

Uncertainty Quantification to Improve the Reliability of LLM-Based Hate Speech Detection

Nehir Idil Yuksel^{1*}

Daryna Dementieva^{1,2}

Alexander Fraser^{1,2,3}

¹Technical University of Munich (TUM), ²Munich Center for Machine Learning (MCML),

³Munich Data Science Institute

idilyuksel@gmail.com, daryna.dementieva@tum.de

Abstract

While Large Language Models (LLMs) have been tested for hate speech detection, their reliability in subjective annotation settings remains underexplored. Existing improved prompting approaches for hate speech classification often rely on continuous context retrieval, which can be computationally expensive and unnecessary for confident predictions. To address this, we propose an uncertainty-aware prompting pipeline that integrates **Uncertainty Quantification (UQ)** into LLM-based hate speech classification. The framework uses uncertainty scores to identify ambiguous and potentially unreliable predictions through threshold-based filtering, enabling selective application of advanced context-enriched prompting only when needed. We evaluate the approach on two benchmark datasets—Measuring Hate Speech (Sachdeva et al., 2022) and CREHate (Lee et al., 2024)—across five LLMs using two advanced prompting strategies: *Persona-Aware* and *Annotation-Grounded Few-Shot Prompting*. Experimental results show that the proposed pipeline consistently improves classification performance and reduces uncertainty, leading to more reliable model behavior with improved computational efficiency.

1 Introduction

LLMs have been tested as hate speech classifiers in recent studies to leverage their generalization capabilities across domains (Choudhary et al., 2025; Nirmal et al., 2024). However, prior work has shown that LLM performance in hate speech domain is lower compared to other tasks (Zhu et al., 2024). One of the main challenges is that hate speech detection is inherently subjective. Even human annotators might disagree on whether a text includes hate speech or not, according to their

*This work was conducted by Nehir Idil Yuksel as part of her Master’s thesis at TUM.

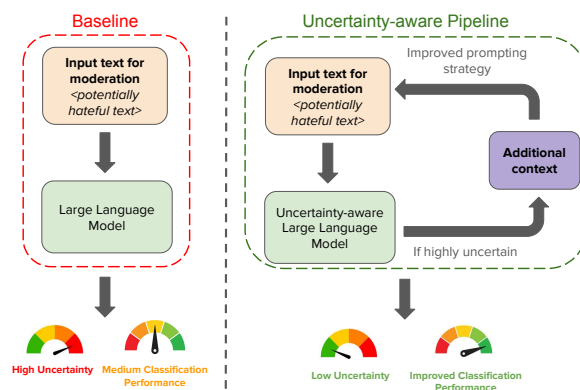


Figure 1: **Uncertainty-aware Pipeline for Hate Speech Classification:** to reduce the overhead of continuous context retrieval, advanced prompting is applied only to highly uncertain samples, improving both classification performance and reliability.

background, such as age, gender, race, and education. Unlike traditional classification tasks, relying solely on deterministic outputs from LLMs may fail to capture the perspectives based on the context. At the same time, online moderation systems often require low-latency responses, whereas continuous context retrieval for every sample can significantly increase inference time. Thus, uncertainty estimation can provide a mechanism to identify only unreliable model predictions that can be improved with additional information added. The relationship between model uncertainty, prediction correctness, and human disagreement remains underexplored in LLM-based hate speech detection.

In this work, we aim to close the gap and evaluate and improve hate speech detection models performance by exploiting uncertainty quantification methods. Our contributions are the following:

- We design an **uncertainty-aware** LLM-based hate speech classification pipeline that includes high-uncertainty samples identification followed by improved prompting (Figure 2).

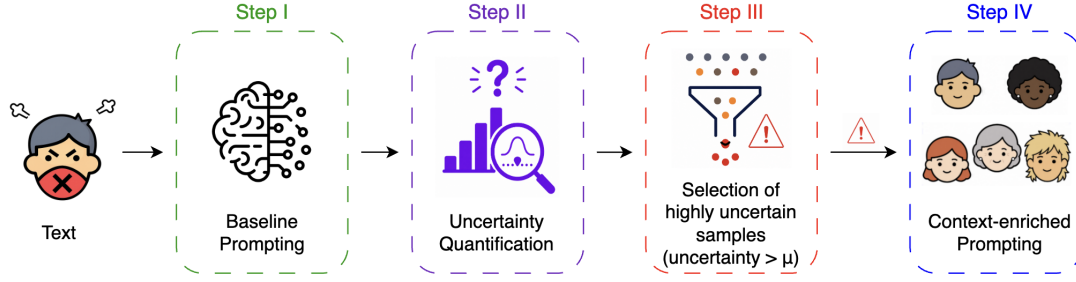


Figure 2: Overview of our proposed Uncertainty-aware Hate Speech Detection with Prompting Framework. In **Step I**, the input text is first processed using baseline prompting to obtain an initial prediction. In **Step II**, uncertainty quantification methods are applied. In **Step III**, samples with uncertainty are identified as high-uncertainty cases. In **Step IV**, these selected samples are re-evaluated using additional context comparing two advanced context-enriched prompting strategies: **Step IV-A** applies persona prompting with annotator demographic attributes; **Step IV-B** applies Annotation-Grounded Few-Shot prompting using retrieved similar labeled examples.

- We then test our new uncertainty-aware hate speech detection pipeline on **two datasets**—Measuring Hate Speech (MHS) (Sachdeva et al., 2022) and CREHate (Lee et al., 2024)—with **two contextual-enriched prompting techniques** providing classification results analysis as well as the uncertainty scores and their impact on the classification results.
- Additionally, we investigate the relationship between **uncertainty metrics, classification performance, and human disagreement**.

Our experiments show that selectively applying context retrieval and advanced prompting to highly uncertain samples consistently improves classification performance and enhances the reliability of model predictions. The code and the evaluation data to replicate the results are available for public usage online.¹

2 Related Work

Hate Speech Detection in NLP As a result of the rapid growth of social media platforms, content that includes hate speech is spreading quickly and targeting people from different identity groups; hence, hate speech detection has become one of the most notable topics in the NLP domain. Hate speech detection methods can be examined under two main categories: task-based fine-tuned classifiers and approaches that leverage LLMs.

Although LLMs are widely used across various classification tasks, (Zhu et al., 2024) finds that their performance on hate speech detection remains

relatively poor. Ghorbanpour et al. (2025) investigated instruction-tuned LLMs for detecting hate speech in eight different non-English languages, where they found that different LLMs and prompt techniques work better for different languages.

Singh et al. (2025) demonstrate that LLaMA3.1-8B outperforms mBERT across different hate speech benchmarks after transfer learning. Choudhary et al. (2025) report that GPT-2 outperforms traditional machine learning models (e.g., SVM+tf-idf) and transformer-based baselines such as BERT, RoBERTa, and HateBERT. Guo et al. (2024) highlights that LLM performance is highly dependent on the prompt format. After experimenting with four different prompting strategies, they found that CoT provides the best F1 score, while few-shot remains the weakest option. Beyond prompting strategies, they also compared BERT, RoBERTa, and ChatGPT, reporting that ChatGPT outperformed on the HateXplain dataset (Mathew et al., 2022).

Annotator Disagreement and Subjectivity

While having different opinions from different annotators is generally viewed as annotation errors, Plank (2022) argues that this is human label variation. Differences among annotators should be seen as valuable information rather than noise.

Wan et al. (2023) examined the impact of demographic information in detecting annotator disagreement using five subjective NLP tasks. They fine-tuned RoBERTa-base ((Liu et al., 2019)) in two different settings: group-level demographics and personal-level demographics. Their findings show us that personal-level settings produce higher F1 scores. Overall, they made a conclusion that annotator demographics data is a powerful infor-

¹<https://github.com/TUM-NLP/uncertainty-aware-hate-classification>

mation on predicting annotator disagreement.

Fleisig et al. (2023) argue that in subjective tasks such as hate speech detection, aggregated labels often fail to capture the opinions of minority groups. Yuan et al. (2025) studied how different personality traits affect hate speech detection using MBTI-based personas. Their survey showed that MBTI types influence how humans label hate speech. Beck et al. (2024) conducted a comprehensive study on sociodemographic prompting using 17 LLMs across seven different subjective classification datasets. According to their findings, sociodemographic prompting changes classifications by up to 80% across different LLMs.

Uncertainty Quantification in LLMs Uncertainty estimation methods are powerful tools for detecting hallucinations to make the usage of LLMs safer. There are multiple attempts to define and measure the uncertainty (Jiang et al., 2021; Manakul et al., 2023; Portillo Wightman et al., 2023). Although uncertainty detection methods for LLMs have strong theoretical contributions, there has been a lack of practical implementations until (Fadeeva et al., 2023) proposed a framework that includes state-of-the-art uncertainty quantification methods. Our study adapts and evaluates a subset of the methods they implemented in a new context to investigate their effectiveness for the subjective hate speech task.

Uncertainty Quantification in Hate Speech Detection There are several works that apply uncertainty quantification in hate speech domain. Earlier work conducted by (Miok et al., 2019) uses Monte Carlo Dropout in recurrent neural networks. (Vazhentsev et al., 2023) found that combining epistemic and aleatoric uncertainty methods performs better than relying on a single method in ambiguous tasks, such as toxicity detection. (Urbiniati et al., 2025) used uncertainty to measure the bias in 11 models on samples from women and non-white annotators. In our work, we perform large-scale comparison of 6 uncertainty quantification methods applied to 5 LLMs.

3 Methodology

We present our uncertainty-aware prompting framework and advanced prompting techniques, which incorporate persona and context into the prompt.

3.1 Uncertainty-Aware Prompting Framework

As illustrated in Figure 2, the framework includes steps below:

Step I: Apply **baseline zero-shot prompting** to all test instances to generate initial hate speech predictions.

Step II: Compute **uncertainty scores** for each prediction.

Step III: Use the **uncertainty score threshold** to identify **high-uncertainty samples** with potentially unreliable predictions, which are subsequently treated as more challenging instances.

Step IV: Apply **advanced context-enriched prompting techniques** to the filtered subset, including: **Step IV-A - Persona Prompting:** Inject annotator demographic context into the prompt; **Step IV-B - Annotation-Grounded Few-Shot Prompting:** Retrieve semantically similar annotated examples and use them as contextual demonstrations.

3.2 Uncertainty as a Indicator of Unreliable Hate Speech Classification

In contrast to prior work (Zhang et al., 2024) that primarily evaluates uncertainty and calibration in hate speech classification, we use uncertainty estimation as a routing signal to identify potentially unreliable LLM predictions for selective re-evaluation. After obtaining the initial predictions from the baseline prompting setup, we compute uncertainty scores for each response and identify instances with high uncertainty. Based on one or multiple aggregated uncertainty measures, these instances can be filtered using a predefined threshold and subsequently re-evaluated with an LLM augmented with additional context. This uncertainty-driven filtering strategy reduces computational overhead and inference time by selectively applying the more expensive secondary prompting stage only to a subset of potentially unreliable examples.

3.3 Advanced Prompting Techniques

For high-uncertainty samples, we then apply two advanced prompting strategies—*Persona Prompting* and *Annotation-Grounded Few-Shot Prompting*—designed to provide richer contextual information for improved hate speech classification.

Full prompt templates and examples are provided in Appendix C.

3.3.1 Persona Prompting

This methodology incorporates annotator demographic information through a persona-based prompt template. Specifically, it extends the baseline by including specific annotator attributes including gender, race, age, and education level to potentially capture demographic-influenced perspectives in hate speech annotation.

Persona-based prompt (simplified)

You are an annotator with the following background:

Gender: {annotator_gender}
Race: {annotator_race}
Age: {annotator_age}
Education: {annotator_education}

Rate the degree of hate speech in the text by generating a score 0, 1, or 2.

0 = Not Hate Speech
1 = Ambiguous
2 = Hate Speech
...

For each text, we used multiple prompts corresponding to different annotators to simulate the real annotation process with LLMs. After obtaining multiple results, we applied majority voting.

3.3.2 Annotation-Grounded Few-Shot Prompting

In contrast to persona prompting, where demographic attributes are explicitly embedded into the prompt, this approach incorporates annotator behavior in an indirect manner. Specifically, representative annotations from similar annotators are retrieved and used as contextual examples.

Given an input sample, we simulate annotation from N annotators with diverse demographic backgrounds. For each simulated annotator, we retrieve examples from the labeled dataset produced by annotators with similar demographic characteristics and provide their annotations for comparable texts as contextual guidance. The model is then prompted N times using this personalized contextual information, and the final label is obtained through aggregation of the individual predictions. This framework approximates a moderation setting with overlapping judgments from multiple annotators. Examples selection process consists of two

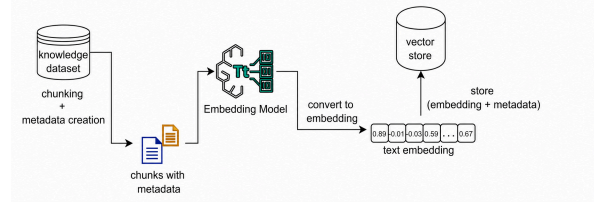


Figure 3: Knowledge Ingestion Pipeline

main stages: (i) knowledge ingestion and (ii) context retrieval.

Knowledge Ingestion The knowledge ingestion pipeline visualized in Figure 3 is designed to transform raw annotations into a structured knowledge base that can later be queried to extract few-shot examples as context.

Annotation-grounded few-shot prompt (simplified)

You are an annotator.

<example_text_and_annotation>
<example_text_and_annotation>
<example_text_and_annotation>

Under consideration of the given examples, rate the degree of hate speech in the text by generating a score 0, 1, or 2.

0 = Not Hate Speech
1 = Ambiguous
2 = Hate Speech
...

Samples in the knowledge dataset are segmented into chunks, with each chunk consisting of a text and its associated annotation reformatted into the style of few-shot examples.

For each text segment, additional contextual information is created as metadata, capturing demographic attributes of the annotator such as age, gender, race, and education level. This ensures that important demographic context is maintained alongside the content and will be used in the filtering stage in context retrieval.

Context Retrieval We designed the context retrieval pipeline, as illustrated in Figure 4, to select the most relevant examples from the knowledge base for each query text. These selected examples are later used as contextual input for prompting.

Each text in the test dataset is converted into an embedding and then used to query the vector store. In the retrieval, we first perform demographic-conditioned filtering to restrict the candidate pool

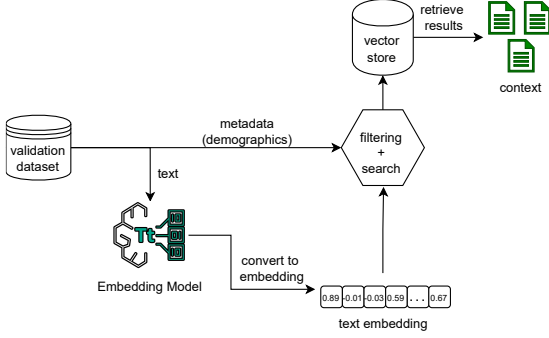


Figure 4: Context Retrieval Pipeline

to annotations provided by annotators with similar demographic profiles.

This differs from standard semantic retrieval approaches, as the retrieval mechanism is designed to capture not only textual similarity but also similarities in annotator behavior and perspective. We then perform semantic similarity search within this demographic-conditioned subset and retrieve the top 10 most relevant texts.

Since some texts appear multiple times due to annotations from different annotators, we aggregate their labels using majority voting to obtain a single representation per text. From the resulting set of unique texts, we select up to 3 closest examples, as this number of examples provides sufficient contextual guidance while reducing noise and limiting prompt length.

4 Experimental Setup

Here, we present the evaluation datasets, uncertainty measures, baseline, and computational setup used in our experiments.

4.1 Dataset

We selected two datasets that contain annotator variation, MHS (Sachdeva et al., 2022) and CREHate (Lee et al., 2024). Both datasets include a three-level annotation scheme, which includes "hate", "not hate", and "ambiguous" classes. The ambiguous class is used when annotators are not sure about the label.

MHS includes comments collected from YouTube, Reddit, and Twitter labeled by annotators from Amazon Mechanical Turk. For each comment, there are different numbers of annotators, and diverse annotators' demographic representation.

On the other hand, CREHate is a dataset designed to address the lack of cultural diversity in

Dataset	Setting	Not Hate	Ambiguous	Hate
MHS	Annotator-level	1,141	106	442
MHS	Majority vote	342	9	149
CREHate	Annotator-level	1,227	149	1,124
CREHate	Majority vote	252	4	244

Table 1: Class distributions for our validation datasets per annotator-level labels and majority-vote aggregation.

hate-speech detection across five English-speaking countries. It comprises 600 country-specific texts collected from YouTube and Reddit using hate-speech keywords elicited via surveys.

Using these datasets that have different data collection and annotation mechanisms allows us to see the effect of dataset characteristics on uncertainty and different prompting strategies. We created a validation dataset consisting of 500 comments from each dataset. Only comments annotated by more than three annotators were included to capture annotator variation. In addition to the validation data, a separate disjoint dataset consisting of 1,000 comments from each dataset was constructed as a knowledge dataset for annotation-grounded few-shot prompting.

Since some samples may include a large number of annotators, we randomly selected 5 annotators for the MHS dataset. For CREHate, one annotator per country was randomly selected to preserve cross-cultural diversity. Class distributions of created validation sets are shown in Table 1.

4.2 Uncertainty Quantification Step

In this work, we applied several uncertainty quantification techniques implemented by (Fadeeva et al., 2023) to assess the uncertainty of LLMs in the hate speech detection task. High levels of uncertainty are treated as a potential indicator of unreliable predictions, including hallucinations.

UQ Methods Considering resource limits, we applied the following methods: Mean Sequence Probability (MSP), Perplexity, Mean Token Entropy (MTE), Mean Pointwise Mutual Information (MPMI), Mean Conditional Pointwise Mutual Information (MCPMI) and, PTrue. Detailed information about their computation logic is presented in Appendix F.

Threshold Selection First, each selected uncertainty quantification metric is independently min-

max normalized to the $[0, 1]$ range. The normalized scores are then averaged to obtain a final uncertainty score for each prediction. As a conservative selection strategy, in this work, the mean uncertainty score across all predictions is used as the threshold for identifying high-uncertainty instances. The distributions of uncertainty scores and the corresponding threshold values are provided in Appendix E.2.

Each uncertainty metric is independently min-max normalized to $[0, 1]$ before averaging.

4.3 Baseline

Firstly, prompting LLM without providing additional context or information serves as a baseline.

Baseline prompt (simplified)

You are an annotator.
Rate the degree of hate speech in the text by generating a score 0, 1, or 2.

0 = Not Hate Speech
1 = Ambiguous
2 = Hate Speech
...

4.4 Models

We leveraged five different LLMs in this work: LLaMA 3 8B Instruct, Mixtral 8x7B Instruct, Mistral 7B Instruct, Qwen3 4B Instruct, and Falcon 3 7B Instruct. They are all open-sourced, mid-sized instruct models with top performance at the time this research was conducted. Detailed licenses information is provided in Appendix A.

4.5 Computational Setup

We implement the knowledge ingestion and retrieval pipeline using Amazon Bedrock Knowledge Bases, Amazon OpenSearch Serverless as the vector store, and Amazon Titan V2 (Stormacq, 2024) for embeddings. All experiments are conducted on Amazon SageMaker GPU instances. Most models are evaluated on an ml.g5.2xlarge instance (24 GiB GPU memory), while Mixtral-8x7B-Instruct requires an ml.g5.12xlarge instance (96 GiB GPU memory) due to its higher computational demands. Apart from instance size, the inference setup remains consistent across models to ensure fair comparison. Following preliminary results (Appendix B), all main experiments use greedy decoding.

5 Results

This section presents the experimental results, including the correlation between uncertainty and reliability metrics, and comparisons of prompting strategies. All experiments are evaluated using 5-fold cross-validation to assess the statistical significance of the results.

5.1 Relationship Between Uncertainty Scores and Classification Correctness

Correlation coefficients between continuous uncertainty scores and binary correctness labels are calculated to evaluate how well model uncertainty reflects prediction correctness as shown in Table 2. Negative correlation values (r) indicate that higher uncertainty is associated with incorrect model predictions which is desirable association.

For both datasets, Mistral-7B-Instruct shows statistically significant strongest negative correlation. This shows that uncertainty estimates produced by this model are more reliable indicators of prediction correctness. Similarly, Falcon3-7B-Instruct shows significant and moderately strong correlations across both datasets. On the other hand other models has weaker correlations.

According to these findings, model uncertainty is moderately associated with classification errors in some settings. Thus, the statistical significance of the correlations suggests that uncertainty indeed captures to certain extend the signal about challenging hate speech classification instances.

Dataset	Model	r	p-value
MHS	Meta-Llama-3-8B	-0.157	0.00042
	Mixtral-8x7B	-0.057	0.201
	Mistral-7B	-0.466	2.90e-28
	Qwen3-4B	-0.026	0.564
	Falcon3-7B	-0.221	6.10e-07
CREHate	Meta-Llama-3-8B	-0.092	0.0393
	Mixtral-8x7B	-0.165	0.000219
	Mistral-7B	-0.231	1.68e-07
	Qwen3-4B	-0.014	0.76
	Falcon3-7B	-0.257	5.37e-09

Table 2: Point-biserial correlations between uncertainty and classification correctness for each LLM. Negative correlations indicate that higher uncertainty is associated with incorrect predictions. Statistically significant p -values ($p < 0.05$) are highlighted in light green. Stronger relationships ($|r| \geq 0.2$) are **bolded**.

		Accuracy	Precision	Recall	F1 Score
<i>MHS</i>					
Meta-Llama-3-8B-Instruct	Baseline	0.460 $\pm$ 0.038	0.802 $\pm$ 0.031	0.460 $\pm$ 0.038	0.447 $\pm$ 0.031
	Persona	0.460 $\pm$ 0.044	0.802 $\pm$ 0.033	0.460 $\pm$ 0.044	0.445 $\pm$ 0.037
	Annotation-Grounded	0.500 $\pm$ 0.038	0.803 $\pm$ 0.027	0.500 $\pm$ 0.038	0.493 $\pm$ 0.032
Mixtral-8x7B-Instruct	Baseline	0.656 $\pm$ 0.076	0.817 $\pm$ 0.029	0.656 $\pm$ 0.076	0.672 $\pm$ 0.070
	Persona	0.650 $\pm$ 0.062	0.822 $\pm$ 0.025	0.650 $\pm$ 0.062	0.668 $\pm$ 0.057
	Annotation-Grounded	0.712 $\pm$ 0.057	0.813 $\pm$ 0.039	0.712 $\pm$ 0.057	0.732 $\pm$ 0.055
Mistral-7B-Instruct	Baseline	0.456 $\pm$ 0.045	0.784 $\pm$ 0.025	0.456 $\pm$ 0.045	0.415 $\pm$ 0.044
	Persona	0.450 $\pm$ 0.044	0.785 $\pm$ 0.025	0.450 $\pm$ 0.044	0.420 $\pm$ 0.044
	Annotation-Grounded	0.516 $\pm$ 0.039	0.771 $\pm$ 0.028	0.516 $\pm$ 0.039	0.501 $\pm$ 0.036
Qwen3-4B-Instruct	Baseline	0.374 $\pm$ 0.030	0.830 $\pm$ 0.030	0.374 $\pm$ 0.030	0.440 $\pm$ 0.045
	Persona	0.378 $\pm$ 0.043	0.824 $\pm$ 0.040	0.378 $\pm$ 0.043	0.433 $\pm$ 0.062
	Annotation-Grounded	0.398 $\pm$ 0.040	0.819 $\pm$ 0.038	0.398 $\pm$ 0.040	0.474 $\pm$ 0.046
Falcon3-7B-Instruct	Baseline	0.568 $\pm$ 0.045	0.774 $\pm$ 0.030	0.568 $\pm$ 0.045	0.584 $\pm$ 0.055
	Persona	0.584 $\pm$ 0.042	0.776 $\pm$ 0.017	0.584 $\pm$ 0.042	0.598 $\pm$ 0.039
	Annotation-Grounded	0.560 $\pm$ 0.023	0.766 $\pm$ 0.035	0.560 $\pm$ 0.023	0.586 $\pm$ 0.032
<i>CREHate</i>					
Meta-Llama-3-8B-Instruct	Baseline	0.576 $\pm$ 0.044	0.701 $\pm$ 0.042	0.576 $\pm$ 0.044	0.613 $\pm$ 0.033
	Persona	0.568 $\pm$ 0.048	0.718 $\pm$ 0.058	0.568 $\pm$ 0.048	0.607 $\pm$ 0.041
	Annotation-Grounded	0.602 $\pm$ 0.059	0.690 $\pm$ 0.066	0.602 $\pm$ 0.059	0.613 $\pm$ 0.062
Mixtral-8x7B-Instruct	Baseline	0.656 $\pm$ 0.026	0.696 $\pm$ 0.021	0.656 $\pm$ 0.026	0.674 $\pm$ 0.019
	Persona	0.676 $\pm$ 0.038	0.730 $\pm$ 0.033	0.676 $\pm$ 0.038	0.699 $\pm$ 0.031
	Annotation-Grounded	0.678 $\pm$ 0.047	0.710 $\pm$ 0.032	0.678 $\pm$ 0.047	0.693 $\pm$ 0.039
Mistral-7B-Instruct	Baseline	0.584 $\pm$ 0.057	0.691 $\pm$ 0.037	0.584 $\pm$ 0.057	0.529 $\pm$ 0.056
	Persona	0.580 $\pm$ 0.054	0.689 $\pm$ 0.043	0.580 $\pm$ 0.054	0.534 $\pm$ 0.057
	Annotation-Grounded	0.602 $\pm$ 0.072	0.681 $\pm$ 0.059	0.602 $\pm$ 0.072	0.559 $\pm$ 0.075
Qwen3-4B-Instruct	Baseline	0.392 $\pm$ 0.033	0.745 $\pm$ 0.071	0.392 $\pm$ 0.033	0.508 $\pm$ 0.033
	Persona	0.418 $\pm$ 0.030	0.736 $\pm$ 0.076	0.418 $\pm$ 0.030	0.526 $\pm$ 0.042
	Annotation-Grounded	0.418 $\pm$ 0.034	0.744 $\pm$ 0.055	0.418 $\pm$ 0.034	0.532 $\pm$ 0.037
Falcon3-7B-Instruct	Baseline	0.580 $\pm$ 0.032	0.632 $\pm$ 0.037	0.580 $\pm$ 0.032	0.603 $\pm$ 0.032
	Persona	0.616 $\pm$ 0.053	0.631 $\pm$ 0.050	0.616 $\pm$ 0.053	0.621 $\pm$ 0.050
	Annotation-Grounded	0.580 $\pm$ 0.055	0.622 $\pm$ 0.053	0.580 $\pm$ 0.055	0.599 $\pm$ 0.054

Table 3: Final classification performance across the MHS and CREHate evaluation datasets. Values are reported as mean $\pm$ standard deviation across 5 folds. For each model and evaluation metric, the best-performing prompting strategy across baseline, persona, and annotation-grounded is highlighted in **bold**. Light green shading indicates the best-performing model-strategy combination for a given metric within each dataset.

5.2 Prompting Strategy Comparison

This section compares the three prompting strategies introduced in Section 3. As illustrated in Figure 2, advanced prompting techniques are selectively applied to high-uncertainty instances identified after the initial inference stage. The comparison therefore evaluates not only the effectiveness of the prompting strategies themselves, but also the impact of uncertainty-based filtering on hate speech classification performance.

5.2.1 Classification Performance Metrics

Firstly, although the prompts explicitly require JSON-formatted outputs, LLM generations may still deviate from the expected structure. Therefore, all outputs are post-processed to extract the predicted labels, while generations that cannot be mapped to valid labels are counted as parsing

failures. Table 9 presents the corresponding failure rates across datasets and prompting strategies. Overall, persona prompting produces fewer parsing failures than annotation-grounded prompting across most models and datasets.

The classification results in Table 3 are evaluated using 5-fold sampling, where performance metrics are computed independently for each fold and reported as mean and standard deviation values. Across both datasets, the enhanced prompting strategies consistently improve classification performance over the baseline in terms of accuracy, precision, recall, and F1 score.

On the MHS dataset, annotation-grounded prompting achieves the strongest overall performance for most models and evaluation metrics. In particular, Mixtral-8x7B-Instruct combined with annotation-grounded prompting obtains the highest

Dataset	Model	Baseline	Persona	Annotation-Grounded
MHS	LLaMA-3-8B	0.466 $\pm$ 0.006	0.484 $\pm$ 0.005	0.424 $\pm$ 0.003
	Mixtral-8x7B	0.535 $\pm$ 0.008	0.514 $\pm$ 0.007	0.474 $\pm$ 0.003
	Mistral-7B	0.574 $\pm$ 0.007	0.479 $\pm$ 0.005	0.532 $\pm$ 0.004
	Qwen3-4B	0.338 $\pm$ 0.007	0.321 $\pm$ 0.005	0.325 $\pm$ 0.005
	Falcon3-7B	0.383 $\pm$ 0.003	0.363 $\pm$ 0.003	0.357 $\pm$ 0.002
CREHate	LLaMA-3-8B	0.496 $\pm$ 0.010	0.413 $\pm$ 0.006	0.443 $\pm$ 0.001
	Mixtral-8x7B	0.517 $\pm$ 0.001	0.514 $\pm$ 0.006	0.506 $\pm$ 0.003
	Mistral-7B	0.518 $\pm$ 0.006	0.434 $\pm$ 0.006	0.495 $\pm$ 0.007
	Qwen3-4B	0.343 $\pm$ 0.003	0.321 $\pm$ 0.005	0.329 $\pm$ 0.003
	Falcon3-7B	0.342 $\pm$ 0.005	0.326 $\pm$ 0.003	0.331 $\pm$ 0.004

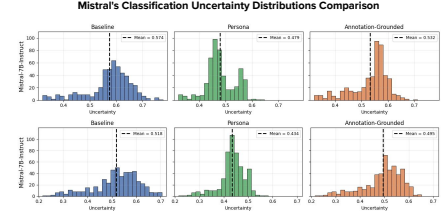


Figure 5: Average uncertainty across prompting strategies on MHS and CREHate (lower is better; distribution plot example is for Mistral across both datasets). Values are reported as mean $\pm$ standard deviation across folds. Light green highlights indicate the lowest uncertainty per model for each dataset. The context-enriched prompting consistently reduces both the mean and variance of model uncertainty, thereby improving prediction reliability.

accuracy, recall, and F1 score among all evaluated configurations.

On the CREHate dataset, annotation-grounded prompting improves the performance of Meta-LLaMA-3-8B-Instruct and Mistral-7B-Instruct, while Falcon3-7B-Instruct performs better with persona prompting. Mixtral-8x7B-Instruct again achieves the strongest overall results, obtaining the highest accuracy and recall with annotation-grounded prompting and the highest F1 score with persona prompting.

Overall, the results indicate that integrating uncertainty-aware filtering with advanced prompting strategies improves hate speech classification performance across both datasets, with annotation-grounded one providing the most consistent gains.

5.2.2 Evaluation of Model Uncertainty

To further analyze the prompting strategies, we evaluate models’ uncertainties distributions. For persona and annotation-grounded prompting, uncertainty scores are aggregated across multiple generations by averaging the uncertainty estimates for each input text. Lower uncertainty values indicate more confident and reliable model predictions.

Table 10 and the distribution plots in Appendix E.2 show that both enhanced prompting strategies generally reduce uncertainty compared to the baseline across most models and datasets, although the optimal strategy varies by model and uncertainty metric. The aggregated results in Table 5 further confirm consistently lower overall uncertainty scores. Annotation-grounded prompting performs best on the MHS dataset, while persona prompting yields lower uncertainty on CREHate, indicating that the effectiveness of uncertainty-aware prompting depends on dataset characteristics and task complexity.

In addition to mean uncertainty values, we analyze the standard deviation of uncertainty estimates to assess uncertainty stability across generations. Lower standard deviation values indicate more consistent uncertainty estimates, whereas higher values reflect greater variability. Results in Table 12 and Table 14 show that annotation-grounded prompting generally produces more stable uncertainty estimates, particularly for Mixtral-8x7B-Instruct and Falcon3-7B-Instruct across both datasets, thus making models’ behavior more stable and reliable.

5.3 Relationship Between Uncertainty and Annotator Disagreement

As annotator disagreement reflects task subjectivity and complexity, we expect samples with higher disagreement to exhibit higher uncertainty. To examine this relationship, we compute Pearson correlation between uncertainty scores and annotator disagreement, where disagreement is measured as normalized Shannon entropy over the label distribution. Results in Table 7 (Appendix E) show mixed correlations across models and datasets. Mistral-7B-Instruct demonstrates the strongest positive correlation on the MHS dataset, while Mistral-7B-Instruct and Falcon3-7B-Instruct exhibit stronger negative correlations on CREHate. Overall, the results suggest that model uncertainty does not consistently align with patterns of human disagreement in our experimental setting.

6 Conclusion

In this work, we proposed a novel uncertainty-aware prompting framework for LLM-based hate speech detection that leverages uncertainty quantification as a signal of prediction reliability. The framework combines uncertainty estimation with context-enhanced prompting strategies to improve

classification for subjective and challenging cases.

We further evaluated two advanced prompting techniques that incorporate additional context retrieval for uncertain samples across two datasets and five LLMs. The proposed pipeline consistently reduces predictive uncertainty and improves classification performance across two datasets and five LLMs. Overall, the results demonstrate that uncertainty-aware filtering enables targeted refinement of unreliable predictions, allowing computationally intensive contextual prompting to be applied selectively where it is most beneficial. This yields an effective and computationally efficient framework for enhancing hate speech classification with LLMs.

7 Limitations

Model Scale The proposed framework is evaluated using mid-sized LLMs to enable computationally efficient experimentation. Future work could extend the analysis to larger and more capable models to further examine the scalability and robustness of uncertainty-aware prompting strategies.

Threshold Selection and Iterative Refinement

Our approach applies a single threshold-based filtering step followed by one additional prompting stage for high-uncertainty samples. Although this setup improves classification performance, future work could investigate iterative refinement with dynamically adjusted thresholds across multiple rounds. In addition, we aggregate multiple uncertainty quantification methods into a single score using a simple averaging strategy. More targeted threshold selection based on a single, task-specific uncertainty metric may provide stronger filtering performance.

Uncertainty Analysis While we analyze the relationship between uncertainty and annotator disagreement, we do not investigate in depth the connection between uncertainty and ambiguous or borderline hate speech cases. Future work could explore whether uncertainty signals align more strongly with semantically ambiguous samples and difficult annotation scenarios.

Embedding Model Selection As the experiments are conducted within the AWS ecosystem, Amazon Titan V2 is used as a general-purpose embedding model for retrieval. Future work could evaluate domain-specific embedding models better

suited for capturing semantic patterns relevant to hate speech detection.

Language and Cultural Scope All experiments are conducted on English-language datasets. Therefore, the applicability of the proposed uncertainty quantification and prompting strategies to multilingual settings remains unexplored. Extending the analysis to other languages would also introduce cultural differences, varying social norms, and different manifestations of hate speech, which may influence both model uncertainty and classification behavior.

Contextual and Demographic Information

The advanced prompting strategies incorporate only limited annotator demographic attributes, including age, gender, education level, and race. However, hate speech interpretation is often influenced by broader contextual and social factors. Additional information, such as characteristics of the content author or prior conversational context, could further improve contextual understanding and uncertainty estimation. The absence of such information limits the ability to fully model factors affecting classification performance and prediction reliability.

8 Ethics Statement

This work studies uncertainty quantification in LLMs for hate speech detection using publicly available datasets. Since the task involves harmful content, there is a risk that such content may be reproduced during model inference. Our experiments are conducted in controlled settings for only research purposes. If these approaches are to be deployed in real-world settings, more extensive testing and safety evaluations should first be conducted.

Our work explicitly models annotator disagreement, which may reflect subjective perspectives influenced by demographic and cultural factors. While we incorporate annotator-related information to improve model reliability, we do not aim to infer or reinforce sensitive personal attributes. Instead, we use these signals to better reflect uncertainty and disagreement.

Finally, we acknowledge that our experiments are limited to English datasets, which may restrict the generalization of our findings across languages and cultural contexts. Future work should consider multilingual and cross-cultural settings to ensure

broader applicability and fairness.

Acknowledgments

This work was co-funded by the European Union (ERC, EPICAL, 101141712). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

References

- Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. [Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615, St. Julian’s, Malta. Association for Computational Linguistics.
- Mahima Choudhary, Basant Agarwal, and Vishnu Goyal. 2025. [Hate speech detection: Leveraging llm-gpt2 with fine-tuning and multi-shot techniques](#). *Procedia Computer Science*, 258:2817–2825. International Conference on Machine Learning and Data Engineering.
- Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. 2023. [LM-polygraph: Uncertainty estimation for language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 446–461, Singapore. Association for Computational Linguistics.
- Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. [When the majority is wrong: Modeling annotator disagreement for subjective tasks](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore. Association for Computational Linguistics.
- Faeze Ghorbanpour, Daryna Dementieva, and Alexander Fraser. 2025. [Can prompting LLMs unlock hate speech detection across languages? a zero-shot and few-shot study](#). In *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*, pages 413–425, Vienna, Austria. Association for Computational Linguistics.
- Keyan Guo, Alexander Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, and Hongxin Hu. 2024. [An investigation of large language models for real-world hate speech detection](#). *Preprint*, arXiv:2401.03346.
- Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. [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, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. [Language models \(mostly\) know what they know](#). *Preprint*, arXiv:2207.05221.
- Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024. [Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4205–4224, Mexico City, Mexico. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#). *Preprint*, arXiv:1907.11692.
- Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. [SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.
- Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2022. [Hatexplain: A benchmark dataset for explainable hate speech detection](#). *Preprint*, arXiv:2012.10289.
- Kristian Miok, Dong Nguyen Doan, Blaz Skrlj, Daniela Zaharie, and Marko Robnik-Sikonja. 2019. [Prediction uncertainty estimation for hate speech classification](#). *CoRR*, abs/1909.07158.
- Ayushi Nirmal, Amrita Bhattacharjee, Paras Sheth, and Huan Liu. 2024. [Towards interpretable hate speech detection using large language model-extracted rationales](#). In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 223–233, Mexico City, Mexico. Association for Computational Linguistics.
- Barbara Plank. 2022. [The “problem” of human label variation: On ground truth in data, modeling and evaluation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

- Gwenyth Portillo Wightman, Alexandra Delucia, and Mark Dredze. 2023. [Strength in numbers: Estimating confidence of large language models by prompt agreement](#). In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 326–362, Toronto, Canada. Association for Computational Linguistics.
- Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. [The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism](#). In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.
- Claude E. Shannon and Warren Weaver. 1949. *The Mathematical Theory of Communication*. University of Illinois Press.
- Daman Deep Singh, Ramanuj Bhattacharjee, and Abhijnan Chakraborty. 2025. [Rethinking hate speech detection on social media: Can llms replace traditional models?](#) *Preprint*, arXiv:2506.12744.
- Sébastien Stormacq. 2024. [Amazon titan text embeddings v2 now available in amazon bedrock, optimized for improving rag](#).
- Junya Takayama and Yuki Arase. 2019. [Relevant and informative response generation using pointwise mutual information](#). In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 133–138, Florence, Italy. Association for Computational Linguistics.
- Alessandra Urbinati, Mirko Lai, Simona Frenda, and Marco Antonio Stranisci. 2025. [Are you sure? measuring models bias in content moderation through uncertainty](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18061–18076, Suzhou, China. Association for Computational Linguistics.
- Liam van der Poel, Ryan Cotterell, and Clara Meister. 2022. [Mutual information alleviates hallucinations in abstractive summarization](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5956–5965, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Artem Vazhentsev, Gleb Kuzmin, Akim Tsvigun, Alexander Panchenko, Maxim Panov, Mikhail Burtsev, and Artem Shelmanov. 2023. [Hybrid uncertainty quantification for selective text classification in ambiguous tasks](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11659–11681, Toronto, Canada. Association for Computational Linguistics.
- Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. [Everyone’s voice matters: quantifying annotation disagreement using demographic information](#). In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence, AAAI’23/IAAI’23/EAAI’23*. AAAI Press.
- Shuzhou Yuan, Ercong Nie, Mario Tawfelis, Helmut Schmid, Hinrich Schütze, and Michael Färber. 2025. [Hateful person or hateful model? investigating the role of personas in hate speech detection by large language models](#). *Preprint*, arXiv:2506.08593.
- Min Zhang, Jianfeng He, Taoran Ji, and Chang-Tien Lu. 2024. [Don’t go to extremes: Revealing the excessive sensitivity and calibration limitations of LLMs in implicit hate speech detection](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12073–12086, Bangkok, Thailand. Association for Computational Linguistics.
- Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2024. [Exploring the capability of chatgpt to reproduce human labels for social computing tasks](#). In *Social Networks Analysis and Mining: 16th International Conference, ASONAM 2024, Rende, Italy, September 2–5, 2024, Proceedings, Part III*, page 13–22, Berlin, Heidelberg. Springer-Verlag.

A Licensing of Resources

Below is an overview of the licenses associated with each resource used in this work (Table 4).

Resource	License	Resource Link
MHS	CC BY 4.0	https://huggingface.co/datasets/ucberkeley-dlab/measuring-hate-speech
CREHate	CC BY-SA 4.0	https://huggingface.co/datasets/nayeon212/CREHate
LLaMA 3 8B Instruct	LLaMA 3 Community License	https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
Mixtral 8x7B Instruct	Apache-2.0	https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1
Mistral 7B Instruct	Apache-2.0	https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
Qwen 3 4B Instruct	Apache-2.0	https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507
Falcon 3 7B Instruct	Falcon LLM License	https://huggingface.co/tiiuae/Falcon3-7B-Instruct
LM-Polygraph	MIT License	https://github.com/HINemo/lm-polygraph

Table 4: Overview of datasets and models used in this work along with their respective licenses.

The licenses associated with the models and datasets utilized in this study are consistent with the intended use of conducting academic research on various NLP application for positive societal impact.

B Selection of Generation Parameters

Before the main experiments, we performed a preliminary evaluation using Mixtral-8x7B-Instruct on the MHS dataset. We tested different temperature values (0, 0.2, 0.5, 0.7, and 1.0), while keeping the top- k parameter fixed at 10 and limiting the maximum number of generated tokens to 100. This evaluation was used to determine the text generation parameters and assess the effect of response diversity on LLM outputs.

F1 score changes across different temperature settings is shown in Figure 6. For all temperature settings, the Annotation-Grounded approach achieves the highest F1 score, and it has a slight decrease as the temperature increases. For all prompting approaches, the best F1 score is achieved in temperature 0.

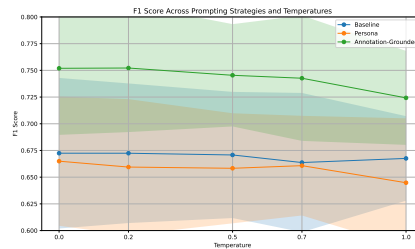
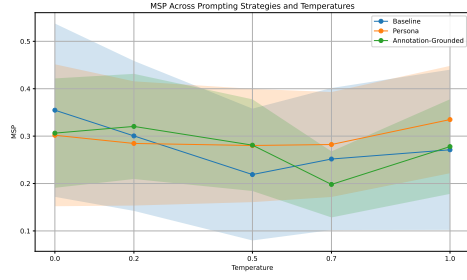


Figure 6: F1 score across prompting strategies and temperatures. Shaded regions indicate standard deviation across folds.

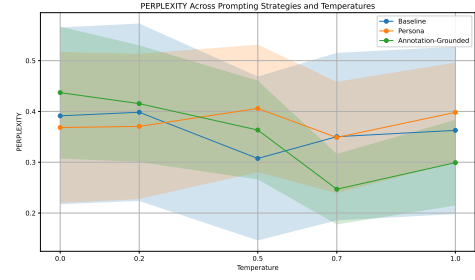
The behavior of multiple uncertainty metrics under different temperature settings is illustrated in Figure 7. Overall, uncertainty values vary with temperature, but there is no single pattern. Across all uncertainty quantification methods, the lowest uncertainty is always achieved with the annotation-grounded approach: at temperature 0.7 for MSP and Perplexity, 1.0 for MTE, 0.2 for PTrue and MCPMI, and 0.0 for MPMI.

In terms of sensitivity, MPMI is the most sensitive uncertainty metric to temperature, exhibiting considerably larger fluctuations compared to the others. Although MPMI exhibits strong fluctuations across temperature values, it consistently shows low standard deviation within each setting.

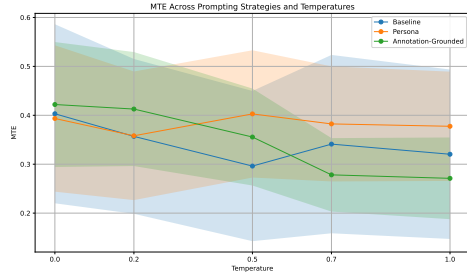
Figure 8 summarizes the average uncertainty across all six uncertainty metrics for different prompting strategies and decoding temperatures by getting averages of individual uncertainties. In contrast to F1



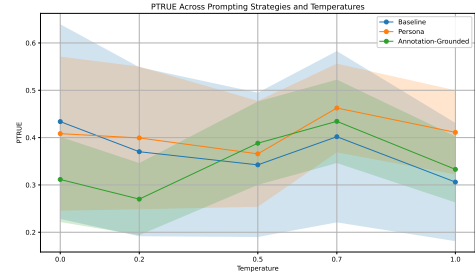
(a) MSP (Maximum Softmax Probability)



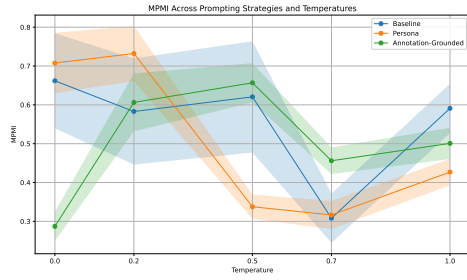
(b) Perplexity



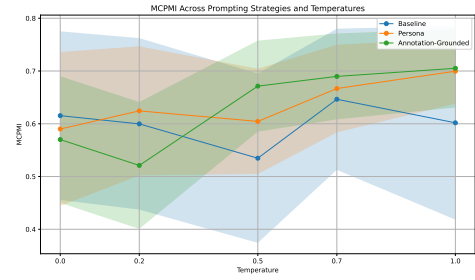
(c) MTE (Model Trust Estimation)



(d) PTrue



(e) MPMI (Mean Pointwise Mutual Information)



(f) MCPMI (Mean Calibrated Pointwise Mutual Information)

Figure 7: Six uncertainty metrics plotted across prompting strategies and decoding temperatures. Shaded regions indicate standard deviation across folds. These metrics contribute to the overall average uncertainty shown in Figure 8.

scores, uncertainty values show greater variation. There is not single temperature that brings lowest uncertainty for all prompting settings.

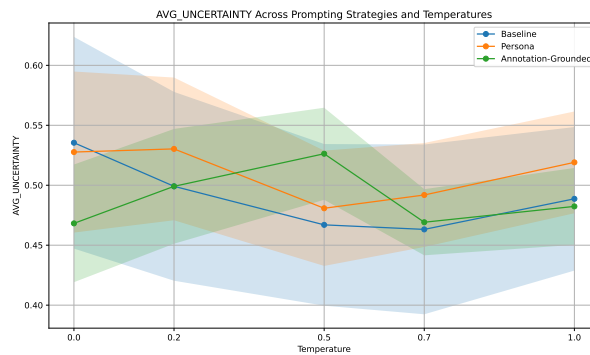


Figure 8: Average uncertainty across prompting strategies and decoding temperatures, summarizing all six uncertainty metrics. Shaded regions indicate standard deviation across folds.

C Prompt Templates

Full prompt templates leveraged in this work are presented in this section.

C.1 Baseline

The baseline prompt uses a generic annotator instruction without additional context as shown below.

Baseline Prompt Template

```
You are an annotator.
Rate the degree of hate speech in the text by generating a score 0, 1, or 2.
0 = Not Hate Speech
1 = Ambiguous/Uncertain
2 = Hate Speech

Provide a concise but complete justification INSIDE the JSON only (2-4 sentences).
Do NOT output any text outside the required block.

Text to evaluate (read-only):
<<<
{text}
>>>

OUTPUT REQUIREMENTS:
- Return ONLY one <output> block containing valid JSON.
- Include BOTH opening <output> and closing </output> tags exactly as shown.
- The JSON must include only two fields: "hate_speech_score" and "justification".
- Do NOT write anything before <output> or after </output>.

<output>
{
  "hate_speech_score": <generated_score>,
  "justification": <explain_your_reasoning_in_2_to_4_sentences>
}
</output>
```

C.2 Persona Prompting

The persona-based prompt uses demographic attributes of annotators to imitate their behavior on LLMs. The prompt template which includes persona background including gender, race, age and education level is presented below.

Persona-based prompt Template

```
You are an annotator with the following background:

Gender: {annotator_gender}
Race: {annotator_race}
Age: {annotator_age}
Education Level: {annotator_education}

Rate the degree of hate speech in the text by generating a score 0, 1, or 2.
0 = Not Hate Speech
1 = Ambiguous
2 = Hate Speech

Provide a concise but complete justification INSIDE the JSON only (2--4 sentences).
Do NOT output any text outside the required block.

Text to evaluate (read-only):
<<<
{text}
>>>

OUTPUT REQUIREMENTS:
- Return ONLY one <output> block containing valid JSON.
```

- Include BOTH opening <output> and closing </output> tags exactly as shown.
- The JSON must include only two fields: "hate_speech_score" and "justification".
- Do NOT write anything before <output> or after </output>.

```
<output>
{
  "hate_speech_score": <generated_score>,
  "justification": <explain_your_reasoning_in_2_to_4_sentences>
}
</output>
```

C.3 Annotation-Grounded Few-Shot Prompting

Annotation-grounded few-shot prompting uses example annotations from similar annotators on semantically similar texts. The prompting template is shown below.

Annotation-grounded Few-shot Prompt Template

You are an annotator.

You must NOT create, modify, or add any examples.

```
<example_text_and_annotation>
<example_text_and_annotation>
<example_text_and_annotation>
```

Under consideration of the given examples, rate the degree of hate speech in the text by generating a score 0, 1, or 2.

0 = Not Hate Speech

1 = Ambiguous

2 = Hate Speech

Provide a concise but complete justification INSIDE the JSON only (2--4 sentences). Do NOT output any text outside the required block.

Text to evaluate (read-only):

```
<<<
{text}
>>>
```

OUTPUT REQUIREMENTS:

- Return ONLY one <output> block containing valid JSON.
- Include BOTH opening <output> and closing </output> tags exactly as shown.
- The JSON must include only two fields: "hate_speech_score" and "justification".
- Do NOT write anything before <output> or after </output>.
- Do NOT create any examples.

```
<output>
{
  "hate_speech_score": <generated_score>,
  "justification": <explain_your_reasoning_in_2_to_4_sentences>
}
</output>
```

We present a single example below, including the input text, the annotator filters, and the corresponding retrieved few-shot examples, to illustrate how the context is constructed.

Text F*ck black people.
Annotator Filters Gender: Female Age: 25–34 Race: Black Education: College Grad. AA
Context Examples Text: F*ck a fake ass b*tch. <i>Hate Speech Score: 1</i> Text: F*cking f*got burn in hell. <i>Hate Speech Score: 2</i> Text: GET RID OF THOSE F*CKING MUSLIMS! the nasty ones that won't assimilate, r*pe women, beat children, threaten democracy, plan t*rror attacks, assault innocent secular people and force their disgusting religion on everyone else by way of violence. Assimilate OR GET THE F*CK OUT OR BE PREPARED TO DIE! <i>Hate Speech Score: 0</i>

Table 5: Example of input text, annotator filters, and context.

D Examples of Improvements with Advanced Prompting

This section presents examples that were predicted incorrectly with baseline prompting but correctly with either Persona or Annotation-Grounded prompting. For each dataset–model pair, three different examples are shown in Table 6.

Model	Text	GT	Baseline	Persona	Annotation-Grounded
<i>MHS</i>					
Meta	Yeah you’re right! Cutting off a dick doesn’t make her a woman. She’s a woman regardless of her genitalia.	0	1	0	0
LLaMA-3	Niggas will literally tell you one thing and then do the complete opposite	2	1	2	2
8B-Instruct	you will be killed for words like that in saudi arabia	0	2	2	0
Mixtral	One more reminder that Asian people can be very racist towards Black people. Please don’t let the “solidarity” confuse you. It’s uncomfortable to hear but it’s very true URL	0	1	1	0
8x7B-Instruct	Allah told me to tell you to go out and get yaself a strong muscley man to retain semen for on ramadan while eating top ramen fam	2	0	2	0
	@VanessaMerrell That is what i call haram	0	2	0	0
Mistral	If you ask any of the 24 Democrat Presidential Candidates the following question: “WHAT IS YOUR PLAN TO STOP ILLEGAL IMMIGRATION?”, what are you most likely to hear?	0	2	2	0
7B-Instruct	6 Jesus Quotes that will Offend Some—and Wake Up Others URL	0	2	2	0
	Looking to exchange nudes with anybody. M/F/etc Snap:tyroned396	0	2	2	0
Qwen3	For them hindi and hindu are inseparable. A hindu who doesn’t speak hindi is not completely hindu.	0	1	1	0
4B-Instruct	I am not racist, but they look like animals when they behave like this, it’s a fact not an opinion.	2	1	2	2
	Israel is not America you sell out. We don’t even have clean water in this country. Who made you the Nazi police of what Americans can say. We don’t like what you say so send Melania and her chain immigrant parents to go back because we don’t want them here.	2	1	2	2
Falcon3	“There are wonderful Muslims in my family. There are wonderful Muslims everywhere. But that doesn’t make Islam wonderful.” @QueenHaram, British-Pakistani #ExMuslim URL	0	2	0	0
7B-Instruct	As a white man growing up in eastern Tennessee I felt the same. Then I moved to Central north Carolina. The black women here are gorgeous across the socioeconomic spectrum and im definitely attracted.	0	2	0	0
	If he’s not lying then why is so much violence these days being caused by white American citizens?	0	2	2	0
<i>CREHate</i>					
Meta	if i had a dollar for every gender ... i 'd have \$ <number> and a pile of counterfeits .	2	1	0	2
LLaMA-3	So with the repeal of 377A, married men are going to church to have gay sex?	2	0	1	2
8B-Instruct	Africans actually have more of some things. Like infant mortality.	2	0	1	2
Mixtral	What’s the hardest part of a vegetable to eat? The wheelchair	2	0	2	0
8x7B-Instruct	Black people are nothing like monkeys because at least monkeys are well behaved	2	0	2	0
	What’s the difference between a women and a computer You can punch information into a computer	2	0	2	0
Mistral	Malay-Muslim people of r/Singapore is this a big issue for you guys? Not being sarcastic. I am genuinely wondering.	0	2	2	0
7B-Instruct	Today I got a joke email about adopting a fast-food delivery robot and that was more believable than the Trump presidency #aprilfoolsday	0	2	0	0
	King curry hut - Bracken ridge!	0	2	0	2
Qwen-3	Heard this from a 12 year old: “Whats the difference between black people and apples?” I like apples	2	0	1	2
4B-Instruct	A rich black man, a educated black man and a poor black man was walking down the street when they saw a drumstick on the ground, who picked it up? The poor black man as the other 2 don’t exist	2	1	2	0
	what ’s the meanest thing you can do to a blind person? give them a cheese grater to read.	2	0	2	2
Falcon3	People told me to stop inviting disabled people to my farm. Well, they’re right. When I buried them, the vegetables didn’t grow!	2	0	0	2
7B-Instruct	He’s right. I have to work with Boers in Australia and they deserve anything they get.	2	0	0	2
	How are people with mental disabilities similar to electronic products? Sometimes they are, “batteries not included”	2	0	0	2

Table 6: Qualitative examples from the MHS and CREHate datasets where persona prompting or annotation-grounded prompting corrects baseline prediction errors. Improved predictions matching the ground truth are shown in **bold**.

E Extended Results

This section provides detailed tables reporting correlation results, failure rates in LLM generations, F1 scores, and uncertainty metrics as additional and supplementary results.

E.1 Relationship Between Uncertainty and Annotator Disagreement

Table 7 presents the Pearson correlation between uncertainty and annotator disagreement scores. Correlations are generally weak and there are also inconsistencies in the direction of the correlation across different datasets and model combinations. Eight settings exhibit negative correlations, while two show positive correlations. This inconsistency across datasets suggests that the uncertainty signals produced by the models do not reliably capture patterns of human disagreement.

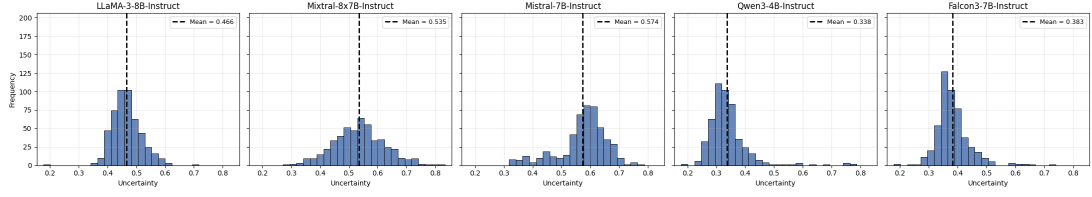
Dataset	Model	r	p-value
MHS	Meta-LLaMA-3-8B-Instruct	-0.066	0.142
	Mixtral-8x7B-Instruct	0.024	0.598
	Mistral-7B-Instruct	0.245	2.77e-08
	Qwen3-4B-Instruct	-0.091	0.0414
	Falcon3-7B-Instruct	-0.045	0.312
CREHate	Meta-LLaMA-3-8B-Instruct	-0.092	0.0393
	Mixtral-8x7B-Instruct	-0.165	0.000219
	Mistral-7B-Instruct	-0.231	1.68e-07
	Qwen3-4B-Instruct	-0.014	0.76
	Falcon3-7B-Instruct	-0.257	5.37e-09

Table 7: Pearson correlation between model uncertainty and annotator disagreement across two datasets. Statistically significant p -values ($p < 0.05$) are highlighted in light green. Stronger relationships ($|r| \geq 0.2$) are **bolded**.

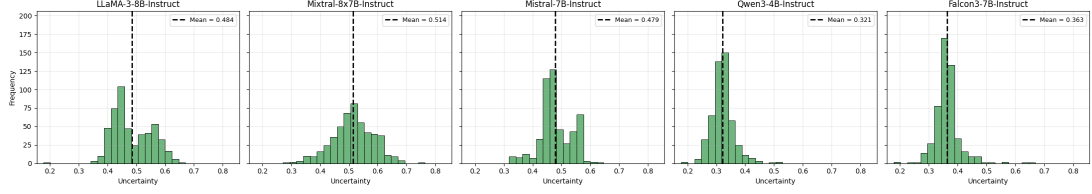
E.2 Uncertainty Score Distributions

In this section, we present the distributions of uncertainty scores aggregated across the six evaluated uncertainty quantification methods, together with their mean values (Figure 9 and Figure 10).

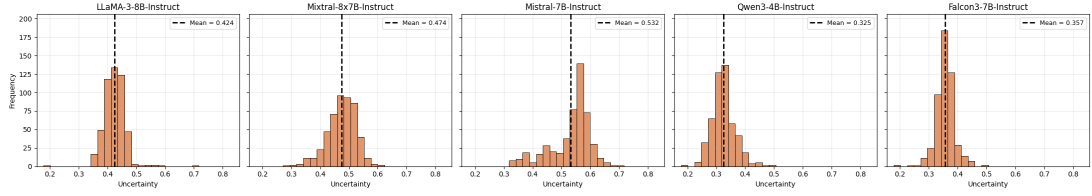
Baseline prompting consistently yields the most spread-out distributions across both datasets, reflecting high variability in model confidence across samples. Both persona and annotation-grounded prompting shift the distributions toward lower uncertainty with reduced spread. The results confirm that applying advanced prompting strategies to high-uncertainty samples generally reduces model uncertainty and consequently improves classification performance.



(a) Baseline prompting

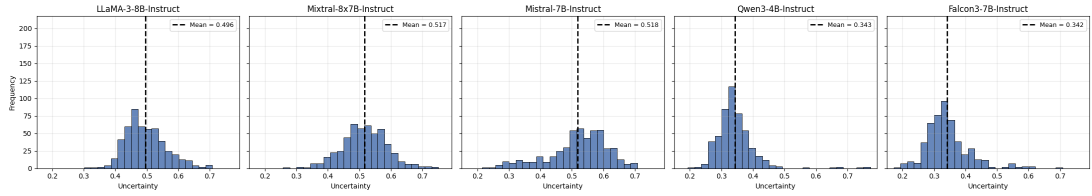


(b) Persona prompting

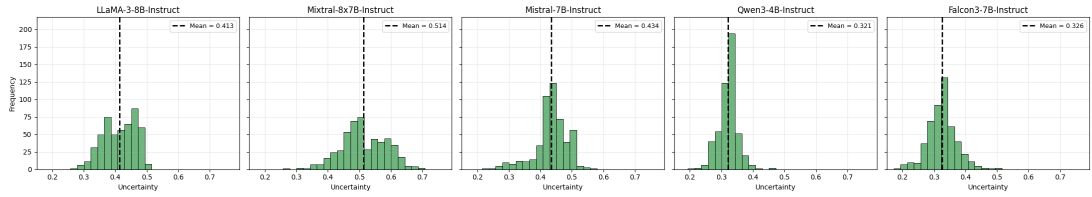


(c) Annotation-Grounded prompting

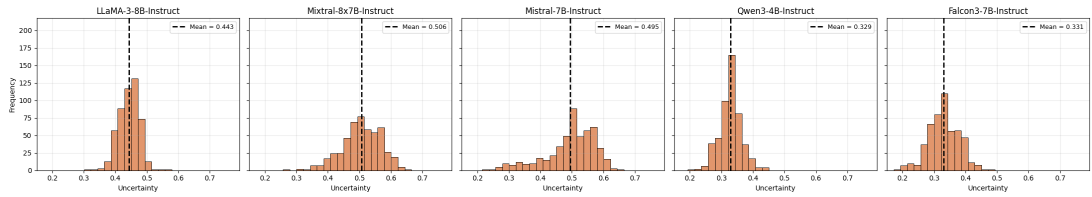
Figure 9: Distribution and mean values of average uncertainty scores across models for different prompting strategies for MHS dataset.



(a) Baseline prompting



(b) Persona prompting



(c) Annotation-Grounded prompting

Figure 10: Distribution and mean values of average uncertainty scores across models for different prompting strategies for CREHate dataset.

E.3 Highly Uncertain LLM Generation Counts

In this section, we report the number of samples identified as highly uncertain and subsequently processed with advanced prompting strategies in Table 8.

Overall, the counts remain in a comparable range across models and datasets. This indicates that the mean-based threshold provides a balanced and consistent selection criterion.

Model	MHS	CREHate
Meta-LLaMA-3-8B-Instruct	229	224
Mixtral-8x7B-Instruct-v0.1	252	250
Mistral-7B-Instruct-v0.3	297	288
Qwen3-4B-Instruct	196	207
Falcon3-7B-Instruct	211	206

Table 8: Number of uncertain LLM generations identified across models for the MHS and CREHate datasets.

E.4 Failure rates in LLM generations

Since some of the LLM outputs do not form the required JSON format, these sentences could not be parsed and the percentage of failure cases is shown in Table 9. For both datasets, Falcon3-7B-Instruct and Qwen3-4B-Instruct produce more failure cases.

Dataset	Total Generations	Model	Persona Failure (%)	Annotation-grounded Failure (%)
MHS	1689	Meta-LLaMA-3-8B-Instruct	0.0	5.3
		Mixtral-8x7B-Instruct	0.0	0.0
		Mistral-7B-Instruct	3.8	0.0
		Falcon3-7B-Instruct	6.0	14.9
		Qwen3-4B-Instruct	28.7	34.8
CREHate	2500	Meta-LLaMA-3-8B-Instruct	0.0	1.8
		Mixtral-8x7B-Instruct	0.0	0.0
		Mistral-7B-Instruct	2.6	0.0
		Falcon3-7B-Instruct	4.1	15.2
		Qwen3-4B-Instruct	22.4	37.4

Table 9: Failure rates of persona-based and annotation-grounded prompting strategies on MHS dataset and CREHate dataset. Failure rate denotes the percentage of invalid or unparseable generations (lower is better). Higher failure rates are highlighted in light red .

E.5 Uncertainty Scores

Table 10 presents uncertainty scores across different quantification metrics for MHS and CREHate datasets. Similar trends are observed across both datasets: Qwen3-4B-Instruct consistently achieves the lowest MPMI, MSP, and perplexity scores. For most models and settings, persona and annotation-grounded prompting yield lower uncertainty than baseline.

Table 11 reports the improvement percentages relative to baseline, where the magnitude of improvement varies across models and settings. In general, uncertainty is decreased across most settings; however, MCPMI represents a notable exception on CREHate, where it is consistently increased under advanced prompting. Overall both persona and annotation-grounded prompting improve some metrics while worsening others for the same model. This shows that uncertainty reduction is metric-dependent.

Standard deviation of uncertainty scores are shown in Table 12. The standard deviation is an important quality measure, as it reflects the variability of uncertainty scores across generated outputs and indicates how consistently a model assigns uncertainty to its predictions. Lower standard deviation values indicate that the model assigns similar uncertainty scores across different generations, while higher values suggest greater variability in uncertainty estimates. Both annotation-grounded and persona prompting make the uncertainty more consistent across samples compared to the baseline, indicating more stable model behavior in most settings. In addition, as shown in the relative improvement results in Table 13, Qwen3-4B

Model	Prompting	MCPMI	MPMI	MSP	MTE	Perplexity	Ptrue
<i>MHS</i>							
Meta LLaMA-3 8B-Instruct	Baseline	0.375 ± 0.008	0.247 ± 0.013	0.330 ± 0.006	0.443 ± 0.010	0.457 ± 0.010	0.418 ± 0.023
	Persona	0.320 ± 0.010	0.260 ± 0.004	0.374 ± 0.008	0.521 ± 0.010	0.517 ± 0.009	0.406 ± 0.017
	Annt-ground	0.358 ± 0.006	0.233 ± 0.008	0.353 ± 0.007	0.305 ± 0.007	0.318 ± 0.008	0.412 ± 0.014
Mixtral 8x7B Instruct	Baseline	0.615 ± 0.011	0.662 ± 0.007	0.355 ± 0.011	0.403 ± 0.013	0.391 ± 0.010	0.434 ± 0.022
	Persona	0.611 ± 0.017	0.669 ± 0.008	0.303 ± 0.011	0.370 ± 0.010	0.353 ± 0.009	0.409 ± 0.022
	Annt-ground	0.611 ± 0.016	0.453 ± 0.022	0.287 ± 0.009	0.366 ± 0.010	0.371 ± 0.012	0.351 ± 0.013
Mistral 7B Instruct	Baseline	0.520 ± 0.007	0.614 ± 0.009	0.327 ± 0.009	0.443 ± 0.014	0.425 ± 0.013	0.729 ± 0.014
	Persona	0.482 ± 0.010	0.402 ± 0.007	0.312 ± 0.010	0.264 ± 0.011	0.252 ± 0.011	0.678 ± 0.006
	Annt-ground	0.580 ± 0.010	0.381 ± 0.009	0.295 ± 0.006	0.412 ± 0.009	0.407 ± 0.006	0.700 ± 0.009
Qwen3 4B Instruct	Baseline	0.382 ± 0.016	0.201 ± 0.007	0.079 ± 0.009	0.254 ± 0.017	0.079 ± 0.009	0.387 ± 0.014
	Persona	0.400 ± 0.011	0.178 ± 0.007	0.041 ± 0.006	0.205 ± 0.012	0.041 ± 0.006	0.401 ± 0.015
	Annt-ground	0.385 ± 0.014	0.210 ± 0.012	0.048 ± 0.004	0.206 ± 0.012	0.046 ± 0.004	0.395 ± 0.014
Falcon3 7B Instruct	Baseline	0.597 ± 0.018	0.245 ± 0.008	0.248 ± 0.004	0.105 ± 0.002	0.128 ± 0.002	0.427 ± 0.021
	Persona	0.583 ± 0.020	0.215 ± 0.012	0.255 ± 0.008	0.110 ± 0.002	0.128 ± 0.001	0.331 ± 0.010
	Annt-ground	0.579 ± 0.024	0.244 ± 0.013	0.212 ± 0.006	0.090 ± 0.004	0.101 ± 0.005	0.360 ± 0.012
<i>CREHate</i>							
Meta LLaMA-3 8B-Instruct	Baseline	0.223 ± 0.005	0.164 ± 0.009	0.442 ± 0.023	0.591 ± 0.013	0.641 ± 0.016	0.414 ± 0.033
	Persona	0.248 ± 0.006	0.184 ± 0.010	0.348 ± 0.005	0.365 ± 0.025	0.389 ± 0.027	0.368 ± 0.014
	Annt-ground	0.278 ± 0.008	0.209 ± 0.010	0.408 ± 0.015	0.366 ± 0.026	0.412 ± 0.024	0.435 ± 0.030
Mixtral 8x7B Instruct	Baseline	0.583 ± 0.015	0.669 ± 0.012	0.283 ± 0.010	0.361 ± 0.012	0.392 ± 0.011	0.430 ± 0.017
	Persona	0.630 ± 0.015	0.652 ± 0.014	0.271 ± 0.008	0.356 ± 0.009	0.385 ± 0.011	0.410 ± 0.018
	Annt-ground	0.637 ± 0.012	0.687 ± 0.011	0.255 ± 0.007	0.332 ± 0.008	0.341 ± 0.008	0.405 ± 0.011
Mistral 7B Instruct	Baseline	0.304 ± 0.012	0.167 ± 0.006	0.421 ± 0.019	0.536 ± 0.005	0.523 ± 0.004	0.711 ± 0.018
	Persona	0.367 ± 0.011	0.166 ± 0.007	0.314 ± 0.019	0.290 ± 0.007	0.276 ± 0.008	0.668 ± 0.019
	Annt-ground	0.431 ± 0.015	0.229 ± 0.006	0.290 ± 0.012	0.442 ± 0.007	0.431 ± 0.007	0.689 ± 0.023
Qwen3 4B Instruct	Baseline	0.313 ± 0.006	0.199 ± 0.005	0.073 ± 0.004	0.264 ± 0.005	0.073 ± 0.004	0.484 ± 0.012
	Persona	0.350 ± 0.012	0.157 ± 0.001	0.035 ± 0.004	0.208 ± 0.006	0.035 ± 0.004	0.478 ± 0.020
	Annt-ground	0.367 ± 0.007	0.230 ± 0.005	0.041 ± 0.001	0.198 ± 0.002	0.041 ± 0.001	0.436 ± 0.017
Falcon3 7B Instruct	Baseline	0.517 ± 0.017	0.229 ± 0.015	0.191 ± 0.015	0.125 ± 0.009	0.121 ± 0.009	0.293 ± 0.006
	Persona	0.519 ± 0.007	0.206 ± 0.006	0.186 ± 0.006	0.095 ± 0.004	0.103 ± 0.003	0.264 ± 0.018
	Annt-ground	0.550 ± 0.014	0.241 ± 0.017	0.191 ± 0.006	0.099 ± 0.004	0.096 ± 0.003	0.237 ± 0.008

Table 10: Mean values of uncertainty metrics across prompting strategies and LLMs on the MHS and CREHate datasets (lower is better). Values are reported as mean ± standard deviation across folds. For each model and uncertainty metric, the best-performing prompting strategy across baseline, persona, and annotation-grounded is highlighted in **bold**. Light green shading indicates the best-performing model–strategy combination for a given uncertainty metric within each dataset.

and Falcon3-7B achieve greater improvements in standard deviation across nearly all metrics compared to other models.

Table 14 reports the standard deviations of average uncertainty scores. The annotation-grounded approach generally yields lower standard deviations across both datasets, achieving the lowest values in 5 out of 8 settings, while persona prompting performs better in the remaining 3. This is consistent with the uncertainty distributions shown in Figure 9 and Figure 10, where baseline prompting exhibits the highest variability in uncertainty scores.

Model	Prompting	MCPMI	MPMI	MSP	MTE	Perplexity	Ptrue
<i>MHS</i>							
Meta-LLaMA-3-8B-Instruct	Persona	+14.7	-5.3	-13.3	-17.6	-13.1	+2.9
Meta-LLaMA-3-8B-Instruct	Annt-ground	+4.5	+5.7	-7.0	+31.2	+30.4	+1.4
Mixtral-8x7B-Instruct	Persona	+0.7	-1.1	+14.6	+8.2	+9.7	+5.8
Mixtral-8x7B-Instruct	Annt-ground	+0.7	+31.6	+19.2	+9.2	+5.1	+19.1
Mistral-7B-Instruct	Persona	+7.3	+34.5	+4.6	+40.4	+40.7	+7.0
Mistral-7B-Instruct	Annt-ground	-11.5	+37.9	+9.8	+7.0	+4.2	+4.0
Qwen3-4B-Instruct	Persona	-4.7	+11.4	+48.1	+19.3	+48.1	-3.6
Qwen3-4B-Instruct	Annt-ground	-0.8	-4.5	+39.2	+18.9	+41.8	-2.1
Falcon3-7B-Instruct	Persona	+2.3	+12.2	-2.8	-4.8	+0.0	+22.5
Falcon3-7B-Instruct	Annt-ground	+3.0	+0.4	+14.5	+14.3	+21.1	+15.7
<i>CREHate</i>							
Meta-LLaMA-3-8B-Instruct	Persona	-11.2	-12.2	+21.3	+38.2	+39.3	+11.1
Meta-LLaMA-3-8B-Instruct	Annt-ground	-24.7	-27.4	+7.7	+38.1	+35.7	-5.1
Mixtral-8x7B-Instruct	Persona	-8.1	+2.5	+4.2	+1.4	+1.8	+4.7
Mixtral-8x7B-Instruct	Annt-ground	-9.3	-2.7	+9.9	+8.0	+13.0	+5.8
Mistral-7B-Instruct	Persona	-20.7	+0.6	+25.4	+45.9	+47.2	+6.0
Mistral-7B-Instruct	Annt-ground	-41.8	-37.1	+31.1	+17.5	+17.6	+3.1
Qwen3-4B-Instruct	Persona	-11.8	+21.1	+52.1	+21.2	+52.1	+1.2
Qwen3-4B-Instruct	Annt-ground	-17.3	-15.6	+43.8	+25.0	+43.8	+9.9
Falcon3-7B-Instruct	Persona	-0.4	+10.0	+2.6	+24.0	+14.9	+9.9
Falcon3-7B-Instruct	Annt-ground	-6.4	-5.2	+0.0	+20.8	+20.7	+19.1

Table 11: Relative improvement (%) in mean uncertainty scores compared to the baseline, computed as $(\text{Baseline} - \text{Advanced Prompting}) / \text{Baseline} \times 100$. Positive values indicate reduced uncertainty and are highlighted in light green .

Model	Prompting	MCPMI	MPMI	MSP	MTE	Perplexity	Ptrue
<i>MHS</i>							
Meta LLaMA-3 8B-Instruct	Baseline	0.113 ± 0.009	0.123 ± 0.014	0.121 ± 0.010	0.079 ± 0.008	0.089 ± 0.008	0.218 ± 0.014
	Persona	0.127 ± 0.012	0.108 ± 0.006	0.149 ± 0.007	0.137 ± 0.010	0.131 ± 0.012	0.171 ± 0.014
	Annt-ground	0.110 ± 0.009	0.104 ± 0.019	0.124 ± 0.007	0.154 ± 0.007	0.149 ± 0.005	0.169 ± 0.007
Mixtral 8x7B Instruct	Baseline	0.160 ± 0.013	0.122 ± 0.011	0.183 ± 0.015	0.183 ± 0.015	0.174 ± 0.014	0.206 ± 0.007
	Persona	0.165 ± 0.014	0.114 ± 0.014	0.146 ± 0.007	0.153 ± 0.008	0.144 ± 0.009	0.183 ± 0.013
	Annt-ground	0.157 ± 0.011	0.187 ± 0.006	0.125 ± 0.009	0.149 ± 0.009	0.155 ± 0.009	0.153 ± 0.011
Mistral 7B Instruct	Baseline	0.131 ± 0.005	0.220 ± 0.012	0.152 ± 0.013	0.134 ± 0.007	0.138 ± 0.007	0.167 ± 0.016
	Persona	0.145 ± 0.012	0.195 ± 0.005	0.132 ± 0.007	0.107 ± 0.014	0.103 ± 0.014	0.157 ± 0.011
	Annt-ground	0.121 ± 0.008	0.197 ± 0.003	0.119 ± 0.003	0.129 ± 0.005	0.139 ± 0.005	0.178 ± 0.014
Qwen3 4B Instruct	Baseline	0.135 ± 0.008	0.129 ± 0.020	0.143 ± 0.022	0.172 ± 0.011	0.143 ± 0.022	0.227 ± 0.012
	Persona	0.116 ± 0.009	0.060 ± 0.005	0.062 ± 0.009	0.099 ± 0.004	0.062 ± 0.009	0.196 ± 0.011
	Annt-ground	0.106 ± 0.009	0.080 ± 0.002	0.062 ± 0.004	0.092 ± 0.007	0.058 ± 0.004	0.189 ± 0.007
Falcon3 7B Instruct	Baseline	0.169 ± 0.003	0.130 ± 0.006	0.139 ± 0.010	0.094 ± 0.004	0.096 ± 0.004	0.217 ± 0.018
	Persona	0.144 ± 0.008	0.093 ± 0.020	0.134 ± 0.013	0.090 ± 0.016	0.082 ± 0.011	0.152 ± 0.015
	Annt-ground	0.136 ± 0.009	0.104 ± 0.008	0.092 ± 0.006	0.049 ± 0.007	0.047 ± 0.006	0.144 ± 0.017
<i>CREHate</i>							
Meta LLaMA-3 8B-Instruct	Baseline	0.101 ± 0.013	0.075 ± 0.019	0.181 ± 0.015	0.133 ± 0.006	0.127 ± 0.003	0.208 ± 0.018
	Persona	0.096 ± 0.016	0.074 ± 0.021	0.123 ± 0.006	0.199 ± 0.008	0.223 ± 0.008	0.139 ± 0.009
	Annt-ground	0.098 ± 0.018	0.091 ± 0.014	0.134 ± 0.009	0.199 ± 0.007	0.202 ± 0.008	0.173 ± 0.009
Mixtral 8x7B Instruct	Baseline	0.154 ± 0.015	0.122 ± 0.019	0.140 ± 0.015	0.155 ± 0.014	0.164 ± 0.016	0.211 ± 0.005
	Persona	0.123 ± 0.016	0.105 ± 0.025	0.119 ± 0.006	0.144 ± 0.006	0.152 ± 0.006	0.174 ± 0.007
	Annt-ground	0.119 ± 0.014	0.114 ± 0.022	0.103 ± 0.006	0.114 ± 0.009	0.109 ± 0.008	0.156 ± 0.010
Mistral 7B Instruct	Baseline	0.125 ± 0.006	0.066 ± 0.016	0.217 ± 0.011	0.192 ± 0.017	0.197 ± 0.014	0.184 ± 0.036
	Persona	0.096 ± 0.010	0.069 ± 0.017	0.130 ± 0.004	0.127 ± 0.010	0.123 ± 0.010	0.168 ± 0.034
	Annt-ground	0.100 ± 0.006	0.098 ± 0.005	0.114 ± 0.006	0.131 ± 0.016	0.136 ± 0.015	0.177 ± 0.033
Qwen3 4B Instruct	Baseline	0.129 ± 0.013	0.109 ± 0.021	0.123 ± 0.020	0.170 ± 0.010	0.123 ± 0.020	0.233 ± 0.010
	Persona	0.109 ± 0.008	0.043 ± 0.004	0.044 ± 0.008	0.092 ± 0.003	0.044 ± 0.008	0.179 ± 0.013
	Annt-ground	0.117 ± 0.010	0.094 ± 0.003	0.048 ± 0.003	0.086 ± 0.002	0.048 ± 0.003	0.178 ± 0.012
Falcon3 7B Instruct	Baseline	0.165 ± 0.014	0.168 ± 0.020	0.125 ± 0.013	0.128 ± 0.015	0.116 ± 0.013	0.241 ± 0.010
	Persona	0.136 ± 0.007	0.092 ± 0.007	0.099 ± 0.008	0.052 ± 0.011	0.056 ± 0.011	0.147 ± 0.018
	Annt-ground	0.128 ± 0.011	0.112 ± 0.013	0.091 ± 0.005	0.049 ± 0.005	0.046 ± 0.005	0.117 ± 0.022

Table 12: Standard deviation of uncertainty metrics across prompting strategies and LLMs on the MHS and CREHate datasets (lower is better). Values are reported as standard deviation ± fold-level standard deviation. For each model and uncertainty metric, the best-performing prompting strategy across baseline, persona, and annotation-grounded is highlighted in **bold**. Light green shading indicates the best-performing model–strategy combination for a given uncertainty metric within each dataset.

Model	Prompting	MCPMI	MPMI	MSP	MTE	Perplexity	Ptrue
<i>MHS</i>							
Meta-LLaMA-3-8B-Instruct	Persona	-12.4	+12.2	-23.1	-73.4	-47.2	+21.6
Meta-LLaMA-3-8B-Instruct	Annt-ground	+2.7	+15.4	-2.5	-94.9	-67.4	+22.5
Mixtral-8x7B-Instruct	Persona	-3.1	+6.6	+20.2	+16.4	+17.2	+11.2
Mixtral-8x7B-Instruct	Annt-ground	+1.9	-53.3	+31.7	+18.6	+10.9	+25.7
Mistral-7B-Instruct	Persona	-10.7	+11.4	+13.2	+20.1	+25.4	+6.0
Mistral-7B-Instruct	Annt-ground	+7.6	+10.5	+21.7	+3.7	-0.7	-6.6
Qwen3-4B-Instruct	Persona	+14.1	+53.5	+56.6	+42.4	+56.6	+13.7
Qwen3-4B-Instruct	Annt-ground	+21.5	+38.0	+56.6	+46.5	+59.4	+16.7
Falcon3-7B-Instruct	Persona	+14.8	+28.5	+3.6	+4.3	+14.6	+30.0
Falcon3-7B-Instruct	Annt-ground	+19.5	+20.0	+33.8	+47.9	+51.0	+33.6
<i>CREHate</i>							
Meta-LLaMA-3-8B-Instruct	Persona	+5.0	+1.3	+32.0	-49.6	-75.6	+33.2
Meta-LLaMA-3-8B-Instruct	Annt-ground	+3.0	-21.3	+26.0	-49.6	-59.1	+16.8
Mixtral-8x7B-Instruct	Persona	+20.1	+13.9	+15.0	+7.1	+7.3	+17.5
Mixtral-8x7B-Instruct	Annt-ground	+22.7	+6.6	+26.4	+26.5	+33.5	+26.1
Mistral-7B-Instruct	Persona	+23.2	-4.5	+40.1	+33.9	+37.6	+8.7
Mistral-7B-Instruct	Annt-ground	+20.0	-48.5	+47.5	+31.8	+31.0	+3.8
Qwen3-4B-Instruct	Persona	+15.5	+60.6	+64.2	+45.9	+64.2	+23.2
Qwen3-4B-Instruct	Annt-ground	+9.3	+13.8	+61.0	+49.4	+61.0	+23.6
Falcon3-7B-Instruct	Persona	+17.6	+45.2	+20.8	+59.4	+51.7	+39.0
Falcon3-7B-Instruct	Annt-ground	+22.4	+33.3	+27.2	+61.7	+60.3	+51.5

Table 13: Relative improvement (%) in standard deviation of uncertainty scores compared to the baseline, computed as $(\text{Baseline} - \text{Advanced Prompting}) / \text{Baseline} \times 100$. Positive values indicate reduced uncertainty variability and are highlighted in light green .

Dataset	Model	Baseline	Persona	Annotation-Grounded
MHS	Meta-LLaMA-3-8B-Instruct	0.050 $\pm$ 0.003	0.068 $\pm$ 0.005	0.036 $\pm$ 0.005
	Mixtral-8x7B-Instruct	0.088 $\pm$ 0.007	0.072 $\pm$ 0.003	0.049 $\pm$ 0.005
	Mistral-7B-Instruct	0.084 $\pm$ 0.003	0.056 $\pm$ 0.003	0.072 $\pm$ 0.003
	Qwen3-4B-Instruct	0.061 $\pm$ 0.011	0.025 $\pm$ 0.004	0.024 $\pm$ 0.002
	Falcon3-7B-Instruct	0.058 $\pm$ 0.001	0.043 $\pm$ 0.004	0.032 $\pm$ 0.002
CREHate	Meta-LLaMA-3-8B-Instruct	0.063 $\pm$ 0.005	0.052 $\pm$ 0.001	0.032 $\pm$ 0.003
	Mixtral-8x7B-Instruct	0.072 $\pm$ 0.009	0.072 $\pm$ 0.005	0.062 $\pm$ 0.006
	Mistral-7B-Instruct	0.094 $\pm$ 0.009	0.052 $\pm$ 0.007	0.080 $\pm$ 0.008
	Qwen3-4B-Instruct	0.061 $\pm$ 0.011	0.025 $\pm$ 0.004	0.024 $\pm$ 0.002
	Falcon3-7B-Instruct	0.069 $\pm$ 0.005	0.046 $\pm$ 0.005	0.048 $\pm$ 0.003

Table 14: Standard deviation of average uncertainty scores across prompting strategies on MHS and CREHate (lower is better). Values are reported as mean $\pm$ standard deviation across folds. Light green highlights denote the lowest standard deviation per model and dataset.

E.6 LLM Disagreement and Annotator Disagreement Coverage

Table 15 shows annotator disagreement coverage percentage across different models and datasets. Overall, annotation-grounded strategy achieves substantially higher coverage than persona prompting.

Considering individual models, Meta-LLaMA-3-8B-Instruct achieves lower coverage for both datasets, and Mixtral-8x7B-Instruct and Falcon3-7B-Instruct outperform the other models. Considering the individual dataset characteristics, CREHate has almost double of the annotator disagreement percentage compared to MHS in some cases and also has higher overall disagreement coverage.

Dataset	# Annotator Disagr.	Annotator Disagr. (%)	Model	Persona Coverage (%)	Annotation-grounded Coverage (%)
MHS	224	44.8	Meta-LLaMA-3-8B-Instruct	3.1	13.8
			Mixtral-8x7B-Instruct	8.0	39.7
			Mistral-7B-Instruct	1.8	16.1
			Falcon3-7B-Instruct	7.1	34.8
			Qwen3-4B-Instruct	9.4	19.2
CREHate	355	71.0	Meta-LLaMA-3-8B-Instruct	10.1	30.7
			Mixtral-8x7B-Instruct	12.7	40.8
			Mistral-7B-Instruct	4.5	25.4
			Falcon3-7B-Instruct	19.7	54.6
			Qwen3-4B-Instruct	22.0	34.9

Table 15: Coverage of annotator disagreements by persona and annotation-grounded prompting strategies on MHS dataset and CREHate dataset (higher is better). Higher coverages per dataset and model pairs highlighted light green .

E.7 Highly Uncertain Samples Results

In this section, we present the results for highly uncertain cases detected by our Uncertainty-aware Prompting Framework (Figure 2).

E.7.1 Classification Correctness

Table 16 presents the classification correctness metrics for only highly uncertain samples. Overall results are similar to the Table 3.

Model	Strategy	Accuracy	Precision	Recall	F1 Score
<i>MHS</i>					
Meta-Llama-3-8B-Instruct	Baseline	0.389 ± 0.043	0.821 ± 0.034	0.389 ± 0.043	0.389 ± 0.060
	Persona	0.389 ± 0.054	0.821 ± 0.035	0.389 ± 0.054	0.383 ± 0.076
	Annt-ground	0.476 ± 0.066	0.824 ± 0.017	0.476 ± 0.066	0.487 ± 0.080
Mixtral-8x7B-Instruct	Baseline	0.612 ± 0.059	0.830 ± 0.044	0.612 ± 0.059	0.645 ± 0.059
	Persona	0.600 ± 0.055	0.839 ± 0.041	0.600 ± 0.055	0.635 ± 0.063
	Annt-ground	0.722 ± 0.022	0.823 ± 0.035	0.722 ± 0.022	0.757 ± 0.024
Mistral-7B-Instruct	Baseline	0.296 ± 0.051	0.090 ± 0.030	0.296 ± 0.051	0.138 ± 0.041
	Persona	0.286 ± 0.070	0.211 ± 0.297	0.286 ± 0.070	0.149 ± 0.070
	Annt-ground	0.397 ± 0.046	0.693 ± 0.131	0.397 ± 0.046	0.336 ± 0.058
Qwen3-4B-Instruct	Baseline	0.403 ± 0.105	0.862 ± 0.021	0.403 ± 0.105	0.475 ± 0.130
	Persona	0.413 ± 0.089	0.850 ± 0.031	0.413 ± 0.089	0.465 ± 0.084
	Annt-ground	0.465 ± 0.053	0.827 ± 0.054	0.465 ± 0.053	0.552 ± 0.050
Falcon3-7B-Instruct	Baseline	0.464 ± 0.061	0.803 ± 0.100	0.464 ± 0.061	0.494 ± 0.101
	Persona	0.502 ± 0.066	0.798 ± 0.107	0.502 ± 0.066	0.529 ± 0.089
	Annt-ground	0.445 ± 0.138	0.758 ± 0.125	0.445 ± 0.138	0.492 ± 0.155
<i>CREHate</i>					
Meta-Llama-3-8B-Instruct	Baseline	0.522 ± 0.063	0.647 ± 0.054	0.522 ± 0.063	0.563 ± 0.053
	Persona	0.504 ± 0.075	0.670 ± 0.067	0.504 ± 0.075	0.547 ± 0.071
	Annt-ground	0.580 ± 0.066	0.625 ± 0.056	0.580 ± 0.066	0.566 ± 0.059
Mixtral-8x7B-Instruct	Baseline	0.588 ± 0.058	0.640 ± 0.057	0.588 ± 0.058	0.606 ± 0.056
	Persona	0.628 ± 0.048	0.702 ± 0.055	0.628 ± 0.048	0.659 ± 0.045
	Annt-ground	0.632 ± 0.041	0.675 ± 0.040	0.632 ± 0.041	0.643 ± 0.042
Mistral-7B-Instruct	Baseline	0.514 ± 0.089	0.278 ± 0.105	0.514 ± 0.089	0.358 ± 0.109
	Persona	0.507 ± 0.082	0.558 ± 0.278	0.507 ± 0.082	0.375 ± 0.109
	Annt-ground	0.545 ± 0.059	0.652 ± 0.113	0.545 ± 0.059	0.439 ± 0.074
Qwen3-4B-Instruct	Baseline	0.434 ± 0.086	0.713 ± 0.135	0.434 ± 0.086	0.531 ± 0.102
	Persona	0.498 ± 0.101	0.706 ± 0.113	0.498 ± 0.101	0.571 ± 0.104
	Annt-ground	0.497 ± 0.041	0.727 ± 0.060	0.497 ± 0.041	0.589 ± 0.046
Falcon3-7B-Instruct	Baseline	0.494 ± 0.166	0.616 ± 0.111	0.494 ± 0.166	0.539 ± 0.149
	Persona	0.582 ± 0.090	0.634 ± 0.073	0.582 ± 0.090	0.594 ± 0.080
	Annt-ground	0.495 ± 0.120	0.656 ± 0.098	0.495 ± 0.120	0.529 ± 0.116

Table 16: Classification performance across the MHS and CREHate datasets for highly uncertain cases. Values are reported as mean ± standard deviation across folds. For each model and evaluation metric, the best-performing prompting strategy across baseline, persona, and annotation-grounded is highlighted in **bold**. Light green shading indicates the best-performing model-strategy combination for a given metric within each dataset.

E.7.2 Uncertainty Scores

Table 17 and Table 19 report the mean uncertainty scores across individual metrics and averaged metrics for highly uncertain samples. As expected, uncertainty values are higher compared to the overall results, since these samples were specifically selected for being above the mean uncertainty threshold.

Table 18 and Table 20 report the corresponding standard deviations. Overall, the findings are consistent with the trends observed in the full dataset results.

Model	Prompting	MCPMI	MPMI	MSP	MTE	Perplexity	Ptrue
<i>MHS</i>							
Meta LLaMA-3 8B-Instruct	Baseline	0.369 ± 0.015	0.293 ± 0.019	0.386 ± 0.005	0.470 ± 0.006	0.490 ± 0.007	0.547 ± 0.032
	Persona	0.248 ± 0.013	0.320 ± 0.021	0.482 ± 0.022	0.639 ± 0.013	0.622 ± 0.014	0.520 ± 0.017
	Annt-grounded	0.330 ± 0.016	0.260 ± 0.015	0.437 ± 0.016	0.168 ± 0.017	0.188 ± 0.015	0.534 ± 0.012
Mixtral 8x7B Instruct	Baseline	0.547 ± 0.024	0.707 ± 0.015	0.470 ± 0.030	0.535 ± 0.011	0.517 ± 0.010	0.485 ± 0.020
	Persona	0.539 ± 0.014	0.722 ± 0.004	0.366 ± 0.013	0.468 ± 0.013	0.441 ± 0.011	0.435 ± 0.021
	Annt-grounded	0.538 ± 0.014	0.292 ± 0.002	0.335 ± 0.009	0.461 ± 0.008	0.477 ± 0.009	0.320 ± 0.015
Mistral 7B Instruct	Baseline	0.469 ± 0.014	0.711 ± 0.010	0.395 ± 0.024	0.522 ± 0.014	0.507 ± 0.015	0.794 ± 0.006
	Persona	0.405 ± 0.019	0.354 ± 0.012	0.370 ± 0.022	0.221 ± 0.023	0.215 ± 0.022	0.708 ± 0.021
	Annt-grounded	0.570 ± 0.006	0.318 ± 0.010	0.340 ± 0.010	0.470 ± 0.009	0.476 ± 0.011	0.745 ± 0.008
Qwen3 4B Instruct	Baseline	0.379 ± 0.021	0.268 ± 0.037	0.162 ± 0.037	0.357 ± 0.037	0.163 ± 0.037	0.428 ± 0.045
	Persona	0.425 ± 0.018	0.210 ± 0.011	0.065 ± 0.013	0.232 ± 0.016	0.065 ± 0.013	0.464 ± 0.024
	Annt-grounded	0.385 ± 0.007	0.291 ± 0.007	0.084 ± 0.008	0.233 ± 0.007	0.078 ± 0.007	0.449 ± 0.019
Falcon3 7B Instruct	Baseline	0.600 ± 0.020	0.292 ± 0.025	0.310 ± 0.026	0.139 ± 0.010	0.168 ± 0.011	0.526 ± 0.043
	Persona	0.567 ± 0.023	0.221 ± 0.011	0.324 ± 0.024	0.153 ± 0.014	0.168 ± 0.012	0.298 ± 0.015
	Annt-grounded	0.558 ± 0.018	0.290 ± 0.017	0.224 ± 0.014	0.104 ± 0.015	0.103 ± 0.013	0.366 ± 0.023
<i>CREHate</i>							
Meta LLaMA-3 8B-Instruct	Baseline	0.208 ± 0.008	0.173 ± 0.008	0.579 ± 0.023	0.672 ± 0.010	0.722 ± 0.014	0.516 ± 0.031
	Persona	0.263 ± 0.020	0.217 ± 0.007	0.370 ± 0.026	0.166 ± 0.008	0.158 ± 0.009	0.413 ± 0.016
	Annt-grounded	0.330 ± 0.005	0.273 ± 0.009	0.503 ± 0.018	0.170 ± 0.012	0.210 ± 0.015	0.563 ± 0.009
Mixtral 8x7B Instruct	Baseline	0.523 ± 0.024	0.707 ± 0.018	0.365 ± 0.009	0.464 ± 0.007	0.503 ± 0.005	0.480 ± 0.020
	Persona	0.616 ± 0.013	0.674 ± 0.014	0.339 ± 0.010	0.455 ± 0.011	0.490 ± 0.013	0.439 ± 0.020
	Annt-grounded	0.630 ± 0.015	0.743 ± 0.009	0.307 ± 0.010	0.407 ± 0.008	0.401 ± 0.009	0.429 ± 0.006
Mistral 7B Instruct	Baseline	0.237 ± 0.017	0.189 ± 0.007	0.552 ± 0.030	0.662 ± 0.013	0.652 ± 0.018	0.783 ± 0.014
	Persona	0.345 ± 0.011	0.187 ± 0.004	0.368 ± 0.012	0.235 ± 0.014	0.225 ± 0.013	0.707 ± 0.005
	Annt-grounded	0.457 ± 0.018	0.297 ± 0.008	0.326 ± 0.017	0.498 ± 0.009	0.494 ± 0.009	0.744 ± 0.012
Qwen3 4B Instruct	Baseline	0.299 ± 0.019	0.253 ± 0.011	0.144 ± 0.014	0.372 ± 0.018	0.144 ± 0.014	0.524 ± 0.015
	Persona	0.387 ± 0.007	0.152 ± 0.007	0.054 ± 0.005	0.236 ± 0.008	0.053 ± 0.005	0.510 ± 0.011
	Annt-grounded	0.429 ± 0.012	0.328 ± 0.007	0.066 ± 0.005	0.212 ± 0.008	0.066 ± 0.005	0.408 ± 0.028
Falcon3 7B Instruct	Baseline	0.512 ± 0.020	0.281 ± 0.032	0.255 ± 0.015	0.182 ± 0.030	0.177 ± 0.026	0.438 ± 0.032
	Persona	0.517 ± 0.017	0.224 ± 0.005	0.242 ± 0.021	0.109 ± 0.006	0.132 ± 0.006	0.367 ± 0.015
	Annt-grounded	0.592 ± 0.006	0.310 ± 0.007	0.253 ± 0.010	0.119 ± 0.006	0.116 ± 0.005	0.300 ± 0.011

Table 17: Mean values of uncertainty metrics across prompting strategies and LLMs on the MHS and CREHate datasets (lower is better) for highly uncertain cases. Values are reported as mean ± standard deviation across folds. For each model and uncertainty metric, the best-performing prompting strategy across baseline, persona, and annotation-grounded is highlighted in **bold**. Light green shading indicates the best-performing model–strategy combination for a given uncertainty metric within each dataset.

Model	Prompting	MCPMI	MPMI	MSP	MTE	Perplexity	Ptrue
<i>MHS</i>							
Meta LLaMA-3 8B-Instruct	Baseline	0.115 ± 0.020	0.130 ± 0.039	0.130 ± 0.010	0.081 ± 0.010	0.089 ± 0.006	0.224 ± 0.022
	Persona	0.106 ± 0.010	0.080 ± 0.015	0.130 ± 0.010	0.095 ± 0.014	0.098 ± 0.011	0.131 ± 0.009
	Annt-grounded	0.103 ± 0.016	0.101 ± 0.026	0.106 ± 0.004	0.104 ± 0.033	0.093 ± 0.026	0.109 ± 0.004
Mixtral 8x7B Instruct	Baseline	0.126 ± 0.013	0.100 ± 0.005	0.158 ± 0.016	0.143 ± 0.025	0.134 ± 0.020	0.207 ± 0.018
	Persona	0.136 ± 0.014	0.067 ± 0.008	0.141 ± 0.017	0.126 ± 0.016	0.124 ± 0.015	0.171 ± 0.007
	Annt-grounded	0.115 ± 0.014	0.039 ± 0.007	0.110 ± 0.011	0.122 ± 0.010	0.123 ± 0.015	0.092 ± 0.007
Mistral 7B Instruct	Baseline	0.114 ± 0.010	0.087 ± 0.010	0.143 ± 0.013	0.098 ± 0.011	0.100 ± 0.011	0.078 ± 0.008
	Persona	0.104 ± 0.022	0.083 ± 0.028	0.117 ± 0.016	0.091 ± 0.039	0.090 ± 0.038	0.087 ± 0.025
	Annt-grounded	0.123 ± 0.006	0.070 ± 0.010	0.108 ± 0.011	0.119 ± 0.014	0.123 ± 0.016	0.133 ± 0.006
Qwen3 4B Instruct	Baseline	0.154 ± 0.033	0.172 ± 0.050	0.192 ± 0.057	0.207 ± 0.034	0.192 ± 0.057	0.248 ± 0.008
	Persona	0.104 ± 0.013	0.063 ± 0.013	0.083 ± 0.015	0.096 ± 0.009	0.083 ± 0.015	0.155 ± 0.023
	Annt-grounded	0.078 ± 0.011	0.046 ± 0.003	0.075 ± 0.004	0.075 ± 0.004	0.070 ± 0.004	0.138 ± 0.009
Falcon3 7B Instruct	Baseline	0.188 ± 0.017	0.169 ± 0.025	0.161 ± 0.027	0.129 ± 0.027	0.128 ± 0.030	0.247 ± 0.018
	Persona	0.128 ± 0.021	0.116 ± 0.024	0.147 ± 0.018	0.118 ± 0.040	0.104 ± 0.032	0.139 ± 0.017
	Annt-grounded	0.108 ± 0.007	0.121 ± 0.012	0.085 ± 0.013	0.060 ± 0.019	0.056 ± 0.017	0.125 ± 0.020
<i>CREHate</i>							
Meta LLaMA-3 8B-Instruct	Baseline	0.095 ± 0.009	0.066 ± 0.007	0.156 ± 0.020	0.112 ± 0.009	0.107 ± 0.006	0.220 ± 0.013
	Persona	0.083 ± 0.009	0.045 ± 0.008	0.131 ± 0.014	0.047 ± 0.006	0.044 ± 0.006	0.088 ± 0.007
	Annt-grounded	0.059 ± 0.005	0.051 ± 0.007	0.091 ± 0.010	0.060 ± 0.008	0.063 ± 0.007	0.082 ± 0.009
Mixtral 8x7B Instruct	Baseline	0.141 ± 0.011	0.109 ± 0.011	0.136 ± 0.009	0.138 ± 0.013	0.142 ± 0.012	0.212 ± 0.025
	Persona	0.101 ± 0.015	0.084 ± 0.011	0.105 ± 0.008	0.118 ± 0.005	0.121 ± 0.006	0.140 ± 0.015
	Annt-grounded	0.092 ± 0.005	0.072 ± 0.008	0.088 ± 0.010	0.084 ± 0.010	0.084 ± 0.009	0.093 ± 0.004
Mistral 7B Instruct	Baseline	0.094 ± 0.024	0.043 ± 0.037	0.172 ± 0.017	0.115 ± 0.013	0.120 ± 0.017	0.086 ± 0.015
	Persona	0.087 ± 0.020	0.053 ± 0.035	0.109 ± 0.013	0.079 ± 0.033	0.078 ± 0.030	0.080 ± 0.010
	Annt-grounded	0.094 ± 0.007	0.046 ± 0.007	0.092 ± 0.007	0.097 ± 0.009	0.101 ± 0.009	0.089 ± 0.020
Qwen3 4B Instruct	Baseline	0.133 ± 0.028	0.147 ± 0.019	0.164 ± 0.024	0.189 ± 0.017	0.164 ± 0.024	0.252 ± 0.014
	Persona	0.069 ± 0.005	0.035 ± 0.005	0.050 ± 0.012	0.073 ± 0.013	0.050 ± 0.012	0.104 ± 0.019
	Annt-grounded	0.067 ± 0.009	0.037 ± 0.004	0.053 ± 0.009	0.061 ± 0.007	0.053 ± 0.009	0.098 ± 0.013
Falcon3 7B Instruct	Baseline	0.190 ± 0.014	0.223 ± 0.030	0.156 ± 0.028	0.178 ± 0.030	0.158 ± 0.027	0.287 ± 0.013
	Persona	0.115 ± 0.024	0.077 ± 0.010	0.105 ± 0.019	0.063 ± 0.015	0.065 ± 0.013	0.109 ± 0.012
	Annt-grounded	0.077 ± 0.006	0.095 ± 0.012	0.080 ± 0.010	0.055 ± 0.009	0.050 ± 0.007	0.061 ± 0.008

Table 18: Standard deviation of uncertainty metrics across prompting strategies and LLMs on the MHS and CREHate datasets (lower is better) for highly uncertain cases. Values are reported as standard deviation ± fold-level standard deviation. For each model and uncertainty metric, the best-performing prompting strategy across baseline, persona, and annotation-grounded is highlighted in **bold**. Light green shading indicates the best-performing model–strategy combination for a given uncertainty metric within each dataset.

Dataset	Model	Baseline	Persona	Annotation-Grounded
MHS	Meta-LLaMA-3-8B-Instruct	0.508 ± 0.005	0.547 ± 0.009	0.415 ± 0.006
	Mixtral-8x7B-Instruct	0.604 ± 0.000	0.562 ± 0.007	0.483 ± 0.005
	Mistral-7B-Instruct	0.626 ± 0.005	0.466 ± 0.008	0.556 ± 0.006
	Qwen3-4B-Instruct	0.393 ± 0.014	0.350 ± 0.005	0.359 ± 0.005
	Falcon3-7B-Instruct	0.430 ± 0.004	0.383 ± 0.006	0.369 ± 0.005
CREHate	Meta-LLaMA-3-8B-Instruct	0.553 ± 0.006	0.367 ± 0.004	0.434 ± 0.007
	Mixtral-8x7B-Instruct	0.573 ± 0.003	0.567 ± 0.003	0.553 ± 0.003
	Mistral-7B-Instruct	0.581 ± 0.007	0.436 ± 0.001	0.540 ± 0.004
	Qwen3-4B-Instruct	0.391 ± 0.006	0.340 ± 0.003	0.357 ± 0.004
	Falcon3-7B-Instruct	0.402 ± 0.013	0.362 ± 0.001	0.375 ± 0.004

Table 19: Average uncertainty across prompting strategies on MHS and CREHate (lower is better) for highly uncertain cases. Values are reported as mean ± standard deviation across folds. Light green highlights indicate the lowest uncertainty per model (row-wise).

Dataset	Model	Baseline	Persona	Annotation-Grounded
MHS	Meta-LLaMA-3-8B-Instruct	0.036 ± 0.005	0.042 ± 0.005	0.042 ± 0.011
	Mixtral-8x7B-Instruct	0.057 ± 0.008	0.054 ± 0.008	0.045 ± 0.004
	Mistral-7B-Instruct	0.038 ± 0.005	0.031 ± 0.008	0.059 ± 0.007
	Qwen3-4B-Instruct	0.077 ± 0.026	0.036 ± 0.006	0.035 ± 0.004
	Falcon3-7B-Instruct	0.054 ± 0.007	0.051 ± 0.009	0.034 ± 0.003
CREHate	Meta-LLaMA-3-8B-Instruct	0.046 ± 0.006	0.032 ± 0.003	0.031 ± 0.004
	Mixtral-8x7B-Instruct	0.045 ± 0.009	0.050 ± 0.004	0.036 ± 0.005
	Mistral-7B-Instruct	0.044 ± 0.003	0.026 ± 0.007	0.046 ± 0.005
	Qwen3-4B-Instruct	0.065 ± 0.010	0.026 ± 0.005	0.025 ± 0.004
	Falcon3-7B-Instruct	0.063 ± 0.011	0.038 ± 0.003	0.029 ± 0.005

Table 20: Standard deviation of average uncertainty across prompting strategies on MHS and CREHate (lower is better) for highly uncertain cases. Values are reported as mean ± standard deviation across folds. Light green highlights indicate the lowest uncertainty per model (row-wise).

F Uncertainty Quantification Techniques

Selected uncertainty quantification methods implemented by (Fadeeva et al., 2023) are described in detail in this section.

F.1 Maximum Sequence Probability

Maximum Sequence Probability (MSP) quantifies model uncertainty by leveraging the probability of the most likely generated sequence. Formally, the probability of generating a sequence $\mathbf{y} = (y_1, \dots, y_L)$ given an input $\mathbf{x}$ and model parameters θ is defined as

$$P(\mathbf{y} \mid \mathbf{x}, \theta) = \prod_{l=1}^L P(y_l \mid y_{<l}, \mathbf{x}, \theta), \quad (1)$$

where $y_{<l}$ denotes all tokens generated before position l .

Based on this, MSP is constructed as

$$\text{MSP}(\mathbf{y} \mid \mathbf{x}, \theta) = 1 - P(\mathbf{y} \mid \mathbf{x}, \theta). \quad (2)$$

Lower MSP indicates higher uncertainty, hence means that model is prone to hallucinations.

MSP is simple to implement and can be computed efficiently. However, it is inherently biased by the length of the input sequence, as it relies on the joint probability of all tokens; consequently, longer sequences tend to receive lower joint probabilities.

F.2 Perplexity

Perplexity is computed from sequence probabilities in a manner similar to MSP discussed in Section F.1. Unlike MSP, it introduces a length normalization step that mitigates the bias arising from longer sequences receiving lower joint probabilities.

Formally, perplexity is defined as the exponentiated average negative log-likelihood of the tokens in the output sequence where $\log P(\mathbf{y} \mid \mathbf{x}, \theta)$ is obtained from the sequence probability defined in Equation 1.

$$\text{PPL}(\mathbf{y}, \mathbf{x}; \theta) = \exp \left\{ -\frac{1}{L} \log P(\mathbf{y} \mid \mathbf{x}, \theta) \right\}, \quad (3)$$

where L denotes the sequence length.

F.3 Mean Token Entropy

Mean Token Entropy (MTE) measures the average uncertainty of the model's predictions at each decoding step, regardless of the token that is actually generated. Entropy quantifies how spread out the probability distribution is and therefore reflects the level of uncertainty. It captures the degree of confidence the model has when predicting the next token on average. It is shown as $\mathcal{H}_T$ below.

$$\mathcal{H}_T(\mathbf{y}, \mathbf{x}; \theta) = \frac{1}{L} \sum_{l=1}^L \mathcal{H}(y_l | y_{<l}, \mathbf{x}, \theta), \quad (4)$$

where $\mathcal{H}(y_l | y_{<l}, \mathbf{x}, \theta)$ is the Shannon entropy (Shannon and Weaver, 1949) of the model's predictive token distribution at step l .

In general, Shannon entropy for a discrete distribution p over outcomes $\{x_i\}$ is

$$H(p) = - \sum_i p(x_i) \log p(x_i) \quad (5)$$

Combining (5) with $p(v) = P(v | y_{<l}, \mathbf{x}, \theta)$ over the vocabulary V yields

$$\mathcal{H}(y_l | y_{<l}, \mathbf{x}, \theta) = - \sum_{v \in V} P(v | y_{<l}, \mathbf{x}, \theta) \log P(v | y_{<l}, \mathbf{x}, \theta), \quad (6)$$

where V denotes the vocabulary.

F.4 Mean Pointwise Mutual Information

Mean Pointwise Mutual Information (MPMI) (Takayama and Arase, 2019) measures how effectively a model considers the given input $\mathbf{x}$ when generating each token in its output. In simpler terms, it shows how well the model's generated tokens are actually connected to or influenced by the original input.

A language model M without access to the input assigns probabilities to the next token. If these assigned probabilities are similar to the ones assigned by a language model N that can access to the inputs, then the language model N is not effectively using the provided instruction. Formally, MPMI is captured by the formula

$$\text{MPMI}(\mathbf{y}, \mathbf{x}; \theta) = \frac{1}{L} \sum_{l=1}^L \log \frac{P(y_l | y_{<l}, \theta)}{P(y_l | y_{<l}, \mathbf{x}, \theta)}, \quad (7)$$

where

- the numerator $P(y_l | y_{<l}, \theta)$ represents the probability of generating token y_l , given only its preceding context $y_{<l}$ under the baseline language model
- the denominator $P(y_l | y_{<l}, \mathbf{x}, \theta)$ corresponds to the probability of the same token when the model is additionally conditioned on the task input $\mathbf{x}$

F.5 Mean Conditional Pointwise Mutual Information

Mean Pointwise Mutual Information (MPMI) explained in Section F.4 is extended to Mean Conditional Pointwise Mutual Information (MCPMI) (van der Poel et al., 2022) by adding an entropy-based threshold that gates the contribution of the unconditional probabilities, defined as

$$P(y_l | y_{<l}, \theta) \quad (8)$$

i.e., the probability of the next token y_l is given only the prefix $y_{<l}$ without conditioning on the external input $\mathbf{x}$.

By selecting situations when input $\mathbf{x}$ has a large impact on the result, the gating enables MCPMI to be more selective and robust.

For a target sequence $y = (y_1, \dots, y_L)$ and conditioning input x , the MPMI is defined in Equation 7 and can be rewritten as

$$\text{MPMI}(y, x; \theta) = \frac{1}{L} \sum_{l=1}^L \log P(y_l | y_{<l}, \theta) - \frac{1}{L} \sum_{l=1}^L \log P(y_l | y_{<l}, x, \theta) \quad (9)$$

Indicator variable $\mathbf{1}_l$ gates the contribution of the unconditional probability depending on the entropy of the conditional distribution

$$\mathbf{1}_l = \begin{cases} 1, & \text{if } \mathcal{H}(y_l | y_{<l}, x, \theta) \geq \tau \\ 0, & \text{otherwise} \end{cases} \quad (10)$$

where $\mathcal{H}(y_l | y_{<l}, x, \theta)$ denotes the entropy of the conditional distribution, τ is the entropy threshold, and $\mathbf{1}_l \in \{0, 1\}$ indicates whether the gate is open ($\mathbf{1}_l = 1$) or closed ($\mathbf{1}_l = 0$). Then, MCPMI is formulated as

$$\text{MCPMI}(y, x; \theta) = -\frac{1}{L} \sum_{l=1}^L \log P(y_l | y_{<l}, x, \theta) + \frac{\lambda}{L} \sum_{l=1}^L \mathbf{1}_l \log P(y_l | y_{<l}, \theta), \quad (11)$$

where $\lambda > 0$ is a tunable parameter controlling the influence of the unconditional distribution.

F.6 PTrue

In PTrue (Kadavath et al., 2022), the model evaluates the correctness of its own generation by using a verification prompt. Given an input question q and a model-generated answer a , the following prompt is used:

Question: q Possible answer: a
 Is the possible answer: (A) True (B) False
 The possible answer is:

After getting the evaluation response, log-probabilities of "True" token is calculated with minus sign.

G Usage of AI Assistants

In this work, we used ChatGPT for writing improvements and generating icons used in visual content.