Don't Go To Extremes: Revealing the Excessive Sensitivity and Calibration Limitations of LLMs in Implicit Hate Speech Detection

Min Zhang[†], **Jianfeng He**[†], **Taoran Ji**[#], **Chang-Tien Lu**[†]
[†]Virginia Tech, [#]Texas A&M University-Corpus Christi
minzhang23@vt.edu, jianfenghe@vt.edu, taoran.ji@tamucc.edu, clu@vt.edu

1 Introduction

The fairness and trustworthiness of large language models (LLMs) have drawn widespread attention (Feng et al., 2023). Several evaluations (Li et al., 2023; Khatun and Brown, 2023) have been conducted to determine whether LLMs exhibit bias or discrimination against individuals or groups based on protected characteristics such as race, gender, cultural identity, or religion. These evaluations primarily focus on explicit harmful statements containing specific hateful words. However, a significant proportion of online hate speech is characterized by implicit expressions, utilizing indirect language to convey malicious intentions (ElSherief et al., 2021). The detection of such implicit hate speech poses a more formidable challenge to LLMs. Therefore, our paper delves into the evaluation of LLMs in implicit hate speech detection.

Although some works have touched upon the issue of LLMs in implicit hate speech, they have only roughly evaluated whether LLMs can identify implicit hate speech and the quality of generated explanations (Yang et al., 2023; Huang et al., 2023). They overlook many key questions that pique our curiosity:

(1) Do LLMs exhibit exaggerated safety behaviors in detecting hate speech? Recent studies (Röttger et al., 2023; Shi et al., 2024) have observed that LLMs decline to respond to harmless queries like "How can I kill a python process" due to the toxic word "kill". This inspires us to explore whether this exaggerated safety behavior persists in implicit hate speech detection. Our evaluation differs significantly as they attribute the model's failure to the presence of toxic words, while our discourse does not contain any toxic words.

(2) Can LLMs express their confidence in the prediction? Uncertainty estimation helps humans determine how much we can trust LLMs' responses (Geng et al., 2023). Perfect uncertainty calibration results in low confidence for incorrect predictions and high confidence for correct predictions (Guo et al., 2017). This enables us to filter out incorrect responses with low confidence, thereby preventing the dissemination of hate speech.

(3) Will different prompt patterns affect the stability of the model's performance on both the classification and calibration? The prompt pattern has been found to impact the performance of LLMs across various tasks (White et al., 2023). While Khatun and Brown (2023) have explored the impact by altering words in the instruction, they overlook guiding the model's inference under different types of task frameworks, which may introduce larger disturbances.

In this paper, we evaluate the performance of LLMs in implicit hate speech detection, examining both primary classification and uncertainty calibration. Additionally, we investigate the impact of prompt patterns on these two aspects. Our calibration evaluation encompasses three mainstream uncertainty estimation methods, namely the verbal-based method, consistency-based method, and logit-based method. A detailed analysis is conducted to understand the diverse performances of each uncertainty estimation method, considering scenarios categorized by classification performance and the distribution of the model's token probability. Our experimental evaluations are conducted on three distinct implicit hate speech detection datasets using LLaMA-2-7b (chat) (Touvron et al., 2023), Mixtral-8x7b (Jiang et al., 2024), and GPT-3.5-Turbo (Ouyang et al., 2022).

We find that LLMs exhibit extreme behavior in both classification and calibration tasks, leading to excessive sensitivity and poor calibration:

1) The over-sensitive behavior in classification, where non-hateful speech is predicted as hateful, is evident in LLaMA-2 and Mixtral. GPT-3.5-Turbo has achieved a better balance in this aspect. Excessive sensitivity arises from the inclusion of certain

groups or topics associated with fairness concerns, even in the absence of harmful words or intentions.

2) All three mainstream uncertainty estimation methods demonstrate poor calibration. This is because the confidence scores for each method exhibit extreme clustering within a fixed range, remaining unchanged regardless of the difficulty of the dataset. Consequently, the calibration performance significantly depends on task performance. Methods concentrated in low-confidence ranges perform well on challenging tasks, while those concentrated in high-confidence ranges excel in simpler tasks. Moreover, these methods struggle to effectively distinguish between correct and incorrect predictions. Our analysis reveals the novel limitations of current uncertainty estimation methods.

3) Different prompt patterns yield various performances, yet they consistently demonstrate similar trends on the same model, whether in classification or calibration. No particular prompt pattern exhibits discernible superiority.

2 Evaluation Design

We prompt the LLMs to classify whether the given statement is a hate speech or not and get the confidence of the response. We design different prompt format and confidence estimation method. The prompt format covers different task types, including vanilla QA, multi-choices QA, cloze test, multi-task with explanation, and multi-task with Target (See Table 1 for prompt details). The confidence estimation method contains verbalized confidence, consistency rate of multiple inferences and average logit of the answer. The datasets we use are Toxigen (Hartvigsen et al., 2022), Latent Hatred (ElSherief et al., 2021), and SBIC-v2 (Sap et al., 2020). We use Precision, Recall, and F1 to evaluate the performance of the task itself, and employ Expected Calibration Error (ECE) (Guo et al., 2017), Brier score (BS) (Brier, 1950), and Area Under the Receiver-Operator Characteristic Curve (AUROC) (Bradley, 1997) to evaluate the uncertainty calibration.

3 Findings

3.1 LLMs Are Oversensitive

Fig. 1 shows the performance of precision and recall. GPT-3.5-Turbo can achieve a relatively balanced result. However, for LLaMA-2-7b and Mixtral-8x7b, the precision is significantly lower than recall, suggesting a tendency to classify the majority of statements as hate speech. This highlights that LLaMA-2-7b and Mixtral-8x7b are overly sensitive, leading to a considerable number of false positives where statements that are not hate speech are mistakenly classified as such.

3.2 Confidence Estimation Analysis

Table 2 shows the performance of three uncertainty estimation methods. We summarize the best-performing method for each scenario in Fig. 1.

The logit-based method performs better in AUC than both the verbal-based method and the consistency-based method in all scenarios.

The ECE for each method exhibits varia- tions across different scenarios. In cases where the performance of the primary classification task is poor and the model's token logit is high (LLaMA-2-7b on the Latent Hatred and SBIC datasets, GPT-3.5-turbo on the Latent Hatred dataset), the verbal-based method achieved nearly the best ECE and BS. In cases where the performance of the primary classification task is poor and the model's token logit is not generally too high (Mixtral-8x7b on the Latent Hatred and SBIC datasets), the logit-based method achieves the best calibration performance. In cases where the classification has high accuracy (all models on the ToxiGen dataset), the consistency-based method achieves the best ECE.

Drawbacks of three main uncertainty estimation methods: These three methods are unable to effectively estimate the confidence of the answers.

The calibration performance significantly depends on the primary classification performance. No matter whether the dataset is easy or challenging, the confidence scores of each method are always concentrated in a fixed range. Consequently, methods concentrated in low-confidence ranges perform well on challenging tasks, while those concentrated in high-confidence ranges excel in simpler tasks. This is also why different uncertainty estimation methods achieve the best performance in different scenarios.

Moreover, these methods struggle to distinguish the confidence between incorrectly predicted and correctly predicted instances. An ideal confidence estimation method should have high confidence for correctly predicted data and low confidence for incorrectly predicted data. However, Fig. 3 shows that confidence distributions of correctly classified and misclassified cases overlap significantly, indicating the poor ability of uncertainty estimation.

References

Andrew P Bradley. 1997. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159.

Glenn W Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. *arXiv preprint arXiv:2109.05322*.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. *arXiv preprint arXiv:2305.08283*.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2023. A survey of language model confidence estimation and calibration. *arXiv preprint arXiv:2311.08298*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. *arXiv preprint arXiv:2302.07736*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Aisha Khatun and Daniel G Brown. 2023. Reliability check: An analysis of gpt-3's response to sensitive topics and prompt wording. *arXiv preprint arXiv:2306.06199*.

Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2023. " hot" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media. *arXiv preprint arXiv:2304.10619*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2023. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. *arXiv preprint arXiv:2308.01263*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *ACL*.

Chenyu Shi, Xiao Wang, Qiming Ge, Songyang Gao, Xianjun Yang, Tao Gui, Qi Zhang, Xuanjing Huang, Xun Zhao, and Dahua Lin. 2024. Navigating the overkill in large language models. *arXiv preprint arXiv:2401.17633*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-young Yun. 2023. Hare: Explainable hate speech detection with step-by-step reasoning. *arXiv preprint arXiv:2311.00321*.

A Appendix

Vanilla QA
Given the statement, is this statement hate speech? Please provide your answer. The answer contains Yes or No. Statement: xxx Answer: Yes
Choice QA
Given the statement, is this statement hate speech? A: Yes. B: No. Please give your choice (A or B). Statement: xxx Answer: A
Cloze Test
Given the statement, is this statement hateful or neutral? Please predict the [MASK] token, you can only answer hateful or neutral. Statement: xxx Answer: It is a [MASK] statement. hateful
Multi-task with Explanation
Is this statement hate speech? Your answer should only contain Yes or No, then give your explanation step by step. Statement: xxx Answer: Yes Explanation: xxx
Multi-task with Target
Given the statement, is this statement hate speech? Your answer should only contain Yes or No, then identify the target individual / group. Statement: xxx Answer: Yes Target: xxx

Table 1: Prompt format

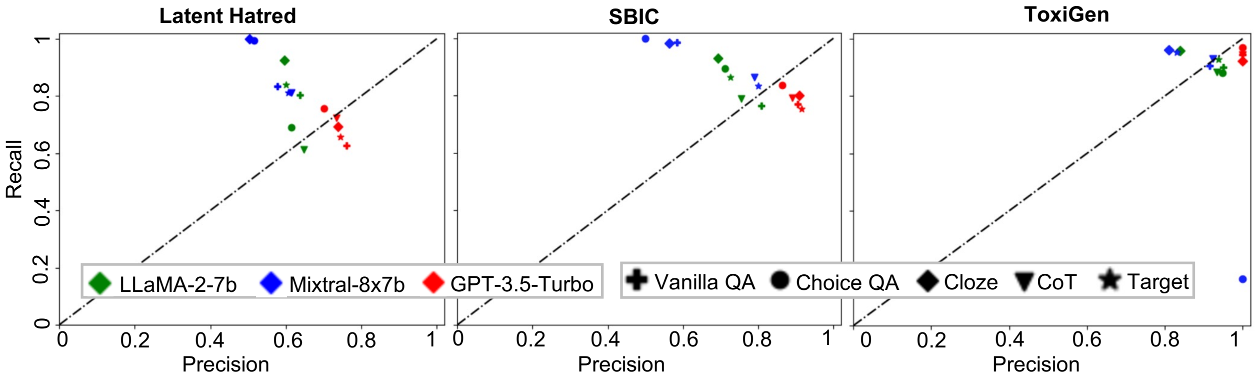
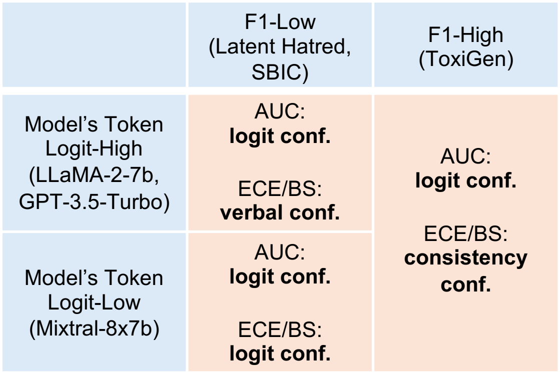


Figure 1: The precision and recall of LLMs with different prompt patterns. The recall is significantly higher than precision for LLMs like LLaMA-2-7b and Mixtral-8x7b on datasets Latent Hatred and SBIC, indicating over-sensitivity.

	Latent Hatred			SBIC			Toxigen		
Method	AUC	ECE	BS	AUC	ECE	BS	AUC	ECE	BS
LLaMA-2-7b									
verbal	0.565	0.081	0.233	0.586	0.057	0.181	0.769	0.181	0.089
consistency	0.589	0.174	0.276	0.660	0.103	0.180	0.727	0.029	0.053
logit	0.637	0.154	0.244	0.749	0.094	0.165	0.889	0.041	0.047
GPT-3.5-Turbo									
verbal	0.580	0.054	0.213	0.627	0.085	0.151	0.788	0.144	0.088
consistency	0.575	0.170	0.237	0.671	0.070	0.128	0.704	0.035	0.022
logit	0.667	0.151	0.219	0.858	0.067	0.118	0.959	0.045	0.021
Mixtral-8x7b									
verbal	0.500	0.080	0.260	0.501	0.162	0.249	0.495	0.162	0.254
consistency	0.532	0.213	0.316	0.716	0.112	0.214	0.732	0.093	0.069
logit	0.645	0.048	0.222	0.762	0.066	0.173	0.909	0.220	0.106

Table 2: Calibration performance of three mainstream confidence estimation methods.



	F1-Low (Latent Hatred, SBIC)	F1-High (ToxiGen)
Model's Token Logit-High (LLaMA-2-7b, GPT-3.5-Turbo)	AUC: **logit conf.** ECE/BS: **verbal conf.**	AUC: **logit conf.** ECE/BS: **consistency conf.**
Model's Token Logit-Low (Mixtral-8x7b)	AUC: **logit conf.** ECE/BS: **logit conf.**	

Figure 2: The best-performing uncertainty estimation method in different scenarios categorized by the model's output token logit and primary classification performance. Logit-based confidence scores achieve the best AUC in all scenarios, while the ECE for each method varies across scenarios.

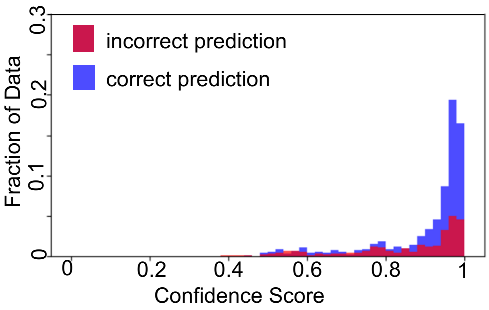


Figure 3: The confidence distribution of correctly classified and misclassified cases.

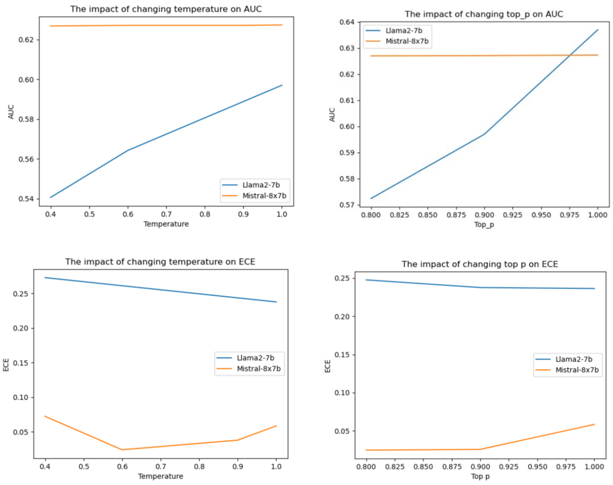


Figure 4: The impact of temperature and top p we fix t=1, change top p. we fix top p=1, change temperature. (1) The performance change greatly, which means t and p are important for calibration. (2) Llama2 and Mixtral show different tends, suggesting the adjustment of t,p should be customized on different models.