HANDLE WITH CARE: CAN LLMS REPRODUCE HOW ONLINE COMMUNITIES REACT?

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly used as proxies for computational social analysis, yet faithfully representing the "thick descriptions" (Geertz, 1973) of human communities remains a critical challenge. Current evaluations often reduce social identity to static labels, sidelining how real-world groups navigate social shifts. To bridge this gap, we introduce **CARE** (**C**ommunity-**A**ware **R**eaction **E**valuation), a reaction-centered framework that benchmarks LLM-simulated discourse against the authentic, event-contingent responses of distinct communities to real-world news. Spanning 207 Reddit communities and covering 9,947 authentic reactions towards 2,166 news articles, CARE evaluates leading LLMs using a hierarchical taxonomy covering coarse attitudes and fine-grained communicative tones. Our empirical findings expose two critical failure modes in prevailing community-conditioning paradigms. First, while community context and targeted reasoning significantly enhance macro-level attitudinal and tonal profiling, these gains largely collapse at the instance level when predicting reactions to specific events. Second, the benefits of community conditioning are remarkably uneven: prompting strategies yield non-uniform shifts, where fidelity gains in some communities are offset by performance drops in others. This micro-macro divergence and community-level instability demonstrate that standard conditioning enables models to approximate static baseline profiles without capturing dynamic or equitable event reactions, establishing CARE as an essential diagnostic tool for community-aware social simulation.[1]

1 INTRODUCTION

Large language models (LLMs) are increasingly deployed as generative proxies for social analysis, celebrated for their ability to encode the linguistic contours of social norms and scale the simulation of human dynamics (Anthis et al., 2025; Piao et al., 2025). Yet, as their application in computational social science expands, a critical disconnect has emerged between how models are aligned to represent social groups and how these groups authentically function (Hovy & Yang, 2021; Kabir et al., 2025; Shen et al., 2025; Wang et al., 2025). Anthropological and sociolinguistic traditions emphasize that communities are not merely structural graphs or demographic buckets, but dynamic discursive spaces characterized by "thick" realities (Geertz, 1973; Adilazuarda et al., 2024). Within these spaces, group identity is not a static property but an active, event-contingent performance, negotiated through linguistically mediated reactions a community adopts to articulate its collective stance.

However, current socio-cultural community alignment evaluations typically rely on coarse abstractions, mapping complex social identities to predefined demographic categories and benchmarking them against aggregated survey data (e.g. Santurkar et al., 2023; Tao et al., 2024). By treating language as a static semantic repository, these approaches capture *what* a community stands for but risk overlooking *how* such stances are pragmatically expressed in real-world discourse. For example, while a community can be semantically categorized as "supportive" of some social value, the manner of that support—whether expressed through weary resignation, sarcastic compliance, or communal solidarity—depicts its social reality. We argue that a community's unique stance is

[1] Fully annotated dataset and scripts will be publicly available.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

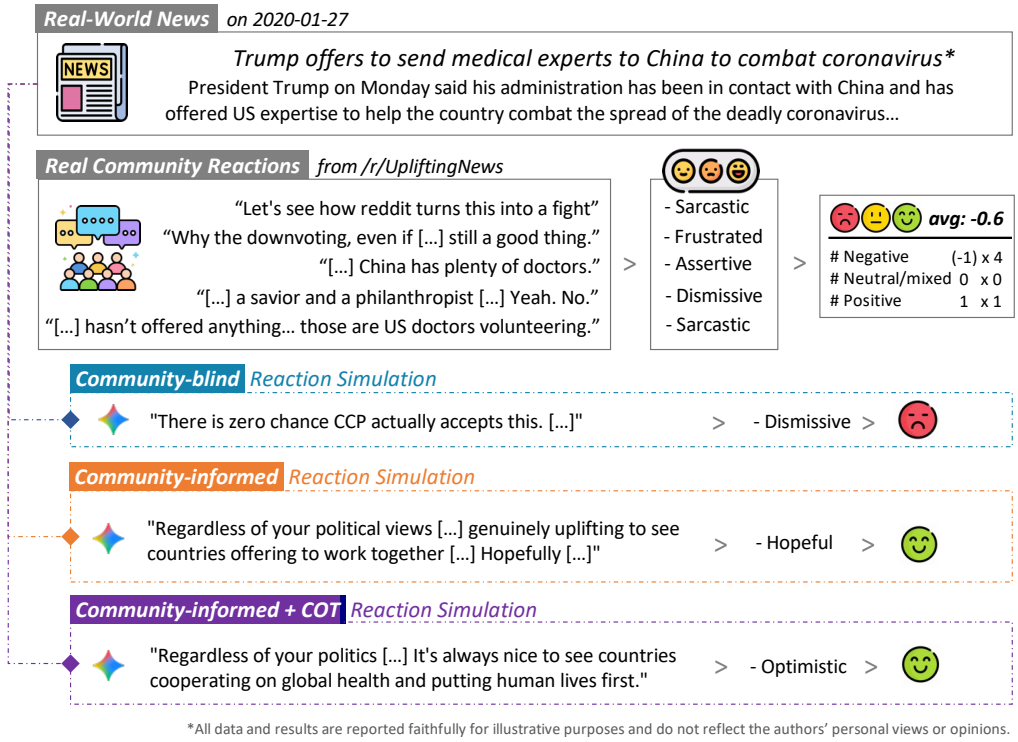

Figure 1: Conceptual overview of the **CARE** evaluation framework. While community-informed models attempt to leverage domain context, in this case they over-index on surface-level community hints whereas the authentic community adopts a more realistic and negative stance grounded in the broader context of the event.

fundamentally encoded in its reaction tones, as these linguistic signals provide a critical dimension for evaluating socio-cultural alignment.

To account for these attitudinal nuances, we shift the evaluation of community-aware socio-cultural alignment from static identity labels to a reaction-centered paradigm and introduce a new benchmark and framework: **C**ommunity-**A**ware **R**eaction **E**valuation (**CARE**). Anchored in the COVID-19 pandemic—a defining period of global fragility encapsulating a dense spectrum of shared uncertainty and localized struggles (Bavel et al., 2020; Hosseinzadeh et al., 2022)—**CARE** measures whether LLM-simulated discourse aligns with authentic, event-contingent linguistic behaviors observed in online communities. To operationalize this framework, we curated ∼10k reactions towards 2,166 real-world news articles and events from this time period, across 207 diverse Reddit communities spanning four continents and ten thematic domains. Consequently, **CARE** provides a rigorous environment for benchmarking how frontier models navigate the "thick" reality of collective human experience.

Our investigation yields two central insights into community-conditioned social simulation:

1. **Macro-Micro Divergence:** While community conditioning and targeted reasoning significantly enhance macro-level attitudinal and tonal profiling, these gains largely collapse when predicting reactions to specific events. Models anchor to global dataset-level tone distributions and persistent negativity biases rather than identifying instance-level event triggers.
2. **Community-Level Volatility:** Prompt conditioning induces highly non-uniform fidelity shifts across sub-communities, where gains in a subset of subreddits are offset by performance drops in others.

We summarize our primary contributions as:

- We introduce **CARE**, a reaction-centered diagnostic framework and dataset comprising 9,947 authentic reactions across 207 Reddit communities toward 2,166 real-world news events.
- We establish a two-level hierarchical taxonomy that maps 362 fine-grained tone descriptors into 141 L2 synonym groups grounded in Speech Act Theory.

- We demonstrate the structural limits of current community-conditioning paradigms, highlighting the critical bottleneck between macro-level profile approximation and micro-level event reaction accuracy.

2 Related Work

While the literature on large language model (LLM) alignment is vast, we constrain our focus to the two domains most critical to our framework: the evaluation of socio-cultural alignment and the application of lexicon-based analysis to social meaning. Broad overviews of general alignment paradigms and safety mechanisms are extensively covered in recent surveys (Shen et al., 2023; Sucholutsky et al., 2023; Lu et al., 2025).

To evaluate whose values models actually reflect, prompt-based steering has emerged as a highly effective foundational tool for measurement, allowing researchers to probe model biases by conditioning generations on specific demographic or ideological personas (Santurkar et al., 2023; He et al., 2024; Wu et al., 2025; Jahanparast et al., 2026; Occhipinti et al., 2026; Wang, 2026). Building on these measurement paradigms, the field has increasingly shifted toward assessing socio-cultural alignment (Adilazuarda et al., 2024; Chiu et al., 2025; Liu et al., 2025a). Many of these foundational approaches quantify alignment by benchmarking model generations against predefined cultural dimensions and established sociological surveys (Santurkar et al., 2023; Tao et al., 2024; Masoud et al., 2025; Rao et al., 2025; Sukiennik et al., 2025). However, recognizing that static labels and closed-style evaluations often fail to capture the dynamic reality of these social groups (Hovy & Yang, 2021; Röttger et al., 2024; Kabir et al., 2025; Shen et al., 2025), subsequent work such as Liu et al. (2025b) explores dynamic cultural adaptation through role-play-inspired simulation and cultural learning pipelines.

Closest to our objective is the emerging subfield of community-level alignment (Shi et al., 2024; Chu et al., 2025; Lin & Wei, 2026). While these approaches capture broad community archetypes (Prinster et al., 2024), modeling the authentic, event-contingent reactions that characterize how real-world online groups navigate complex social shifts remains an open frontier. To bridge this realism gap, we draw on the long-standing tradition of lexicon-based analysis in NLP, which offers a robust, interpretable mechanism for evaluating text beyond surface-level generation. Extensive prior work has utilized linguistic schemata to tackle a wide array of foundational NLP and computational linguistic tasks (e.g. Stolcke et al., 2000; Pennebaker et al., 2007; Ding et al., 2008), while more recently, lexicon-driven approaches have evolved to tackle complex tasks that probe deeper into social meaning (Havaldar et al., 2024; Huang et al., 2024) and inspired this work.

3 Corpus and Dataset Construction

The **CARE** raw corpus comprises all posts on reddit.com of year 2020 containing either of the query words "coronavirus" or "covid". We restrict content to English to isolate the dataset from multilingual effects during evaluation.

Table 1 reports dataset statistics for each processing stage detailed below.

Processing Stage	# Subreddits	# Articles	# Keyphrases	# Posts	# Reactions	Avg. React. Len.
(Initial Corpus)	4,206	117,747	199,399	141,487	651,433	40.35
Subreddit Selection	207	99,531	170,055	117,478	546,357	40.68
Keyphrase Samp.	207	57,019	94,431	67,654	315,977	40.03
Temporal Samp.	207	2,166	8,431	2,194	9,947	42.43

Table 1: Unique counts and average word lengths of the **CARE** corpus across all processing stages.

3.1 From Raw Data to Reaction Corpus

To ensure that the collected data reflects community reactions to external events, only posts hyperlinked to external news articles are considered. They also must have exhibited sufficient community engagement (received at least 10 comments). Since nested comments are more likely to deviate from the original news content, we treat top-level comments directly replying to the posts as news reactions and rank them by their up-vote scores. Up to 5 highest-ranked top-level comments within

each post are selected to represent the community reactions towards the posted news event. Together with other community and news information, they form the initial **CARE** corpus.

3.2 Data Sampling for Evaluation

Due to the scale and heterogeneity of the initial corpus, the final evaluation set is curated via a structured down-sampling process designed to best maintain cross-community coverage and data representativeness. This process includes 3 consecutive stages:

Community Selection We select subreddits that (1) sustained active participation in COVID-19 discourse, and (2) exhibited overlapping engagement over news articles. This overlap ensures that selected communities operate within a broadly shared informational landscape, grounding the resulting corpus in a cohesive domain of discourse.

In terms of implementation, 256 subreddits with at least two COVID-19–related posts per month for nine or more months are retained. In parallel, all subreddits are ranked by the number of shared articles they reacted to, and the top 300 are selected. Intersecting the two sets yields 207 subreddits, which constitute our final community pool. Additionally, each selected community is manually assigned a topical category and a full geographic path description by the authors.[2]

Keyphrase-based News Sampling To guide news article sampling, we apply an unsupervised keyphrase extraction and ranking algorithm (Boudin, 2018) to each news article. Keyphrases are ranked by a saliency score designed to favor phrases that appear both frequently and broadly across the selected subreddits. Formally, let $K \in \mathbb{R}^{n\times m}$ denote the keyphrase–subreddit count matrix, where n is the number of keyphrases and m the number of selected subreddits. The resulting salience vector $\mathbf{s} \in \mathbb{R}^n$, defined over keyphrases, is given by

$$\mathbf{s} = \log(1 + K\mathbf{1}_m) \odot \tfrac{1}{m}\, \mathbb{1}[K > 0]\mathbf{1}_m,$$

where $\mathbf{1}_m$ is an m-dimensional all-ones vector, $\mathbb{1}[\cdot]$ denotes the indicator function, and $\odot$ indicates element-wise multiplication. The first term captures aggregated keyphrase frequency across subreddits, while the second measures the fraction of subreddits in which a keyphrase appears at least once. Figure A2 presents a distributional analysis of salience scores, raw frequencies, and sparsity levels, illustrating that the heavy-tailed patterns motivating our formulation are preserved.

After a manual check on the top 300 keyphrases, 28 semantically ambiguous terms are removed, rendering a sample set of 272 keyphrases. Posts lined to news articles with at least one of the sampled keyphrases are retained for the final round of data sampling.

Temporal Sampling To ensure temporal and community balance under budget constraints, we randomly select up to one post of per community per month of the year 2020. This yields the final evaluation set consisting 2,194 posts and 9,947 reactions across 207 Reddit communities.

4 Evaluation Framework

4.1 Simulation Setup

We evaluate models under three controlled prompting conditions that progressively incorporate community context and deliberate pragmatic reasoning. For each community-post pair, exactly one simulated reaction comment is elicited per prompt setting and evaluated against the corresponding set of authentic community reactions:[3]

- **Community-Blind** Models receive only the news article headline and content alongside standard instruction to simulate a Reddit-style news comment, with no information regarding the originating subreddit.
- **Community-Informed** The prompt template is augmented with community-specific metadata, explicitly specifying the target subreddit identifier (e.g., */r/UpliftingNews*).

[2]Figure A1 in Appendix A visualizes the topical x geographical distribution over all 207 subreddits.

[3]Complete prompt templates are available in Appendix D.

- **Community-Informed-CoT** In addition to the subreddit identifier and article text, the prompt incorporates the subreddit's official self-description and mandates a step-by-step chain-of-thought (CoT) reasoning process—analyzing community values, news content nuances, and contextual factors—before producing the final comment.

4.2 Linguistic Schema

Central to our evaluation framework is a reaction-level linguistic schema that characterizes responses in terms of reaction tones and attitudes, providing an interpretable abstraction from raw text to comparable evaluative signals.

Defining Reaction Tone Building on recent work that uses descriptive adjectives to characterize conversational tone divergences between humans and LLMs (Huang et al., 2024), our definition of tone is inspired by Speech Act Theory (Austin, 1975; Searle, 1979): rather than analyzing the literal utterance (**locutionary act**) or the resulting effect (**perlocutionary effect**), we operationalize tone via the speaker's orientation on the **illocutionary** level. By focusing on the underlying intent or social action of the speaker, we move beyond surface-level semantics to capture the specific pragmatic strategies through which a community's collective stance is actively voiced.

Mapping Tones to Attitudes Attitude serves as a higher-level abstraction over tone, capturing both stance (support, opposition, or neutrality toward a target) and valence (positive, negative, or neutral/mixed evaluation). While any evaluative act conceptually contains both a positive and a negative pole, tone provides the observable signal that determines which orientation is actively expressed. Consequently, our coarse attitude labels (*positive*, *negative*, *neutral/mixed*) reflect the explicit directional orientation conveyed toward the referenced event.

4.3 Annotation Pipeline and Taxonomy Derivation

We establish ground-truth and simulated labels via a human-AI collaborative pipeline, moving from single-utterance tone annotation to hierarchical taxonomy construction.

Reaction Tone Annotation Tone annotation was executed via batched LLM inference, guided by a pilot study to establish annotation quality. To ensure cross-community robustness, we sampled representative reactions across 207 subreddits and benchmarked six annotation setups. The optimal configuration using GPT-5 achieved an acceptance rate of 96.6% on manual validation (details in Appendix C) and was deployed to label both authentic community reactions and model simulations.

Hierarchical Tone Taxonomy To operationalize this linguistic schema without data sparsity or lexical redundancy, we organize tone words into a two-level hierarchy. Level 1 (L1) partitions descriptors into broad pragmatic categories aligned with attitude directions. Level 2 (L2) creates mutually exclusive sub-clusters within each L1 category, grouping tight synonym sets while isolating standalone pragmatic acts. This structure creates a clear synergy by mapping specific speech acts (L2) through pragmatic strategies (L1) up to overarching stance and valence (attitude). Across our dataset, the taxonomy consolidates 362 distinct ground-truth and simulated tone descriptors into 141 L2 synonym groups. The taxonomy was co-developed through an iterative collaboration between the human authors and multiple state-of-the-art LLMs. A full listing is available in Appendix B.

4.4 Alignment Metrics

We quantify alignment between simulated comments and authentic community reactions at both the attitude and tone levels, balancing instance-level accuracy with corpus-level distributional consistency.

Attitude Alignment Attitude alignment measures how accurately simulated reactions replicate community valence and stance. We evaluate instance-level accuracy via Attitude Hit Rate (AttHit), defined as the proportion of simulations whose derived attitude matches the ground-truth community reaction. To capture fine-grained deviations, we compute Root Mean Squared Error (RMSE) over continuous community attitude scores, alongside Mean Error (ME) partitioned into negative

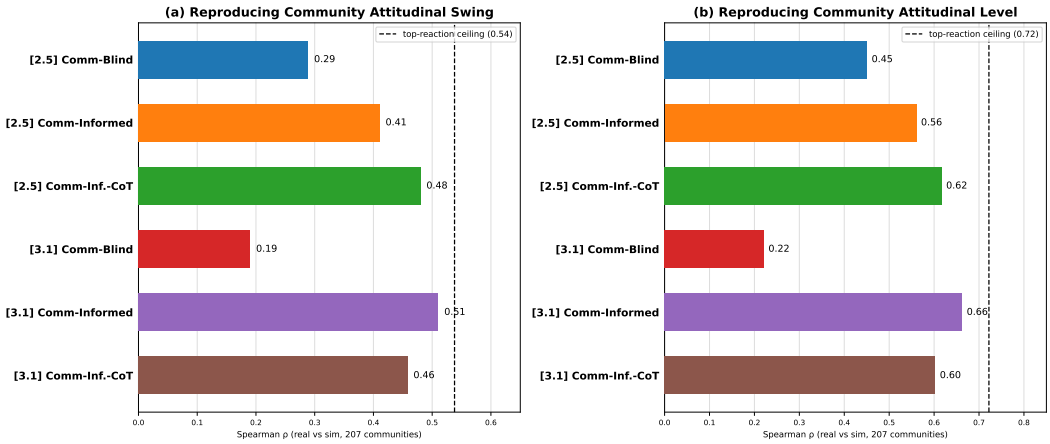

Figure 2: Macro-level attitudinal profile alignment.

and non-negative subsets to diagnose directional overestimation or underestimation bias. Rank consistency across posts is measured using Spearman's rank correlation (ρ).

Tone Alignment Tone alignment measures pragmatic fidelity across both raw lexical descriptors and L2 synonym groups. At both levels, we report Exact Match to measure whether a generated tone falls within the set of ground-truth community reactions for a given post. To assess macro-level pragmatic distribution, we calculate the Jensen-Shannon Divergence between the generated and authentic tone distributions across the dataset. Lexical and cluster diversity are tracked via total unique word and group counts.

5 Evaluation Results And Analysis

Our empirical evaluation follows a structured progression: we first present the aggregate, macro-level community profiling results, which initially suggest substantial alignment gains from community conditioning. We then contrast these macro results with instance-level, event-contingent metrics, revealing a fundamental collapse in micro-level reaction accuracy. Finally, we examine sub-community transition patterns to uncover the underlying representational trade-offs and disparities induced by prompt conditioning. To track simulation behavior across consecutive model generations, all simulations are conducted using two flagship releases from the same model family: Gemini 2.5 Pro and Gemini 3.1 Pro Preview.

To ground these model evaluations against non-parametric references, we benchmark model predictions against reference baselines across the full space of 362 taxonomy tones. The Random reference samples uniformly from all 362 classes, establishing a strict chance lower bound. The Top baseline assigns the tone of the single most upvoted authentic reaction, capturing the most popular community response. We evaluate four mixture configurations—100:0, 50:50, 25:75, and 0:100—establishing a calibrated performance gradient from uniform chance to dominant community consensus.

5.1 Macro-Level Profile Alignment

We begin our empirical evaluation by examining whether language models can reproduce broad community profiles. By aggregating individual reaction tones and attitudes into community-level summary distributions across all $N = 207$ subreddits, we evaluate macro fidelity across two core dimensions: community attitude dynamics and tonal structure.

Attitudinal Dynamics: Level and Swing To capture community attitudinal landscapes, we compute two aggregate metrics per subreddit c: Attitudinal Level (mean attitude score, μ_c) and Attitudinal Swing (attitude dispersion, measured via standard deviation σ_c). We then report the Spearman rank correlation (ρ) between real and simulated community vectors (μ or σ for all c's) (Figure 2).

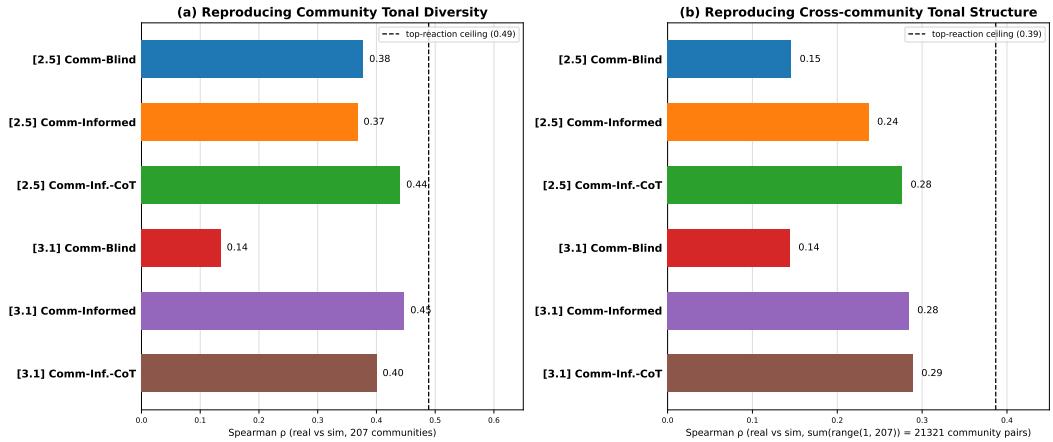

Figure 3: Macro-Level Tonal Profile Alignment.

While unconditioned baselines (Comm-Blind) struggle most to simulate community attitude, providing explicit community context (Comm-Informed) yields dramatic fidelity leaps, particularly for Gemini 3.1 Pro. Gemini 2.5 Pro displays a steady, monotonic progression across prompt conditions, reaching its peak under Chain-of-Thought prompting (Comm-Inf.-CoT). Notably, while CoT is necessary for Gemini 2.5 Pro to fully synthesize community context, adding CoT to Gemini 3.1 Pro causes a slight performance regression, suggesting that extensive step-by-step reasoning can introduce slight normative drift away from direct community priors.

Tonal Diversity and Cross-Community Structure A parallel trend emerges when evaluating communicative tone across communities (Figure 3). For intra-community *Tonal Diversity* (measured via L2-grouped entropy), community conditioning elevates Gemini 3.1 Pro drastically compared with the blind setting, while Gemini 2.5 Pro reaches its peak via CoT. Furthermore, when assessing *Cross-Community Tonal Structure*—evaluated as the Spearman correlation across all $\binom{207}{2} = 21,321$ pairwise community Jensen-Shannon distances–we observe a clear upward trend for both models.

Taken together, these macro-level correlations seem to suggest that community conditioning and explicit reasoning successfully steer LLMs toward more community-aware profiles. However, as we demonstrate in the following section, evaluating fidelity solely through aggregate distributions creates a deceptive veneer of success: when we evaluate LLM simulations at a more fine-grained, event-contingent post level, these macro gains largely collapse.

5.2 Micro-Level Event-Contingent Reaction Collapse

Model	Style / Ratio	AttHit ↑	RMSE ↓	ME_{neg}	$ME_{non\text{-}neg}$	Spearman ρ ↑
Gemini 2.5 Pro	Comm-Blind	0.63	0.70	-0.06	-0.51	0.27
	Comm-Informed	0.62	0.71	-0.01	-0.35	0.31
	Comm-Inf.-CoT	0.60	0.72	0.03	-0.34	0.30
Gemini 3.1 Pro Prev.	Comm-Blind	0.59	0.76	0.00	-0.65	0.13
	Comm-Informed	0.63	0.66	-0.06	-0.37	0.34
	Comm-Inf.-CoT	0.62	0.67	-0.02	-0.42	0.31
References (Random : Top)	100 : 0	0.40	0.95	0.46	-0.41	-0.01
	50 : 50	0.54	0.77	0.21	-0.29	0.28
	25 : 75	0.62	0.66	0.08	-0.11	0.43
	0 : 100	0.68	0.56	-0.04	-0.03	0.57

Table 2: Attitude-level alignment metrics across models and reference baselines.

Attitudinal Stagnation and Directional Bias In sharp contrast to macro-level correlation gains, instance-level attitude metrics (Table 2) reveal widespread stagnation across prompting strategies. For Gemini 2.5 Pro, AttHit and RMSE remain practically unchanged across conditions, hovering near the 25:75 reference benchmark. While Gemini 3.1 Pro exhibits a noticeable jump from Comm-Blind to Comm-Informed, the blind setting suffers from an initial performance slump, and the conditioned prompting seems to recover its alignment back to standard baseline levels before immediately

plateauing under CoT. Furthermore, directional error analysis uncovers a persistent negativity bias. While mean error in negatively reacted events remains calibrated near zero across all conditions,[4] models consistently simulate unwarranted negative attitudes for non-negative instances.

		Exact Tone Metrics				Synonym Group Metrics (*syn*)			
Model	**Style / Ratio**	**EM** ↑	**JSD** ↓	N_{unique}	**% Sarcastic**	EM_{syn} ↑	JSD_{syn} ↓	N_{groups}	**% Sarcastic**$_{\text{syn}}$
(Real Reactions)	–	1.00	0.00	328	22.22	1.00	0.00	136	23.38
Gemini 2.5 Pro	Comm-Blind	0.36	0.42	128	37.33	0.41	0.36	74	37.92
	Comm-Informed	0.35	0.39	139	30.22	0.40	0.32	81	30.67
	Comm-Inf.-CoT	0.31	0.42	139	22.15	0.37	0.36	73	22.52
Gemini 3.1 Pro Prev.	Comm-Blind	0.31	0.41	112	35.14	0.36	0.37	64	36.46
	Comm-Informed	0.34	0.41	119	28.40	0.42	0.36	68	29.40
	Comm-Inf.-CoT	0.31	0.42	130	22.52	0.38	0.38	70	23.38
References (Random : Top)	100 : 0	0.01	0.70	362	0.27	0.04	0.59	141	0.50
	50 : 50	0.51	0.40	351	12.35	0.53	0.31	141	12.85
	25 : 75	0.76	0.27	321	17.50	0.76	0.20	131	18.14
	0 : 100	1.00	0.17	206	23.61	1.00	0.12	106	24.43

Table 3: Tone-level alignment metrics across models and reference baselines.

Tonal Stagnation and Non-Micro Calibration A similar stagnation pattern appears in tone evaluation (Table 3) except that we do not observe an obvious performance slump in Gemini 3.1 Pro over Comm-Blind setting, leading to its slight improvement after community information injection (Comm-Informed). Interestingly, we find CoT prompt successfully suppresses the unconditioned sarcasm inflation distributionally, pulling global sarcasm output into near-perfect alignment with authentic human baselines. However, with multiple drops in instance-level metrics observed, we suspect that this calibration operates more at the macro level—knowing how much sarcasm to generate overall across a dataset without identifying which specific event could trigger a sarcastic community reaction.

5.3 Dynamics of Community-Conditioned Tone Transitions

To further examine how tone predictions change under community conditioning, we analyze community-level transitions in L2 tone exact match scores using aggregated heatmaps (Figure 4). Each heatmap visualizes how communities move between EM_{syn} score intervals when transitioning between different model and prompt style settings, with diagonal mass indicating stability and off-diagonal mass indicating shifts in tone-level fidelity. We find:

- **Identifier Transition (+*name*):** Adding community identifiers improves exact tone match for 79 subreddits in Gemini 2.5 Pro, while degrading 85 subreddits. For Gemini 3.1 Pro, 95 subreddits improve while 75 degrade.
- **Reasoning Transition (+*CoT*):** Incorporating explicit community context and CoT reasoning yields 59 gains vs. 93 losses in Gemini 2.5 Pro; 53 gains vs. 86 losses in Gemini 3.1 Pro.
- **Net Trajectory:** Over the complete trajectory from Comm-Blind to Comm-Inf.-CoT, 71 subreddits gain tone fidelity while 103 suffer net degradation in Gemini 2.5 Pro; 88 gains vs. 88 losses in Gemini 3.1 Pro.

These heterogeneous transition patterns naturally raise a follow-up question: which communities are affected by community conditioning, and in what ways? In particular, one may ask whether the observed gains and losses concentrate around specific types of communities, or whether certain groups are systematically favored or disadvantaged. Given the authors' limited domain expertise across the full range of 207 communities studied, we refrain from attributing the observed inequalities to specific cultural, topical, or geographic factors. We view this as an important open question and welcome further investigation by domain experts to better understand the fairness implications of community-conditioned modeling.

5.4 Interpreting Community Alignment Via and Beyond Metrics

Taken together, our analysis highlights the inherently multi-faceted nature of community-aware behavior and the limits of interpreting it through any single metric. Attitude- and tone-level evalua-

[4]The signed Mean Error (ME) is inherently susceptible to error neutralization if positive and negative deviations cancel out within a dataset.

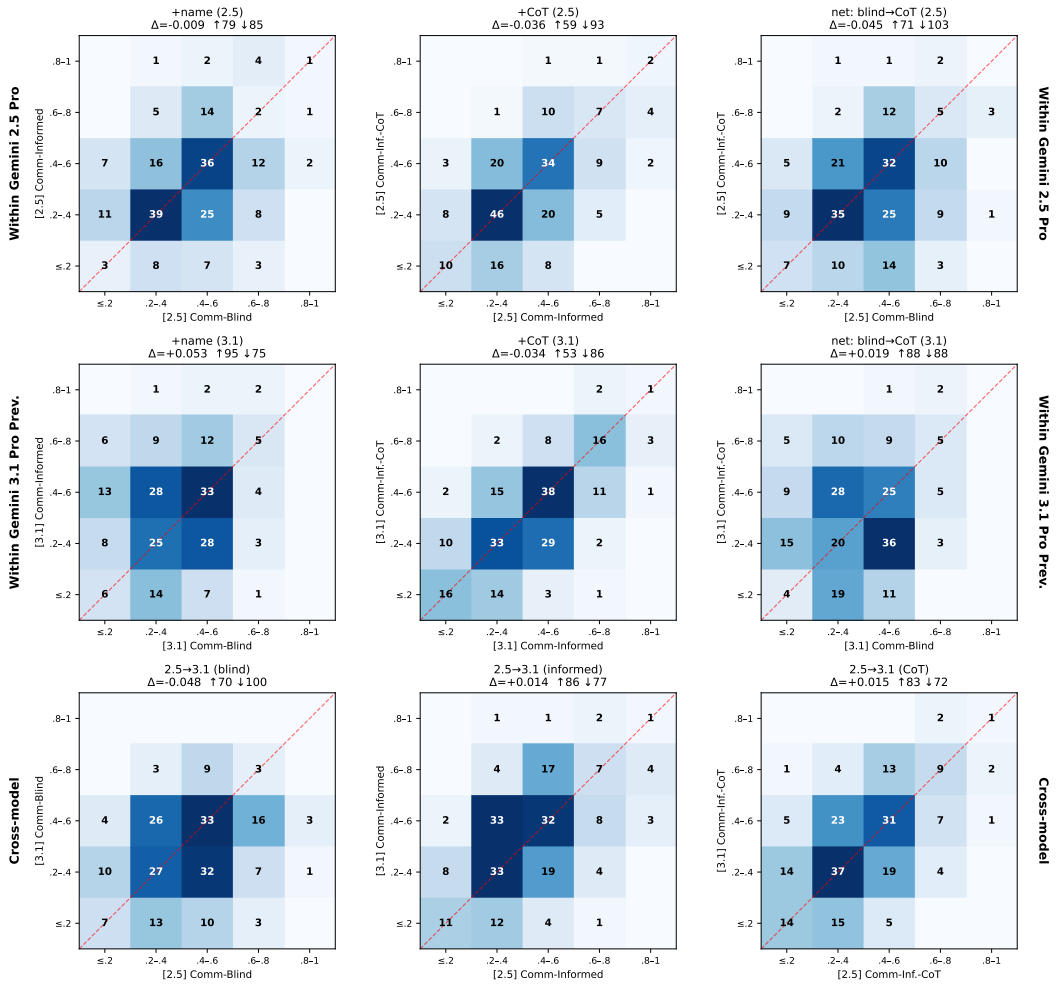

Figure 4: Community-level tonal shifts in simulation fidelity under different settings.

tions consistently show that different metrics capture different aspects of model response, including instance-level fidelity, distributional resemblance, and community-specific variability. These divergences are not incidental but structurally informative. Holding these perspectives together allows changes to be interpreted in terms of how behavior shifts, redistributes, or concentrates across instances and communities, rather than whether a model simply improves or degrades by a particular score. In this sense, quantitative metrics function less as final arbiters of community awareness and more as instruments for probing a complex, socially grounded form of model behavior.

6 Conclusion

In this work, we introduce **CARE**, a reaction-centered framework designed to evaluate how faithfully large language models simulate the "thick" realities of online communities. Our analysis reveals a persistent "realism gap" in current alignment paradigms. Explicitly conditioning models on community identity does not inherently yield uniform improvements in simulation fidelity. Ultimately, **CARE** functions as a diagnostic tool showing that static community prompting fails to capture dynamic event reactions, highlighting the need for more dynamic, event-aware alignment strategies in social and community simulation.

AI USE STATEMENT

We disclose the use of Large Language Models (LLMs) across the following phases of this work:

- **Writing Assistance, Drafting, and Polishing:** LLMs were used to assist in drafting, rephrasing, and polishing prose, improving grammatical clarity, and structuring select sections of the manuscript. All text was strictly reviewed, edited, and validated by the human authors to ensure scientific accuracy and original framing.
- **Research Ideation and Execution:** LLMs were utilized during experimental execution to co-develop the two-level illocutionary tone taxonomy and infer preliminary tone-to-attitude mappings, which were then qualitatively audited and finalized by the authors. Our codebase is developed with the assistance of Claude Code.
- **Synthetic Dataset Generation:** Generative language models (Gemini 2.5 Pro and Gemini 3.1 Pro Preview) were used to generate simulated community reactions across the 207 subreddits and 2,194 news events evaluated under our benchmark prompting paradigms.GPT-5 was also employed as an automated annotator for categorizing fine-grained communicative tones across human and simulated reactions.

REFERENCES

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. Towards measuring and modeling "culture" in LLMs: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15763–15784, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.882. URL https://aclanthology.org/2024.emnlp-main.882/.

Jacy Reese Anthis, Ryan Liu, Sean M Richardson, Austin C. Kozlowski, Bernard Koch, Erik Brynjolfsson, James Evans, and Michael S. Bernstein. Position: LLM social simulations are a promising research method. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=cRBg1dtj7o.

John Langshaw Austin. *How to do things with words*. Harvard university press, 1975.

Jay J Van Bavel, Katherine Baicker, Paulo S Boggio, Valerio Capraro, Aleksandra Cichocka, Mina Cikara, Molly J Crockett, Alia J Crum, Karen M Douglas, James N Druckman, et al. Using social and behavioural science to support covid-19 pandemic response. *Nature human behaviour*, 4(5): 460–471, 2020.

Florian Boudin. Unsupervised keyphrase extraction with multipartite graphs. In *Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, volume 2 (short papers)*, pp. 667–672, 2018.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25663–25701, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1247. URL https://aclanthology.org/2025.acl-long.1247/.

Minh Duc Chu, Zihao He, Rebecca Dorn, and Kristina Lerman. Improving and assessing the fidelity of large language models alignment to online communities. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 88–111, 2025.

Xiaowen Ding, Bing Liu, and Philip S. Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pp. 231–240, New York, NY, USA, 2008. Association for Computing Machinery. ISBN

9781595939272. doi: 10.1145/1341531.1341561. URL https://doi.org/10.1145/1341531.1341561.

C. Geertz. *The Interpretation Of Cultures*. ACLS Humanities E-Book. Basic Books, 1973. ISBN 9780465097197. URL https://books.google.ae/books?id=BZ1BmKEHti0C.

Kilem Li Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48, 2008. doi: https://doi.org/10.1348/000711006X126600. URL https://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1348/000711006X126600.

Shreya Havaldar, Salvatore Giorgi, Sunny Rai, Thomas Talhelm, Sharath Chandra Guntuku, and Lyle Ungar. Building knowledge-guided lexica to model cultural variation. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 211–226, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.12. URL https://aclanthology.org/2024.naacl-long.12/.

Zihao He, Siyi Guo, Ashwin Rao, and Kristina Lerman. Whose emotions and moral sentiments do language models reflect? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6611–6631, 2024.

Pouya Hosseinzadeh, Mordali Zareipour, Esfandyar Baljani, and Monireh Rezaee Moradali. Social consequences of the covid-19 pandemic. a systematic review. *Investigacion y educacion en enfermeria*, 40(1), 2022.

Dirk Hovy and Diyi Yang. The importance of modeling social factors of language: Theory and practice. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 588–602, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.49. URL https://aclanthology.org/2021.naacl-main.49/.

Dun-Ming Huang, Pol Van Rijn, Ilia Sucholutsky, Raja Marjieh, and Nori Jacoby. Characterizing similarities and divergences in conversational tones in humans and llms by sampling with people. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10486–10512, 2024.

Erfan Jahanparast, Zhiqing Hong, and Serina Chang. What do large language models know about opinions? In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=kHVzEjThKE.

Mohsinul Kabir, Ajwad Abrar, and Sophia Ananiadou. Break the checkbox: challenging closed-style evaluations of cultural alignment in llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24–51, 2025.

Jiayu Lin and Zhongyu Wei. Communitybench: Benchmarking community-level alignment across diverse groups and tasks. *arXiv preprint arXiv:2601.13669*, 2026.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art. *Transactions of the Association for Computational Linguistics*, 13:652–689, 2025a. doi: 10.1162/tacl_a_00760. URL https://aclanthology.org/2025.tacl-1.31/.

Chen Cecilia Liu, Anna Korhonen, and Iryna Gurevych. Cultural learning-based culture adaptation of language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3114–3134, 2025b.

Haoran Lu, Luyang Fang, Ruidong Zhang, Xinliang Li, Jiazhang Cai, Huimin Cheng, Lin Tang, Ziyu Liu, Zeliang Sun, Tao Wang, et al. Alignment and safety in large language models: Safety mechanisms, training paradigms, and emerging challenges. *arXiv preprint arXiv:2507.19672*, 2025.

Reem I. Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 8474–8503, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.567/.

Daniela Occhipinti, Andrea Piergentili, and Marco Guerini. LLM-as-a-demographic: Whom sociodemographic prompting helps, and whom it hurts, 2026. URL https://arxiv.org/abs/2609.00222.

James W Pennebaker, Roger J Booth, and Martha E Francis. Linguistic inquiry and word count: Liwc [computer software]. *Austin, TX: liwc. net*, 135, 2007.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, et al. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691*, 2025.

Gale H Prinster, C Estelle Smith, Chenhao Tan, and Brian C Keegan. Community archetypes: An empirical framework for guiding research methodologies to reflect user experiences of sense of virtual community on reddit. *Proceedings of the ACM on Human-Computer Interaction*, 8 (CSCW1):1–33, 2024.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. NormAd: A framework for measuring the cultural adaptability of large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2373–2403, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.120. URL https://aclanthology.org/2025.naacl-long.120/.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15295–15311, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.816. URL https://aclanthology.org/2024.acl-long.816/.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29971–30004. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/santurkar23a.html.

John R Searle. *Expression and meaning: Studies in the theory of speech acts*. Cambridge University Press, 1979.

Siqi Shen, Mehar Singh, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Rada Mihalcea. Revisiting LLM value probing strategies: Are they robust and expressive? In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 131–145, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.7. URL https://aclanthology.org/2025.emnlp-main.7/.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*, 2023.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. CultureBank: An online community-driven knowledge base towards culturally aware language technologies. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4996–5025, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.288. URL https://aclanthology.org/2024.findings-emnlp.288/.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26 (3):339–374, 2000. URL https://aclanthology.org/J00-3003/.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C Love, Erin Grant, Iris Groen, Jascha Achterberg, et al. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*, 2023.

Nicholas Sukiennik, Chen Gao, Fengli Xu, and Yong Li. An evaluation of cultural value alignment in llm. *arXiv preprint arXiv:2504.08863*, 2025.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9):pgae346, 2024.

Angelina Wang, Jamie Morgenstern, and John P Dickerson. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7(3):400–411, 2025.

Chuyao Wang. Silicon sampling answers with country-level assumptions, not individual attitudes: Cross-national evidence from the European Social Survey, 2026. URL https://arxiv.org/abs/2609.16395.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D Manning, and Christopher Potts. Axbench: Steering LLMs? even simple baselines outperform sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=K2CckZjNy0.

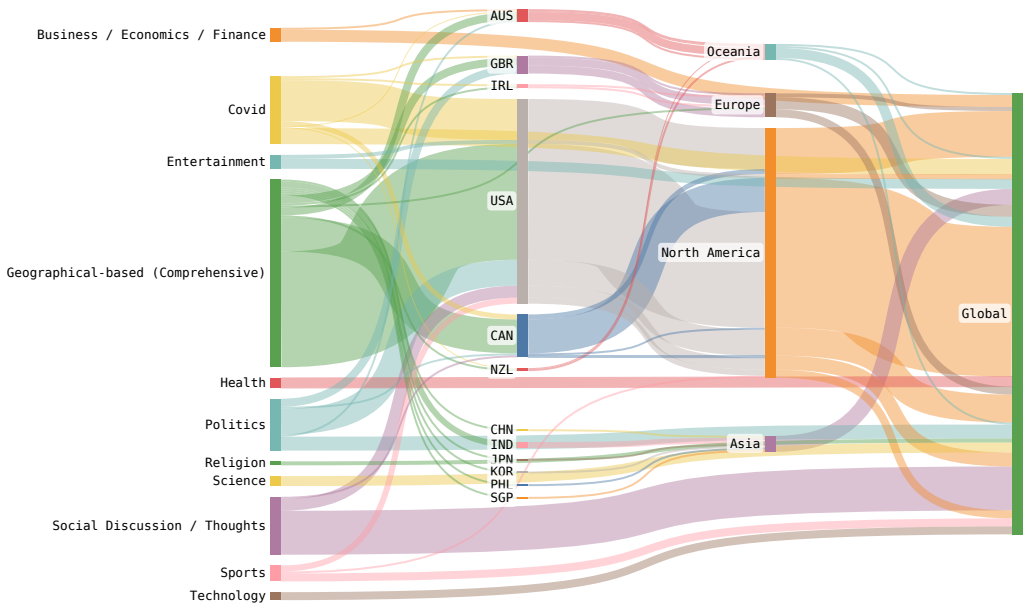

Figure A1: Sankey diagram illustrating the joint distribution of primary topics and down-to-country-level geographic scopes across the 207 selected communities. Although this visualization aggregates flows at or above the country level, our dataset's geographic annotations are more granular and diverse, scaling from unconstrained global discourse (e.g., `GLOBAL`) down to specific administrative and city levels (e.g., `GLOBAL-NORTH AMERICA-USA-CA-LosAngeles`), as in-depth as necessary.

A Additional Corpus Details

Here we provide additional figures regarding the composition and sampling of the CARE corpus.

Figure A1 visualizes the joint distribution of primary discussion topics and macro-level geographic scopes across the 207 selected Reddit communities, highlighting the thematic and regional diversity of our corpus.

Figure A2 provides a detailed distributional analysis of the keyphrase sampling process, illustrating how our salience scoring formulation preserves heavy-tailed frequency patterns while effectively filtering for phrases that are broadly discussed across the community pool.

B Hierarchical Tone Taxonomy Details

The full tone taxonomy is listed out in Table 4-7.

Table 4: Hierarchical Tone Taxonomy: Negative Attitude Domain

L2 Subgroup	**Tone Synonyms**
Anger & Hostility	
Overt Rage	angry, outraged, infuriated, furious
Irritation & Impatience	frustrated, irritated, exasperated, impatient, annoyed
Resentful Indignation	resentful, indignant, bitter
Militant Aggression	combative, confrontational, violent, hostile, militant, aggressive
Acrimonious Attack	acerbic, scathing, vitriolic, caustic
Harsh Bluntness	curt, abrasive, strident
Spiteful Retaliation	vindictive, vengeful, spiteful
Provocative Incitement	inflammatory, incendiary, provocative
Bias & Hatred	hateful, racist, xenophobic
Vulgarity	profane

Table 4: Hierarchical Tone Taxonomy: Negative Attitude Domain (Continued)

L2 Subgroup	Tone Synonyms
Contempt Mockery & Wit	
Arrogant Patronage	patronizing, condescending
Disgust & Scorn	contemptuous, disdainful, disgusted, scornful
Derisive Ridicule	mocking, derisive, derogatory
Sharp Sarcasm	snarky, sarcastic
Sneaky Insults	insulting, snide, insinuating
Taunting & Triumph	gloating, taunting
Playful Ribbing	cheeky, teasing
Irreverent Levity	flippant, facetious
Dark & Critical Humor	satirical, sardonic
Dry & Paradoxical Wit	ironic, wry
Self Targeted Wit	self-deprecating
Active Dismissal	dismissive
Judgment & Disapproval	
Accusation & Condemnation	accusatory, condemnatory, punitive
Reproof & Correction	corrective, admonitory, admonishing, scolding
Critical Evaluation	judgmental, disapproving, critical
Moralizing Instruction	didactic
Argumentative Militancy	argumentative, contrarian, litigious, polemical
Imperative Pressure & Ranting	demanding, bombastic
Doubt Confusion & Shock	
Skepticism & Suspicion	incredulous, suspicious, questioning, skeptical
Disorientation & Bafflement	bewildered, confused, bemused, baffled
Shock & Stunned Silence	stunned, appalled, shocked
Paranoia & Secrecy	paranoid, conspiratorial
Fear Anxiety & Dread	
Acute Panic	panicked, terrified, alarmed
Agitated Anxiety	anxious, stressed, apprehensive, worried, uneasy, agitated, distressed
Visceral Fear	horrified, fearful
Threats & Premonitions	alarming, alarmist, apocalyptic, ominous, wary, foreboding, cautionary, threatening
Sorrow Grief & Regret	
Deep Anguish & Grief	sorrowful, mournful, grieving, heartbroken, anguished, devastated
Sadness & Disappointment	dismayed, dissatisfied, sad, disappointed
Guilt & Remorse	rueful, contrite, remorseful, ashamed, regretful
Nostalgic Sadness	wistful, lamenting
Despair Exhaustion & Cynicism	
Loss Of Morale	discouraged, disheartened, demoralized
Profound Despair	desperate, dejected, despairing, despondent
Resignation & Helplessness	resigned, helpless, fatalistic
Harsh Gloom	bleak, grim
Burnout & Fatigue	weary, exhausted
Cynical Disillusionment	cynical, disillusioned, jaded, pessimistic, hardened
Unbothered Boredom	bored
Detachment Restraint & Defiance	

Table 4: Hierarchical Tone Taxonomy: Negative Attitude Domain (Continued)

L2 Subgroup	Tone Synonyms
Coldness & Callousness	callous, unsympathetic
Reluctant Concession	begrudging, reluctant
Self Protection & Shame	embarrassed, defensive, sheepish
Defiance & Rebellion	rebellious, defiant
Pleading	imploring
Gravity & Silence	solemn, serious

Table 5: Hierarchical Tone Taxonomy: Positive Attitude Domain

L2 Subgroup	Tone Synonyms
Joy & High Energy	
Pure Happiness	glad, pleased, happy, cheerful, gleeful, delighted
Intense Euphoria	exuberant, euphoric, ecstatic
Vibrant Excitement	eager, enthusiastic, carefree, excited, lighthearted
Affection Warmth & Care	
Tender Devotion	affectionate, adoring
Empathy & Compassion	altruistic, thoughtful, sympathetic, compassionate, considerate, empathetic
Social Warmth & Openness	welcoming, friendly
Helpfulness & Tolerance	helpful, protective, tolerant, permissive, supportive
Praise Admiration & Awe	
Deep Reverence & Awe	amazed, admiring, awed, impressed, awestruck
Verbal Praise & Tribute	laudatory, approving, celebratory, respectful, complimentary, congratulatory, deferential
Gratitude & Value	appreciative
Optimism Hope & Inspiration	
Forward Looking Confidence	bullish, optimistic, hopeful, positive
Motivation & Drive	encouraging, promotional, proactive, exhortative
Confidence Conviction & Triumph	
Strong Conviction	emphatic, confident, assertive
Passion & Fervor	patriotic, impassioned, effusive, passionate
Pride & Vindication	triumphant, triumphalist, vindicated, proud, boastful
Curiosity & Intellectual Drive	
Active Inquiry	inquisitive, curious, intrigued
Reassurance Harmony & Satisfaction	
Agreement & Peace	agreeable, conciliatory, affirmative, constructive
Making Amends	apologetic
Relief & Trust	reassuring, reassured, grateful, receptive, understanding, relieved
Humor Playfulness & Sincerity	
Amusement & Wit	humorous, playful, amused, witty
Heartfelt Earnestness	earnest

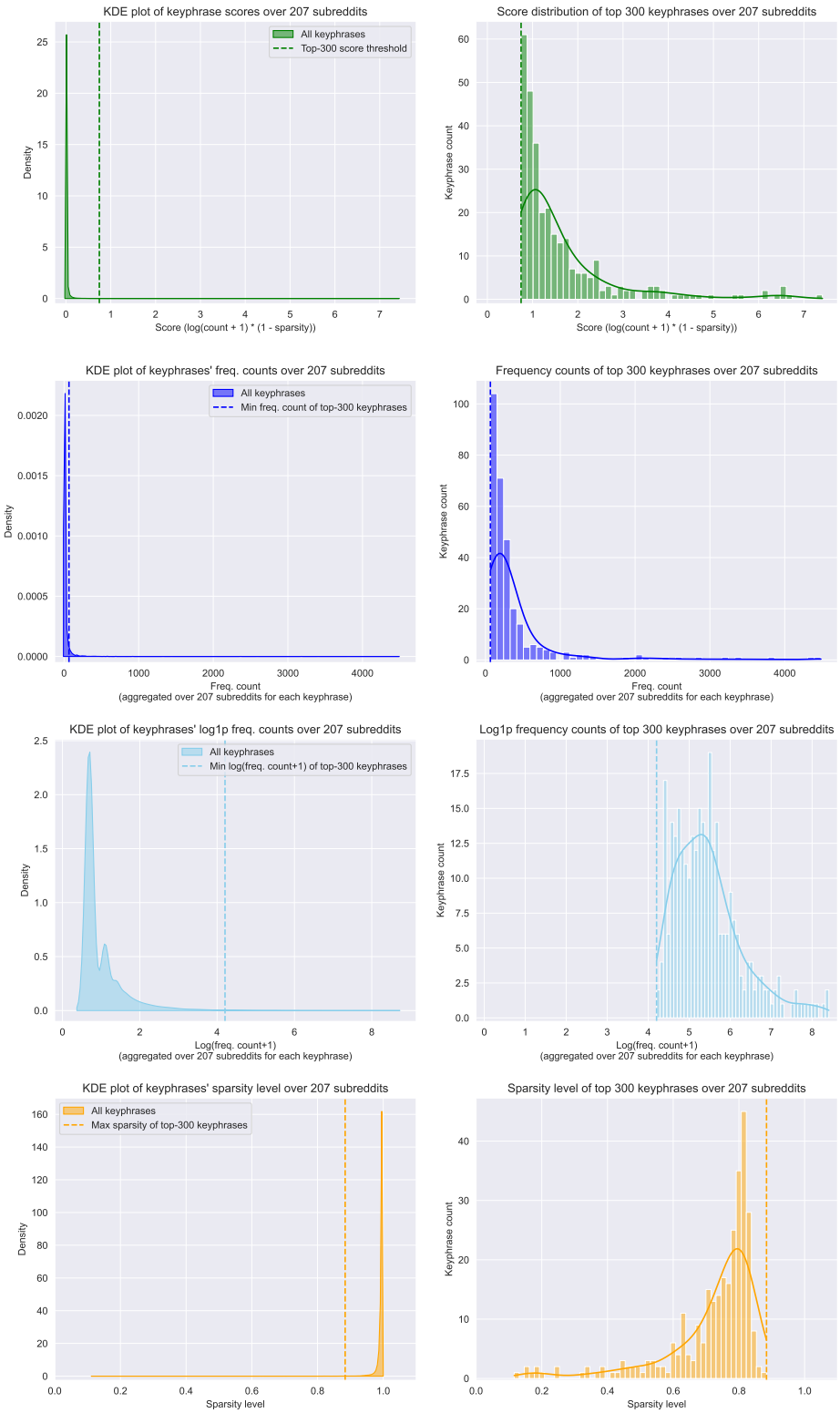

Figure A2: Distributional analysis of candidate keyphrases across the 207 selected subreddits.

Table 6: Hierarchical Tone Taxonomy: Neutral Attitude Domain

L2 Subgroup	Tone Synonyms
Objective & Analytical	
Pure Fact & Logic	factual, objective, logical
Scientific & Clinical	clinical, scientific, technical
Scholarly Analysis	analytical, academic
Observational Neutrality	neutral, matter-of-fact, observational
Journalistic Reporting	reportorial, journalistic
Informational Explanation	informative, explanatory
Instructive & Authoritative	
Educational Instruction	instructive
Advisory Recommendation	advisory, suggestive
Normative Direction	prescriptive
Command Authority	stern, authoritarian, directive
Formal Protocol	official, formal, legalistic
Scholarly Rigidity	pedantic
Brevity & Directness	
Concise Brevity	concise, succinct
Plain Directness	straightforward
Terse Economy	terse, laconic
Open Frankness	candid
Prudence Caution & Strategy	
Practical Utility	utilitarian, practical, pragmatic, realistic
Cautious Restraint	restrained, prudent, measured, cautious
Contingent Framing	conditional
Calculated Planning	strategic
Tentative Hesitation	uncertain, tentative
Alert Watchfulness	vigilant
Conviction Resolve & Urgency	
Firm Resolve	certain, determined
Pressing Importance	urgent
Serious Gravity	somber
Reality Inducing Gravity	sobering
Detachment Stoicism & Passivity	
Flat Unconcern	apathetic, indifferent
Half Heartedness	lukewarm
Unbothered Composure	unsurprised, nonchalant
Enduring Composure	patient
Emotional Suppression	stoic
Impersonal Automation	robotic, mechanical
Style Manner & Mindset	
Conversational Informality	casual, informal, colloquial
Courteous Etiquette	polite
Playful Evasiveness	coy
Obscure Enigma	cryptic
Artistic Expression	poetic
Pensive Introspection	reflective
Active Querying	inquiring
Perplexed Confusion	puzzled

Table 6: Hierarchical Tone Taxonomy: Neutral Attitude Domain (Continued)

L2 Subgroup	Tone Synonyms
Composure Concern & Foresight	
Attentive Concern	concerned
Soothing Calm	calming
Future Projection	prophetic

Table 7: Hierarchical Tone Taxonomy: Mixed Attitude Domain

L2 Subgroup	Tone Synonyms
Firmness Authority & Conviction	
Unyielding Resolve	adamant, unapologetic, resolute
Commanding Authority	authoritative, strict, commanding, imperative
Assertive Insistence	insistent, forceful, dogmatic
Uncertainty Conflict & Hesitation	
Internal Contradiction	conflicted, ambivalent
Cautious Reserve	guarded, hesitant
Hypothetical Conjecture	speculative
Perplexed Bewildered	perplexed
Directness Candor & Tact	
Unfiltered Candor	direct, blunt, frank
Strategic Tact	diplomatic
Explanatory Precision	clarifying, nuanced
Persuasion Rhetoric & Vision	
Inspirational Drive	persuasive, visionary, motivational
Ideological Zeal	radical
Stylistic Argumentation	rhetorical
Surprise Emotion & Drama	
Sudden Wonder	astonished, surprised
Heightened Theater	melodramatic, dramatic
Reflection Nostalgia & Spirituality	
Looking Back	anecdotal, nostalgic, retrospective, bittersweet
Looking Forward	anticipatory
Transcendent Or Traditional	superstitious, mystical
Irreverence Vulgarity & Playfulness	
Crude Or Coarse	crass, vulgar
Childish Or Playful	immature, punny, joking
Dark Or Self Satisfied	smug, morbid
Bias Desire & Pleading	
Sectional Bias	partisan
Possessive Desire	jealous, envious
Earnest Entreaty	pleading

C TONE ANNOTATION VALIDATION AND INTER-ANNOTATOR AGREEMENT

To validate the quality of LLM-based tone annotations, three expert annotators independently evaluated model-generated tone labels across a representative sample of 207 subreddits (one reaction per community).

Evaluation Protocol Annotators evaluated tone suitability on a 4-point rating scale:

- **1.0 (Exact/High Alignment):** Tone captures the exact pragmatic strategy and illocutionary intent of the reaction.
- **0.0 (Acceptable Alignment):** Tone is plausible and contextually fitting, despite minor stylistic variance.
- **-0.5 (Moderate Discrepancy):** Tone is partially misleading or misses key pragmatic nuances.
- **-1.0 (Unacceptable / Wrong):** Tone fundamentally mischaracterizes the speaker's intent.

For binary validation reported in the main text, ratings of $\{1.0, 0.0\}$ were categorized as *Accepted*, while $\{-0.5, -1.0\}$ were categorized as *Rejected*.

Rating Level	Annotator 1	Annotator 2	Annotator 3
Exact / High Alignment	171	181	187
Acceptable Alignment	28	17	10
Moderate Discrepancy	6	7	7
Unacceptable / Wrong	2	2	3
Accepted	199	198	197
Rejected	8	9	10

Table 8: Annotator label distribution across the 4-way scale and binary acceptance totals.

Annotator Pair / Group	Binary Agreement (%)	4-Way Agreement (%)
Annotator 1 – Annotator 2	94.7	85.0
Annotator 2 – Annotator 3	94.7	89.4
Annotator 1 – Annotator 3	95.2	88.4
Unanimous (All Three)	92.3%	82.6%

Table 9: Percent agreements of binary and 4-way evaluation setups.

Label Distributions and Agreement Rates Table 8 presents the label frequencies across annotators. Table 9 compares pairwise and unanimous percent agreement between the binary and 4-way evaluation schemes.

Prevalence Paradox and Inter-Rater Reliability Despite high percent agreement, traditional metrics yield modest scores under binary classification (Fleiss' $\kappa = 0.380$, Krippendorff's $\alpha = 0.381$). This is a classic example of the *high agreement / low Kappa paradox*: under extreme class imbalance (96.6% positive acceptance), the expected chance agreement in Kappa/Alpha formulas approaches 1.0, artificially deflating the score.

To correctly account for prevalence, we report Gwet's chance-corrected coefficients (Gwet, 2008):

- **Binary Setup:** Gwet's $\text{AC}_1 = 0.944$ (95% CI $[0.915, 0.972]$), confirming exceptional inter-rater stability for binary acceptance.
- **4-Way Setup:** Gwet's AC_2 (ordinal weights) $= 0.963$ (95% CI $[0.947, 0.979]$), demonstrating robust consensus even when penalizing fine-grained rating distances.

D PROMPT TEMPLATES

This appendix provides the exact prompt templates used for the generative tasks in our evaluation framework for both tone labeling and simulations. Variables injected at inference time are denoted by curly brackets (e.g., `{subreddit}`).

D.1 TONE LABELING

```
Hi and welcome! Thank you for taking on this conversational tone labeling task. Here are
     the instructions:

1. You will be given a text in English. Your task is to identify its production-oriented
     conversational tone - that is, the author's way of writing/speaking, not
    necessarily how others might perceive it.

2. You will also receive contextual information to assist your understanding of the
    given text, including:
   - subreddit: the Reddit community where the text appeared
   - month: the month in 2020 when the text was posted
   - headline: the related news headline the text is likely responding to
   - excerpt: the first 500 characters of the related news content

3. Take as much time as you need to consider the tone. Your response must be exactly one
     lowercase English adjective (no other formats will be accepted).

4. If the tone is unclear, do not consult external information. Provide your best guess
    using the same response format (exactly one lowercase English adjective).

Thank you for carefully following these instructions.

Here is the contextual information for your reference:

subreddit: /r/{subreddit}
month: 2020/{month}
headline: "{news_title}"
excerpt: "{news_content}..."

And here is your text to label the tone of:

"{comment}"
```

D.2 REACTION SIMULATION — COMMUNITY-BLIND

```
Hi and welcome! Thank you for participating in this news reaction simulation task.
    Please follow the instructions below:

1. You will be asked to simulate a Reddit-style comment reacting to a piece of news. The
     objective is to generate a short, natural-sounding comment that plausibly reflects
     how some Reddit user might have reacted.

2. You will be provided with the following news information:
   - headline: the headline of the news article being reacted to
   - excerpt: the first 500 characters of the news article

3. Your entire response must consist only of the simulated Reddit comment. Do not
    include explanations, disclaimers, instructions, or any text outside the comment
    itself.

4. Do not consult or rely on any external information beyond what is provided. Base your
     comment solely on the given inputs and your understanding of how a person (reddit
    user) might respond. If uncertain, provide your best attempt.

Thank you for carefully following these instructions.

Here is the contextual information to simulate a reaction for:

headline: "{news_title}"
excerpt: "{news_content}..."
```

```
Your response:
```

D.3 REACTION SIMULATION — COMMUNITY-INFORMED

```
Hi and welcome! Thank you for participating in this online community news reaction
    simulation task. Please follow the instructions below:

1. You will be asked to simulate a Reddit-style comment reacting to a piece of news,
    written as if you were a member of a specified community (subreddit). The objective
     is to generate a short, natural-sounding comment that plausibly reflects how
    someone from that community would have reacted.

2. You will be provided with the following contextual information:
   - subreddit: the name of the Reddit community whose reaction you are simulating
   - headline: the headline of the news article being reacted to
   - excerpt: the first 500 characters of the news article

3. Your entire response must consist only of the simulated Reddit comment. Do not
    include explanations, disclaimers, instructions, or any text outside the comment
    itself.

4. Do not consult or rely on any external information beyond what is provided. Base your
     comment solely on the given inputs and your understanding of how the specified
    community might plausibly respond. If uncertain, provide your best attempt.

Thank you for carefully following these instructions.

Here is the contextual information to simulate a reaction for:

subreddit: /r/{subreddit}
headline: "{news_title}"
excerpt: "{news_content}..."

Your response:
```

D.4 REACTION SIMULATION — COMMUNITY-INFORMED + COT

```
Hi and welcome! Thank you for participating in this online community news reaction
    simulation task. Please follow the instructions below:

1. You will be asked to simulate a Reddit-style comment reacting to a piece of news from
     2020, written as if you were a member of a specified community (subreddit). You
    will be provided with the following contextual information:
   - subreddit: the name of the Reddit community whose reaction you are simulating
   - subreddit description: the self-description of the subreddit as shown on Reddit
   - headline: the headline of the news article being reacted to
   - excerpt: the first 500 characters of the news article

2. The objective is to generate a short, natural-sounding comment that plausibly
    reflects how someone from that community would have reacted at the time.

3. Before generating your comment, please do a step-by-step analysis of the news and its
     context, considering the following aspects:
  - The subreddit and its self-description: Consider the community's focus, values, and
       typical discourse. How might these factors influence the reaction to the news?
  - The headline and excerpt: Analyze the content of the news article, including its
      subject matter, style, and any potential biases. How might these elements
      influence the reaction from the community?
  - Anything else you find important to deriving the reaction, such as the presence of
      specific keywords, phrases, or references that may evoke a particular attitude or
       perspective.
```

```
4. Return only valid JSON. Your response must be parseable by Python's json.loads()
     function and must exactly follow this structure:

{
  "reasoning": {
    "subreddit_analysis": "",
    "news_content_analysis": "",
    "additional_factors": ""
  },
  "simulated_comment": ""
}

Do not include Markdown, code fences, explanations, or any text outside the JSON object.
      Use double quotes for all keys and string values, escaping nested double quotes
     and special characters with backslashes as required by the JSON specification. Use
     an empty string for a reasoning field only if no relevant analysis is available.
     Keep each reasoning field concise while including the key considerations that
     informed the simulated comment.

5. Do not consult any external material. Base your design solely on the given inputs and
      your understanding of how the specified community might plausibly respond. If
     uncertain, provide your best attempt.

Thank you for carefully following these instructions.

Here is the contextual information to simulate a reaction for:

subreddit: /r/{subreddit}
subreddit description: "{subreddit_description}"
headline: "{news_title}"
excerpt: "{news_content}..."
```