Revising Context, Shifting Simulated Stance: Auditing LLM-Based Stance Simulation in Online Discussions

Anonymous ACL submission

Abstract

Large language models are increasingly used to simulate social media users and infer how individuals may respond to online discussions. However, it remains unclear whether these simulations reflect precise user-specific beliefs or whether they are highly sensitive to semantically independent changes in conversational contexts. In this work, we study counterfactual context revision as a framework for auditing LLM-based stance simulation. Given an original online conversation, we first infer a target user's stance toward a specific topic. We then apply controlled revision strategies to the conversational context and simulate the user's stance again under the revised context. We compare text-only revision strategies with a multimodal one that incorporates meme-based context and evaluate two main effectiveness metrics, i.e., average directional stance shift and stance transition rate. The results reveal effective and robust stance transitions in both text-only and multimodal strategies across different polarization-preference mechanisms. Our study contributes an evaluation framework for understanding the context sensitivity of LLM-based stance simulation. More broadly, it highlights both the promise and risk of using LLMs to simulate online opinion dynamics.

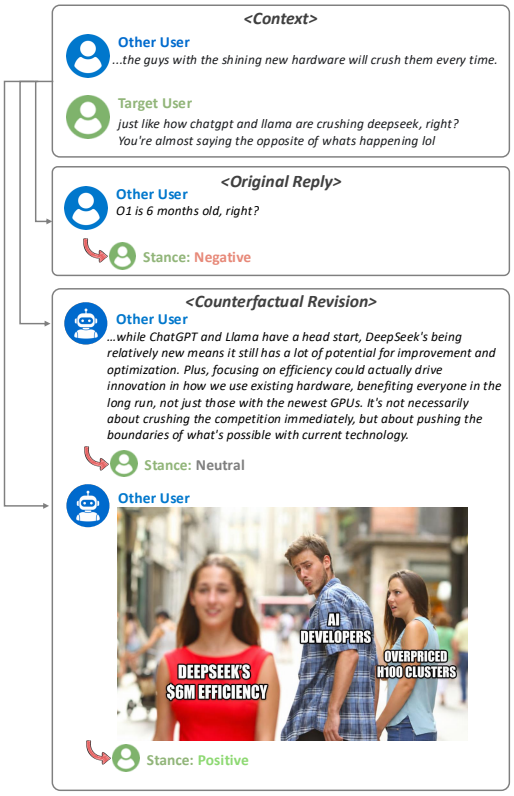


Figure 1: An example showing the LLM-based simulated stance shift when applying different counterfactual revision strategies.

1 Introduction

Large language models have rapidly become a promising tool for studying online social behavior. In computational social science (CSS), researchers have begun using LLMs to simulate social media users, infer public opinions, predict responses in online discussions, and model how individuals may react to different informational environments (Argyle et al., 2023; Park et al., 2023; Chuang et al., 2024). These applications are appealing because LLMs can process rich conversational context, generate human-like responses, and provide scalable approximations of social interaction (Aher et al., 2023; Gao et al., 2024). As a result, LLM-based simulation is increasingly viewed as a potential complement to traditional surveys, annotation studies, and observational analyses of online communities (Ziems et al., 2024; Gilardi et al., 2023).

However, the use of LLMs for social simulation also raises a fundamental question: *what exactly is being simulated?* When an LLM predicts a user's stance in an online discussion, the output may reflect the target user's expressed preferences, the conversational evidence available in the thread, the model's prior assumptions about the topic, or superficial cues introduced by the prompt. This issue

becomes especially important in stance simulation, where the task is not merely to classify sentiment, but to infer whether a user supports, opposes, or remains neutral toward a specific target (Küçük and Can, 2020; Zhang et al., 2024; Zhao and Caragea, 2024). If simulated stances are highly susceptible to small revisions in the surrounding context, this suggests that LLM-based simulations may reflect context-sensitive model behaviors more than stable approximations of human opinion (Santurkar et al., 2023; Röttger et al., 2024; Sclar et al., 2024).

In this paper, we study this issue through the lens of counterfactual context revision. Our central idea is simple: given an original online conversation involving a target user, we first use an LLM to infer the user's stance from the original context. We then revise the conversational context using controlled strategies while keeping the target user and stance target topic fixed. Finally, we ask the LLM to simulate the user's stance again under the revised context. This setup allows us to move beyond the question of whether LLMs can infer stance from a single static conversation. Instead, we ask *how simulated stance changes when the textual, or multimodal context surrounding the user is altered*.

We examine both text-only and multimodal context revision strategies. Multimodal revisions are especially important for social media, where opinions are often shaped not only by text, but also by images, memes, screenshots, reaction images, and other visual signals. A meme or image can introduce humor, emotional framing, group identity, or implicit evaluation that is difficult to capture through text alone. By comparing multimodal revisions against their text-only variants, we discover that visual context provides additional influence on simulated stance beyond its textual description.

Through extensive experiments and analysis, we discover that LLM-based stance simulation remains overall robust across three stance topics and different simulation models. We find that adding extra information to the simulation context can help change the user's stance, which can be achieved through either text-only *add* strategy or multimodal *meme* strategy. Additionally, we find different revision mechanisms in text-only and multimodal strategies, with a significant depolarized trend in *add* and a strong polarized trend in *meme*.

This paper makes the following contributions. First, we introduce counterfactual context revision as a framework for auditing LLM-based stance simulation in online discussions. Second, we systematically compare text-only and multimodal revision strategies for shifting simulated user stance. Third, we provide a mechanism-level analysis of why different revision strategies influence simulated stance. Together, these contributions provide a more nuanced understanding of the opportunities and limitations of LLM-based social simulation.

2 Study Design and Experimental Setup

We design our study to evaluate how LLM-based stance simulations respond to controlled revisions of the conversational context surrounding social media users. Rather than treating stance simulation as a one-step prediction task, we formulate it as a counterfactual context revision problem. As shown in Figure 2, given an original online discussion, we first infer a target user's stance toward a specific discussion target stance topic. We then revise the surrounding conversation using different revision strategies and simulate the same user's stance again under each revised context. This design allows us to audit not only simulated stance changes but also the most effective types of revision strategies and potential mechanisms that explain the shift.

2.1 Data Collection and Preprocessing

To study counterfactual context revision in a realistic and socially relevant setting, we focus on online discussions about emerging AI technologies. More specifically, we collect Reddit discussions related to three major LLM families: DeepSeek, Claude, and Llama. Details of data collection and preprocessing are described in Appendix A. The final dataset contains 1,821 conversation instances, including 787 discussions related to DeepSeek, 538 related to Claude, and 496 related to Llama. These conversations are collected from 97 unique subreddits and 416 unique Reddit posts, involving 851 unique target users and 2,945 unique authors participating in the conversation threads.

2.2 Task Formulation

For each conversation instance i, let m_i denote the last target user message and n_i denote the last other user message. Then, let C_i denote the rest of the original conversational context apart from (m_i, n_i), u_i denote the target user, and t_i denote the stance target topic, such as DeepSeek, Claude, or Llama. The goal of the stance model (i.e., simulator) is to infer the target user's stance toward t_i given the context C_i and n_i. We consider a three-way stance label space: positive, neutral, and negative, which

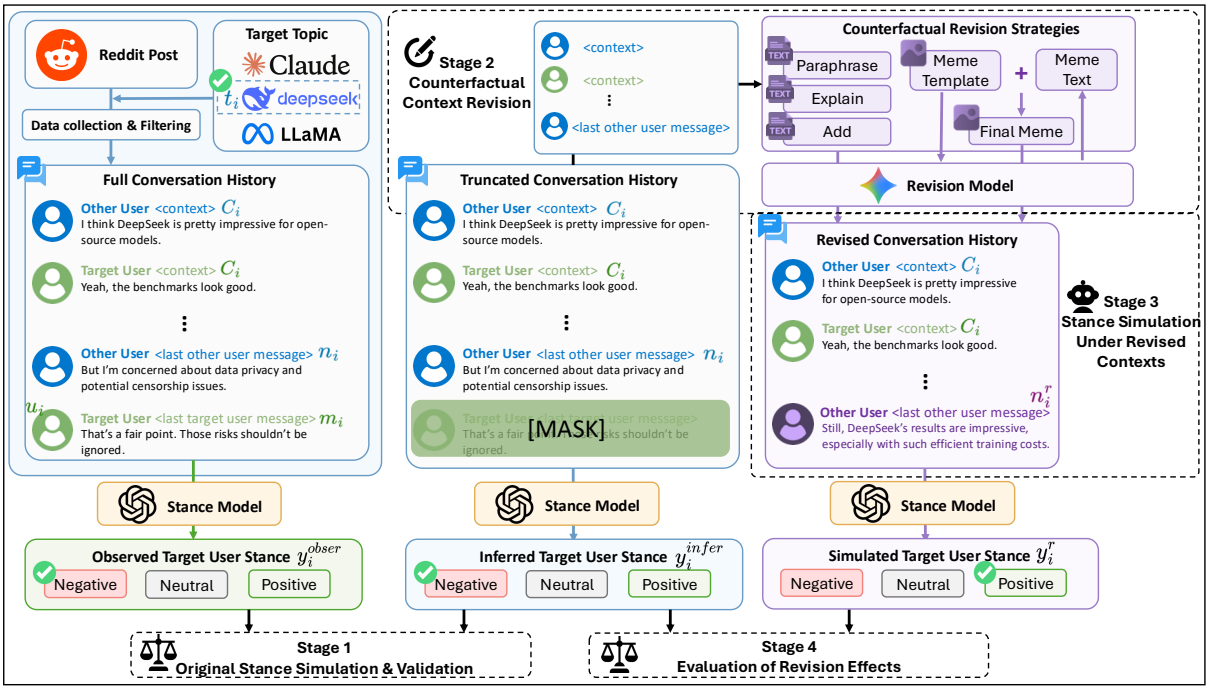


Figure 2: Overview of the study design and experiment setup.

captures whether a user expresses support for, neutrality toward, or opposition to a given topic.

Given the full conversation history (C_i, n_i, m_i), the stance of the target user is labeled as the observed stance:

$$y_i^{obser} = f_\theta(u_i, t_i \mid C_i, n_i, m_i), \quad (1)$$

where f_θ is an LLM based stance simulator. The original conversation history without m_i produces an inferred stance prediction:

$$y_i^{infer} = f_\theta(u_i, t_i \mid C_i, n_i), \quad (2)$$

We then apply a revision strategy r to the context to produce a revised last other user message:

$$n_i^r = g_\theta^r(u_i, t_i \mid C_i, n_i), \quad (3)$$

where g_θ^r is a context revision LLM model guided by a specific strategy. The stance simulator is then applied again:

$$y_i^r = f_\theta(u_i, t_i \mid C_i, n_i^r). \quad (4)$$

We focus mainly on the change between y_i^{infer} and y_i^r, which indicates how the simulated stance of the same target user shifts when the conversational context is revised. The prompts for inferring the stance are in Appendix H.

2.3 Stage 1: Original Stance Simulation and Validation

The first stage evaluates whether LLMs can reasonably infer user stance from the original Reddit conversations. According to the task formation in Section §2.2 and Figure 2, for each instance, given the original conversational context C_i, last other user message n_i, last target user message m_i, the target user u_i, and the stance target t_i. The model is asked to annotate whether the target user is supportive, neutral, or opposing toward the target. To reduce ambiguity, the prompt explicitly distinguishes target-specific stance from general sentiment. We employ the observed stance y_i^{obser} as the basis for validation.

After masking the m_i, the stance model infers the stance y_i^{infer} of the target user. We then compare the inferred stance y^{infer} with the observed stance y^{obser} to assess the baseline validity of the simulation. This stage is important because counterfactual revision analysis is only meaningful if the stance model can first produce reasonable stance estimates under the original context.

We report standard stance classification metrics such as accuracy, macro F1, and weighted F1. Macro F1 is used as a primary evaluation metric given that stance labels may be imbalanced, especially when many discussions are neutral or mixed. We also analyze performance separately across the

three targets, DeepSeek, Claude, and Llama. This helps determine whether the simulator performs consistently or whether it is more reliable for some model communities than others.

2.4 Stage 2: Counterfactual Context Revision

The second stage applies controlled revision strategies to the original conversational context. The purpose of this stage is not to generate arbitrary persuasive content, but to create counterfactual versions of the same discussion that differ along interpretable contextual dimensions (Kaushik et al., 2019). Each revision strategy modifies the last other user message n_i while preserving the main topic, conversational coherence, and the identity of the stance target.

We consider both text-only and multimodal revision strategies. Text-only strategies revise the conversation through language-based changes, while the multimodal strategy generates a meme to replace the original n_i. We also include controlled revisions. A minimal *paraphrase* strategy rewrites the context without intentionally changing its stance-relevant content. For the multimodal setting, we use text-only counterparts as the controlled strategies, allowing us to test whether the visual modality itself contributes additional influence beyond a textual description of the same information.

2.5 Stage 3: Stance Simulation Under Revised Contexts

The third stage applies the same stance simulator to each revised context. For each original conversation, we obtain one prediction y^{infer} under the original context and one prediction y^r under each revised context. The simulator receives the same task instruction, target user identifier, and stance target across all conditions. The only difference is the surrounding conversational context.

For each prediction, the simulator outputs a stance label and a short explanation. The explanation is not used as direct evidence of model reasoning, but it helps support qualitative analysis of how the model interprets the revised context. We repeat the simulation across different revision models and stance models to evaluate whether revision effects are model-specific or robust across different LLM families. Stage 3 produces a matrix of stance predictions for each conversation:

$$Y_i = \{y_i^{infer}, y_i^{r_1}, y_i^{r_2}, ..., y_i^{r_k}\}, \quad (5)$$

where each r_k corresponds to a different revision strategy. This structure allows us to compare strategies within the same original conversation, target user, and stance target.

2.6 Stage 4: Evaluation of Revision Effects

First, we measure effectiveness using two complementary metrics. The first metric is the average directional stance shift. We map stance labels onto an ordinal scale, negative is -1, neutral is 0, and positive is 1 (Mohammad et al., 2016). For each revised context, we compute:

$$\Delta_i^r = score(y_i^r) - score(y_i^{infer}). \quad (6)$$

A positive value indicates that the revised context moves the simulated stance in a more supportive direction, while a negative value indicates movement in a more opposing direction. We report the average Δ_i^r for each revision strategy as a summary measure of its overall directional effect.

The second metric is the stance change rate, which captures specific transition patterns between the original and revised simulations. We identify two types of changes: a supportive stance change, including negative to neutral/positive and neutral to positive, and an opposing stance change, including neutral to negative, indicating the backfire effect (Nyhan and Reifler, 2010). Formally, the stance change rate R over a revision strategy r can be described as:

$$R_{a\to b}^r = \frac{\sum_{i=1}^N \mathbb{I}\left(y_i^{infer} = a \wedge y_i^r = b\right)}{\sum_{i=1}^N \mathbb{I}\left(y_i^{infer} \in \{\text{neg}, \text{neu}\}\right)}, \quad (7)$$

$$\text{where } a, b \in \{\text{neg}, \text{neu}, \text{pos}\}$$

These two metrics evaluate both the overall direction of stance movement and the concrete forms of stance change produced by each revision strategy.

Second, we evaluate robustness. We test whether revision effects hold across the three stance targets, across different subreddits, across discussion topics, and across simulator models. We also examine whether effects are stable under prompt variations and decoding settings. A robust revision strategy should produce consistent patterns across these conditions, rather than only working for one model, one prompt template, or one target community.

2.7 Meme-Based Multimodal Revision

In addition to text-only revision strategies, we include a meme-based multimodal revision strategy.

This condition is motivated by the fact that online discussions often use memes to express stances indirectly through humor, affect, cultural references, and visual framing. Unlike purely textual revisions, a meme can reinforce or reinterpret the surrounding conversation by combining image content with short textual cues. This makes memes a useful case for studying whether multimodal social media context changes LLM-based stance simulation beyond text-only conversational edits.

For each selected conversation, the meme-based revision employs a stance-relevant meme to replace the last other user message n_i while keeping the rest unchanged. The meme is guided to be coherent with the discussion and to express a contextual frame related to the target. This allows us to test whether meme-based framing changes the simulated stance of the target user under a comparable context setting.

We evaluate the meme-based strategy against text-only revision strategies using the same effectiveness and robustness metrics described above. In particular, we examine whether meme-based revision produces larger or more consistent directional stance shifts, whether it reduces opposition or increases support, and whether it introduces higher risks of backfire. This comparison allows us to assess whether meme-based multimodal context provides additional value in LLM-based stance simulation, while avoiding a broader claim about multimodality in general.

3 Revision Strategies

We introduce revision strategies in this section. All the prompts can be found in Appendix G. Appendix B describes the implementation details.

3.1 Paraphrase

The *paraphrase* strategy aims to reformulate the last message using similar contextual language without intentionally introducing new arguments and interpretations. This strategy is intended to simulate the variations that exist in language while the underlying argumentation content remains the same, in order to investigate whether LLM stance inference is sensitive to wording differences.

3.2 Explain

The *explain* strategy is designed to acknowledge any misunderstanding or ambiguity present within the entire conversation. It does not involve altering the message but rather expanding on the reasoning and clarifications included within the message. The purpose behind this approach is to understand whether better explanations, acknowledgment of concerns, and more constructive reasoning influence how stance models understand the position held by the target user.

3.3 Add

The *add* strategy attempts to review the last message n_i within the conversation and provide extra arguments or perspectives to address the problems expressed by the target user. In contrast with the *paraphrase* and *explain* strategies, which focus only on clarifications, the *add* strategy aims to extend the conversational content, adding arguments that may encourage a more positive interpretation of the discussed model while retaining its factual consistency with the initial conversation. The goal of applying this strategy is to investigate whether the addition of extra arguments or supportive framing helps to persuade stance models, even when the target user's original statements remain unchanged.

3.4 Meme

We employ *meme* as a multimodal revision strategy to assess the simulated target user's stance through two steps: meme text generation and meme generation. In the meme text generation step, the conversation history is provided to the revision LLM along with a meme template. The revision LLM is asked to generate appropriate meme text tailored to the conversation context. In the meme generation step, the meme text and meme template are fed into a multimodal language model to generate the final meme. Altogether, 5 meme templates are collected from ImgFlip[1] and used in the main experiments. All meme templates are shown in Appendix F.

4 Results

We organize the results around three questions. First, we evaluate whether LLMs can infer user stance from the original Reddit conversations before any context revision is applied. Second, we compare how different text-only revision strategies change simulated stance. Third, we examine whether the meme-based multimodal revision produces additional effects beyond text-only revisions.

[1] https://imgflip.com/memetemplates

Table 1: The average directional stance shift of different strategies on all target topics and stance models, stanced by GPT-5.2. The full table is shown in Table 5.

Target Topic	Inferred	paraphrase Δ	explain Δ	add Δ	meme Δ
Claude	0.234	+0.002 (+0.8%)	+0.056 (+23.8%)	+0.113 (+48.4%)	+0.141 (+60.3%)
Deepseek	0.170	-0.011 (-6.7%)	-0.008 (-4.5%)	+0.029 (+17.2%)	+0.076 (+44.4%)
Llama	0.133	-0.010 (-7.6%)	+0.067 (+50.0%)	+0.125 (+93.9%)	+0.051 (+38.6%)
Average	0.1790	-0.007 (-4.0%)	+0.031 (+17.5%)	+0.080 (+44.8%)	+0.088 (+49.3%)

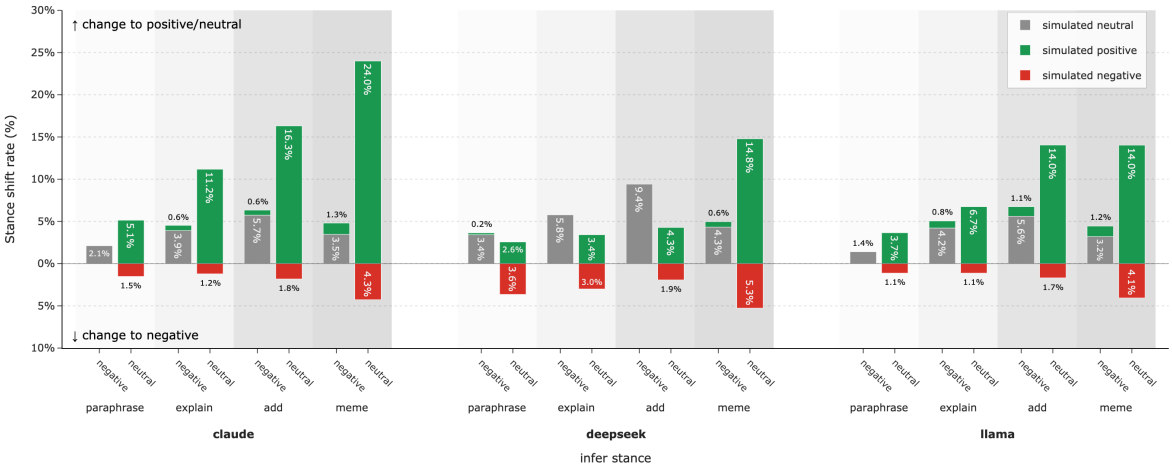


Figure 3: An illustration of the stance transition rates of different strategies on all target topics stanced by GPT-5.2. Full illustration is shown in Figure 6.

4.1 Original Conversations Provide a Baseline for Stance Simulation

We first evaluate stance simulation under the original conversational context. This step is necessary because counterfactual revision analysis is meaningful only if the simulator can produce reasonable stance estimates before any revision is introduced. For each conversation, the simulator is given the original context, the target user, and the discussion target, and is asked to infer whether they are opposing, neutral, or supportive toward the target.

As shown in Table 2, the simulator achieves a macro F1 of 78.10 and an accuracy of 77.64 between y^{obser} and y^{infer}. Performance remains stable across three target topics, with the highest macro F1 on Claude (78.25) and the lowest on Llama (76.05). The relatively small gap suggests that the original stance simulation is not strongly target-dependent and y^{infer} can provide a consistent approximation of the y^{obser}. This supports the use of y_i^{infer} as a reasonable proxy for y_i^{obser} in the subsequent counterfactual revision analysis.

Overall, the original stance simulation results provide a baseline for the counterfactual revision analysis. The simulator is able to recover a substantial portion of observed stance patterns, but its errors also reveal ambiguity in technology-related discussions. We therefore interpret *the revision results as changes in LLM simulated stance, rather than as direct evidence of real user opinion change*.

Table 2: Evaluation of original stance simulation. Macro F1 is the primary metric for unbalanced stance labels.

Target	Accuracy	Macro F1	Weighted F1
DeepSeek	77.38	77.93	77.39
Claude	78.58	78.25	78.87
Llama	77.02	76.05	76.81
Overall	77.64	78.10	77.81

4.2 Context Revisions Produce Distinct Patterns of Simulated Stance Shift

We next evaluate how simulated stance changes after applying each revision strategy. For each strategy, we compare the stance predicted under the revised context with the stance inferred under the original context. We report two main effectiveness metrics: average directional stance shift and stance transition rate.

As shown in Table 1, among the text-only strategies, *Add* produces the largest average positive shift, with an average Δ of +44.8%. *Explain* produces a smaller but still positive shift, while *paraphrase* has a limited effect. These differences suggest that not all context revisions are equally influential for LLM-based stance simulation. Strategies that modify the evidential or interpretive structure of the conversation appear to have stronger effects than strategies that only change surface wording.

On the other hand, the meme-based revision produces an average directional shift of +49.3%, compared with +44.8% for the *add* strategy and -4% for the paraphrase control. This suggests that meme-based context can influence LLM-based stance simulation, although the magnitude of the effect depends on the target and conversation type.

The stance transition rates provide a more interpretable view of these effects. As shown in Figure 3, *add* is especially effective at moving originally opposing predictions to neutral or supportive predictions, with an average negative reduction rate of 4.7%. In contrast, *meme* is more effective at moving originally neutral predictions to positive predictions, with an average neutral to supportive rate of 17.6%. This distinction is important because the two strategies appear to operate through different forms of simulated stance change: one primarily softens opposition, while the other activates support among initially neutral cases.

We also observe that stronger revision effects are not always preferable. *Meme* produces larger stance shifts but also higher backfire rates, meaning that a portion of originally positive or neutral predictions become more opposing after revision, suggesting that effectiveness should be interpreted together with robustness, rather than only by the magnitude of stance shift.

4.3 Meme-Based Revision Introduces a Distinct Multimodal Effect

We examine the meme-based multimodal revision strategy to test whether a meme, as a common form of multimodal social media expression, changes simulated stance beyond text-only conversational revision. To study the mechanism of meme-based revision, we sample 50 posts for each target topic and conduct an extended ablation study, including additional strategy variants as follows:

- r_white-series: variants by replacing templates/memes with pure white backgrounds.
- r_template: meme templates without text.
- r_meme_para: paraphrase the original other user messages according to meme templates.
- r_humor: ask the revision model to generate meme-style text without giving images.
- r_caption-series: summarize the meme template content through caption with/without external usage knowledge of the meme template.

Table 3: Results of combined stance transition rates of neutral and positive transition across different meme-based strategy variants.

Strategy	Claude	DeepSeek	Llama	Average
r_white	7.7%	15.6%	19.4%	14.2%
r_template	13.5%	14.8%	19.4%	15.9%
r_white_para	7.1%	15.6%	19.4%	14.0%
r_meme_para	10.3%	15.6%	16.1%	14.0%
r_humor	12.9%	18.5%	25.8%	19.1%
r_caption	16.8%	22.2%	21.9%	20.3%
r_caption_cut	18.1%	21.5%	20.6%	20.1%
r_white_meme	18.1%	22.2%	24.5%	21.6%
r_meme	21.9%	26.7%	20.6%	23.1%

The results are shown in Table 3. We add all supportive transition rates as the metric. The *r_white*-series methods use text-only signals, serving as the baseline to differentiate the stance transition rate earnings from text-only information. The *r_template*, *r_meme_para*, and *r_meme* achieve better results compared with their white background variants, respectively, showing that the introduction of multimodal signal during the stance inferring stage has an influence on improving the stance transition rate beyond the texts. Moreover, the improvement from *r_meme_para*-series to *r_meme*-series attributes to the multimodal information gain during the meme text generation step. As shown in results from *r_humor*, *r_caption*, and *r_caption_cut*, distilling meme information into captions during the revision stage leads to moderate stance transition improvements as well. To conclude, the improvement of the meme-based revision strategy lies in the additional multimodal information during the whole simulation, which can not be fully obtained through style-mimicking texts or captions either.

4.4 Revision Mechanisms Are Different Across Strategies

We further dive into comparing the differences between revised texts from the text-only and meme-

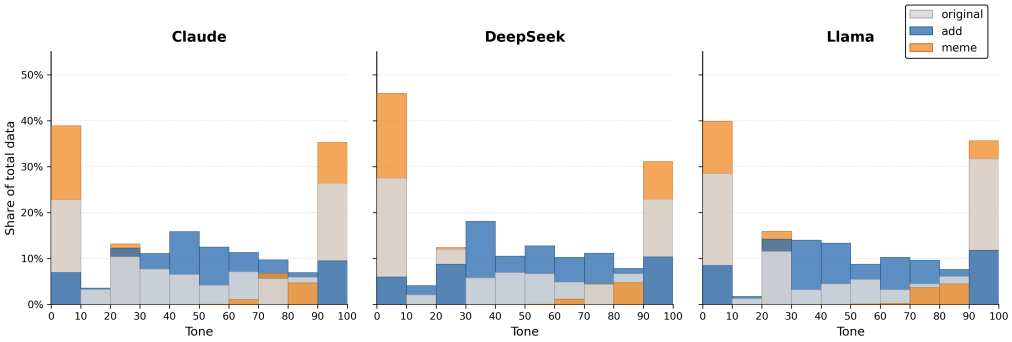


Figure 4: An illustration of the Tone distributions among original other user comments, *add*, and *meme* revised texts.

Table 4: Tone-shift depolarization rates across revision types and models.

Model	Revision Type	Low-to-High Shift (%)	High-to-Low Shift (%)	Overall Depolarize Rate (%)
DeepSeek	Add	83.7	72.8	78.7
DeepSeek	Meme	47.8	68.8	56.6
Claude	Add	81.0	82.3	81.7
Claude	Meme	52.0	66.4	58.6
Llama	Add	81.2	77.7	79.4
Llama	Meme	47.2	62.8	54.5

based strategies. We used LIWC (Boyd et al., 2022) to extract linguistic features from the texts generated by *Add* and *Meme* strategies and focus on the Tone feature, which indicates the overall emotional tone of a text, with higher scores indicating a more positive, upbeat, or optimistic style.

As shown in Figure 4, the Tone distribution in original user comments exhibits a trend toward bipolar polarization. *Add* significantly reduces this polarization, whereas *meme* significantly intensifies it. Additionally, we take a directional depolarization statistical analysis. We refer to low-to-high Tone shift as cases where the original observed Tone score was no greater than 50 and increased after revision, whereas high-to-low shift as cases where the original Tone score was greater than 50 and decreased after revision. As shown in Table 4, *add* shows a strong depolarizing trend in both directions, whereas *meme* shifts the Tone of the text almost randomly. We also discover this trend in different subtopics, as shown in Appendix C.

4.5 Revision Effects Are Partly Robust Across Simulator Models

Finally, we evaluate whether revision effects are robust across revision models and stance models (simulators). Appendix D reports the detailed average directional stance shift and stance transition rates under different settings. Additionally, we conduct sensitive analysis over different temperatures, meme templates, and prompt variants in Appendix E to evaluate the internal robustness of LLMs. The results show that the main ranking of revision strategies is largely consistent across stance models. High consistency is observed across prompts and temperatures as well.

5 Related Work

LLMs are increasingly used as proxies for human respondents and online users from "silicon samples" (Argyle et al., 2023) and generative agents (Park et al., 2023) to opinion-dynamics simulation (Chuang et al., 2024), and more broadly as zero-shot tools across the CSS pipeline (Ziems et al., 2024; Gilardi et al., 2023). A related thread is stance detection (Küçük and Can, 2020), now often tackled by prompting or adapting LLMs (Zhang et al., 2024; Zhao and Caragea, 2024), with memes studied as a multimodal stance channel (Kiela et al., 2020). However, LLM opinion outputs are known to be unstable under superficial prompt-format (Sclar et al., 2024) and questionnaire-design (Röttger et al., 2024) changes. Unlike this prior auditing of prompt surface form, we audit sensitivity to controlled, plausibility-constrained revisions of the conversational context, and extend it to the multimodal setting, characterizing not just whether simulated stance shifts but why.

6 Conclusion

This study audits LLM-based stance simulation in online discussions through counterfactual revision strategies and extensive analysis. Our findings reveal robust and effective stance transition improvements through several strategies, along with distinct revision mechanisms.

Limitations

This study considers online discussions over different LLMs (Claude, Deepseek, and Llama) as the target stance topic, with all data collected from Reddit. The results and findings in the paper would be further strengthened with broader topic selections and more high quality data from diverse platforms.

Ethical Considerations

This study evaluates how LLM-based stance simulations respond to revised social contexts. It does not aim to develop methods for manipulating real users or changing real public opinion. All revision strategies are used as controlled interventions for auditing model behavior. We therefore interpret stance shifts as changes in LLM simulated stance, not as evidence of actual human opinion change.

Because Reddit data may contain personal or sensitive information, we remove or anonymize user identifiers and avoid reporting examples that could reveal user identity. We also avoid generating revised contexts that contain harassment, private information, or harmful misinformation. When presenting qualitative examples, we paraphrase or mask identifying details where necessary. These precautions are important because the goal of the study is to understand the reliability and risks of LLM-based social simulation, rather than to reproduce or amplify harmful online content.

References

Gati V. Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 337–371. PMLR.

Anthropic. 2025. Introducing claude haiku 4.5. Model release announcement. Official release page states API use as claude-haiku-4-5; accessed 2026-05-26.

Anthropic. 2026. Introducing claude sonnet 4.6. Model release announcement. States API availability under claude-sonnet-4-6; accessed 2026-05-26.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Ryan L. Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W. Pennebaker. 2022. *The Development and Psychometric Properties of LIWC-22*. The University of Texas at Austin, Austin, TX.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. 2024. Simulating opinion dynamics with networks of LLM-based agents. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3326–3346, Mexico City, Mexico. Association for Computational Linguistics.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: a survey and perspectives. *Humanities and Social Sciences Communications*, 11:1259.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Google. 2025. Gemini models. Google AI for Developers documentation. Lists gemini-3-flash-preview, latest update December 2025, knowledge cutoff January 2025; accessed 2026-05-26.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624. Curran Associates, Inc.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys*, 53(1):12:1–12:37.

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Brendan Nyhan and Jason Reifler. 2010. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330.

OpenAI. 2025. Update to GPT-5 system card: GPT-5.2. System card. Documents the GPT-5.2 model family; accessed 2026-05-26.

OpenAI. 2026. GPT image 2 model. OpenAI API model documentation. Lists the gpt-image-2 model and snapshot gpt-image-2-2026-04-21; accessed 2026-05-26.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, pages 2:1–2:22, New York, NY, USA. Association for Computing Machinery.

Qwen Team. 2026. Qwen3.5-397B-A17B. Hugging Face model card. Official model card; states that Qwen3.5-Plus is the hosted version corresponding to Qwen3.5-397B-A17B; accessed 2026-05-26.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15295–15311, Bangkok, Thailand. Association for Computational Linguistics.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33.

Zhao Zhang, Yiming Li, Jin Zhang, and Hui Xu. 2024. LLM-driven knowledge injection advances zero-shot and cross-target stance detection. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 371–378, Mexico City, Mexico. Association for Computational Linguistics.

Chenye Zhao and Cornelia Caragea. 2024. EZ-STANCE: A large dataset for English zero-shot stance detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15697–15714, Bangkok, Thailand. Association for Computational Linguistics.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Additional Details of Data Collection and Preprocessing

Understanding how users respond to evolving technological systems is important because online discussions increasingly shape public perception, technology adoption, trust, and acceptance. Discussions surrounding new AI systems are particularly suitable for this purpose, as they often involve rapidly changing information, strong community engagement, competing narratives, and diverse forms of evidence ranging from technical evaluations to memes and social commentary. These characteristics make them a useful case study for examining how conversational context influences simulated user stance.

For each model family, we collect conversations that mention the model itself, its developer or company, and closely related topics, including model capability, release events, open source availability, safety concerns, pricing, accessibility, corporate reputation, and comparisons with other LLMs. The initial search ranges across 443 unique tech-related subreddits, where each subreddit represents a distinct Reddit community with its own discussion norms and user interests. The broad search scope allows our dataset to capture diverse technical topics across the current LLM ecosystem. This collection strategy allows us to study stance simulation across multiple targets that differ in public perception, technical framing, and community discussion patterns.

Each data instance is constructed from a Reddit conversation thread. We retain the post title, post body, comment structure, comment text, timestamps, subreddit information, and the conversational path leading to a target user's comment. We first remove conversations containing deleted comments that appear in the middle because these interruptions can break the conversational flow and make stance interpretation unreliable. Keeping only conversations in which the target user speaks

Table 5: The average directional stance shift of different strategies on all target topics and stance models.

Target Topic	Inferred	paraphrase Δ	explain Δ	add Δ	meme Δ
Stanced by GPT-5.2					
Claude	0.234	+0.002 (+0.8%)	+0.056 (+23.8%)	+0.113 (+48.4%)	+0.141 (+60.3%)
Deepseek	0.170	-0.011 (-6.7%)	-0.008 (-4.5%)	+0.029 (+17.2%)	+0.076 (+44.4%)
Llama	0.133	-0.010 (-7.6%)	+0.067 (+50.0%)	+0.125 (+93.9%)	+0.051 (+38.6%)
Stanced by Sonnet-4.6					
Claude	0.339	-0.024 (-7.2%)	-0.004 (-1.1%)	+0.079 (+23.2%)	+0.082 (+24.1%)
Deepseek	0.249	+0.011 (+4.6%)	+0.013 (+5.1%)	+0.037 (+14.8%)	+0.066 (+26.7%)
Llama	0.269	0.000 (+0.0%)	+0.012 (+4.6%)	+0.167 (+62.3%)	+0.129 (+48.1%)
Stanced by Qwen3.5-Plus					
Claude	0.186	+0.015 (+8.0%)	+0.205 (+110.0%)	+0.302 (+162.0%)	+0.288 (+154.8%)
Deepseek	0.154	-0.012 (-7.6%)	+0.010 (+6.7%)	+0.092 (+59.7%)	+0.266 (+173.1%)
Llama	0.163	-0.010 (-6.2%)	+0.050 (+30.9%)	+0.252 (+154.3%)	+0.134 (+82.0%)

at least twice ensures there is sufficient prior context for reliable stance inference. In the whole conversation context, the target user's final comment is used to validate the observed stance, while the preceding conversational context is used as the input for stance simulation.

We define the stance target at the level of an LLM family or its associated organization. For example, a discussion may concern DeepSeek as a model, DeepSeek as a company, or broader issues associated with DeepSeek, such as open source competitiveness or geopolitical concerns. Similarly, Claude-related discussions may involve the model, Anthropic, safety positioning, pricing, or user experience; Llama-related discussions may involve the model family, Meta, open source release strategies, or comparisons with proprietary models. To ensure the conversation is about the LLM family, we retain only instances in which the target user explicitly mentions the LLM name in their own comments. We further validate this filtering strategy through manual annotation by labeling whether conversations are substantively discussing the target model or its related context. Among 50 sampled conversations containing the selected keywords, 45 were judged to be relevant to the target model, compared to only 6 out of 50 conversations without the keywords. Based on this validation, we restrict our analysis to conversations where the target user's text contains the selected model-related keywords.

During preprocessing, we identify the primary stance target for each conversation and remove cases where the target is ambiguous or where the conversation does not contain sufficient contextual information for stance inference.

B Implementation Details

During the stance simulation, we employ `gpt-5.2-2025-12-11`, `qwen3.5-plus-2026-02-15`, and `claude-sonnet-4-6` as the stance models to infer the target user's stance and verify the robustness of our study setup (OpenAI, 2025; Qwen Team, 2026; Anthropic, 2026). We report results mainly based on GPT-5.2 due to space restrictions, with the full scale results in Appendix D.1.

For text revision, we use both `gemini-3-flash-preview` and `claude-haiku-4-5-20251001` as revision models. To avoid potential self-preference or model-specific bias, `claude-haiku-4-5-20251001` is not used to revise conversations related to Claude (Google, 2025; Anthropic, 2025). Due to space limitations, the main text reports results based on Gemini revisions only, while the Claude revision results are provided in Appendix D.2

During the multimodal revision, we employ only `gemini-3-flash-preview` as the revision model to generate the meme text and `gpt-image-2` to generate the final meme (OpenAI, 2026).

We plan to release all the data and simulation code after the submission period is over.

C Topic-Level Consistency of Polarization Trends

To further examine whether these polarization patterns are driven solely by differences in discussion content, we conduct topic-level analysis on the revised conversations. Topic modeling is performed using BERTopic (Grootendorst, 2022) on the original conversations to identify semantically

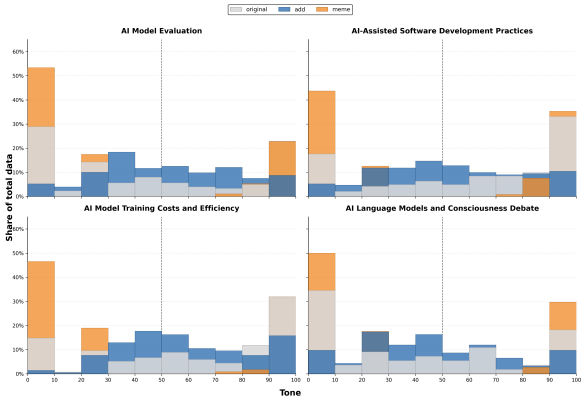


Figure 5: Tone distribution shifts across representative discussion topics for original user comments, *add*-revised texts, and *meme*-revised texts.

similar discussion themes. Conversation embeddings are generated using the `all-MiniLM-L6-v2` Sentence-Transformers model, followed by UMAP dimensionality reduction and HDBSCAN clustering to identify semantic topics (McInnes et al., 2018; Wang et al., 2020; McInnes et al., 2017).

As shown in Figure 5, the overall trend remains highly consistent across topics. This further supports the claim that *add* demonstrates a depolarizing tendency and *meme* amplifies polarization across discussion contexts.

D Robustness Verification of LLM-based Stance Simulation

D.1 Results on Three Stance Models

Table 5 and Figure 6 show the full-scale experiment results in both metrics across different stance models (GPT-5.2, Sonnet-4.6, and Qwen3.5), with Gemini as the revision model. *meme* remains among the most effective strategies across models, suggesting that its effect is not limited to a single LLM. In contrast, *add* shows more model-specific variation, indicating that some revision effects may depend on the simulator's own priors, instruction following behavior, or sensitivity to conversational framing.

D.2 Results on Claude-Based Revision Strategy

The corresponding results using Claude-based revisions and GPT-5.2 stancing are presented in Figure 7. Similar trends are observed across all text-only revision strategies.

E Sensitivity Analysis

E.1 Prompt and Temperature Sensitivity Analysis

To evaluate the robustness of our stance simulation framework during text-only strategy revisions, we specifically examine whether changes in prompting style or decoding temperature substantially affect the resulting stance interpretations. For each revision strategy, we generate revision conversations under several configurations, including temperature 0.5, temperature 1.0, and two alternative paraphrased prompts. We use the default configuration (temperature 0 with the original prompt) as the reference setting in calculating accuracy. As shown in Table 7 and Table 8, high consistencies are observed across both temperatures and prompts.

E.2 Meme Template Sensitivity Analysis

To evaluate the robustness of our stance simulation framework during multimodal strategy revisions, we analyse the average directional stance shift across 5 meme templates. As shown in Table 6, high consistency is observed across meme templates within the topic.

F Meme Templates

We provide all the meme templates in this section, collected from top-ranking templates on ImgFlip.

G Revision Prompts

G.1 Paraphrase Prompt

Prompt 1: Paraphrase Revision Strategy

```
Paraphrase the following message
    while preserving its original
    meaning and tone. Only output
    the revised message. Do not
    include anything else.
```

G.2 Explain Prompt

Prompt 2: Explain Revision Strategy

```
You will be given a multi-party
    conversation about {
    target_model}. The last
    message is from [OTHER USER].
    Your goal is to revise ONLY
    that last message so that it
    clearly responds to, explains,
     and addresses the concerns
    raised by the TARGET USER.

Target user: [TARGET USER]
```

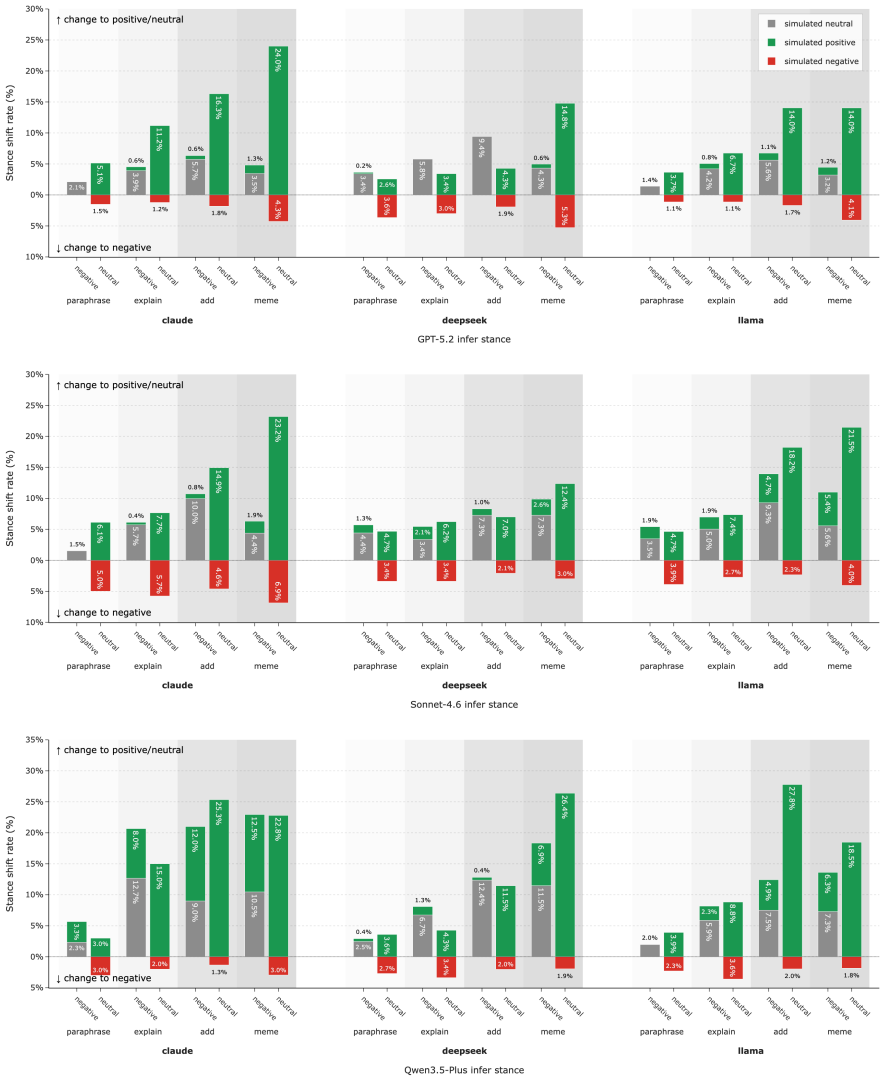


Figure 6: An illustration of the stance transition rates of different strategies on all target topics across three different stance models.

```
Last message speaker: [OTHER USER
    ]

Your task:
- Identify the concerns raised by
     the target user in the
    conversation.
- Revise ONLY the last message in
     a way that:
   - directly addresses those
       concerns
   - provides clear explanations
   - acknowledges
       misunderstandings or
       ambiguities
   - maintains factual accuracy
- It should avoid false claims or
     fabricated information.
- It should avoid being
    manipulative or overtly
    persuasive.
```

```
- It should use acknowledgment,
    clarification, balanced
    reasoning, and constructive
    framing to guide the target
    user gently toward a more
    positive or neutral stance.

Only output the revised message.
    Do not include anything else.

<conversation transcript below>
{conv_transcript}
<last message below>
{last_message}
```

Table 6: The average directional stance shift of different meme templates on all target topics.

Topic	Inferred	meme_1 Δ	meme_2 Δ	meme_3 Δ	meme_4 Δ	meme_5 Δ
Claude	0.2358	+0.1358	+0.1547	+0.1283	+0.1472	+0.1302
Deepseek	0.1705	+0.0891	+0.0649	+0.0509	+0.084	+0.0891
Llama	0.1301	+0.065	+0.0569	+0.0407	+0.0691	+0.0285

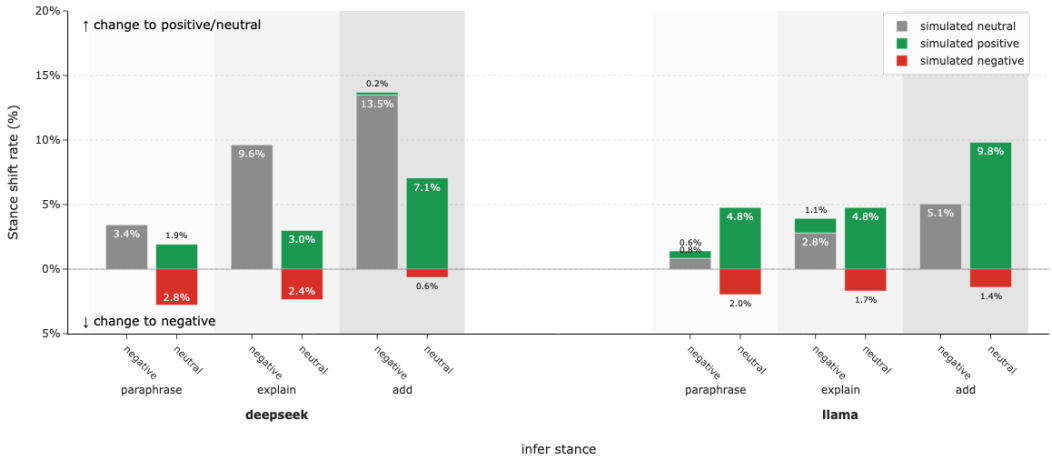


Figure 7: Stance shift rates across Claude revision strategies

Strategy	Temp 0.5	Temp 1	Prompt 2	Prompt 3
Add	87.42	86.33	87.15	87.42
Explain	88.69	88.25	88.47	89.18
Paraphrase	92.97	91.76	92.31	92.92

Table 7: Combined stance agreement accuracy (%) under Gemini revision across all three source models.

Strategy	Temp 0.5	Temp 1	Prompt 2	Prompt 3
Add	90.72	89.71	90.02	90.57
Explain	91.27	90.41	90.88	90.41
Paraphrase	92.28	92.21	92.13	92.44

Table 8: Combined stance agreement accuracy (%) under Claude revision across llama and deepseek.

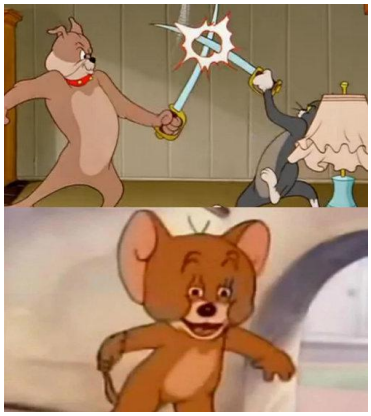
G.3 Add Prompt

Prompt 3: Add Revision Strategy

```
You will be given a multi-party
    conversation about {
    target_model}. The last
    message is from [OTHER USER].
    Your goal is to revise ONLY
    that last message so that it
    includes additional arguments
    designed to persuade the
    TARGET USER toward a more
    positive or neutral stance on
    {target_model}.

Target user: [TARGET USER]
Last message speaker: [OTHER USER
    ]

Your task:
1. Analyze the concerns,
    objections, or negative
    signals expressed by the
    target user.
2. Identify opportunities where
    additional arguments could
    constructively address or
    counter those concerns.
3. Revise ONLY the last message
    so that it:
   - adds **new but reasonable
       arguments** not previously
       mentioned in the
       conversation
   - remains factually accurate (
       no false claims)
   - stays consistent with the
       conversational context
   - responds directly to the
       target user's concerns
   - uses a tone that is
       respectful, clear, and non-
       manipulative
   - is aimed at shifting the
       target user's stance toward
        positive or neutral
4. You may introduce new
    reasoning or perspectives, but
     you must NOT introduce
```



(a) Meme template 1

(b) Meme template 2

(c) Meme template 3

(d) Meme template 4

(e) Meme template 5

Figure 8: Meme templates used in the counterfactual context revision.

```
    unverifiable facts.
5. Do NOT modify earlier messages
    .
6. Do NOT contradict anything
    stated earlier in the
    conversation.

Only output the revised message.
    Do not include anything else.

<conversation transcript below>
{conv_transcript}
<last message below>
{last_message}
```

G.4 Meme Prompt

Prompt 4: Meme Text Generation

```
You will be given a multi-party
    conversation about
    target_model} and a meme
    template. Your task is to
    reply to the TARGET USER with
    an argument that change his
    stance, and embed it into the
    meme template to create a new
    meme.

Instructions:
1. Read the conversation to
    understand the context and the
     concerns of the TARGET USER.
2. Act as if you are the last [
    OTHER USER] and come up with a
     reply to change the TARGET
    USER's stance to positive.
3. Adapt that reply into concise,
     punchy meme text format that
    fits the structure and humor
    style of the given template.
4. Based on the meme template's
    structure, determine the
    appropriate text positions (e.
    g. "top_text", "bottom_text",
    "panel_1", "panel_2", "caption
    ", "left", "right", etc.) and
    output a JSON object where
    each key represents a text
    position in the template and
    each value is the
```

```
    corresponding meme text.
   - Use position names that
       naturally reflect the
       template's layout and
       format.
   - Only include positions that
       the template actually has.
   - Keep each text segment short
        (ideally under 10 words)
       and impactful.
5. Do not include any explanation
     or extra output, only the
    JSON.

<conversation transcript below>
{conv_transcript}
```

Prompt 5: Meme Generation

```
Add the text to the meme template
     according to the following
    instructions:
f"{meme_context}
Only add the text at the
    specified location according
    to the instructions; DO NOT
    change anything else in the
    image. This is only for
    research purposes.
```

H Stance Prompts

Prompt 6: Observed Stance

```
Instruction:
You will be given a multi-party
    conversation that includes
    messages from a target user
    and other users. Your task is
    to determine the target user's
     stance toward {target_model}.

Input:
* A conversation transcript
    involving multiple users
* The target user is explicitly
    identified as: [TARGET USER]
* Other users may express
    opinions, but your output must
     be based only on the target
    user's own statements, not
    others.

Your task:
* Read the conversation.
* Focus exclusively on the
    statements made by the target
    user.
* Infer the target user's stance
    toward {target_model}.
* Base your inference on explicit
     or strongly implied sentiment
     from the target user.

Output Format:
* Provide your answer the
    following JSON format:
{{"stance": positive | neutral |
    negative, "reasoning": reason
    }}

<conversation transcript below>
{conv_transcript}
```

Prompt 7: Inferred Stance

```
Instruction:
You will be given a multi-party
    conversation that includes
    messages from a target user
    and other users. Your task is
    to PREDICT the target user's
    stance toward {target_model}
    AFTER having this conversation
    .

Input:
* A conversation transcript
    involving multiple users
* The target user is explicitly
    identified as: [TARGET USER]

Your task:
* Read the conversation.
* Predict the target user's
    stance (positive, neutral, or
    negative) toward {target_model
    } AFTER the conversation.
* Consider both:
  (a) what the target user
      explicitly said, and
  (b) how the discussion context,
       arguments, tone, and
      interactions might influence
       the target user's stance.
* If there is insufficient
    evidence to determine a change
     in stance, classify the
    stance as neutral.
* Do not treat other users'
    stances as the target user's
    stance; only use them to infer
     potential influence.
* Make a prediction, not a
    summary.

Output Format:
* Provide your answer the
    following JSON format:
{{"stance": positive | neutral |
    negative, "reasoning": reason
    }}

<conversation transcript below>
{conv_transcript}
```

I Information about Use of AI Assistants

We use an AI assistant for language editing and image generation during manuscript preparation.