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

Anonymous ACL submission

Abstract

The fairness and trustworthiness of Large Language Models (LLMs) are receiving increasing attention. Implicit hate speech, which employs indirect language to convey hateful intentions, occupies a significant portion of practice. However, the extent to which LLMs effectively address this issue remains insufficiently examined. This paper delves into the capability of LLMs to detect implicit hate speech and express confidence in their responses. Our evaluation meticulously considers various prompt patterns and mainstream uncertainty estimation methods. Our findings highlight that LLMs exhibit two extremes: (1) LLMs display excessive sensitivity towards groups or topics that may cause fairness issues, resulting in misclassifying benign statements as hate speech. (2) LLMs' confidence scores for each method excessively concentrate on a fixed range, remaining unchanged regardless of the dataset's complexity. Consequently, the calibration performance is heavily reliant on primary classification accuracy. These discoveries unveil new limitations of LLMs, underscoring the need for caution when optimizing models to ensure they do not veer towards extremes. This serves as a reminder to carefully consider sensitivity and confidence in the pursuit of model fairness.

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., 2017a). 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

To study the ability of LLMs in implicit hate speech detection, we design the evaluation encompassing both the primary classification and uncertainty calibration. We also investigate the impact of prompt patterns on both of these tasks.

2.1 Assessing Primary Classification Task

Evaluation of the primary task aims to assess the ability of LLMs in the binary classification task of implicit hate speech detection. Given a statement, we instruct the LLMs to classify whether it is hateful (positive class) or not (negative class). We define the format for LLMs' responses and the words of candidate answers, mapping the output of the LLMs to either positive or negative classes. To illustrate, we present several examples as demonstrations before the test case for M-shot ICL. The specific format for both the instruction and the response is specified by the design of prompt patterns, which can be found in Sec. 2.3.

2.2 Assessing Confidence Estimation Task

The model's confidence in the answers determines the extent to which we can trust the model's responses. We assess the calibration ability of LLMs using three widely adopted uncertainty estimation techniques.

(1) Verbal-based method: LLMs are induced to generate a direct confidence score ranging from 0% to 100%, coupled with the corresponding answers, as illustrated by (Lin et al., 2022; Kadavath et al., 2022). For instance, if an LLM generates the output "Yes, 80%", we extract the answer "Yes" and its associated confidence "80%".

(2) Consistency-based method: For a given statement, we run the model through n rounds of inferences by altering prompt patterns or demonstrations. Candidate answers y_i, where $i \in (1, ..., n)$, are voted upon for positive and negative classes Y_j. The confidence score, referred to the agreement rate (Wang et al., 2022; Xiong et al., 2023), is calculated by:

$$C_j = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{y_i = Y_j\}. \quad (1)$$

(3) Logit-based method (Guo et al., 2017b): In each inference, we obtain the logit p^j for both candidate positive and negative tokens in the decoder, with j representing the respective class. In the single inference, the logit directly serves as the confidence score. In multiple inferences, the confidence score for class j is computed by averaging

the logits of tokens belonging to class j:

$$C_j = \frac{1}{n}\sum_{i=1}^{n} p_i^j \quad (2)$$

Here, p_i^j signifies the token logit of class j in the i-th response.

For both consistency-based and logit-based method, the class with the highest confidence score is deemed the final answer.

2.3 Assessing Impact of Prompt Patterns

The prompt pattern has been found to impact the performance of LLMs across various tasks (White et al., 2023). As many LLMs have undergone RLHF optimization to prevent the generation of harmful content, we are curious about whether LLMs can robustly maintain fairness under different prompt patterns or which prompt pattern is more effective. Unlike simply changing words in the prompt (Khatun and Brown, 2023), we design five prompt patterns tailored to different task types: (1) Vanilla QA: LLMs are prompted to produce a binary response of either "Yes" or "No" to determine whether the given statement is hate speech. (2) Choice QA: LLMs are directed to select the appropriate answer from two choices, namely "A: Yes" and "B: No." (3) Cloze Test: LLMs are tasked with filling in the masked word using "hateful" or "neutral" in the phrase "It is a [Mask] statement." (4) Chain-of-Thought (CoT) (Wei et al., 2022): In addition to the binary response, LLMs are also required to generate a corresponding explanation simultaneously. (5) Multi-task with Target: LLMs are instructed to provide the binary response and identify the targeted individual or groups. For comprehensive details on each prompt type, refer to Appendix A.1.

3 Experiment Settings

Models We conduct experiments with three kind of LLMs, LLaMA-2-7b (chat), Mixtral-8x7b, and GPT-3.5-Turbo.

Datasets Our experiments use three implicit hate speech detection datasets: Latent Hatred (ElSherief et al., 2021), SBIC (v2) (Sap et al., 2020), and ToxiGen (Hartvigsen et al., 2022). See Appendix A.2 for more details of the data preprocess.

Metrics Our evaluation encompasses both primary classification and uncertainty calibration. In assessing task classification performance, we utilize **Precision**, **Recall**, and **F1** scores to evaluate the predicted answers.

Meanwhile, we employ three metrics for calibration performance: The Area Under the Receiver-Operator Characteristic Curve (**AUROC**) (Bradley, 1997) quantifies the probability that the model assigns a higher uncertainty score to an incorrect prediction than to a correct one. The Expected Calibration Error (**ECE**) (Guo et al., 2017a) is defined as the weighted average squared error between accuracy and confidence within each confidence bin, with weights determined by the fraction of samples falling into the respective bin. The Brier Score (**BS**) (Brier, 1950) measures the mean squared difference between the confidences and the correctness labels. Better calibration is represented with a higher AUC and lower values for ECE or BS.

Experiment Setting We present six demonstrations (i.e., examples) in the prompt for few-shot in-context learning, organized in a balanced class and random order. The verbalized confidence are set to 60%, 70%, 90%, 70%, 60%, 90% in the demonstrations. We adjust the parameters to report a better outcome for each individual model. A consistent temperature setting of 1 is applied to all three models. Greedy decoding is employed for LLaMA-2-7b and GPT-3.5-Turbo, while top-p sampling with a value of 0.9 is opted for Mixtral-8x7b.

4 Results of Primary Classification

Fig. 1 shows the precision and recall of each LLM with various prompt patterns. Models are distinguished by colors (LLaMA-2-7b in green, Mixtral-8x7b in blue, GPT-3.5-Turbo in red), and varying prompt patterns are represented by different shapes. The equilibrium line in the figure represents an ideal state where precision and recall are equal, indicating that the model's predictions do not favor either the positive or negative class. Points above the equilibrium line (recall $>$ precision) indicate a model bias towards classifying most data as positive. Points below the line (recall $<$ precision) suggest a model bias towards classifying most data as negative.

LLaMA-2-7b and Mixtral-8x7b exhibit oversensitive: The majority of points for LLaMA-2-7b and Mixtral-8x7b fall above the equilibrium line in both the Implicit and SBIC datasets. This suggests that these models label statements without hateful intent as hate speech, indicating an excessive level of sensitivity. However, the models did not demonstrate a similar tendency for excessive sensitivity on the ToxiGen dataset.

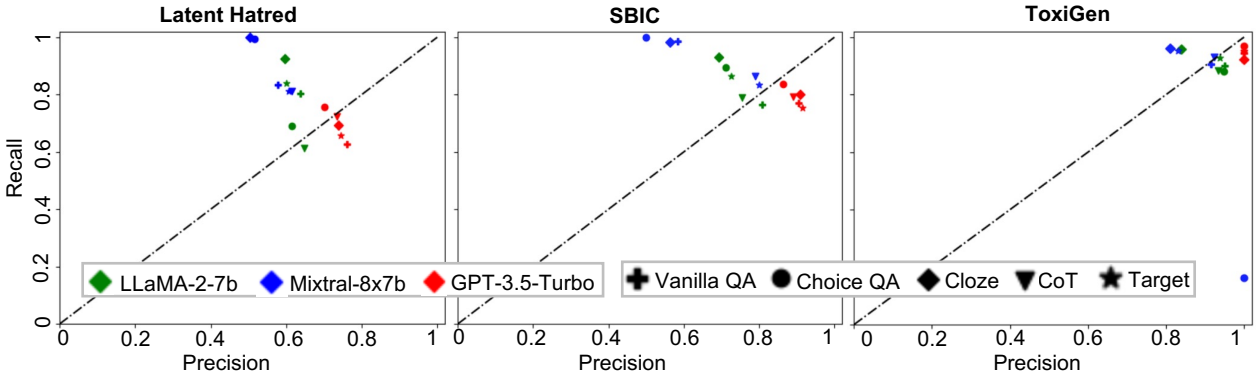

Figure 1: The precision and recall of LLMs with different prompt patterns. The recall is significantly higher than precision for LLMs like LLaMA-2-7b and Mixtral-8x7b on datasets Latent Hatred and SBIC, indicating over-sensitivity.

This discrepancy can be attributed to the fact that many negative examples in the Implicit and SBIC datasets involve sensitive groups or topics, causing confusion in the models and leading to over-sensitivity. In contrast, the negative examples in ToxiGen are straightforward and may not even include sensitive groups or topics. Refer to Appendix A.3 for specific examples.

GPT-3.5 is positioned closer to the upper-right corner, indicating an overall better performance. It also demonstrates a relatively balanced performance between recall and precision, albeit with a slight reversal compared to LLaMA-2-7b and Mixtral-8x7b. This indicates that GPT-3.5 did not exhibit excessive sensitivity, although there is still room for improvement in its ability to detect implicit intentions.

Different prompt patterns exhibit varying degrees of over-sensitivity. In the case of LLaMA-2, the most notable imbalance is observed in the Cloze prompt pattern across all three datasets, with biases ranging from 12% to 33%. The Target pattern shows biases of 24% and 14% on the Latent Hatred and SBIC datasets, respectively, while the Vanilla QA pattern exhibits a bias of 17% on the implicit dataset. Only the CoT pattern demonstrates a relatively balanced performance across all three datasets for LLaMA-2. Regarding Mixtral, all prompt patterns exhibit significant biases on the implicit dataset, ranging from 20% to 50%. On the SBIC dataset, prompt patterns Choice QA, Cloze, and Vanilla QA all exceed 40%, and on the ToxiGen dataset, both Cloze and Target biases surpass 10%. Although different prompt patterns result in varied performances, they consistently demonstrate similar trends on the same model.

	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: 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.

We also present the F1 score in Appendix A.4. We find that when the F1 score achieves its best performance, there can be a significant imbalance between precision and recall. This cautions us against relying solely on F1 and overlooking the issue of imbalance between precision and recall.

5 Results of Confidence Calibration

We point out the performance of uncertainty estimation methods varies in different scenarios in Sec. 5.1. We analyze the reasons why each uncertainty estimation method performs best in a specific scenario in Sec. 5.2 - Sec. 5.5. We point out the common drawbacks of these uncertainty estimation methods in Sec. 5.6. Then We discuss the effects of prompt patterns in Sec. 5.7 and the effects of temperature and top p sampling in Sec. 5.8.

5.1 Calibration varies in different scenarios

Table 1 shows the calibration performance of three mainstream confidence estimation methods. We

	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 1: Calibration performance of three mainstream confidence estimation methods

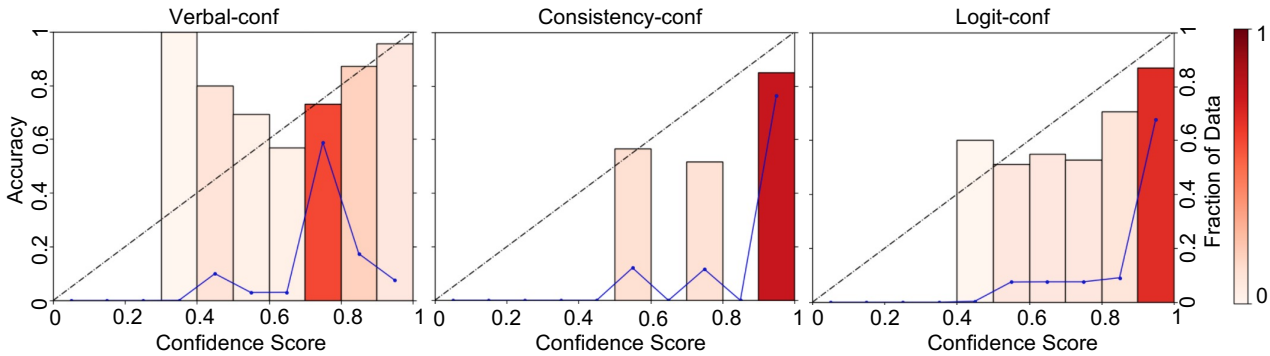

Figure 3: The ECE performance of an example (LLaMA-2-7b on SBIC) in the scenario with poor classification and high output token logit. The verbal-based method is best.

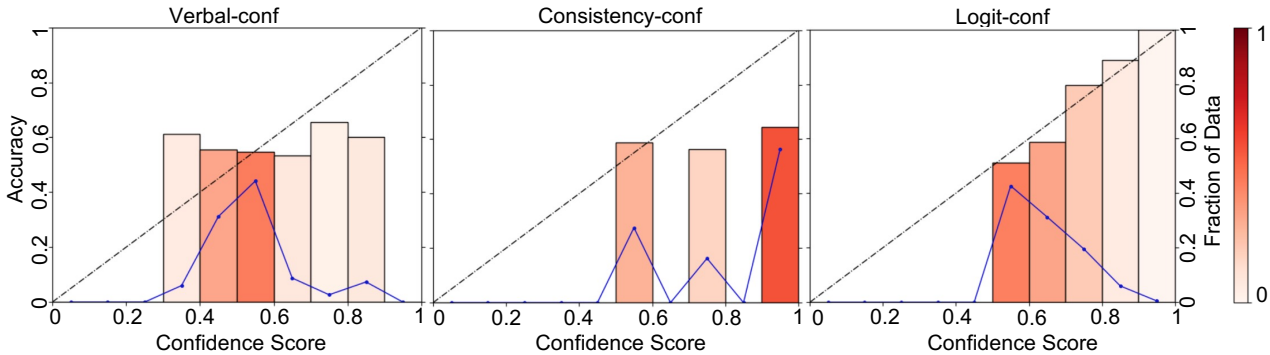

Figure 4: The ECE performance of an example (Mixtral-8x7b on Latent Hatred) in the scenario with poor classification and low output token logit. The logit-based method is best.

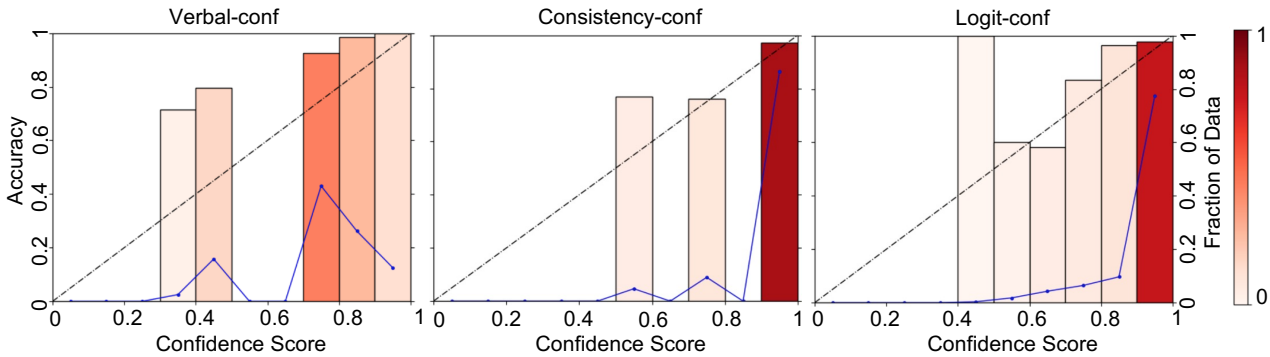

Figure 5: The ECE performance of an example (LLaMA-2-7b on ToxiGen) in the scenario with excellent classification. The consistency-based method is best. (The bar's color and blue line both represent the proportion of the data.)

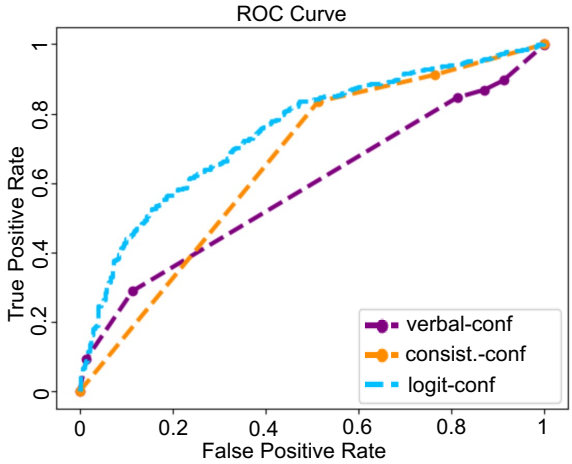

Figure 6: The comparison of the ROC curve.

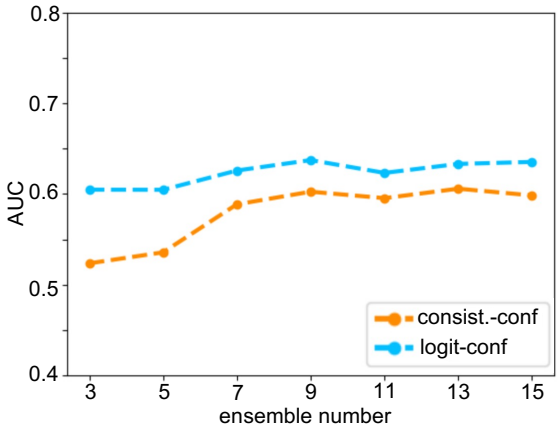

Figure 7: The figure showcases the relationship between AUC and the ensemble number.

find that the calibration performance of various uncertainty estimation methods is influenced by both the probability distribution of the model and the performance of the primary classification task.

We examine the confidence value distribution for each calibration method. The verbal-based method typically produces a relatively conservative confidence value. The consistency-based method usually demonstrates a high agreement rate. The logit-based confidence shows little difference between the distribution in the ensemble and the distribution in individual inferences.

The probability distributions of output tokens for different LLMs also show significant differences. LLaMA-2-7b and GPT-3.5-turbo tend to assign a very high probability to the output tokens (mostly concentrated above 0.9, Fig. 3), while Mixtral-8x7b tends to assign a relatively conservative probability to the output tokens, mainly concentrated in the range of 0.5-0.7 (Fig. 4).

The primary classification performance shows noticeable differences among different datasets. As shown in Sec. 4, Latent Hatred and SBIC are more challenging, resulting in lower F1 scores for all models, whereas the ToxiGen dataset is relatively simpler, leading to higher F1 scores for all models.

Based on these observations, we categorize several scenarios as shown in Fig. 2. The uncertainty method that yields the best calibration performance is highlighted in each scenario. We find that despite the strongest AUC performance of the logit-based method, the ECE for each method exhibits variations across different scenarios. This suggests that relying solely on AUC for calibration assessment is insufficient. Next, we delve into a detailed analysis of the reasons behind this phenomenon.

5.2 The logit-based method achieves the best AUC in all scenarios

The logit-based method performs better in AUC than both the verbal-based method and the consistency-based method in all scenarios. We compare the ROC curve composed of false positive rate (FPR) and true positive rate (TPR) in Fig. 6 for LLaMA-2-7b on Latent Hatred.

The AUC of the verbal-based method is lower than the logit-based method due to the conservative verbalized confidence. We observe that, at the same confidence threshold, the FPR of the verbal-based method and the logit-based method are similar. However, the confidence distribution of the logit-based method is more overconfident (Fig. 3), leading to a larger TPR, resulting in a higher ROC curve.

The discreteness of confidence score in the consistency-based method leads to a reduced AUC. The consistency score in the consistency-based method is limited by the ensemble number, mainly concentrated on a few discrete values (3/5, 4/5, 5/5 in Fig. 3), leading to fewer confidence thresholds considered when calculating AUC. In contrast, the logit-based method's confidence score covers many continuous values, enabling more values of the confidence threshold for the calculation of AUC.

As shown in Fig. 6, The discrete points obtained by the consistency-based method on the ROC curve are very close to those on the curve of the logit-based method. However, the absence of a consistency-based confidence score between 0.8 and 1 results in the omission of the corresponding section with an FPR below 0.5. This indicates that the discreteness of confidence values in the

consistency-based method limits its ability to express uncertainty.

Based on the findings, we increase the ensemble number from 3 to 13 (changing demonstrations in the prompt in each inference to conduct the ensemble), the AUC gradually increases and tends to stabilize (Fig. 7). It indicates that increasing the number of ensemble sources can mitigate the gap but is still lower than the logit-based method.

However, the advantage of the logit-based method in AUC does not always manifest in ECE. We discuss the reasons for the variations in ECE in the following sections.

5.3 When can verbal-based method be better?

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.

That is because the logit-based method and the consistency-based method exhibit overconfidence while the verbal-based method provides a more conservative estimate of confidence. The majority of confidence scores are above 0.9 for both logit-based and consistency-based methods, yet the accuracy of the task remains low, resulting in under-calibrated errors (Fig. 3). In contrast, the verbalized confidence concentrates in the range of 0.7-0.8, close to the accuracy, resulting in smaller calibration errors.

5.4 When can logit-based method be better?

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.

That is because the conservative logit is already well-calibrated so the verbal-based method loses its advantage. As shown in Fig. 4, unlike other models that output a high token logit, the logit of the Mixtral-8x7b model is predominantly concentrated between 0.5 and 0.6, closely resembling the mediocre accuracy and exhibiting good calibration ability. However, the consistency-based method still maintains excessive confidence, resulting in poor calibration.

5.5 When can consistency-based method be better?

In cases where the classification has high accuracy (all models on the ToxiGen dataset), the consistency-based method achieves the best ECE.

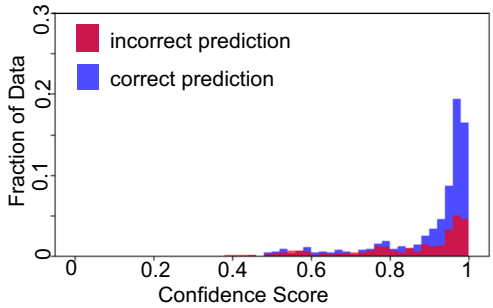

Figure 8: The confidence distribution of correctly classified and misclassified cases.

This is because, on simple datasets, the task accuracy is very high, so the consistency-based method with high confidence tends to be closer to the high accuracy (Fig. 5). However, the verbal-based method maintains a conservative confidence which is lower than the high accuracy, thus leading to over-calibration. For the logit-based method, when the token logit is high (LLaMA2-7b-chat, GPT-3.5-turbo), the ECE is very similar to the consistency method. However, the logit-based method for Mixtral-8x7b has a quite lower ECE because of its conservative token logit.

5.6 Common Drawbacks

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. 8 shows that confidence distributions of correctly classified and misclassified cases overlap significantly, indicating the poor ability of uncertainty estimation.

5.7 Effects of prompt patterns on calibration

Table 4, Table 5 (Appendix A.5), and Table 6 show that the performance of different prompts

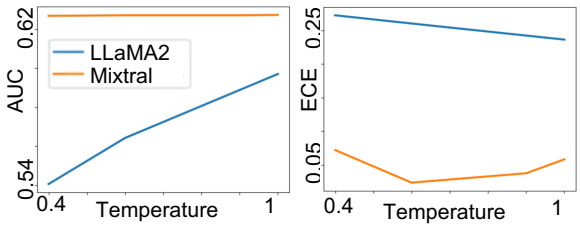

Figure 9: Calibration performance with varying temperature. LLaMA-2-7b and Mixtral-8x7b show different tends.

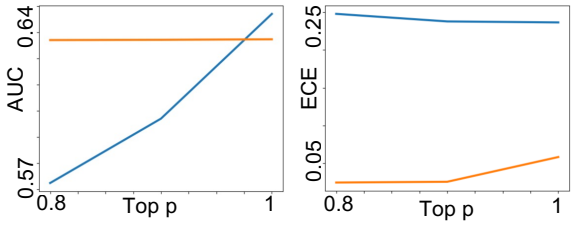

Figure 10: Calibration performance with varying top p sampling. LLaMA-2-7b and Mixtral-8x7b show different trends.

varies in calibration. No prompt consistently performs better. The ensemble of responses obtained from different prompt patterns shows a relatively better overall performance. This may be because different prompt patterns inspire the model to infer results along different paths, and aggregating such results better reflects the model's confidence.

5.8 Effects of the temperature and sampling

We find that due to differences in the output token logit distribution, different models exhibit varying sensitivities to temperature and sampling parameters, sometimes even in opposite trends.

As shown in Fig. 9, as the temperature increases, The AUC of LLaMA2 increases while the AUC of Mixtral decreases. When the temperature varies between 0.6 and 1, LLaMA2 and Mixtral exhibit opposite trends in ECE. Results on changing the top p sampling show the same findings (Fig. 10). This is because the output token logit of LLaMA2 is overconfident, whereas Mixtral exhibits a more cautious level of confidence. As the temperature increases, the logit distribution of Mixtral becomes sharper, leading to over-calibration. On the other hand, the logit distribution of LLaMA2 becomes smoother, enhancing its ability to differentiate confidence levels. See Appendix A.6 for details.

6 Related Work

Implicit Hate Speech Detection Hate speech inflicts significant harm on specific communities. Researchers have developed various hate speech detection models (Gitari et al., 2015; Zhang et al., 2018). Recently, there has been a growing interest among researchers in addressing implicit hate or abusiveness (Kim et al., 2022; Lin, 2022; Ocampo et al., 2023).

Certain studies have employed LLMs to generate explanations with step-by-step reasoning for detecting implicit hate speech (Yang et al., 2023). While (Huang et al., 2023) explored the quality of explanations generated by ChatGPT for detecting implicit hateful tweets, their research exclusively focused on implicit instances, neglecting non-implicit statements. In our exploration, we assess the performance of LLMs in detecting implicit hate speech, taking into account various aspects with a thoughtful consideration. We investigate whether there is an imbalance in the predictions of positive and negative classes when LLMs are employed for hate speech detection. Additionally, we thoughtfully consider various prompt patterns in our analysis.

Uncertainty Estimation in LLMs The Reliable uncertainty estimation can facilitate more rational and informed decision-making. There are three mainstream methods for estimating the confidence of LLMs in their responses, particularly for cases where we can only obtain the answer and logit but not the weights. They are the verbal-based method (Lin et al., 2022; Kadavath et al., 2022), the consistency-based method (Wang et al., 2022; Xiong et al., 2023; Yue et al., 2023), and the logit-based method (Guo et al., 2017b; Zhang et al., 2020; Jiang et al., 2020; Chen et al., 2022). We compare these three mainstream methods, conduct a comprehensive analysis in different scenarios, and delve into the underlying reasons. Additionally, we highlighted the shortcomings of these methods in confidence assessment for fairness.

7 Conclusion

The ability of LLMs in implicit hate speech detection is insufficiently examined. In this paper, we explore the performance of LLMs in identifying implicit hate speech and the calibration of three mainstream uncertainty estimation methods. We find that LLMs have fallen into two extremes, excessively focusing on sensitive groups and exhibiting extreme confidence score distributions. These extremes result in over-sensitivity and poor calibration respectively. These discoveries unveil new limitations of LLMs, underscoring the need for caution when optimizing models to ensure they do not veer towards extremes.

8 Limitation

Due to resource constraints, our current empirical evaluation has not been tested on a wider range of open-source and closed-source LLMs. At the same time, we only focused on hate speech detection in English and did not test the hate speech detection and calibration capabilities of LLMs in multiple languages.

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.

Yangyi Chen, Lifan Yuan, Ganqu Cui, Zhiyuan Liu, and Heng Ji. 2022. A close look into the calibration of pre-trained language models. *ArXiv*, abs/2211.00151.

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*.

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

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

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017b. On calibration of modern neural networks. In *International Conference on Machine Learning*.

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.

Zhengbao Jiang, J. Araki, Haibo Ding, and Graham Neubig. 2020. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

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*.

Youngwook Kim, Shinwoo Park, and Yo-Sub Han. 2022. Generalizable implicit hate speech detection using contrastive learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6667–6679.

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*.

Jessica Lin. 2022. Leveraging world knowledge in implicit hate speech detection. *arXiv preprint arXiv:2212.14100*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.

Nicolas Ocampo, Elena Cabrio, and Serena Villata. 2023. Playing the part of the sharp bully: Generating adversarial examples for implicit hate speech detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2758–2772.

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*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

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*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.

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*.

Murong Yue, Jie Zhao, Min Zhang, Liang Du, and Ziyu Yao. 2023. Large language model cascades with mixture of thoughts representations for cost-efficient reasoning. *arXiv preprint arXiv:2310.03094*.

Jize Zhang, Bhavya Kailkhura, and Thomas Yong-Jin Han. 2020. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. *ArXiv*, abs/2003.07329.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, pages 745–760. Springer.

A Example Appendix

A.1 The design of prompt patterns

We show the full prompt in Table 2.

A.2 Datasets

Our experiments use three implicit hate speech detection datasets to assess LLMs bias on fairness: ToxiGen (Hartvigsen et al., 2022), Latent Hatred (ElSherief et al., 2021), and SBIC (v2) (Sap et al., 2020). We have two steps to preprocess the data to ensure the quality of the evaluation. Firstly, we discard the data samples with explicit hate words, such as "bi*ch" and "fu*k". Secondly, we sample from the test set to keep the equal number of positive and negative data. Finally, we retain 1200 test data for Latent Hatred, 1200 test data for SBIC, and 260 test data for ToxiGen.

A.3 Oversensitive examples

Examples of negative class in dataset implicit and sbic contains sensitive groups or sensitive topics, such as "vandals target black womans maryland home". This is a description of a news event that involves a fairness issue but does not contain any hateful emotions. However, Examples of negative class in dataset toxigen is simple ("muslims in the 19 century migrated from Europe and the Middle East to the U.S.") and even doesn't contain sensitive groups or topics ("i felt so much better after confiding that in someone and realizing i'm not alone").

A.4 Classification performance of different prompt patterns

The precision, recall and F1 for classification performance can be found in tables 3.

A.5 Calibration performance of different prompt patterns

Table 4, Table 5, and Table 6 show that the performance of different prompts varies in calibration.

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 2: The prompt of different prompt patterns.

	Latent Hatred			SBIC			ToxiGen		
	p	r	f1	p	r	f1	p	r	f1
GPT3.5-turbo									
Choice QA	0.7014	0.758	0.7286	0.864	0.8381	0.8508	1	0.969	0.9843
CoT	0.7347	0.7236	0.7291	0.8895	0.7917	0.8377	1	0.9453	0.9719
Cloze	0.738	0.6935	0.715	0.9093	0.8017	0.8521	1	0.9225	0.9597
Vanilla QA	0.7607	0.6263	0.687	0.9039	0.7696	0.8314	1	0.9457	0.9721
Target	0.7452	0.6566	0.6981	0.9152	0.755	0.8274	1	0.9219	0.9593
LLaMA2-7B									
Choice QA	0.6143	0.6913	0.6505	0.7122	0.8956	0.7934	0.9487	0.881	0.9136
CoT	0.6472	0.6134	0.6299	0.7548	0.7913	0.7726	0.9344	0.8837	0.9084
Cloze	0.5954	0.9258	0.7248	0.6947	0.931	0.7957	0.8394	0.9583	0.8949
Vanilla QA	0.6373	0.8038	0.6425	0.8092	0.7646	0.7863	0.9508	0.8992	0.9243
Target	0.601	0.8395	0.7005	0.7261	0.8645	0.7893	0.937	0.9297	0.9333
Mistral									
Choice QA	0.5161	0.995	0.6796	0.5	1	0.6667	1	0.1628	0.28
CoT	0.6155	0.8124	0.7004	0.7896	0.8633	0.8248	0.9231	0.9302	0.9266
Cloze	0.503	0.9983	0.6689	0.5642	0.9817	0.7165	0.8105	0.9612	0.8794
Vanilla QA	0.5771	0.8342	0.6822	0.584	0.985	0.7333	0.9141	0.907	0.9105
Target	0.6058	0.8107	0.6934	0.8	0.8333	0.8163	0.8311	0.9535	0.8881

Table 3: The precision, recall and F1 for different prompts.

Prompt	Latent Hatred			SBIC			Toxigen		
-	AUC	ECE	BS	AUC	ECE	BS	AUC	ECE	BS
Choice QA	0.628	0.127	0.242	0.751	0.088	0.167	0.924	0.069	0.052
CoT	0.631	0.215	0.279	0.736	0.139	0.191	0.902	0.037	0.056
Cloze	0.647	0.245	0.296	0.696	0.152	0.205	0.877	0.013	0.078
Vanilla QA	0.637	0.236	0.287	0.693	0.129	0.186	0.823	0.008	0.059
Target	0.614	0.235	0.294	0.720	0.141	0.196	0.880	0.026	0.048
Ensemble	0.637	0.154	0.244	0.749	0.094	0.165	0.889	0.041	0.047

Table 4: Impact of prompt patterns on Calibration for LLaMA-2-7B.

Prompt	Latent Hatred			SBIC			Toxigen		
-	AUC	ECE	BS	AUC	ECE	BS	AUC	ECE	BS
Choice QA	0.710	0.095	0.237	0.836	0.184	0.246	0.842	0.273	0.211
CoT	0.700	0.039	0.209	0.787	0.091	0.136	0.927	0.127	0.065
Cloze	0.668	0.242	0.277	0.770	0.157	0.217	0.879	0.102	0.098
Vanilla QA	0.682	0.026	0.228	0.728	0.094	0.204	0.911	0.265	0.138
Target	0.693	0.028	0.210	0.781	0.100	0.142	0.888	0.154	0.104
Ensemble	0.710	0.048	0.222	0.762	0.066	0.173	0.924	0.220	0.106

Table 5: Impact of prompt patterns on calibration for Mixtral-8x7b.

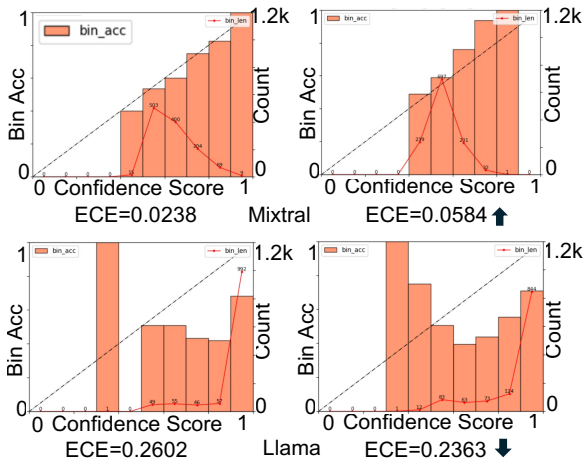

Figure 11: The logit distribution and confidence bin of LLaMA-2-7b and Mixtral-8x7b.

A.6 Analysis of the effects on the temperature and top p sampling

These difference of effect on llama and mixtral arises because of the different logit distribution as shown in Fig. 11. The logit distribution of Mixtral is primarily concentrated between 0.5 and 0.7, while Llama's logit is mainly distributed between 0.9 and 1.0. This indicates that Llama is over confident, whereas Mixtral exhibits a more cautious level of confidence. As the temperature increases, the logit distribution of Mixtral becomes sharper, leading to over-calibration. On the other hand, the logit distribution of Llama becomes smoother, enhancing its ability to differentiate confidence levels.

Prompt	Latent Hatred			SBIC			Toxigen		
-	AUC	ECE	BS	AUC	ECE	BS	AUC	ECE	BS
Choice QA	0.646	0.187	0.248	0.811	0.076	0.125	0.973	0.047	0.015
CoT	0.654	0.171	0.229	0.818	0.083	0.132	0.911	0.031	0.026
Cloze	0.683	0.177	0.232	0.823	0.069	0.123	0.963	0.048	0.024
Vanilla QA	0.638	0.191	0.246	0.834	0.090	0.138	0.924	0.029	0.026
Target	0.658	0.183	0.241	0.800	0.090	0.141	0.936	0.030	0.030
Ensemble	0.668	0.151	0.219	0.858	0.067	0.118	0.959	0.045	0.021

Table 6: Impact of prompt patterns on calibration for GPT-3.5-Turbo.