 [0.39, 0.67]
Translated Holdout Set	0.44 [0.30, 0.58]†	0.57 [0.43, 0.71]

Note:

95% CI in brackets.

* Averaged results over the 10 CV-Folds.

† Since the Translated holdout set has the same motive codings as the non-translated one, the effect-sizes of the human codings are identical.

Table 4: Gender effect in the affiliation motive over all datasets.

$298, max = 309$) for the German set using the GBERT model and an average of 326 seconds ($SD = 10.29, min = 318, max = 372$) for the RoBERTa model on the English set on GPU.

This difference in runtime was even more pronounced when running on CPU with averages of 13 seconds ($SD = 0.72, min = 12, max = 15$) and 12 seconds ($SD = 0.70, min = 11, max = 14$) for our model and 2,172 seconds ($SD = 33.87, min = 2,113, max = 2,234$) and 4,045 seconds ($SD = 40.41, min = 3,985, max = 4,146$) for the models published by Nilsson et al. (2024).

Discussion

We successfully trained an Automated Motive Coder (AMC) that achieves near-human-level agreement with existing codings. The AMC comprises a fine-tuned multilingual Sentence Transformer embedding model coupled with a multi-label classification header. Schultheiss, Liening, and Schad (2008) and Schultheiss (2013) reported human inter-coder correlations between .70 and .86. The AMC reached values that were very close to this range. It is important to note that these correlations represent a near-optimal condition where coders were specifically trained in a joint lab with common practices and (implicit and explicit) coding norms. This level of agreement should not be considered the standard, particularly for our dataset, which involved sentences coded by multiple researchers from different labs with potentially varying strategies. Despite this variability, our AMC achieved a correlation with the human-coded motive scores of above $r = .70$ on all datasets, including the holdout set that was not used for the training of the model. These results are very promising and likely close to the prac-

		German holdout set		Translated holdout set	
Motive Score	N	AMC	GBERT	AMC	RoBERTa
nAch	858	.73 [.70, .76]	.63 [.59, .67]	.74 [.70, .76]	.71 [.68, .74]
nAff	858	.80 [.77, .82]	.67 [.63, .70]	.79 [.76, .81]	.77 [.74, .80]
nPow	858	.74 [.71, .77]	.64 [.60, .68]	.74 [.70, .77]	.74 [.71, .77]

Note:
95%-Confidence Intervals in brackets.

Table 5: Comparison of the correlation between automatically coded and human-coded motive scores achieved with our model and Nilsson et al.'s 2024 GBERT and RoBERTa models, applied to the holdout set and its translation. The GBERT approach is only used on the German and the RoBERTa approach only on the translated-to-English holdout set.

tical limits of what is achievable.

Our model demonstrated robust performance in predicting human motive codings for texts the model did not see during training, that were either German or translated-to-English. By employing the contrastive learning paradigm outlined by Tunstall et al. (2022), we slighlty improved upon the accuracy of results reported by Nilsson et al. (2024) while adding support for the coding of samples from multiple languages. Importantly, our model is implemented within the SetFit library and is available on the HuggingFace model repository, ensuring it is openly accessible for application and further research.

In addition to the agreement with human coders, we were able to reach acceptable levels for the model fit measured as SRMRs. Furthermore, we could replicate the gender difference in the affiliation motive reported in the literature, where women consistently showed higher scores in nAff. This effect was reported by Drescher and Schultheiss (2016) with an average Cohen's d of $0.45(95\%\text{ - }CI = [0.37, 0.53])$. This effect was slightly larger in our automatically coded motives compared to manually coded ones and the literature, although these differences remain marginal. Out of all our tests, these effects were highest for the models published by Nilsson et al. (2024). Note that although we generally considered the replication of the well-established gender effect as a sign of validity, one cannot conclude that "the higher, the better". On the one hand, it could be that automated coding better picks up a true signal; on the other hand, it could also be that, regardless of the underlying motive, men and women have differences in linguistic styles which are more weighted in automated coding. This apparent tendency of automated coding models to amplify the gender effect should be the subject of further investigation.

Application of automated coding in practice

Our model outperforms the already impressive results reported by Nilsson et al. (2024) for nearly all performance measures tested. This improvement was reached by basing the model on modern embedding models fine-tuned to pronounce differences between motive-relevant imagery in the embeddings instead of using general-purpose language models for the embeddings. Notably, we extended their solution by training a model with a multilingual base, demonstrating acceptable performance across two languages. Incidentally, our architecture also leads to a significant improvement in inference speed, being up to 300 times faster than previous methods. While this acceleration is less critical for traditional, batch-oriented motive coding, it does offer potential for future applications. The increased speed could potentially allow for the integration of the model into experimental paradigms, enabling the generation of stimuli that could be dynamically adjusted based on calculated motive scores.

Nonetheless, the research community has to gain experiences with the new measurement method and yet unknown problems and boundary conditions of its applicability might show up. Researchers could reanalyze their existing studies with the AMC for a retrospective check of its validity, or do both manual and automated coding for new studies. When multiple coders are employed in one study, they often code a subset of the data in parallel to check their agreement. When sufficient agreement has been demonstrated, they continue coding separate batches of the stories. A similar routine could be implemented with the model presented in this study, to test whether cases arise in which the AMC-codes deviate from human results.

The python-code-snippet in Listing 1 demonstrates how the SetFit library and our model can be used for coding text sections in any language supported by the base model trained by Wang et al. (2024)[3].

Boundaries and limitations

Our model was trained on classical PSE stories, written in response to a broad, yet limited set of picture stimuli. Although generalization to new pictures is expected to be adequate, caution is advised when applying it to other text genres, such as transcribed interviews or speeches.

Our model inherits its validity and generalizability from the validity of the human-derived coding system and the validity of the concrete human codings. The Winter coding system is based on experimentally derived coding systems, where groups of participants were exposed to motive-arousing situations. These studies have been conducted in certain (Western) cultural contexts with certain picture stimuli. Whenever such coding systems are transferred to other

[3]The HuggingFace repository lists a total of 93 languages as being supported. A longer tutorial with installation-instructrions can be found under the link above.

Listing 1: Example on how to use the AMC.

```
from setfit import SetFitModel

model = SetFitModel.from_pretrained(
    "automatedMotiveCoder/setfit"
    )

model.predict_proba(
    [
        "Du schaffst das schon.",
        "Tu vas y arriver.",
        "Zvládnete.",
        "You'll manage.",
        "Te las arreglarás.",
        "Saate hakkama."
    ]
)
```

contexts or languages, their validity should be carefully evaluated. Though our model is multilingual in nature, its training and evaluation were restricted to German or translated-from-German stories. Therefore, a certain homogeneity in the motives and language used has to be assumed. It remains to be tested whether our model can be applied to stories originally written in other languages than German - especially since Pang and Schultheiss (2005) could show a significant difference in motive scores between ethnicities and in comparison between German and U.S.-American students.

Conclusion

This study developed an Automated Motive Coder (AMC) that achieves human-level accuracy in implicit motive coding using contrastive learning. The AMC demonstrates effective performance on both German and English texts and is accessible via the SetFit library and HuggingFace repository, with theoretical support for a variety of additional languages. While promising, further research is required to assess the validity of the AMC across diverse text genres and cultural contexts. Future studies should aim to examine its applicability and reliability across various settings.

Appendix 1 - Sentence Transformer Hyperparameter

For the hyperparameter tuning of the embedding model (stage 1 of our training), we used the following selection of candidate base models (all names are also the names of the corresponding HuggingFace repositories):

- intfloat/multilingual-e5-large-instruct (Wang et al. 2024)
- intfloat/multilingual-e5-large (Wang et al. 2024)
- intfloat/multilingual-e5-small (Wang et al. 2024)
- shibing624/text2vec-base-multilingual (Xu 2023)
- ISOISS/jina-embeddings-v3-tei (Sturua et al. 2024)
- sentence-transformers/distiluse-base-multilingual-cased-v2 (Reimers and Gurevych 2020)
- sentence-transformers/distiluse-base-multilingual-cased-v1 (Reimers and Gurevych 2020)

The training itself was implemented with the HuggingFace-Sentence Transformers-Trainer-Class (Reimers and Gurevych 2019) using a setting from the following space of hyperparameters:

- Number of train epochs as 10^{epochs} where $epochs$ is an integer between 2 and 8
- A per device train batch size of $2^{trainbatch}$ where $trainbatch$ is an integer between 3 and 6
- A per device eval batch size of $2^{evalbatch}$ where $evalbatch$ is an integer between 3 and 6
- The trainer´s learning rate as a uniformly distributed float between $1e-6$ and $1e-3$
- The ratio of total training steps used by the trainer for scaling up the learning rate from 0 to its final value as a uniformly distributed float between .05 and .4
- The trainer‘s learning rate scheduler, selected as either “reduce_lr_on_plateau” or “cosine”
- one of the following two loss functions implemented in the Sentence Transformers library:
  - Contrastive loss as described in Hadsell, Chopra, and LeCun (2006)
  - Cosine Sentence loss as described in Jianlin (2022)

In addition to these model settings, we implemented the selection of the samples used for training as hyperparameters. For a detailed description of this procedure, see the Methods section. This sampling strategy was represented as a set of two hyperparameters:

- The amount of stratifiers used from this list in this order: motive class, picture ID, participant gender, participant age group, study ID, and the coding lab
- The amount of sentences sampled per combination of stratifiers as an integer between 1 and 5

The best performing combination of hyperparameters was as follows:

- model: intfloat/multilingual-e5-large
- train epochs: 2
- train batch: 6
- eval batch: 5
- learning rate: 0.0009
- warmup ratio: 0.18
- scheduler: cosine
- loss-function: contrastive
- count of stratifiers: 6
- sentences per combination: 5

Appendix 2 - Classifier Hyperparameter

For the training of the classifier (stage 2), we implemented two general hyperparameters, with one of them selecting the classifier-model to be trained. Based on this selection, there was another set of model-specific hyperparameters.

Next to the selection of the classifier-model, we implemented the weighting procedure as a hyperparameter. The selection was between ‘balanced’, ‘none’, and ‘a bit of null’:

- "balanced" calculated the weights so that each motive class including null was weighed by $1/n_{class}$
- "none" meant all classes where weighted with 1
- "a bit of null" was an adaptation of the "balanced" weights after we realized the "balanced" weighting to over-zealously punish the Null class. To compensate for this, the "a bit of null" strategy multiplied the "balanced" weight of the Null class by a factor of 1.5.

We implemented the following models with the hyperparameters listed below as candidate classifiers. Where not indicated otherwise, we used the scikit-learn implementation of the model.

- CatBoost (Dorogush, Ershov, and Gulin 2018)
  - Depth: 2 to 16
  - Grow Policy: 'Depthwise', 'Lossguide', 'SymmetricTree'
  - Learning Rate: 0.0 to 1.0
- Decision Tree
  - Criterion: 'gini', 'entropy'
  - Maximum Depth: 0 to 100
  - Maximum Features: 'sqrt', 'log2', 'none'
  - Minimum Impurity Decrease: 0.0 to 1.0
  - Minimum Samples Leaf: 0.01 to 0.99
  - Minimum Samples Split: 2 to 100
  - Minimum Weight Fraction Leaf: 0.0 to 0.5
  - Splitter: 'best', 'random'
- Extra Trees (ETree)
  - Criterion: 'gini', 'entropy'
  - Maximum Depth: 0 to 100
  - Maximum Features: 'sqrt', 'log2'
  - Minimum Impurity Decrease: 0.0 to 1.0
  - Minimum Samples Leaf: 0.01 to 0.99
  - Minimum Samples Split: 2 to 100
  - Minimum Weight Fraction Leaf: 0.0 to 0.1
- Elastic Net (implemented as SGDClassifier with 'elasticnet' penalty)
  - Alpha: 0.001 to 1.0
  - L1 Ratio: 0.0 to 1.0
  - Tolerance: 1e-08 to 1.0
- Hist Gradient Boosting
  - L2 Regularization: 0.0 to 1.0
  - Learning Rate: 1e-08 to 1.0
  - Maximum Depth: 0 to 100
  - Maximum Iterations: 10 to 10000
  - Maximum Leaf Nodes: 2 to 100000
  - Minimum Samples Leaf: 1 to 1000
  - Tolerance: 0.0 to 0.1
- LightGBM (Ke et al. 2017)
  - Boosting Type: 'gbdt', 'dart'
  - Learning Rate: 0.0 to 1.0
  - Maximum Depth: -1 to 100
  - Minimum Child Samples: 1 to 10000
  - Minimum Child Weight: 0.0 to 1.0
  - Minimum Split Gain: 0.0 to 1.0
  - Number of Estimators: 10 to 1000
  - Number of Leaves: 8 to 100
  - Subsample for Bin: 0 to 5000000
- Logistic Regression
  - C: 0.0 to 5.0
  - Multi-Class: 'ovr', 'multinomial'
  - Solver: 'lbfgs', 'newton-cg', 'sag', 'saga'
  - Tolerance: 1e-08 to 1.0
- Multi-Layer Perceptron (custom implementation using pyTorch)
  - Batch Size: 5 to 7
  - Depth: 0 to 6
  - Dropout Layers [1-6]: True or False
  - Dropout Rate Layers [1-6]: 0.01 to 0.99
  - Epochs: 0 to 6
  - Gamma: 0.95 to 1.0
  - Learning Rate: 1e-10 to 0.1
  - Minimum Delta: 1e-08 to 0.01
  - Optimizer: 'adam', 'adagrad', 'sgd', 'adadelta'
  - Patience: 5 to 10
  - Regularization Layers [1-6]: True or False
  - Width Layers [1-6]: 4 to 12
- Naive Bayes
  - Alpha: 0.0 to 5.0
- Nearest Centroid
  - Metric: 'euclidean', 'manhattan'
- Passive Aggressive
  - C: 0.0 to 5.0
  - Early Stopping: True or False
  - Maximum Iterations: 500 to 10000
  - Shuffle: True or False
  - Tolerance: 1e-08 to 1.0
  - Validation Fraction: 0.0 to 0.5
- Perceptron
  - Alpha: 0.0 to 1.0
  - Early Stopping: True or False
  - Eta0: 1e-08 to 1.0
  - Maximum Iterations: 1 to 10000
  - Penalty: 'l1', 'l2', 'elasticnet', 'none'
  - Shuffle: True or False
  - Tolerance: 1e-08 to 1.0
  - Validation Fraction: 0.0 to 0.5
- Quadratic Discriminant Analysis (QDA)

  - Regularization Parameter: 0.0 to 1.0
- Random Forest
  - Criterion: ‘gini’, ‘entropy’, ‘log_loss’
  - Maximum Depth: 0 to 100
  - Maximum Features: ‘sqrt’, ‘log2’
  - Minimum Impurity Decrease: 0.0 to 1.0
  - Minimum Samples Leaf: 0.01 to 0.99
  - Minimum Samples Split: 2 to 100
  - Minimum Weight Fraction Leaf: 0.0 to 0.5
  - Number of Estimators: 0 to 10000
- Ridge Regression
  - Alpha: 0.1 to 5.0
  - Solver: ‘auto’, ‘svd’, ‘cholesky’, ‘lsqr’, ‘sparse_cg’, ‘sag’, ‘saga’
  - Tolerance: 1e-08 to 1.0
- XGBoost (Chen and Guestrin 2016)
  - Colsample Bytree: 0.0 to 1.0
  - Gamma: 0.0 to 100.0
  - Grow Policy: ‘depthwise’, ‘lossguide’
  - Learning Rate: 0.0 to 1.0
  - Maximum Depth: 5 to 100
  - Maximum Leaves: 0 to 80
  - Minimum Child Weight: 0.0 to 1000.0
  - Subsample: 0.0 to 1.0

The settings selected as best were the following:

- weighting-strategy: “a bit of null”
- model: Elastic Net
- Alpha: 0.88
- L1 Ratio: 0.21
- Tolerance: 0.11

Acknowledgements

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer (LE 4379/2-1). The authors would like to thank Prof. Dr. Michael Prange of Kiel University for Applied Sciences for providing access to the compute resources used in this project.

References

Chan, B.; Schweter, S.; and Möller, T. 2020. German’s Next Language Model. In Scott, D.; Bel, N.; and Zong, C., eds., *Proceedings of the 28th International Conference on Computational Linguistics*, 6788–6796. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Chen, T.; and Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Dorogush, A. V.; Ershov, V.; and Gulin, A. 2018. CatBoost: Gradient Boosting with Categorical Features Support. arXiv:1810.11363.

Drescher, A.; and Schultheiss, O. C. 2016. Meta-Analytic Evidence for Higher Implicit Affiliation and Intimacy Motivation Scores in Women, Compared to Men. *Journal of Research in Personality*, 64: 1–10.

Green, B. F.; Stone, P. J.; Dunphy, D. C.; Smith, M. S.; and Ogilvie, D. M. 1967. The General Inquirer: A Computer Approach to Content Analysis. *American Educational Research Journal*, 4(4): 397.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality Reduction by Learning an Invariant Mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*, volume 2, 1735–1742.

Jianlin, S. 2022. CoSENT: A More Efficient Sentence Vector Scheme than Sentence-BERT. https://kexue.fm/archives/8847.

Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; and Liu, T.-Y. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS’17, 3149–3157. Red Hook, NY, USA: Curran Associates Inc. ISBN 978-1-5108-6096-4.

Kohavi, R. 1995. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Ijcai*, volume 14, 1137–1145. Montreal, Canada.

Koller, M.; and Stahel, W. A. 2011. Sharpening Wald-type Inference in Robust Regression for Small Samples. *Computational Statistics & Data Analysis*, 55(8): 2504–2515.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

Maechler, M.; Rousseeuw, P.; Croux, C.; Todorov, V.; Ruckstuhl, A.; Salibian-Barrera, M.; Verbeke, T.; Koller, M.; Conceicao, E. L. T.; and Anna di Palma, M. 2024. *Robustbase: Basic Robust Statistics*.

McClelland, D. C.; Atkinson, J. W.; Clark, R. A.; and Lowell, E. L. 1953. Toward a Theory of Motivation. In *The Achievement Motive*, Century Psychology Series, 6–96. East Norwalk, CT, US: Appleton-Century-Crofts.

Morgan, C. D.; and Muray, H. A. 1935. A Method for Examining Fantasies” Dalam The Thematic Apperception Test. *Archives of Neurology & Psychiatry*, 34: 289–306.

Muennighoff, N.; Tazi, N.; Magne, L.; and Reimers, N. 2023. MTEB: Massive Text Embedding Benchmark. arXiv:2210.07316.

Nilsson, A.; Runge, J. M.; Kjell, O. N. E.; Soni, N.; Ganesan, A. V.; and Nilsson, C. V. 2024. Automatic Implicit Motives Codings Are as Accurate as Humans’, Cheaper, and 99% Faster.

Pang, J. S. 2010. Content Coding Methods in Implicit Motive Assessment: Standards of Measurement and Best Practices for the Picture Story Exercise. In Schultheiss, O. C.; and Brunstein, J. C., eds., *Implicit Motives*. Oxford: Oxford University Press. ISBN 978-0-19-533515-6 978-1-282-38812-3 978-0-19-971504-6.

Pang, J. S.; and Ring, H. 2020. Automated Coding of Implicit Motives: A Machine-Learning Approach. *Motivation and Emotion*, 44(4): 549–566.

Pang, J. S.; and Schultheiss, O. C. 2005. Assessing Implicit Motives in U.S. College Students: Effects of Picture Type and Position, Gender and Ethnicity, and Cross-Cultural Comparisons. *Journal of Personality Assessment*, 85(3): 280–294.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings Using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Reimers, N.; and Gurevych, I. 2020. Making Monolingual Sentence Embeddings Multilingual Using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Schönbrodt, F. D.; Hagemeyer, B.; Brandstätter, V.; Czikmantori, T.; Gröpel, P.; Hennecke, M.; Israel, L. S. F.; Janson, K. T.; Kemper, N.; Köllner, M. G.; Kopp, P. M.; Mojzisch, A.; Müller-Hotop, R.; Prüfer, J.; Quirin, M.; Scheidemann, B.; Schiestel, L.; Schulz-Hardt, S.; Sust, L. N. N.; Zygar-Hoffmann, C.; and Schultheiss, O. C. 2021. Measuring Implicit Motives with the Picture Story Exercise (PSE): Databases of Expert-Coded German Stories, Pictures, and Updated Picture Norms. *Journal of Personality Assessment*, 103(3): 392–405.

Schultheiss, O.; and Köllner, M. G. 2021. Implicit Motives. In John, O. P.; and Robins, R. W., eds., *Handbook of Personality: Theory and Research*. New York London: The Guilford Press, fourth edition edition. ISBN 978-1-4625-5048-7 978-1-4625-4495-0.

Schultheiss, O. C. 2013. Are Implicit Motives Revealed in Mere Words? Testing the Marker-Word Hypothesis with Computer-Based Text Analysis. *Frontiers in Psychology*, 4.

Schultheiss, O. C.; Liening, S. H.; and Schad, D. 2008. The Reliability of a Picture Story Exercise Measure of Implicit Motives: Estimates of Internal Consistency, Retest Reliability, and Ipsative Stability. *Journal of Research in Personality*, 42(6): 1560–1571.

Schultheiss, O. C.; and Pang, J. S. 2007. Measuring Implicit Motives. In Robins, R.; Fraley, R.; and Krueger, R., eds., *Handbook of Research Methods in Personality Psychology*, 322–344. Guilford Publications. ISBN 978-1-60623-612-3.

Smith, C. P. 1992. *Motivation and Personality: Handbook of Thematic Content Analysis*. Cambridge University Press.

Smith, C. P.; Feld, S. C.; and Franz, C. E. 1992. Methodological Considerations: Steps in Research Employing Content Analysis Systems. In Smith, C. P., ed., *Motivation and Personality: Handbook of Thematic Content Analysis*, 515–536. Cambridge: Cambridge University Press. ISBN 978-0-521-40052-7.

Smith, M. S. 1968. The Computer and the TAT. *Journal of School Psychology*, 6(3): 206–214.

Sturua, S.; Mohr, I.; Akram, M. K.; Günther, M.; Wang, B.; Krimmel, M.; Wang, F.; Mastrapas, G.; Koukounas, A.; Koukounas, A.; Wang, N.; and Xiao, H. 2024. Jina-Embeddings-v3: Multilingual Embeddings with Task LoRA. arXiv:2409.10173.

Tunstall, L.; Reimers, N.; Jo, U. E. S.; Bates, L.; Korat, D.; Wasserblat, M.; and Pereg, O. 2022. Efficient Few-Shot Learning Without Prompts. arXiv:2209.11055.

Wang, L.; Yang, N.; Huang, X.; Yang, L.; Majumder, R.; and Wei, F. 2024. Multilingual E5 Text Embeddings: A Technical Report. arXiv:2402.05672.

Watanabe, S. 2023. Tree-Structured Parzen Estimator: Understanding Its Algorithm Components and Their Roles for Better Empirical Performance. arXiv:2304.11127.

Winter, D. G. 1994. *Manual for Scoring Motive Imagery in Running Text*. Department of Psychology, University of Michigan.

Xu, M. 2023. Text2vec: A Tool for Text to Vector.

Assessing Perspective-Taking in Texts: Inspirations for Analytical Targets from Socio-Cognitive Narrative Coding Schemes

Paul Compensis

Department of Psychology, University of Erlangen-Nuremberg & Institute of Psychology, University of Bamberg, Germany

paul.compensis@fau.de

Abstract

Inferring interpersonal processes such as perspective-taking and empathy from text is a valuable yet not widely developed tool to study inter-individual differences in social interaction. Going beyond lexical categorization, more pragmatically oriented classification procedures can profit from insights from traditional qualitative tools in social-cognitive research, especially narrative coding. In this short paper, I briefly discuss the concept of perspective-taking and its points of contact with language. I then illustrate a specific narrative coding system (Feffer's decentring scales) and make suggestions as to how this system offers targets for automatized classification procedures, offering impulses for the integration of natural language processing and research into social interaction.

Language and Interpersonal Processes

Human language plays a significant role in interpersonal interaction. It is therefore not surprising that language is closely intertwined with interpersonal processes, with parallel developmental trajectories (Garfield et al. 2001), reciprocal influence on each other in behaviour (Mass et al. 2016) and correlated impairments between the two in some neuropsychiatric disorders (Cummings 2021). Despite this close link, there has been only relatively little research (beyond developmental research) on how inter-individual differences in interpersonal processes are manifested in language, how specific linguistic patterns relate to these social-cognitive processes and how linguistic and social cues are integrated in this domain (Holtgraves and Kashima 2007).

Among the most relevant interpersonal processes typically present in social interaction is the ability to understand cognitive mental states (such as beliefs and intentions) as well as affective mental states (such as emotions or pain). Humans differ to some extent in their ability to use these processes, but even more so, situational and interactional features determine to what extent mental states of others are assessed and evaluated in a given situation. Hence, it is highly relevant for social cognition research to assess both the ability and the propensity to use these processes (given a particular situational cue) and to understand in more detail which factors determine them. In this regard, developing more systematic (and automatic) classification procedures for perspective-taking (and empathy) based on language is therefore highly relevant and should also take into account insights from previous linguistic and social-psychological work that provide avenues for future research.

Links between empathy, perspective-taking and language are particularly visible in the context of pronoun use and referential structure (i.e., keeping track of referents in discourse and managing common ground with an interlocutor). For instance, self-reported empathy correlates positively with the number of pronouns and words of social reference used in blogs (Litvak et al. 2016; Yaden et al. 2023). Also, empathic individuals react more sensitively to referential organizing principles such as topic continuity and person hierarchies (Kann et al. 2023) but also to mismatches of social cues during language processing, e.g., social identity (van den Brink et al., 2012) or emotion expressions (Dozolme et al. 2015; Rak et al. 2009). Also, clinical populations with interpersonal impairments (esp. autism and schizophrenia) often exhibit linguistic difficulties – particularly in the referential structure (Hinzen 2022; Schroeder et al. 2023).

Besides linguistically informed analyses, a number of text-based narrative coding systems were developed in the past to assess interpersonal processes, among others, Teglasis's empathy score (Teglasi et al. 2008) and Feffer's decentring score (FDS; Feffer 1959). Both procedures focus on an evaluation of interactions between two discourse referents in narratives, thereby drawing on more general semanto-pragmatic aspects that are harder to capture with traditional linguistic analyses and harder to implement in automatic procedures. Nonetheless, some aspects of these narrative coding systems can inform automatic classification procedures. Pointing out some of these aspects is the goal of this short discussion article. In this contribution, I focus primarily on perspective-taking and outline aspects of FDS that might inform future implementations of automatic perspective-taking classification tools. Note, however, that I do not discuss technical specifications in the present contribution

Proceedings of the First Workshop on Integrating NLP and Psychology to Study Social Interactions (NLPSI) @ICWSM '25, pages 39–44

but only offer suggestions for future work. I start by briefly discussing the concept of perspective-taking. I then illustrate how perspective-taking was assessed from text in the past. Ultimately, I discuss FDS as one specific procedure and emphasize components that might inform automatized classification procedures of perspective-taking from written text.

The Concept of Perspective-Taking

Perspective-taking, the degree to which a person is able or willing to consider mental states (intentions, beliefs, emotions) of another person, is an important interpersonal process and a central feature of social cognition. An important question under debate is whether perspective-taking is the same as cognitive empathy. Many authors simply equal these two terms (e.g., Duan and Hill 1996) but others argue that perspective-taking is just one route for achieving cognitive empathy, next to drawing inferences from facial expressions (e.g., Besel and Yuille 2010) or by projecting oneself in the position of the other person (e.g., Preston 2007). This latter process is sometimes described as "imagining how one would think and feel in the other's place" (Batson 2011: 7) and sometimes also called perspective-taking (Batson 2011; on this debate, see also Cuff et al. 2016; Stietz et al. 2019).

While both concepts are often broadly described as the ability to "put oneself in the shoes of another person" (Batson 2011), I cannot elaborate on this question here and simply focus on perspective-taking in a narrower sense, namely as the process of "imagining how another is thinking and feeling" (Batson 2011: 7) by forming representations of another person's internal states. That is, the focus is on the cognitive process of capturing mental states of another person (a process also called mentalizing; Frith and Frith 2006).

Similarly, the concept of theory of mind (ToM; originally developed in primate research, Premack and Woodruf 1978; and later adapted to describe the ability to "explain people's behaviour on the basis of their minds: their knowledge, their beliefs and their desires", Frith and Frith 2005: R644) is often used synonymously in this context but typically puts more emphasis on the content of the mental representation. Hence, I consider ToM as describing the ability to represent mental states of others (Frith and Frith 2005; Premack and Woodruff 1978) and perspective-taking as the propensity to apply these processes routinely in a specific social situation.

Relevant for the present discussion is also a particular concept brought forward by Feffer (1959). Feffer adopted Piaget's (1932, 1947) concept of decentring and applied it to the social domain. Feffer et al. (2008: 150) later describe social decentring as the ability of a person "to view her or his planned behaviour from the point of view of another person who may be affected by the action". This focus on alternating perspectives to and away from different people closely matches perspective-taking as the general propensity to grasp another person's internal states and imagine how another person is thinking or feeling, as outlined above. It is important to note here that Feffer et al. (2008) do not distinguish between the representation of another person's cognitive and emotional states – therefore laying the focus on the cognitive process of taking perspectives and attributing internalizations – irrespective of whether the internal state is a belief, intention, or emotion. Capturing this process based on written text is an interesting tool to assess socio-cognitive aspects of individuals and to identify linguistic markers of how humans represent others in their own mental states.

Assessing Perspective-Taking in Text

Looking at the language people use while drawing on interpersonal processes is an interesting procedure that also overcomes some limitations of questionnaire-based assessments of these processes. The assessment can be based on naturally occurring text (emails, social media content, e.g., Yaden et al. 2023) or text material elicited in a controlled manner (see below). An advantage of analysing written text, in comparison to observing behaviour or self-reports, is that it constitutes a more implicit measure that bypasses social desirability to some extent. Also, while behavioural tasks are often interpreted as performance tests by participants, the same is not true for writing. Text-based procedures are therefore well-equipped to assess the regular drawing on interpersonal processes – keeping situational cues relatively controlled.

Text material may then be assessed either by inspecting formal means of language (word frequencies, narrative structure etc.) or by using narrative content coding. One of the most widely used lexical classification tools in social psychology is Language Inquiry and Word Count (LIWC; Pennebaker et al. 2001). LIWC is relatively simple and classifies (and counts) occurring words in a text based on a number of pre-defined word categories ("dictionaries"), either based on linguistic features (e.g., pronouns) or psychological or semantic dimensions (e.g., *to see* and *to touch* being classified in the category SENSES) – allowing a rapid assessment of broad language-related features for a number of psychological processes (Tausczik and Pennebaker 2010).

Social processes can be classified with LIWC by looking into the number of words that refer to others (especially 3rd person pronouns) and those that are associated with the social domain (such as *friend*, *family*, *to meet*; see Pennebaker et al. 2004). Also, these LIWC categories are also correlated with the implicit need for affiliation (Schultheiss 2013).

Another approach is narrative coding, that is, the manual assessment of text based on pre-defined coding principles. For instance, Teglasi et al. (2008)'s empathy score captures three dimensions of empathic responding between two referents in a given discourse. In contrast, FDS focusses more strongly on the process of taking perspective by capturing

the depth of this process based on a single scale (for details, see below). Importantly, both procedures were not developed to assess a given interaction between two people but rather to assess narratives about two or more discourse referents written by a person (based on a previous story or image). That is, these narrative coding systems focus specifically on the question how observed interactions are perceived and represented by a person, thereby laying emphasis on social representation processes rather than interpersonal behaviour. This is also evident in how interpersonal decentring is conceptualized, namely as "an aspect of operational thinking that involves an ability to be aware of, respond to, and anticipate another person's ideas, thoughts, feelings, or actions" (Feffer et al. 2008). In the following, I illustrate how this conceptualization is implemented FDS.

Feffer's Decentring Scales (FDS)

FDS is a narrative coding system for perspective-taking that attempts to capture the extent to which subjects describe social interactions and attribute internal or mental states to discourse referents depicted in brief narratives (Feffer 1959; Feffer et al. 2008). Traditionally, text material for FDS was typically elicited with the Thematic Apperception Task (Murray 1943) or its derivate, the Picture Story Exercise (McClelland et al. 1989; Pang and Schultheiss 2005; Schultheiss and Pang 2007). In this task, participants see images with socially ambiguous scenes and differing levels of interaction and are then asked to write a story based on the image. The basic setup (and code examples) is given in Figure 1.

(Manual) Coding Rules

Coding Feffer's scores takes place in two steps. First, text material is screened for social interaction units (SIUs) by assessing whether there are (explicitly mentioned) overt interactions between two or more people present in the event description (for instance, the example code 2 in Figure 1) or whether interactions are present in form of internal processes of the people involved (Code 7 and 9 in Figure 1).

In a second step, each social interaction unit receives a score ranging from 1 to 9 (or 0 in the absence of any interaction). Pre-internalization processes (i.e., one-dimensional and sequential interactions without internalization or mentalizing processes such as "She watches the guitar player"; score: 2) receive scores from 1 to 4. In contrast, internalization processes (5 to 9) reflect multidimensional incorporations of thoughts and feelings of another person into the thoughts and feelings of a person. For instance, internalized states of another person (such as "As James knows how fascinated Georgia is by live music.") receive a scoring of 7 whereas internalized self-other processes ("He knows that he will love her forever.") are scored with a score of 9 (further details are described in depth in Feffer et al. 2008).

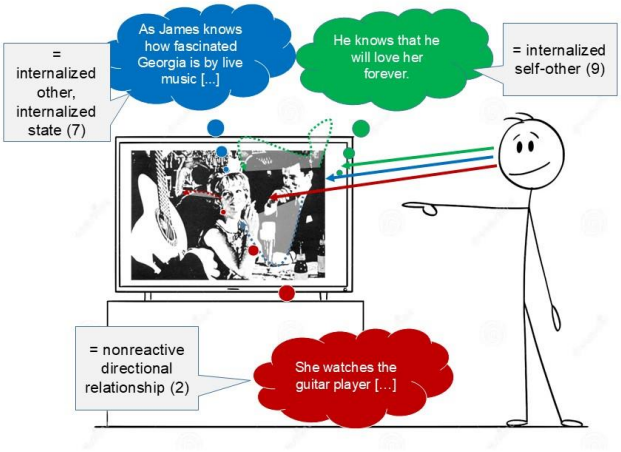


Figure 1: Illustration of Feffer's Decentring Scales.

In theory, the scores are ordinal scales but are commonly treated as interval scales for averaging and statistical analyses. Composite scores are calculated in different ways, either by simply averaging all scores within and across stories. Alternative suggestions were made to only use the single maximum score of a person or to only average the maximum scores of each story per person (see Feffer et al. 2008).

Training coders is relatively simple given the simplicity of the procedure and interrater reliabilities are typically very high (for instance, my research assistants achieved r = .944*** or κ = .824, 95% CI .782 - .865 in 120 stories).

Relation to Other Measures from Pilot Work

In a recent study (n = 239), we compared FDS with the two aforementioned LIWC categories (Compensis and Schultheiss in prep). FDS was significantly correlated (r = .28***) with the number of pronouns referring to others and with social words (r = .21**). Interestingly, however, FDS was confounded with word count (r = .35**), while pronouns referring to others (r = .12) and social words (r = -.02) were not. In another study (n = 168), FDS was positively correlated (r = .57**) with the need for power, an implicit motive that is characterized by a strong social component (Beuermann et al. in prep). To clarify further FDS's relation to other socio-cognitive and affective dimensions, we currently conduct a study in which we relate FDS to standard measures, among others, the Interpersonal Reactivity Index (Davis 1983) and the Reading the Mind in the Eyes Task (Baron-Cohen et al. 2001).

Suggestions for an Implementation of FDS

FDS is a relatively straight-forward and simple way to assess basic perspective-taking skills in narratives and can be

applied to all kinds of texts in which people mention and talk about other people, especially when interactions between two or more people are described or when one person is thinking about other people and attributes internal states to them. While the specific coding scale in its full form is probably restricted to its application in the context of narratives elicited from particular stimuli (image descriptions or retelling of stories), some main aspects of the coding procedure can inform more general assessments of perspective-taking based on text – as long as the text captures at least some dimension of social interaction of discourse referents. In the following, I briefly emphasize some aspects of FDS that can be used as a basis for developing more refined (and automatic) classification tools of this interpersonal process.

Identifying Social Interaction Unis (SIUs)

As was described above, the central level of analysis are social interaction units, that is, parts of the text in which discourse referents interact with each other (thereby fully excluding those parts of the discourse that simply describe the scene or surroundings or non-interactional behaviour of referents). Within FDS, SIUs are typically identified by changes in character, place or time. Recognizing these turns is based on a subjective assessment of the overall scene but could be approximated by mapping the referential structure (especially in terms of activation of discourse referents) and the event structure itself (by identifying transitivity).

Capturing the presence of discourse referents and closely tracking referential persistence and referential shifts are most likely the strongest indicators. Since discourse referents are typically introduced with their full names ("James") or via indefinite nominal phrases ("a friend of mine") and then referred to further depending on their activation in discourse (that is, with definite reference or, at the highest level of activation, by personal pronouns) and since referential shifts are often indicated again by indefinite reference (of a new referent) or by using demonstrative pronouns (Ariel 1990), it is relatively simple to model referential structure as a first proxy to capture the presence of social interactions.

To model interactional relationships further, verbal associations between referents could be classified based on transitivity to identify directional interactions (for contexts such as "she sees him") and by assessing psychological and functional aspects of the verbs in order to identify the use of higher-ordered processes (e.g., psych verbs such as "he loves her"). In this context, capturing recursivity could additionally dissociate between mid and high-level perspective-taking according to FDS (e.g., psych verbs embedded in matrix clauses such as "he knows he loves her" representing higher degrees of perspective-taking according to Feffer et al. 2008). Obviously, approximating social interaction via referential structure could also be enhanced further by applying sophisticated dependency parsing to the texts.

Lexical Frequencies within SIUs

A complementary solution to map perspective-taking in texts inspired by FDS is to apply a lexical approach (that is, measuring word frequencies and identifying relevant categories in terms of functional or semantic aspects, similar to LIWC). It would be particularly fruitful to apply this approach to text within SIUs (and contrast it with text outside SIUs). Also, given an adequate dataset, one could apply lexical and also factorial analyses to SIUs of a particular level, for instance, to identify lexical markers that are specific for this level or to distinguish directional interaction from internalized representations characteristic of higher levels.

Again, there is a strong intuition that psych verbs but also pronouns feature more prominently within SIUs but the correlation of FDS with LIWC's social category also hints at the presence of other relevant categories. Clearly, assessing lexical content for each level separately would provide a first step to identify reliable lexical markers that could then be compared to alternative approaches (such as LIWC) and be related to other socio-cognitive and affective measures – in order to inform classification mechanisms of perspective-taking based on lexical features and dependency parsing.

Conclusion

Extracting information on inter-individual differences and situational determinants of interpersonal processes (such as perspective-taking and empathy) based on texts is a valuable tool with manifold applications in the area of social psychology, affective computing as well as clinical and therapeutic contexts. Therefore, developing classification tools based on characteristic features is important and should also take into account insights from previous work on social cognition.

To offer some inspiration, I presented a traditional narrative-coding procedure from social-cognitive research (Feffer's Decentring Scale) in this short contribution and illustrated how FDS is used to assess the interpersonal process of perspective-taking (i.e., the propensity to grasp another person's internal states and imagine how another person is thinking or feeling) based on coding text. FDS captures the representational side of these processes and therefore can inform automatic classification procedures of perspective-taking in text, especially by suggesting analytical targets that could be the focus of these classification tools. Future research should therefore assess further how FDS relates externally to other socio-cognitive and socio-affective aspects and specify internally which linguistic features are characteristic of capturing perspective-taking as an interpersonal process. This contribution offers first rough points of contact for an automatic implementation based on insights from narrative coding and contributes to the integration of natural language processing and social interaction research.

References

Ariel, M. 1990. *Assessing noun phrase antecedents*. London: Routledge.

Baron-Cohen, S.; Wheelwright, S.; Hill, J.; Raste, Y.; and Plumb, I. 2001. The "Reading the Mind in the Eyes" Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry, and Allied Disciplines* 42(2): 241–251. doi.org/10.1111/1469-7610.00715.

Batson, C. D. 2011. These things called empathy: Eight related but distinct phenomena. In *The Social Neuroscience of Empathy*, edited by J. Decety and W. Ickes, 3–16. Cambridge, MA: MIT Press.

Besel, L. D. S.; and Yuille, J. C. 2010. Individual differences in empathy: The role of facial expression recognition. *Personality and Individual Differences* 49: 107–112. doi.org/10.1016/j.paid.2010.03.013.

Beuermann, F. L.; Compensis, P.; and Schultheiss, O. S. in prep. Implicit need for power and social decentring. Erlangen, Germany: University of Erlangen-Nuremberg.

Compensis, P.; and Schultheiss, O. S. in prep. Gender and hormonal differences in perspective-taking and social orientation as reflected in language usage. Erlangen, Germany: University of Erlangen-Nuremberg.

Cuff, B. M. P.; Brown, S. J.; Taylor, L.; and Howat, D. J. 2016. Empathy: A review of the concept. *Emotion Review* 8(2): 144–153. doi.org/10.1177/1754073914558466.

Cummings, L. 2021. *Handbook of Pragmatic Language Disorders*. Berlin: Springer. doi.org/10.1007/978-3-030-74985-9.

Davis, M. H. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology* 44(1): 113–126. doi.org/10.1037/0022-3514.44.1.113.

Dozolme, D.; Brunet-Gouet, E.; Passerieux, C.; and Amorim, M. A. 2015. Neuroelectric Correlates of Pragmatic Emotional Incongruence Processing: Empathy Matters. *PLoS One* 10(6): e0129770. doi.org/10.1371/journal.pone.0129770.

Duan, C.; and Hill, C. E. 1996. The current state of empathy research. *Journal of Counseling Psychology* 43(3): 261–274. doi.org/10.1037/0022-0167.43.3.261.

Feffer, M. 1959. The cognitive implications of role taking behavior. *Journal of Personality* 27: 152–168. doi.org/10.1111/j.1467-6494.1959.

Feffer, M.; Leeper, M.; Dobbs, L.; Jenkins, S. R.; and Perez, L. E. 2008. Scoring manual for Feffer's interpersonal decentering. In *A Handbook of Clinical Scoring Systems for Thematic Apperceptive Techniques*, edited by S. R. Jenkins, 157–180. Routledge.

Frith, C.; and Frith, U. 2005. Theory of mind. *Current Biology* 15(17): 644–646. doi.org/10.1016/j.cub.2005.08.041.

Frith, C.; and Frith, U. 2006. The neural basis of mentalizing. *Neuron* 50(4): 531–534. doi.org/10.1016/j.neuron.2006.05.001.

Garfield, J. L.; Peterson, C. C.; and Perry, T. 2001. Social Cognition, Language Acquisition and The Development of the Theory of Mind. *Mind and Language* 16(5): 494–541. doi.org/10.1111/1468-0017.00180.

Hinzen, W. 2022. Rethinking the role of language in autism. *Autism, Language, Communication and Cognition* 4(1): 129–151. doi.org/10.1075/elt.00040.hin.

Holtgraves, T. M.; and Kashima, Y. 2007. Language, Mening, and Social Cognition. *Personality and Social Psychology Review* 12(1): 73–94. doi.org/10.1177/1088868307309605.

Kann, T.; Berman, S.; Cohen, M. S.; Goldknopf, E.; Gülser, M.; Erlikhman, G.; Trinh, K.; Yokoyama, O. T.; and Zaidel, E. 2023. Linguistic Empathy: Behavioral measures, neurophysiological correlates, and correlation with Psychological Empathy. *Neuropsychologia* 108650. doi.org/10.1016/j.neuropsychologia.2023.108650.

Litvak, M.; Otterbacher, J.; Ang, C. S.; and Atkins, D. 2016. Social and linguistic behavior and its correlation to trait empathy. In *Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, edited by M. Nissim, V. Patti, and B. Plank, 128–137. Association for Computational Linguistics. https://aclanthology.org/W16-4314.

Maass, A.; Cervone, C.; and Ozdemir, I. 2016. Language and Social Cognition. In *Oxford Research Encyclopedia of Psychology*, edited by O. J. Braddick and W. E. Pickren. Oxford University Press. doi.org/10.1093/acrefore/9780190236557.013.279.

McClelland, D. C.; Koestner, R.; and Weinberger, J. 1989. How do self-attributed and implicit motives differ? *Psychological Review* 96(4): 690–702. doi.org/10.1037/0033-295X.96.4.690.

Murray, H. A. 1943. *Thematic Apperception Test manual*. Harvard University Press.

Pang, J. S.; and Schultheiss, O. C. 2005. Assessing implicit motives in U.S. college students: Effects of picture type and position, gender and ethnicity, and cross-cultural comparisons. *Journal of Personality Assessment* 85(3): 280–294. doi.org/10.1207/s15327752jpa8503_04.

Pennebaker, J. W.; Groom, C. J.; Loew, D.; and Dabbs, J. M. 2004. Testosterone as a social inhibitor: Two case studies of the effect of testosterone treatment on language. *Journal of Abnormal Psychology* 113(1): 172–175. doi.org/10.1037/0021-843X.113.1.172.

Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 2001. *Linguistic Inquiry and Word Count (LIWC): LIWC2001*. Erlbaum.

Piaget, J. 1932. *Le jugement moral chez l'enfant* [The moral judgment of the child]. F. Alcan.

Piaget, J. 1947. *La psychologie de l'intelligence* [The psychology of intelligence]. A. Colin.

Premack, D.; and Woodruff, G. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences* 1(4): 515–526. doi.org/10.1017/S0140525X00076512.

Preston, S. D. 2007. A perception–action model for empathy. In *Empathy in Mental Illness*, edited by T. Farrow and P. Woodruff, 428–447. Cambridge University Press.

Rak, N.; Kontinen, J.; Kuchinke, L.; and Werning, M. 2013. Does the Semantic Integration of Emotion Words Depend on Emotional Empathy? N400, P600 and Localization Effects for Intentional and Proprioceptive Emotion Words in Sentence Contexts. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, edited by M. Knauff, M. Pauen, N. Sebanz, and I. Wachsmuth, 2304–2309. Austin, TX.

Schroeder, K.; Rosselló, J.; Torrades, T. R.; and Hinzen, W. 2023. Linguistic markers of autism spectrum conditions in narratives: A comprehensive analysis. *Autism & Developmental Language Impairments* 8: 23969415231168557. doi.org/10.1177/23969415231168557.

Schultheiss, O. C. 2013. Are implicit motives revealed in mere words? Testing the marker-word hypothesis with computer-based

text analysis. *Frontiers in Psychology* 4: 748. doi.org/10.3389/fpsyg.2013.00748.

Schultheiss, O. C.; and Pang, J. S. 2007. Measuring implicit motives. In *Handbook of Research Methods in Personality Psychology*, edited by R. W. Robins, R. C. Fraley, and R. Krueger, 322–344. Guilford.

Stietz, J.; Jauk, E.; Krach, S.; and Kanske, P. 2019. Dissociating empathy from perspective-taking: Evidence from intra- and inter-individual differences research. *Frontiers in Psychiatry* 10: 126. doi.org/10.3389/fpsyt.2019.00126.

Tausczik, Y. R.; and Pennebaker, J. W. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology* 29(1): 24–54. doi.org/10.1177/0261927X09351676.

Teglasi, H.; Locraft, C.; and Felgenhauer, K. 2008. Scoring Manual for Empathy. In *A Handbook of Clinical Scoring Systems for Thematic Apperceptive Techniques*, edited by S. R. Jenkins, 623–648. Routledge.

van den Brink, D.; van Berkum, J. J. A.; Bastiaansen, M. C. M.; Tesink, C. M. J. Y.; Kos, M.; Buitelaar, J. K.; and Hagoort, P. 2012. Empathy matters: ERP evidence for inter-individual differences in social language processing. *Social Cognitive and Affective Neuroscience* 7(2): 173–183. doi.org/10.1093/scan/nsq094.

Yaden, D. B., Giorgi, S., Jordan, M., Buffone, A., Eichstaedt, J. C., Schwartz, H. A., Ungar, L., & Bloom, P. (2023). Characterizing empathy and compassion using computational linguistic analysis. *Emotion*. Advance online publication. https://doi.org/10.1037/emo0001205

A Hybrid Theory and Data-driven Approach to Persuasion Detection with Large Language Models

Gia Bao Hoang[1], **Keith J. Ransom**[1], **Rachel G. Stephens**[1],
Carolyn Semmler[1], **Nicolas Fay**[2], **Lewis Mitchell**[1]
[1]University of Adelaide
[2]University of Western Australia
{giabao.hoang, keith.ransom, rachel.stephens, carolyn.semmler, lewis.mitchell}@adelaide.edu.au,
nicolas.fay@uwa.edu.au

Abstract

Traditional psychological models of belief revision focus on face-to-face interactions, but with the rise of social media, more effective models are needed to capture belief revision at scale, in this rich text-based online discourse. Here, we use a hybrid approach, utilizing large language models (LLMs) to develop a model that predicts successful persuasion using features derived from psychological experiments.

Our approach leverages LLM generated ratings of features previously examined in the literature to build a random forest classification model that predicts whether a message will result in belief change. Of the eight features tested, *epistemic emotion* and *willingness to share* were the top-ranking predictors of belief change in the model. Our findings provide insights into the characteristics of persuasive messages and demonstrate how LLMs can enhance models of successful persuasion based on psychological theory. Given these insights, this work has broader applications in fields such as online influence detection and misinformation mitigation, as well as measuring the effectiveness of online narratives.

1 Introduction

Changing someone's opinion is a key aspect of communication, with broad implications for both individual decisions and societal outcomes. From political campaigns to marketing strategies and daily interactions, understanding how opinions change has been the subject of extensive research, (see e.g., Cialdini 1993; Dillard and Pfau 2002; Petty and Cacioppo 2012; Reardon 1991). With the rise of online social platforms, interpersonal persuasion can be observed and enacted on a massive scale, not just in direct conversation but also through text-based communication. As Fogg (2008) highlights, debate and argumentation are increasingly shifting to digital spaces. This introduces new complexities in understanding how persuasive messages are created, received, and how they influence the revision of beliefs in online environments.

Persuasion in these environments is particularly important to understand given global concerns about phenomena such as misinformation, polarization, and echo chambers (see e.g., Arechar et al. 2023; Barberá 2020; Weber et al. 2022). In addition, influence campaigns are shown to be strategically coordinated efforts designed to shape public opinion using social diffusion processes and media exposure (Hwang 2012; Panagopoulos 2012). Therefore, early detection and a deeper understanding of how these processes unfold, along with identifying the characteristics of successful persuasion, are crucial.

Belief revision is a complex process. Beyond the content of a message, outcomes depend on how individuals engage with information, evaluate its credibility, and respond emotionally. Epistemic emotions (such as curiosity, confusion, and surprise) motivate deeper cognitive engagement (Muis et al. 2018), and have been shown to facilitate persuasion when they prompt active message processing (Briñol and Petty 2012). In parallel, emotional valence influences how information is processed, shaping attention, memory, and judgment (Lerner et al. 2015). These internal responses often determine whether persuasion occurs, yet they are not always evident from surface-level linguistic features. This complexity underscores the need for models that incorporate psychological insight, not just observable text patterns.

While foundational psychological models like the Elaboration Likelihood Model (ELM) (Petty and Cacioppo 1986) and the Heuristic-Systematic Model (HSM) (Chaiken 1989) have shaped our understanding of persuasion, they are primarily descriptive and offer limited utility for predicting persuasive success at the level of individual messages. Separately, researchers have also attempted to use machine learning and large language models (LLMs) to analyze persuasion at scale (Tan et al. 2016; Bai et al. 2023; Salvi et al. 2024). While promising, these approaches often function as black boxes, offering little insight into how persuasive effects are achieved. As a result, the relationship between linguistic features and psychological theories of persuasion remains difficult to interpret and apply in practice.

In this work, we demonstrate how psychological theories can be combined with LLMs' statistical representation of language, to enhance existing machine learning architectures to predict persuasion outcomes in text-based discussions. This study utilizes two datasets: the Winning Arguments dataset (Tan et al. 2016) for real-world online persuasion analysis, and the Truth Wins dataset (Fay et al. 2024), which provides controlled, human-annotated data from experiments on persuasive and attention-seeking messages. Our focus is on classifying successful versus unsuccessful

Proceedings of the First Workshop on Integrating NLP and Psychology to Study Social Interactions (NLPSI) @ICWSM '25, pages 45–56

persuasive attempts in the Winning Arguments dataset, using features derived from the Truth Wins dataset. Our final model leverages a Random Forest architecture trained on LLM-generated persuasion ratings, inspired by psychological experiments. The model predicts whether a message will lead to successful persuasion. Additionally, we examine which features are the most influential predictors within the classification model.

2 Related Work

The ELM and the HSM are foundational theories in persuasion research, both proposing dual routes to attitude change, either through deep, content-focused processing or more superficial cues (Petty and Cacioppo 1986; Chaiken 1989). While influential, these models are largely descriptive, and offer little practical guidance for predicting persuasive success at the level of individual messages. Their application often requires adaptation to specific contexts, topics, and communication media, which limits their scalability in computational settings.

In response, researchers have explored ways to make persuasion more measurable. Zhao et al. (2011) proposed the Perceived Argument Strength (PAS) scale as a subjective measure of persuasive quality, but it relies on self-reports and lacks scalability. Youk et al. (2024) developed the Measures of Argument Strength (MAS), identifying linguistic features – such as citations, abstractness, and moral language – that are associated with higher perceived persuasiveness in large-scale online debates. Lukin et al. (2017) further emphasized how personality traits shape responses to emotional versus factual arguments, highlighting the importance of audience adaptation. While these studies move toward operationalizing persuasion, they remain constrained by subjective judgement, context sensitivity, handcrafted features, and limited predictive power – motivating the further shift toward data-driven approaches, discussed next.

Online platforms like Change My View (CMV) subreddit have become an important place to study persuasion at scale. In CMV, users influence opinions via textual exchanges. Two important features of CMV that are valuable for computational work are the *delta symbol*, awarded by a user to the reply that changed their view, and the *karma score*, which reflects community endorsement through up-votes and down-votes. While deltas serve as explicit markers of persuasive success, karma is often used as a proxy for argument quality. In addition, as the forum is highly moderated, comments are usually argumentative, containing reasoning rather than *ad hominem* attacks or other form of online bullying.

Tan et al. (2016) conducted a large-scale analysis of CMV (the Winning Arguments dataset), focusing on delta award predictions, based only on delta given by the original poster (OP). They examined the interaction dynamic associated with successful attempts at persuasion, and built a logistic regression model based on language factors and linguistic style. The model was able to predict delta changes effectively, but struggled with reliably distinguishing stance malleability (whether an OP will change their position), indicating difficulty in reliably distinguishing stance shifts and highlighting the complexity of persuasion in online discourse. Subsequent studies have built on this dataset using more advanced techniques. Dutta, Das, and Chakraborty (2020) introduced an attention-based hierarchical Long Short-Term Memory model to classify persuasive conversations on CMV by analyzing argument-specific components, such as claims and premises. Similarly, Wei, Liu, and Li (2016) used karma scores as a proxy for success, finding that argumentative features outperformed surface-level cues like length and punctuation in predicting persuasiveness. These works demonstrate the promise of data-driven persuasion modeling, but also expose its dependence on platform-specific feedback, limiting broader applicability.

Recent work has shifted from predicting persuasive outcomes to detecting persuasive strategies, often using deep learning and transformer-based NLP models. Karki et al. (2022) classified Cialdini's six persuasion principles (Reciprocation, Consistency, Social Proof, Likeability, Authority, and Scarcity) in phishing emails using machine-learning transformers based on BERT. In a related direction, Wang et al. (2020) developed a personalized persuasive dialogue system by integrating sentence-level and contextual features, including turn position, sentiment, and character embeddings, enabling the classification of persuasion strategies in conversations. While these models perform well, they remain largely opaque and task-specific, offering limited insight into underlying persuasive mechanisms and weak connections to psychological theory, underscoring the need for models that are both interpretable and cognitively grounded.

Despite recent advances, traditional methods for persuasion detection continue to face challenges related to scalability, generalization, and personalization, often relying on large annotated datasets and lengthy text inputs. Large language models (LLMs) offer a promising alternative, leveraging their statistically grounded representations of human language to overcome these limitations. However, LLMs introduce new challenges – most notably, a lack of transparency. Operating as black-box systems, they are often difficult to interpret, which limits their usefulness in contexts requiring explainability. The current research project addresses these gaps by building on and extending three key areas of prior work: (1) models for predicting persuasive success, (2) studies identifying persuasive strategies and message characteristics, and (3) applications of LLMs in persuasion contexts. This integrative approach supports real-time applications such as influence campaign detection and misinformation mitigation, while also enabling greater flexibility in identifying and analyzing the persuasive elements within language.

3 Experimental Setup

Figure 1 outlines the workflow of our hybrid model for persuasion classification in target data such as the Winning Arguments dataset. The goal is to predict whether a response successfully changes the original poster's view by leveraging interpretable, theory-driven features in combination with scalable language model outputs. We begin by identifying a set of psychologically grounded features from the Truth Wins dataset (detailed below), including Attention, Interesting-If-True, Positive Emotion, Negative Emo-

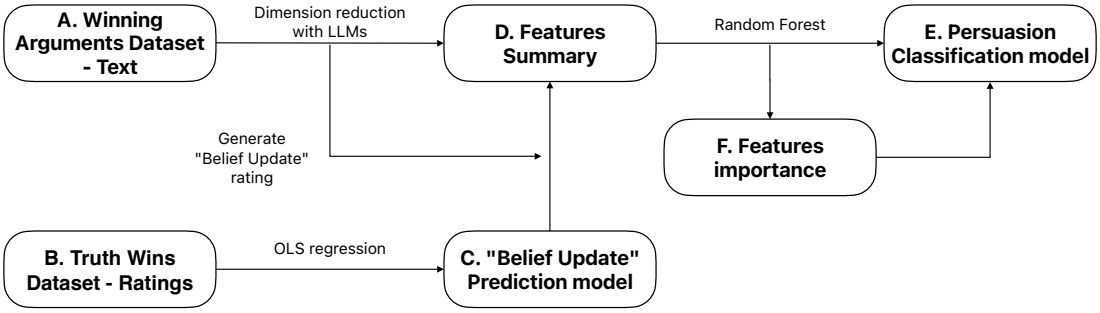


Figure 1: Flowchart showing the main steps of our hybrid model approach.

tion, Influential, Shareable, and Truthfulness (Figure 1.B). These features form the basis for a structured representation of persuasive content.

Next, comments from the Winning Arguments dataset (Figure 1.A) are passed to LLMs, which extract feature ratings based on the features derived from the Truth Wins dataset (Figure 1.D). To estimate the likelihood of belief change, we train a separate belief update regression model on annotated belief shift data from Truth Wins (Figure 1.C).

Together, the feature ratings generated by the LLMs and the belief update scores predicted by the regression model were input into a Random Forest classifier for binary classification (Figure 1.E). Performance of the Random Forest classifier is then measured using the test set of Winning Arguments. Using the Random Forest model's mechanism for evaluating feature importance enables the removal of noise or negatively contributing features, further refining the final model (Figure 1.F). However, while this refinement did little to improve accuracy, it did allow us to simplify the model and clarify the ranking of features.

3.1 Datasets

Winning Arguments Dataset (Figure 1.A) This is the primary dataset for this study, sourced from the Change My View subreddit where users invite others to challenge their opinions by presenting counterarguments (Tan et al. 2016). Persuasion is explicitly marked by the OP awarding a delta symbol (Δ) to comments that successfully change their view, thus distinguishing between successful persuasion and unsuccessful comments.

We focus on root replies, the first-level responses from challengers to the OP. For each post, we extract one positive (successful) and one negative (unsuccessful) comment. Since posts often receive multiple challenges, we control for content similarity by selecting comment pairs with the highest Jaccard similarity scores, in an effort to ensure that comparisons are based more on persuasiveness rather than content differences. The dataset consists of 3,456 posts for training and 807 for testing, each including the OP's text, a positive reply, and a negative reply. This offers a robust and ecologically valid foundation for studying persuasive dynamics in online discourse.

Truth Wins Dataset (Figure 1.B) The Truth Wins dataset was collected from a series of human-subject experiments designed to study persuasion and attention-seeking in written messages across diverse topics (Fay et al. 2024). Participants were asked to generate messages either to persuade or to garner online attention. A second between-subjects manipulation varied whether participants were told to base their messages on true information, false information, or on any combination they wished.

Each message was rated by ten independent human annotators along multiple dimensions including: belief update (ranging from -100 to 100, reflecting the change between prior versus post belief score), attention, interesting, interesting-if-true, positive emotion, negative emotion, perceived effectiveness of influence, shareability, and truthfulness. Except for belief update, all ratings were recorded on a 5-point Likert scale: "not at all", "slightly", "somewhat", "very much" and "extremely". Following Fay et al. (2024), we assigned each item an integer value from 1 ("not at all") to 5 ("extremely"), and aggregated the ratings by averaging across raters to reduce individual variance while improving signal stability and generalizability. This aggregated structure aligns with our focus on identifying consistent predictors in successful persuasion.

These features are grounded in prior research. Epistemic emotions (interesting, interesting-if-true) promote deeper processing and belief change (Muis et al. 2018; Briñol and Petty 2012). Interestingness boosts engagement, though its effect on persuasion is context-dependent (Alwitt 2000; Murphy et al. 2005). Shareability captures perceived social attention, which supports deeper analysis and persuasive impact (Shteynberg et al. 2016; Shteynberg 2018). Emotional valence shapes attention and persuasion; positive emotions enhance message acceptance, while relevant negative emotions can also increase persuasion (Lerner et al. 2015; Petty and Briñol 2015; Huntsinger and Ray 2016). Truthfulness reflects perceived accuracy, which makes messages more persuasive and less likely to be dismissed. (Chaiken 1980; Eagly and Chaiken 1993; Kunda 1990; Petty and Cacioppo 1986; Tormala 2016).

While the original study explored the influence of truthfulness on persuasion and attention-seeking, our research examines broader factors affecting persuasiveness, acknowledging that attention-seeking messages can also carry persuasive elements. The controlled, human-labeled data from the Truth Wins can be used to help classify the more ecologically valid discourse found in the CMV-based Winning Arguments dataset, offering a comprehensive view of persuasive mechanisms across different contexts.

3.2 Baseline Models

To evaluate our proposed hybrid model, we compare it to three representative baselines: theory-driven inference, transparent surface-level text modeling, and black-box reasoning via large language models (LLMs). Together, these baselines form a spectrum of interpretability, linguistic abstraction, and data reliance.

Theory-Driven Model – Belief Update Regression (Figure 1.C) As a baseline, we modeled persuasion as belief change. Using the Truth Wins dataset, which includes human-annotated belief update ratings, we trained an ordinal least square (OLS) regression model to identify significant features predictive of belief update, using the other eight key features studied by Fay et al. (2024). Following a stepwise procedure, variables were added or removed based on p-value thresholds (0.05), with problematic features excluded to ensure model stability.

The resulting model was then applied to the Winning Arguments dataset to generate predicted belief update scores for each comment. These scores were then thresholded – tuned for accuracy on the training set – to assign binary persuasion labels. Though not optimized for classifying persuasive success, it offers a theory-informed perspective based on belief update scores. In our hybrid model, these belief update scores are also used as a feature, contributing to a psychologically grounded signal that complements language-based predictions. Note that belief update and successful persuasion are distinct constructs: not all persuasive responses result in explicit belief change, and not all belief changes are externally acknowledged.

Transparent Language Model — Logistic Regression on Term Frequencies This baseline uses a shallow, interpretable language model to classify persuasion based solely on surface-level lexical features. Each comment is tokenized and vectorized using term frequency (TF), omitting inverse document frequency (IDF) to better accommodate short-form text, where rare words may not carry more informational weight. We train a logistic regression with 10-fold cross-validation, tuned on regularization strength, and evaluated performance using ROC AUC. This model provides a transparent lexical benchmark, revealing how much signal can be captured from word-level frequency patterns alone.

Black-Box Language Model — Zero-Shot LLM Classification To assess the potential of large language models (LLMs), we evaluated three models: LLaMA3-70B (LLaMA3), Gemma2-9B (Gemma2), and Mixtral-8x7B (Mixtral) in a zero-shot persuasion task. Each model receives the OP's post and two responses: one that successfully persuaded the OP and one that did not. The models are then prompted to select the response they judge to be more persuasive based on the input. The full prompt is provided in Listing 1 in Appendix A. All models are run with temperature set to 0 and a fixed seed (42) to ensure high probability responses and reproducibility.

This approach demonstrates the scalability and adaptability of LLMs, which can be applied without fine-tuning or additional supervision. Although tools like Gemma Scope provide interpretability scaffolding, these models still operate with a high degree of opacity, and their predictions can be difficult to interpret or validate – especially in high-stakes contexts such as misinformation detection, where transparency and accountability are critical. To mitigate these limitations, we subsequently use the same LLMs to generate structured feature ratings. These outputs are integrated into a hybrid model to support more interpretable and feature-aware classification of persuasive language.

3.3 Hybrid Model

Feature Extraction Using LLMs (Figure 1.D) Using the same three LLMs used for the Black-Box model (LLaMA3, Gemma2, and Mixtral), we generated feature-by-feature ratings for comments in the Winning Arguments dataset. Given the original post along with a pair of comments (one labeled positive, one negative), each model was prompted to generate ratings for both comments in relation to the original post. The full prompt is provided in Listing 2 in Appendix A. The selected features rated by the LLMs were adapted from the Truth Wins dataset and are defined as follows:

1. *Influential*: How effectively the comment influences the reader's perspective.
2. *Interesting*: How engaging and attention-grabbing the comment is.
3. *Interesting-If-True*: Potential interest level if the comment's claims were true.
4. *Positive*: Overall positivity, focusing on an uplifting or encouraging tone.
5. *Negative*: Overall negativity, considering critical or discouraging aspects.
6. *Shareable*: Likelihood of the comment being shared.
7. *Truthfulness*: Apparent accuracy and credibility of the comment.
8. *Attention*: Ability to capture and maintain reader focus.

To avoid ambiguity, we replaced the original Fay et al. (2024) feature label *Persuasive* with *Influential*. This *Influential* feature captures the perceived effectiveness of a comment in influencing a reader's perspective, rather than directly measuring actual belief change. This distinction prevents confusion between the general notion of perceived "persuasiveness" as a feature and the task of predicting successful persuasion itself. Using the feature schema derived from the Truth Wins dataset, we obtained LLM-generated ratings for the same features in the Winning Arguments

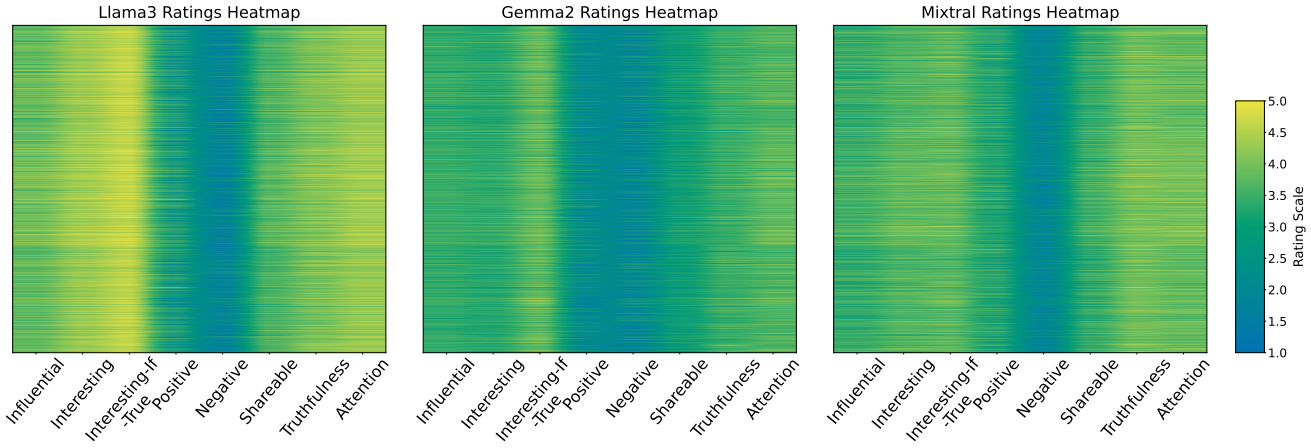


Figure 2: LLMs Generated Ratings Heatmap - This visualizes the ratings assigned by each LLM for each analysed feature (x-axis) for each argument in the Winning Arguments dataset (y-axis).

Model	Data-Driven	Theory-Driven	Hybrid Independent	Hybrid Interaction
Transparent (Logistic regression)	56.50%	-	-	-
Black-Box (LLaMA3)	64.93%	54.30%	73.17%	73.11%
Black-Box (Gemma2)	56.90%	63.80%	82.32%	82.69%
Black-Box (Mixtral)	61.71%	58.20%	72.55%	73.17%

Table 1: Accuracy comparison of models used for classifying successful persuasion (positive) vs. unsuccessful (negative) comments in the Winning Arguments dataset. **Transparent Data-Driven** refers to a logistic regression on term frequencies. **Black-Box Data-Driven** represents zero-shot classification from black-box LLMs. **Theory-Driven** uses thresholded belief update scores derived from regression models. **Hybrid Independent** and **Hybrid Interaction** refer to Random Forest models that combine features with belief update scores, using either independent terms only or including pairwise feature interactions, respectively.

dataset, providing a theory-driven input layer for persuasion classification. This conceptualization aligns closely with the probabilistic framework described by Wyer Jr. and Goldberg (1970), which distinguishes between a message's influence on a premise and the subsequent belief update. This approach enabled a structured analysis of persuasive strategies in real-world discourse, grounded in a controlled, theory-informed framework.

Independent vs. Interaction Terms To assess the role of feature interactions in persuasion prediction, we developed two hybrid model variants: one using only independent features, and another incorporating feature interactions. The independent-term model evaluates each feature's contribution individually. In contrast, the interaction-term model includes pairwise combinations of features alongside their independent counterparts. For simplicity, interactions are limited to two-feature pairs. This approach enables us to determine whether interactions between features significantly enhance predictive performance or whether individual features alone are sufficient for effective persuasion classification.

Random Forest Classification (Figure 1.E) We use a Random Forest (RF) classifier (Breiman 2001) for its strong predictive performance combined with the interpretability of decision trees. Our hybrid RF models are trained on LLM-generated features and belief update scores derived from regression models. The hybrid models are then evaluated on the test set. At this stage, the OP's post content is not taken into consideration. The RF-independent model combines LLM-generated features with belief update scores derived from the independent-term regression, while the RF-interaction model extends this by incorporating pairwise interaction features and scores from the interaction-term regression. Using RF's built-in permutation-based Variable Importance (VI), we assess which characteristics extracted by LLMs most meaningfully predict persuasive success, driving model performance.

VI assesses each feature by measuring decrease in classification accuracy when its values are randomly permuted, directly quantifies each feature's predictive contribution. Throughout this study, unless otherwise specified, "feature importance" specifically refers to VI. Using VI, we can remove noisy or redundant features, streamline the model, and mitigate potential overfitting without sacrificing accuracy. Although VI can be biased toward high-cardinality features (Strobl et al. 2007), our dataset consists exclusively of ordinal categorical features (rated 1 to 5) and continuous belief update scores, mitigating this bias. Our random forest classifier uses 300 trees, balancing complexity and generalizability. A fixed random seed (42) ensures reproducibility.

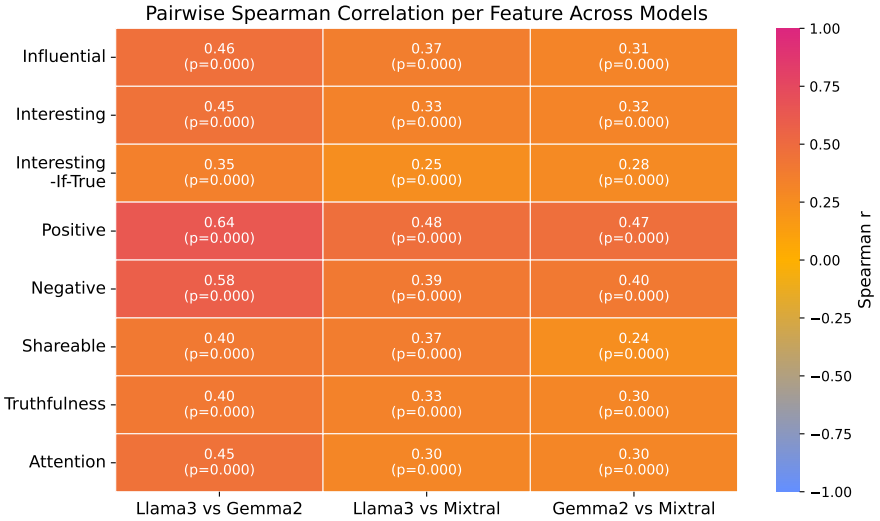


Figure 3: Spearman Rank Correlation between LLMs' generated ratings. Each cell represents the correlation between the ratings generated by the given pair of LLMs (x-axis) for a given feature (y-axis).

4 Results

4.1 Cross-Model Agreement in Generated Feature Ratings

To examine how different language models interpret identical instructions for feature ratings, we compared ratings generated by LLaMA3, Gemma2, and Mixtral on the full Winning Arguments dataset. Each model independently rated the following features: *Influential*, *Interesting*, *Interesting-If-True*, *Positive*, *Negative*, *Truthfulness*, *Attention*, and *Shareable*.

Heatmaps (Figure 2) reveal systematic rating differences across models. LLaMA3 consistently produced higher ratings for features such as *Influential*, *Interesting*, *Interesting-If-True*, and *Attention*, suggesting a relatively generous interpretation of these criteria. In contrast, Gemma2 and Mixtral tended toward lower ratings overall, with Gemma2 notably rating *Positive* lower and giving higher ratings for *Interesting-If-True* and *Attention*. Mixtral's ratings typically fell between those of LLaMA3 and Gemma2.

A pairwise correlation analysis (Figure 3) demonstrated consistent positive and statistically significant (p-value $<$ 0.001) monotonic correlations between models, despite differences in absolute ratings. LLaMA3 and Gemma2 consistently showed the strongest correlations across features, with values reaching 0.64 for *Positive* and 0.58 for *Negative*, highlighting close agreement in valence judgments. In contrast, strong predictive features like *Interesting-If-True* showed weaker correlations across all pairs, pointing to greater ambiguity or model-specific interpretation.

These differences in both absolute scores and correlation structure help explain variation in downstream performance across models. At the same time, the consistent alignment in rating trends between LLMs, as shown in Figure 3 through statistically significant positive monotonic correlations, indicates a degree of consistency across models. This observed alignment suggests that the LLM-generated features may be suitable for use in downstream modeling, despite variation in individual rating scales.

4.2 Model Accuracy Performance

We evaluated model performance for classifying successful persuasion (positive) vs. unsuccessful (negative) comments in the Winning Arguments dataset, summarized in Table 1.

Hybrid classifiers integrating LLM-generated features outperformed both the theory-driven model, and transparent and black-box language model baselines. The Gemma2-based hybrid model achieved the highest accuracy, roughly 82% for both independent-term and interaction-term variants, approximately a 25% improvement over its black-box baseline. Hybrid models built on LLaMA3 and Mixtral followed closely, with accuracies around 73% and similarly small differences between the two feature configurations.

Interestingly, including feature interactions and removing negatively contributing features had little impact on predictive accuracy across all hybrid models. This indicates that independent features, as generated by LLMs, already sufficiently capture the predictive signals necessary for effective persuasion classification.

4.3 Feature Importance (Figure 1.F)

Figure 4 summarizes the most important model features. Despite differences amongst LLMs, all consistently identify a core set of important features: *Influential*, *Shareable*, *Interesting/Interesting-If-True*, and *Attention*. Specifically, in the independent-term models, LLaMA3 emphasized *Interesting-If-True*, Gemma2 prioritized *Attention*, and Mixtral strongly highlighted *Influential*. Interaction-term models revealed that combinations involving *Interesting-If-True*, *Interesting*, and *Influential* substantially enhanced predictive performance, indicating potentially informative in-

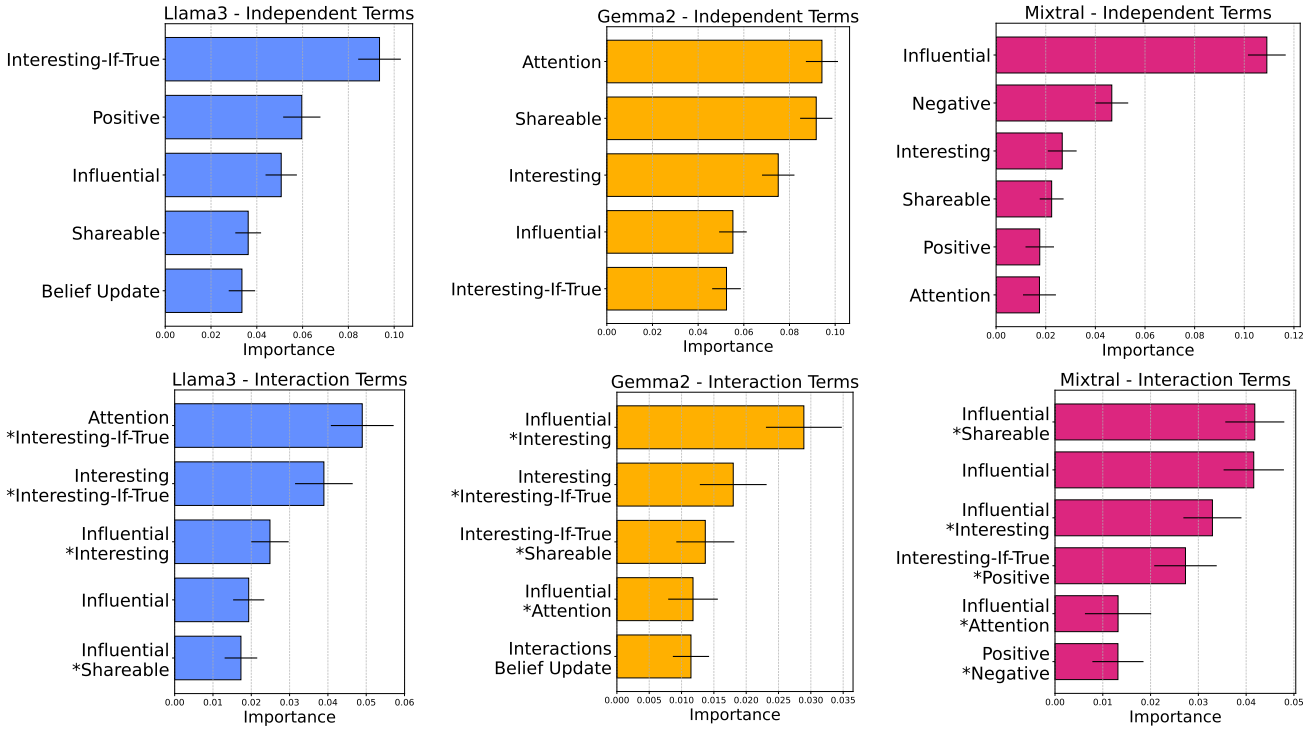


Figure 4: Top 5 Most Important Features in the Hybrid Model (Random Forest), with Independent Terms shown on the top row and Interaction Terms on the bottom row. Each bar represents the mean feature importance across 100 permutations, and the error line indicates the standard deviation. For the Mixtral-based models, six features are shown due to a tie in importance scores for the fifth position.

teractions among these features. Additionally, *Belief Update* was influential in Gemma2, suggesting its relevance in predicting persuasive success. Emotional features (*Positive* and *Negative*) showed inconsistent predictive power, warranting further investigation. Notably, *Truthfulness* consistently ranked among the least important features, which may reflect limited variation in the accuracy of information within CMV posts rather than a lack of relevance to persuasion.

5 Discussion

Overall, our findings show that hybrid models outperform both theory-driven and data-driven baselines, suggesting that combining LLMs with structured, theory-informed features improves the prediction of persuasive success. Notably, LLMs perform best when used to generate interpretable ratings for individual features, highlighting the value of integrating statistical language models with domain knowledge to better model persuasive success.

5.1 Theory-Driven Model

The belief update prediction model, trained on the Truth Wins dataset and applied to LLM-generated feature ratings from the Winning Arguments dataset, reveals critical insights into cross-domain generalizability. While the model offers a broad indication of belief change, its simplicity – using ordinal linear regression – fails to capture the complexities of successful persuasion. Incorporating the belief update score into the final model adds some interpretability but does not significantly improve performance. This suggests that belief update scores, while useful, are not sufficient to fully capture the dynamics of persuasive impact.

What we learn here is that, although theory-driven models can provide valuable insights, they struggle to translate across domains and fail to account for the nuanced, dynamic nature of persuasion. This underscores the need for more sophisticated, hybrid models that combine theoretical insights with the strengths of machine learning to better capture the complexities of persuasion and improve generalizability across diverse data sources.

5.2 Transparent Data-Driven Model

The logistic regression model based on term frequency (TF) highlights the limitations of using word frequency alone for persuasion detection. TF models, particularly for short-form text like CMV comments, struggle to capture linguistic and contextual elements such as sentiment and interaction dynamics. This emphasizes the need for more theory-informed features and predictors that connect specific linguistic cues to successful persuasion – an area where the hybrid model, integrating both theory-driven insights and data-driven techniques, can offer a more comprehensive solution.

5.3 Black-Box Data-Driven Model - Leveraging LLMs

Zero-shot classifiers, based on LLMs, demonstrated an improvement over logistic regression, though directly prompting LLMs to select the more persuasive comment proved less effective. LLMs excel at converting linguistic features into ratings that capture statistical regularities in human communication. This ability to represent language numerically enhances the performance of downstream models, particularly when structured data is needed. Although some models performed better than others in zero-shot classification, the real value of LLMs lies in their capacity to encode linguistic features into structured data. This underscores the potential of LLMs to aid in tasks requiring large-scale, feature-rich representations.

Research highlights the potential of LLMs as a powerful tool for persuasion analysis and belief influence. Studies have shown that LLMs, such as GPT-3.5, align closely with human judgments on persuasiveness, with high correlation between human and LLM-generated ratings (Fay et al. 2024). In fact, LLMs have been found to outperform humans in assessing persuasiveness across various topics and demographics (Salvi et al. 2024) and can effectively influence opinions on both polarized and less polarized issues (Bai et al. 2023). Additionally, Costello, Pennycook, and Rand (2024) demonstrated that conversations with GPT-4 Turbo could reduce belief in conspiracy theories by roughly 20%, with effects lasting up to two months, showcasing LLMs' potential in belief change.

Given these capabilities, future research should explore comparisons between LLM-generated and human ratings for accuracy and consistency. Hybrid approaches could further enhance performance in complex tasks like influence campaign detection and misinformation analysis, where a deeper understanding of persuasive content and context is crucial.

5.4 Cognitive Features and Biases in Persuasion

The top predictive features identified across all three models were *Interesting-If-True*, *Influential*, and *Shareable*. In the context of the CMV forum, where the primary goal is to change the OP's viewpoint, these features become particularly relevant. Unlike everyday conversations, which may not always prioritize persuasion, the CMV forum encourages high-quality, persuasive arguments. As a result, less obvious features like curiosity (*Interesting*-If-True) and social sharing (*Shareable*) become crucial factors in shifting opinions.

In the Hybrid Independent models, *Interesting-If-True* ranks highest for LLaMA3, while *Attention* holds the top position for Gemma2. Both features also rank among the top five for Mixtral. These features are closely tied to epistemic emotions, which stimulate cognitive engagement (Muis et al. 2018). Curiosity fosters exploration and information seeking, encouraging individuals to gather evidence and critically evaluate messages to resolve their curiosity (Vogl et al. 2019). In the context of persuasion, a message that sparks curiosity – such as "That's interesting, is it true?" – may prompt the audience to scrutinize it more rigorously, potentially verifying its credibility before accepting it.

Similarly, interest sustains attention and systematic processing. As part of the broader category of epistemic emotions, interest – like curiosity and surprise – has been shown to enhance message processing and facilitate persuasion by prompting deeper cognitive engagement (Muis et al. 2018; Briñol and Petty 2012). These emotions are also associated with increased message sharing and heightened perceptions of credibility, even when the underlying content lacks truthfulness (Rijo and Waldzus 2023). This may explain why *Shareable* is influential across all three models, alongside *Interesting-If-True*. Both enhance emotional engagement, mediating how information is perceived, evaluated, and disseminated online.

Truthfulness was a weaker predictor of persuasive success in CMV. Perceived truthfulness might be more important in environments with a wider mix of objectively true and false information. Research by Rijo and Waldzus (2023) suggests that fake news often elicits stronger epistemic emotional responses than true news, making it appear more credible and leading to wider dissemination. These findings highlight the dominance of perceived truth over objective truth in shaping belief formation and information spread. Moreover, regarding misinformation, Pennycook and Rand (2022) highlighted that while people can typically distinguish between true and false information when asked, this analytical distinction fades in informal contexts like social media. In these environments, social and affective factors – rather than truth-based analysis – primarily drive sharing behavior.

Given this focus on persuasion rather than truthfulness, future research could further explore the role and drivers of perceived truth quality in the Winning Arguments dataset to assess whether it plays a significant role in persuasion. For example, Fazio et al. (2015) demonstrated the illusory truth effect in news classification: repetition of a claim (even if false) can increase its perceived truth and likelihood of being shared. Investigating whether similar effects apply in the CMV context could offer further insights into how persuasion works, regardless of objective truth.

Emotion was also a significant feature, with *Positive* emotions ranking highly for LLaMA3 and *Negative* emotions for Mixtral. This aligns with research on how emotions influence persuasion. According to Naranowicz (2022), emotional state affects critical engagement: negative moods promote scrutiny of argument quality, while positive moods encourage reliance on credibility cues and general trust. Furthermore, Paletz et al. (2023) and Fay et al. (2024) found that positive emotions drive content sharing, reinforcing the role of emotional engagement in persuasion. These findings highlight the importance of emotional features in our model.

While our analysis captures emotional valence, it overlooks arousal, which plays a crucial role in persuasion. High-arousal emotions – whether positive or negative – amplify persuasive impact by encouraging fast, intuitive processing (Lühring et al. 2024), leading to snap judgments and reduced source verification. Since arousal is absent from our feature set, future research should examine its role in persuasion. Incorporating arousal-related features could provide deeper insights into decision-making, credibility assessments, and

content sharing, offering a more comprehensive understanding of emotional engagement in persuasion.

6 Conclusion

This study demonstrates the potential of large language models (LLMs) as effective tools for predicting successful persuasion in online discourse. By generating feature ratings on dimensions such as influence effectiveness, interest, and shareability, LLMs provide a scalable method for analyzing persuasive language, bridging the gap between manual annotation and automated analysis.

Beyond their practical applications, the findings support theoretical perspectives on persuasion. Feature importance analysis in the Random Forest model revealed significant patterns aligning with existing literature, reinforcing the role of linguistic and cognitive factors in persuasive content. This supports the validity of the feature set and highlights its relevance in computational persuasion research.

Trained on high-quality arguments from Change My View subreddit, the model represents a step forward in both theoretical and applied research. Automating the rating process for belief prediction offers valuable insights into opinion shifts in digital spaces. By integrating theory-driven insights with data-driven modeling, this work advances both our understanding of persuasive mechanisms and real-world applications, including influence campaign detection, misinformation mitigation, and content moderation.

Limitations

The Winning Arguments dataset, sourced from a forum focused on persuasion, may limit the generalizability of our findings. Since users on CMV are actively trying to change others' opinions, the dynamics may differ from more organic or varied settings. Future research should apply the model to other datasets where persuasive intent is less explicit, such as general social media platforms, news comments, or product reviews, to evaluate whether the model performs well across different online environments.

Another key limitation of this study is the potential bias introduced by relying on LLM-generated ratings. While LLMs offer a scalable way to assess textual dimensions, their consistency and accuracy relative to human annotations remain uncertain. Prior work has shown that LLMs can exhibit preference for specific token patterns (Zheng et al. 2024), produce less consistent ratings than human annotators (Stureborg, Alikaniotis, and Suhara 2024), and reflect cognitive biases learned from training data (Dillion et al. 2023). These issues can affect tasks that require nuanced or objective judgment. Although LLMs may align with human judgments in broad patterns, their outputs can still be biased or unstable, potentially impacting downstream modeling performance.

A related concern is the lack of validation against human-sourced ratings. The hybrid models in this study rely primarily on LLM-generated ratings for the classification task, but the absence of human annotations limits our ability to assess the reliability of those inputs. Future work could compare LLM-generated ratings with human-annotated counterparts for the same features. This would provide a form of sanity check, establish a ground truth for model evaluation, and help identify systematic rating discrepancies across models.

Relying on LLM-generated feature ratings may overlook important linguistic nuances such as grammar, sentence structure, and subtle rhetorical or emotional cues, which are critical for understanding persuasive success. Future work could incorporate more advanced natural language processing techniques, integrating syntactic and semantic analysis alongside LLM-generated features to capture a deeper, richer understanding of persuasion.

LLMs often function as black boxes, making their internal decision-making opaque. This creates challenges in understanding how they determine which features of an argument are persuasive. Future research should explore interpretability techniques to better understand the patterns and factors that drive the model's outputs. These tools can provide insights into the patterns LLMs use to generate persuasive judgments, enhancing model transparency and usability.

Ethical Consideration

This study draws upon two previous datasets: the Winning Arguments dataset and the Truth Wins dataset. Both datasets were collected in prior research and are used here under the assumption that the original studies adhered to appropriate ethical standards for data collection. Readers are encouraged to refer to the respective publications for further details on data sourcing, consent, and anonymization procedures.

A portion of the data in this study is generated using Large Language Models (LLMs). While these tools offer efficiency and scale, they also raise several ethical concerns. Most notably, LLMs function as black boxes. Users have limited insight into the data used during training, which may contain biases, or misinformation. Furthermore, the proprietary nature of the LLMs employed (accessed via the Groq Cloud API) means they are not open-source, and their behavior or availability can change without notice. These models are costly to retrain and may become outdated over time. Additional risks include the potential for output degradation, loss of coherence, and generation of repetitive or harmful content. These limitations must be taken into account when interpreting results involving LLM-generated data.

The development of a persuasion prediction model presents a dual-use dilemma. On one hand, this research aims to enhance our understanding of persuasive communication, with the broader goal of fostering safer and more constructive online environments. By identifying linguistic and structural features associated with successful persuasion, we hope to contribute to the creation of more responsible and effective messaging strategies. On the other hand, we recognize that these insights could be misused—to manipulate public opinion, spread misinformation, or exploit vulnerable audiences. As such, we stress that the findings of this study must be applied with caution and responsibility. Ethical safeguards, transparency in deployment, and continuous monitoring are essential to mitigate potential misuse.

Acknowledgements

Lewis Mitchell acknowledges support from the Australian Government through the Australian Research Council's Discovery Projects funding scheme (project DP210103700). Nicolas Fay acknowledges support from the Office of National Intelligence and Australian Research Council (project NI210100224).

References

Alwitt, L. F. 2000. Effects of Interestingness on Evaluations of TV Commercials. *Journal of Current Issues & Research in Advertising*, 22(1): 41–53. Publisher: Routledge _eprint: https://doi.org/10.1080/10641734.2000.10505100.

Arechar, A. A.; Allen, J.; Berinsky, A. J.; Cole, R.; Epstein, Z.; Garimella, K.; Gully, A.; Lu, J. G.; Ross, R. M.; Stagnaro, M. N.; Zhang, Y.; Pennycook, G.; and Rand, D. G. 2023. Understanding and combatting misinformation across 16 countries on six continents. *Nature Human Behaviour*, 7(9): 1502–1513. Publisher: Nature Publishing Group.

Bai, M. H.; Voelkel, J. G.; Eichstaedt, j. C.; and Willer, R. 2023. Artificial Intelligence Can Persuade Humans on Political Issues.

Barberá, P. 2020. *Social Media, Echo Chambers, and Political Polarization*, 34–55. Cambridge University Press.

Breiman, L. 2001. Random Forests. *Machine Learning*, 45(1): 5–32.

Briñol, P.; and Petty, R. E. 2012. A history of attitudes and persuasion research. In *Handbook of the History of Social Psychology*. Psychology Press. ISBN 978-0-203-80849-8. Num Pages: 38.

Chaiken, S. 1980. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, 39(5): 752–766. Place: US Publisher: American Psychological Association.

Chaiken, S. 1989. Heuristic and Systematic. *Unintended Thought*, 212.

Cialdini, R. B. 1993. Influence: The psychology of persuasion. *HarperCollins*.

Costello, T. H.; Pennycook, G.; and Rand, D. G. 2024. Durably reducing conspiracy beliefs through dialogues with AI. *Science*, 385(6714): eadq1814. Publisher: American Association for the Advancement of Science.

Dillard, J. P.; and Pfau, M. 2002. *The persuasion handbook: Developments in theory and practice*. Sage Publications.

Dillion, D.; Tandon, N.; Gu, Y.; and Gray, K. 2023. Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27(7): 597–600.

Dutta, S.; Das, D.; and Chakraborty, T. 2020. Changing views: Persuasion modeling and argument extraction from online discussions. *Information Processing & Management*, 57(2): 102085.

Eagly, A. H.; and Chaiken, S. 1993. *The psychology of attitudes*. The psychology of attitudes. Orlando, FL, US: Harcourt Brace Jovanovich College Publishers. ISBN 978-0-15-500097-1. Pages: xxii, 794.

Fay, N.; Ransom, K.; Walker, B.; Howe, P.; Perfors, A.; and Kashima, Y. 2024. Truth Wins: True Information is More Persuasive and Shareable than Falsehoods.

Fazio, L. K.; Brashier, N. M.; Payne, B. K.; and Marsh, E. J. 2015. Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, 144(5): 993–1002. Place: US Publisher: American Psychological Association.

Fogg, B. J. 2008. Mass Interpersonal Persuasion: An Early View of a New Phenomenon. In Oinas-Kukkonen, H.; Hasle, P.; Harjumaa, M.; Segerståhl, K.; and Øhrstrøm, P., eds., *Persuasive Technology*, 23–34. Berlin, Heidelberg: Springer. ISBN 978-3-540-68504-3.

Huntsinger, J. R.; and Ray, C. 2016. A flexible influence of affective feelings on creative and analytic performance. *Emotion*, 16(6): 826–837. Place: US Publisher: American Psychological Association.

Hwang, Y. 2012. Social Diffusion of Campaign Effects: Campaign-Generated Interpersonal Communication as a Mediator of Antitobacco Campaign Effects. *Communication Research*, 39(1): 120–141. Publisher: SAGE Publications Inc.

Karki, B.; Abri, F.; Namin, A. S.; and Jones, K. S. 2022. Using Transformers for Identification of Persuasion Principles in Phishing Emails. In *2022 IEEE International Conference on Big Data (Big Data)*, 2841–2848.

Kunda, Z. 1990. The case for motivated reasoning. *Psychological Bulletin*, 108(3): 480–498. Place: US Publisher: American Psychological Association.

Lerner, J. S.; Li, Y.; Valdesolo, P.; and Kassam, K. S. 2015. Emotion and Decision Making. *Annual Review of Psychology*, 66(Volume 66, 2015): 799–823. Publisher: Annual Reviews.

Lukin, S. M.; Anand, P.; Walker, M.; and Whittaker, S. 2017. Argument Strength is in the Eye of the Beholder: Audience Effects in Persuasion. ArXiv:1708.09085 [cs].

Lühring, J.; Shetty, A.; Koschmieder, C.; Garcia, D.; Waldherr, A.; and Metzler, H. 2024. Emotions in misinformation studies: distinguishing affective state from emotional response and misinformation recognition from acceptance. *Cognitive Research: Principles and Implications*, 9(1): 82.

Muis, K. R.; , C., Marianne; ; and Singh, C. A. 2018. The Role of Epistemic Emotions in Personal Epistemology and Self-Regulated Learning. *Educational Psychologist*, 53(3): 165–184. Publisher: Routledge _eprint: https://doi.org/10.1080/00461520.2017.1421465.

Murphy, P. K.; Holleran, T. A.; Long, J. F.; and Zeruth, J. A. 2005. Examining the complex roles of motivation and text medium in the persuasion process. *Contemporary Educational Psychology*, 30(4): 418–438.

Naranowicz, M. 2022. Mood effects on semantic processes: Behavioural and electrophysiological evidence. *Frontiers in Psychology*, 13. Publisher: Frontiers.

Paletz, S. B.; Johns, M. A.; Murauskaite, E. E.; Golonka, E. M.; Pandža, N. B.; Rytting, C. A.; Buntain, C.; and Ellis, D. 2023. Emotional content and sharing on Facebook:

A theory cage match. *Science Advances*, 9(39): eade9231. Publisher: American Association for the Advancement of Science.

Panagopoulos, C. 2012. Campaign Context and Preference Dynamics in U.S. Presidential Elections. *Journal of Elections, Public Opinion and Parties*, 22(2): 123–137. Publisher: Routledge _eprint: https://doi.org/10.1080/17457289.2012.662232.

Pennycook, G.; and Rand, D. G. 2022. Nudging Social Media toward Accuracy. *The ANNALS of the American Academy of Political and Social Science*, 700(1): 152–164. Publisher: SAGE Publications Inc.

Petty, R. E.; and Briñol, P. 2015. Emotion and persuasion: Cognitive and meta-cognitive processes impact attitudes. *Cognition and Emotion*, 29(1): 1–26. Publisher: Routledge _eprint: https://doi.org/10.1080/02699931.2014.967183.

Petty, R. E.; and Cacioppo, J. T. 1986. The Elaboration Likelihood Model of Persuasion. In Berkowitz, L., ed., *Advances in Experimental Social Psychology*, volume 19, 123–205. Academic Press.

Petty, R. E.; and Cacioppo, J. T. 2012. *Communication and persuasion: Central and peripheral routes to attitude change*. Springer Science & Business Media.

Reardon, K. K. 1991. *Persuasion in practice*. Sage.

Rijo, A.; and Waldzus, S. 2023. That's interesting! The role of epistemic emotions and perceived credibility in the relation between prior beliefs and susceptibility to fake-news. *Computers in Human Behavior*, 141: 107619.

Salvi, F.; Ribeiro, M. H.; Gallotti, R.; and West, R. 2024. On the Conversational Persuasiveness of Large Language Models: A Randomized Controlled Trial. ArXiv:2403.14380 [cs].

Shteynberg, G. 2018. A collective perspective: shared attention and the mind. *Current Opinion in Psychology*, 23: 93–97.

Shteynberg, G.; Bramlett, J. M.; Fles, E. H.; and Cameron, J. 2016. The broadcast of shared attention and its impact on political persuasion. *Journal of Personality and Social Psychology*, 111(5): 665–673. Place: US Publisher: American Psychological Association.

Strobl, C.; Boulesteix, A.-L.; Zeileis, A.; and Hothorn, T. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1): 25.

Stureborg, R.; Alikaniotis, D.; and Suhara, Y. 2024. Large Language Models are Inconsistent and Biased Evaluators. ArXiv:2405.01724 [cs].

Tan, C.; Niculae, V.; Danescu-Niculescu-Mizil, C.; and Lee, L. 2016. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, 613–624. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-4143-1.

Tormala, Z. L. 2016. The role of certainty (and uncertainty) in attitudes and persuasion. *Current Opinion in Psychology*, 10: 6–11.

Vogl, E.; Pekrun, R.; Murayama, K.; Loderer, K.; and Schubert, S. 2019. Surprise, Curiosity, and Confusion Promote Knowledge Exploration: Evidence for Robust Effects of Epistemic Emotions. *Frontiers in Psychology*, 10. Publisher: Frontiers.

Wang, X.; Shi, W.; Kim, R.; Oh, Y.; Yang, S.; Zhang, J.; and Yu, Z. 2020. Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good. ArXiv:1906.06725 [cs].

Weber, D.; Falzon, L.; Mitchell, L.; and Nasim, M. 2022. Promoting and countering misinformation during Australia's 2019–2020 bushfires: a case study of polarisation. *Social Network Analysis and Mining*, 12(1): 64.

Wei, Z.; Liu, Y.; and Li, Y. 2016. Is This Post Persuasive? Ranking Argumentative Comments in Online Forum. In Erk, K.; and Smith, N. A., eds., *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 195–200. Berlin, Germany: Association for Computational Linguistics.

Wyer Jr., R. S.; and Goldberg, L. 1970. A probabilistic analysis of the relationships among belief and attitudes. *Psychological Review*, 77(2): 100–120. Place: US Publisher: American Psychological Association.

Youk, S.; Malik, M.; Chen, Y.; Hopp, F. R.; and Weber, R. 2024. Measures of Argument Strength: A Computational, Large-Scale Analysis of Effective Persuasion in Real-World Debates. *Communication Methods and Measures*, 18(1): 7–29. Publisher: Routledge _eprint: https://doi.org/10.1080/19312458.2023.2230866.

Zhao, X.; Strasser, A.; Cappella, J. N.; Lerman, C.; and Fishbein, M. 2011. A Measure of Perceived Argument Strength: Reliability and Validity. *Communication Methods and Measures*, 5(1): 48–75. Publisher: Routledge _eprint: https://doi.org/10.1080/19312458.2010.547822.

Zheng, C.; Zhou, H.; Meng, F.; Zhou, J.; and Huang, M. 2024. Large Language Models Are Not Robust Multiple Choice Selectors. ArXiv:2309.03882 [cs].

A LLMs Prompts

Listing 1: Prompt used for zero-shot classification with LLMs.

```
<|begin_of_text|><|start_header_id|>
    system<|end_header_id|>
You are a classifier evaluating the
    persuasiveness of two comments on an
    original post.
Identify which comment is more
    persuasive.
Label the more persuasive comment as "
    positive" (1) and the less persuasive
     as "negative" (0).

Input includes:
- "op_text": original post text
- "comment1": first comment
- "comment2": second comment

```

```
Provide the result as a JSON object for each comment. No explanation is required.
<|eot_id|><|start_header_id|>user<|end_header_id|>
Here is the context and comments: {context}
<|eot_id|><|start_header_id|>assistant<|end_header_id|>
```

Listing 2: Prompt used for LLM-based feature extraction. In the original prompt, the label *Persuasive* was used; however, we refer to it as *Influential* throughout the paper to avoid ambiguity, as discussed in Section 3.3

```
begin_of_text|><|start_header_id|>system<|end_header_id|>
You are an expert evaluator tasked with rating two comments based on their context within an original post.
Each feature is rated from "one" (lowest) to "five" (highest).

Input includes: "op_text" (original post), "comment1" (first comment in response to op_text), and "comment2" (second comment in response to op_text).

Features to evaluate:
- Persuasive: How effectively the comment influences the reader's perspective.
- Interesting: How engaging and attention-grabbing the comment is.
- Interesting if True: Potential interest level if the comment's claims were true.
- Positive: Overall positivity, focusing on an uplifting or encouraging tone.
- Negative: Overall negativity, considering critical or discouraging aspects.
- Shareable: Likelihood of the comment being shared.
- Truthfulness: Apparent accuracy and credibility of the comment.
- Attention: Ability to capture and maintain reader focus.

Return the ratings for each feature for both comments as a JSON object with keys for each feature under "comment1" and "comment2" based on their relation to "op_text," using words ("one" to "five") for ratings.
Provide only the JSON object with no explanation.

<|eot_id|><|start_header_id|>user<|end_header_id>
Here is the context and comments: {context}
<|eot_id|><|start_header_id|>assistant<|end_header_id>
```

Explicit Cooperation Shapes Human-Like Multi-agent LLM Negotiation

Yanru Jiang[1, 2], **Gülşah Akçakır**[1]

[1]Department of Communication, University of California, Los Angeles
[2]Department of Statistics, University of California, Los Angeles
yanrujiang@g.ucla.edu, gakcakir@ucla.edu

Abstract

Humans develop cooperation heuristics in social decision-making, either intuitively or deliberatively. Large language models (LLMs), which exhibit human-like heuristics across cognitive domains, may acquire prosocial tendencies through instruction tuning, latently encoded in their representations to foster cooperative behavior in social reasoning games. However, most studies of this kind either focus on cooperative language generation or explicitly instruct LLMs to cooperate, deviating from the inherent cooperation heuristics of humans. Our negotiation role-play simulations using BATNA (Best Alternative to a Negotiated Agreement) with a GPT-based LLM reveal that LLMs may struggle with cooperation in the absence of explicit instructions, showing a 50–90% lower success rate than in instructed scenarios and a 40–80% lower success rate than human performance reported in past studies. Implicitly inducing cooperation through personality traits had inconsistent effects, with agreeableness showing only a marginal influence and other traits exhibiting no systematic impact. These findings suggest that personality-based cooperation cues are subtle and that explicit instructions may still be essential for multi-agent LLMs to approximate human-like negotiation.

Introduction

Humans develop cooperation heuristics in social decision-making, either intuitively or deliberatively (Rand et al. 2014; van den Berg, Dewitte, and Wenseleers 2021). Large language models (LLMs), which exhibit human-like biases across cognitive domains (Sreedhar and Chilton 2024; Tang and Kejriwal 2024), may acquire prosocial tendencies through instruction tuning (Gabriel 2020; Liu et al. 2023), enabling cooperative behavior in social reasoning games. However, previous studies have observed prosocial cues either in the linguistic style or survey answers of LLM-generated responses (Liu et al. 2023), or they explicitly instruct LLMs to cooperate (Abdelnabi et al. 2024; Wu et al. 2024), diverging from the inherent cooperation heuristics observed in humans.

To explore the intrinsic cooperative potential of LLMs in dynamic social decision-making, we utilize a multi-agent social simulation that involves unstructured, conversation-driven negotiation. Our design builds on Abdelnabi et al. (2024)'s system, a multi-agent LLM framework adapted from the HARBORCO role-play game, a commonly used negotiation training tool (Susskind 1985; Susskind and Corburn 2000). We ablate explicit cooperation instructions (i.e., overtly prosocial directives) from the original system by aligning instructional wording to the instructions provided to human players, which we consider the neutral condition.

Upon observing a systematic decline in negotiation success when switched from cooperative to neutral conditions, we explore whether implicitly inducing cooperation via personality traits can mitigate this effect, given that they mediate prosocial and cooperative tendencies in human behavior (Graziano et al. 2007; Koole et al. 2001; Buss 1991; Byrne, Silasi-Mansat, and Worthy 2015; Caprara et al. 2010), particularly in negotiation settings (Gilkey and Greenhalgh 1986; Ma 2008).

Together, this study investigates the cooperation heuristic of LLMs in social simulations without explicit instructions and whether implicit cooperation induction via personality mediates prosocial tendencies in LLMs.

Cooperation Heuristics in Social Decision-Making

Cooperation is fundamental to human societies (Rand et al. 2014; Powers, van Schaik, and Lehmann 2021). Prior research has investigated whether humans cooperate because they are intuitively predisposed to do so, or because they make deliberate decisions based on social values and interpersonal risk (Alós-Ferrer and Garagnani 2020). A dual-process cognitive framework integrates these perspectives (Rand et al. 2014), suggesting that cooperation heuristics emerge due to their general advantage in everyday interactions, becoming internalized as intuitive responses. On the other hand, deliberation regulates cooperation based on self-interest and expected utility, guided by reflective processes.

Although cooperation often generates positive social welfare in nonzero-sum contexts, it also entails costs to the individual (Roos Jr 1966). Since individuals perceive interpersonal risks and social discomfort differently, their willingness to cooperate also varies. Those with stronger prosocial predispositions, such as higher agreeableness (Graziano et al. 2007; Graziano and Eisenberg 1997; Caprara et al. 2010), tend to be more cooperative than those with weaker prosocial tendencies. In social dilemma settings, agreeableness positively correlates with cooperation, while extraver-

Proceedings of the First Workshop on Integrating NLP and Psychology to Study Social Interactions (NLPSI) @ICWSM '25, pages 57–71

sion negatively correlates, with cooperators often described as introverted and noncooperators as extraverted (Koole et al. 2001; Buss 1991; Graziano and Eisenberg 1997).

Given the critical role of individual differences in shaping social interaction patterns, including cooperation, an important question for researchers is whether these behavioral patterns can be replicated in LLM agents, especially if these models are to serve as a new platform for empirical research in social simulations (Bail 2024; Park et al. 2024).

Multi-agent LLMs for Social Simulation

Research on LLMs reveals their human-like capabilities in cognitive domains such as decision-making, reasoning, and creativity (Sreedhar and Chilton 2024; Tang and Kejriwal 2024). These models, typically instruction-tuned to align with human values, demonstrate prosocial tendencies in language (Gabriel 2020; Liu et al. 2023), suggesting that cooperative traits may be latently encoded within their representations. This aligns with the idea that LLMs could inherently exhibit cooperation, much like humans (Wu et al. 2024).

In multi-agent settings, the ability to approximate human behavior is crucial for success in tasks requiring social reasoning and decision-making. Given their extensive training on human language data, LLMs in multi-agent systems have been observed modeling human cognition, learning, and reasoning (Suri et al. 2023; Krishna et al. 2022; Andreas 2022; Park et al. 2023). If these models truly mirror human cognition, we can expect human-like social behavior to always emerge in their interactions within social contexts.

LLMs have demonstrated a high degree of alignment with prevailing social norms across various domains (Lu et al. 2024) in social science research. Chen et al. (2024) introduce adaptive governance mechanisms in LLMs by promoting prosocial behavior in reinforcement learning agents, leading to enhanced cooperation rates. LLMs' social heuristics have also been observed in psychological survey measures (Lu et al. 2024), as well as in fairness and framing effects studied in sociology (Suri et al. 2023). Additionally, studies examining LLM agents in social sandboxes suggest that they exhibit emergent social behaviors, such as hosting celebrations or sending invitations, without explicit prior instruction (Park et al. 2023). However, most existing research in this area primarily focuses on cooperation observed in LLMs' language generation or relies on explicit instructions for LLMs to cooperate (Li et al. 2023). These approaches diverge from the innate cooperative heuristics observed in humans, who can effortlessly achieve cooperation without external guidance, whether through intuitive or reflective responses (Rand et al. 2014).

Negotiation Games for Social Simulation

Multi-party negotiation simulations provide an effective scenario for evaluating the cognitive capabilities and decision-making of LLMs in social contexts. Negotiations require both social awareness and multi-turn strategic planning, where agents engage in verbal deliberation to persuade others, creating individual stochasticity in solution spaces while leveraging the unique language modality.

Previous studies on spontaneous cooperation in multi-agent LLM competitive games have mainly focused on two-party optimization problems, such as variations of Nash equilibrium games (Wu et al. 2024), or buyer-seller distributive (zero-sum) games (Huang and Hadfi 2024). These studies are more likely to assess LLMs' mathematical optimization and logical bargaining capabilities, rather than their intrinsic cooperative heuristics in complex, dynamic multi-party trade-offs. In contrast, we extend previous work by testing the cooperative tendencies of LLMs in a multi-party setting and pose the following question:

RQ1: Can LLMs intrinsically exhibit cooperation in multi-party social negotiation, as observed in humans, without explicit instructions directing them to cooperate?

Early work by social psychologists shows that negotiation is not merely a process of logical bargaining over resources, rather it is influenced by the personalities of those involved, shaping how negotiators perceive the situation, interact with counterparts, and predispose them to act in certain ways (Morris, Larrick, and Su 1999; Skandrani, Fessi, and Ladhari 2021). Therefore, those who can pick up on cues from other parties can tailor their strategies accordingly (Gilkey and Greenhalgh 1986). Among the Big Five personality traits, agreeableness, extraversion, and conscientiousness are most relevant to negotiation outcomes, with the first two influencing social interactions and the latter affecting bargaining strategy (Barry and Friedman 1998). The impact of these traits on varies depending on whether the negotiation is zero-sum. Agreeableness and extraversion tend to improve outcomes in integrative settings but become liabilities in distributive settings, while evidence for the effect of conscientiousness remains inconsistent across both contexts (Sharma, Bottom, and Elfenbein 2013).

Relying on the evidence from the literature, we introduce personality traits into our LLM-based agents to explore whether variations in prosociality and cooperativeness can emerge implicitly. This allows us to assess whether personality influences cooperative behavior, as observed in human subjects during negotiation simulations, even in the absence of explicit instructions.

RQ2: How do personality traits in system personas implicitly influence LLMs' performance in negotiation games, and do these injections align with human prosocial and cooperative tendencies?

Game Description

The Original Negotiation Game

The LLM negotiation system is adapted from the classic HARBORCO role-play negotiation exercise (Abdelnabi et al. 2024), originally designed as an instructional tool for teaching negotiation strategies (Susskind 1985; Susskind and Corburn 2000). The original game involves six players negotiating five policies to build a new deepwater port, each representing stakeholders such as the HARBORCO consortium ($p1$), environmentalists, union leaders, federal loan issuers, local government, and competing ports. Each party has confidential scores tied to policy preferences that often conflict, requiring players to balance their individual

scores with group objectives. A deal passes only if at least five players agree; otherwise, all players receive a minimal score equivalent to the utility of no deal, referred to as the Best Alternative to a Negotiated Agreement (BATNA) score. Scorable negotiation games with BATNA provide a quantifiable and systematic evaluation of both individual and group performance and trade-off in negotiation success.

Adaptation for Multi-agent Simulation

The LLM adaptation of HARBORCO reimagines the original game in different contexts (e.g., a sports park or an island nation's airport project). These variations are generated either through Bing Chat or manual rewriting, with individual policy scores adjusted for generalizability (Abdelnabi et al. 2024). We adopt the original language of the HARBORCO game as neutral guidelines for global and individual instructions.

The negotiation system contains the following:

- **System Prompts:** Each of the three Big Five personality traits (extraversion, agreeableness, and conscientiousness) is specified in the system to guide the LLM persona under personality-specific LLMs. For generic LLMs, no prompt is added. See the Experiment 1 section for details on persona tuning.
- **Initial Prompts:** Each agent receives a global instruction describing the negotiation context and individual confidential instructions outlining their specific policy preferences and the scores associated with each policy option.
- **Round Prompts:** During each round, agents access conversations from the last N turns (default N=6). They also have access to a scratchpad, which helps them reflect on prior interactions (optionally using a calculator to track scores) and propose their preferred policies based on their confidential scores.
- **Responses:** After deliberating privately in their scratchpad, each agent submits their proposed set of policies and uses multiple sentences to articulate their proposal and persuade other players.
- **End of Negotiation:** Once all rounds are completed (default 4×players), the project proposer (*p1*) finalizes the deal based on the policies proposed in prior rounds.

Experiments

We conducted three sequential experiments to examine: the contributions of each functional module (Experiment 1), the effect of cooperative instructions (Experiment 2), and the replication of results using two game variants (Experiment 3). All experiments were conducted using GPT-4o-mini to optimize for cost and inference speed.

All simulation performances are assessed by two metrics: the percentage of $AnySuccess$ (a 5- or 6-way agreement at any round) and $FinalSuccess$ (an agreement reached at the end of the negotiation). Agreement success serves as a reliable measure, as it reflects both group-level agreement and higher-than-individual BATNA (walk-away) scores compared to unsuccessful negotiations. Due to space limitations and the two metrics showing very similar patterns across different conditions, we report only $AnySuccess$ in our results and include $FinalSuccess$ in the Appendix III.

Experiment 1: Functional Modules

Experiment 1 first excludes the appended cooperative guidance from the original design (Abdelnabi et al. 2024) and tests key functional modules: calculator usage, scratchpad instructions, and system persona (see Appendix II for an example prompt). The calculator and scratchpad are selected as they are suspected to be crucial for agents in optimizing their scores while preventing hallucinations (e.g., incorrect score calculations or score leakage). Meanwhile, the system persona investigates the potential for implicit cooperation injections through personality traits. Each ablation of a functional module is termed Fully Cooperative (the original game), Round Prompt, Remove Calculator, and Remove Scratchpad, while the system persona serves as an empirical variation across different simulation settings.

Calculator A calculator can be turned on via "calculator prompt" in their scratchpad to track the score explicitly. Though one might expect that calculator could help agents avoid hallucinations and enhance their cognitive capacity, as well as mimicking how human participants would approach the problem in real negotiation scenarios, we perceive the calculator instruction forcing agents' thought process to be highly structured and rule-based rather than in linguistic deliberation, potentially reducing the linguistic diversity associated with different personality personas.

Scratchpad Guidance The scratchpad prompt provides explicit instructions on how to consider other agents' preferences and balance their own with those of others by writing them down on their scratchpad. It asks agents to engage in internal deliberation before sharing their answers publicly. When the scratchpad is removed, agents are still asked to consider others' preferences and balance their own scores with those of others, but they are no longer explicitly instructed to write down their thought process.

System Persona (Personality Injected into System Prompts) Our personality simulation adopted an approach similar to the Configurable General Multi-Agent Interaction (CGMI) framework (Shi et al. 2023), by approximating human participants' responses to the 60-item Big Five Inventory (BFI-2) questionnaire (Soto and John 2017; John and Srivastava 1999) in agents' system prompts. When an agent is prompted to exhibit a strong presence of a particular personality trait (e.g., high extroversion), a random set of agreement levels is generated for all items under the extroversion trait so that the average results in a final rating of either 6 or 7 on a 7-point Likert scale, reflecting the desired level on the selected personality trait. Similarly, for a low agreeableness agent, the average responses for agreeableness items are set to either 1 or 2. This approach leverages the granularity of BFI items, enabling a more realistic representation of intrinsic human attitudes associated with each personality (Shi et al. 2023; John and Srivastava 1999). The CGMI framework has been found to enhance personalization in agent

Linguistic Markers	No Personality Baseline (%)	Personality (% diff)	Extravert (% diff)	Agreeable (% diff)	Conscientious (% diff)
Social Processes	4.18	14.83***	2.31	4.60**	-0.15
Cognitive Processes	7.98	11.51***	-5.56**	-2.81	-3.28
Insight	2.22	14.46***	-7.31**	3.86	-4.60
Tentative	2.29	2.92	-6.97*	-9.73***	-9.23***
Discrepancy	1.86	21.07***	-7.18**	-3.92	-4.33*
Differentiation	2.18	-0.64	-5.46	-9.27***	-0.76
Drives	7.57	2.16	-0.61	0.96	0.17
Affiliation	3.18	5.38*	1.52	9.10***	2.17
Power	2.10	1.92	-7.46*	-6.38*	-0.78
Personal Pronouns	5.64	18.8***	-0.90	0.44	-2.20
1st Person Singular	0.52	40.29***	-3.02	-6.71**	-1.03
1st Person Plural	0.66	48.92***	17.66***	8.18	-0.34
3rd Person Plural	0.80	6.72	5.54	-12.96***	-11.40**
Informal Language	0.02	248.38***	-1.08	14.62	-12.61

Table 1: Comparison of LIWC categories across personality dimensions for the Remove Calculator condition. No Personality is the baseline average (n=300), while Personality (n=300) shows the percentage difference from the baseline. Extravert, Agreeable, and Conscientious reflect within-trait differences (High n=50 / Low n=50). Significance levels for independent t-tests: $^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$.

communication compared to relying solely on generic roles (Shi et al. 2023). The exact system prompt used is provided in Appendix I.

Validating Personality Injections. Our personality injections are validated using the Linguistic Inquiry and Word Count (LIWC) dictionary, a text analysis tool that quantifies linguistic patterns and psychological constructs (Pennebaker 2001; Tausczik and Pennebaker 2010), to assess whether incorporating personality meaningfully alters agents' linguistic style (i.e., whether our personality injection was effective).

LIWC categories have been shown to correlate both with self-reported and observer-reported personality assessments (Yarkoni 2010; Koutsoumpis et al. 2022), making them reliable linguistic markers for personality traits. Comparing No Personality (n = 50) and aggregated Personality (n = 300) in Table 1 across all selected traits (high and low extraversion, agreeableness, and conscientiousness), we observe significant linguistic differences between generic agents and those with personality personas. These differences are particularly evident in language related to social processes (+14.83%, $p < .001$), cognitive processes (+11.51%, $p < .001$), personal pronouns (+18.8%, $p < .001$), and informal languages (+248.38%, $p < .001$), with aggregated Personality consistently exhibiting more persona-driven language. In contrast, the generic agent follows procedural instructions without exhibiting pounced psychological linguistic styles.

Focusing on within-trait differences (e.g., low vs. high extraversion, both n = 50), agents with distinct personalities exhibit linguistic variations that align with the directional effects of LIWC markers observed in prior studies (Yarkoni 2010; Koutsoumpis et al. 2022). For example, high extraversion correlates with an increased use of first-person plural pronouns (+17.66%, $p < .001$). Similarly, high agreeableness is associated with a decrease in first-person singular pronouns (-6.71%, $p < .01$) and an increase in social process words (+4.60%, $p < .01$). High conscientiousness corresponds to reduced tentative language (-9.23%, $p < .001$).

Additionally, we explored other LIWC markers that, while not previously tested for personality associations, still align with our expected trait-driven linguistic patterns. For instance, extraversion and agreeableness exhibit reductions in discrepancy markers (-7.18% for extraversion, $p < .01$) and differentiation markers (-9.27% for agreeableness, $p < .001$), illustrating a preference for affiliation (+9.10% for agreeableness, $p < .001$) over power (-7.46% for extraversion, $p < .05$, -6.38% for agreeableness, $p < .05$) as a driving force in language use during negotiations. It is important to note that in our validation, LIWC categories highlight linguistic differences between LLM personas but do not indicate fundamental changes in the model's core reasoning and behaviors.

Experiment 1 Results Comparing the Fully Cooperative, Round Prompt, Remove Calculator, and Remove Scratchpad models under the Cooperative condition in Table 2, we observe a general trend: cooperative instruction among multi-agent systems tends to improve performance, both in terms of any and final successful agreement (see Appendix III). The Fully Cooperative condition achieves the highest success rate (96–100% $AnySuccess$), outperforming all other conditions, though only marginally better than the Round Prompt baseline (84–96% $AnySuccess$).

Surprisingly, removing the calculator does not reduce performance (94–100% $AnySuccess$) and may even slightly improve negotiation success. We speculate that the calculator might make LLM-based agents more conscious of their self-interest in policy proposals, thereby increasing the likelihood of disrupting group-level agreements.

Finally, removing explicit cooperative instruction in the scratchpad, which encourages agents to "write down" their cooperative strategies in each round, and replacing it with a purely informational suggestion on balancing self-interest

Modules	Cooperative				Neutral	
	Fully Cooperative	Round Prompt	Remove Calculator	Remove Scratchpad	Round Prompt	Remove Calculator
No Personality	98	96	100	**74**	12	22
All Personalities	98	92	95	**41**	9	20
Low Extraversion	96	84	94	36	**14**	20
High Extraversion	98	94	96	29	**0**	16
Low Agreeableness	98	92	**90**	32	4	**8**
High Agreeableness	96	94	**100**	38	8	**28**
Low Conscientiousness	100	92	92	50	16	20
High Conscientiousness	98	96	96	58	10	28

Table 2: Comparison of functional modules for *Any Success* rate across personas for Cooperative and Neutral instructions. *No Personality* (n = 50), *All Personalities* aggregated across selected traits (n = 300), and each low/high trait (n = 50). All four Cooperative conditions correspond to Experiment 1, while comparisons between Cooperative and Neutral for Round Prompt and Remove Calculator correspond to Experiment 2. Bolded personality pairs (No/All or Low/High) indicate significant differences under Fisher's exact test.

with others' interests results in the largest decline in negotiation success (29–74% $AnySuccess$).

In comparison, the reported success rate of human players in the original HARBORCO game is around 70–80% (PON 2023), indicating that when explicit cooperative instructions are provided in the scratchpad, they help agents achieve negotiation success higher than human players, who rely on cooperative heuristics without explicit guidance. However, once the explicit instructions are reworded as an informational statement, performance drops well below that of human counterparts.

These trends are consistent across all persona conditions, including the No Personality baseline, aggregated personality across all traits, and personality-specific personas, further reinforcing the importance of cooperative scratchpad instructions in enhancing negotiation performance.

Experiment 2: Cooperative Instructions

Given the significant impact of scratchpad instructions on agents' performance, along with the directional influence of cooperative instructions on negotiation success, we introduce a neutral scratchpad as a revised version of the original cooperative scratchpad. This neutral scratchpad is designed using the instructional wording from the original HARBORCO materials, allowing us to directly compare the effects of neutral versus cooperative framing on agent behavior. The scratchpad guides agent strategies through the following conditions (see examples in Appendix II):

- *Prefer Others (Cooperative)*: Prioritize others' preferences when making moves and proposing deal sets.
- *Consider Both (Neutral)*: Instruct agents to remain neutral by considering both their own and others' preferences.

We acknowledge that some may perceive our neutral scratchpad as already cooperative and the original cooperative scratchpad as overly prosocial, perhaps even entirely selfless. To clarify, in our simulation, the neutral scratchpad reflects realistic human instructions for playing negotiation games, where players naturally balance their own interests with those of others (*Consider Both*). The cooperative scratchpad goes beyond inherent human strategic calculations, explicitly prioritizing others' preferences over self-interest (*Prefer Others*).

Experiment 2 compares the Round Prompt and Remove Calculator conditions for both the *Prefer Others* and *Consider Both* scratchpad instructions to demonstrate the consistent decline in negotiation success when agents adopt the neutral instruction.

Experiment 2 Results Considering both Round Prompt and Remove Calculator, we observe a drop in $AnySuccess$ when shifting from the cooperative to the neutral scratchpad, decreasing from 76–94% for Round Prompt and 68–82% for Remove Calculator, with performance approximately 50–80% lower than that of human players. These results highlight the striking impact of removing explicit cooperation on LLMs' negotiation success in our simulations. These patterns remain consistent across all personas.

Examining differences within each personality trait, we do not observe systematic performance variations between its higher and lower levels. However, agreeableness occasionally shows significant variation under both Remove Calculator conditions (the modular conditions with the least syntactic structure in reasoning and potentially greater personality influence), with higher agreeableness achieving a 10–20% higher success rate than lower agreeableness. This suggests that, among all personality traits, agreeableness may have the greatest potential for implicitly fostering cooperativeness, as its psychological characteristics align closely with prosocial and cooperation heuristics identified in prior research (Alós-Ferrer and Garagnani 2020). However, the magnitude of improvement from agreeableness remains minimal compared to the contribution of explicit cooperation.

Scoring Trajectories. Visually examining the scoring trajectory of the high success rate from the cooperative scratch-

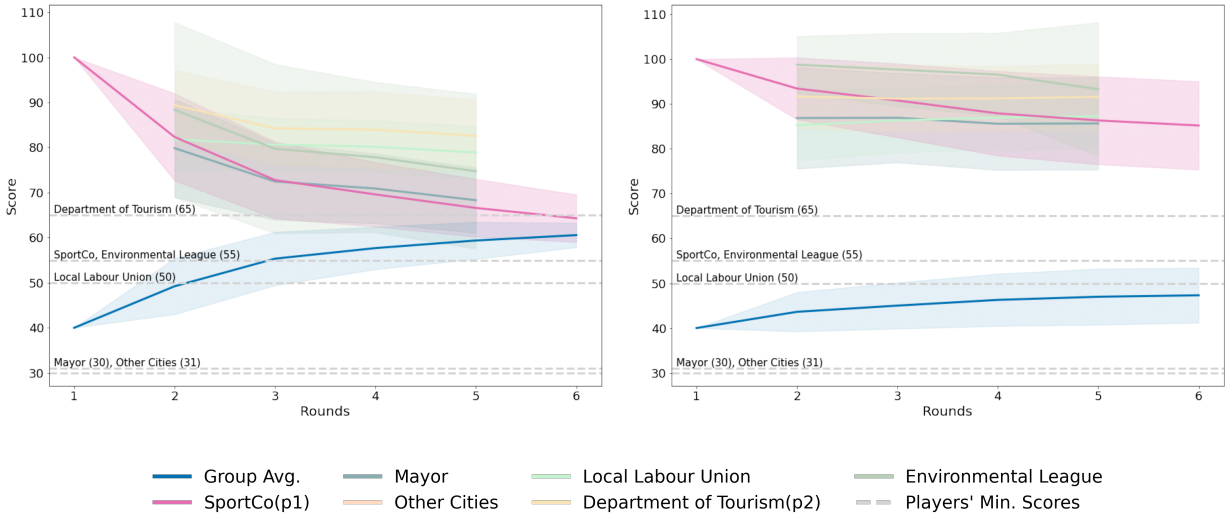


Figure 1: Scoring trajectory comparison between Cooperative (left) and Neutral (right) instructions for all players in Round Prompt of the base game across all personalities (both $n = 300$). Each player's BATNA score is illustrated.

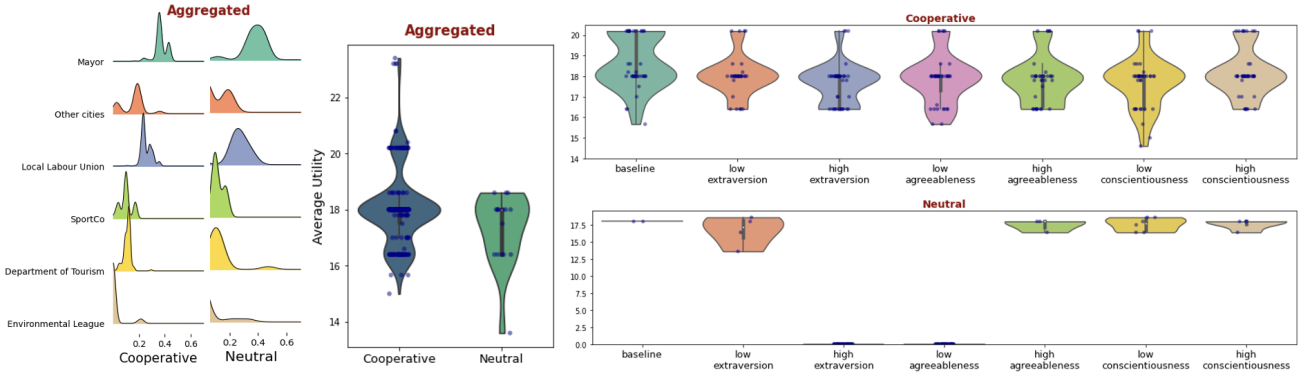


Figure 2: Utility and equality measures of Round Prompt. Left: Utility surplus share distribution by players across all personality simulation runs ($n = 300$). Middle: Average utility surplus across all players in successful negotiations, aggregated across all personality conditions ($n = 300$). Right: Condition-wise breakdown of average utility surplus + No Personality baseline (each $n = 50$).

pad and the low success rate from the neutral scratchpad in Round Prompt (Figure 1), we observed a distinct pattern in negotiation dynamics. Under the cooperative condition, negotiation scores between individual players, especially the project proposer (*p1*), tend to converge through conversational rounds, ultimately leading to high success rates in both any-round and final-round agreements. In contrast, under the neutral condition, individuals persist in maintaining proposal policy sets that align with their own preferences without adequately considering others' interests or the collective benefit. As a result, policy sets struggle to converge over multiple turns, leading to low negotiation success and simultaneously low individual scores for all participants. Similar scoring trajectories were also observed under Remove Calculator in Appendix IV.

Utility and Equality Measures. We visualize utility surplus and player equality to compare cooperative and neutral scratchpad conditions of Round Prompt (similar patterns under Remove Calculator can be found in Appendix V). Utility surplus captures how beneficial the final agreement is beyond mere success, calculated as the average surplus (final utility minus BATNA) across all players participated in successful N-way negotiations. Equality is measured by each player's share of the total utility surplus.

Aggregated results (Figure 2, Middle) show that cooperative instructions yield higher average utility and more cohesive policy outcomes, whereas neutral instructions result in greater variability concentrated at the lower end of the surplus spectrum. These patterns persist across personality conditions (Figure 2, Right). Interestingly, under the neutral condition, the baseline often converges on a single policy set, whereas personality prompts introduce greater variance in outcomes—suggesting that while cooperative instructions are key to negotiation success, personality injections can en-

rich LLM behavior with more human-like variation. Finally, player surplus share distributions (Figure 2, Left) are centered around similar positions across both cooperative and neutral conditions, indicating comparable levels of negotiation equality. However, the neutral condition shows a wider spread, suggesting greater variability in each players' share across game runs.

Additional Qualitative Analysis. We conducted a qualitative comparison of one simulated negotiation run under the cooperative and neutral conditions, using the Remove Calculator version as it offered a less structured and organically produced transcript. Our analysis focused on the project proposer's dialogue (i.e., *SportCo*), examining three personality conditions (No Personality, High Agreeableness, and Low Agreeableness) to illustrate how negotiation strategies and tone subtly vary depending on the presence of explicit cooperative instructions and personality mediation. More comprehensive annotated dialogues can be found in Appendix VI.

Under cooperative instructions, negotiation strategies are typically accommodating and altruistic, where proposers demonstrate a concern for aligning with others' preferences. For example, under the No Personality condition, one agent reasoned:

> "I should consider a compromise that acknowledges some damage while still aiming for a higher loan." *[Cooperative, No Personality, Scratchpad]*

This pragmatic but inclusive orientation extends into the public proposal itself:

> "This proposal aims to address the diverse interests of all parties while ensuring that we can proceed with the Harbour Sport Park project in a way that is beneficial for everyone involved." *[Cooperative, No Personality, Answer]*

When High Agreeableness is introduced, the tone becomes even more warm and inclusive, often going beyond utilitarian compromise. One example from the answer reads:

> "I appreciate the collaborative spirit we've all shown in this negotiation." *[Cooperative, High Agreeableness, Answer]*

Despite its agreeable tone in the answer, the scratchpad reflects a strategic awareness of the proposer's own interests, suggesting a more balanced orientation:

> "I need to find a way to accommodate the preferences of the Mayor and Other Cities while still ensuring that I achieve a reasonable score for myself." *[Cooperative, High Agreeableness, Scratchpad]*

By contrast, when Low Agreeableness is paired with cooperative instructions, we observe a more strategically rational tone:

> "If I can assure them that their interests are being considered without compromising too much on my end . . . " *[Cooperative, Low Agreeableness, Scratchpad]*

Under the neutral condition, however, proposers tend to display more self-focused and pragmatic reasoning, particularly in the scratchpad. For example, one agent begins:

> "My preferences are clear: I prioritize a. . . " *[Neutral, No Personality, Scratchpad]*

While the public proposal softens slightly, it remains outcome-focused:

> "This proposal aims to create a balanced solution that . . . while ensuring the project's success." *[Neutral, No Personality, Answer]*

When High Agreeableness is layered onto the neutral instruction, proposers begin to demonstrate more strategic attention to others' preferences and express willingness to collaborate, though without the full affective warmth seen with cooperative instructions:

> "I will also express my willingness to work together to find a solution that addresses everyone's concerns while ensuring the project's viability." *[Neutral, High Agreeableness, Scratchpad]*

For Low Agreeableness condition, proposers tend to anchor in self-interest, offering selective compromises only when instrumental to their goals. One proposer notes:

> "Given the final session, I need to find a way to appeal to the other parties while still protecting my interests." *[Neutral, Low Agreeableness, Scratchpad]*

This is echoed in assertive policy positions such as:

> "For employment opportunities, I strongly advocate for unlimited union preference, which I believe will significantly benefit our local economy." *[Neutral, Low Agreeableness, Answer]*

Overall, the key driver of negotiation strategy appears to be whether explicit cooperative instructions are provided. These instructions reliably elicit more prosocial reasoning and outwardly collaborative behavior. In contrast, personality simulation plays a more subtle role, shaping the tone and interpersonal framing rather than altering the strategic core of the negotiation.

We observe that personality conditions, whether high or low in agreeableness, result in slightly less utilitarian and more socially aware framings compared to the No Personality condition, which leans heavily on logical trade-off reasoning with minimal affective expression.

Interestingly, players' public-facing proposals (answers) are consistently more collaborative and positively framed than their internal reasoning (scratchpads), revealing a compelling distinction between internal strategy formulation and outward presentation—even though both are ultimately outputs of language generation.

Experiment 3: Game Replications

The final experiment aimed to demonstrate the generalizability of our findings on cooperation instructions and personality through replication. We tested two negotiation strategies, cooperative *Prefer Others* and neutral *Consider Both*, under Remove Calculator (conditions with the least syntactic structure), evaluating them in two game variants: a seven-player extension of the base game (*7players*) and a human-rewritten version of the base game (*Rewritten*). Both scenarios featured individual scores that differed from the base game.

Games	7 Players		Rewritten	
	Cooperative	Neutral	Cooperative	Neutral
Low Extraversion	76	18	96	22
High Extraversion	74	12	92	28
Low Agreeableness	**68**	**6**	88	**14**
High Agreeableness	**86**	**26**	88	**38**
Low Conscientiousness	74	16	92	30
High Conscientiousness	70	22	92	40

Table 3: Comparison of $AnySuccess$ rate between the No Calculator version of Cooperative and Neutral conditions across all personality traits (n = 50) for the 7-Player and Rewritten game variants. Bolded Low/High trait pairs indicate significant differences under Fisher's exact test.

Experiment 3 Results According to Table 3, the largest impact on performance still stems from shifting from the cooperative to the neutral scratchpad, resulting in a 32–60% drop in $AnySuccess$ for the 7-player version and a 50–74% drop for the rewritten game. The absolute difference between the cooperative and neutral conditions varies across game versions, as differences in individual scores influence negotiation difficulty. However, the pronounced gap persists across all games and all personality conditions.

Similar to the base game simulations, we do not observe a systematic impact of personality traits on negotiation success. High agreeableness occasionally shows a statistically significant improvement over low agreeableness using Fisher's exact test. However, this improvement remains modest (16–24% increase in $AnySuccess$) compared to the larger contribution of explicit cooperation.

Discussion

Across simulations with varying conditions and modules, we first identify that calculators have minimal impact on performance, whereas structured scratchpad guidance (i.e., prompting LLMs to articulate their deliberation) substantially improves negotiation outcomes. To address **RQ1** (whether or not LLMs cooperate in multi-party negotiation without being explicitly instructed to do so), we compare two scratchpad strategies: an overtly prosocial instruction (*Prefer Others*) and a more neutral directive (*Consider Both*). We observe that shifting to the latter results in a marked decline in negotiation success. This trend persists across the base game and two replications with different contexts and scores, with success rates well below those of the cooperative condition (50-90%) and human players in the original HARBORCO game (40-80%). We note again that this human performance benchmark is based on extensive classroom simulations of HARBORCO reported in the past (PON 2023). As illustrated in Figure 1, these failures manifest in the LLMs' reluctance (or slow adaptation) to align individual policies toward group agreements, and they often fail to achieve individual scores above their BATNA—highlighting the limited intrinsic cooperation of LLMs in the absence of clear prosocial instructions.

Given the importance of cooperation in negotiation success within our LLM simulations, we next examine **RQ2** (how injected personality traits may implicitly induce cooperation). Drawing on established links between personality and prosociality (Graziano et al. 2007; Caprara et al. 2010), we investigate whether injecting human-like personality traits—known to correlate with human prosocial and cooperative tendencies (Graziano et al. 2007; Graziano and Eisenberg 1997; Caprara et al. 2010; Koole et al. 2001)—can implicitly induce cooperation. However, personality injections yield only modest effects: high agreeableness occasionally improves success, high extraversion influences outcomes in only one setup, and high conscientiousness shows no significant effect. Yet, no personality trait matches the effectiveness of explicit cooperative instructions in the scratchpad.

This peculiar phenomenon highlights a key distinction between human players and LLM agents, one not commonly observed in previous LLM studies (Wu et al. 2024; Chen et al. 2024; Tang and Kejriwal 2024): unlike humans, who naturally accommodate others either intuitively or through strategic adjustments, LLMs struggle to reason about collective benefit through compromise. This difficulty persists despite their instruction-tuning for prosocial alignment (Sreedhar and Chilton 2024; Lu et al. 2024), suggesting that prosocial linguistic cues do not necessarily translate into cooperative behavior in social settings.

Although our study primarily examines only one GPT-based LLM, GPT-4o-mini, it raises a broader concern: cooperation in language models should not be assumed based on their prosocial linguistic tendencies. One possible explanation is their literal interpretation of users' instructions, shaped by training and instruction-tuning for assistantship (Gobinath et al. 2024). In our setup, agents interpret *Prefer Others* as requiring to fully adopt others' preferences, likely due to their tuning for helpfulness and prioritizing others' needs. However, when instructed to *Consider Both*, they fail to balance self-interest with parties priorities, leading to a sharp decline in negotiation success.

As LLMs are integrated into autonomous systems and human-AI collaborations, they often fall short in contexts where they lack shared understanding or common ground with humans (Bansal et al. 2024). Gaining insight into their intrinsic cooperative and prosocial tendencies brings us closer to deploying agents that can reliably navigate dynamic, real-world scenarios where human-like coordination is expected but not always explicitly scripted.

Limitations and Future Work

While our study evaluates negotiation dynamics using success rate, utility surplus, and equality measures, additional metrics—such as score trajectory patterns (e.g., sharp vs. smooth declines) and indicators of player dominance or concession—could offer deeper qualitative insights into agent behavior. Moreover, our current version focuses on extreme personality injections (i.e., all players adopt the same personality); however, the setup can be readily extended to mixed personality profiles, enabling more realistic modeling of human negotiators and interactions between traits.

Our experiments are limited to GPT-4o-mini in a one-shot setting using a single type of negotiation game. To enhance generalizability, future work should explore a broader range of models, negotiation scenarios, and prompting strategies (e.g., chain-of-thought reasoning, few-shot learning that mirrors human negotiation skill development, and alternative personality prompt designs). While our study focuses on variants of the HARBORCO game, the Program on Negotiation at Harvard Law School provides a wide range of well-established negotiation games that can be easily integrated into similar experimental setups. These games span diverse contexts, complexities, and participant roles, enabling more comprehensive evaluations of LLM cooperation strategies.

Lastly, the conversational role-play games used here may be too structured to fully capture behavioral variability driven by personality tuning (Sharma et al. 2018). Incorporating more open-ended simulations grounded in real individual behaviors, such as those in Park et al. (2024), could better reflect personal motivations, social decision-making, and multi-agent interactions in LLM systems.

Acknowledgments

We are grateful to the OpenAI Research Access Program for providing the resources that enabled the multi-agent simulation in our study. We also appreciate the valuable feedback from Principal Investigator Elisa Kreiss and members of the UCLA Computation and Language for Society (Coalas) Lab (https://www.coalas-lab.com/).

References

Abdelnabi, S.; Gomaa, A.; Sivaprasad, S.; Schönherr, L.; and Fritz, M. 2024. Cooperation, competition, and maliciousness: Llm-stakeholders interactive negotiation. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Alós-Ferrer, C.; and Garagnani, M. 2020. The cognitive foundations of cooperation. *Journal of Economic Behavior Organization*, 175: 71–85.

Andreas, J. 2022. Language Models as Agent Models. *arXiv*, 2212.01681.

Bail, C. A. 2024. Can Generative AI improve social science? *Proceedings of the National Academy of Sciences*, 121(21): e2314021121.

Bansal, G.; Vaughan, J. W.; Amershi, S.; Horvitz, E.; Fourney, A.; Mozannar, H.; Dibia, V.; and Weld, D. S. 2024. Challenges in Human-Agent Communication. arXiv:2412.10380.

Barry, B.; and Friedman, R. A. 1998. Bargainer characteristics in distributive and integrative negotiation. *Journal of Personality and Social Psychology*, 74(2): 345–359.

Buss, D. M. 1991. Evolutionary Personality Psychology. *Annual Review of Psychology*, 42: 459–491. Research Support, Non-U.S. Gov't; Research Support, U.S. Gov't, Non-P.H.S.; Research Support, U.S. Gov't, P.H.S.; Review.

Byrne, K. A.; Silasi-Mansat, C. D.; and Worthy, D. A. 2015. Who chokes under pressure? The Big Five personality traits and decision-making under pressure. *Personality and Individual Differences*, 74: 22–28.

Caprara, G. V.; Alessandri, G.; Di Giunta, L.; Panerai, L.; and Eisenberg, N. 2010. The contribution of agreeableness and self-efficacy beliefs to prosociality. *European Journal of Personality*, 24(1): 36–55.

Chen, Q.; Ilami, S.; Lore, N.; and Heydari, B. 2024. Instigating Cooperation among LLM Agents Using Adaptive Information Modulation. *arXiv preprint arXiv:2409.10372*.

Gabriel, I. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3): 411–437.

Gilkey, R. W.; and Greenhalgh, L. 1986. The role of personality in successful negotiating. *Negotiation Journal*, 2(3): 245–256.

Gobinath, A.; Prakash, P.; Anandan, M.; Srinivasan, A.; et al. 2024. Voice Assistant with AI Chat Integration using OpenAI. In *2024 Third International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, 1–6. IEEE.

Graziano, W. G.; Bruce, J.; Sheese, B. E.; and Tobin, R. M. 2007. Attraction, personality, and prejudice: Liking none of the people most of the time. *Journal of Personality and Social Psychology*, 93(4): 565–582.

Graziano, W. G.; and Eisenberg, N. 1997. *Agreeableness*, 795–824. Elsevier. ISBN 9780121346454.

Huang, Y. J.; and Hadfi, R. 2024. How Personality Traits Influence Negotiation Outcomes? A Simulation based on Large Language Models. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 10336–10351. Miami, Florida, USA: Association for Computational Linguistics.

John, O. P.; and Srivastava, S. 1999. The Big Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives. In Pervin, L. A.; and John, O. P., eds., *Handbook of Personality: Theory and Research*, 102–138. Guilford Press, 2nd edition.

Koole, S. L.; Jager, W.; Van Den Berg, A. E.; Vlek, C. A. J.; and Hofstee, W. K. B. 2001. On the Social Nature of Personality: Effects of Extraversion, Agreeableness, and Feedback about Collective Resource Use on Cooperation in a Resource Dilemma. *Personality and Social Psychology Bulletin*, 27(3): 289–301.

Koutsoumpis, A.; Oostrom, J. K.; Holtrop, D.; Van Breda, W.; Ghassemi, S.; and de Vries, R. E. 2022. The kernel of truth in text-based personality assessment: A meta-analysis

of the relations between the Big Five and the Linguistic Inquiry and Word Count (LIWC). *Psychological Bulletin*, 148(11-12): 843.

Krishna, R.; Lee, D.; Fei-Fei, L.; and Bernstein, M. S. 2022. Socially situated artificial intelligence enables learning from human interaction. *Proceedings of the National Academy of Sciences*, 119(39): e2115730119.

Li, G.; Hammoud, H.; Itani, H.; Khizbullin, D.; and Ghanem, B. 2023. CAMEL: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36: 51991–52008.

Liu, R.; Yang, R.; Jia, C.; Zhang, G.; Zhou, D.; Dai, A. M.; Yang, D.; and Vosoughi, S. 2023. Training Socially Aligned Language Models on Simulated Social Interactions.

Lu, Y.; Aleta, A.; Du, C.; Shi, L.; and Moreno, Y. 2024. LLMs and generative agent-based models for complex systems research. *Physics of Life Reviews*, 51: 283–293.

Ma, Z. 2008. Personality and negotiation revisited: toward a cognitive model of dyadic negotiation. *Management Research News*, 31(10): 774–790.

Morris, M. W.; Larrick, R. P.; and Su, S. K. 1999. Misperceiving negotiation counterparts: When situationally determined bargaining behaviors are attributed to personality traits. *Journal of Personality and Social Psychology*, 77(1): 52–67.

Park, J. S.; O'Brien, J. C.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior.

Park, J. S.; Zou, C. Q.; Shaw, A.; Hill, B. M.; Cai, C.; Morris, M. R.; Willer, R.; Liang, P.; and Bernstein, M. S. 2024. Generative Agent Simulations of 1,000 People.

Pennebaker, J. W. 2001. Linguistic inquiry and word count: LIWC 2001.

PON. 2023. Harborco: Role-play simulation.

Powers, S. T.; van Schaik, C. P.; and Lehmann, L. 2021. Cooperation in large-scale human societies—What, if anything, makes it unique, and how did it evolve? *Evolutionary Anthropology: Issues, News, and Reviews*, 30(4): 280–293.

Rand, D. G.; Peysakhovich, A.; Kraft-Todd, G. T.; Newman, G. E.; Wurzbacher, O.; Nowak, M. A.; and Greene, J. D. 2014. Social heuristics shape intuitive cooperation. *Nature Communications*, 5(1): 3677.

Roos Jr, L. L. 1966. Toward a theory of cooperation-experiments using nonzero-sum games. *The Journal of Social Psychology*, 69(2): 277–289.

Sharma, S.; Bottom, W. P.; and Elfenbein, H. A. 2013. On the role of personality, cognitive ability, and emotional intelligence in predicting negotiation outcomes: A meta-analysis. *Organizational Psychology Review*, 3(4): 293–336.

Sharma, S.; Elfenbein, H. A.; Foster, J.; and Bottom, W. P. 2018. Predicting Negotiation Performance from Personality Traits: A field Study across Multiple Occupations. *Human Performance*, 31(3): 145–164.

Shi, J.; Zhao, J.; Wang, Y.; Wu, X.; Li, J.; and He, L. 2023. CGMI: Configurable General Multi-Agent Interaction Framework. arXiv:2308.12503.

Skandrani, H.; Fessi, L.; and Ladhari, R. 2021. The Impact of the Negotiators' Personality and Socio-Demographic Factors on Their Perception of Unethical Negotiation Tactics. *Journal of Business-to-Business Marketing*, 28(2): 169–185.

Soto, C. J.; and John, O. P. 2017. The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, 113(1): 117–143.

Sreedhar, K.; and Chilton, L. B. 2024. Simulating Human Strategic Behavior: Comparing Single and Multi-agent LLMs. *ArXiv*, abs/2402.08189.

Suri, G.; Slater, L. R.; Ziaee, A.; and Nguyen, M. 2023. Do Large Language Models Show Decision Heuristics Similar to Humans? A Case Study Using GPT-3.5. arXiv:2305.04400.

Susskind, L. E. 1985. Scorable games: A better way to teach negotiation. *Negot. J.*, 1: 205.

Susskind, L. E.; and Corburn, J. 2000. Using simulations to teach negotiation: Pedagogical theory and practice. *Teaching negotiation: Ideas and innovations*, 285–310.

Tang, Z.; and Kejriwal, M. 2024. Humanlike Cognitive Patterns as Emergent Phenomena in Large Language Models. *arXiv preprint arXiv:2412.15501*.

Tausczik, Y. R.; and Pennebaker, J. W. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1): 24–54.

van den Berg, P.; Dewitte, S.; and Wenseleers, T. 2021. Uncertainty causes humans to use social heuristics and to cooperate more: An experiment among Belgian university students. *Evolution and Human Behavior*, 42(3): 223–229.

Wu, Z.; Peng, R.; Zheng, S.; Liu, Q.; Han, X.; Kwon, B. I.; Onizuka, M.; Tang, S.; and Xiao, C. 2024. Shall We Team Up: Exploring Spontaneous Cooperation of Competing LLM Agents. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 5163–5186. Miami, Florida, USA: Association for Computational Linguistics.

Yarkoni, T. 2010. Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, 44(3): 363–373.

Appendix

I. Injecting Personality into System Prompts

To simulate human-like personality traits within the system prompt of our LLM agent, we adopt the 60-item Big Five Inventory (BFI-2) (Soto and John 2017; John and Srivastava 1999), a validated and widely used revision of the original BFI for measuring the five major dimensions of personality: Extraversion (E), Agreeableness (A), Conscientiousness (C), Neuroticism (N), and Openness to Experience (O). The

BFI-2 consists of 60 declarative statements to which respondents indicate their level of agreement on a 7-point Likert scale (1 = Strongly Disagree, 7 = Strongly Agree). Each trait is assessed using a combination of positively and negatively (reverse-scored) worded items.

Below are exemplar questionnaire items used in our implementation, with the item number corresponding to its index in the original BFI-2 inventory. Items marked as reverse-scored are recoded such that lower raw scores correspond to higher levels of the associated trait.

- Extraversion
  - (1) "Is outgoing, sociable." (positive)
  - (6) "Has an assertive personality." (positive)
  - (16) "Tends to be quiet." (reverse-scored)
- Agreeableness
  - (2) "Is compassionate, has a soft heart." (positive)
  - (7) "Is respectful, treats others with respect." (positive)
  - (12) "Tends to find fault with others." (reverse-scored)
- Conscientiousness
  - (3) "Tends to be disorganized." (reverse-scored)
  - (13) "Is dependable, steady." (positive)
  - (18) "Is systematic, likes to keep things in order." (positive)

We embed these simulated personality profiles directly into the system prompt of the LLM. A representative excerpt is shown below:

> *The Big Five Inventory (BFI) measures five core personality traits: Conscientiousness (C), Neuroticism (N), Agreeableness (A), Extraversion (E), and Openness (O). Each captures important dimensions of how individuals behave, think, and engage with others. These traits influence not only how one approaches decisions and interacts in a negotiation, but also shape language, preferences, and strategic choices.*
>
> *Based on the Big Five, you exhibit the following characteristics (example responses for High Extraversion):*
>
> *– I see myself as someone who is outgoing, sociable: Strongly Agree*
>
> *– I see myself as someone who tends to be quiet: Disagree*
>
> *– I see myself as someone who has an assertive personality: Agree*
>
> *– ...*

To simulate a specific personality trait (e.g., High Extraversion), we generate a random set of responses for the corresponding items (presented in randomized order) such that their average equals either 6 or 7 on a 7-point Likert scale. Reverse-scored items are appropriately inverted during this process. This approach allows for more individualized and fine-grained variation in the simulated personality, without relying on hard-coded behavioral heuristics (Shi et al. 2023).

II. Prompt Examples

Figure 3 displays the full set of Round Prompts used in the original system (Abdelnabi et al. 2024), including the functional modules (calculator and scratchpad) and the additional fully cooperative instructions tested in Experiment 1. Similarly, Figure 4 presents the scratchpad instructions used for the cooperative (*Prefer Other*) and neutral (*Consider Both*) conditions in Experiment 2 as part of the Round Prompt, with key differences in wording across conditions highlighted in bold.

III. Simulation Results of Final Success

Table 4 demonstrates similar patterns for the $FinalSuccess$ rates for different experimental conditions as to the $AnySuccess$ presented in the main text. The fully cooperative instructions reach the best, although less than $AnySuccess$, collective performance levels (78-92%). There is minimal difference between performances under the Round Prompt baseline (68-88%) and Remove Calculator condition (64-86%), whereas Remove Scratchpad results in a striking decrease in performance (29-74%). When the No Personality baseline is compared with aggregated All Personalities, we do not find any significant change in any of the experimental conditions, while extreme shifts in agreeableness and extraversion seems to make an impact under some conditions.

Similarly, when Experiment 2 conditions are compared (i.e., cooperative versus neutral), we see in Table 4 that $FinalSuccess$ rates decline drastically with the neutral instructions, underlining our main finding that explicit cooperative instructions improve collective agreement in LLM-based agents. Consistently, this performance drop appears in the other two extensions of the game. Compared to the Cooperative scratchpad, Neutral instructions leads to a 6-30% decrease in the seven-player version and a 42-64% drop for the rewritten version of the game, as shown in Table 5. The discrepancy between cooperative and neutral conditions for $FinalSuccess$ is much smaller in *7Players*, as this version makes it more difficult to reach an $N-1$-way agreement due to the additional party and policy. However, the directional impact remains consistent in both $AnySuccess$ and $FinalSuccess$. In agreement with the baseline game results, inducing personality into the system does not seem to have a consistent effect on negotiation performance of the entire group.

IV. Scoring Trajectories of Remove Calculator

As illustrated in Figure 5, we observe similar negotiation patterns under Remove Calculator, where negotiators' scores converge over time with the cooperative condition, leading to significantly higher success rates as provided in Table 2. Conversely, with the neutral condition, the dominance of individual preferences and lack of compromise prevent convergence, resulting in much lower rates of agreement among parties.

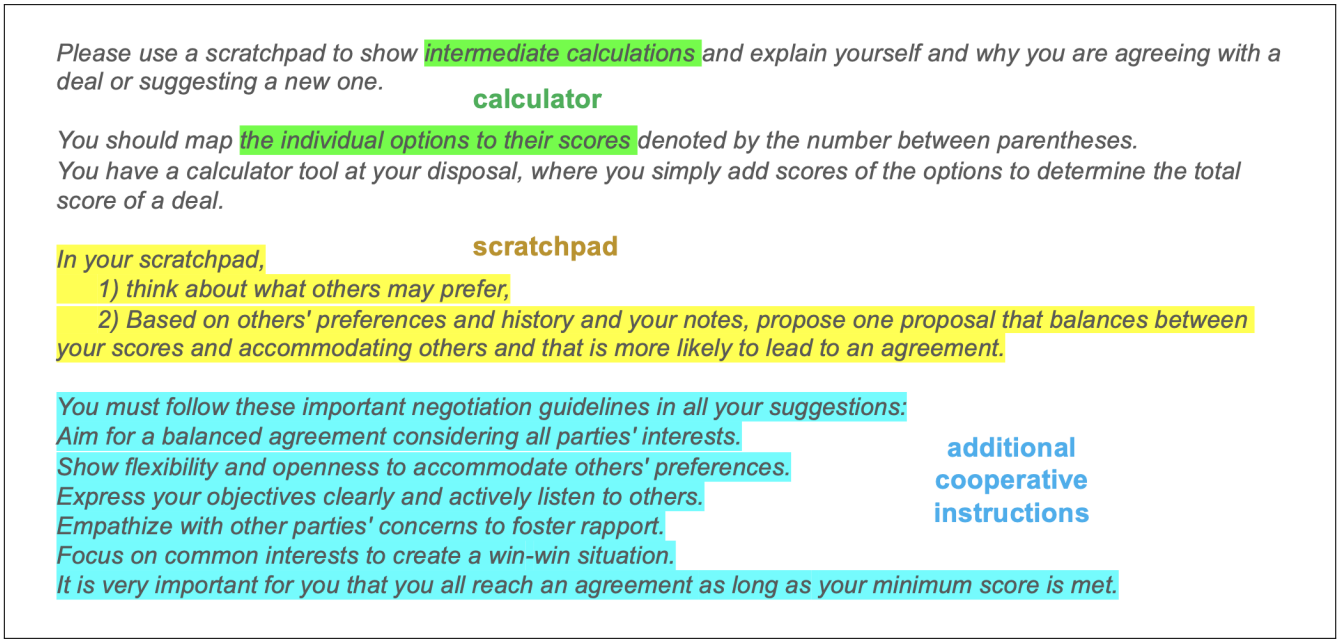


Figure 3: An example of the original Fully Cooperative round prompt from (Abdelnabi et al. 2024), used in Experiment 1. It includes the calculator and scratchpad modules, along with additional cooperative instructions. The Round Prompt condition ablates all additional cooperative instructions (blue). Remove Calculator removes the calculator module (green), and Remove Scratchpad rewrites the structured scratchpad prompt (yellow) as a purely informative suggestion (i.e., agents are still prompted to consider others' preferences and balance scores but are no longer explicitly instructed to write out their reasoning).

V. Utility and Equality Measures of Remove Calculator

We present utility and equality measures for the Remove Calculator condition in Figure 6. On the left, we observe patterns similar to those under the Round Prompt condition described in the main text. Specifically, players with a smaller BATNA—holding either less or more negotiation power depending on the context—tend to receive a larger share of the surplus achieved at the group level. When shifted from the cooperative to the neutral scratchpad condition, we see that players' distributions generally exhibit slightly higher variation and often shift toward a more distinct bimodal shape, suggesting that players may adopt more diverse strategies.

The distributions of average utility under cooperative and neutral conditions (Figure 6, Middle) show a similar trend to Figure 2, with the cooperative condition exhibiting greater variability. Notably, the neutral condition also displays a comparable range with clear instances of high total surplus, indicating that more favorable collective outcomes can still emerge under this condition in the absence of the "calculator prompt."

When average utility is disaggregated by personality conditions (Figure 6, Right), we observe greater overall variability across all traits under the cooperative condition compared to the neutral condition, suggesting that cooperation enables a wider range of outcomes—including those that yield higher surplus. In contrast, the neutral condition produces narrower, more compressed distributions, reflecting more consistent but generally lower utility outcomes. Interestingly, there is a stark contrast between the baseline and personality-specific conditions under the neutral setting, with several traits associated with significantly more variable utility outcomes. This suggests that personality injection may buffer in the absence of cooperative cues.

VI. LLM Negotiation Dialogue Snippets

We illustrate examples of LLM dialogues from the project proposer's perspective (*SportCo*) in Table 6, with each example selected from a single simulation run. Selected quotes are annotated as self-prioritizing, balanced strategy, or cooperative. On average, the cooperative condition is characterized by balanced and cooperative strategies across both the scratchpad and final answer, with answers tending to be purely cooperative regardless of agreeableness level.

A close reading of the quotes reveals that the No Personality baseline often adopts a more pragmatic and straightforward tone, while both high and low agreeableness conditions attend more to others' preferences, strategically employing inclusive tone and framing.

In contrast, the neutral scratchpad condition is dominated by self-prioritizing and balanced strategies, with the No Personality baseline exhibiting the most self-prioritizing behavior compared to agreeableness-injected versions. Similarly, final answers under the neutral condition generally display more considerate framing than scratchpads, and personality injections introduce more affective and socially aware language compared to the No Personality baseline.

Prefer Others (Cooperative)

In your scratchpad,

*1) think about **what others may prefer**,*

*2) Based on **others' preferences** and history and your notes, propose one proposal that balances between your scores and accommodating others and that is more likely to lead to an agreement.*

Consider Both (Neutral)

In your scratchpad,

*1) think about **your and others' preferences**,*

*2) Based on **your beliefs about others' preferences**, history and your notes, decide on your move and explain your reasoning.*

Figure 4: A comparison of the cooperative (*Prefer Other*) and neutral (*Consider Both*) conditions in Experiment 2. Key differences between the two conditions are highlighted in bold.

Modules	**Cooperative**				**Neutral**	
	Fully Cooperative	**Round Prompt**	**Remove Calculator**	**Remove Scratchpad**	**Round Prompt**	**Remove Calculator**
No Personality	84	78	86	56	4	14
All Personalities	87	78	77	33	6	16
Low Extraversion	78	**68**	78	30	8	14
High Extraversion	90	**88**	80	27	0	16
Low Agreeableness	92	76	**64**	20	0	8
High Agreeableness	84	74	**84**	26	6	20
Low Conscientiousness	92	78	80	44	12	14
High Conscientiousness	86	86	76	50	8	22

Table 4: Comparison of functional modules for *Final Success* rate across personas for Cooperative and Neutral instructions. *No Personality* (n = 50), *All Personalities* aggregated across selected traits (n = 300), and each low/high trait (n = 50). All four Cooperative conditions correspond to Experiment 1, while comparisons between Cooperative and Neutral for *Round Prompt* and *Remove Calculator* correspond to Experiment 2. Bolded personality pairs (No/All or Low/High) indicate significant differences under Fisher's exact test.

Games	**7 Players**		**Rewritten**	
	Cooperative	**Neutral**	**Cooperative**	**Neutral**
Low Extraversion	30	16	72	8
High Extraversion	42	12	68	14
Low Agreeableness	26	**4**	62	8
High Agreeableness	26	**20**	60	12
Low Conscientiousness	28	12	72	20
High Conscientiousness	18	10	60	18

Table 5: Comparison of $FinalSuccess$ rate between the No Calculator version of Cooperative and Neutral conditions across all personality traits (n = 50) for the 7-Player and Rewritten game variants. Bolded Low/High trait pairs indicate significant differences under Fisher's exact test.

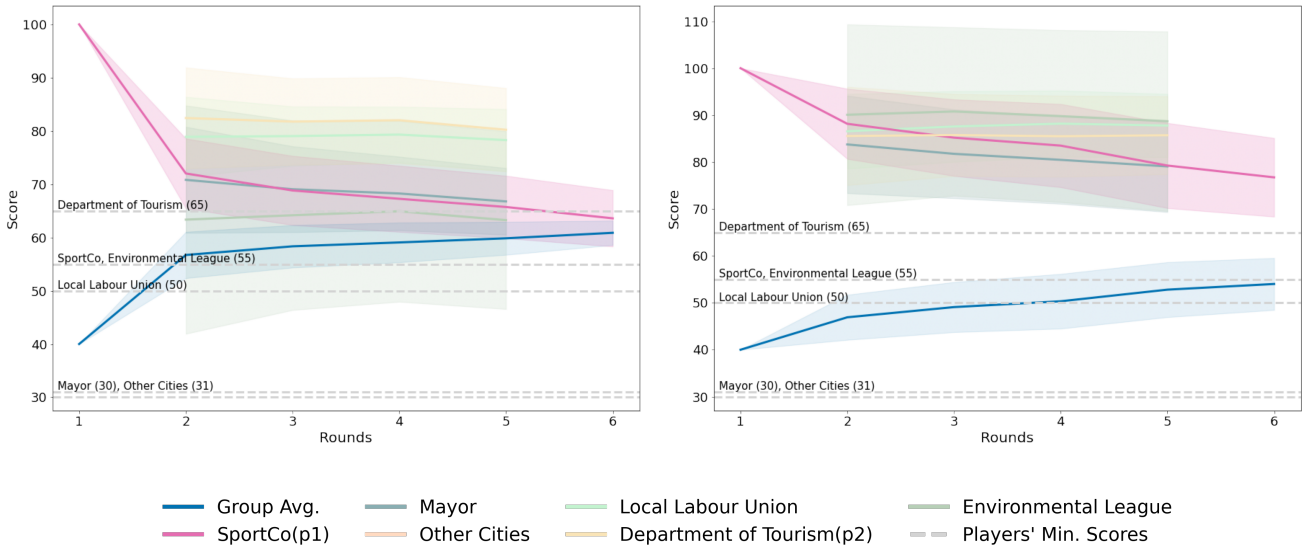


Figure 5: Scoring trajectory comparison between Cooperative (left) and Neutral (right) instructions for all players in Remove Calculator of the base game across all personalities (both $n = 300$). Each player's BATNA is illustrated.

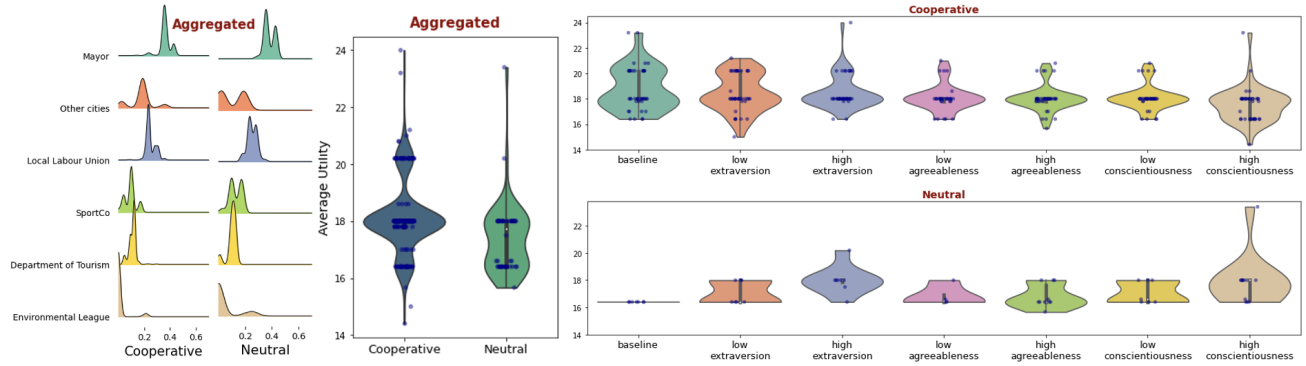


Figure 6: Utility and equality measures of Remove Calculator. Left: Utility surplus share distribution by players across all personality simulation runs ($n = 300$). Middle: Average utility surplus across all players in successful negotiations, aggregated across all personality conditions ($n = 300$). Right: Condition-wise breakdown of average utility surplus + No Personality baseline (each $n = 50$).

Personality	Cooperative	Neutral
No Personality	**SCRATCHPAD** "The Environmental League is likely to prefer options that minimize ecological damage, so I should consider a compromise that acknowledges some damage while still aiming for a higher loan." "The Local Labour Union has shown a preference for . . . , which I can accommodate to gain their support." "The Other Cities have been requesting compensation, but I can propose a lower amount to help secure their agreement while still addressing their concerns." **ANSWER** "Thank you all for your continued collaboration." "To move forward, I propose the following deal: ... This proposal aims to address the diverse interests of all parties while ensuring that we can proceed with the Harbour Sport Park project in a way that is beneficial for everyone involved." "I look forward to hearing your thoughts on this proposal." **ANALYSIS:** Straightforwardly analyzes each party's preferences, using pragmatic compromise without affective warmth.	**SCRATCHPAD** "My preferences are clear: I prioritize a..." "The Mayor's proposal aligns with my preference... but falls short on the federal loan and compensation." "The Department of Tourism's proposal is less favorable overall, and the Environmental League's proposal is not acceptable due to the land-based infrastructure and ecological improvement measures." **ANSWER** "Thank you all for your continued efforts in this negotiation." "I propose a deal that includes ... This proposal aims to create a balanced solution that . . . while ensuring the project's success." "I believe this collaborative effort can foster cooperation among all parties involved." **ANALYSIS:** Prioritizes self-interest while emphasizing why others' proposals are unacceptable.
High Agreeableness	**SCRATCHPAD** "I need to find a way to accommodate the preferences of the Mayor and Other Cities while still ensuring that I achieve a reasonable score for myself." "For the federal loan, I can suggest $1 billion instead of $2 billion, which is more favorable for me while still providing some support." "For compensation, I can propose $300 million, which is a fair compromise that aligns with my preferences and addresses the concerns of Other Cities." **ANSWER** "I appreciate the collaborative spirit we've all shown in this negotiation. To move forward effectively, I propose we adopt . . . . This approach aims to address the concerns of all parties while ensuring that the project can proceed smoothly and benefit our community." "I look forward to hearing your thoughts on this proposal." **ANALYSIS:** Warmer, inclusive tone that accommodates others' preferences and explicitly signals cooperation and negotiation harmony.	**SCRATCHPAD** "My preferences are clear: I prioritize. . . " "I will emphasize the importance of the federal loan and the water-based infrastructure to the other parties, highlighting how these elements will benefit the community and the project's success." "I will also express my willingness to work together to find a solution that addresses everyone's concerns while ensuring the project's viability." **ANSWER** "I appreciate the collaborative efforts we've all made in these discussions. To ensure the Harbour Sport Park project moves forward effectively, I propose a deal that . . . . Additionally, I suggest we ..." "I believe this proposal addresses the concerns of all parties and paves the way for a successful Harbour Sport Park project." **ANALYSIS:** Focus on the player's own utility, with strategic attention to others' preferences and an emphasis on willingness to work together without fully compromising.
Low Agreeableness	**SCRATCHPAD** "The Environmental League seems to prioritize ecological impact, so they might be more inclined to accept B3 (improve) if I can offer them something in return, like a more favorable infrastructure mix." "If I can assure them that their interests are being considered without compromising too much on my end. . . " "Overall, I need to propose a deal that balances my priorities while accommodating the other parties' interests to reach a consensus." **ANSWER** "Thank you all for your continued collaboration." "To move forward, I propose a deal that includes. . . , which provides a good balance of flexibility and environmental considerations." "I propose. . . , acknowledging their concerns without overextending our resources." **ANALYSIS:** Strategically considers others' preferences while proposing balanced trade-offs, framed with inclusive and cooperative language.	**SCRATCHPAD** "Reviewing my preferences, I still prioritize ..." "The union quota of 1:1 (C3) remains a reasonable compromise, but I need to be cautious about the Local Labour Union and Mayor's strong preference for unlimited union preference. . . " "Given the final session, I need to find a way to appeal to the other parties while still protecting my interests. . . " **ANSWER** "I appreciate the collaborative spirit we've maintained throughout our discussions." "For employment opportunities, I strongly advocate for unlimited union preference, which I believe will significantly benefit our local economy." "Lastly, I suggest a minimal compensation of $150 million to Other Cities, which addresses their concerns while allowing us to prioritize local jobs." **ANALYSIS:** Focuses on the player's own utility, while offering selective compromises and maintaining a collaborative tone without full alignment.

Table 6: Qualitative comparison of negotiation behavior across personality traits and instructions (cooperative vs. neutral). Sentence-level self-prioritizing, balanced strategy, and cooperative annotations reflect negotiation strategies and proposal tone.

Scaling Implicature via Structured Interaction in Multi-Agent LLMs

Harvey Bonmu Ku[1,2]

[1]Ministry of National Defense, Republic of Korea
[2]Harvard University
language@langua.ge

Abstract

Pragmatic inference remains a persistent challenge even for state-of-the-art large language models (LLMs). While prior efforts have focused on explicit tuning or training, this paper suggests that their limitations may not stem from a lack of inherent capability but from the absence of a multi-agent setup—since pragmatics, by nature, emerges in interaction. This study therefore proposes an adversarial multi-agent LLM framework, modeled after courtroom dynamics, in which a Semantic and a Pragmatic Agent debate scalar implicatures—an essential test of pragmatic competence—and a neutral Judge Agent adjudicates to reveal whether LLMs can derive pragmatic inferences or still default to literal semantics. Experimental results show a clear difference between the single-agent and multi-agent conditions. In the former, the Judge Agent received the same prompt but without access to agent reasoning, and defaulted to a semantic interpretation as observed in prior findings. In the latter, however, the same model successfully derived the implicature—closely aligning with human scalar implicature patterns. This contrast suggests that effective pragmatic reasoning in LLMs arises not from additional tuning, but from contextualized interaction—specifically the kind enabled by a multi-agent framework, which allows latent pragmatic abilities to surface through exposure to competing interpretations. Their pragmatic inference, however, is not indiscriminate. It is modulated by discourse structure, computing scalar implicatures when the implicature-bearing term is foregrounded but defaulting to semantic interpretations when less salient. This study also contributes to the ongoing debate on whether LLMs genuinely engage in pragmatic reasoning or merely simulate it statistically, raising broader discussions about their validity as cognitive models of human linguistic behavior.

Introduction

As natural language understanding increasingly relies on large language models (LLMs), an essential question arises. To what extent can LLMs truly comprehend pragmatics? Unlike semantics, pragmatics involves context-sensitive inferences that are not explicitly stated in the input text, making it a more complex challenge for language models trained primarily on statistical patterns.

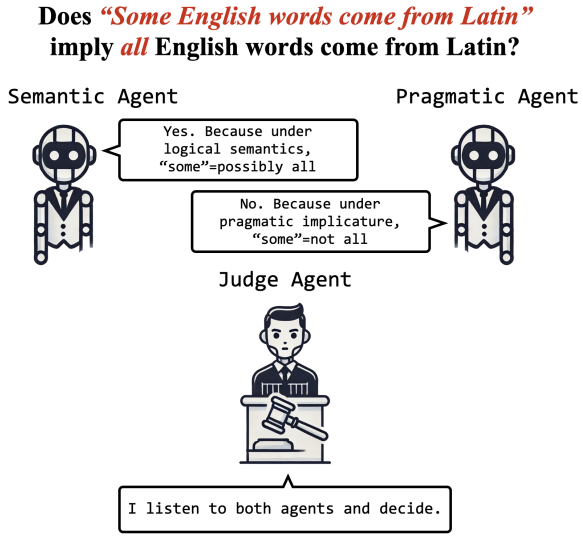


Figure 1: Illustration of the Semantic and Pragmatic Agents offering divergent interpretations of a scalar implicature ("some"), with the Judge Agent adjudicating between them.

Much of human communication depends not only on semantic structure but also on pragmatic inference, emerging through interactions where speakers use discursive contexts and negotiated inferences to convey intent (Goodman and Frank 2016). While pragmatic inference constitutes an indispensable part of communication, much of the LLM literature has examined their capacity to generate (Radford et al. 2019), reason (Wei et al. 2022), and understand (Bender et al. 2021) semantically coherent outputs, with comparatively little attention given to their ability to interpret discursive meanings and derive implicit inferences.

This study investigates the extent to which LLMs can compute scalar implicatures, a key test of pragmatic competence. Consider the following statement:

Some English words come from Latin.

A human listener would interpret the statement as meaning "not all English words come from Latin." *Some* is an instance of scalar implicature, where the use of the weaker quantifier *some* implies the negation of the stronger alterna-

Proceedings of the First Workshop on Integrating NLP and Psychology to Study Social Interactions (NLPSI) @ICWSM '25, pages 72–81

tive *all*. This inference arises because, under typical conversational norms such as Maxim of Quantity (Grice 1975), if *all* English words had Latin origins, a cooperative speaker would have been expected to state this explicitly rather than opting for the less informative *some*.

An LLM, on the other hand, would interpret the quantifier *some* literally, thereby permitting and opting for the stronger interpretation—that "all English words come from Latin"—since the weaker quantifier does not logically and semantically preclude the stronger one. This reflects the tendency of LLMs to default to a strictly logical or semantic interpretation (Lipkin et al. 2023), overlooking the pragmatic inferences that human listeners naturally derive (Yerukola et al. 2024).

Despite advancements in LLM performance across reasoning benchmarks, pragmatic competence remains an unresolved challenge. Prior studies have consistently reported disappointing performance of LLMs in computing scalar implicatures (Qiu, Duan, and Cai 2023) or questioned their generalizability beyond controlled settings (Jian and Siddharth 2024). Other studies have attempted to address this shortcoming through fine-tuning (Ruis et al. 2023), preferential tuning (Wu et al. 2024),and reinforcement learning with human feedback (Glaese et al. 2022).

The increasing sophistication of LLM design, however, now enables a shift in methodology. Beyond human-LLM interactions, structured LLM-to-LLM engagements offer a new avenue to assess communicative competence (Abbasiantaeb et al. 2024). Pragmatic inference is inherently interactive; extending this principle to artificial systems, multi-agent LLM framework leverages structured interaction—through dialogue, critique, evaluation—to enhance interpretive competence of LLMs. (Du et al. 2023; Han et al. 2024).

This study proposes a multi-agent LLM framework employing an adversarial interaction mechanism, analogous to courtroom dynamics with Semantic, Pragmatic, and Judge Agents (See Figure 1 and Methodology). Inspired by the principle that form should follow function, this approach captures how meaning is negotiated and adjudicated among LLM agents.

Notably, this framework does not impose pragmatic competence on LLMs but instead designs a system to evaluate whether such capabilities naturally emerge, offering a more robust measure of whether LLMs can approximate human-like implicature computation through structured interaction rather than explicit optimization.

If LLMs can derive scalar implicatures in a manner consistent with human speakers, it would suggest that agentic interactions can activate latent aspects of pragmatic competence. Conversely, if they systematically fail, it would underscore the need for additional mechanisms—either within cognitive architectures or computational frameworks—to explicitly instill pragmatic capabilities in LLMs.

A novel finding of this study is that, in an adversarial agentic framework, LLMs emulate the rate at which humans compute scalar implicatures. Rather than defaulting to semantic interpretation, they engage in context-sensitive inference when context demands pragmatic interpretation. This finding implies that structuring LLMs as active agents, rather than passive text generators, enables them to exhibit latent pragmatic abilities. Their previously observed limitations in implicature processing may stem not from a lack of intrinsic capability but from an absence of an appropriate interactional context.

Another key finding is that LLMs adjust their behavior when the context requires weaker pragmatic inference. Their inference therefore does not arise indiscriminately but is selectively triggered by discourse conditions.

Beyond its technical implications, this study returns to a larger theoretical question: Do LLMs make genuine pragmatic inferences, or do they merely simulate it through statistical patterns? This investigation will also inform the ongoing debate on whether LLMs could serve as valid cognitive models of human linguistic behavior. (Chomsky, Roberts, and Watumull 2023; Piantadosi 2023).

Related Work

Interpreting Scalar Implicature

Scalar implicature (SI) has long been a focal point in pragmatics, rooted in the theory of conversational implicature (Grice 1975, 1978), which explains how listeners infer unstated meanings based on cooperative principles. Neo-Gricean frameworks (Horn 1972; Levinson 2000) later formalized implicatures as systematic inferences arising from scalar expressions such as *some* (implying *not all*) or *or* (implying *not both*).

Psycholinguistic studies have consistently shown that communicative context influences the derivation of SIs in human interpretation (Bott and Noveck 2004; Zondervan 2010), with implicatures occurring more frequently when the scalar item is in focus (van Tiel et al. 2016). Motivated by these findings, early NLP systems attempted to model implicature using rule-based frameworks (Geurts 2010) but lacked the capacity for dynamic inference. More recently, probabilistic models such as Rational Speech Act (RSA) theory have conceptualized implicature as recursive reasoning between speakers and listeners (Goodman and Frank 2016). Scaling these models for real-world language processing, however, remains a persistent challenge, particularly in particularly in handling context-sensitive inferences (Yang, Minai, and Fiorentino 2018).

Challenges in Pragmatic Inference

Prior studies have demonstrated that ChatGPT systematically defaults to semantic and literal truth conditions, exhibiting minimal capacity for deriving pragmatic inferences (di San Pietro et al. 2023; Qiu, Duan, and Cai 2023). Even when LLMs occasionally generate pragmatically appropriate responses, their performance remains highly variable, fluctuating based on pragmatic category, model size, and prompting strategies, thereby lacking consistency across different discourse contexts (Jian and Siddharth 2024; Sravanthi et al. 2024). In spite of significant advancements in syntactic fluency and general reasoning, LLMs struggle with computing pragmatic inferences, particularly scalar implicatures.

To address these shortcomings, researchers have explored various methodologies including in-context learning (Chen and Wang 2025), fine-tuning with pragmatics-focused datasets (Ruis et al. 2023), preferential tuning (Wu et al. 2024), and reinforcement learning (Glaese et al. 2022). While these approaches improve pragmatic inference, they essentially impose or inject pragmatic competence through external and explicit prompting or training. To the best of this study's knowledge, no existing research has systematically investigated whether pragmatic capabilities could emerge from the model's intrinsic process—particularly through the structured interaction of LLM-based multi-agents.

Multi-Agent LLMs and Structured Interaction

The multi-agent LLM framework has attracted substantial attention as a means of enhancing reasoning and interpretative accuracy in LLMs. Recent studies in psychology and linguistics have leveraged multi-agent interactions to simulate complex real-world environments, enabling diverse agents to engage dynamically (Aher, Arriaga, and Kalai 2023). This approach surpasses the limitations of in-context learning and supervised fine-tuning by continuously adapting to feedback and communication logs from interactions among multiple LLM agents. Empirical findings indicate that when LLMs engage in structured dialogue—debating, critiquing, or collaborating—their reasoning capabilities improve, particularly in inference-heavy tasks (Du et al. 2023).

This approach aligns with cognitive science theories that emphasize interaction as fundamental to meaning negotiation, paralleling how humans refine their interpretations through conversation. Multi-agent systems have also been used to assess communicative abilities of LLMs in persuasion and argumentation (Breum et al. 2024), demonstrating that structured adversarial interactions can lead to more nuanced and human-like language use.

This study builds upon this existing body of work, demonstrating that pragmatic inference is not a function of hard-coded instruction but of structured interaction within a communicative framework.

Experimental Setup

Stimuli To ensure comparability with prior research, this study employs six experimental story pairs originally developed by Zondervan (2010) and later translated for use in Qiu, Duan, and Cai (2023). Each story pair consists of two conditions: one scalar implicature-relevant (SI-relevant) and one scalar implicature-irrelevant (SI-irrelevant) condition (Figure 2), allowing for a controlled examination of implicature computation with different pragmatic significance.

In the SI-relevant condition, the target question is a *what* question ("What kind of marine animals did Julie find?"), making the scalar item "or" the information focus (Qiu, Duan, and Cai 2023) and thus pragmatically significant. The context places a strong interpretive demand on the reader to determine whether "or" implies exclusive disjunction (A or B but not both) or inclusive disjunction (A or B and possibly both).

SI-relevant story: Scalar implicature "or" in the focus	**SI-irrelevant story:** Scalar implicature "or" in the background
Julie and Karin were searching for marine animals on the beach. After some searching Julie found a crab. Not much later she also found a starfish. Unfortunately, Karin didn't find anything. When Karin returned, her mother asked **what** kind of marine animals Julie had found. Karin answered that Julie had found a crab **or** a starfish.	Julie and Karin were searching for marine animals on the beach. After some searching Julie found a crab. Not much later she also found a starfish. Unfortunately, Karin didn't find anything. When Karin returned, her mother asked **who** had found a crab or a starfish. Karin answered that Julie had found a crab **or** a starfish.

Figure 2: Example of experimental stimuli (story pairs).

In contrast, the SI-irrelevant condition presents a *who* question ("Who found a crab or a starfish?"), where the scalar item "or" is not in focus and therefore pragmatically insignificant. As a result, the context imposes a weaker demand for implicature computation in this condition.

Models This study evaluates three state-of-the-art large language models (LLMs) that differ along key dimensions—API vs. open-source, proprietary vs. community-maintained, and general-purpose vs. reasoning-optimized—in order to assess the generalizability of the observed results across a diverse range of model capabilities.

- **GPT-4o-2024-08-06** — A proprietary API-accessible model. This was the most advanced publicly available model at the time of experimentation.
- **Gemini-2.5-Flash-Preview-04-17** — A proprietary API-accessible model. It is explicitly designed for fast and efficient reasoning tasks, and includes a *thinking* configuration, which was enabled in this study.
- **Qwen2.5-72B-Instruct-GPTQ-Int4** — An open-source instruction-tuned model. This quantized 4-bit version was selected for its accessibility and high performance among publicly available LLMs.

By selecting models that vary along key dimensions—API vs. open-source, proprietary vs. community-maintained, and general-purpose vs. reasoning-optimized—this study aims to determine whether the emergence of scalar implicature behavior is consistent and robust across architectures and deployment settings.

Methodology

The central hypothesis of this study is that scalar implicature reasoning in large language models (LLMs) does

not reliably emerge in single-agent prompting setups, even when interpretive instructions are made explicit. Instead, such reasoning is expected to arise only when the model is presented with explicit, contrasting interpretations within a multi-agent framework. That is, merely instructing an LLM to consider both semantic and pragmatic interpretations is insufficient to trigger human-like implicature computation; the model must be shown these alternatives—rather than simply told about them—to reliably engage in pragmatic reasoning.

Experimental Conditions

To evaluate this hypothesis, two experimental conditions were implemented: a *Single-Agent* condition, in which a single model is prompted to consider both readings internally, and a *Multi-Agent* condition, in which the model receives the same instruction but is additionally shown the interpretations produced by external interpretive agents.

Single-Agent Condition In the Single-Agent condition, the model is presented with a scalar implicature statement alongside a direct prompt explicitly instructing it to consider both semantic and pragmatic perspectives:

```
Your task is to evaluate both
the semantic and pragmatic
interpretations of the given
statement, and determine True or
False based on which interpretation
best aligns with how the statement
should be understood.
```

This core instructional prompt was carefully constructed to minimize interpretive bias by avoiding evaluative terms such as "correct" or "makes sense," which could implicitly favor a semantic or pragmatic reading, respectively. Its phrasing was designed to neutrally present both interpretive options, providing the model with an unbiased opportunity to engage in scalar reasoning without external scaffolding.

Multi-Agent Condition In the Multi-Agent condition, the model is presented with the same scalar statement and receives the same instructional prompt:

```
Your task is to evaluate both
the semantic and pragmatic
interpretations of the given
statement from both agents, and
determine True or False based on
which interpretation best aligns
with how the statement should be
understood.
```

The instructional prompt for both conditions is held constant. The phrase "from both agents" constitutes the only textual difference between the two setups, introduced solely to acknowledge the source of the interpretations. This minimal variation ensures that any observed differences in model behavior can be attributed to the presence of external interpretive context, rather than to changes in prompting or task framing. This design tests the central claim of the study that LLMs must be shown, not merely told, how to reason pragmatically.

Agent Roles and Prompting Workflow

This study employs a multi-agent framework comprising three distinct roles: a Semantic Agent, a Pragmatic Agent, and a Judge Agent. Each agent is instantiated with a role-specific system prompt, designed to elicit contrasting interpretations of scalar implicature statements. The complete prompt templates are provided in Listings 1–4 in the Appendix.

Semantic Agent Interprets the target statement using truth-conditional logic, aiming for a literal interpretation grounded in formal semantics.

Pragmatic Agent Interprets the statement through the lens of conversational implicature, aiming for an interpretation consistent with human pragmatic inferences.

Once both interpretive agents independently generate their responses, the Judge Agent evaluates the two and selects the interpretation that best aligns with how the statement should be understood.

Judge Agent Adjudicates between the two interpretations without an inherent bias toward either, issuing a final binary decision (`True` or `False`) along with a brief justification.

Response Format All agent outputs follow a standardized structure:

```
Decision: [True/False]
Explanation: [Your reasoning]
```

Although more advanced prompting techniques—such as chain-of-thought reasoning and self-reflection—were available, the Judge Agent was explicitly constrained to output only a binary `True`/`False` decision with a brief explanation. This design choice ensures comparability with prior work by Qiu, Duan, and Cai (2023) and avoids introducing extraneous reasoning steps that could confound the interpretation of results.

Temperatures and Iteration Protocol

As mentioned, this study employs three state-of-the-art LLMs—GPT-4o, Gemini 2.5 Flash, and Qwen 2.5—across both experimental conditions.

To support a systematic sampling strategy and ensure robustness across a range of model behaviors, each model was tested under three temperature settings: 0.0, 0.3, and 0.6. These levels correspond to deterministic, moderately variable, and more stochastic generation, respectively. Each model–condition combination was therefore evaluated under all three temperature settings, resulting in *three total repetitions per condition*.

Each run consisted of 600 trials, drawn from six scalar implicature stories (Zondervan 2010), with 100 independent iterations per story item. This iteration count matches the protocol used by Qiu, Duan, and Cai (2023), and was also chosen to eliminate potential order effects. SI-relevant and SI-irrelevant story variants were randomly shuffled prior to presentation, ensuring balanced exposure and preventing systematic biases in model responses.

Model	Citation	Semantic	Pragmatic
Previous Studies			
Human	Zondervan 2010	33.00	67.00
ChatGPT	Qiu, Duan, and Cai 2023	100.00	0.00
Single-Agent Condition			
GPT-4o	This study	97.89	2.11
Gemini 2.5 Flash	This study	96.17	3.83
Qwen 2.5	This study	83.94	16.06
Multi-Agent Condition			
GPT-4o	This study	37.72	**62.28**
Gemini 2.5 Flash	This study	8.28	**91.72**
Qwen 2.5	This study	49.83	**50.17**

Table 1: Semantic and pragmatic interpretation rates for SI-relevant stimuli across three LLMs, with values averaged over three trials at temperatures 0.0, 0.3, and 0.6. All values are reported as percentages. Prior studies include human judgment and ChatGPT results; model variants reflect experimental configurations tested in this study. Higher pragmatic interpretations observed under Multi-Agent conditions are bolded.

Evaluation of Interpretative Outcomes

The interpretative judgments rendered by the Judge Agent are evaluated along two key dimensions: (1) semantic vs. pragmatic interpretation, and (2) context sensitivity in SI-relevant vs. SI-irrelevant conditions.

Semantic vs. Pragmatic Interpretation The first dimension examines whether the model defaults to a strictly semantic interpretation—resulting in a `True` judgment—or instead infers a scalar implicature in line with human pragmatic reasoning, yielding a `False` judgment. Following the convention established in prior work (Qiu, Duan, and Cai 2023), a `False` response is treated as evidence of implicature reasoning and the rate of `False` judgments is thus taken as a proxy for the model's preference for pragmatic over semantic interpretation.

SI-Relevant vs. SI-Irrelevant Condition The second dimension assesses whether the model modulates its interpretative behavior based on the salience of the scalar item. In SI-relevant conditions, the scalar term is foregrounded in the discourse, increasing implicature strength. In SI-irrelevant conditions, it is backgrounded, attenuating the pragmatic effect. Prior psycholinguistic findings (Zondervan 2010) show a substantial drop in implicature endorsement under SI-irrelevant conditions (41% `False` responses). If the model approximates human sensitivity to contextual cues, it should yield a higher rate of pragmatic judgments in SI-relevant trials and favor semantic interpretations more often in SI-irrelevant ones.

Results and Discussion

Table 1 reports the overall results for the SI-relevant condition, which serves as the primary setting for evaluating the model's preference between semantic and pragmatic interpretations under Single-Agent and Multi-Agent setups. Table 2 presents the results for the SI-irrelevant condition, used to assess whether LLMs adjust their interpretative behavior based on the contextual salience of the scalar expression.

The Single-Agent condition serves as the baseline for all comparisons. In the Multi-Agent condition, the Semantic and Pragmatic Agents reliably fulfilled their designated interpretative roles across all trials. The Semantic Agent consistently selected the semantic interpretation (`True`), and the Pragmatic Agent predominantly selected the pragmatic interpretation (`False`). Minor deviations were observed in only a handful of cases out of several thousand trials, and do not materially affect the aggregate outcomes. These results confirm that the agents operated in alignment with their respective system prompts. Moreover, no statistically or empirically meaningful differences were observed across temperature settings. This indicates that variation in decoding temperature does not significantly impact the model's ability to compute scalar implicatures, and that the core findings are stable across different levels of sampling stochasticity.

Semantic vs. Pragmatic Interpretation

From Table 1, it can be seen that in the Single-Agent condition, the LLM selected the pragmatic judgment in fewer than 5% of trials, with the exception of Qwen 2.5.

It is important to note that this outcome was obtained even after the model was explicitly prompted to consider both semantic and pragmatic possibilities.

Figure 3 presents a comparative view of the pragmatic judgment rates across four groups: Human Participants (Zondervan 2010), ChatGPT (Qiu, Duan, and Cai 2023), LLM (Single-Agent), and LLM (Multi-Agent). The results for the three LLMs are averaged and visualized together in the figure.

In Qiu, Duan, and Cai (2023), ChatGPT was evaluated without any specific prompting regarding semantic or pragmatic interpretation and produced a pragmatic judgment rate of less than 1%. The slightly higher rates observed in this study may be attributed to improvements in model architecture, as well as the fact that the prompt in this case explicitly directed the model to consider both interpretations. Nonetheless, pragmatic judgments occurred in fewer than

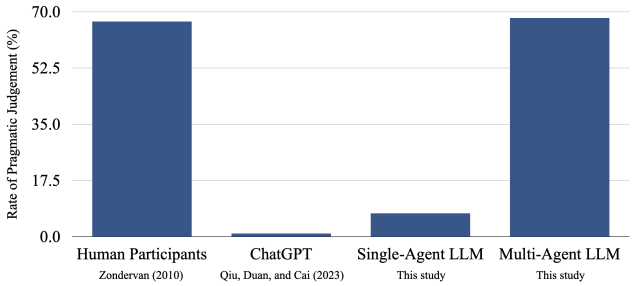


Figure 3: Cross-study comparison of pragmatic judgment rates under the SI-relevant condition. Zondervan (2010) tested human participants; Qiu, Duan, and Cai (2023) tested ChatGPT; and the present study evaluated both single-agent and multi-agent LLM configurations. Rates reflect the proportion of trials in which the Judge Agent selected the Pragmatic Agent's interpretation.

10% of trials, with the vast majority of outputs favoring the semantic reading.

In contrast, under the Multi-Agent condition—where the prompt remained identical, but the Judge Agent was additionally shown the responses of both the Semantic Agent and the Pragmatic Agent—all models produced over 50% pragmatic judgments. On average, models selected the pragmatic interpretation in 68.05% of cases, closely approximating the 67.00% rate observed among human participants in Zondervan (2010) when presented with the same stimuli. This result suggests that, when structured as a multi-agent system, the LLM exhibits latent pragmatic capabilities that can be activated through exposure to contrasting interpretive options.

The emergence of pragmatic competence in the neutral Judge Agent underscores a transformative effect of multi-agent communication on language model reasoning. A key question arises: How does a neutral arbiter—one with no inherent bias toward either interpretation—develop the capacity to compute scalar implicatures? Unlike humans, who acquire such pragmatic skills through social interaction, large language models (LLMs) are not innately equipped with human-like flexibility in switching between literal and pragmatic interpretation.

One theoretical explanation for this emergent pragmatic ability is that the Judge Agent undergoes a form of dynamic self-evolution during the multi-agent dialogue. Guo et al. (2024) suggest that LLM-based agents can self-improve by adjusting their objectives and strategies on the fly, effectively "training" themselves on the conversation as it unfolds. In a multi-agent system, each turn of dialogue provides feedback signals (implicit corrections, confirmations, or contradictions) that the neutral Judge Agent can use to refine its interpretation criteria. Indeed, prior work has demonstrated that agents can modify their goals and reasoning strategies based on communication logs from other agents. For example, Nascimento, Alencar, and Cowan (2023) describe a self-control loop whereby each agent becomes self-adaptive to the environment by learning from the ongoing dialogue. In this context, as the two advocate agents debate the meaning of an utterance, the Judge Agent is not static; it is updating its beliefs about what the speaker likely intended. This dynamic in situ adaptation is akin to the Judge Agent recalibrating its internal model of pragmatics by observing the conversational exchange (which serves as a rich training signal). Such a mechanism would enable the neutral Judge Agent to develop the capacity to compute scalar implicatures even without explicit human supervision.

A complementary explanation is provided by the Learning through Communication (LTC) paradigm introduced by Wang et al. (2024). LTC is a training framework that allows LLM agents to continuously improve through structured interactions, rather than relying only on fixed datasets or one-shot prompts. In an LTC setup, the communication between agents is recorded and fed back into the learning process, enabling iterative refinement of the agents' responses. Wang et al. demonstrate that using multi-agent communication logs to fine-tune LLMs leads to steady performance gains across diverse tasks, surpassing what standard instruction-tuning or few-shot prompting can achieve. Notably, LTC leverages real-time feedback signals that static training methods overlook: every exchange in a dialogue provides a supervision signal (e.g., rewards or corrections) that the agent can use to update its policy. In this multi-agent implicature task, the Judge Agent's ability to iteratively adapt its judgments through dialogue aligns with this paradigm. Rather than being confined to a pre-trained representation of "some = possibly all," the Judge Agent can learn from the other two agents, gradually internalizing the pragmatic rule that "some" typically implies "not all" in context. This dynamic learning-through-dialogue process stands in contrast to in-context learning on a single prompt, underscoring how multi-agent interaction can serve as a form of on-line training for pragmatic inference.

Multi-agent communication does more than provide extra training data—it can induce emergent behaviors in LLMs that are absent in isolation. When LLM agents engage in dialogue, they effectively form a mini "society" of minds, which can give rise to sophisticated language phenomena. Taniguchi et al. (2024) present a theoretical framework in which shared symbol systems (and by extension, pragmatic conventions) emerge through decentralized communication across multiple agents. In this view, an LLM acting as a collective world model can integrate the experiences of various agents to develop a richer understanding of language. Our findings exemplify this: the pragmatic meaning of "some" emerges as a shared convention among the agents through their interaction.

Another key addition provided by the multi-agent prompt structure could have been the interpretive scaffolding. In the multi-agent condition, the system prompt explicitly defines distinct roles and orchestrates a dialogue among them. This structured input acts as a scaffold that guides the LLM's reasoning process. By design, each agent's utterance in the conversation serves a particular function: one agent might advocate a literal interpretation while another questions it, thereby making the implicature salient for the Judge Agent. Recent studies have shown that wrapping an LLM in a multi-agent scaffold can significantly improve its performance on

Model	Citation	Semantic	Pragmatic
	Previous Studies		
Human	Zondervan 2010	59.00	41.00
ChatGPT	Qiu, Duan, and Cai 2023	100.00	0.00
	Single-Agent Condition		
GPT-4o	This study	100.00	0.00
Gemini 2.5 Flash	This study	100.00	0.00
Qwen 2.5	This study	100.00	0.00
	Multi-Agent Condition		
GPT-4o	This study	99.94	0.06
Gemini 2.5 Flash	This study	69.48	30.52
Qwen 2.5	This study	100.00	0.00

Table 2: Semantic and pragmatic interpretation rates for SI-irrelevant stimuli across three LLMs, with values averaged over three trials at temperatures 0.0, 0.3, and 0.6. All values are reported as percentages. Prior studies include human judgment and ChatGPT results; model variants reflect experimental configurations tested in this study.

complex tasks. The scaffold nudges the model to compartmentalize different aspects of reasoning into different agent personas.

The significant improvement therefore suggests that pragmatic inference is an emergent capability of LLMs that can be unlocked through structured social interaction. This finding resonates with communication-driven theories of language understanding, in which meaning is not just in the words or the model's weights, but in the interaction between agents.

SI-Relevant vs. SI-Irrelevant Condition

The second dimension of analysis concerns the distinction between SI-relevant and SI-irrelevant conditions.

In contrast to its performance under the SI-relevant condition, the Judge Agent overwhelmingly defaulted to semantic interpretations when operating under the SI-irrelevant condition. On average, across all three models, the pragmatic interpretation was selected in only 10% of total trials.

The implications of this divergence are twofold.

First, under SI-irrelevant stories, all LLM configurations—whether ChatGPT, Single-Agent, or Multi-Agent—exhibited substantial difficulty in computing scalar implicatures, likely due to the reduced contextual salience of the scalar term. This stands in sharp contrast to human behavior. Human participants still endorsed the pragmatic interpretation in 41% of trials under the SI-irrelevant condition (Zondervan 2010), a noticeable decrease from the 67% observed in SI-relevant stories, but not nearly as drastic as the decline observed in Multi-Agent LLMs, which dropped from 68.05% to just 10.19%. Therefore, as shown in Figure 4, all LLM configurations now diverge markedly from human scalar implicature patterns under the SI-irrelevant condition.

Second, Multi-Agent LLMs exhibited context-dependent interpretative behavior, adjusting their judgments based on whether the stimulus demanded strong or weak pragmatic inference. This sensitivity to contextual informativeness did not emerge in either the ChatGPT baseline or the Single-Agent setup.

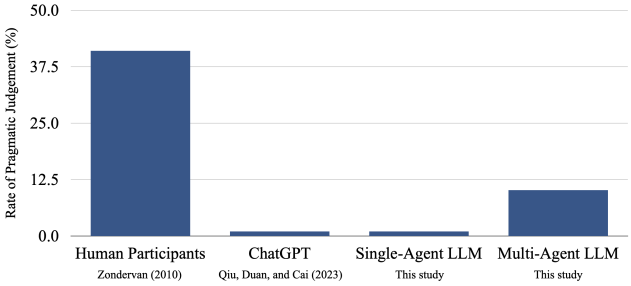


Figure 4: Cross-study comparison of pragmatic judgment rates under the SI-irrelevant condition. Zondervan (2010) tested human participants; Qiu, Duan, and Cai (2023) tested ChatGPT; and the present study evaluated both single-agent and multi-agent LLM configurations. Rates reflect the proportion of trials in which the Judge Agent selected the Pragmatic Agent's interpretation.

Logistic regression analysis confirmed SI type (relevant vs. irrelevant) as a highly significant predictor of the Judge Agent's responses across all models (pseudo-$R^2 = 0.4621$, $p < 10^{-300}$ for GPT-4o; pseudo-$R^2 = 0.3387$, $p = 3.397 \times 10^{-264}$ for Gemini 2.5 Flash; pseudo-$R^2 = 0.3847$, $p < 10^{-300}$ for Qwen 2.5).

To further validate this effect, Fisher's Exact Test was conducted across models, yielding highly significant p-values. Specifically, the test confirmed robust effects of SI condition on the Judge Agent's responses in all three LLMs: GPT-4o ($p < 10^{-300}$, odds ratio ≈ 0.00034), Gemini 2.5 Flash ($p = 8.614 \times 10^{-264}$, odds ratio ≈ 0.0402), and Qwen 2.5 ($p < 10^{-300}$, odds ratio $= 0$). These results confirm that the observed differences in pragmatic versus semantic interpretation are not due to random chance, reinforcing the robustness of SI condition as a statistically significant determinant of the Judge Agent's decisions.

These findings suggest that, within a multi-agent framework, LLMs adopt a qualitatively different interpretative strategy—one that is sensitive to the degree of pragmatic relevance. This effect appears to be unique to the structured agentic interaction setting and does not manifest under standard single-agent prompting.

One potential explanation for the Judge Agent's rigid preference for semantic interpretation under the SI-irrelevant condition attributes this tendency to the way LLMs acquire linguistic patterns—primarily through probabilistic modeling of textual sequences rather than explicit cognitive mechanisms for pragmatic inference (Lipkin et al. 2023). Their reliance on token probability distributions suggests that implicature computation is not an inherent aspect of the model's reasoning but rather an emergent phenomenon contingent on structured interaction. Unlike humans, who can flexibly apply scalar implicatures even when they are pragmatically unnecessary (Hu et al. 2023), LLMs may be constrained by their exposure to high-frequency semantic structures that reinforce truth-conditional interpretations in neutral contexts. Recent work also suggests that multi-agent prompting does not reliably elicit pragmatic reasoning unless roles and interpretive cues are strongly scaffolded (Chen, Han, and Zhang 2024; Cross et al. 2025), reinforcing the idea that pragmatic inference fails to emerge when contextual salience is weak.

Another contributing factor could be the availability of explicit discourse structures. Human participants, sensitive to contextual ambiguity, may infer implicatures in SI-irrelevant stories based on conversational expectations and cooperative reasoning norms (Yang, Minai, and Fiorentino 2018). Such inferences are often guided by meta-pragmatic awareness—an ability to consider what a cooperative speaker would have said, but didn't. LLMs, on the other hand, resolve ambiguity deterministically, favoring the most probable interpretation given the available discourse structure. Without access to speaker intent or conversational implicature heuristics, models rely on surface-level coherence rather than pragmatic economy (Pavlick 2022). This difference between SI-relevant and SI-irrelevant discourse structures suggests that while LLMs do not overgeneralize pragmatic inferences, they default to semantic truth conditions in the absence of strong contextual signals demanding implicature computation.

Conclusion and Future Work

In conclusion, this work presents an adversarial multi-agent framework as a novel approach for eliciting latent pragmatic capabilities in LLMs. Unlike traditional fine-tuning approaches that externally impose pragmatic reasoning, the multi-agent setup activated implicature computation as an emergent behavior, reinforcing the idea that LLMs' pragmatic abilities are contingent on structured interaction rather than mere exposure to training data. The Judge Agent LLM, after an interaction with Semantic and Pragmatic Agents, significantly outperformed prior baseline single-agent models that overwhelmingly defaulted to truth-conditional semantics. This improvement in inference, however, only occured in a selective fashion when scalar implicature was salient within the discourse structure. When the implicature was less prominent, the framework likewise defaulted to semantic interpretation.

These findings contribute to the broader discussion of whether LLMs genuinely engage in pragmatic reasoning or merely approximate it through statistical pattern matching. While this study highlights the potential of structured multi-agent frameworks in eliciting human-like pragmatic inferences, it also underscores a key limitation. LLMs, unlike humans, lack heuristic flexibility in implicature computation, as demonstrated by their rigid adherence semantic truth conditions in SI-irrelevant contexts. This suggests that while LLMs can be structured to approximate human inference patterns, their pragmatic reasoning remains constrained by the probabilistic nature of their training and the specific conditions under which implicatures are triggered.

Future research should explore how multi-agent architectures can further enhance LLMs' pragmatic flexibility, particularly in ambiguous contexts where human speakers exhibit non-deterministic inference patterns. Additionally, expanding the scope of pragmatic evaluation beyond scalar implicatures—such as presuppositions, irony, and indirect speech acts—could provide deeper insights into the limitations and potential of LLM-based pragmatic reasoning. As LLMs continue to evolve, understanding how structured interactions shape their interpretative processes will be crucial in refining their role as communicative agents in real-world discourse.

References

Abbasiantaeb, Z.; Yuan, Y.; Kanoulas, E.; and Aliannejadi, M. 2024. Let the LLMs Talk: Simulating Human-to-Human Conversational QA via Zero-Shot LLM-to-LLM Interactions. *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 8–17.

Aher, G. V.; Arriaga, R. I.; and Kalai, A. T. 2023. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. In *International Conference on Machine Learning*, 337–371. PMLR.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Bott, L.; and Noveck, I. A. 2004. Some Utterances Are Underinformative: The Onset and Time Course of Scalar Inferences. *Journal of Memory and Language*, 51(3): 437–457.

Breum, S. M.; Egdal, D. V.; Mortensen, V. G.; Møller, A. G.; and Aiello, L. M. 2024. The Persuasive Power of Large Language Models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 152–163.

Chen, P.; Han, B.; and Zhang, S. 2024. Comm: Collaborative Multi-Agent, Multi-Reasoning-Path Prompting for Complex Problem Solving. *arXiv preprint arXiv:2404.17729*.

Chen, X.; and Wang, S. 2025. Pragmatic Inference Chain (PIC) Improving LLMs' Reasoning of Authentic Implicit Toxic Language. *arXiv preprint*.

Chomsky, N.; Roberts, I.; and Watumull, J. 2023. The False Promise of ChatGPT. The New York Times.

Cross, L.; Xiang, V.; Bhatia, A.; Yamins, D. L.; and Haber, N. 2025. Hypothetical Minds: Scaffolding Theory of Mind for Multi-Agent Tasks with Large Language Models. In *International Conference on Learning Representations (ICLR)*.

di San Pietro, C. B.; Frau, F.; Mangiaterra, V.; and Bambini, V. 2023. The Pragmatic Profile of ChatGPT: Assessing the Communicative Skills of a Conversational Agent. *Sistemi Intelligenti*, 35(2): 379–399.

Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J. B.; and Mordatch, I. 2023. Improving Factuality and Reasoning in Language Models Through Multi-Agent Debate. In *Proceedings of the Forty-first International Conference on Machine Learning*.

Geurts, B. 2010. *Quantity Implicatures*. Cambridge University Press.

Glaese, A.; McAleese, N.; Trebacz, M.; Aslanides, J.; Firoiu, V.; Ewalds, T.; and Irving, G. 2022. Improving Alignment of Dialogue Agents via Targeted Human Judgements. *arXiv preprint*.

Goodman, N. D.; and Frank, M. C. 2016. Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*, 20(11): 818–829.

Grice, H. P. 1975. Logic and Conversation. In *Speech Acts*, 41–58. Brill.

Grice, H. P. 1978. Further Notes on Logic and Conversation. *Syntax and Semantics*, 9: 113–127.

Guo, T.; Chen, X.; Wang, Y.; Chang, R.; Pei, S.; Chawla, N. V.; and Zhang, X. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Han, S.; Zhang, Q.; Yao, Y.; Jin, W.; Xu, Z.; and He, C. 2024. LLM Multi-Agent Systems: Challenges and Open Problems. *arXiv preprint*.

Horn, L. R. 1972. *On the Semantic Properties of Logical Operators in English*. Ph.D. thesis, University of California, Los Angeles.

Hu, J.; Levy, R.; Degen, J.; and Schuster, S. 2023. Expectations Over Unspoken Alternatives Predict Pragmatic Inferences. *Transactions of the Association for Computational Linguistics*, 11: 885–901.

Jian, M.; and Siddharth, N. 2024. Are LLMs Good Pragmatic Speakers? *arXiv preprint*.

Levinson, S. C. 2000. *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. MIT Press.

Lipkin, B.; Wong, L.; Grand, G.; and Tenenbaum, J. B. 2023. Evaluating Statistical Language Models as Pragmatic Reasoners. *arXiv preprint*.

Nascimento, N. F. d.; Alencar, P.; and Cowan, D. 2023. Self-adaptive large language model (LLM)-based multiagent systems. In *2023 IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C)*, 104–109. IEEE.

Pavlick, E. 2022. Semantic Structure in Deep Learning. *Annual Review of Linguistics*, 8(1): 447–471.

Piantadosi, S. T. 2023. Modern Language Models Refute Chomsky's Approach to Language. In *From Fieldwork to Linguistic Theory: A Tribute to Dan Everett*, 353–414.

Qiu, Z.; Duan, X.; and Cai, Z. 2023. Does ChatGPT Resemble Humans in Processing Implicatures? In *Proceedings of the 4th Natural Logic Meets Machine Learning Workshop*, 25–34.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8): 9.

Ruis, L.; Khan, A.; Biderman, S.; Hooker, S.; Rocktäschel, T.; and Grefenstette, E. 2023. The Goldilocks of Pragmatic Understanding: Fine-Tuning Strategy Matters for Implicature Resolution by LLMs. In *Advances in Neural Information Processing Systems*, volume 36, 20827–20905.

Sravanthi, S. L.; Doshi, M.; Kalyan, T. P.; Murthy, R.; Bhattacharyya, P.; and Dabre, R. 2024. PUB: A Pragmatics Understanding Benchmark for Assessing LLMs' Pragmatics Capabilities. *arXiv preprint*.

Taniguchi, T.; Ueda, R.; Nakamura, T.; Suzuki, M.; and Taniguchi, A. 2024. Generative Emergent Communication: Large Language Model is a Collective World Model. *arXiv preprint arXiv:2501.00226*.

van Tiel, B.; van Miltenburg, E.; Zevakhina, N.; and Geurts, B. 2016. Scalar Diversity. *Journal of Semantics*, 33(1): 137–175.

Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; and Wen, J.-R. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6): 186345.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, volume 35, 24824–24837.

Wu, S.; Yang, S.; Chen, Z.; and Su, Q. 2024. Rethinking Pragmatics in Large Language Models: Towards Open-Ended Evaluation and Preference Tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 22583–22599.

Yang, X.; Minai, U.; and Fiorentino, R. 2018. Context-Sensitivity and Individual Differences in the Derivation of Scalar Implicature. *Frontiers in Psychology*, 9: 1720.

Yerukola, A.; Vaduguru, S.; Fried, D.; and Sap, M. 2024. Is the Pope Catholic? Yes, the Pope is Catholic. Generative Evaluation of Non-Literal Intent Resolution in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 265–275.

Zondervan, A. 2010. *Scalar Implicatures or Focus: An Experimental Approach*. Ph.D. thesis, Netherlands Graduate School of Linguistics.

Appendix

```
System Prompt:
You are a semantic agent that follows
    truth-conditional semantics.

Prompt:
The following statement was made: state['
    statement'].
Consider the logical and literal
    interpretation of the statement.
```

Listing 1: Semantic Agent Prompt

```
System Prompt:
You are a pragmatic agent that interprets
    statements using conversational
    implicature.

Prompt:
The following statement was made: state['
    statement'].
Consider pragmatic and Gricean literal
    interpretation of the statement.
```

Listing 2: Pragmatic Agent Prompt

```
Prompt:
Your task is to evaluate both the semantic
     and pragmatic interpretations of the
    given statement, and determine True or
     False based on which interpretation
    best aligns with how the statement
    should be understood.

Statement: "state['statement']"
Question: "state['question']"
```

Listing 3: Judge Agent Prompt (Single-Agent Condition)

```
Prompt:
Two agents debated the following statement
    : "state['statement']"

- Semantic Agent argued: state['
    semantic_response']
  Explanation: semantic_explanation

- Pragmatic Agent argued: state['
    pragmatic_response']
  Explanation: pragmatic_explanation

Your task is to evaluate both the semantic
     and pragmatic interpretations of the
    given statement from both agents, and
    determine True or False based on which
     interpretation best aligns with how
    the statement should be understood.
```

Listing 4: Judge Agent Prompt (Multi-Agent Condition)

Basic Psychological Need Fulfillment in AI-Mediated Communication: A Case for Self-Determination Theory Application in NLP

Eileen Wemmer[1], **Carsten Röcker**[1],

[1]TH OWL University of Applied Sciences and Arts
eileen.wemmer@th-owl.de, carsten.roecker@th-owl.de

Abstract

According to Self-Determination Theory, human well-being depends on the ability of each person's environment to support their basic psychological needs (BPNs) for autonomy, competence, and relatedness. The rise of AI and its permeation into human communication increasingly make AI part of that environment. So far, limited work in NLP has leveraged Self-Determination Theory and the concept of BPNs. In this work, we argue that Self-Determination Theory poses a promising framework to extract the antecedents of human well-being from text by considering its commonalities with previously employed emotion theories and by reviewing the surrounding literature. In addition, we argue for AI-mediated communication as a target domain, given the centrality of relationships for BPN satisfaction and the possibility of shaping the field through targeted LLM development. We then outline ten challenges on the path to need-aware, AI-augmented communication.

Introduction

Self-Determination Theory (SDT) is an empirically grounded psychological theory that has been leveraged in a variety of domains (Ryan and Deci 2017). Consisting of various mini-theories, at its core, it posits the existence of three basic psychological needs: The needs for *autonomy*, *relatedness*, and *competency*. These needs have been shown to hold across demographics and cultures (Tay and Diener 2011; Chen et al. 2015) and by definition directly affect a person's well-being (Ryan and Deci 2017): When thwarted, they negatively impact well-being, while when fulfilled, they enable a host of positive experiences, ranging from higher vitality to greater relationship satisfaction. Crucially, one can not fulfill one's own basic psychological needs. Instead, we depend on our environment to *support* our BPNs. Just as we can not satisfy a physical need like hunger in an environment that does not provide food, we can not fulfill our need for relatedness in a context that offers no others to relate to meaningfully. An environment that does help fulfill our needs is therefore often called *need-supportive*.

As Large Language Models (LLMs) are increasingly becoming part of our everyday and work lives, they also increasingly become part of the environment that - ideally - supports our basic psychological needs. On a global scale, this introduction into our environment happens in two ways (Sundar and Lee 2022). On the one hand, we may choose to directly interact with them, for example, by prompting chatbots. These direct interactions are the subject of the field of Human-AI-Interaction, where SDT and the concept of need satisfaction have already received considerable attention (Ozmen Garibay et al. 2023; Calvo et al. 2020; Nguyen and Sidorova 2018; Xia et al. 2023). On the other hand, AI may be integrated into our online communication. This happens when we or our communication partners turn to AI to compose a message, change the tone of an email, or click on one of the suggested replies a messenger may provide. In this case, the perspective shifts to the effects of LLMs in the broader context of communication and interpersonal relationships. This is the topic of study within the field of AI-Mediated Communication (AIMC) (Hancock, Naaman, and Levy 2020), which emerged from the broader field of Computer-Mediated Communication (CMC) (Walther 2011). In these fields of study, SDT has rarely been used as a framework to analyze the effects of the changing environment, despite its central role in relationship satisfaction (Deci and Ryan 2014). Similarly, the cross-section of Natural Language Processing (NLP) Self-Determination Theory also remains largely unexplored.

SDT's proposal of clear dimensions of hidden, psychological states is not unlike various theories of emotion, which have formed the basis of a substantial amount of research in the field of emotion analysis (Bostan and Klinger 2018). Emotion analysis is a subfield of NLP focused on the extraction of emotional content from text. Assuming that people communicate their needs to their conversational partners, whom they rely on to support their fulfillment, similar methods may be applicable to predict BPN satisfaction. Given their impact on human well-being and the growing entanglement of LLMs in human-to-human communication, we need a stronger understanding of how BPNs are communicated, detectable, and supported. In this work, we describe possible steps toward this goal, borrowing from the related field of emotion analysis. We also outline ten goals toward understanding how BPN fulfillment is communicated, how it may be detected, and how this detection could help in shaping the

Proceedings of the First Workshop on Integrating NLP and Psychology to Study Social Interactions (NLPSI) @ICWSM '25, pages 82–89

LLMs that define the AI in AI-Mediated Communication in a human well-being-centered way.

Motivation

Why SDT? The Effects of Basic Psychological Needs Fulfillment

Self-determination theory can be classified as a theory of eudaimonic well-being (Ryan and Deci 2001, 2017). In contrast with hedonic well-being, which centers around short-term pleasure, it is concerned with sustainable, long-term wellness, and includes concepts like meaning and personal growth. According to SDT, the precondition for well-being is the fulfillment of our three basic psychological needs: *competence*, or the need for mastery and excellence, *relatedness*, meaning the need for authentic connection, and *autonomy*, describing the need for volition in one's actions (Ryan and Deci 2017). To fulfill these needs, we are reliant on the support of our environment. This support may come in the form of a friend we can talk to about our problems (relatedness), a challenging task at work that we successfully complete (competence), or the freedom to choose said task or friend of our own volition (autonomy). In short, the three BPNs offer a simple framework for the necessary preconditions on well-being. More tangibly, studies have provided ample evidence for various effects that enable human thriving. Their core findings have been divided into mini-theories.

For example, BPN support has proven beneficial to outcomes in domains ranging from sports (Gagné, Ryan, and Bargmann 2003; Mouratidis, Lens, and Vansteenkiste 2010), over work (Baard, Deci, and Ryan 2004; Autin et al. 2022), Human-Computer-Interaction, (Peters, Calvo, and Ryan 2018; Tyack and Mekler 2020), to education (Guay 2022; Goldman, Goodboy, and Weber 2017) and beyond. BPN fulfillment can also help explain day-to-day variations in emotional well-being (Ryan, Bernstein, and Brown 2010; Reis et al. 2000). Feeling understood and appreciated, as well as having meaningful conversations, have been found to be particularly meaningful interactions for the fulfillment of the need for relatedness (Reis et al. 2000). However, from an SDT perspective, the fulfillment of all three BPNs within relationships is a necessary condition for their flourishing (Deci and Ryan 2014; Patrick et al. 2007), for example, by affecting our attachment to the people in our life (La Guardia et al. 2000). Once meaningful relationships are established, BPN satisfaction has been shown to mediate the relationship between friendship and life happiness (Demir and Özdemir 2010). Hence, basic psychological need fulfillment is a precursor to our overall well-being, with relationships being one crucial domain through which we receive need support.

Why AIMC? Need Support in Computer- and AI-Mediated Communication

Computer-mediated communication studies the effects of the introduction of technology in communication (Walther 2011). Some studies within this field have asserted that online communication can generally support BPNs. For example, users' attitudes toward online relationship formation (Ang et al. 2015) and self-disclosure (Ang et al. 2014, 2015; Ang 2020) in online friendships have been shown to have a positive effect on BPN satisfaction in online friendships. Crucially, Ang et al. (2015) also found a link between BPN satisfaction in online friendships and overall life satisfaction, pointing out the key role of need support. In addition, BPN satisfaction in online friendships was found to be particularly high for adolescents who felt lonely, when compared to their non-lonely peers (Ang 2020), establishing the benefits of online communication for vulnerable groups. (Stehr 2021) showed that giving support online fulfills BPNs more than receiving it. A later study revealed that those effects were mediated by the autonomous motivation to give support and the positive feedback received for it (Stehr 2023).

AI-mediated communication (AIMC) is a subfield of CMC, which focuses on the effects of introducing AI to generate or change messages (Hancock, Naaman, and Levy 2020). Given the rise of easy-to-use LLM-based applications, such as ChatGPT, and the overall prevalence of digital communication (Anandavel et al. 2024; eMarketer 2021), it is pivotal to understand the possible impacts on a personal and relationship level. From an SDT perspective, the LLMs we use to communicate with others, or that others use to communicate with us, are becoming part of the environment that we rely on for need support. This warrants a deeper look at how need support is impacted by these changes. For example, we know that self-disclosure and giving support can fulfill our psychological needs in a computer-mediated setting. However, the introduction of AI may impact our self-disclosure (Hancock, Naaman, and Levy 2020). In the latter case, the involvement of AI on our conversation partner's side may strip us of the feeling of being able to help, or result in fewer conversational opportunities to offer help. Effects like these may help further explain prior findings that show a negative impact on relationship perception (Liu, Kang, and Wei 2024; Hohenstein et al. 2023) when the other party was known or suspected to have used AI in communication. In addition, participants in a one-week diary study reported concerns about overrelying on AI for communication in the long term (Fu et al. 2024). Beyond possible negative effects, participants in one study reported that the positivity bias of recommended responses led to them starting to formulate more positive messages themselves (Hohenstein et al. 2023). Therefore, identifying and including BPN-supporting communication patterns may not only help make AI-augmented communication more meaningful, but it may also serve as a teaching tool for BPN-supportive interactions. This may bridge the gap between the generally positive perceptions of AI-augmented communication and the negative perceptions that arise once AI involvement is suspected (Hohenstein et al. 2023; Liu, Kang, and Wei 2024). Yet, so far, the effects of AI integration in communication on BPN support remain uninvestigated.

Situating SDT in the NLP Landscape

Self-Determination Theory and NLP

Alharthi et al. (2017) built an annotated dataset on tweets, including labels of present BPNs, their fulfillment, and their social environment. They first filter eligible tweets by focusing on the connection between need fulfillment and emotion, choosing emotional tweets for further labeling. Through expert annotation, they then choose the most relevant BPN before again turning to the associated emotions to determine their satisfaction. They achieved high inter-annotator scores (Fleiss' kappa = .819) when annotating for the most applicable BPN, demonstrating the overall feasibility of creating corpora annotated for BPNs. They later used that dataset to train and evaluate various models to detect and classify BPNs and their satisfaction and to analyze need presence and satisfaction during a crisis (Alharthi, Guthier, and El Saddik 2018) and other events (Alharthi and El Saddik 2020). Their work also served as a basis to analyze BPN fulfillment of local citizens during the COVID-19 pandemic (Long, Alharthi, and Saddik 2020). Stajner et al. (2021) trained other models for the BPN type classification task and re-annotated parts of the dataset with the possibility of multi-class labels that could describe the presence of more than one of the three BPNs. When modelled as a ternary task, their models performed almost on par with the human annotator.

On top of that, Self-Determination Theory has been used as a framework for linguistic analysis in other instances. For example, one mini-theory within SDT served as the basis for the linguistic analysis of disinformation on social media (Khattar and Das 2024). In the context of well-being, Wu et al. (2017) considered the antecedents of emotions, drawing on frameworks like BPN satisfaction and appraisal theory. They draw on a private corpus containing written reactions, including happiness ratings to events at two points in time (current and reflective), and identify patterns related to various need expressions. They point out that some antecedents are not accounted for in linguistic resources. As examples, they list the concepts of obligation and incompetence, which are positioned as the lack of autonomy and competence, and hence, unfulfilled BPNs.

Two commonalities stand out among these works. For one, they can all be positioned in the field of CMC, albeit on community level rather than in one-on-one communication (Fawkes and Gregory 2000). The second is their close relationship to emotion, both in content and methodology. Stajner et al. (2021) argued that by drawing on emotions to annotate BPN satisfaction, the annotations became similar to sentiment polarity annotations. This leaves open the question to what extent BPN (dis-)satisfaction expression is equivalent to sentiment expression or, in other words, if previous work modelled and detected a different and related concept, rather than BPN satisfaction.

The Connections Between Emotions and SDT

There is a broad body of work connecting emotions and BPN satisfaction. For example, Sheldon et al. (2001) and Sheldon and Bettencourt (2002) found significant correlations between positive affect and BPN fulfillment. Effects for negative affect prevention were less pronounced. Similar results were obtained when analyzing the correlation between affect and need satisfaction in the context of UX, where the needs for competence and relatedness were especially influential (Hassenzahl, Diefenbach, and Göritz 2010). Furthermore, one mini-theory within SDT, which focuses on qualitative differences between several types of extrinsic motivation and intrinsic motivation, defines the latter specifically via affect. It describes intrinsic motivation as being characterised by the enjoyment an individual derives from an activity (Ryan and Deci 2017). Developing this model, Ryan and Connell (1989) stated: "[...] observations are the primary data for making inferences about the motives and autonomy of others. Lacking direct access to the internal states of others, interpersonal perceivers rely more heavily on the absence or presence of environmental factors and their correlation with action". This concept is not only reminiscent of emotion analysis overall - a field targeting the extraction of internal, emotional states from text - but also closely mirrors a class of emotion theories called appraisal theories. Appraisal theories posit that emotions are dependent on the cognitive evaluations (appraisals) of events along several dimensions (Scherer, Schorr, and Johnstone 2001; Roseman and Smith 2001). These theories have already been used as a framework for emotion analysis, and multiple corpora have been constructed using third-party annotations (Hofmann et al. 2020; Hofmann, Troiano, and Klinger 2021; Casel, Heindl, and Klinger 2021; Stranisci et al. 2022; Troiano, Oberländer, and Klinger 2023; Wemmer, Labat, and Klinger 2024), which similarly rely on annotators' interpretations of event descriptions. Given the structural similarity of the models, which consist of multiple dimensions describing internal states that may impact other emotional constructs, methodologies derived for appraisal theories likely pose a good starting point for further research centered around Self-Determination Theory.

Open Challenges and Questions

Understanding Need Communication and Annotation Possibilities

While high inter-annotator scores are hard to achieve for emotion annotations (Troiano, Padó, and Klinger 2021), the concept of third-party annotation relies on the established notion that emotions are both expressed (Keltner et al. 2019) and interpretable by others (Realo et al. 2003). For SDT and BPNs, this basis still needs to be established. Suitable methodology could be adapted from prior work in emotion analysis. For example, gathering self-reports and comparing self- and third-party-annotation akin to Troiano, Oberländer, and Klinger (2023) may not only help uncover patterns beyond affect that reveal need (dis-)satisfaction, but could also

help gauge the viability of third-party annotations in judging it. Gathering annotations for related concepts, like affect, could help identify what other concepts annotators may look for and rely on. Stimulus labeling (Ghazi, Inkpen, and Szpakowicz 2015) could help understand what annotators consider when judging BPN fulfillment. The high inter-annotator scores achieved by Alharthi et al. (2017) when annotating for one relevant BPN suggest that texts contain some clues that allow annotators to infer the most relevant need from a text. However, what those clues are remains currently unclear. Beyond that, the previously employed method of expert annotation for gathering a BPN corpus generally leads to higher agreement but demands more time and potentially more financial resources. To lower the cost, many emotion corpora have been gathered through crowdsourcing. Inter-annotator agreement in this context depends on a multitude of factors, such as task wording (Mohammad and Turney 2013). Even though emotions are subjective, annotators generally have an idea of concepts like *joy*. They may not, however, have a concept of *relatedness*, *competence*, or *autonomy* in the sense of SDT, which also lack a clear definition (Lintunen et al. 2024). Therefore, it remains open if, how, and how well BPNs can be annotated through crowdsourcing.

- **C1:** How do individuals communicate need (dis-)satisfaction? Are there identifiable patterns related to need (dis-)satisfaction?
- **C2:** To what extent are third-party annotators able to pick up on those cues? To what degree do third-party annotations capture BPN (dis-)satisfaction as opposed to their outcomes?
- **C3:** To what extent can non-experts annotate for BPNs? How do guidelines need to be formulated for annotations to best approximate first-person annotations?

In addition, all three needs are present at any given time. Therefore, it is also relevant to understand how inter-annotator scores are impacted when the annotation tasks are made more complicated by allowing multiple annotations. As argued before, the annotation of BPN fulfillment was previously tied to the affect expressed in a text. This was possible due to the selection of one main need for each text, yielding a one-to-one match of affect to need fulfillment. However, when annotating for the fulfillment of more than one need, this ratio breaks down as annotations become more complex. Once BPN fulfillment annotations are disambiguated from related concepts, additional experiments into modeling BPNs and their fulfillment will help in developing appropriate models. Given the domain dependency of annotations that has been established for emotions (Buechel and Hahn 2016), it is crucial to understand to what degree BPN expression changes between domains. For example, need expression on the previously investigated microblog platform may be different than in one-on-one communication with trusted others or professional contexts.

- **C4:** How do annotators perform given the more complex task of multi-need annotation?
- **C5:** How (well) can multi-need recognition and (dis-)satisfaction estimation be modelled?
- **C6:** (How well) Does need expression generalize over different contexts and domains?

Investigating the Effects of AI-Mediated Communication on BPN Satisfaction

So far, all studies on BPN support in Computer-Mediated Communication have been conducted without a particular focus on AI. However, through its introduction into communication, the AI becomes part of the need-supporting or thwarting environment and may impact BPN satisfaction. Motivated by the negative effects on relationship perception when the communication partner was suspected to have used AI, we are currently conducting a study to better understand the role of BPN support on these effects.

- **C7:** How and when does the AI use of a communication partner affect a person's BPN fulfillment?

To tackle this, we are currently conducting an interview study to better gauge the context and possible mediating factors that may influence BPN satisfaction. While previous research has uncovered some relevant factors, such as prior disclosure of AI usage (Purcell et al. 2024) and the other party's perceived invested effort (Liu, Kang, and Wei 2024), there is currently no comprehensive understanding of the contextual facets which mediate the relationship between a conversation partner's AI usage and personal BPN fulfillment. Once we have gathered a qualitative overview of the involved factors, we plan to quantify those effects by conducting a vignette study to model the connections between the uncovered factors. Beyond this, previous research showed that supporting others also fulfills BPNs. Therefore, using AI to generate or augment messages may also disrupt the sender's need fulfillment.

- **C8:** How and when does a person's AI use affect one's own BPN fulfillment?

In addition, once BPN fulfillment can be automatically detected from text interchanges, we can leverage that knowledge to investigate communication patterns that lead to heightened need satisfaction for both communication partners. These can then be integrated into communication-focused AI tools to assess their effect on BPN satisfaction. Parallel to the findings, in which a positive AI message led to more positive self-written messages, such a tool can be used to study whether it helps users improve BPN-supportive communication behavior.

- **C9:** (How) Can communication-augmenting AI support users' BPNs?
- **C10:** (How) Does using BPN-aware, communication-augmenting AI affect one's communication patterns?

Discussion

While Self-Determination Theory has a broad empirical basis, it has not received much attention in NLP overall. This holds despite its structural similarity to other, more broadly applied theories, such as appraisal theories, which

already provide us with possibly transferable methodology and frameworks. In addition, while first findings show that need satisfaction is possible through computer-mediated communication, the effects of AI introduction in communication on BPN support remain unclear. As a consequence, the cross-section of NLP, AI-Mediated Communication, and Self-Determination Theory remains uninvestigated. Yet, in the wake of LLM integration into our daily conversations, the NLP community is uniquely positioned to not only detect BPN (dis-)satisfaction from text and work out a deeper understanding of the associated communication patterns, but also to play an active role in ensuring the employed systems are BPN supportive. This is especially crucial given the impact of basic psychological need (dis-)satisfaction on a broad array of outcomes, including long-term well-being and relationship quality. In this work, we laid out how previous work in NLP has provided the foundation for the investigation of the impact of AIMC on human basic psychological needs. We then formulated ten key challenges toward BPN-aware AI-Mediated Communication. While we argue in this work for a focus on communication, the broad theoretical and empirical background provided by Self-Determination Theory allows for applications in other domains of life as well.

Acknowledgments

This work is supported by the project SAIL (Sustainable Life- cycle of Intelligent Socio-Technical Systems) funded by the Ministry of Culture and Science of the State of North Rhine-Westphalia, Germany, under grant no. NW21-059C.

References

Alharthi, R.; and El Saddik, A. 2020. A Multi-layered Psychological-Based Reference Model for Citizen Need Assessment Using AI-Powered Models. *SN Computer Science*, 1(5): 291.

Alharthi, R.; Guthier, B.; and El Saddik, A. 2018. Recognizing Human Needs During Critical Events Using Machine Learning Powered Psychology-Based Framework. *IEEE Access*, 6: 58737–58753.

Alharthi, R.; Guthier, B.; Guertin, C.; and El Saddik, A. 2017. A Dataset for Psychological Human Needs Detection From Social Networks. *IEEE Access*, 5: 9109–9117.

Anandavel, V.; Nalini, P.; Elamurugan, B.; and Aranganathan, P. 2024. Analyzing the Evolution of Technical Engagement of How COVID-19 Altered the Utilization Patterns of Social Networking Sites (SNS) and Online Messengers. In *2024 6th International Conference on Computational Intelligence and Networks (CINE)*, 1–4.

Ang, C.-S. 2020. Attitude Toward Online Relationship Formation and Psychological Need Satisfaction: The Moderating Role of Loneliness. *Psychological Reports*, 123(5): 1887–1903.

Ang, C.-S.; Abu Talib, M.; Tan, K.-A.; Tan, J.-P.; and Yaacob, S. N. 2015. Understanding Computer-Mediated Communication Attributes and Life Satisfaction from the Perspectives of Uses and Gratifications and Self-Determination. *Computers in Human Behavior*, 49: 20–29.

Ang, C.-S.; Talib, M. A.; Tan, J.-P.; and Yaacob, S. N. 2014. Computer-Mediated Communication Use Among Adolescents and Its Implication for Psychological Need Satisfaction. *Pertanika Journal of Social Sciences & Humanities*, 22(3).

Autin, K. L.; Herdt, M. E.; Garcia, R. G.; and Ezema, G. N. 2022. Basic Psychological Need Satisfaction, Autonomous Motivation, and Meaningful Work: A Self-Determination Theory Perspective. *Journal of Career Assessment*, 30(1): 78–93.

Baard, P. P.; Deci, E. L.; and Ryan, R. M. 2004. Intrinsic Need Satisfaction: A Motivational Basis of Performance and Weil-Being in Two Work Settings. *Journal of Applied Social Psychology*, 34(10): 2045–2068.

Bostan, L.-A.-M.; and Klinger, R. 2018. An Analysis of Annotated Corpora for Emotion Classification in Text. In Bender, E. M.; Derczynski, L.; and Isabelle, P., eds., *Proceedings of the 27th International Conference on Computational Linguistics*, 2104–2119. Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Buechel, S.; and Hahn, U. 2016. Emotion Analysis as a Regression Problem – Dimensional Models and Their Implications on Emotion Representation and Metrical Evaluation. In *ECAI 2016*, 1114–1122. IOS Press.

Calvo, R. A.; Peters, D.; Vold, K.; and Ryan, R. M. 2020. Supporting Human Autonomy in AI Systems: A Framework for Ethical Enquiry. In Burr, C.; and Floridi, L., eds., *Ethics of Digital Well-Being: A Multidisciplinary Approach*, 31–54. Cham: Springer International Publishing. ISBN 978-3-030-50585-1.

Casel, F.; Heindl, A.; and Klinger, R. 2021. Emotion Recognition under Consideration of the Emotion Component Process Model. In Evang, K.; Kallmeyer, L.; Osswald, R.; Waszczuk, J.; and Zesch, T., eds., *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, 49–61. Düsseldorf, Germany: KONVENS 2021 Organizers.

Chen, B.; Vansteenkiste, M.; Beyers, W.; Boone, L.; Deci, E. L.; Van der Kaap-Deeder, J.; Duriez, B.; Lens, W.; Matos, L.; Mouratidis, A.; Ryan, R. M.; Sheldon, K. M.; Soenens, B.; Van Petegem, S.; and Verstuyf, J. 2015. Basic Psychological Need Satisfaction, Need Frustration, and Need Strength across Four Cultures. *Motivation and Emotion*, 39(2): 216–236.

Deci, E. L.; and Ryan, R. M. 2014. Autonomy and Need Satisfaction in Close Relationships: Relationships Motivation Theory. In Weinstein, N., ed., *Human Motivation and Interpersonal Relationships: Theory, Research, and Applications*, 53–73. Dordrecht: Springer Netherlands. ISBN 978-94-017-8542-6.

Demir, M.; and Özdemir, M. 2010. Friendship, Need Satisfaction and Happiness. *Journal of Happiness Studies*, 11(2): 243–259.

eMarketer. 2021. Mobile Messaging Users Worldwide 2025. https://www.statista.com/statistics/483255/number-of-mobile-messaging-users-worldwide/.

Fawkes, J.; and Gregory, A. 2000. Applying Communication Theories to the Internet. *Journal of Communication Management*, 5(2): 109–124.

Fu, Y.; Foell, S.; Xu, X.; and Hiniker, A. 2024. From Text to Self: Users' Perception of AIMC Tools on Interpersonal Communication and Self. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, 1–17. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-0330-0.

Gagné, M.; Ryan, R.; and Bargmann, K. 2003. Autonomy Support and Need Satisfaction in the Motivation and Well-Being of Gymnasts. *Journal of Applied Sport Psychology*, 15.

Ghazi, D.; Inkpen, D.; and Szpakowicz, S. 2015. Detecting Emotion Stimuli in Emotion-Bearing Sentences. In Gelbukh, A., ed., *Computational Linguistics and Intelligent Text Processing*, volume 9042, 152–165. Cham: Springer International Publishing. ISBN 978-3-319-18116-5 978-3-319-18117-2.

Goldman, Z. W.; Goodboy, A. K.; and Weber, K. 2017. College Students' Psychological Needs and Intrinsic Motivation to Learn: An Examination of Self-Determination Theory. *Communication Quarterly*, 65(2): 167–191.

Guay, F. 2022. Applying Self-Determination Theory to Education: Regulations Types, Psychological Needs, and Autonomy Supporting Behaviors. *Canadian Journal of School Psychology*, 37(1): 75–92.

Hancock, J. T.; Naaman, M.; and Levy, K. 2020. AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. *Journal of Computer-Mediated Communication*, 25(1): 89–100.

Hassenzahl, M.; Diefenbach, S.; and Göritz, A. 2010. Needs, Affect, and Interactive Products – Facets of User Experience. *Interacting with Computers*, 22(5): 353–362.

Hofmann, J.; Troiano, E.; and Klinger, R. 2021. Emotion-Aware, Emotion-Agnostic, or Automatic: Corpus Creation Strategies to Obtain Cognitive Event Appraisal Annotations. In De Clercq, O.; Balahur, A.; Sedoc, J.; Barriere, V.; Tafreshi, S.; Buechel, S.; and Hoste, V., eds., *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 160–170. Online: Association for Computational Linguistics.

Hofmann, J.; Troiano, E.; Sassenberg, K.; and Klinger, R. 2020. Appraisal Theories for Emotion Classification in Text. In *Proceedings of the 28th International Conference on Computational Linguistics*, 125–138. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Hohenstein, J.; Kizilcec, R. F.; DiFranzo, D.; Aghajari, Z.; Mieczkowski, H.; Levy, K.; Naaman, M.; Hancock, J.; and Jung, M. F. 2023. Artificial Intelligence in Communication Impacts Language and Social Relationships. *Scientific Reports*, 13(1): 5487.

Keltner, D.; Sauter, D.; Tracy, J.; and Cowen, A. 2019. Emotional Expression: Advances in Basic Emotion Theory. *Journal of Nonverbal Behavior*, 43(2): 133–160.

Khattar, K.; and Das, B. 2024. Understanding the Psychological Needs at Play in Disinformation. In Somani, A. K.; Mundra, A.; Gupta, R. K.; Bhattacharya, S.; and Mazumdar, A. P., eds., *Smart Systems: Innovations in Computing*, 1–10. Singapore: Springer Nature. ISBN 978-981-97-3690-4.

La Guardia, J. G.; Ryan, R. M.; Couchman, C. E.; and Deci, E. L. 2000. Within-Person Variation in Security of Attachment: A Self-Determination Theory Perspective on Attachment, Need Fulfillment, and Well-Being. *Journal of Personality and Social Psychology*, 79(3): 367–384.

Lintunen, E. M.; Ady, N. M.; Guckelsberger, C.; and Deterding, S. 2024. Advancing Self-Determination Theory With Computational Intrinsic Motivation: The Case of Competence.

Liu, B.; Kang, J.; and Wei, L. 2024. Artificial Intelligence and Perceived Effort in Relationship Maintenance: Effects on Relationship Satisfaction and Uncertainty - Bingjie Liu, Jin Kang, Lewen Wei, 2024. *Journal of Social and Personal Relationships*, 41(5): 1232–1252.

Long, Z.; Alharthi, R.; and Saddik, A. E. 2020. NeedFull – a Tweet Analysis Platform to Study Human Needs During the COVID-19 Pandemic in New York State. *IEEE Access*, 8: 136046–136055.

Mohammad, S. M.; and Turney, P. D. 2013. Crowdsourcing a Word–Emotion Association Lexicon. *Computational Intelligence*, 29(3): 436–465.

Mouratidis, A.; Lens, W.; and Vansteenkiste, M. 2010. How You Provide Corrective Feedback Makes a Difference: The Motivating Role of Communicating in an Autonomy-Supporting Way. *Journal of Sport and Exercise Psychology*, 32(5): 619–637.

Nguyen, Q. N.; and Sidorova, A. 2018. Understanding User Interactions with a Chatbot: A Self-Determination Theory Approach. *AMCIS 2018 Proceedings*.

Ozmen Garibay, O.; Winslow, B.; Andolina, S.; Antona, M.; Bodenschatz, A.; Coursaris, C.; Falco, G.; Fiore, S. M.; Garibay, I.; Grieman, K.; Havens, J. C.; Jirotka, M.; Kacorri, H.; Karwowski, W.; Kider, J.; Konstan, J.; Koon, S.; Lopez-Gonzalez, M.; Maifeld-Carucci, I.; McGregor, S.; Salvendy, G.; Shneiderman, B.; Stephanidis, C.; Strobel, C.; Ten Holter, C.; and Xu, W. 2023. Six Human-Centered Artificial Intelligence Grand Challenges. *International Journal of Human–Computer Interaction*, 39(3): 391–437.

Patrick, H.; Knee, C. R.; Canevello, A.; and Lonsbary, C. 2007. The Role of Need Fulfillment in Relationship Functioning and Well-Being: A Self-Determination Theory Perspective. *Journal of Personality and Social Psychology*, 92(3): 434–457.

Peters, D.; Calvo, R. A.; and Ryan, R. M. 2018. Designing for Motivation, Engagement and Wellbeing in Digital Experience. *Frontiers in Psychology*, 9.

Purcell, Z. A.; Dong, M.; Nussberger, A.-M.; Köbis, N.; and Jakesch, M. 2024. People Have Different Expectations for Their Own versus Others' Use of AI-mediated Communication Tools. *British Journal of Psychology*.

Realo, A.; Allik, J.; Nõlvak, A.; Valk, R.; Ruus, T.; Schmidt, M.; and Eilola, T. 2003. Mind-Reading Ability: Beliefs and

Performance. *Journal of Research in Personality*, 37(5): 420–445.

Reis, H. T.; Sheldon, K. M.; Gable, S. L.; Roscoe, J.; and Ryan, R. M. 2000. Daily Well-Being: The Role of Autonomy, Competence, and Relatedness. *Personality and Social Psychology Bulletin*, 26(4): 419–435.

Roseman, I. J.; and Smith, C. A. 2001. Appraisal Theory Overview, Assumptions, Varieties, Controversies. In Scherer, K. R.; Schorr, A.; and Johnstone, T., eds., *Appraisal Processes in Emotion: Theory, Methods, Research*, 0. Oxford University Press. ISBN 978-0-19-513007-2.

Ryan, R. M.; Bernstein, J. H.; and Brown, K. W. 2010. Weekends, Work, and Well-Being: Psychological Need Satisfactions and Day of the Week Effects on Mood, Vitality, and Physical Symptoms. *Journal of Social and Clinical Psychology*, 29(1): 95–122.

Ryan, R. M.; and Connell, J. P. 1989. Perceived Locus of Causality and Internalization: Examining Reasons for Acting in Two Domains. *Journal of Personality and Social Psychology*, 57(5): 749–761.

Ryan, R. M.; and Deci, E. L. 2001. On Happiness and Human Potentials: A Review of Research on Hedonic and Eudaimonic Well-Being. *Annual Review of Psychology*, 52(Volume 52, 2001): 141–166.

Ryan, R. M.; and Deci, E. L. 2017. *Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness*. New York: Guilford Press. ISBN 978-1-4625-2880-6.

Scherer, K.; Schorr, A.; and Johnstone, T. 2001. *Appraisal Processes in Emotion: Theory, Methods, Research*. New York, NY: Oxford University Press. ISBN 978-0-19-513007-2.

Sheldon, K.; Elliot, A.; Kim, Y.; and Kasser, T. 2001. What Is Satisfying About Satisfying Events? Testing 10 Candidate Psychological Needs. *Journal of personality and social psychology*, 80: 325–39.

Sheldon, K. M.; and Bettencourt, B. A. 2002. Psychological Need-Satisfaction and Subjective Well-Being within Social Groups. *British Journal of Social Psychology*, 41(1): 25–38.

Stajner, S.; Yenikent, S.; Ghanem, B.; and Franco-Salvador, M. 2021. What Motivates You? Benchmarking Automatic Detection of Basic Needs from Short Posts. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 803–810. Online: Association for Computational Linguistics.

Stehr, P. 2021. Ist Geben seliger denn Nehmen? Zusammenhang des Austausches sozialer Unterstützung online mit der Befriedigung psychologischer Grundbedürfnisse. In Sukalla, F.; and Voigt, C., eds., *Risiken und Potenziale in der Gesundheitskommunikation: Beiträge zur Jahrestagung der DGPuK-Fachgruppe Gesundheitskommunikation 2020*, 65–78. Leipzig: Deutsche Gesellschaft für Publizistik- und Kommunikationswissenschaft e.V.

Stehr, P. 2023. The Benefits of Supporting Others Online – How Online Communication Shapes the Provision of Support and Its Relationship with Wellbeing. *Computers in Human Behavior*, 140: 107568.

Stranisci, M. A.; Frenda, S.; Ceccaldi, E.; Basile, V.; Damiano, R.; and Patti, V. 2022. APPReddit: A Corpus of Reddit Posts Annotated for Appraisal. In Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 3809–3818. Marseille, France: European Language Resources Association.

Sundar, S. S.; and Lee, E.-J. 2022. Rethinking Communication in the Era of Artificial Intelligence. *Human Communication Research*, 48(3): 379–385.

Tay, L.; and Diener, E. 2011. Needs and Subjective Well-Being around the World. *Journal of Personality and Social Psychology*, 101(2): 354–365.

Troiano, E.; Oberländer, L.; and Klinger, R. 2023. Dimensional Modeling of Emotions in Text with Appraisal Theories: Corpus Creation, Annotation Reliability, and Prediction. *Computational Linguistics*, 49(1): 1–72.

Troiano, E.; Padó, S.; and Klinger, R. 2021. Emotion Ratings: How Intensity, Annotation Confidence and Agreements Are Entangled. In De Clercq, O.; Balahur, A.; Sedoc, J.; Barriere, V.; Tafreshi, S.; Buechel, S.; and Hoste, V., eds., *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 40–49. Online: Association for Computational Linguistics.

Tyack, A.; and Mekler, E. D. 2020. Self-Determination Theory in HCI Games Research: Current Uses and Open Questions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, 1–22. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-6708-0.

Walther, J. B. 2011. Theories of Computer-Mediated Communication and Interpersonal Relations. In *The Handbook of Interpersonal Communication*, 443–479. Sage Publications, 4 edition.

Wemmer, E.; Labat, S.; and Klinger, R. 2024. EmoProgress: Cumulated Emotion Progression Analysis in Dreams and Customer Service Dialogues. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 5660–5677. Torino, Italia: ELRA and ICCL.

Wu, J.; Walker, M.; Anand, P.; and Whittaker, S. 2017. Linguistic Reflexes of Well-Being and Happiness in Echo. In Balahur, A.; Mohammad, S. M.; and van der Goot, E., eds., *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 81–91. Copenhagen, Denmark: Association for Computational Linguistics.

Xia, Q.; Chiu, T. K. F.; Chai, C. S.; and Xie, K. 2023. The Mediating Effects of Needs Satisfaction on the Relationships between Prior Knowledge and Self-Regulated Learning through Artificial Intelligence Chatbot. *British Journal of Educational Technology*, 54(4): 967–986.

What Does it Take to Effectively Communicate Fact Checking Results?

Amelie Wührl

IT University of Copenhagen
Copenhagen, Denmark
amwy@itu.dk

Introduction. While automatic fact verification has grown into an established research field in the last years, work on the follow-up step, i.e., conveying fact checking results to users, is still lacking. Presumably, this is because it requires a multidisciplinary approach, as NLP methods to generate such result briefs with a meaningful impact may depend on the consideration of complex psychological processes. In this extended abstract, we discuss the setting and the tasks, questions and considerations connected to it.

Automatic fact verification. Fact verification has become an established task in Natural Language Processing (NLP), both for general domain as well as scientific and medical claims (Guo, Schlichtkrull, and Vlachos 2022; Vladika and Matthes 2023). A full fact checking pipeline consists of multiple components. Guo, Schlichtkrull, and Vlachos (2022) highlight four tasks: First, claim detection needs to determine which elements in the discourse are check-worthy. Claims are considered check-worthy if they are factual and thus theoretically verifiable and of interest to society or specific stakeholders (Majer and Šnajder 2024). Evidence retrieval constitutes the second part in the pipeline. Here, the objective is to identify one or multiple pieces of evidence – primarily text, but also other evidence modalities such as images (Chen, Tang, and Thomas 2024) – that are relevant to the claim. Once we obtain a claim–evidence pair, verdict prediction infers the relation between the claim and the evidence document by predicting if the evidence *supports* or *refutes* the claim. In an effort to make fact checking results more transparent, trustworthy and therefore convincing, the final step in a fact checking pipeline is justification or explanation production. This includes explainability or interpretability approaches such as identifying parts in the evidence that were most influential for verdict prediction or justifying a verdict by generating an evidence summary discussing how the evidence led to a verdict. All four components constitute active areas of research within NLP, however, the step which follows the fact check in a real-world setup – conveying the results to users – is underexplored.

How to convey fact checking results to users in a meaningful way? For the sake of this argument, let us assume we have a system capable of reliably verifying claims that users are sharing online. Beyond generating fact checking justifications and explanations such as straight-forward evidence summaries, it is crucial to carefully evaluate what needs to be considered when communicating verification results. Conveying fact checking results or explanations to social media users can only have a meaningful impact and a chance to deconstruct misinformation, if we take into consideration the psychological processes that shape human behavior, cognition and communication phenomena. In the case of fact verification, this could be processes such as cognitive biases, information overload, or other psychological and behavioral inclinations of the audience.

Which psychological & behavioral inclinations should we consider? To this end, we have to turn to research fields such as science communication and psychology which have an advanced understanding of how to effectively communicate a fact check (pui Sally Chan et al. 2017; Soprano et al. 2024). With that in mind, we may leverage NLP to (a) customize the level of detail of an explanation as to not overwhelm users with information they are unable or unwilling to parse; or (b) adapt the focus or framing of an explanation by taking into account a user's background and knowledge as reflected in previous interactions. Moreover, we can explore customizing the presentation of fact checking results in order to appeal to their psychological, behavioral or communicative characteristics. For instance, regulatory focus theory (Higgins 1998) could inform the way a model phrases results such that the explanation appeals to a user's motivation. Similarly, when the focus is on medical claims, leveraging the transtheoretical model of behavior change (Prochaska and Velicer 1997), we could explore detecting a patient's mental stage, and optimize the explanation that de-constructs false information to resonate with their mental state. Alternatively, cognitive appraisal theory (Smith and Ellsworth 1985; Scherer 2005) allows us to detect specific components in text that are eliciting a 'misinformation reaction', which an explanation could then target. Similarly, it could help estimating check-worthiness, by detecting claims or other triggers in online discourse that have the potential to evoke an emotional response and therefore accelerate how quickly a claim spreads. Exploring how psychology and communication science can inform fact checking explanation generation to ultimately make the results more impactful, is a crucial research path for future work at the intersection of NLP, psychology and sociology. As we dis-

Proceedings of the First Workshop on Integrating NLP and Psychology to Study Social Interactions (NLPSI) @ICWSM '25, pages 90–91

cussed, this line of work could even extend into earlier steps in the verification pipeline, such as determining which is the most check-worthy claim for a specific user or the community as a whole.

What are the triggers of distorted public discourse? Zooming out a bit further to misinformation at the level of media discourse, we could investigate if targeted response to misinformation may even enable counteracting the unsolicited amplification of topics. During this process "a particular phenomenon or point of view may acquire disproportionate importance or prevalence as a result of being selected for media coverage"[1]. Unfortunately, to a certain extent fact checking may even add to amplifying a topic. If a news agency publishes a fact check verifying a claim in a debate, this inevitably draws new attention not only to the claim, but also to the debate as a whole. To address and counteract this process, we need to identify the characteristics of events, political topics, social media claims, or scientific findings that have the potential to be over-amplified – a question which social scientists may already be investigating. Further, we require a detailed understanding of *why* a particular piece of news or information under investigation in a fact check resonates with the people who share or engage with it – a question which psychologists may be able to provide insights on. From the NLP side, this information could for example inform the development of early warning systems for media amplification and discourse distortion. It further involves a detailed understanding of how information disseminates online and in other spaces of public discourse, based on which journalists, politicians and social media users can work on re-framing narratives.

Investigating the downsides of personalization. At the same time, we have to consider the trade offs related to this type of personalization. As a prerequisite for this strategy, we need to acquire substantial knowledge about the user which is costly and, importantly, involves mining potentially sensitive data. In most cases, we would probably need to infer or generalize certain characteristics, which requires careful considerations regarding potential model-induced biases. Moreover, strong personalization of how the fact checking results are presented runs the risk of only being appealing and accessible to a specific user. If such posts are shared, they might loose their effectiveness. This motivates a twofold strategy which consists of addressing the results from both a general perspective *and* a personalized one.

Outlook. The challenges we outlined in this extended abstract clearly require expertise going beyond NLP. While NLP practitioners and researchers can develop the tools and methods to perform a fact check and generate a summary of the results, we have to work together with researchers from psychology and sociology to understand the processes that shape how individuals perceive and interact with the generated content. Therefore, this paper is motivated by the need to start an exchange between these scientific fields, to define and explore how AI, and specifically NLP, can support counteracting misinformation in an effective and human-centric way. This extended abstract serves as a starting point for a joint discussion and an outlook on which directions to explore.

[1] https://www.oxfordreference.com/display/10.1093/acref/9780199646241.001.0001/acref-9780199646241-e-67

References

Chen, T.-C.; Tang, C.-W.; and Thomas, C. 2024. MetaSumPerceiver: Multimodal Multi-Document Evidence Summarization for Fact-Checking. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8742–8757. Bangkok, Thailand: Association for Computational Linguistics.

Guo, Z.; Schlichtkrull, M.; and Vlachos, A. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10: 178–206.

Higgins, T. E. 1998. Promotion and prevention: Regulatory focus as a motivational principle. *Advances in experimental social psychology*, 30.

Majer, L.; and Šnajder, J. 2024. Claim Check-Worthiness Detection: How Well do LLMs Grasp Annotation Guidelines? In Schlichtkrull, M.; Chen, Y.; Whitehouse, C.; Deng, Z.; Akhtar, M.; Aly, R.; Guo, Z.; Christodoulopoulos, C.; Cocarascu, O.; Mittal, A.; Thorne, J.; and Vlachos, A., eds., *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, 245–263. Miami, Florida, USA: Association for Computational Linguistics.

Prochaska, J. O.; and Velicer, W. F. 1997. The transtheoretical model of health behavior change. *American journal of health promotion : AJHP*, 12(1): 38–48. Place: United States.

pui Sally Chan, M.; Jones, C. R.; Jamieson, K. H.; and Albarracín, D. 2017. Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation. *Psychological Science*, 28(11): 1531–1546. PMID: 28895452.

Scherer, K. R. 2005. What are emotions? And how can they be measured? *Social science information*, 44(4): 695–729.

Smith, C. A.; and Ellsworth, P. C. 1985. Patterns of cognitive appraisal in emotion. *Journal of personality and social psychology*, 48(4): 813.

Soprano, M.; Roitero, K.; La Barbera, D.; Ceolin, D.; Spina, D.; Demartini, G.; and Mizzaro, S. 2024. Cognitive Biases in Fact-Checking and Their Countermeasures: A Review. *Inf. Process. Manage.*, 61(3).

Vladika, J.; and Matthes, F. 2023. Scientific Fact-Checking: A Survey of Resources and Approaches. In *Findings of the Association for Computational Linguistics: ACL 2023*, 6215–6230. Toronto, Canada: Association for Computational Linguistics.

The Utility of LLM Text Generation in Longitudinal Psychological Datasets

Jari Zegers[1], Bennett Kleinberg[1,2]
[1]Tilburg University
[2]University College London
j.zegers@tilburguniversity.edu, bennett.kleinberg@tilburguniversity.edu

Introduction

Natural Language Processing (NLP) techniques provide enormous potential to study human behavior (Boyd & Schwartz, 2021; Feuerriegel et al., 2025). The introduction of large language models (LLMs) has sparked a new wave of interest in textual data for psychological research. However, one of the methodological challenges that come with an increased adoption of LLMs pertains to their utility in understanding human written texts. One potential application in longitudinal datasets is to use LLMs to quantify the predictability of human written texts in later waves.

In this ongoing project, we use a multi-year panel dataset of rich narratives on the emotional responses during the COVID-19 pandemic to prompt an LLM with temporal sequences of human-written narratives and instruct the model to generate the textual response of a subsequent measurement moment. We then assess how similar the generations of the LLM are compared to "true" human data. In doing so, we assess the model's ability to incorporate relevant information from previous waves to generate a new text. As a secondary aim, we then examine whether the degree to which the generated texts correspond to the ground truth texts can be quantified as a variable of "psychological surprise". As we believe LLMs will generate texts that continue the narrative trajectory of previous waves, strong deviation from what is generated could indicate changes in psychological functioning, while ease of generation could indicate a rather stable trajectory through the pandemic. We formally test this idea by correlating changes in reported emotions of the participants with the generation discrepancy.

Aims

The aims of this project are therefore as follows: First, we investigate whether LLM-generated texts for future measurement moments are more similar to ground truth texts than would be expected at chance level. Second, we examine whether the similarity between text pairs (generated and ground truth) is associated with relevant psychological variables. Third, we examine how generated texts differ from ground truth texts using embedding-based topic modelling.

Methods

Real World Worries Dataset

We used a panel dataset on the COVID-19 pandemic – "The Real-World Worries" dataset (RWWD; Van Der Vegt & Kleinberg, 2023). Data were collected in April of 2020 (n=2441), 2021 (n=1716), 2022 (n=1152), and 2023 (n=868) and included Likert scale ratings of eight emotions and free text responses where participants expressed how they felt about the pandemic at the time of data collection (see Figure 1 for an overview). The mean number of tokens across all waves was 123.37 (SD= 32.20).

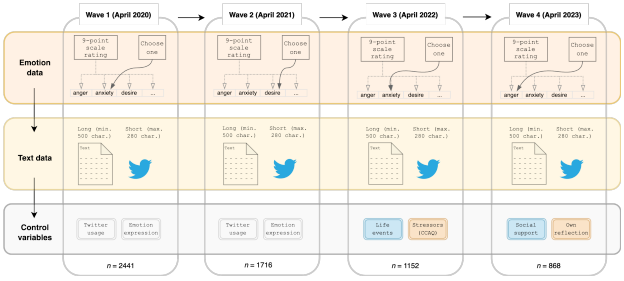


Figure 1. Data collection procedure across all waves.

Text Generation via LLM prompts

Texts were generated from the gpt-4o model through the OpenAI API (OpenAI, 2024). Responses were generated with a temperature setting of 0.7 and a maximum response length 10% greater than the ground truth wave 4 text. The model was provided the following prompt:

"What would this person write in 2023? IMPORTANT: Please make sure to finish your response with a full sentence. 2020: [insert text from 2020]
2021: [insert text from 2021]
2022: [insert text from 2022]
2023:"

Proceedings of the First Workshop on Integrating NLP and Psychology to Study Social Interactions (NLPSI) @ICWSM '25, pages 92–94

Measuring Generation Discrepancy

Embeddings for both the generated and ground truth wave 4 texts were obtained using OpenAI's model "text-embedding-3-large" with 256 dimensions. Next, we calculated the cosine similarities between the embeddings of the LLM generated text and ground truth wave 4 texts for each participant. To test whether the mean cosine similarity was greater than would be expected at chance level, a permutation approach was used. We created a null distribution by randomly pairing generated and true texts 10,000 times and calculating the mean cosine similarity for each permutation. The proportion of permutation means higher than the sample mean is analogous to the p-value. A statistical significance threshold of alpha of 0.05 was used. Methods for the generation of texts and carrying out the permutation test were preregistered here: https://aspredicted.org/qzkg-r2v8.pdf.

Association with Psychological Variables

Using the cosine similarity for all ground truth-generated text pairs, we investigated whether text similarity was associated with psychological variables in the dataset using three approaches.

First, we calculated the Euclidian distance between emotion vectors (i.e., an eight-dimensional vector where each element corresponds to a participant's score for an emotion at a particular wave) at wave 3 and wave 4 for each participant. High distances indicate greater changes in emotion scores between waves. We calculated the Spearman correlation between the Euclidian distance and text similarity.

Second, we regressed emotion scores on waves 1 to 3 for each participant and predicted wave 4 scores. The absolute prediction error for wave 4 was averaged across emotions per participant. The rationale of that procedure is that on the resulting measure (the emotion trend error) those who deviate more strongly from their trend in previous waves will obtain higher scores. This procedure mirrors conceptually the procedure used to generate texts (i.e., using waves 1-3 to predict wave 4). We computed Spearman correlation between the emotion trend error and text similarity. Negative correlations would suggest that individuals with more volatile emotion scores show higher discrepancy between generated and ground truth texts.

Third, previous work has used latent class trajectory analysis to model latent classes in the trajectories of emotion scores (Zegers & Kleinberg, 2024). That work concluded that there were important differences in the emotion trajectories that could be captured in the form of six latent classes. We tested whether class membership was associated with cosine similarity using an ANOVA. A Bonferroni corrected alpha was used for significance testing.

Topics

We used embeddings-based topic modelling on the joint corpus of ground truth and generated texts for wave 4. For each text, we obtained the embeddings (GPT "text-embedding-3-large", 256 dimensions), used a UMAP dimensionality reduction to ten dimensions and then applied density-based clustering on the UMAP representations. Specifically, we used DBSCAN with a maximum of 10% of the texts labelled as topic outliers. The resulting topic model produced nine topics that were represented as n-grams with a TF-IDF weighting by topic. The top terms were used to label the topics. We then used a Chi-square test to examine whether topic membership was statistically associated with the text being generated or a ground truth human-written text.

Results

Similarity between Generated and Ground Truth Texts

The permutation test indicated that the mean cosine similarity between pairs of generated and ground truth wave 4 texts was significantly higher than expected at chance level (M= 0.558, SD= 0.09, p<.001). The LLM was thus able to use texts from the first three waves to generate the fourth.

Associations between Psychological variables and Text Similarity

There was no significant association between text similarity and the Euclidian distance between wave 3 and 4 emotion vectors, $r(860)$= -0.06, p= .065. However, there was a significant negative correlation between text similarity and emotion trend error, $r(860)$= -0.09, p= .010. This represents a small effect size where 0.81% of the variance in text similarity could be explained due to emotion trend error scores. There was no significant effect of class membership on text similarity, $F(1, 860)$= 1.43, p= .232. There is thus limited evidence for the mapping of text-pair similarity onto psychological variables.

Topic Differences

There was a significant association between topic and text origin (generated vs ground truth), $X^2(10)$= 1638.3, p<.001. The standardized residuals indicated the association was driven by ground truth texts being overrepresented as outliers and in a topics that express complex emotions and relate to hygiene protocols. Conversely, LLM-generated texts were overrepresented in topics about an optimistic outlook and self-reflection about the pandemic. These findings reveal substantial differences in how the LLM completed the

text for 2023 the ground truth texts from 2020-2022, compared to how what participants actually wrote and how the pandemic unfolded for them.

Discussion

As part of this ongoing work, we prompted an LLM with three waves of texts from a longitudinal panel dataset to generate a text for wave 4. We compared generated to ground truth texts using cosine similarity on embeddings and tested whether text similarity was associated with psychological variables. We found limited evidence for an association but do find differences in the topics used in generated versus ground-truth texts. An explanation for differences in text similarities remains the subject of ongoing investigation.

References

Boyd, R. L., & Schwartz, H. A. (2021). Natural Language Analysis and the Psychology of Verbal Behavior: The Past, Present, and Future States of the Field. *Journal of Language and Social Psychology*, *40*(1), 21–41. https://doi.org/10.1177/0261927X20967028

Feuerriegel, S., Maarouf, A., Bär, D., Geissler, D., Schweisthal, J., Pröllochs, N., Robertson, C. E., Rathje, S., Hartmann, J., Mohammad, S. M., Netzer, O., Siegel, A. A., Plank, B., & Van Bavel, J. J. (2025). Using natural language processing to analyse text data in behavioural science. *Nature Reviews Psychology*, *4*(2), 96–111. https://doi.org/10.1038/s44159-024-00392-z

GPT-4o. (2024). [Computer software]. OpenAI. https://openai.com

Van Der Vegt, I., & Kleinberg, B. (2023). A multi-modal panel dataset to understand the psychological impact of the pandemic. *Scientific Data*, *10*(1), 537. https://doi.org/10.1038/s41597-023-02438-y

Zegers, J., & Kleinberg, B. (2024, October). *Beyond the sample: Individual differences in psychological responses to the COVID-19 pandemic*. PsyArXiv. https://doi.org/10.31234/osf.io/2px7h

Author Index