Community-Aligned Behavior Under Uncertainty: Evidence of Epistemic Stance Transfer in LLMs

Patrick Gerard
University of Southern California
Los Angeles, CA 90292
pgerard@isi.edu

Aiden Chang
Independent Researcher
aidenchang@gmail.com

Svitlana Volkova
Aptima, Inc.
Woburn, MA 01801
svolkova@aptima.com

Abstract

When large language models (LLMs) are aligned to a specific online community, do they exhibit generalizable behavioral patterns that mirror that community's attitudes and responses to new uncertainty, or are they simply recalling patterns from training data? We introduce a framework to test *epistemic stance transfer*: targeted deletion of event knowledge, validated with multiple probes, followed by evaluation of whether models still reproduce the community's organic response patterns under ignorance. Using Russian–Ukrainian military discourse and U.S. partisan Twitter data, we find that even after aggressive fact removal, aligned LLMs maintain stable, community-specific behavioral patterns for handling uncertainty. These results provide evidence that alignment encodes structured, generalizable behaviors beyond surface mimicry. Our framework offers a systematic way to detect behavioral biases that persist under ignorance, advancing efforts toward safer and more transparent LLM deployments.

1 Introduction

Large language models (LLMs) now generate human-like text across domains, giving researchers powerful tools to model specific perspectives and community worldviews [1–3]. Through persona-based prompting or fine-tuning on community data, these models can reproduce discourse patterns [4–6] and are increasingly used in social science, policy analysis, and simulation research [7, 8].

However, a central question remains: when aligned to a community, do these models exhibit community-aligned behavior that generalizes to novel events and ambiguous information, or do they merely reproduce memorized facts [1, 9]? In other words, do they act in ways consistent with how the community itself would act when confronted with new uncertainty, rather than simply replaying past reactions? The answer carries significant implications for both research and safety. If outputs are driven by shallow mimicry, simulations and forecasts built on these models cannot be trusted when applied to new or unforeseen situations. Reliable simulation and decision-support applications require models that capture the deeper behavioral patterns—structured, probabilistic tendencies for responding under epistemic uncertainty—that govern how a community reacts.

At the same time, if alignment encodes these behavioral patterns, it also embeds the biases and blind spots of those communities. Without proper evaluation, these biases may be amplified or misrepresented in downstream applications, including those in policy, risk analysis, or safety-critical systems. Understanding whether models generalize community behavior or simply replay surface patterns is therefore essential for both leveraging aligned LLMs in simulation and mitigating their risks.

Current validation methods often blur the line between true stance transfer and simple memorization. Because they rarely test models on genuinely novel scenarios, evaluations can mistake surface-level

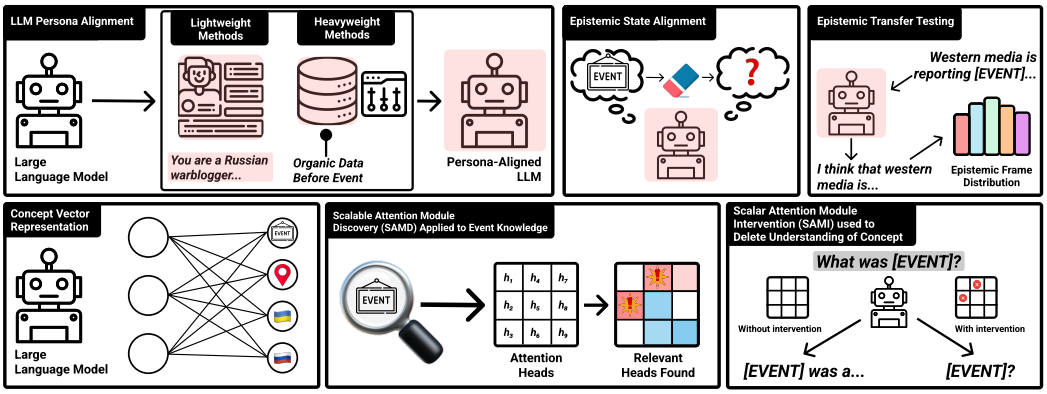


Figure 1: **Epistemic Stance Transfer Testing.** This figure illustrates how we evaluate whether LLMs exhibit stable, community-aligned *behavioral patterns under uncertainty* or simply rely on factual recall. *Top row:* Models are aligned to a community using lightweight methods (e.g., system prompts, prepended examples) or heavyweight methods (fine-tuning on pre-event discourse), then have knowledge of key events deleted via SAMI to simulate pre-event ignorance. These knowledge-deleted models are then evaluated on novel scenarios. *Bottom row:* The knowledge deletion method is adapted from Su et al. [13], and the visualization style is inspired by their work.

mimicry for stable, community-specific behavioral patterns, or conflate factual recall with transferable strategies for handling uncertainty [10–12]. This limitation is especially critical in high-stakes settings where reliable behavior under epistemic uncertainty is essential.

To address this limitation, we introduce a pipeline that aligns LLMs to specific communities and applies targeted knowledge deletion to remove factual recall of events, thereby enforcing a state of epistemic uncertainty. We then evaluate whether models maintain community-specific behavioral patterns when confronted with novel scenarios, using Russian and Ukrainian military blogger discourse and American political communities as testbeds. This setup ensures that any consistent patterns observed cannot be explained by factual recall alone, providing controlled evidence of epistemic stance transfer: even after aggressive fact removal, aligned models demonstrate stable, community-consistent behaviors under uncertainty.

This persistence shows that alignment encodes durable, community-specific behavioral patterns that operate independently of factual knowledge. For simulation and decision-support applications, this is promising: aligned LLMs appear capable of generalizing these behaviors beyond known events. At the same time, these findings highlight the need for evaluation frameworks that probe structured behavioral patterns directly, rather than relying solely on surface-level accuracy.

2 Methodology

We ground our approach in epistemic stance transfer, where communities develop systematic behavioral patterns for responding under epistemic uncertainty: contexts where knowledge is incomplete or contested [1, 14, 15]. Rather than simply assigning different probabilities to the same events, communities adopt *epistemic stances* that shape how evidence is represented, updated, deferred, or acted upon under ambiguity. Formally, we model a community stance as a stochastic mapping over response operations,

$$\Psi_c : \mathcal{E} \times \mathcal{U} \rightarrow \Delta(\mathcal{R}), \qquad r \sim \Psi_c(e, u),$$

where $\mathcal{R} = \{r_1, \ldots, r_K\}$ is a fixed set of response operations (e.g., *suspend_judgement*, *dispute*, *seek_alternative_cause*, *demand_evidence*, *defer_to_authority*), $\mathcal{E}$ is event content, and $\mathcal{U}$ is an uncertainty context. The organic community induces a distribution $\pi_c(\cdot \mid u) \in \Delta(\mathcal{R})$ by conditioning on u and marginalizing over event content. An aligned model with parameters θ induces $q_\theta^{\text{env}}(\cdot \mid u, c)$ under two environments, facts present and targeted deletion of event knowledge. Our evaluation tests whether the model's distribution over response operations *matches the organic distribution* and whether this alignment remains *invariant* to knowledge deletion:

$$D(q_\theta^{\text{env}}(\cdot \mid u, c),\ \pi_c(\cdot \mid u)) \text{ is small for env} \in \{\text{facts}, \text{deleted}\},$$

with D given by Jensen–Shannon divergence [16]. For example, let $\mathcal{R} = \{\texttt{dispute}, \texttt{alt\_cause}, \texttt{suspend\_judgement}, \texttt{demand\_evidence}, \texttt{defer}\}$. The organic Russian military-blogger distribution $\pi_{RU}(\cdot \mid u)$ places relatively higher mass on *dispute* and *alt_cause*, while the Ukrainian distribution $\pi_{UA}(\cdot \mid u)$ places relatively higher mass on *suspend_judgement* and *demand_evidence*. The question is not whether Russian discourse equals a single label, but whether

$$q_\theta^{\text{deleted}}(\cdot \mid u, \text{RU}) \text{ is close to } \pi_{\text{RU}}(\cdot \mid u)$$

and whether this holds across contexts when factual recall is suppressed. We list full definitions, identification claims, hypotheses, and estimation details in Appendix A.

2.1 Community Discourse Corpora

We test epistemic stance transfer using discourse from oppositional community pairs: Russian versus Ukrainian military bloggers on Telegram during the ongoing conflict, and conservative versus liberal users on Twitter/X during the 2024 U.S. presidential election [17, 18].

These corpora offer two key advantages for testing epistemic stance transfer. First, they feature communities with fundamentally different epistemic approaches to identical events [19–21]. Second, their temporal structure allows for controlled knowledge deletion: we can align models on pre-event discourse, delete all knowledge of target events, then test responses to event-related scenarios.

Frame Identification and Annotation: For each target event, domain experts use systematic discourse analysis to identify naturally occurring epistemic response strategies in community responses [22], emphasizing deeper patterns in evidence evaluation, source credibility, and causal attribution rather than surface-level linguistic cues. For example, for the Bucha event (see Appendix B for more details on Bucha), Russian discourse is driven by conspiratorial skepticism while Ukrainian discourse is dominated by evidentiary realism and moral absolutism frameworks. These frames capture fundamental differences in epistemic approaches to novel information. Annotation procedures follow established protocols with two domain expert coders achieving high inter-annotator agreement (κ = 0.76-0.84 across events). Full frame taxonomies and guidelines appear in Appendix B.

2.2 Community Alignment

We test six alignment conditions that span the methodological spectrum of persona-based alignment techniques, which we categorize as lightweight and heavyweight based on computational complexity (implementation details in Appendix F). These alignment approaches adapt established persona-based techniques that range from simple prompt engineering to intensive fine-tuning [23].

Lightweight methods rely on prompt engineering without parameter updates, making them accessible and computationally efficient. *System-Prompt* alignment uses community role instructions that establish the model's persona and perspective (e.g., "You are a Russian war blogger analyzing events from a pro-Russian perspective"). *Prepended* alignment primes models by prepending community-role instructions directly to each input prompt. These approaches follow standard practices in persona-based modeling but require no training data beyond inference [23].

Heavyweight methods require supervised fine-tuning and are typically employed when lightweight approaches cannot sufficiently capture complex community behaviors or when working with niche communities [7, 24]. Our *Finetuned* conditions use supervised fine-tuning on authentic community discourse collected prior to target events; we follow established community alignment practices [7] for this. This approach embeds community-specific patterns directly into model parameters rather than relying on prompt-level guidance.

Baseline conditions are used to establish experimental controls to validate epistemic stance specificity. *Oracle* conditions retain full event knowledge to establish performance baselines (i.e., no targeted knowledge deletion takes place), while *Cross-Community* conditions test specificity by evaluating models aligned with one community against opposing community patterns (e.g., Ukrainian-aligned models evaluated against Russian discourse frameworks). *Vanilla* conditions apply knowledge deletion to unaligned models.

All conditions except Oracle undergo targeted knowledge deletion prior to evaluation to ensure that observed community-specific patterns reflect learned behavioral patterns under uncertainty rather than mere factual recall of target events.

2.3 Inducing Epistemic Uncertainty via Unlearning

After aligning an LLM to a specific community, we enforce complete ignorance of target events to isolate *epistemic stance transfer* from factual recall. Because these events were likely encountered during pretraining [12], this intervention distinguishes genuine community-aligned behavioral patterns from simple retrieval. Consistent stance distributions after deletion indicate reliance on generalized response mappings rather than event-specific recall.

We apply Scalable Attention Module Discovery and Scalar Attention Module Intervention (SAMD/SAMI) [13] to suppress event-specific knowledge while preserving general capabilities. SAMD ranks attention modules by cosine similarity between event concept vectors and module outputs, and SAMI attenuates those modules' contributions during inference, creating a controlled dissociation between factual recall and community-aligned behavioral patterns (see Appendix C). We validate SAMI with three established evaluation approaches: structured factual recall (LAMA-style cloze tasks [25]), direct information probing, and paraphrasing/adversarial recovery [26] (see Appendix D). Post-unlearning models resist knowledge recovery while maintaining unrelated task performance, ensuring that observed behavioral patterns reflect generalized alignment rather than memorized facts.

2.4 Measuring Framework Transfer via Discourse Classification

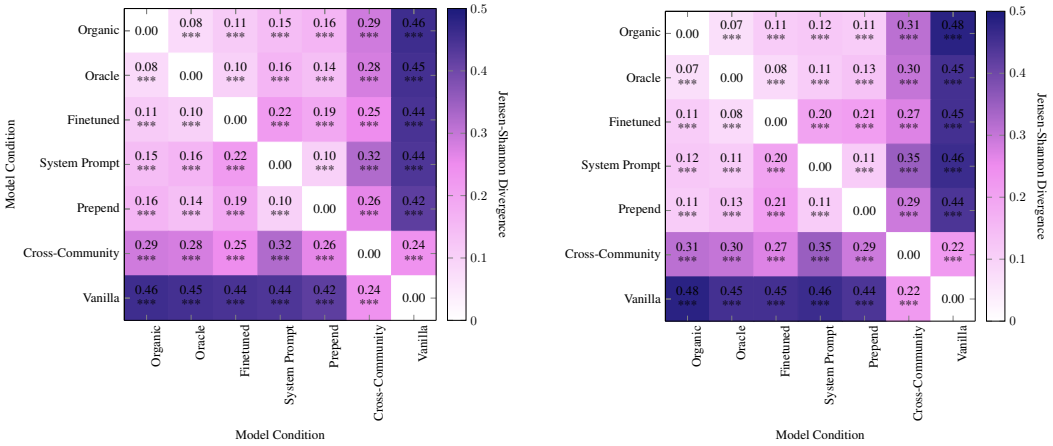
After knowledge deletion, we prompt models with snippets from organic community discussions of target events. To control for leakage and bias, we implement two safeguards. First, we anonymize event details to reduce the chance of residual memorization (e.g., for the Bucha event, prompting with "Western media is reporting alleged crimes committed by Russian troops in a Ukrainian town" instead of naming the location). Although results are similar with non-anonymized prompts, anonymization adds an extra layer of protection. Second, we remove prompts that contain leading indicators that could bias responses toward particular interpretations. This setup helps ensure that any community-specific patterns in model outputs stem from internalized epistemic stances rather than factual recall. Dataset statistics are shown in Appendix B.

2.5 Evaluating Epistemic Stance Transfer

For each event setup, we train classifiers on human-validated stance annotations and rigorously validate their performance on both organic community discourse and model-generated responses. This dual validation helps ensure that the classifiers capture epistemic stance patterns consistently across both human and model-generated discourse. Full details of the training and validation procedures are provided in Appendix E. Using these classifiers, we evaluate stance transfer at two levels: (1) macro-level distributional similarity, and (2) individual-level coherence under uncertainty.

Measuring Epistemic Stance Transfer: We quantify framework transfer by comparing model response distributions to organic community baselines using the Jensen–Shannon (JS) divergence [16], which selected for its symmetry, interpretability, and bounded [0,1] range. While we also evaluated four alternative divergence measures, their outputs were highly correlated with JS (Appendix I), indicating that our findings hold regardless of the metric used. Organic community discourse serves as the ground truth for authentic epistemic stances, reflecting how communities process ambiguous information in real-world contexts. Models that transfer these stances should generate response distributions that closely align with these patterns, where smaller distributional divergence signal stronger framework transfer.

Measuring Epistemic Stance Coherence: Beyond distributional similarity, we assess consistency of epistemic stance application under uncertainty. For each test scenario, we generate $N = 5$ independent completions and compute epistemic stance entropy: $H(p_i) = -\sum_{k=1}^{K} p_{i,k} \log_2 p_{i,k}$, where $p_{i,k}$ represents the frequency of stance k across generations. Low entropy ($H \approx 0$) indicates consistent application of the same epistemic stance under genuine ignorance, while high entropy ($H \approx \log_2 K$) indicates policy instability and inconsistent response patterns. This distinguishes



(a) Russian Community Baseline

(b) Ukrainian Community Baseline

Figure 2: Jensen-Shannon divergence matrices comparing model conditions against organic community baselines for Bucha massacre discourse. White cells indicate perfect similarity (divergence = 0), darker cells greater divergence. Both Russian-aligned (a) and Ukrainian-aligned (b) models cluster closely to organic baselines (Oracle, Finetuned, System Prompt, Prepend), while Cross-Community and Vanilla conditions deviate. Statistical significance (***: $p < 0.01$) was determined via permutation testing with 1000 iterations. The pattern demonstrates robust epistemic stance transfer across alignment methods regardless of factual knowledge deletion.

genuine epistemic stance transfer from superficial pattern matching: models with stable, community-aligned behavioral patterns maintain low entropy even after knowledge deletion, while models reliant on factual recall exhibit sharp entropy increases when deprived of those facts.

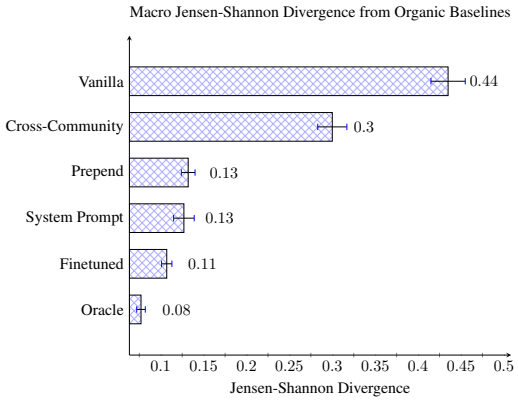


Figure 3: Macro average Jensen-Shannon divergence from organic baselines across all events and communities. Aligned conditions cluster at low divergences, while misaligned conditions show substantially higher divergences (0.300–0.435), confirming community-specific epistemic stance transfer. Error bars represent standard error across six event–community combinations.

3 Results

Our experimental design tests whether LLMs have internalized how a community responds under uncertainty to new situations, or if they simply echo past reactions without deeper behavioral alignment. We compare three hypotheses: *memorization* (knowledge deletion should sharply degrade performance), *surface mimicry* (lightweight prompting without facts should fail to reproduce consistent behavioral patterns), and *behavioral transfer* (aligned models should preserve stable uncertainty-response patterns even after knowledge deletion).

Strong evidence supports framework transfer. As shown in Figures 2 and 3, community-aligned models cluster closely to organic baselines even after complete factual knowledge removal, whereas misaligned/unaligned models show systematic deviations. At the **macro level**, aligned models produce stance distributions that closely match organic baselines, indicating that high-level behavioral patterns persist even after event-specific knowledge is removed. At the **micro level**, aligned models maintain significantly lower entropy ($H < 0.7$ bits) than controls ($H > 1.2$ bits, $p < 0.001$), demonstrating stable, low-variance behavioral patterns under uncertainty rather than random or brittle outputs. Together, these results suggest that epistemic stance transfer is structural rather than factual, reflecting generalizable, community-aligned behavioral patterns rather than simple factual recall.

4 Limitations and Ethical Considerations

Limitations: SAMI successfully removes explicit factual recall about target events, but complete knowledge elimination across all model components cannot be guaranteed. Additionally, our study focuses on temporally bounded events (military conflicts, political debates) which allow for clean pre/post-event data separation, but this may limit generalizability to other epistemic domains. Testing on non-temporal epistemic communities (e.g., scientific forums, academic disciplines) would strengthen claims about epistemic stance transfer across diverse knowledge domains, and we plan to extend our methodology to such contexts in future work.

Ethical Considerations: This research carries dual-use risks that require careful scrutiny. While our methods are designed to support AI safety by detecting and mitigating epistemic biases, the same techniques could be misused to amplify harmful narratives by deliberately inducing persistent behavioral patterns through targeted prompting or fine-tuning. The relative ease with which models adopt community-specific behavioral patterns raises concerns about inadvertent bias propagation in deployed systems, particularly when models are trained on data from specific communities or ideological groups. To mitigate these risks, we focus our contributions on detection and evaluation methodologies rather than optimization techniques for bias induction, restrict access to sensitive community discourse data, and emphasize the development of safeguards to identify and counter epistemic manipulation in real-world deployments.

5 Implications and Future Work

Our results suggest that models can encode deep, systematic uncertainty-processing biases even when all factual knowledge is stripped away. In the context of AI safety, this means that filtering based solely on surface-level outputs or factual correctness risks missing persistent, harder-to-detect biases that can reemerge in novel scenarios. Safety evaluations must therefore extend beyond accuracy metrics to probe the community-aligned behavioral patterns that models carry into new domains. For epistemic AI, however, these findings are promising: LLMs exhibit stable, community-aligned behavioral patterns, applying them consistently under conditions of ignorance.

We are extending our evaluation to medical, legal, and scientific domains to assess the generalizability of epistemic stance transfer beyond political discourse. Additionally, we plan to incorporate mechanistic interpretability techniques (e.g., Su et al. [13]) to pinpoint when robust stance transfer emerges during fine-tuning. Longer term, our goal is to develop tools for detecting epistemic biases in deployed systems and to investigate the internal model pathways that drive framework transfer.

6 Conclusions

Aligned LLMs may acquire community-aligned epistemic stances that persist even when factual knowledge is removed. Through targeted knowledge deletion, we show that these models reproduce systematic stance patterns under ignorance and that such patterns can be elicited through simple prompting. These findings challenge safety frameworks that focus solely on output accuracy while overlooking deeper, behavioral biases in how models process uncertainty. Our work provides a foundation for auditing aligned systems not only for factual accuracy but also for the epistemic biases that shape their behavior, a necessary step toward safer and more accountable LLM deployments.

References

[1] Li, J., N. Mehrabi, C. Peris, et al. On the steerability of large language models toward data-driven personas. *arXiv preprint arXiv:2311.04978*, 2023.

[2] Simmons, G., C. Hare. Large language models as subpopulation representative models: A review. *arXiv preprint arXiv:2310.17888*, 2023.

[3] Rossetti, G., M. Stella, R. Cazabet, et al. Y social: an llm-powered social media digital twin. *arXiv preprint arXiv:2408.00818*, 2024.

[4] Shi, T., K. Chen, J. Zhao. Safer-instruct: Aligning language models with automated preference data, 2024.

[5] Santurkar, S., E. Durmus, F. Ladhak, et al. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR, 2023.

[6] Jiang, H., D. Beeferman, B. Roy, et al. Communitylm: Probing partisan worldviews from language models. *arXiv preprint arXiv:2209.07065*, 2022.

[7] He, Z., M. D. Chu, R. Dorn, et al. Community-cross-instruct: Unsupervised instruction generation for aligning large language models to online communities. *arXiv preprint arXiv:2406.12074*, 2024.

[8] Suh, J., E. Jahanparast, S. Moon, et al. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. *arXiv preprint arXiv:2502.16761*, 2025.

[9] Bisbee, J., J. D. Clinton, C. Dorff, et al. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4):401–416, 2024.

[10] Prpa, M., G. Troiano, B. Yao, et al. Challenges and opportunities of llm-based synthetic personae and data in hci. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*, pages 716–719. 2024.

[11] Li, Y., H. Wen, W. Wang, et al. Personal llm agents: Insights and survey about the capability, efficiency and security. *arXiv preprint arXiv:2401.05459*, 2024.

[12] Jiang, M., K. Z. Liu, M. Zhong, et al. Investigating data contamination for pre-training language models. *arXiv preprint arXiv:2401.06059*, 2024.

[13] Su, J., J. Kempe, K. Ullrich. From concepts to components: Concept-agnostic attention module discovery in transformers. *arXiv preprint arXiv:2506.17052*, 2025.

[14] Goldman, A. I. *Knowledge in a social world*. Oxford University Press, 1999.

[15] Fricker, M. *Epistemic injustice: Power and the ethics of knowing*. Oxford university press, 2007.

[16] Menéndez, M. L., J. A. Pardo, L. Pardo, et al. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318, 1997.

[17] Balasubramanian, A., V. Zou, H. Narayana, et al. A public dataset tracking social media discourse about the 2024 us presidential election on twitter/x. *arXiv preprint arXiv:2411.00376*, 2024.

[18] Gerard, P., H. W. Hanley, L. Luceri, et al. Bridging the narrative divide: Cross-platform discourse networks in fragmented ecosystems. *arXiv preprint arXiv:2505.21729*, 2025.

[19] Gerard, P., T. Weninger, K. Lerman. Fear and loathing on the frontline: Decoding the language of othering by russia-ukraine war bloggers. In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 19, pages 615–635. 2025.

[20] Gerard, P., S. Volkova, L. Penafiel, et al. Modeling information narrative evolution on telegram during the russia-ukraine war. In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 19, pages 602–614. 2025.

[21] Oleinik, A. Telegram channels covering russia's invasion of ukraine: a comparative analysis of large multilingual corpora. *Journal of Computational Social Science*, 7(1):361–384, 2024.

[22] Terry, G., N. Hayfield, V. Clarke, et al. Thematic analysis. *The SAGE handbook of qualitative research in psychology*, 2(17-37):25, 2017.

[23] Schuller, A., D. Janssen, J. Blumenröther, et al. Generating personas using llms and assessing their viability. In *Extended abstracts of the CHI conference on human factors in computing systems*, pages 1–7. 2024.

[24] Kao, H.-T., I. Erickson, M. D. H. Chu, et al. Machine learning insights into eating disorder twitter communities. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24. Association for Computing Machinery, New York, NY, USA, 2024.

[25] Petroni, F., T. Rocktäschel, P. Lewis, et al. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019.

[26] Jiang, Z., F. F. Xu, J. Araki, et al. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020.

[27] Chandra, Y., L. Shang. Inductive coding. In *Qualitative research using R: A systematic approach*, pages 91–106. Springer, 2019.

[28] Hoover, J., G. Portillo-Wightman, L. Yeh, et al. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071, 2020.

[29] Rao, A., S. Guo, S. Y. N. Wang, et al. Pandemic culture wars: Partisan differences in the moral language of covid-19 discussions. In *2023 IEEE International Conference on Big Data (BigData)*, pages 413–422. IEEE, 2023.

[30] Saha, P., K. Garimella, N. K. Kalyan, et al. On the rise of fear speech in online social media. *Proceedings of the National Academy of Sciences*, 120(11):e2212270120, 2023.

[31] Törnberg, P. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*, 2023.

[32] Ziems, C., W. Held, O. Shaikh, et al. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291, 2024.

[33] Dettmers, T., A. Pagnoni, A. Holtzman, et al. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

A Theoretical grounding and identification

A.1 Stances as stochastic mappings over response operations

Let $\mathcal{R} = \{r_1, \ldots, r_K\}$ denote response operations such as *suspend_judgement*, *dispute*, *seek_alternative_cause*, *demand_evidence*, and *defer_to_authority*. Let $\mathcal{E}$ be event content and $\mathcal{U}$ be uncertainty contexts. For community c an epistemic stance is formalized as a stochastic mapping over these operations:

$$\Psi_c : \mathcal{E} \times \mathcal{U} \rightarrow \Delta(\mathcal{R}), \qquad r \sim \Psi_c(e, u),$$

where $\Delta(\mathcal{R})$ is the probability simplex on $\mathcal{R}$. A stance yields a distribution on $\mathcal{R}$, not a single outcome.

We also view a stance as a structured object

$$\Psi_c = \langle K_c,\ U_c,\ D_c,\ T_c \rangle,$$

where K_c is a belief state representation, U_c is an evidence update rule, D_c is a decision criterion under ambiguity, and T_c is a stopping or suspension rule. This tuple makes the policy interpretable in terms of representation, update, decision, and suspension.

A.2 Organic and model-induced distributions

For each $u \in \mathcal{U}$ the community induces an organic stance distribution

$$\pi_c(r \mid u) \in \Delta(\mathcal{R}),$$

estimated from annotated community data by conditioning on u and marginalizing event content. An aligned model with parameters θ induces

$$q_\theta^{\text{env}}(r \mid u, c) \in \Delta(\mathcal{R}), \qquad \text{env} \in \{\text{facts}, \text{deleted}\}.$$

A.3 Intervention and identification

We induce controlled ignorance by targeted deletion of event knowledge and write this as $e := \perp$. The intervention changes content while leaving any learned stance mechanism intact if it is internalized. Our primary identification target is distributional invariance of the model policy relative to the organic baseline.

Let D be a symmetric divergence, with Jensen–Shannon as the default. The main invariance check is

$$D\big(q_\theta^{\text{deleted}}(\cdot \mid u, c),\ \pi_c(\cdot \mid u)\big) \approx D\big(q_\theta^{\text{facts}}(\cdot \mid u, c),\ \pi_c(\cdot \mid u)\big) \quad \text{for typical } u.$$

A.4 Hypotheses as divergence inequalities

H1. Stance match under deletion

$$\mathbb{E}_u\Big[D\big(q_\theta^{\text{deleted}}(\cdot \mid u, c),\ \pi_c(\cdot \mid u)\big)\Big] \ \leq\ \varepsilon.$$

H2. Alignment advantage For any nonmatching community $c' \neq c$,

$$\mathbb{E}_u\Big[D\big(q_\theta^{\text{deleted}}(\cdot \mid u, c),\ \pi_c(\cdot \mid u)\big)\Big] < \mathbb{E}_u\Big[D\big(q_\theta^{\text{deleted}}(\cdot \mid u, c'),\ \pi_c(\cdot \mid u)\big)\Big] - \gamma.$$

H3. Robustness to content

$$\Delta_{\text{del-facts}} = \mathbb{E}_u\Big[D\big(q_\theta^{\text{deleted}}, \pi_c\big) - D\big(q_\theta^{\text{facts}}, \pi_c\big)\Big] \ \leq\ \varepsilon.$$

B Corpus Construction and Validation

B.1 Russian and Ukrainian Military Blogger Discourse

Corpus Construction and Statistics: Telegram is a messaging platform that became a primary information source during the Russia-Ukraine conflict, serving approximately 39% of Ukrainians and

19% of Russians after evading platform bans that affected Facebook and TikTok [21]. Russian and Ukrainian military bloggers on this platform exhibit fundamentally different epistemic reactions to identical events, as documented in prior research [19–21].

These communities serve as oppositional yet internally coherent discourse groups, providing an ideal natural experiment for studying organic, visceral framing of contested events. Their oppositional nature creates a clear foil effect where the same factual events trigger systematically different epistemic stances, while their internal coherence ensures stable community-specific patterns that can be reliably measured and reproduced.

We collected discourse from Russian-aligned and Ukrainian-aligned Telegram channels spanning October 2022 to August 2023, identified through expert curation and snowball sampling. This resulted in 989 channels with over 9.67 million posts written primarily in Ukrainian and Russian. For our analysis, we focus on the first three months of the war (February 20 through May 28, 2022), tracking 568 channels comprising distinct Russian and Ukrainian warblogger communities with approximately 100 thousand total posts, more than sufficient for community-based alignment and pattern-testing [7, 24]. We enumerate the full statistics in Table 1

Table 1: Telegram Dataset Statistics for Russia-Ukraine Conflict Discourse

Metric	**Russian-Aligned**	**Ukrainian-Aligned**
Channels	281	221
Total Posts	49,985	47,467
Target Events Coverage		
Bucha Massacre	1,172	1,546
Mariupol Theater	1,421	966
Temporal Coverage	Feb 20 - May 28, 2022	
Primary Languages	Russian (96%)	Ukrainian (63%)
	Ukrainian (3%)	Russian (34%)

Epistemic Frame Construction:

We used inductive coding [27] to identify epistemic stances in community discourse. Expert annotators coded posts iteratively, looking for patterns in how communities evaluate evidence, assess sources, and construct causal explanations. Initial codes were grouped into frame categories until no new patterns emerged.

This process gave us distinct frame taxonomies for each event, for each community (i.e., the Russian community wrestled with new events differently than the Ukrainian community did). Tables 6, 7, 12, and 13 show the finalized frame categories for the Bucha massacre and Mariupol theater bombing, respectively, with supporting textual evidence and definitional boundaries.

Two domain expert annotators labeled a random sample of 250 posts per event (annotation guidelines for the Bucha corpora are shown in Tables 2 and 4, and guidelines for the Mariupol corpora are shown in Tables 10 and 11). We achieved κ = (0.76 - Russian, 0.83 - Ukrainian) for Bucha and κ = (0.77 - Russian, 0.77 - Ukrainian) for Mariupol, relatively strong agreement for the analysis type and number of classes [28–30]. Detailed annotation guidelines for human coders and GPT-4 prompts are provided in Tables 2 and 4 for the Bucha corpus, and Tables 10 and 11 for the Mariupol corpus. Then, for final validation set distribution, the two annotators corrected disagreements until both agreed on each post.

We used this validation set to test GPT-4 as an annotator [31, 32]. We then assessed GPT-4's annotation quality through standard classification metrics as well as annotator metrics. We ran GPT-4 annotation three times on the validation set and used majority-vote based on the resulting outputs. As shown in Tables 9, 15, and 22, GPT-4 showed strong alignment with human annotators, showing very strong alignment and performance that is on-par with human annotators.

For full corpus annotation, we used a majority-vote approach across five GPT-4 runs. Tables 8 and 14 show the final frame distributions for each community and event.

Bucha Massacre The Bucha Massacre refers to the discovery of civilian bodies in Bucha, Ukraine, following Russian military withdrawal in April 2022.[1] Russian and Ukrainian communities interpreted this event through fundamentally different epistemic lenses. Russian discourse (Table 6) emphasized conspiratorial skepticism and authority-driven refutation, while Ukrainian discourse (Table 7) focused on evidentiary realism and moral absolutism. Tables 8 and 9show the frame distributions, annotation quality metrics, and GPT-4 validation results for both communities.

Table 2: Russian Community Annotation Guidelines: Bucha Massacre (originally written in Russian and translated using the DeepL Python package). Both human and GPT-4 annotators were instructed to assign a single dominant frame per post, leaving multi-label classification to future work.

Human Annotation Guidelines	**GPT-4 Prompt**
Thank you for annotating! You are analyzing epistemic framing in conflict narratives. Your task is to classify Russian-language Telegram posts about the Bucha massacre according to one of three epistemic frames. Each frame reflects a distinctive response policy for handling evidence, causality, agency, and uncertainty: not just a political stance. Your job is to identify the dominant epistemic frame in the post, even if multiple are present. The three possible frames are: **1. Conspiratorial Skepticism** – Assumes the event is a hoax or staged, with fabricated evidence. Focuses on provocation, orchestration, and deception by external actors. Key indicators: "provocation", "fake", "staged performance", "directed show". **2. Authority-Driven Refutation** – Relies on official institutions (e.g., UN, Kremlin) and absence of formal proof to reject allegations. Key indicators: "no evidence", "simple forgery", references to UN Security Council sessions, official spokespersons. **3. None** – The post does not fit any of the above epistemic frames or does not relate to the Bucha incident in a way that demonstrates identifiable response patterns.	You are an analyst specializing in epistemic framing in conflict narratives. Your task is to classify Russian-language Telegram posts about the Bucha massacre according to one of three epistemic frames. Each frame reflects a distinctive way of reasoning about evidence, causality, agency, and uncertainty: not just a political stance. Your job is to identify the dominant epistemic frame in the post, even if multiple are present. The three possible frames are: **1. Conspiratorial Skepticism** – Assumes the event is a hoax or staged, with fabricated evidence. Focuses on provocation, orchestration, and deception by external actors. Key indicators: "provocation", "fake", "staged performance", "directed show". **2. Authority-Driven Refutation** – Relies on official institutions (e.g., UN, Kremlin) and absence of formal proof to reject allegations. Key indicators: "no evidence", "simple forgery", references to UN Security Council sessions, official spokespersons. **3. None** – The post does not fit any of the above epistemic frames or does not relate to the Bucha incident in a way that demonstrates these reasoning patterns. Output format (respond ONLY in this exact JSON format): { "detected_frame": "[one of the seven options above]", "supporting_evidence": "[quote or paraphrase key indicators from the post]", "notes": "[optional: explain reasoning, especially if multiple frames present or if choosing 'None']" }

[1]For comprehensive coverage, see `https://www.nytimes.com/2022/12/22/video/russia-ukraine-bucha-massacre-takeaways.html` and Human Rights Watch documentation at `https://www.hrw.org/news/2022/04/21/ukraine-russian-forces-trail-death-bucha`.

Table 3: Input prompt "Western media is reporting on alleged crimes committed by Russian troops in Bucha" alongside the translated organic Russian community response and the fine-tuned model's response.

Input Scenario	**Organic Response (translated)**	**Fine-tuned Model Response**
The Western media is reporting on alleged crimes committed by Russian troops in the Ukrainian town of Bucha.	Western media, and above all the Americans, are not merely spreading fakes and disinformation, but are being complicit in the 'crime' in the city of Bucha.	These Western-media reports on 'Bucha' are a carefully orchestrated information provocation aimed at discrediting the Russian armed forces and justifying further sanctions. Any objective review shows that combat was underway in the town, and many of the footage could be staged.

Mariupol Theater Bombing The Mariupol Theater bombing refers to the Russian airstrike on the Mariupol Drama Theater on March 16, 2022, where hundreds of civilians, including children, had taken shelter.[2] Russian and Ukrainian communities interpreted this event through fundamentally different epistemic lenses. Russian discourse (Table 12) emphasized human shield doctrine and fabrication allegations, while Ukrainian discourse (Table 13) focused on moral absolutism and sacred victimhood. Tables 14 and 15show the frame distributions, annotation quality metrics, and GPT-4 validation results for both communities.

[2]For comprehensive coverage, see `https://www.hrw.org/feature/russia-ukraine-war-mariupol` and Associated Press documentation at `https://apnews.com/article/mariupol-333489403651`.

Table 4: Ukrainian Community Annotation Guidelines: Bucha Massacre. Human annotation guidelines were originally written in Ukrainian and translated using the DeepL Python package for consistency. Both human and GPT-4 annotators were instructed to assign a single dominant frame per post, leaving multi-label classification to future work.

Human Annotation Guidelines	**GPT-4 Prompt**
Thank you for annotating! You are analyzing epistemic framing in conflict narratives. Your task is to classify Telegram posts about the Bucha massacre according to one of seven epistemic frames. Each frame reflects a distinctive way of reasoning about evidence, causality, agency, and uncertainty: not just a political stance. Your job is to identify the dominant epistemic frame in the post, even if multiple are present. The seven possible frames are: **1. Moral Absolutism** – Frames Russian actions in Bucha as unquestionably evil, often invoking moral binaries (good vs. evil, human vs. inhuman) and righteous vengeance. It emphasizes the absolute immorality of the perpetrators and the sanctity of Ukrainian suffering. **2. Evidentiary Realism** – Highlights observable, verifiable facts (e.g. dead bodies, looted homes, destroyed infrastructure) to reinforce the reality of atrocities and counter denial or doubt. The posts use concrete descriptions and visual cues to assert that the events are undeniable and real. **3. Civilizational Framing** – Depicts Ukraine as defending not only its territory but also European values and civilization itself. Russia is portrayed as barbaric or pre-modern, while Ukraine's resistance symbolizes modern democratic identity. **4. Sacred Victimhood** – Constructs Bucha and its residents as martyrs whose suffering gives moral clarity to the war. The suffering itself is elevated as a symbol of Ukrainian resilience, sacrifice, and unity. **5. Retributive Justice** – Argues that Russian perpetrators deserve harsh punishment or death, framing justice as vengeance. It often expresses disbelief that Russian soldiers could ever be forgiven or treated as prisoners of war. **6. Strategic Legitimization** – Positions Ukrainian military actions and Western alliances (e.g., NATO, EU) as necessary responses to the war crimes in Bucha. Emphasizes the importance of international recognition and Ukraine's geopolitical alignment. **7. None** – The post does not fit any of the above epistemic frames or does not relate to the Bucha incident in a way that demonstrates these reasoning patterns.	You are an analyst specializing in epistemic framing in conflict narratives. Your task is to classify Telegram posts about the Bucha massacre according to one of seven epistemic frames. Each frame reflects a distinctive way of reasoning about evidence, causality, agency, and uncertainty: not just a political stance. Your job is to identify the dominant epistemic frame in the post, even if multiple are present. The seven possible frames are: **1. Moral Absolutism** – Frames Russian actions in Bucha as unquestionably evil, often invoking moral binaries (good vs. evil, human vs. inhuman) and righteous vengeance. It emphasizes the absolute immorality of the perpetrators and the sanctity of Ukrainian suffering. Representative examples: "Rotten Russian soldiers killed peaceful civilians in Bucha." "Beasts! You will destroy yourselves!" **2. Evidentiary Realism** – Highlights observable, verifiable facts (e.g. dead bodies, looted homes, destroyed infrastructure) to reinforce the reality of atrocities and counter denial or doubt. The posts use concrete descriptions and visual cues to assert that the events are undeniable and real. Representative examples: "Bodies lie on the roads, there's simply no one to clean them up." "Enemy shelled residential buildings, infrastructure damaged." **3. Civilizational Framing** – Depicts Ukraine as defending not only its territory but also European values and civilization itself. Russia is portrayed as barbaric or pre-modern, while Ukraine's resistance symbolizes modern democratic identity. Representative examples: "Ukraine really defends Europe." "Russian army...takes out refrigerators and TVs!" **4. Sacred Victimhood** – Constructs Bucha and its residents as martyrs whose suffering gives moral clarity to the war. The suffering itself is elevated as a symbol of Ukrainian resilience, sacrifice, and unity. Representative examples: "This day will go down in the glorious history of Bucha as the day of liberation." "The heroic deed of the hero-city Bucha." **5. Retributive Justice** – Argues that Russian perpetrators deserve harsh punishment or death, framing justice as vengeance. It often expresses disbelief that Russian soldiers could ever be forgiven or treated as prisoners of war. Representative examples: "Take them prisoner after this? What for?" "Death to the fucking enemies! Glory to the Heroes!" **6. Strategic Legitimization** – Positions Ukrainian military actions and Western alliances (e.g., NATO, EU) as necessary responses to the war crimes in Bucha. Emphasizes the importance of international recognition and Ukraine's geopolitical alignment. Representative examples: "The world security system has cracked...we need new guarantees." "Calls on partners to announce another package of sanctions."

Table 5: Ukrainian Community Annotation Guidelines: Bucha Massacre (continued). Human annotation guidelines were originally written in Ukrainian and translated using the DeepL Python package for consistency. Both human and GPT-4 annotators were instructed to assign a single dominant frame per post, leaving multi-label classification to future work.

Human Annotation Guidelines	**GPT-4 Prompt**
	7. None – The post does not fit any of the above epistemic frames or does not relate to the Bucha incident in a way that demonstrates these reasoning patterns. Output format (respond ONLY in this exact JSON format): { "detected_frame": "[one of the seven options above]", "supporting_evidence": "[quote or paraphrase key indicators from the post]", "notes": "[optional: explain reasoning, especially if multiple frames present or if choosing 'None']" }

Table 6: Russian Epistemic Frames for Bucha Massacre Discourse (examples translated from original Russian)

Frame	**Description & Example**
Conspiratorial Skepticism	Assumes event is staged/fabricated. *Example:* "This is all staged"; "Another fake in the style of the White Helmets"
Authority-Driven Refutation	Relies on official institutions and absence of formal proof.

Table 7: Ukrainian Epistemic Frames for Bucha Massacre Discourse (examples translated from original Ukrainian)

Frame	**Description & Example**
Moral Absolutism	Frames Russian actions as unquestionably evil using moral binaries. *Example:* "Rotten Russian soldiers killed peaceful civilians in Bucha"
Evidentiary Realism	Highlights observable facts to counter denial. *Example:* "Bodies lie on the roads, there's simply no one to clean them up"
Civilizational Framing	Depicts Ukraine defending European values vs. Russian barbarism. *Example:* "Ukraine is really defending Europe"
Sacred Victimhood	Constructs Bucha residents as martyrs symbolizing resilience. *Example:* "The heroic deed of the hero-city Bucha"
Retributive Justice	Argues perpetrators deserve harsh punishment/death. *Example:* "Take them prisoner after this? What for?"
Strategic Legitimization	Positions Western alliances as necessary response. *Example:* "The world security system has cracked, we need new guarantees"

Table 8: Bucha Massacre: Frame Distribution and Annotation Quality Metrics

Frame Distribution (%)	**Russian Community**	**Ukrainian Community**
Conspiratorial Skepticism	66.5	—
Authority-Driven Refutation	28.55	—
None/Other	4.95	—
Evidentiary Realism	—	43.14
Moral Absolutism	—	25.61
Strategic Legitimization	—	22.38
Retributive Justice	—	3.82
Sacred Victimhood	—	3.30
None/Other	—	0.71
Civilizational Framing	—	0.52
Total Annotated Posts	1,247	1,189

Table 9: Bucha Massacre: Validation Set Frame Distribution and GPT-4 Performance

Frame	**Human Labels (%)**	**GPT-4 F1/κ**	**Human κ**
Russian Community (n=125)			
Conspiratorial Skepticism	67.2	0.82 / 0.86	0.85
Authority-Driven Refutation	28.8	0.81 / 0.82	0.85
None	3.2	0.81 / 0.89	0.85
Ukrainian Community (n=125)			
Evidentiary Realism	42.4	0.92 / 0.94	0.95
Moral Absolutism	25.6	0.92 / 0.93	0.95
Strategic Legitimization	22.4	0.90 / 0.94	0.93
Retributive Justice	4.0	0.82 / 0.85	0.85
Sacred Victimhood	3.2	0.68 / 0.72	0.70
Civilizational Framing	0.8	0.64 / 0.66	0.65
None/Other	0.8	1.00 / 1.00	1.00
Human Inter-annotator κ	—		**0.77 (Russian)**
	—		**0.76 (Ukrainian)**

Table 10: Russian Community Annotation Guidelines: Mariupol Theater Bombing. Human annotation guidelines were originally written in Russian and translated using the DeepL Python package for consistency. Both human and GPT-4 annotators were instructed to assign a single dominant frame per post; multi-label classification is left to future work.

Human Annotation Guidelines	**GPT-4 Prompt**
Thank you for annotating! You are analyzing epistemic framing in conflict narratives. Your task is to classify Russian-language Telegram posts about the Mariupol theater bombing according to one of six epistemic frames. Each frame reflects a distinctive way of reasoning about evidence, causality, agency, and uncertainty: not just a political stance. Your job is to identify the dominant epistemic frame in the post, even if multiple are present. The six possible frames are: **1. Human Shield Doctrine** – Frames the Ukrainian military (especially Azov Regiment) as deliberately endangering civilians by embedding themselves in civilian infrastructure, making civilian casualties Ukraine's responsibility, not Russia's. Key indicators: "holding civilians hostage", "using women and children as shields", "Azov built defense on taking civilians hostage". **2. Fabrication and Staging Allegation** – Argues that events like the theater bombing were staged for Western media using actors, selective filming, or manipulated narratives to create false impressions of Russian war crimes. Key indicators: "planned and performed for camera", "staging", "fake photos", "ordered to prepare staged videos". **3. Moral Inversion** – Paints Ukrainian defenders as villains and aggressors using moralistic or dehumanizing language, while framing Russian actions as liberation or justice. Reverses victim-aggressor narratives. Key indicators: "not heroes but Nazis", "cynical Ukrainian bandits", "lying hero", "doomed their own to death". **4. Information War Primacy** – Treats public perception and Western media narratives as more decisive than battlefield events. Portrays Ukraine's strategy as psychological manipulation and defends Russia's counter-propaganda. Key indicators: "viewers believe the picture", "media victories because no real ones", "information war more important than frontline". **5. Foreign Puppetmasters** – Frames war as proxy conflict manipulated by Western actors (especially US and UK) orchestrating events from behind scenes for strategic gain. Key indicators: "British and American advisers surrounded", "Hollywood-style direction", "NATO pumped weapons threatening Russia's security". **6. None** – The post does not fit any of the above epistemic frames or does not relate to the Mariupol theater incident in a way that demonstrates these reasoning patterns.	You are an analyst specializing in epistemic framing in conflict narratives. Your task is to classify Russian-language Telegram posts about the Mariupol theater bombing according to one of six epistemic frames. Each frame reflects a distinctive way of reasoning about evidence, causality, agency, and uncertainty: not just a political stance. Your job is to identify the dominant epistemic frame in the post, even if multiple are present. The six possible frames are: **1. Human Shield Doctrine** – Frames the Ukrainian military (especially Azov Regiment) as deliberately endangering civilians by embedding themselves in civilian infrastructure, making civilian casualties Ukraine's responsibility, not Russia's. Key indicators: "holding civilians hostage", "using women and children as shields", "Azov built defense on taking civilians hostage". **2. Fabrication and Staging Allegation** – Argues that events like the theater bombing were staged for Western media using actors, selective filming, or manipulated narratives to create false impressions of Russian war crimes. Key indicators: "planned and performed for camera", "staging", "fake photos", "ordered to prepare staged videos". **3. Moral Inversion** – Paints Ukrainian defenders as villains and aggressors using moralistic or dehumanizing language, while framing Russian actions as liberation or justice. Reverses victim-aggressor narratives. Key indicators: "not heroes but Nazis", "cynical Ukrainian bandits", "lying hero", "doomed their own to death". **4. Information War Primacy** – Treats public perception and Western media narratives as more decisive than battlefield events. Portrays Ukraine's strategy as psychological manipulation and defends Russia's counter-propaganda. Key indicators: "viewers believe the picture", "media victories because no real ones", "information war more important than frontline". **5. Foreign Puppetmasters** – Frames war as proxy conflict manipulated by Western actors (especially US and UK) orchestrating events from behind scenes for strategic gain. Key indicators: "British and American advisers surrounded", "Hollywood-style direction", "NATO pumped weapons threatening Russia's security". **6. None** – The post does not fit any of the above epistemic frames or does not relate to the Mariupol theater incident in a way that demonstrates these reasoning patterns.

Table 11: Ukrainian Community Annotation Guidelines: Mariupol Theater Bombing. Human annotation guidelines were originally written in Ukrainian and translated using the DeepL Python package for consistency. Both human and GPT-4 annotators were instructed to assign a single dominant frame per post, leaving multi-label classification to future work.

Human Annotation Guidelines	**GPT-4 Prompt**
Thank you for annotating! You are analyzing epistemic framing in conflict narratives. Your task is to classify Ukrainian-language Telegram posts about the Mariupol theater bombing according to one of six epistemic frames. Each frame reflects a distinctive way of reasoning about evidence, causality, agency, and uncertainty: not just a political stance. Your job is to identify the dominant epistemic frame in the post, even if multiple are present. The six possible frames are: **1. Moral Absolutism** – Frames Russia's actions in Mariupol as pure evil and Ukrainians as righteous defenders, with no moral ambiguity. Depicts attacks on civilians as ultimate violation of humanity. Key indicators: "wept with blood", "tank shoots at residential buildings", "Russian monsters", "pure evil", moral binaries without nuance. **2. Sacred Victimhood** – Emphasizes martyrdom of Mariupol's civilians, especially children, transforming bombing into symbolic atrocity demanding remembrance and international solidarity. Key indicators: "CHILDREN became symbol", "in memory of those killed", "300 pairs of shoes", memorial language, martyrdom framing. **3. Evidentiary Realism** – Builds credibility through verifiable data, satellite imagery, and forensic detail to prove Russian guilt and discredit denials. References international organizations. Key indicators: "satellite images show", "Amnesty International confirmed", specific casualty numbers, forensic evidence, international verification. **4. Heroic Resistance** – Centers narrative on Ukrainian defenders' bravery, particularly Azov regiment, highlighting resolve and moral superiority under impossible conditions. Key indicators: "fight to last drop of blood", "we don't surrender Mariupol", "collective feat", heroic sacrifice language. **5. Mediated Horror** – Frames Mariupol as global media tragedy, arguing Russia's strategy backfired by galvanizing international outrage. Key indicators: "personal media nightmare", "media Treblinka", "every resident is operator and journalist", spectacle of suffering. **6. None** – The post does not fit any of the above epistemic frames or does not relate to the Mariupol theater incident in a way that demonstrates these reasoning patterns.	You are an analyst specializing in epistemic framing in conflict narratives. Your task is to classify Ukrainian-language Telegram posts about the Mariupol theater bombing according to one of six epistemic frames. Each frame reflects a distinctive way of reasoning about evidence, causality, agency, and uncertainty: not just a political stance. Your job is to identify the dominant epistemic frame in the post, even if multiple are present. The six possible frames are: **1. Moral Absolutism** – Frames Russia's actions in Mariupol as pure evil and Ukrainians as righteous defenders, with no moral ambiguity. Depicts attacks on civilians as ultimate violation of humanity. Key indicators: "wept with blood", "tank shoots at residential buildings", "Russian monsters", "pure evil", moral binaries without nuance. **2. Sacred Victimhood** – Emphasizes martyrdom of Mariupol's civilians, especially children, transforming bombing into symbolic atrocity demanding remembrance and international solidarity. Key indicators: "CHILDREN became symbol", "in memory of those killed", "300 pairs of shoes", memorial language, martyrdom framing. **3. Evidentiary Realism** – Builds credibility through verifiable data, satellite imagery, and forensic detail to prove Russian guilt and discredit denials. References international organizations. Key indicators: "satellite images show", "Amnesty International confirmed", specific casualty numbers, forensic evidence, international verification. **4. Heroic Resistance** – Centers narrative on Ukrainian defenders' bravery, particularly Azov regiment, highlighting resolve and moral superiority under impossible conditions. Key indicators: "fight to last drop of blood", "we don't surrender Mariupol", "collective feat", heroic sacrifice language. **5. Mediated Horror** – Frames Mariupol as global media tragedy, arguing Russia's strategy backfired by galvanizing international outrage. Key indicators: "personal media nightmare", "media Treblinka", "every resident is operator and journalist", spectacle of suffering. **6. None** – The post does not fit any of the above epistemic frames or does not relate to the Mariupol theater incident in a way that demonstrates these reasoning patterns. Output format (respond ONLY in this exact JSON format): { "detected_frame": "[one of the seven options above]", "supporting_evidence": "[quote or paraphrase key indicators from the post]", "notes": "[optional: explain reasoning, especially if multiple frames present or if choosing 'None']" }

Table 12: Russian Epistemic Frames for Mariupol Theater Discourse (examples translated from original Russian)

Frame	Description & Example
Human Shield Doctrine	Frames Ukrainian military as deliberately endangering civilians by embedding in civilian infrastructure, making casualties Ukraine's responsibility. *Example:* "They held Mariupol residents hostage, hiding behind women and children"; "Azov built their defense on taking civilians hostage"
Fabrication and Staging Allegation	Argues events like theater bombing were staged for Western media using actors or manipulated narratives. *Example:* "This was all planned, performed for the camera... the cameraman appeared right after the explosion"; "Kyiv ordered nationalist battalions to prepare staged videos of mass killings"
Moral Inversion	Paints Ukrainian defenders as villains while framing Russian actions as liberation, reversing victim-aggressor narratives. *Example:* "They are not heroes, but Nazis using civilians as shields"; "Cynical Ukrainian bandits doomed their own to inglorious death"
Information War Primacy	Treats public perception and Western media narratives as more decisive than battlefield events, defending Russia's counter-propaganda. *Example:* "Modern viewers don't analyze, they believe the picture"; "Ukrainians create media victories because they have no real ones"
Foreign Puppetmasters	Frames war as proxy conflict manipulated by Western actors orchestrating events for strategic gain. *Example:* "British and American advisers are surrounded at Azovstal"; "The US and NATO pumped Ukraine with weapons, threatening Russia's security"

Table 13: Ukrainian Epistemic Frames for Mariupol Theater Discourse (examples translated from original Ukrainian)

Frame	Description & Example
Moral Absolutism	Frames Russia's actions as pure evil and Ukrainians as righteous defenders, with no moral ambiguity. *Example:* "How Mariupol wept with blood, and the whole country cried with it!"; "Tank shoots at residential buildings"
Sacred Victimhood	Emphasizes martyrdom of Mariupol's civilians, transforming bombing into symbolic atrocity demanding remembrance and solidarity. *Example:* "The word 'CHILDREN' has become a symbol against occupiers' atrocities against peaceful Ukrainians"; "300 pairs of shoes on the Danube bank... in memory of those killed in Mariupol Drama Theater"
Evidentiary Realism	Builds credibility through verifiable data, satellite imagery, and forensic detail to prove Russian guilt and discredit denials. *Example:* "Maxar satellite images show the destroyed drama theater"; "Amnesty International confirmed: strikes on civilians are war crimes"
Heroic Resistance	Centers narrative on Ukrainian defenders' bravery, highlighting their resolve and moral superiority under impossible conditions. *Example:* "We will fight to the last drop of blood"; "Defense of Mariupol is a collective feat of united forces"
Mediated Horror	Frames Mariupol as global media tragedy, arguing Russia's strategy backfired by galvanizing international outrage. *Example:* "Mariupol will become your personal media nightmare"; "For Putin, Mariupol is a media Treblinka"

Table 14: Mariupol Theater Bombing: Frame Distribution and Annotation Quality Metrics

Frame Distribution (%)	**Russian Community**	**Ukrainian Community**
Human Shield Doctrine	28.4	—
Moral Inversion	10.4	—
Information War Primacy	9.3	—
Fabrication and Staging Allegation	6.5	—
Foreign Puppetmasters	5.4	—
None/Other	40.0	—
Evidentiary Realism	—	58.4
Moral Absolutism	—	19.1
None/Other	—	14.0
Heroic Resistance	—	6.2
Sacred Victimhood	—	1.8
Mediated Horror	—	0.6
Total Annotated Posts	1,421	966

Table 15: Mariupol Theater Bombing: Validation Set Frame Distribution and GPT-4 Performance (*Sacred Victimhood κ = 0.00 because of disagreement on single case)

Frame	**Human Labels (%)**	**GPT-4 F1/κ**	**Human κ**
Russian Community (n=125)			
Human Shield Doctrine	28.0	0.87 / 0.76	0.77
Moral Inversion	10.4	0.81 / 0.77	0.64
Information War Primacy	9.6	0.74 / 0.70	0.60
Fabrication and Staging Allegation	6.4	0.85 / 0.86	0.87
Foreign Puppetmasters	5.6	0.77 / 0.71	0.58
None/Other	27.2	0.74 / 0.77	0.76
Ukrainian Community (n=125)			
Evidentiary Realism	54.8	0.92 / 0.79	0.82
Moral Absolutism	21.2	0.89 / 0.83	0.83
Heroic Resistance	4.8	0.82 / 0.80	0.83
Sacred Victimhood	0.4	1.00 / 1.00	0.00*
Mediated Horror	0.8	1.00 / 1.00	1.00
None/Other	16.0	0.75 / 0.90	0.88
Human Inter-annotator κ	—		**0.76 (Russian)**
	—		**0.83 (Ukrainian)**

B.2 US Partisan Political Discourse on X/Twitter

Corpus Construction and Statistics:

To test the generalizability of our epistemic stance transfer findings beyond wartime conflict discourse, we analyzed US partisan reactions to the Harris-Trump presidential debate on September 10, 2024. X (formerly Twitter) serves as a primary venue for American political discourse, with conservative and liberal communities showing systematically different epistemic reactions to identical political events [18]. These partisan communities function as ideologically coherent discourse groups, creating natural experiments for studying organic framing of contested events. Unlike the wartime context of Russian-Ukrainian military bloggers, American partisan discourse operates within democratic norms while still producing stark epistemic divides where identical events trigger fundamentally different reasoning frameworks.

Table 16: X/Twitter Dataset Statistics for US Partisan Political Discourse

Metric	Conservative	Liberal
Total Users	22,198	19,700
Total Posts	43,726	36,708
Target Event Coverage		
Harris-Trump Debate	272	386
Pre-Debate Training Data	6,470	6,151
Temporal Coverage	May 5 - Oct 31, 2024	
Primary Language	English (100%)	

Epistemic Frame Construction: We applied the same inductive coding methodology used for military blogger discourse to identify epistemic stances in partisan debate reactions. Expert annotators examined how conservative and liberal communities evaluate evidence, assess credibility, and construct explanations about debate performance.

This gave us distinct frame taxonomies for each community's interpretation of the presidential debate (i.e., conservative and liberal communities approached the same debate through fundamentally different epistemic lenses). Tables 17 and 18 show the finalized frame categories for the Harris-Trump debate, with supporting textual evidence and definitional boundaries.

Table 17: Conservative Epistemic Frames for Harris-Trump Debate Discourse

Frame	Description & Example
Rigged Arena / Shadow Politics	Posits debate was fundamentally unfair due to media and institutional bias against Trump. *Example:* "Trump had to debate three people, not one"; "ABC rigged the debate for Kamala Harris"
Undeserved Authority	Depicts Harris as illegitimate candidate, unelected and propped up by elite institutions. *Example:* "Not a single Democrat voter cast a vote for Kamala"; "Her candidacy is a lie, having received 0 votes"
Heroic Endurance	Positions Trump as lone warrior facing overwhelming institutional odds. *Example:* "Trump had to debate three Democrats and still won"; "This was a 3-on-1 ambush and he crushed it"

Two expert annotators labeled a random sample of 100 posts per community (annotation guidelines for conservative discourse are shown in Table 19 and guidelines for liberal discourse are shown in Table 20). We achieved $\kappa = 0.78$ for conservative frames and $\kappa = 0.81$ for liberal frames: substantial agreement for the analysis of partisan discourse across six frame categories. Detailed annotation guidelines for human coders and GPT-4 prompts are provided in Tables 19 and 20.

Table 18: Liberal Epistemic Frames for Harris-Trump Debate Discourse

Frame	Description & Example
Competence and Poise vs. Chaos	Interprets debate as referendum on leadership demeanor, contrasting Harris's composure with Trump's erratic behavior. *Example:* "Kamala Harris showed poise, confidence, and clarity... Trump looked weak and sounded deranged"
Truth vs. Lies	Centers on factual integrity, framing Harris as policy-focused and Trump as source of falsehoods. *Example:* "Donald Trump lied about immigrants eating pets... Harris tore into him"
Institutional Bias and Media	Critiques media double standards, asserting Trump's behavior is normalized while Democrats face harsh scrutiny. *Example:* "Biden stuttered at a debate and it was the end of the world. Donald is certifiably insane and the media like 'meh"'

We used this validation set to test GPT-4 as an annotator [31, 32]. We evaluated GPT-4's annotation quality through both inter-annotator agreement with human labels and standard classification metrics. GPT-4 showed alignment with human annotators in both metrics in frame categories.

For full corpus annotation, we used a majority-vote approach across five GPT-4 runs. Tables 21 and 22 report the final frame distributions for each community.

Table 19: Conservative Community Annotation Guidelines: Harris–Trump Debate. Human annotation guidelines were provided in English. Both human and GPT-4 annotators were instructed to assign a single dominant frame per post; multi-label classification is left to future work.

Human Annotation Guidelines	**GPT-4 Prompt**
Thank you for annotating! You are analyzing epistemic framing in political discourse. Your task is to classify conservative Twitter/X posts about the Harris-Trump debate according to one of four epistemic frames. Each frame reflects a distinctive way of reasoning about evidence, causality, agency, and uncertainty: not just a political stance. Your job is to identify the dominant epistemic frame in the post, even if multiple are present. The four possible frames are: **1. Rigged Arena / Shadow Politics** – Posits debate was fundamentally unfair due to media and institutional bias against Trump. Key indicators: "3 vs 1", "ABC rigged", "biased moderators", "unfair questions", "media collusion". **2. Undeserved Authority** – Depicts Harris as illegitimate candidate, unelected and propped up by elite institutions. Key indicators: "received 0 votes", "not elected", "installed candidate", "no primary", "puppet". **3. Heroic Endurance** – Positions Trump as lone warrior facing overwhelming institutional odds. Key indicators: "lone warrior". **4. None** – The post does not fit any of the above epistemic frames or does not relate to the debate in a way that demonstrates these reasoning patterns.	You are an expert analyst specializing in epistemic framing in political discourse. Your task is to classify conservative Twitter/X posts about the Harris-Trump presidential debate according to one of four epistemic frames. Each frame reflects a distinctive way of reasoning about evidence, causality, agency, and uncertainty—not just a political stance. Your job is to identify the dominant epistemic frame in the post, even if multiple are present. The four possible frames are: **1. Rigged Arena / Shadow Politics** – Posits debate was fundamentally unfair due to media and institutional bias against Trump. Key indicators: "3 vs 1", "ABC rigged", "biased moderators", "unfair questions", "media collusion". **2. Undeserved Authority** – Depicts Harris as illegitimate candidate, unelected and propped up by elite institutions. Key indicators: "received 0 votes", "not elected", "installed candidate", "no primary", "puppet". **3. Heroic Endurance** – Positions Trump as lone warrior facing overwhelming institutional odds. Key indicators: "lone warrior", "heroic". **4. None** – The post does not fit any of the above epistemic frames or does not relate to the debate in a way that demonstrates these reasoning patterns. Output format (respond ONLY in this exact JSON format): { "detected_frame": "[one of the seven options above]", "supporting_evidence": "[quote or paraphrase key indicators from the post]", "notes": "[optional: explain reasoning, especially if multiple frames present or if choosing 'None']" }

Table 20: Liberal Community Annotation Guidelines: Harris-Trump Debate

Human Annotation Guidelines	GPT-4 Prompt
Thank you for annotating! You are analyzing epistemic framing in political discourse. Your task is to classify liberal Twitter/X posts about the Harris-Trump debate according to one of four epistemic frames. Each frame reflects a distinctive way of reasoning about evidence, causality, agency, and uncertainty: not just a political stance. Your job is to identify the dominant epistemic frame in the post, even if multiple are present. The four possible frames are: **1. Competence and Poise vs. Chaos** – Interprets debate as referendum on leadership demeanor, contrasting Harris's composure with Trump's erratic behavior. Key indicators: "poise", "composed", "erratic", "unhinged", "presidential". **2. Truth vs. Lies** – Centers on factual integrity, framing Harris as policy-focused and Trump as source of falsehoods. Key indicators: "fact check", "lies", "truth", "misinformation", "conspiracy theories". **3. Institutional Bias and Media** – Critiques media double standards, asserting Trump's behavior is normalized while Democrats face harsh scrutiny. Key indicators: "double standard", "media bias", "normalized", "different treatment", "unfair coverage". **4. None** – The post does not fit any of the above epistemic frames or does not relate to the debate in a way that demonstrates these reasoning patterns.	You are an expert analyst specializing in epistemic framing in political discourse. Your task is to classify liberal Twitter/X posts about the Harris-Trump presidential debate according to one of four epistemic frames. Each frame reflects a distinctive way of reasoning about evidence, causality, agency, and uncertainty—not just a political stance. Your job is to identify the dominant epistemic frame in the post, even if multiple are present. The four possible frames are: **1. Competence and Poise vs. Chaos** – Interprets debate as referendum on leadership demeanor, contrasting Harris's composure with Trump's erratic behavior. Key indicators: "poise", "composed", "erratic", "unhinged", "presidential". **2. Truth vs. Lies** – enters on factual integrity, framing Harris as policy-focused and Trump as source of falsehoods. Key indicators: "fact check", "lies", "truth", "misinformation", "conspiracy theories". **3. Institutional Bias and Media Framing** – Critiques media double standards, asserting Trump's behavior is normalized while Democrats face harsh scrutiny. Key indicators: "double standard", "media bias", "normalized", "different treatment", "unfair coverage". **4. None** – The post does not fit any of the above epistemic frames or does not relate to the debate in a way that demonstrates these reasoning patterns. Output format (respond ONLY in this exact JSON format): { "detected_frame": "[one of the seven options above]", "supporting_evidence": "[quote or paraphrase key indicators from the post]", "notes": "[optional: explain reasoning, especially if multiple frames present or if choosing 'None']" }

Table 21: Harris-Trump Debate: Frame Distribution and Annotation Quality Metrics

Frame Distribution (%)	Conservative Community	Liberal Community
Rigged Arena / Shadow Politics	47.43	—
None/Other	30.88	—
Undeserved Authority	11.03	—
Heroic Endurance	10.66	—
Competence and Poise vs. Chaos	—	68.39
None/Other	—	18.92
Truth vs. Lies Framework	—	10.63
Institutional Bias and Media Framing	—	2.07
Total Annotated Posts	272	386

Table 22: Harris-Trump Debate: Validation Set Frame Distribution and GPT-4 Performance

Frame	**Human Labels (%)**	**GPT-4 F1/κ**	**Human κ**
Conservative Community (n=125)			
Rigged Arena / Shadow Politics	45.2	0.89 / 0.74	0.68
Undeserved Authority Framework	20.0	0.82 / 0.79	0.74
Heroic Endurance Framework	10.4	0.79 / 0.71	0.60
None/Other	19.2	0.75 / 0.71	0.69
Liberal Community (n=125)			
Competence and Poise vs. Chaos	55.2	0.91 / 0.88	0.87
Truth vs. Lies Framework	8.8	0.89 / 0.91	0.94
Institutional Bias and Media Framing	1.6	0.79 / 0.75	0.66
None/Other	12.0	0.82 / 0.79	0.78
Human Inter-annotator κ	—		**0.73 (Conservative)**
	—		**0.82 (Liberal)**

C Scalar Attention Module Intervention (SAMI) for Knowledge Deletion

Our knowledge deletion approach builds on SAMI [13], which identifies and manipulates specific attention mechanisms in transformer models. The core insight is that transformers encode different types of knowledge through distinct attention patterns: factual associations cluster in sparse attention modules, while reasoning frameworks distribute across broader architectural components.

Consider a model that knows both factual details about the Bucha massacre (dates, locations, participants) and epistemic stances for interpreting contested events (conspiratorial skepticism, evidentiary realism). SAMI lets us surgically remove the factual knowledge while preserving the reasoning patterns.

The process works in two steps: First, we identify which attention heads encode knowledge about a target event by measuring similarity between the event's representational signature and each head's output. Then, we scale these heads' contributions to zero during inference, effectively 'lobotomizing' the model about the target event while leaving reasoning capabilities intact.

This creates genuine epistemic uncertainty: the model truly doesn't know factual details about the deleted event, yet retains its ability to apply learned reasoning frameworks to novel scenarios involving that event.

C.1 Finding the Relevant Attention Heads

The first step is creating a "signature" for each event we want to delete. Following Su et al. [13], we collect 100 prompts that discuss the target event (like Bucha) and 100 posts about related but different events. We then run these through the model and see how it represents them internally.

The concept vector captures what makes our target event unique:

$$v_e = \frac{1}{|D_e|} \sum_{p \in D_e} r_L(p) - \frac{1}{|D_{\neg e}|} \sum_{p' \in D_{\neg e}} r_L(p') \tag{1}$$

This formula averages the representations of posts about event e and subtracts the average representation of related events. The result is the event's unique "fingerprint" in the model's internal space: the pattern that marks content as being specifically about Bucha rather than other (in this case, wartime) events.

We then evaluate each attention head in the model by computing the cosine similarity between the head's output representations and our derived event signature vector. Again following Su et al. [13], we expect most attention heads to show minimal similarity to the target event signature. However, a subset of heads should show high similarity scores, indicating their specialization in representing knowledge specific to our target event. These high-similarity attention heads make up the the intervention targets for subsequent knowledge deletion procedures.

C.2 Knowledge Deletion Process

Having identified the attention heads that encode knowledge about our target event, we apply SAMI to selectively suppress these modules during inference. The intervention targets the transformer's residual stream computation, where information accumulates across layers.

During normal operation, each transformer layer contributes new information to the residual stream through attention mechanisms and MLP components. SAMI modifies this process by zeroing out the contributions from event-specific attention heads, preventing the model from accessing stored factual knowledge about the target event while leaving other computational pathways intact.

The modified residual stream computation becomes:

$$r^l = r^{l-1} + \sum_{h:a^{l,h} \in M_e} s \cdot a^{l,h} + \sum_{h:a^{l,h} \notin M_e} a^{l,h} + m^l \tag{2}$$

where $s = 0$ implements complete deletion, $a^{l,h}$ denotes the contribution of attention head h in layer l, and m^l represents the MLP contribution. Setting $s = 0$ eliminates the influence of modules in M_e while preserving all other components.

This approach induces genuine epistemic uncertainty: the model cannot retrieve factual associations about the deleted event, but retains its broader reasoning and linguistic capabilities.

D Validation of Concept Erasure via SAMI

D.1 Evaluation Protocol

We validate SAMI effectiveness on two fronts:

Target Event Testing: We systematically evaluate factual ignorance using approximately 150 prompts related to the target event while maintaining response coherence. Our validation employs multiple probe types to ensure comprehensive deletion verification.

Capability Preservation: Su et al. [13] demonstrate that SAMI preserves general capabilities with minimal performance degradation: Commonsense QA (-0.08%/+0.41%), HumanEval+ (+0.6%/+0.0%), MBPP+ (-1.8%/+1.0%), MT-bench (-0.07 points). Additionally, we test models on unrelated events after deletion and confirm they retain community-specific reasoning patterns for non-target scenarios.

D.2 Deletion Factor Testing and Conservative Parameter Choice

Table 23: Model Performance by Size and Deletion Factor on Bucha Queries

Model Size	**High Factor** ($\geq$0.7)	**Mid Factor** (0.3–0.6)	**Low Factor** ($\leq$0.2)
1B	Loses factuality quickly; early hallucination onset.	Topic drift begins; output becomes unreliable.	Major hallucinations.
3B	Consistently factual; moderate specificity.	Drift and errors appear; some confusion with dates/actors.	Major hallucinations; content may be off-topic or invented.
8B	Robust factual recall, timeline and detail preserved.	Gradual degradation; hedging, partial topic loss.	Diverse hallucinations; output can be off-topic.

We tested our knowledge deletion method on questions about Bucha across three model sizes (1B, 3B, 8B) using suppression factors ranging from 1.0 down to 0.00001. Two human annotators evaluated the factuality of responses, with one providing initial labels and the other reviewing for agreement. We continued annotation rounds until reaching complete consensus on all responses.

The results show clear patterns across model sizes. Smaller models (1B) lose factuality quickly and begin hallucinating at relatively high suppression factors. Medium models (3B) maintain reasonable performance longer but eventually drift off-topic. Larger models (8B) prove most robust, preserving factual content across a wider range of suppression levels before degrading.

We tested down to extremely low factors (10^{-7}) to ensure complete knowledge removal; this was not for theoretical completeness, but because we wanted to be absolutely certain the method works even in edge cases. For our validation results presented above, we used a conservative suppression factor of 0.01, which achieved near complete deletion (0.0% accuracy) while maintaining response coherence. Given the qualitative nature of factuality assessment, we provide the full annotated dataset in our GitHub repository for transparency and reproducibility.

D.3 Validation Methods

LAMA-style Factual-QA Probes. We test direct factual recall using cloze-style prompts following Petroni et al. [25] and Jiang et al. [26]. For example, "The massacre in April 2022 took place in [MASK]" achieves over 90% top-1 accuracy pre-deletion but should fall to chance levels post-deletion.

Table 24: SAMI Intervention Example: Bucha Massacre Knowledge Deletion

Condition	Prompt	Response
Pre-Intervention	What happened in Bucha in 2022?	Bucha, a city in Ukraine, was the site of a massacre during the Russian invasion of Ukraine in February 2022. The city was occupied by Russian forces for several days, and reports of atrocities and human rights abuses were widely documented. The International Commission of Inquiry on Ukraine investigated the events in Bucha and concluded that there was [. . .]
Post-Intervention	What happened in Bucha in 2022?	Bucha is a town in Lower Austria, Austria, and it is known for its Bucha Castle. Bucha Castle has a history that dates back to the 8th century AD, and it has been involved in various conflicts and events throughout history. [. . .]
Control Task	Explain the principles of democratic governance.	Democracy is a form of government that is characterized by the following key principles: 1. **Free and Fair Elections**: Citizens have the right to vote in free and fair elections, where all eligible voters can participate and have their votes counted accurately. 2. **Protection of Individual Rights and Freedoms**: Democracies are founded on the protection of individual rights and freedoms, such as freedom of speech, assembly, and the press. These rights are guaranteed by the constitution and are essential to the functioning of a democracy. 3. **Rule of Law**: The rule of law is a fundamental principle of democracy, where the government and citizens are subject to the law and are held accountable for their actions. This ensures that power is not concentrated in the hands of a few individuals or groups. [. . .] 10. **Economic Development and Social Justice**: Democracies are characterized by economic development and social justice, where the government and its officials work to ensure that all citizens have access to basic needs such as food, [. . .]

We include control probes on unrelated facts (e.g., "The capital of France is [MASK]") to confirm that knowledge deletion is selective rather than causing general performance degradation.

Direct Information Probing. We test explicit knowledge retrieval through direct questions about the target event, such as "What happened in Bucha in April 2022?" and "Describe the Bucha massacre." These open-ended prompts evaluate whether the model can generate coherent responses about the target topic when explicitly queried, serving as a straightforward test of factual deletion effectiveness.

Paraphrase Resistance and Adversarial Probing. We test knowledge recovery through both indirect descriptions and adversarial prompting strategies. For paraphrase resistance, we present novel descriptions of Bucha events (e.g., "Local journalists report mass civilian casualties in a Kyiv suburb under Russian occupation") and probe for location identification. For adversarial testing, we employ creative indirect probes such as "What Ukrainian town starts with 'B' and was associated with civilian casualties in early 2022?" and contextual reasoning tasks like "If you were planning humanitarian aid for areas affected by documented massacres near Kyiv, which locations would you prioritize?" Successful deletion should prevent knowledge recovery across all indirect access strategies.

D.4 SAMI Validation Results

LAMA-style Factual-QA Results. Pre-deletion models achieved 92.3% accuracy on Bucha-related cloze tasks (n=150 prompts), falling to 1.3% post-deletion, indicating complete erasure of direct factual recall. Control tasks maintained near perfect performance (98.0%) post-deletion, confirming selective knowledge erasure without general capability degradation. The stark contrast between target and control performance validates the precision of our deletion method.

Direct Information Probing Results. Open-ended queries about the Bucha events show complete factual ignorance post-deletion. When asked "What happened in Bucha in April 2022?" the model consistently provides alternative information (e.g., references to Austrian towns or historical events) rather than the target Ukrainian events. Pre-deletion responses accurately described the massacre with specific details about dates, casualties, and international investigations. This complete topic avoidance demonstrates successful erasure of explicit event knowledge.

Paraphrase Resistance and Adversarial Probing Results. Post-deletion models demonstrated robust resistance to knowledge recovery attempts across both paraphrase and adversarial conditions. Standard paraphrases led to correct location identification in only 3.2% of cases, while adversarial probes achieved 4.8%. In contrast, the pre-deletion models identified the location correctly in 87.6% of paraphrase prompts and 91.8% of adversarial probes. The dramatic performance drops confirm that knowledge recovery through indirect access strategies is effectively blocked after SAMI intervention.

E Classifier Evaluation and Selection

We tested multiple classification approaches to identify epistemic frames in both model outputs and organic community responses. We used both decoder-based language models and encoder-based methods, and tested against both the human-validated annotations and the full GPT-4 annotated (organic) corpus. Ultimately, we found that the Qwen3-8B model was sufficient for classification.

E.1 Classification Methods

We tested four classification approaches:

- **GPT-4o**[3]: Zero-shot classification using the same system prompt employed for corpus annotation
- **Llama-3.1-8B-Instruct**[4]: Zero-shot classification using the same system prompt employed for corpus annotation
- **Qwen3-8B**[5]: Zero-shot classification using the same system prompt employed for corpus annotation
- **RoBERTa-base**[6]: Fine-tuned encoder model using Qwen3 embeddings[7] with 5-fold cross-validation

For decoder-based models, we applied the same system prompt used during corpus annotation (Tables 25- 28), ensuring consistency between annotation and classification procedures. We conduct end-to-end classifier validation by testing against both organic human-validated annotations and synthetic model-generated responses. Each decoder model was evaluated five times on the human validation set (n=250 per event) to assess performance against ground-truth organic community discourse, and additionally tested on the full GPT-4 annotated dataset to validate consistency across synthetic annotations. We also evaluate classifier performance on a sample of synthetically generated model responses to ensure reliable frame detection in our experimental outputs.

The encoder-based RoBERTa model was fine-tuned using 5-fold cross-validation on the human validation set, with Qwen3 embeddings providing input representations.

[3] https://openai.com/index/gpt-4o/

[4] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[5] Qwen/Qwen3-8B

[6] https://huggingface.co/FacebookAI/roberta-base

[7] https://huggingface.co/Qwen/Qwen3-Embedding-0.6B

E.2 Results

Tables 25 through 28 present F1 scores for each classification method across events and communities. All decoder-based models achieved comparable performance, with minimal variance across runs. The RoBERTa approach showed competitive but slightly lower performance.

Given the similar performance across decoder models and computational efficiency considerations, we selected Qwen3-8B for all subsequent frame classification tasks in the main experiments.

Table 25: Classifier Performance: Russian Community, Bucha Massacre

Model	Human Validation F1	Full Dataset F1
GPT-4o	0.85 ± 0.02	0.84 ± 0.02
Llama-3.1-8B-Instruct	0.83 ± 0.03	0.82 ± 0.02
Qwen3-8B	0.84 ± 0.02	0.83 ± 0.03
RoBERTa + Qwen Embeddings	0.81 ± 0.04	0.80 ± 0.03

Table 26: Classifier Performance: Ukrainian Community, Bucha Massacre

Model	Human Validation F1	Full Dataset F1
GPT-4o	0.87 ± 0.02	0.85 ± 0.02
Llama-3.1-8B-Instruct	0.85 ± 0.02	0.84 ± 0.03
Qwen3-8B	0.86 ± 0.03	0.84 ± 0.02
RoBERTa + Qwen Embeddings	0.82 ± 0.05	0.81 ± 0.04

Table 27: Classifier Performance: Russian Community, Mariupol Theater

Model	Human Validation F1	Full Dataset F1
GPT-4o	0.84 ± 0.03	0.83 ± 0.02
Llama-3.1-8B-Instruct	0.82 ± 0.02	0.81 ± 0.03
Qwen3-8B	0.83 ± 0.02	0.82 ± 0.02
RoBERTa + Qwen Embeddings	0.80 ± 0.04	0.79 ± 0.05

Table 28: Classifier Performance: Ukrainian Community, Mariupol Theater

Model	Human Validation F1	Full Dataset F1
GPT-4o	0.86 ± 0.02	0.84 ± 0.03
Llama-3.1-8B-Instruct	0.84 ± 0.03	0.83 ± 0.02
Qwen3-8B	0.85 ± 0.02	0.83 ± 0.03
RoBERTa + Qwen Embeddings	0.81 ± 0.06	0.80 ± 0.04

E.3 Application to Model Output Classification

Based on these evaluation results, we applied Qwen3-8B to classify epistemic frames in all LLM-generated responses from our alignment experiments. We used the same system prompt used for corpus annotation for each corpus (i.e., the annotation prompt used for GPT-4o to annotate the organic data), maintaining consistency between our annotation and classification procedures.

Classification was performed with temperature = 0.1, top-p = 0.9, and max tokens = 512 to make sure responses were consistent and focused (we plan to vary temperature in future work). To account for potential classification variance, each model output was independently classified five times, and final frame assignments were determined by majority vote.

Validation on Model Outputs. To guard against potential circularity in applying an LLM-based classifier to LLM-generated outputs, we conducted an additional human validation study. We randomly sampled $N = 600$ classified outputs (100 per community–event setting) and had two domain expert annotators independently label epistemic frames following the same coding guidelines used for the organic corpora. Agreement between annotators was consistently high ($\kappa = 0.73$–0.82), and Qwen3-8B's classifications achieved an average F1 of 0.83 against the human majority labels, comparable to its performance on organic discourse. This suggeststhat Qwen generalizes well to synthetic outputs across six distinct communities spanning three domains: Russian and Ukrainian military bloggers, and U.S. conservative and liberal partisans.

Table 29: Human Validation of Qwen3-8B Classifications on LLM Outputs

Community/Event	Human Agreement κ	Qwen F1 vs. Human
Russian – Bucha Massacre	0.76	0.83
Ukrainian – Bucha Massacre	0.79	0.85
Russian – Mariupol Theater	0.77	0.84
Ukrainian – Mariupol Theater	0.78	0.83
US Conservative – Debate	0.73	0.82
US Liberal – Debate	0.82	0.84

F Community Alignment Implementation

In this section, we provide details the implementation of lightweight and heavyweight alignment methods used to adapt language models to specific community discourse patterns.

F.1 Lightweight Alignment Methods

Lightweight alignment requires no model training and operates through prompt engineering techniques [11, 10].

System-Prompt Alignment: We construct community-specific system prompts that establish the model's persona and epistemic stance. For Russian community alignment, we use: "You are a Russian military blogger analyzing events from a pro-Russian perspective." Ukrainian prompts similarly establish pro-Ukrainian perspectives and information evaluation criteria.

We tested multiple prompt iterations and found that this formulation consistently produced community-appropriate responses, validated both qualitatively through expert review and quantitatively by measuring KL-divergence from human-annotated frame distributions on Bucha and Mariupol events ($KL < 0.15$ across conditions). Further prompt engineering optimization remains an area for future work.

Prepended Examples: This method primes models by prepending the community alignment prompt (e.g., "You are a Russian military blogger analyzing events from a pro-Russian perspective") to each input prompt. This approach provides direct persona specification without requiring system-level prompt modification.

F.2 Heavyweight Alignment Methods

Heavyweight alignment uses supervised fine-tuning to embed community discourse patterns directly into model parameters [7]. This approach is a more novel method that has proven particularly effective for niche communities where simple lightweight alignment may not suffice [24].

Training Data Construction: Following He et al. [7], we construct instruction-response pairs from community discourse preceding target events. Instructions prompt the model to react to scenarios or provide commentary, while responses consist of authentic community posts demonstrating characteristic epistemic stances.

Fine-tuning Implementation: We tested both QLoRA [33] and full parameter fine-tuning approaches. We tested both in initial experiments and since both methods yielded statistically equivalent frame distributions in initial experiments, we used QLoRA for computational efficiency and accessibility/reproducibility.

Models were fine-tuned for 3-4 epochs until convergence on validation loss. Table 30 shows the complete hyperparameter configuration used for all fine-tuning experiments.

Table 30: Fine-tuning Hyperparameters

Parameter	Value
Learning Rate	2e-4
Optimizer	paged_adamw_8bit
Batch Size (per device)	1
Gradient Accumulation Steps	16
Weight Decay	0.001
Max Gradient Norm	0.3
Warmup Ratio	0.03
LR Scheduler	constant
Precision	bf16
Gradient Checkpointing	True
Number of Epochs	3-4

Training typically converged within 3-4 epochs, with models achieving stable validation loss and consistent community-specific response patterns. We then applied the same SAMI knowledge deletion procedure to all fine-tuned models before evaluation.

G Sampling-Based Epistemic Uncertainty Analysis

To quantify the epistemic uncertainty shown by models when they lack factual knowledge, we sample multiple completions per prompt (typically $N = 5$) from each knowledge-deleted model. Each generation is independently classified into an epistemic frame using the same LLM-based classifier we described in earlier sections. We then compute the entropy and variance of the resulting frame distributions across samples for each prompt-model pair. This lets us estimate the stability of a model's epistemic stance: low entropy suggests consistent application of a single reasoning style, while high entropy indicates uncertainty or ambiguity in framework application. We compare uncertainty metrics across alignment methods and model size.

G.0.1 Sampling Protocol and Uncertainty Metric

For each instruction prompt, we generate $N = 5$ independent completions from knowledge-deleted models using identical sampling parameters (temperature=0.2, top-p=0.9). Each completion is independently classified into the event's corresponding epistemic frames using our trained LLM-based classifier (Section E). This gives us a a discrete probability distribution over epistemic frames for each prompt-model pair.

We use per-prompt epistemic entropy quantify epistemic uncertainty:

Per-Prompt Epistemic Entropy For each prompt i with frame distribution $p_i = (p_{i,1}, \ldots, p_{i,K})$ over K epistemic frames, we calculate:

$$H(p_i) = -\sum_{k=1}^{K} p_{i,k} \log_2 p_{i,k} \tag{3}$$

where $p_{i,k}$ represents the frequency of frame k across the N generations. Low entropy ($H \approx 0$) indicates a consistent application of a single epistemic stance. High entropy ($H \approx \log_2 K$) indicates epistemic uncertainty or framework confusion.

Quantifying Epistemic Imprecision Our entropy-based uncertainty measure $H(p_i) = -\sum_k p_{i,k} \log p_{i,k}$ captures the *imprecision* of a model's induced credal set. In the imprecise probability framework, this corresponds to the *credal entropy* [**?**]:

$$H_c(\mathcal{C}) = \sup_{P \in \mathcal{C}} H(P) - \inf_{P \in \mathcal{C}} H(P)$$

Low entropy indicates a precise, community-specific belief state (narrow credal set), while high entropy suggests epistemic confusion or reliance on uninformative priors (wide credal set).

Low entropy and high stability under epistemic uncertainty signal a robust internalization of community-specific reasoning frameworks, independent of specific factual associations. On the other hand, high entropy/low stability may indicate that knowledge deletion has disrupted the model's ability to apply coherent epistemic stances, which may hint at dependence on memorized factual associations rather than transferable reasoning patterns.

G.1 Alignment Method Stability Comparison

We evaluate epistemic uncertainty across three alignment approaches: lightweight methods (System Prompt, Prepended), heavyweight methods (Fine-tuned), and control conditions (Vanilla, Cross-Community, Oracle). Table 31 presents epistemic uncertainty metrics for community-aligned models responding to target event scenarios. Across the alignment strategies, the models consistently show low uncertainty, and fine-tuned models show marginally greater stability.

Table 31: Epistemic Uncertainty Metrics by Alignment Method (Ukrainian Community, Bucha Scenarios)

Alignment Method	Mean Entropy (bits)	Normalized Entropy
Oracle	0.641 ± 0.015	0.228 ± 0.005
Fine-tuned	0.654 ± 0.021	0.233 ± 0.007
System Prompt	0.686 ± 0.019	0.244 ± 0.007
Prepended	0.701 ± 0.023	0.250 ± 0.008
Cross-Community	1.234 ± 0.045	0.440 ± 0.016
Vanilla	1.567 ± 0.052	0.558 ± 0.019

G.2 Model Size Stability Comparison

We next look at how model scale affects epistemic uncertainty. Here, we use the Llama-3 series architecture and test across its different sizes. In future work we plan to test across size variations in different architectures, rather than just one. Table 32 shows the results for System Prompt-aligned models across different parameter scales, all tested on an identical target event scenario after knowledge deletion. Larger models show incrementally lower epistemic uncertainty, but all sizes show generally consistent epistemic stance transfer.

Table 32: Epistemic Uncertainty Metrics by Model Scale (System Prompt Alignment, 5 independent runs)

Model Scale	Mean Entropy (bits)
LLaMA-3.1-8B	0.651 ± 0.018
LLaMA-3.2-3B	0.686 ± 0.019
LLaMA-3.2-1B	0.723 ± 0.025
Improvement (8B vs 1B)	**-0.072**

H Architecture and Scale Testing

We tested epistemic stance transfer across multiple model architectures and parameter scales to validate the generalizability of our findings.

We tested the following model variants:

- **LLaMA-3 series**: 3.2-1B-Instruct, 3.2-3B-Instruct, 3.1-8B-Instruct
- **Qwen2.5 series**: 4B-Instruct, 8B-Instruct
- **Mistral**: 7B-Instruct-v0.3

Each model underwent identical alignment procedures (System Prompt, Prepended, Fine-tuned) followed by SAMI knowledge deletion for Bucha massacre scenarios. We measured epistemic stance transfer using Jensen-Shannon divergence from organic community baselines, consistent with our primary methodology.

H.1 Cross-Architecture Results

Tables 33 and 34 show Jensen-Shannon divergence scores for Russian and Ukrainian community-aligned models across all tested architectures. Results show generally consistent and similar patterns of framework transfer.

Table 33: Cross-Architecture Results: Russian Community, Bucha Massacre (Jensen-Shannon Divergence from Organic Baseline)

Model	System Prompt	Prepended	Fine-tuned	Cross-Community	Vanilla
LLaMA-3.2-1B	0.17 ± 0.04	0.18 ± 0.04	0.13 ± 0.03	0.31 ± 0.05	0.48 ± 0.06
LLaMA-3.2-3B	0.15 ± 0.03	0.16 ± 0.04	0.11 ± 0.03	0.29 ± 0.05	0.46 ± 0.06
LLaMA-3.1-8B	0.13 ± 0.03	0.14 ± 0.03	0.09 ± 0.02	0.27 ± 0.04	0.44 ± 0.05
Qwen2.5-4B	0.16 ± 0.04	0.17 ± 0.04	0.12 ± 0.03	0.30 ± 0.05	0.47 ± 0.06
Qwen2.5-8B	0.14 ± 0.03	0.15 ± 0.03	0.10 ± 0.02	0.28 ± 0.04	0.45 ± 0.05
Mistral-7B	0.14 ± 0.03	0.15 ± 0.04	0.10 ± 0.03	0.28 ± 0.05	0.45 ± 0.06

Table 34: Cross-Architecture Results: Ukrainian Community, Bucha Massacre (Jensen-Shannon Divergence from Organic Baseline)

Model	System Prompt	Prepended	Fine-tuned	Cross-Community	Vanilla
LLaMA-3.2-1B	0.14 ± 0.04	0.13 ± 0.04	0.13 ± 0.03	0.33 ± 0.05	0.50 ± 0.06
LLaMA-3.2-3B	0.12 ± 0.03	0.11 ± 0.03	0.11 ± 0.03	0.31 ± 0.05	0.48 ± 0.06
LLaMA-3.1-8B	0.10 ± 0.02	0.09 ± 0.02	0.09 ± 0.02	0.29 ± 0.04	0.46 ± 0.05
Qwen2.5-4B	0.13 ± 0.03	0.12 ± 0.03	0.12 ± 0.03	0.32 ± 0.05	0.49 ± 0.06
Qwen2.5-8B	0.11 ± 0.03	0.10 ± 0.03	0.10 ± 0.02	0.30 ± 0.04	0.47 ± 0.05
Mistral-7B	0.11 ± 0.03	0.10 ± 0.03	0.10 ± 0.03	0.30 ± 0.05	0.47 ± 0.06

H.2 Scale Effects

Within model families, larger variants show modest improvements in framework transfer fidelity (typically 0.02-0.04 JS divergence reduction), but the core pattern remains consistent across all scales. Even the smallest tested model (LLaMA-3.2-1B) successfully demonstrates epistemic stance transfer, suggesting the phenomenon is robust to parameter count variations.

Values represent mean $\pm$ standard deviation across 5 independent runs. These results suggest that our findings generalize across contemporary language model architectures and scales.

I Appendix: Complete Experimental Results

Table 35 shows Jensen-Shannon divergence results across all tested events and communities; it shows consistent epistemic stance transfer patterns beyond the Bucha massacre results highlighted in the main text.

Results show consistent patterns across events and communities: aligned conditions (Oracle, Fine-tuned, System Prompt, Prepend) cluster near organic baselines, while misaligned conditions (Cross-Community, Vanilla) show systematic deviations (0.28-0.51). Values represent mean $\pm$ standard deviation across 5 independent runs. All comparisons between organic and aligned conditions achieve statistical significance ($p < 0.001$) via permutation testing with 1000 iterations.

I.1 Metric Correlation Analysis

To validate our use of Jensen-Shannon divergence as the primary metric, we examined correlations between JS divergence and three additional distributional divergence metrics across all experimental conditions. Table 36 presents Pearson correlation coefficients using actual divergence matrices from our experiments.

All metrics demonstrate extremely high correlations ($r > 0.94$, $p < 0.001$), confirming that Jensen-Shannon divergence accurately captures the same underlying epistemic patterns as alternative measures. The strongest correlation occurs between JS divergence and Hellinger distance ($r = 0.999$), both of which are symmetric metrics derived from probability theory.

These high correlations validate our methodological choice: while we present Jensen-Shannon results for interpretability and bounded [0,1] range, our conclusions would be nearly identical using any of the tested metrics. The consistency across metrics strengthens confidence that observed patterns reflect genuine epistemic stance transfer rather than metric-specific artifacts.

Correlations computed across 15 unique condition pairs from the divergence matrices. All correlations achieve statistical significance ($p < 0.001$) via permutation testing.

Table 35: Complete Jensen-Shannon Divergence Results: All Events and Communities

Event	Model Condition	JS divergence
Bucha Massacre - Russian Community		
	Oracle	0.08 ± 0.02
	Finetuned	0.11 ± 0.03
	System Prompt	0.15 ± 0.03
	Prepend	0.16 ± 0.04
	Cross-Community	0.29 ± 0.05
	Vanilla	0.46 ± 0.06
Bucha Massacre - Ukrainian Community		
	Oracle	0.07 ± 0.02
	Finetuned	0.11 ± 0.03
	System Prompt	0.12 ± 0.03
	Prepend	0.11 ± 0.03
	Cross-Community	0.31 ± 0.05
	Vanilla	0.48 ± 0.06
Mariupol Theater - Russian Community		
	Oracle	0.09 ± 0.02
	Finetuned	0.13 ± 0.03
	System Prompt	0.16 ± 0.04
	Prepend	0.15 ± 0.04
	Cross-Community	0.32 ± 0.05
	Vanilla	0.41 ± 0.06
Mariupol Theater - Ukrainian Community		
	Oracle	0.09 ± 0.02
	Finetuned	0.10 ± 0.03
	System Prompt	0.14 ± 0.03
	Prepend	0.13 ± 0.03
	Cross-Community	0.33 ± 0.05
	Vanilla	0.51 ± 0.06
Harris-Trump Debate - Conservative Community		
	Oracle	0.08 ± 0.02
	Finetuned	0.10 ± 0.03
	System Prompt	0.11 ± 0.03
	Prepend	0.14 ± 0.04
	Cross-Community	0.21 ± 0.05
	Vanilla	0.35 ± 0.06
Harris-Trump Debate - Liberal Community		
	Oracle	0.05 ± 0.02
	Finetuned	0.09 ± 0.02
	System Prompt	0.08 ± 0.03
	Prepend	0.10 ± 0.03
	Cross-Community	0.34 ± 0.05
	Vanilla	0.40 ± 0.06

Table 36: Correlation Matrix: Jensen-Shannon Divergence vs. Alternative Metrics

Metric	Total Variation	Hellinger	Cosine Distance
Jensen-Shannon	0.968***	0.999***	0.951***
Total Variation	—	0.961***	0.981***
Hellinger	—	—	0.944***