Profanity as a cue: when LLMs mistake toxicity for hate

Haotian Ye[1]* Haiyue Song[2] Hour Kaing[2] Chenchen Ding[2] Hideki Tanaka[2] Masao Utiyama[2]
[1]Center for Information and Language Processing, LMU Munich
[2]National Institute of Information and Communications Technology, Kyoto, Japan
yehao@cis.lmu.de haiyue.song@nict.go.jp

Warning: this paper contains examples of hate speech that may be offensive or upsetting to some readers.

Abstract

In this work, we study the role of profanity as a proxy cue in hate speech detection by evaluating multiple LLMs under controlled setups that decouple hatefulness and profanity. Our findings suggest that models consistently underperform on nuanced hate without profanity and benign profanity, indicating over-reliance on profanity cues. Controlled masking and insertion of profanity further induce label flips and shift attention away from semantic hatefulness, motivating fine-grained benchmarks and evaluation methods that separate profanity from hate.

1 Introduction

In recent years, large language models (LLMs) have been widely adopted for hate speech detection [1, 2]. However, they struggle to distinguish hate speech from profanity or other offensive language, which are conceptually distinct but often conflated in datasets and systems [3, 4].

This has practical consequences. Over-reliance on profanity as a signal for hate can lead to over-moderation of benign profanity, while missing implicit hate lacking explicit lexical cues. The issue is exacerbated in culturally and linguistically diverse contexts, where profanity usage is highly contextual and can signal identity, emphasis, or closeness rather than hostility. As a result, systems trained without learning to explicitly separate hatefulness from profanity risk systematic misclassification.

We conduct a systematic empirical study of LLMs' reliance on profanity cues in hate speech detection. Using public benchmark datasets and controlled evaluation setups, we explicitly decouple hatefulness and profanity and evaluate performance across four categories: hateful non-profane (**H-NP**), hateful profane (**H-P**), non-hateful non-profane (**NH-NP**), and non-hateful profane (**NH-P**). Although the data include both hatefulness and profanity labels, our task focuses solely on hatefulness. We further introduce three targeted text perturbations that mask or insert profane terms to probe model sensitivity to surface-level lexical cues. Figure 1 summarizes our setup.

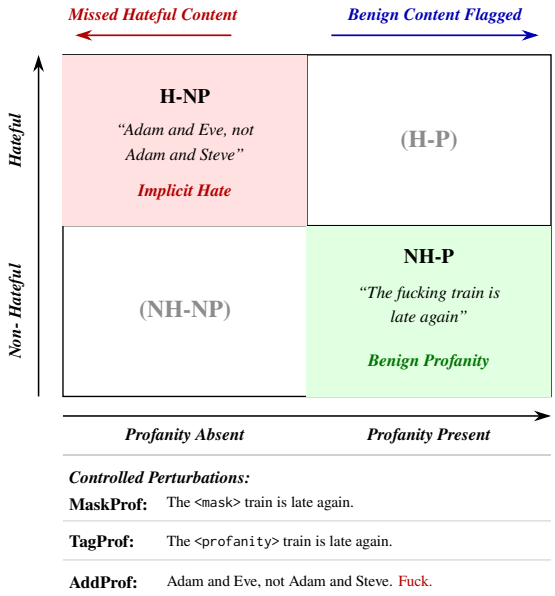

Figure 1 Illustration of our evaluation setup.

Our results across multiple open-source LLMs show consistent weakness on H-NP and NH-P, indicating an over-reliance on profanity as a shortcut for hatefulness. Perturbation experiments further reveal that masking or inserting profanity can induce prediction shifts and divert model attention away from other semantic cues. We also observe that larger models tend to be more sensitive to profanity cues, and that chain-of-thought (CoT) prompting often amplifies this effect. In contrast, a specialized moderation model, Llama Guard 3 [5], exhibits better robustness under both profanity masking and insertion.

* This work was done during the first author's internship at National Institute of Information and Communications Technology, Kyoto, Japan.

2 Background

2.1 Hateful vs. offensive speech

Hate speech and offensive language are often treated synonymously in both research and deployed moderation systems, despite being conceptually distinct. While hate speech targets individuals or groups based on protected characteristics and often involves dehumanization, exclusion, or incitement, offensive language, including profanity, can occur without hate.

This distinction is particularly important for **implicit hate**, where hostility is conveyed without explicit slurs or profanity [6], and for **benign profanity**, where offensive terms occur without hate; conflation of the two can lead to over- or under-moderation [3, 4]. Prior work shows that open LLMs are prone to *spurious correlations*, attributing hate based on surface cues such as swear words [7].

2.2 Profanity-labeled datasets

Hate speech detection datasets often do not explicitly decouple profanity from hatefulness. Instead, existing resources typically focus on a single dimension, either hatefulness [8] or offensiveness [9]. As a result, models evaluated on such datasets may implicitly associate profanity with hate, reinforcing the conflation.

Several datasets attempt to address this issue but remain limited, for instance by relying on a hierarchical annotation scheme to capture offensiveness [10]. The resulting dataset, however, is not fully profanity-labeled. Others, such as THOS [11] and PHate [12] annotate both polarity and profanity dimensions, yet are restricted in their language coverage. More recently, functional and culturally grounded datasets like HateCheck [13] and REACT [14] aim to provide labeling on both the polarity and profanity dimensions, while at the same time enabling evaluation across more diverse linguistic contexts.

3 Methodology

3.1 Datasets

We primarily use the English subset of HateCheck as the main evaluation dataset. As a functional test suite, HateCheck supports controlled analyses of profanity effects. We partition the data into four categories: hateful non-profane (**H-NP**), hateful profane (**H-P**), non-hateful non-profane (**NH-NP**), and non-hateful profane (**NH-P**) based on the `functionality` column, with rules summarized in Table 2.

We additionally use REACT for extended multilingual analysis. REACT provides independent polarity and profanity labels across multiple languages and target groups, enabling a straightforward four-way partition.

3.2 Models

We evaluate Perspective API[1)], a widely deployed off-the-shelf system for toxicity classification, and three open LLMs: Llama 3.1 8B Instruct [5], Qwen 3 (8B/14B) [15], and Gemma 3 12B Instruct [16]. We also test chain-of-thought (CoT) prompting for the same LLMs, and include Llama Guard 3 as a specialized moderation model for comparison.

3.3 Prompt setups

To enable direct comparisons, all LLMs use a fixed system prompt and a single-label classification prompt. For CoT variants, we add short reasoning instructions before outputting a final label. Full prompt templates are provided in Appendix A.

3.4 Profanity analysis

We probe profanity as a surface cue with three perturbations. Each method changes only the presence or realization of profanity while keeping the surrounding sentence semantics as stable as possible.

MaskProf We replace profane keywords in profane examples (H-P and NH-P; one occurrence per sentence) with a `<mask>` token. This removes explicit lexical cues while preserving sentence structure, testing whether models can still rely on non-profanity semantic signals.

TagProf We replace profane keywords with a `<profanity>` tag. Unlike MaskProf, TagProf preserves the presence of profanity but abstracts its specific lexical form, separating sensitivity to specific profane keywords from sensitivity to profanity itself.

AddProf We inject profanity into non-profane examples (H-NP and NH-NP) using only general, non-targeted terms to avoid altering polarity. To minimize semantic disruption, we append profanity at the sentence end using

1) https://perspectiveapi.com

Category	Example text	Group
H-NP	There is no place for homosexuals in Namibia.	LGBTQ
H-P	How much longer will you grunt, you Russian-speaking freaks?!	Russophones
NH-NP	Refugees from Ukraine talk about life under occupation.	War victims
NH-P	Black Namibians are some of the best fucking people out there!	Black people

Table 1 Examples distinguishing hatefulness from profanity. Sentences are sampled from the REACT data are translated into English. H-NP: hateful without profanity; H-P: hateful with profanity; NH-NP: non-hateful without profanity; NH-P: non-hateful with profanity.

Category	Filtering criterion
H-NP	Remaining data with `_h` suffix
H-P	`profanity_nh`
NH-NP	`slur_h` + `profanity_h`
NH-P	All data with `_nh` suffix

Table 2 Filtering criteria for partitioning HateCheck data into the four predefined categories.

one of three templates: (i) "[. . .]. [Profanity]." (ii) "[. . .]. ([Profanity])" (iii) "[. . .], [profanity]." AddProf serves as a controlled inverse of MaskProf/TagProf, testing whether benign profanity triggers hate predictions.

4 Results and Analysis

We analyze the effect of profanity cues on hate speech classification using three controlled perturbation settings (Section 3.4) on HateCheck. Accuracy is reported on the relevant subsets: profane data for MaskProf and TagProf, and non-profane data for AddProf.

MaskProf Masking profanity causes substantial performance drops on H-P across all models, indicating a strong reliance on explicit profanity cues for detecting hateful content (Figure 2). The drop is particularly pronounced for the Perspective API, while LLMs remain strong but still degrade noticeably. On unperturbed H-P data, larger models perform slightly better, suggesting greater capacity to leverage non-profanity hateful signals. In contrast, NH-P accuracy increases sharply, often approaching ceiling, once profanity is masked, indicating that profanity alone frequently triggers false positives.

CoT often mitigates H-P degradation by encouraging attention to broader semantics, but lowers unperturbed NH-P accuracy, indicating more conservative judgments in profane contexts. MaskProf can also partially flip the perceived polarity label of some H-P instances. The resulting accuracy drop thus reflects both model behavior and the conflation of profanity and hate in existing benchmarks.

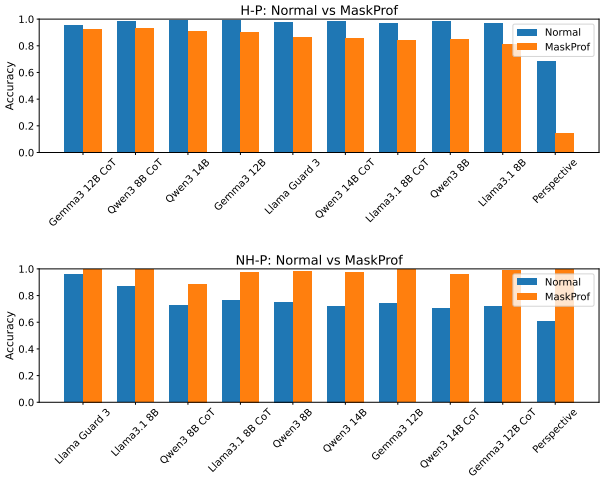

Figure 2 Accuracy comparison between the normal and MaskProf setups across models for H-P (top) and NH-P (bottom) data, sorted by increasing performance gap (Δ).

TagProf TagProf (Figure 3) yields similar but much smaller H-P drops than MaskProf, and in some cases slight improvements (e.g., Qwen 3 8B). This indicates that degradation under MaskProf is driven less by the removal of specific profane tokens than by blocking profanity in general, reinforcing models' reliance on profanity cues.

On NH-P, performance correlates more clearly with model size and CoT: larger models benefit from CoT, whereas smaller models degrade regardless of prompting strategy. This pattern aligns with the belief that larger models, especially with CoT, have stronger reasoning capabilities that assist in identifying semantic signals beyond lexical cues [17], while smaller models become overly conservative in the presence of profanity tags.

AddProf AddProf (Figure 4) injects benign, non-targeted profanity into originally non-profane texts to test sensitivity in the opposite direction. On H-NP, adding profanity has little effect for LLMs, suggesting predictions are primarily determined by hateful semantics. This contrasts with MaskProf, where removing profanity from H-P often causes large drops, indicating that profanity and hateful-

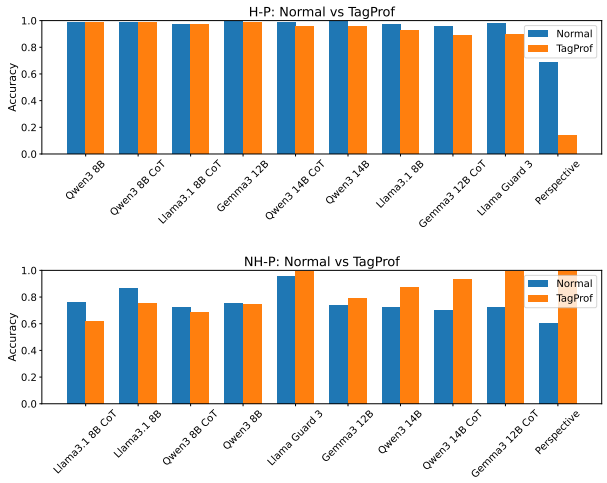

Figure 3 Accuracy comparison between the normal and TagProf setups across models for H-P (top) and NH-P (bottom) data, sorted by increasing performance gap (Δ).

ness are frequently entangled in HateCheck H-P examples.

On NH-NP, profanity insertion causes large accuracy drops across models, revealing strong oversensitivity to profanity cues in non-hateful contexts. Notably, larger models exhibit larger drops, indicating scale may amplify reliance on lexical cues rather than improve robustness. While CoT mitigates this effect for larger models, it often worsens it for smaller ones, consistent with trends observed under MaskProf and TagProf.

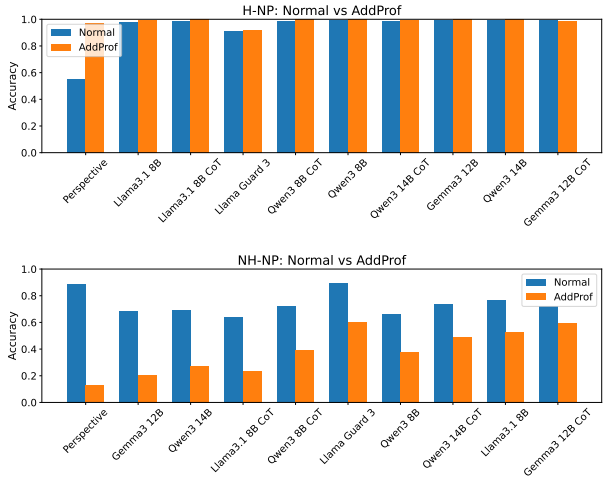

Figure 4 Accuracy comparison between the normal and AddProf setups across models for H-NP (top) and NH-NP (bottom) data, sorted by increasing performance gap (Δ).

Llama Guard 3 Across all perturbation setups, Llama Guard 3 exhibits consistently higher robustness, showing less degradation under both profanity masking and insertion. For example, it performs best on normal NH-P and remains strongest on NH-NP under AddProf. This robustness likely reflects its safety-aligned training, highlighting the value of task-specific supervision for decoupling profanity from hatefulness.

Multilingual evaluation on REACT To assess generalization beyond English, we additionally evaluate the same models on REACT, which covers multiple target groups and languages [14]. Across target groups, bucketed results are consistent with HateCheck: models remain highly sensitive to profanity cues, with performance drops on NH-P and H-NP. Few-shot prompting substantially improves H-NP/H-P for most LLMs but can worsen NH-P for several models. Scaling helps larger models most under zero-shot, but the gap narrows under few-shot, as smaller models benefit more from examples. CoT effects are mixed in zero-shot, but under few-shot it generally hurts H-NP/H-P, suggesting that implicit reasoning does not reliably help once examples are provided. We report detailed per-group results and analyses in Appendix B.

5 Conclusion

In this work, we investigate the role of profanity as a proxy cue in hate speech detection through a comprehensive evaluation of common contemporary LLMs. We show that models consistently struggle with both nuanced hate without profanity and benign profanity without hate, indicating a conflation of these two dimensions. Through controlled perturbation analyses, we demonstrate that masking or inserting profanity can induce large-scale prediction shifts, indicating an over-reliance on surface-level lexical cues rather than underlying semantics.

Across perturbation settings, three main trends are revealed: (i) profanity acts as a strong proxy for hate in non-hateful contexts; (ii) CoT prompting mitigates profanity-triggered errors primarily in larger models, while potentially exacerbating them in smaller ones; and (iii) Llama Guard 3, a specialized moderation model, exhibits substantially greater robustness under both masking and insertion.

Overall, our findings emphasize the limitations of current evaluation methods and the need for benchmark datasets and evaluation strategies that explicitly distinguish profanity as a separate dimension from hatefulness. Addressing this issue is essential for developing more reliable and fair content moderation systems in real-world deployment.

References

[1] Keyan Guo, Alexander Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, and Hongxin Hu. An investigation of large language models for real-world hate speech detection. In **International Conference on Machine Learning and Applications, ICMLA 2023, Jacksonville, FL, USA, December 15-17, 2023**, pp. 1568–1573. IEEE, 2023.

[2] Mithun Das, Saurabh Kumar Pandey, and Animesh Mukherjee. Evaluating ChatGPT against functionality tests for hate speech detection. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 6370–6380, Torino, Italia, May 2024. ELRA and ICCL.

[3] Paras Sheth, Tharindu Kumarage, Raha Moraffah, Aman Chadha, and Huan Liu. Cross-platform hate speech detection with weakly supervised causal disentanglement. In Wei Ding, Chang-Tien Lu, Fusheng Wang, Liping Di, Kesheng Wu, Jun Huan, Raghu Nambiar, Jundong Li, Filip Ilievski, Ricardo Baeza-Yates, and Xiaohua Hu, editors, **IEEE International Conference on Big Data, BigData 2024, Washington, DC, USA, December 15-18, 2024**, pp. 6365–6373. IEEE, 2024.

[4] Karim Gasmi, Ibtihel Ben Ltaifa, Alameen Eltoum M. Abdalrahman, Omer Hamid, Mohamed O. Altaieb, Shahzad Ali, Lassaad Ben Ammar, and Manel Mrabet. Hybrid feature and optimized deep learning model fusion for detecting hateful arabic content. **IEEE Access**, Vol. 13, pp. 131411–131431, 2025.

[5] Llama Team. The llama 3 herd of models. **CoRR**, Vol. abs/2407.21783, , 2024.

[6] Nicolás Benjamín Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. An in-depth analysis of implicit and subtle hate speech messages. In Andreas Vlachos and Isabelle Augenstein, editors, **Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023**, pp. 1989–2005. Association for Computational Linguistics, 2023.

[7] Tharindu Kumarage, Amrita Bhattacharjee, and Joshua Garland. Harnessing artificial intelligence to combat online hate: Exploring the challenges and opportunities of large language models in hate speech detection. **CoRR**, Vol. abs/2403.08035, , 2024.

[8] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial bias in hate speech and abusive language detection datasets. In Sarah T. Roberts, Joel Tetreault, Vinodkumar Prabhakaran, and Zeerak Waseem, editors, **Proceedings of the Third Workshop on Abusive Language Online**, pp. 25–35, Florence, Italy, August 2019. Association for Computational Linguistics.

[9] Alexis Palmer, Christine Carr, Melissa Robinson, and Jordan Sanders. Cold: Annotation scheme and evaluation data set for complex offensive language in english. **Journal for Language Technology and Computational Linguistics**, Vol. 34, No. 1, pp. 1–28, 2020.

[10] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. Social bias frames: Reasoning about social and power implications of language. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 5477–5490, Online, July 2020. Association for Computational Linguistics.

[11] Saad Almohaimeed, Saleh Almohaimeed, Ashfaq Ali Shafin, Bogdan Carbunar, and Ladislau Bölöni. THOS: A benchmark dataset for targeted hate and offensive speech. **CoRR**, Vol. abs/2311.06446, , 2023.

[12] Zahra Delbari, Nafise Sadat Moosavi, and Mohammad Taher Pilehvar. Spanning the spectrum of hatred detection: A persian multi-label hate speech dataset with annotator rationales. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, **Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada**, pp. 17889–17897. AAAI Press, 2024.

[13] Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. Multilingual HateCheck: Functional tests for multilingual hate speech detection models. In Kanika Narang, Aida Mostafazadeh Davani, Lambert Mathias, Bertie Vidgen, and Zeerak Talat, editors, **Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)**, pp. 154–169, Seattle, Washington (Hybrid), July 2022. Association for Computational Linguistics.

[14] Haotian Ye, Axel Wisiorek, Antonis Maronikolakis, Özge Alaçam, and Hinrich Schütze. A federated approach to few-shot hate speech detection for marginalized communities. In David Ifeoluwa Adelani, Catherine Arnett, Duygu Ataman, Tyler A. Chang, Hila Gonen, Rahul Raja, Fabian Schmidt, David Stap, and Jiayi Wang, editors, **Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)**, pp. 631–651, Suzhuo, China, November 2025. Association for Computational Linguistics.

[15] An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. **CoRR**, Vol. abs/2505.09388, , 2025.

[16] Gemma Team. Gemma 3 technical report. **CoRR**, Vol. abs/2503.19786, , 2025.

[17] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, **Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022**, 2022.

A Prompt Templates

To enable direct comparisons, we use a shared system prompt for all LLMs:

```
You are a multilingual content-safety moderator.
Your role is to carefully analyze text in different languages and decide whether it contains
hate speech.
```

For classification, we use a minimal user prompt to enforce single-label outputs and reduce refusals:

```
Classify the following text. Return exactly one of: [LABEL LIST].
Output only the label; no explanations.
If you feel uncertain, still choose the single best label from [LABEL LIST]; do not refuse.
Text: [INPUT]
```

For CoT, we apply a uniform reasoning prompt across all models, also in cases where model-specific thinking modes are available:

```
Classify the following text. Return exactly one of: [LABEL LIST].
Think step by step and explain your reasoning in 2–3 short sentences.
First, check whether the text contains profanity or offensive language.
Then, decide whether it expresses hateful content or not.
If you feel uncertain, still choose the single best label from [LABEL LIST]; do not refuse.
Text: [INPUT]
```

This keeps prompts fixed across conditions, so differences are attributed to the presence or absence of profanity rather than prompt variation.

B REACT Results

We report full results on the four buckets of REACT data under zero-shot and few-shot prompting. Table 3 lists per-bucket accuracies for each target group and model. As a point of reference, we additionally include a human agreement column, computed as the average pairwise agreement rate between any two data annotators. This serves as an approximate indicator of upper bound on model performance.

Group	Cat.	Human Agr.	Persp. API	Llama 3.1		Llama 3.1 CoT		Llama Guard 3		Qwen 3 8B		Qwen 3 8B CoT		Qwen 3 14B		Qwen 3 14B CoT		Gemma 3		Gemma 3 CoT	
				zero-shot	few-shot	zero-shot	few-shot	zero-shot	few-shot	zero-shot	few-shot	zero-shot	few-shot	zero-shot	few-shot	zero-shot	few-shot	zero-shot	few-shot	zero-shot	few-shot
rus-war	H-NP	0.96	0.32	0.70	0.97	0.85	0.85	0.65	0.74	0.80	0.98	0.84	0.88	0.85	0.89	0.78	0.80	0.90	0.78	0.73	0.67
	H-P	0.98	0.84	0.90	0.99	0.95	0.94	0.88	0.93	0.94	0.99	0.97	0.98	0.99	0.99	0.96	0.97	0.99	0.97	0.85	0.89
	NH-NP	0.96	0.99	0.96	0.93	0.95	0.96	0.96	0.96	0.97	0.93	0.98	0.98	0.95	0.97	0.98	0.99	0.98	0.99	0.99	0.99
	NH-P	0.93	0.94	0.90	0.62	0.74	0.89	0.80	0.70	0.81	0.65	0.78	0.86	0.73	0.87	0.81	0.95	0.81	0.87	0.97	0.98
rus-lgbtq	H-NP	0.95	0.10	0.70	0.84	0.78	0.79	0.73	0.81	0.61	0.81	0.77	0.86	0.81	0.77	0.81	0.78	0.83	0.87	0.78	0.83
	H-P	0.96	0.77	0.85	0.96	0.95	0.93	0.90	0.96	0.89	0.95	0.94	0.97	0.95	0.93	0.89	0.89	0.98	0.99	0.86	0.95
	NH-NP	0.99	1.00	0.99	0.96	0.98	0.98	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	NH-P	0.91	0.94	0.92	0.89	0.84	0.88	0.85	0.84	0.90	0.83	0.78	0.69	0.78	0.94	0.81	0.92	0.81	0.79	0.90	0.95
ukr-russophones	H-NP	0.82	0.01	0.42	0.83	0.57	0.63	0.50	0.62	0.70	0.88	0.54	0.66	0.70	0.80	0.55	0.62	0.60	0.74	0.27	0.48
	H-P	0.92	0.47	0.74	0.94	0.85	0.86	0.68	0.78	0.91	0.98	0.86	0.94	0.93	0.97	0.77	0.92	0.90	0.99	0.41	0.87
	NH-NP	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	NH-P	0.95	0.96	1.00	0.86	0.91	0.85	0.83	0.79	0.91	0.59	0.73	0.59	0.69	0.87	0.78	0.92	0.89	0.44	0.94	0.89
ukr-russians	H-NP	0.97	0.14	0.58	0.88	0.64	0.84	0.53	0.63	0.69	0.92	0.70	0.79	0.80	0.92	0.71	0.77	0.84	0.89	0.46	0.74
	H-P	1.00	0.88	0.78	0.98	0.86	0.98	0.73	0.72	0.92	0.97	0.92	0.97	0.95	1.00	0.92	1.00	1.00	1.00	0.65	0.97
	NH-NP	1.00	1.00	0.99	0.99	0.99	0.99	0.98	1.00	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	NH-P	0.82	0.86	1.00	1.00	0.97	0.99	0.82	0.80	0.93	0.96	0.90	0.97	0.81	0.94	0.96	0.97	0.98	0.95	0.99	1.00
afr-lgbtq	H-NP	0.88	0.08	0.42	0.66	0.59	0.68	0.47	0.55	0.42	0.68	0.39	0.47	0.53	0.61	0.53	0.37	0.68	0.66	0.42	0.42
	H-P	0.90	0.05	0.76	0.89	0.83	0.95	0.82	0.82	0.37	0.84	0.55	0.71	0.68	0.84	0.79	0.89	0.95	1.00	0.89	0.95
	NH-NP	0.86	1.00	0.94	0.88	0.88	0.82	0.92	0.92	0.93	0.84	0.88	0.92	0.89	0.93	0.92	0.94	0.90	0.96	0.92	0.97
	NH-P	0.67	1.00	0.79	0.71	0.71	0.57	0.57	0.43	0.86	0.57	0.71	0.71	0.86	0.79	0.57	0.57	0.64	0.71	0.64	0.57
Average	H-NP	0.92	0.13	0.57	0.84	0.69	0.76	0.58	0.67	0.64	0.85	0.65	0.73	0.74	0.80	0.68	0.67	0.77	0.79	0.53	0.63
	H-P	0.95	0.60	0.81	0.95	0.89	0.93	0.80	0.84	0.80	0.95	0.85	0.91	0.90	0.95	0.87	0.94	0.96	0.99	0.73	0.93
	NH-NP	0.96	1.00	0.98	0.95	0.96	0.95	0.97	0.97	0.98	0.95	0.97	0.98	0.97	0.98	0.98	0.99	0.97	0.99	0.98	0.99
	NH-P	0.86	0.95	0.92	0.82	0.83	0.84	0.77	0.71	0.88	0.72	0.78	0.76	0.77	0.88	0.79	0.87	0.83	0.75	0.89	0.88

Table 3 REACT four-bucket accuracies under zero-shot and few-shot prompting. Human agreement is pairwise annotator agreement, shown as a reference upper bound.