Unveiling the Capabilities of Large Language Models in Detecting Offensive Language with Annotation Disagreement

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have become essential for offensive language detection, yet their ability to handle annotation disagreement remains underexplored. Disagreement samples, which arise from subjective interpretations, pose a unique challenge due to their ambiguous nature. Understanding how LLMs process these cases, particularly their confidence levels, can offer insight into their alignment with human annotators. This study systematically evaluates the performance of multiple LLMs in detecting offensive language at varying levels of annotation agreement. We analyze binary classification accuracy, examine the relationship between model confidence and human disagreement, and explore how disagreement samples influence model decision-making during few-shot learning and instruction fine-tuning. Our findings reveal that LLMs struggle with low-agreement samples, often exhibiting overconfidence in these ambiguous cases. However, utilizing disagreement samples in training improves both detection accuracy and model alignment with human judgment. These insights provide a foundation for enhancing LLM-based offensive language detection in real-world moderation tasks.

Disclaimer: The paper contains content that may be profane, vulgar, or offensive.

1 Introduction

Motivation. A fundamental challenge in offensive language detection is annotation disagreement—cases where human annotators provide conflicting labels for the same text. Disagreement arises due to differences in individual perception, cultural context, and linguistic ambiguity, making offensive language detection inherently subjective (Aroyo et al., 2019; Basile, 2020; Uma et al., 2021b). However, prior research predominantly treats this task as a binary classification problem, assuming consensus among annotators and failing to account for the inherent subjectivity in offensive language perception.

While large language models (LLMs) have been extensively applied to offensive language detection (Kumar et al., 2024; Huang et al., 2023), existing studies primarily evaluate their performance on datasets with binary labels, overlooking their ability to handle cases where annotators disagree. This oversimplification limits our understanding of how well LLMs align with human judgment in ambiguous cases. Moreover, models may exhibit overconfidence in cases where human annotators themselves are uncertain, raising concerns about their reliability for real-world moderation.

Research Objectives. To bridge this gap, we systematically evaluate LLMs' ability to process disagreement samples, analyzing both classification accuracy and model confidence. Through this, we seek to determine whether LLMs can effectively navigate subjective offensive language judgments and align with human reasoning. An ideal model should express high confidence for unanimously labeled cases and lower confidence for ambiguous samples, reflecting their inherent uncertainty (Weerasooriya et al., 2023; Baumler et al., 2023; Uma et al., 2021a; Leonardelli et al., 2023). Our study provides insights into whether LLMs capture these nuances or exhibit overconfidence in disagreement cases, which could undermine their trustworthiness in content moderation.

This paper systematically investigates how LLMs handle annotation disagreement in offensive language detection. Specifically, we address the following research questions: (**RQ1**) To what extent can LLMs accurately detect offensive language in cases of human annotation disagreement? (**RQ2**) How do disagreement samples shape LLM learning and influence decision-making?

To answer RQ1, we evaluate multiple open-source and closed-source LLMs in a zero-shot set-

Agreement Degree	Example	Hard Label	Soft Label
N^{++}	Stand for something or else fall for anything. #BlackLivesMatter Hello world! What a great day to be alive #Trump2020 #MAGA	0	0
N^{+}	Come on man! Lock'em up!!! #maga Not the first time. You all misspelled #blacklivesmatter. Speak up! @user	0	0.2
N^{0}	Set fire to Fox News (metaphorically) @user is outing #BLACK-LIVES-MATTER as a cult! HE IS CORRECT!	0	0.4
O^{0}	#DISGUSTING #Democrats terrorize old folks just before #elections2020 I love this shit! #BlackLivesMatter	1	0.6
O^{+}	@user You're a bumbling fool #elections2020 Elections 2020: Red Rapist v. Blue Racist	1	0.8
O^{++}	Y'all trending about kpop stans instead of #BlackLivesMatter big fack you Crazy idiots. This is batshit bullshit. #elections2020	1	1

Table 1: Eamples with different degrees of annotation agreement, where N denotes non-offensive language and O denotes offensive language. The superscripts ++, +, and 0 represent unanimous, medium, and low agreement, respectively. These samples are collected from the MD-Agreement dataset (Leonardelli et al., 2021).

ting, analyzing both classification accuracy and the relationship between annotation agreement and model confidence. For RQ2, we examine the impact of disagreement samples in few-shot learning and instruction fine-tuning, assessing how different agreement levels affect model performance.

Contributions. We summarize the contributions of this paper as follows:

- We provide the first systematic evaluation of LLMs' performance in offensive language detection under annotation disagreement, revealing key insights into model reliability and human-AI alignment.
- We conduct an extensive empirical study on LLMs' handling of disagreement cases, examining models' performance, confidence, and alignment with human judgment.
- We analyze the impact of training on disagreement samples, demonstrating how few-shot learning and instruction fine-tuning on these samples influence LLM decision-making in offensive language detection.

2 Preliminary

2.1 Dataset

Since annotation disagreements can stem from both intrinsic linguistic ambiguity and labeling error, selecting an appropriate benchmark dataset requires meeting two key criteria: (1) high annotation quality to ensure reliability, and (2) open access to unaggregated annotations to facilitate fine-grained analysis. To ensure a robust evaluation, we employ the MD-Agreement dataset (Leonardelli et al., 2021), a high-quality corpus for offensive language detection. It contains 10,753 tweets, each labeled by five trained human annotators, ensuring a reliable annotation process.

The dataset provides both hard labels (majority-voted labels) and soft labels, which indicate the level of agreement among annotators. The soft labels are categorized into three levels:

- *Unanimous agreement* (A^{++}): All five annotators agree on the label.
- *Mild agreement* (A^{+}): Four out of five annotators agree.
- *Weak agreement* (A^{0}): Only three annotators agree, while two disagree.

Each sample in the dataset is also classified as either *offensive* or *non-offensive*, following the same agreement-level framework:

- **Offensive samples** (O^{++}, O^{+}, O^{0}): Instances labeled as offensive, where the agreement level corresponds to the unanimous, mild, or weak agreement, respectively.
- **Non-offensive samples** (N^{++}, N^{+}, N^{0}): Instances labeled as non-offensive, with the same agreement-level distinctions.

Thus, the agreement notation ($++$, $+$, 0) applies uniformly across both offensive and non-offensive categories, ensuring consistency in the dataset's annotation schema. To facilitate subsequent research, we convert the soft labels into floating-point numbers in the range [0, 1] by averaging the hard labels from five annotators (offensive as 1, non-offensive as 0) for each sample.

Split	A^{++}	A^{+}	A	N^{++}	N^{+}	N	O	O^{+}	O^{++}	N-Off.	Off.	Total
Train	2,778	1,930	1,884	2,303	1,295	1,032	852	635	475	4,630	1,962	6,592
Dev	464	317	322	346	199	171	151	118	118	1,103	387	1,103
Test	1,292	909	803	1,020	549	470	386	360	272	3,004	1,018	3,057
MD-Agreement	4,535	3,156	3,062	3,669	2,043	1,673	1,389	1,113	866	7,385	3,368	10,753

Table 2: Statistics of the MD-Agreement dataset, where N denotes non-offensive language and O denotes offensive language. The superscripts ++, +, and 0 represent unanimous, medium, and low agreement, respectively.

Examples of samples across different agreement levels are provided in Table 1, and the overall dataset statistics are presented in Table 2. The reliability of MD-Agreement's annotations has been independently validated in prior studies (Sandri et al., 2023), confirming that the disagreement samples are caused by their inherent ambiguity, rather than labeling errors. MD-Agreement serves as the official corpus for SemEval 2023 Task 11 (Leonardelli et al., 2023) and has been widely utilized by researchers in the field (Deng et al., 2023; Mokhberian et al., 2024). Further dataset details are provided in Appendix A.1.

2.2 Models

To ensure a comprehensive evaluation, we include both closed-source and open-source LLMs, covering a range of architectures and parameter sizes. For closed-source models, we evaluate widely used proprietary LLMs, including GPT-3.5, GPT-4, GPT-4o, GPT-o1, Claude-3.5, and Gemini-1.5. We also evaluate three families of open-source LLMs at different scales: LLaMa-3 (8B, 70B), Qwen-2.5 (7B, 72B), and Mixtral (8x7B, 8x22B). Further details on the model versions are provided in Appendix A.4.

3 RQ1: Evaluating LLMs on Offensive Language with Annotation Disagreement

In this section, we evaluate the ability of LLMs to detect offensive language in a zero-shot setting. We focus on two key aspects: (1) binary classification accuracy, assessing how effectively models distinguish offensive from non-offensive language across varying annotation agreement levels, and (2) model confidence, analyzing whether LLMs exhibit appropriate uncertainty in ambiguous cases. These aspects are essential for determining whether LLMs can reliably perform offensive language detection in real-world scenarios, where human annotators often disagree.

3.1 Evaluation of Binary Classification Performance

We assess binary classification performance by evaluating LLMs in a zero-shot setting without additional fine-tuning. To ensure deterministic predictions, we set the temperature coefficient of the LLMs to 0, forcing the model to select the most probable category. We use accuracy and F1 score as evaluation metrics to measure classification performance. The prompt template used for offensive language detection is provided in Appendix A.3. LLM outputs are converted into hard predictions, where 1 indicates offensive and 0 indicates non-offensive. We utilize all the samples from MD-Agreement for a comprehensive evaluation. The classification results are presented in Table 3. Based on the results, we observe the following key findings:

(1) LLMs achieve high accuracy for unanimous agreement (A^{++}) samples. In the zero-shot setting, LLMs consistently accurately classify unanimously agreed-upon (A^{++}) samples, achieving 88.28% accuracy for closed-source models and 86.07% for open-source models. Notably, LLaMa3-70B now performs comparably to proprietary models. These results suggest that LLMs perform well on clear-cut cases, driven by their background knowledge and reasoning capabilities.

(2) LLM performance declines sharply for ambiguous cases. As annotation agreement decreases, LLMs struggle to classify offensive language consistently. GPT-4o's F1 score drops from 85.24% on A^{++} samples to 74.6% on A^{+} and 57.06% on A^{0}. Similarly, all models score below 65% on A^{0} samples. This highlights LLMs' inability to resolve subjective cases in the real world, where human disagreement often stems from cultural, contextual, or linguistic nuances that models fail to capture.

(3) Larger models improve accuracy but do not resolve annotation disagreement. While larger models generally perform better, their improvement shrinks for ambiguous cases. For example,

Model	Overall		A^{++}		A^{+}		A^{o}	
	Acc. ↑	F1 ↑	Acc. ↑	F1 ↑	Acc. ↑	F1 ↑	Acc. ↑	F1 ↑
Closed-Source Large Language Models (CS-LLMs)								
GPT-o1	78.35	69.03	91.95	81.29	77.50	72.03	59.08	58.63
GPT-4o	**80.36**	70.33	**93.96**	**85.24**	**80.67**	**74.60**	59.90	57.06
GPT-4	74.18	69.07	88.64	76.75	70.12	68.91	56.96	**64.63**
GPT-3.5	67.07	63.45	78.99	64.02	62.39	63.28	54.25	63.18
Claude-3.5	78.56	70.93	92.59	83.13	76.39	72.02	**60.03**	62.61
Gemini-1.5	69.50	66.07	83.53	69.70	64.48	65.73	53.89	64.07
Avg. of CS-LLM	74.67	68.15	88.28	76.69	71.93	69.43	57.35	61.70
Open-Source Large Language Models (OS-LLMs)								
LLaMa3-70B	76.93	**71.06**	91.40	81.36	74.46	71.96	58.03	64.37
LLaMa3-8B	71.82	65.31	85.56	70.72	68.22	66.06	55.19	61.26
Qwen2.5-72B	72.08	66.86	84.74	70.92	68.41	67.36	57.12	63.76
Qwen2.5-7B	71.10	67.14	85.34	72.02	66.92	67.25	54.31	64.06
Mixtral-8x22B	73.46	67.82	87.12	74.27	69.93	68.21	56.86	63.44
Mixtral-8x7B	70.57	65.59	82.27	67.58	67.14	66.32	56.76	63.63
Avg. of OS-LLM	72.66	67.30	86.07	72.81	69.18	67.86	56.38	63.42

Table 3: Binary classification performance of LLMs on the MD-Agreement dataset and its three subsets A^{++}, A^{+}, and A^{0}. *Avg. of CS-LLM* and *OS-LLM* respectively denote the average performance of the close-source and open-source LLMs. Results show the accuracy (*Acc.*) and F_1 in percentage (%). The **bold** and underline scores respectively represent the optimal and suboptimal values.

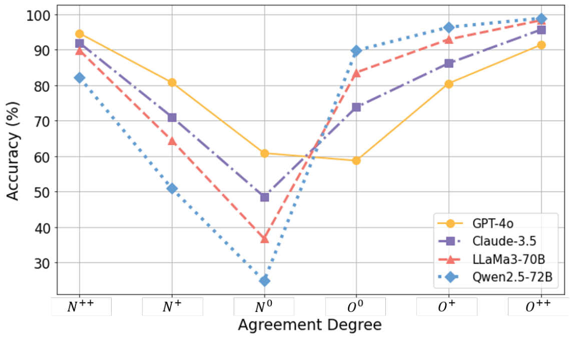


Figure 1: Accuracy of LLMs on detecting offensive and non-offensive language with different degrees of annotation agreement.

LLaMa3-70B outperforms LLaMa3-8B by 10.64% on A^{++} samples but only by 3.11% on A^{0}. Similarly, Mixtral and Qwen2.5 show no substantial gain in detecting disagreement samples despite increased parameters. Model scaling alone does not resolve ambiguity, suggesting that larger models lack the nuanced human reasoning required to navigate subjective cases. Alternative training strategies, such as human-in-the-loop approaches or fine-tuning on disagreement samples, may be necessary.

(4) LLMs are biased toward classifying uncertain cases as offensive. We evaluate the accuracy for offensive and non-offensive language across different agreement levels, as shown in Figure 1. We observed that across all agreement levels, LLMs demonstrate higher accuracy in identifying offensive language than non-offensive language. In particular, for low-agreement non-offensive samples (N^{0}), accuracy drops to 45.77%, indicating a strong tendency to misclassify ambiguous content as offensive. This over-sensitivity could lead to false positives in automated moderation systems, increasing the risk of justified content removal and restricting legitimate speech.

3.2 Evaluation of Relationship between Agreement Degree and LLM Confidence

We analyze how well LLM confidence aligns with human annotation agreement, as a well-calibrated model should exhibit high confidence for clear cases and lower confidence for ambiguous cases. If LLMs assign high confidence to disagreement samples, this may indicate overconfidence, limiting their ability to reflect human-like uncertainty. To evaluate this, we apply the self-consistency method (Chen and Mueller, 2024; Wang et al., 2023b), which resamples model outputs under varying temperature settings to estimate confidence.

To measure confidence, we evaluate models under five temperature settings: 0, 0.25, 0.5, 0.75, and 1. Higher temperatures introduce more randomness in predictions, helping assess the model's certainty across varying conditions. The final confidence score is computed by averaging the hard predictions across these temperature settings.

We use Mean Squared Error (MSE) to measure

Model	Overall		A^{++}		A^{+}		A^{o}	
	MSE ↓	ρ ↑	MSE ↓	ρ ↑	MSE ↓	ρ ↑	MSE ↓	ρ ↑
Closed-Source Large Language Models (CS-LLMs)								
GPT-4o	**0.1138**	0.6535	**0.0514**	**0.8098**	**0.1268**	**0.6298**	**0.1928**	0.2332
GPT-4	0.1716	0.6819	0.1131	0.7175	0.2064	0.5323	0.2224	**0.2478**
GPT-3.5	0.2163	0.5889	0.1878	0.6021	0.2430	0.4236	0.2309	0.1820
Claude-3.5	0.1306	0.6780	0.0657	0.7590	0.1544	0.5818	0.2022	0.2379
Gemini-1.5	0.2137	0.6305	0.1638	0.6970	0.2505	0.4517	0.2498	0.1877
Avg. of CS-LLMs	0.1692	0.6466	0.1164	0.7171	0.1962	0.5238	0.2196	0.2177
Open-Source Large Language Models (OS-LLMs)								
LLaMa3-70B	0.1400	**0.6990**	0.0753	0.7634	0.1680	0.5856	0.2072	0.2369
LLaMa3-8B	0.1803	0.5912	0.1316	0.6533	0.2068	0.4572	0.2251	0.1667
Qwen2.5-72B	0.1909	0.6588	0.1380	0.6817	0.2235	0.5001	0.2359	0.2119
Qwen2.5-7B	0.1962	0.6024	0.1480	0.6638	0.2237	0.4756	0.2393	0.2056
Mixtral-8x22B	0.1810	0.6287	0.1251	0.6858	0.2112	0.4944	0.2326	0.2107
Mixtral-8x7B	0.1978	0.5921	0.1578	0.6267	0.2218	0.4709	0.2323	0.2132
Avg. of OS-LLMs	0.1810	0.6287	0.1293	0.6791	0.2092	0.4973	0.2287	0.2075

Table 4: Estimation of relationship between annotators and LLMs on MD-Agreement and its three subsets. Results show Mean Squared Error (MSE) and Spearman's Rank Correlation Coefficient (ρ).

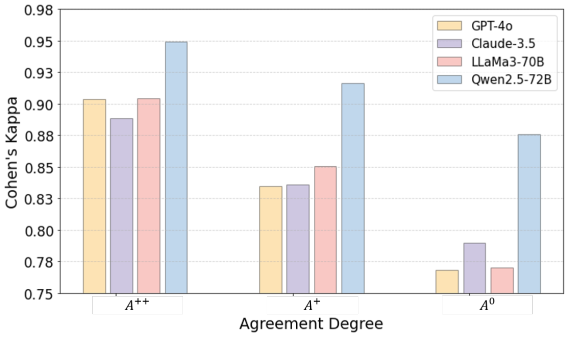
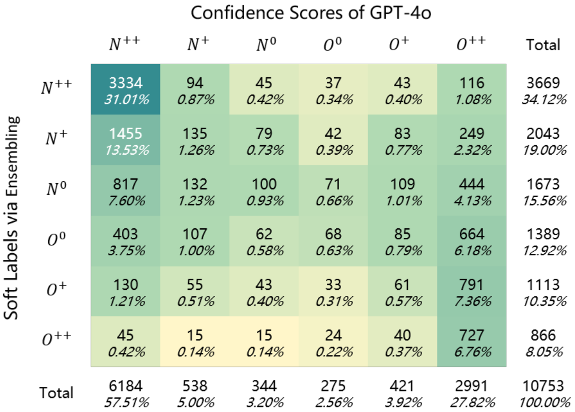


Figure 2: Self-consistency of several LLMs across varying degrees of annotation agreement with Cohen's Kappa (κ) as the metric.



	N^{++}	N^{+}	N^{0}	O^{0}	O^{+}	O^{++}	Total
N^{++}	3334 31.01%	94 0.87%	45 0.42%	37 0.34%	43 0.40%	116 1.08%	3669 34.12%
N^{+}	1455 13.53%	135 1.26%	79 0.73%	42 0.39%	83 0.77%	249 2.32%	2043 19.00%
N^{0}	817 7.60%	132 1.23%	100 0.93%	71 0.66%	109 1.01%	444 4.13%	1673 15.56%
O^{0}	403 3.75%	107 1.00%	62 0.58%	68 0.63%	85 0.79%	664 6.18%	1389 12.92%
O^{+}	130 1.21%	55 0.51%	43 0.40%	33 0.31%	61 0.57%	791 7.36%	1113 10.35%
O^{++}	45 0.42%	15 0.14%	15 0.14%	24 0.22%	40 0.37%	727 6.76%	866 8.05%
Total	6184 57.51%	538 5.00%	344 3.20%	275 2.56%	421 3.92%	2991 27.82%	10753 100.00%

Figure 3: Confusion matrix (raw counts and percentage) between confidence scores of GPT-4o (x-axis) and soft labels (y-axis).

the alignment between LLM confidence and annotation agreement, where a smaller MSE indicates closer alignment (Uma et al., 2021a; Leonardelli et al., 2023). Additionally, we employ Spearman's Rank Correlation Coefficient (ρ) to assess statistical correlation. The detailed metric definitions are provided in Appendix A.2. The results are presented in Table 4.

(1) As annotation agreement decreases, the alignment between model confidence and human agreement weakens. As annotation agreement decreases, LLMs become less reliable in assessing their own uncertainty. GPT-4o, which performs best overall, has an MSE of 0.05 for A^{++} samples but sees this error rise to 0.2 for A^{0} samples. Additionally, Spearman's correlation (ρ) between confidence and agreement weakens from above 0.7 for unanimous samples to below 0.3 for disagreement cases. This suggests that LLMs do not effectively recognize uncertainty in ambiguous cases. In real-world moderation, this could lead to overconfident misclassifications, where the model assigns a high confidence score to an incorrect label, making it harder to detect errors and apply human oversight.

(2) LLMs demonstrate high self-consistency but may be overconfident in disagreement cases. We assess self-consistency using Cohen's Kappa (κ), measuring how stable LLM outputs remain across multiple sampling attempts. As shown in Figure 2, self-consistency decreases for lower agreement samples but remains above 0.75 even for A^{0} cases, indicating strong internal agreement. While high self-consistency is desirable for clear-cut cases, it becomes problematic in ambiguous

	Test Set		A^{++}		A^{+}		A^{o}	
Model	Acc. ↑	MSE ↓	Acc. ↑	MSE ↓	Acc. ↑	MSE ↓	Acc. ↑	MSE ↓
GPT-4o (*zero-shot*)	80.11	0.1133	93.11	0.0560	79.76	0.1242	60.86	**0.1923**
$w/\ A^{++}$	80.93	0.1260	93.65	0.0647	81.41	0.1332	61.21	0.2107
$w/\ A^{+}$	81.26	0.1144	93.65	0.0515	81.96	0.1157	61.80	0.2080
$w/\ A^{0}$	81.22	0.1138	93.73	0.0558	82.18	0.1171	61.33	0.1979
$w/\ A^{++/+}$	83.74	0.1054	95.28	**0.0361**	86.03	0.1120	**63.90**	0.2032
$w/\ A^{++/0}$	**83.87**	0.1063	**95.74**	0.0416	**86.14**	0.1079	63.55	0.2022
$w/\ A^{+/0}$	82.07	0.1171	94.66	0.0500	82.18	0.1290	62.97	0.2059
$w/\ A^{++/+/0}$	83.51	**0.1045**	95.51	0.0367	85.48	**0.1078**	63.32	0.2035

Table 5: Performance of GPT-4o on the test set of MD-Agreement in few-shot learning: Accuracy (Acc.) for binary classification and MSE for evaluating alignment degree between annotation agreement and model's confidence. The first row shows GPT-4o's performance in the zero-shot scenario, while the second and third sections evaluate the model with prompts containing a single level and combinations of agreement, respectively.

cases, as it suggests that LLMs remain overconfident even when human annotators disagree. This rigidity limits the model's ability to adjust for nuanced linguistic or contextual differences.

(3) Even high-performing models exhibit overconfidence, limiting their ability to reflect human-like uncertainty. We construct a confusion matrix of GPT-4o to visually analyze the relationship between the model's confidence score and the soft labels of samples, as shown in Figure 3. The result reveals that even GPT-4o, the best-performing model, assigns high confidence to its predictions regardless of annotation agreement, indicating a lack of adaptability to disagreement cases. This overconfidence highlights a critical flaw in LLM-based moderation: their inability to reflect the diversity of human judgment. Overconfident models are more likely to make systematic errors in handling subjective content, leading to unreliable moderation outcomes. Instead of relying on LLMs as sole decision-makers, future research should explore ensemble methods, uncertainty-aware training, or human-AI collaboration to mitigate biases and improve disagreement resolution.

We further analyze consistency across different models in Appendix B.2, and reveal low agreement among LLMs on disagreement samples. This highlights the potential of ensemble models to handle these nuanced cases.

4 RQ2: Impact of Disagreement Samples on LLM Learning

In this section, we examine how samples with varying annotation agreements influence LLM performance during the learning phase. We focus on two key learning paradigms: few-shot learning and instruction fine-tuning. Specifically, we explore the impact of both single-category agreement samples and different agreement-level combinations on model performance.

4.1 Impact of Disagreement Samples on Few-Shot Learning

We evaluate the effect of disagreement samples on GPT-4o's binary classification accuracy and its confidence alignment with human annotations during few-shot learning.

Few-Shot Learning Setup. We evaluate both single-category agreement samples and combinations of agreement levels in few-shot learning, following (Leonardelli et al., 2021). We first construct prompts using positive and negative sample pairs randomly drawn from the MD-Agreement training set, with each prompt including pairs corresponding to the respective agreement level. For example, the simplest setup $w/\ A^{++}$ consists of only unanimous agreement (A^{++}) samples, containing one offensive and one non-offensive example. Furthermore, we examine mixed setups with different agreement configurations, consisting of sample pairs from their respective single categories for reliable evaluation. For instance, $w/\ A^{++/0}$ and $w/\ A^{++/+}$ combine unanimous agreement samples with one level of disagreement, respectively. Additionally, we assess a broader configuration, $w/\ A^{++/+/0}$, which includes samples from all three agreement levels. The template details are provided in Appendix A.3.

We evaluate model performance on the MD-Agreement test set, analyzing both overall results and performance across different agreement levels. Table 6 summarizes the key findings.

	Test Set		A^{++}		A^{+}		A^{o}	
Model	Acc. ↑	MSE ↓	Acc. ↑	MSE ↓	Acc. ↑	MSE ↓	Acc. ↑	MSE ↓
LLaMa3-8B (zero-shot)	70.92	0.1856	85.22	0.1350	66.45	0.2167	54.09	0.2288
$w/\ A^{++}$	75.79	0.1671	89.01	0.1064	73.60	0.1919	58.18	0.2366
$w/\ A^{+}$	77.04	0.1552	90.40	0.0898	74.81	0.1815	59.70	0.2291
$w/\ A^{0}$	73.99	0.1348	86.53	0.1020	70.74	0.1537	56.31	**0.1665**
$w/\ A^{++/+}$	80.27	0.1340	93.34	0.0643	78.33	0.1582	60.86	0.2232
$w/\ A^{++/0}$	79.29	0.1292	92.49	0.0641	78.11	0.1503	60.16	0.2138
$w/\ A^{+/0}$	$\underline{82.53}$	**0.1075**	$\underline{95.20}$	$\underline{0.0404}$	$\underline{83.94}$	**0.1160**	$\underline{61.80}$	$\underline{0.1978}$
$w/\ A^{++/+/0}$	**84.23**	$\underline{0.1106}$	**95.98**	**0.0379**	**85.81**	$\underline{0.1186}$	**64.37**	0.2150

Table 6: Performance of LLaMa3-8B on the test set of MD-Agreement under instruction fine-tuning.

(1) Few-shot learning improves classification accuracy but may increase overconfidence in ambiguous samples. Few-shot learning enhances classification accuracy, particularly in the medium agreement subset (A^{+}), where accuracy increases by an average of 3.87%. However, for detection of low-agreement samples (A^{0}), few-shot learning increases the MSE, suggesting that models become overconfident and misaligned with ambiguous human annotations. This occurs because few-shot learning reinforces model consistency, making it less adaptable to subjective disagreements.

(2) Learning from disagreement samples improves model generalization. Using disagreement samples (A^{+} and A^{0}) in few-shot learning leads to greater performance improvements across all evaluation metrics compared to using only unanimous agreement samples (A^{++}). Disagreement samples often capture borderline or ambiguous cases, which challenge the model to refine its decision boundaries. Learning from these samples enhances the model's ability to differentiate nuanced offensive language from non-offensive content.

(3) Combining different agreement levels enhances performance, but excessive variation reduces accuracy. Incorporating both unanimous agreement samples and disagreement samples (e.g., $w/\ A^{++/0}$ or $w/\ A^{++/+}$) improves model performance compared to using only disagreement samples ($w/\ A^{+/0}$). However, including too many agreement categories ($w/\ A^{++/+/0}$) does not further enhance accuracy and may even decrease performance. The increased variation makes it harder for the model to establish clear decision boundaries, potentially leading to inconsistent classifications.

These results indicate that strategically balancing agreement levels is critical in few-shot learning. A well-chosen mix of clear and ambiguous cases helps the model generalize effectively, whereas excessive variation may introduce confusion and decrease performance.

To verify the robustness of our findings, we replicate the experiment using the open-source LLM Qwen2.5-72B. The results align closely with those of GPT-4o, suggesting that these insights generalize across different LLM architectures. Detailed results are provided in Appendix B.3.

4.2 Impact of Disagreement Samples on Instruction Fine-tuning

We analyze how instruction fine-tuning with different annotation agreement levels affects model performance, using LLaMa3-7B as the backbone.

Instruction Fine-tuning Setup. We fine-tune an equal number of instances from each agreement level in the MD-Agreement dataset. Specifically, we extract 1,800 samples each from A^{++}, A^{+}, and A^{0}, based on the least-represented A^{0} category. The instruction template remains consistent with that used in the zero-shot setting (see Appendix A.3). We also evaluate combinations of multiple agreement levels, using the same experimental markers as in Section 4.1. Table 5 presents the results, leading to the following conclusions:

(1) Medium-agreement (A^{+}) samples yield the best balance in fine-tuning. Fine-tuning with high-agreement (A^{++}) samples improves classification accuracy, while low-agreement (A^{0}) samples enhance confidence alignment with human annotations, reducing MSE. However, exclusive reliance on A^{0} samples may lead to catastrophic forgetting, where the model becomes overly attuned to ambiguous cases at the cost of general classification accuracy. A^{+} samples offer the best trade-off, allowing the model to capture nuanced decision boundaries while maintaining robust performance.

(2) Combining multiple agreement levels further enhances performance. Fine-tuning with all three agreement levels ($w/\ A^{++/+/0}$) achieves the best overall results, yielding performance comparable to GPT-4o in few-shot learning (see Table 5). Among two-category combinations, mixing disagreement samples ($w/\ A^{+/0}$) provides the most improvement, reinforcing the importance of disagreement-aware learning.

These results confirm that strategically selecting disagreement samples is essential for instruction fine-tuning. A well-balanced combination enhances both classification performance and confidence calibration, ensuring better alignment with human judgments.

We replicate the instruction fine-tuning experiment with Qwen2.5-7B using the same training and test data. The results closely align with those of LLaMa3-7B, confirming that these insights generalize across different model architectures. See Appendix B.4 for details.

5 Related Work

Large Language Model. In recent years, large language models (LLMs) have rapidly emerged, showcasing extensive world knowledge and strong reasoning capabilities (Kojima et al., 2022; Ouyang et al., 2022; OpenAI, 2023). Many researchers have proposed diverse tasks to deeply analyze the relationship between the model's outputs and human judgments (Xu et al., 2024; Fan et al., 2024). In addition, the confidence of LLMs in their outputs has also attracted attention from researchers, which is often used to assess the reliability and robustness of the generated content (Jiang et al., 2021). Various methods for estimating confidence have been proposed (Zhang et al., 2020; Wang et al., 2023b; Tian et al., 2023; Lin et al., 2022). In this study, we employ the most straightforward approach, *self-consistency*, to estimate the model's confidence.

Offensive Language Detection. Researchers have developed various methods for detecting offensive language (Founta et al., 2018; Davidson et al., 2017; Mathew et al., 2021). As research advances, many studies argue that treating offensive language detection as a binary classification is an idealized assumption (Basile et al., 2021; Basile, 2020; Plank, 2022), as annotation disagreement are inherent in datasets for such subjective task (Pavlick and Kwiatkowski, 2019; Uma et al., 2021b). Using majority voting for annotation agreement leads to information loss (Davani et al., 2022), as these disagreements arise from the subtlety of the samples, not labeling errors (Uma et al., 2022). Leonardelli et al. (2023) emphasizes that detection models should recognize this disagreement, rather than just improving classification performance.

Recently, several studies have begun evaluating the potential of LLMs for detecting offensive language (Kumar et al., 2024; Roy et al., 2023), and designing detection methods based on them (Park et al., 2024; Wen et al., 2023). Some studies (Wang et al., 2023a; Huang et al., 2023) leverage the generative capabilities of LLMs to provide explanations for offensive language, assisting human annotation. Furthermore, Giorgi et al. (2024); Zhang et al. (2024) assess the sensitivity of LLMs to demographic information in the context of offensive language. Though great efforts have been made, these studies lack focus on the phenomenon of offensive language with annotation disagreement. In this paper, we aim to fill this research gap.

6 Conclusion

This study examines how LLMs handle annotation disagreement in offensive language detection, a critical challenge in real-world moderation. We evaluate multiple LLMs in a zero-shot setting and find that while they perform well on unanimously agreed-upon samples, their accuracy drops significantly for disagreement cases. Moreover, their overconfidence leads to rigid predictions, misaligning them with human annotations.

To address this, we investigate the impact of disagreement samples in few-shot learning and instruction fine-tuning. Our results show that incorporating these samples improves detection accuracy and human alignment, enabling LLMs to better capture the subjective nature of offensive language. We further find that balancing agreement levels in training data prevents overfitting to ambiguous cases, ensuring model robustness.

Key findings of this work include: (1) a systematic evaluation of LLMs on annotation disagreement, (2) insights into how disagreement samples improve learning, and (3) guidelines for leveraging disagreement-aware training strategies. These results emphasize the need for model calibration techniques to mitigate overconfidence and for training strategies that incorporate disagreement to improve generalization. Future research should explore dynamic fine-tuning approaches and confidence-aware moderation systems to bridge the gap between LLM decisions and human subjectivity.

Limitations

(1) Due to the scarcity of high-quality offensive language datasets with unaggregated labels, we only utilize the MD-Agreement dataset for experiments, which has been widely used in the field. Considering that relying on a single dataset may introduce bias or randomness, we mitigate this by conducting experiments with multiple closed-source and open-source LLMs to ensure the consistency and reliability of our findings, reducing the impact of bias. In future work, we plan to further explore the performance of LLMs in other subjective text analysis tasks, such as humor detection and misogyny detection, particularly in understanding samples with annotation disagreement.

(2) Due to usage restrictions, we are unable to evaluate the detection performance of several emerging LLMs, such as GPT-o3. We plan to further assess these more advanced models as soon as experimental conditions allow. Additionally, due to space limitations, the potential of certain techniques for detecting offensive language with annotation disagreement, such as reinforcement learning methods, are not discussed. We plan to explore these methods in future work and investigate effective strategies for enabling LLMs to fully leverage disagreement samples, thereby enhancing their detection capabilities.

(3) In evaluating the confidence of LLMs, we adopt a straightforward approach based on temperature resampling. We have noted another common method, the Logit-based approach (Guo et al., 2017; Zhang et al., 2020), which involves using the logits of category-specific tokens to compute statistical probabilities within the model's output. This method may provide deeper insights into the decision-making mechanisms of LLMs when handling disagreement samples. We plan to explore and evaluate this method in future work.

Ethics Statement

The opinions and findings contained in the samples of this paper should not be interpreted as representing the views expressed or implied by the authors. Accessing the MD-Agreement dataset requires users to agree to the creators' usage agreements. The usage of these samples in this study fully complies with these agreements.

References

Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. 2019. Crowdsourcing subjective tasks: The case study of understanding toxicity in online discussions. In Companion of The 2019 World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019, pages 1100–1105. ACM.

Valerio Basile. 2020. It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks. In Proceedings of the AIxIA 2020 Discussion Papers Workshop co-located with the the 19th International Conference of the Italian Association for Artificial Intelligence (AIxIA2020), Anywhere, November 27th, 2020, volume 2776 of CEUR Workshop Proceedings, pages 31–40. CEUR-WS.org.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future, pages 15–21, Online. Association for Computational Linguistics.

Connor Baumler, Anna Sotnikova, and Hal Daumé III. 2023. Which examples should be multiply annotated? active learning when annotators may disagree. In Findings of the Association for Computational Linguistics: ACL 2023, pages 10352–10371, Toronto, Canada. Association for Computational Linguistics.

Jiuhai Chen and Jonas Mueller. 2024. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 5186–5200. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. Trans. Assoc. Comput. Linguistics, 10:92–110.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017, pages 512–515. AAAI Press.

Naihao Deng, Xinliang Frederick Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through annotator representations. In Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023, pages 12475–12498. Association for Computational Linguistics.

Yihe Fan, Yuxin Cao, Ziyu Zhao, Ziyao Liu, and Shaofeng Li. 2024. Unbridled icarus: A survey of the potential perils of image inputs in multimodal large language model security. CoRR, abs/2404.05264.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018, pages 491–500. AAAI Press.

Tommaso Giorgi, Lorenzo Cima, Tiziano Fagni, Marco Avvenuti, and Stefano Cresci. 2024. Human and LLM biases in hate speech annotations: A socio-demographic analysis of annotators and targets. CoRR, abs/2410.07991.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, volume 70 of Proceedings of Machine Learning Research, pages 1321–1330. PMLR.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In Companion Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023, pages 294–297. ACM.

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. Trans. Assoc. Comput. Linguistics, 9:962–977.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022.

Deepak Kumar, Yousef AbuHashem, and Zakir Durumeric. 2024. Watch your language: Investigating content moderation with large language models. In Proceedings of the Eighteenth International AAAI Conference on Web and Social Media, ICWSM 2024, Buffalo, New York, USA, June 3-6, 2024, pages 865–878. AAAI Press.

Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. Semeval-2023 task 11: Learning with disagreements (lewidi). In Proceedings of the The 17th International Workshop on Semantic Evaluation, SemEval@ACL 2023, Toronto, Canada, 13-14 July 2023, pages 2304–2318. Association for Computational Linguistics.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, pages 10528–10539. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. Trans. Mach. Learn. Res., 2022.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, pages 14867–14875. AAAI Press.

Negar Mokhberian, Myrl G. Marmarelis, Frederic R. Hopp, Valerio Basile, Fred Morstatter, and Kristina Lerman. 2024. Capturing perspectives of crowdsourced annotators in subjective learning tasks. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024, pages 7337–7349. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. CoRR, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022.

Someen Park, Jaehoon Kim, Seungwan Jin, Sohyun Park, and Kyungsik Han. 2024. PREDICT: multi-agent-based debate simulation for generalized hate speech detection. In Proceedings of the

2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024, pages 20963–20987. Association for Computational Linguistics.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. Trans. Assoc. Comput. Linguistics, 7:677–694.

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.

Sarthak Roy, Ashish Harshavardhan, Animesh Mukherjee, and Punyajoy Saha. 2023. Probing llms for hate speech detection: strengths and vulnerabilities. In Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023, pages 6116–6128. Association for Computational Linguistics.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why don’t you do it right? analysing annotators’ disagreement in subjective tasks. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023, pages 2420–2433. Association for Computational Linguistics.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, pages 5433–5442. Association for Computational Linguistics.

Alexandra Uma, Dina Almanea, and Massimo Poesio. 2022. Scaling and disagreements: Bias, noise, and ambiguity. Frontiers Artif. Intell., 5:818451.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021a. SemEval-2021 task 12: Learning with disagreements. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 338–347, Online. Association for Computational Linguistics.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021b. Learning from disagreement: A survey. J. Artif. Intell. Res., 72:1385–1470.

Han Wang, Ming Shan Hee, Md. Rabiul Awal, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2023a. Evaluating GPT-3 generated explanations for hateful content moderation. In Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China, pages 6255–6263. ijcai.org.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net.

Tharindu Cyril Weerasooriya, Alexander Ororbia, Raj Bhensadadia, Ashiqur KhudaBukhsh, and Christopher Homan. 2023. Disagreement matters: Preserving label diversity by jointly modeling item and annotator label distributions with DisCo. In Findings of the Association for Computational Linguistics: ACL 2023, pages 4679–4695, Toronto, Canada. Association for Computational Linguistics.

Jiaxin Wen, Pei Ke, Hao Sun, Zhexin Zhang, Chengfei Li, Jinfeng Bai, and Minlie Huang. 2023. Unveiling the implicit toxicity in large language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, pages 1322–1338. Association for Computational Linguistics.

Rongwu Xu, Xuan Qi, Zehan Qi, Wei Xu, and Zhijiang Guo. 2024. Debateqa: Evaluating question answering on debatable knowledge. CoRR, abs/2408.01419.

Jize Zhang, Bhavya Kailkhura, and Thomas Yong-Jin Han. 2020. Mix-n-match : Ensemble and compositional methods for uncertainty calibration in deep learning. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, volume 119 of Proceedings of Machine Learning Research, pages 11117–11128. PMLR.

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), ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 12073–12086. Association for Computational Linguistics.

A Experimental Details

A.1 Details of Dataset

In this section, we provide a detailed introduction to the annotation quality control process of our used MD-Agreement dataset (Leonardelli et al., 2021). The researchers implemented a two-stage annotation process: First, three linguists annotated a subset of the samples, and those with unanimous agreement were used as the gold standard for the annotation process. Following this, trained annotators from Amazon Mechanical Turk were employed to annotate the complete samples based on the established gold standard. After the task was completed, annotations from workers who did not achieve at least 70% accuracy were discarded. Additionally, it was ensured that each sample in the final dataset received five annotations. These measures help ensure the accuracy of the annotations. Sandri et al. (2023) further manually reviewed a random selection of 2,570 samples with annotation disagreement from the MD-Agreement dataset. The results showed that only 12 samples contained annotation errors, accounting for less than 0.5%, demonstrating the high quality and reliability of the dataset.

A.2 Description of Metrics

This section introduces the metrics used to assess the relationship between LLM confidence and the degree of human annotation agreement.

Mean Squared Error (MSE): The MSE is a widely used evaluation metric in regression tasks, measuring the difference between predicted and actual values. In this study, we adopt MSE for alignment estimation, as described by Leonardelli et al. (2023), where a smaller MSE indicates closer alignment between LLM confidence and agreement degree. We first obtain soft labels y and soft predictions $\hat{y}$ of samples by averaging their discrete 0-1 annotation sequences Y and the LLM outputs $\hat{Y}$ across different samplings, as follows:

$$y_i = \frac{1}{n}\sum_{i=1}^{n} Y_i, \quad \hat{y}_i = \frac{1}{n}.\sum_{i=1}^{n} \hat{Y}_i, \tag{1}$$

where n is the number of observations, set to $n = 5$ in this paper, representing the number of annotators and LLM outputs. Then, the MSE is calculated as:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2, \tag{2}$$

Range of Coefficient ((ρ)	Correlation Degree
$(0.7, 1.0]$	High Correlation
$(0.3, 0.7]$	Medium Correlation
$(0.0, 0.3]$	Low Correlation
0.0	No Correlation
$[-1.0, 0.0)$	Negative Correlation

Table A1: Correlation degree corresponding to different coefficient values (ρ).

Range of Kappa (κ)	Agreement Degree
$[0.8, 1.0]$	High Agreement
$[0.6, 0.8)$	Good Agreement
$[0.4, 0.6)$	Moderate Agreement
$(0.0, 0.4)$	Poor Agreement
0.0	No Agreement
$[-1.0, 0.0)$	Negative Correlation

Table A2: Agreement degree corresponding to different kappa values (κ).

where m is the total number of samples.

Spearman's Rank Correlation Coefficient (ρ): The Spearman's Rank Correlation Coefficient is a non-parametric test that quantifies the degree of monotonic relationship between two variables. Unlike Pearson correlation, which assumes normally distributed variables, Spearman's correlation does not require this assumption and can be applied to discrete data. This makes it an ideal choice for assessing the statistical correlation between annotation agreement and LLM confidence, which is computed as follows:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}, \tag{3}$$

where di is the difference between the ranks of corresponding values of soft labels y and predictions $\hat{y}$. The correlation degrees corresponding to different ρ are shown in Table A1.

Cohen's Kappa (κ): Cohen's Kappa is widely used to evaluate the consistency between annotators, especially in classification tasks. Compared to simple percentage agreement, which directly calculates the proportion of agreement between two evaluators, Cohen's Kappa provides a more precise measure as it accounts for the influence of random agreement. In this paper, we use Cohen's Kappa to estimate the consistency of LLM outputs $\hat{Y}$. We first compute the kappa value between the i-th and the j-th LLM output of samples, i.e., $\hat{Y}_i$ and $\hat{Y}_j$, using the following formula:

Offensive language refers to words or expressions intended to insult, harm, or belittle individuals or groups. This includes:
· Insulting, derogatory, or combative language
· Discriminatory language based on race, religion, gender, or other personal attributes
· Profanity, violence, or vulgar expressions
· Sarcastic or harmful humor
Based on the above definition, determine if the following sentence is offensive language. Return only "*yes*" or "*no*".

<*If in the Few-shot Learning*>
Example_1:
Input: `[text of example_1]`
Output: `[label_1]`
Example_2:
Input: `[text of example_2]`
Output: `[label_2]`
<*Other Examples*>

Here is the sample to be detected:
Input: `[sample to be detected]`
Output: `[prediction]`

Table A3: Prompt template of the LLM, consisting primarily of three parts: task definition, examples (only for the few-shot scenario), and the sample to be detected.

$$\kappa_{i,j} = \frac{P_{o_{i,j}} - P_{e_{i,j}}}{1 - P_{e_{i,j}}}, \tag{4}$$

where $P_{o_{i,j}}$ represents the observed agreement, which is the proportion of agreement between Y_i and Y_j, and $P_{e_{i,j}}$ refers to the expected agreement, calculated based on the probability of selecting categories, namely 0 or 1. We then calculate the average value of the kappas as follows, which is used as the metric:

$$\kappa = \frac{1}{C(n,2)} \sum_{1 \le i < j \le 5} \kappa_{i,j} \tag{5}$$

The agreement degrees corresponding to different κ are shown in Table A2.

A.3 Design of Prompt Template

To enhance the reproducibility of our study, we avoided conducting complex prompt engineering. Instead, we directly referenced (Roy et al., 2023) to design a straightforward prompt template, as shown in Table A3. The template includes three parts: first, the definition of offensive language, which aligns with that used in the MD-Agreement dataset (Leonardelli et al., 2021) to ensure the accuracy of the evaluation; second, examples of varying degrees of disagreement in a few-shot scenario; and finally, the sample to be detected.

Model	Version
GPT-o1	`o1-preview-2024-09-12`
GPT-4o	`gpt-4o-2024-08-06`
GPT-4	`gpt-4-turbo-2024-04-09`
GPT-3.5	`gpt-3.5-turbo-0125`
Claude-3.5	`claude-3-5-sonnet-20240620`
Gemini-1.5	`gemini-1.5-pro`
LLaMa3-70B	`Meta-Llama-3-70B-Instruct`
LLaMa3-8B	`Meta-Llama-3-8B-Instruct`
Qwen2.5-72B	`Qwen2.5-72B-Instruct`
Qwen2.5-7B	`Qwen2.5-7B-Instruct`
Mixtral-8x22B	`Mixtral-8x22B-Instruct-v0.1`
Mixtral-8x7B	`Mixtral-8x7B-Instruct-v0.1`

Table A4: Specific versions of used LLMs.

A.4 Other Experimental Settings

We access closed-source LLMs via their official APIs and deploy open-source LLMs with parameters downloaded from Hugging Face. To ensure a fair comparison, we use model versions released around the same time, as detailed in Table A4. Since GPT-o1 only has a default temperature of 1 and does not allow adjustments, we present its binary performance in this setting. Except for the temperature coefficient, other hyperparameters, such as top-p and top-k, are set to their default values for each model. For instruction fine-tuning, we adopt the efficient Qlora fine-tuning method. The learning rate is set to 2e-4, with a per-device batch size of 36. We train the model for 15 epochs using the AdamW optimizer, applying an early stopping mechanism. We reserve the parameters of best-performing models based on the development set and evaluate their performance on the test set. The models are trained on two NVIDIA H100 80GB GPUs. All the few-shot learning and instruction fine-tuning experiments are repeated five times with different random seeds to minimize error, and the average results are reported.

A.5 Handling of Refusal Behavior

Handling offensive language can trigger the refusal behavior of LLMs, as they are designed with ethical and safety considerations (Kumar et al., 2024). Nevertheless, in our experiments, refusal occurred only in the zero-shot evaluation setting, where Claude-3.5, with a temperature coefficient set to 1, failed to generate responses for 23 samples. When the experiment was repeated with the same settings, the model successfully provided predictions for these samples. This phenomenon also highlights the model's sensitivity to offensive language.

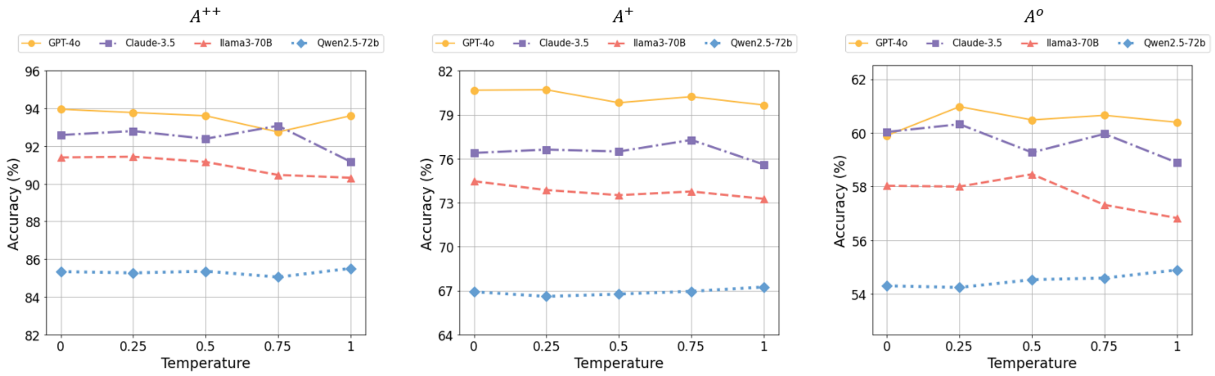


Figure B1: Accuracy of LLMs on detecting offensive language with different degrees of annotation agreement under different temperature sampling settings.

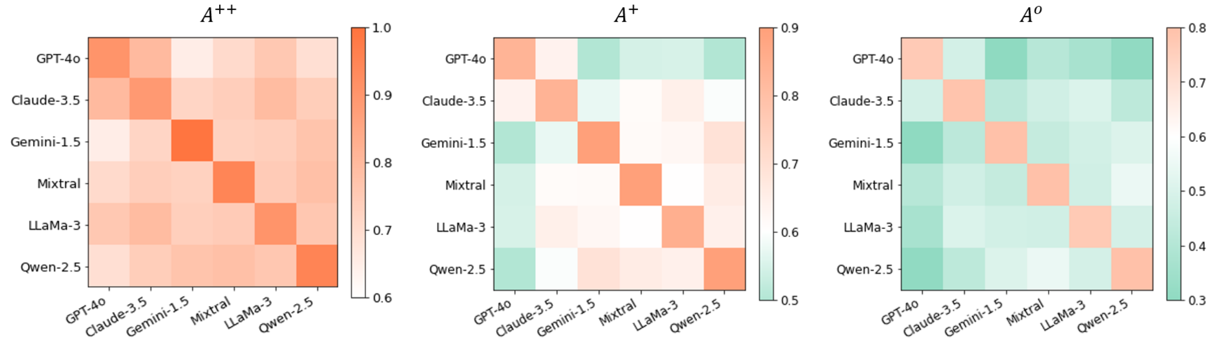


Figure B2: Consistency of outputs from different LLMs across varying degrees of annotation agreement with Cohen's Kappa as the metric. The color scale represents different Kappa values.

B Supplementary Experiments

B.1 Impact of Temperature Sampling on Detection Performance of LLMs

In this section, we analyze the impact of temperature sampling on the accuracy of detecting offensive language by LLMs. We select four representative models for comparison: the closed-source models GPT-4 and Claude-3.5, as well as the open-source models LLaMa3-70B and Qwen2.5-72B. The experimental results are shown in Figure B1. Based on these results, we conclude that after adjusting the temperature coefficient, the detection accuracy of each LLM remains generally stable, although some fluctuations are observed, with varying degrees of sensitivity to the temperature coefficient across different models. As the temperature increases, the accuracy of most models shows a declining trend, with the sole exception being Qwen2.5-72B, which exhibits an increase in accuracy. This may be due to differences in the models' training mechanisms. Nevertheless, the performance ranking between the models remains stable, indicating that changes in the temperature coefficient do not notably affect the performance differences among the models.

B.2 Consistency Analysis Across Different LLMs

Building upon Section 3.2, we further explore the consistency of hard predictions across different LLMs when processing samples with varying degrees of annotation agreement. We select six representative models, including the close-source models GPT-4o, Claude-3.5, and Gemini-1.5, as well as the open-source models Mixtral-8x22B, LLaMa3-70B, and Qwen2.5-72B. Cohen's Kappa is used as the metric. The results are presented in Figure B2. Based on the results, we can observe that:

As annotation agreement decreases, cross-model consistency in detecting offensive language declines more significantly compared to each model's self-consistency. For unanimous agreement samples (A^{++}), cross-model consistency generally exhibits good agreement, with $\kappa > 0.6$. However, for low agreement samples A^{0}, consistency drops explicitly, with many models showing poor agreement ($\kappa < 0.4$), despite many of these models exhibiting similar overall performance in terms of both binary classification accuracy and alignment with human annotations (see Table 3 and 4). Notably, the lowest prediction consistency Kappa is

	Overall		A^{++}		A^{+}		A^{o}	
Model	Acc. ↑	MSE ↓	Acc. ↑	MSE ↓	Acc. ↑	MSE ↓	Acc. ↑	MSE ↓
Qwen2.5-72B (*zero-shot*)	72.08	0.1962	84.74	0.1480	68.41	0.2237	57.12	0.2393
$w/\ A^{++}$	77.92	0.1321	90.94	0.0809	75.91	0.1484	60.40	0.1920
$w/\ A^{+}$	79.10	0.1275	91.95	0.0702	78.66	0.1414	60.16	0.1993
$w/\ A^{0}$	$\underline{82.56}$	$\underline{0.1054}$	$\underline{94.12}$	$\underline{0.0514}$	83.50	**0.1088**	**64.14**	**0.1832**
$w/\ A^{++/+}$	81.42	0.1127	93.19	0.0561	82.51	0.1199	62.50	$\underline{0.1905}$
$w/\ A^{++/0}$	82.43	0.1099	93.58	0.0530	$\underline{84.38}$	0.1108	$\underline{63.55}$	0.1950
$w/\ A^{+/0}$	**82.96**	**0.1044**	**94.97**	**0.0427**	**84.71**	$\underline{0.1090}$	62.97	0.1927
$w/\ A^{++/+/0}$	82.04	0.1101	93.42	0.0544	84.05	0.1095	62.73	0.1949

Table B1: Performance of Qwen2.5-72B on the test set of MD-Agreement in few-shot learning.

	Overall		A^{++}		A^{+}		A^{o}	
Model	Acc. ↑	MSE ↓	Acc. ↑	MSE ↓	Acc. ↑	MSE ↓	Acc. ↑	MSE ↓
Qwen2.5-7B (*zero-shot*)	69.77	0.1998	83.44	0.1542	66.12	0.2289	53.04	0.2379
$w/\ A^{++}$	80.18	0.1407	92.96	0.0703	79.76	0.1572	61.21	0.2309
$w/\ A^{+}$	80.41	0.1347	93.11	0.0649	80.64	0.1497	62.73	0.2211
$w/\ A^{0}$	80.08	0.1395	92.57	0.0704	80.09	0.1533	60.86	0.2238
$w/\ A^{++/+}$	$\underline{82.30}$	0.1261	$\underline{95.36}$	$\underline{0.0457}$	82.95	0.1406	62.38	0.2311
$w/\ A^{++/0}$	82.17	0.1192	94.50	0.0486	$\underline{84.27}$	$\underline{0.1243}$	$\underline{63.90}$	$\underline{0.2082}$
$w/\ A^{+/0}$	81.88	$\underline{0.1185}$	94.43	0.0522	83.39	0.1283	$\underline{63.90}$	**0.2060**
$w/\ A^{++/+/0}$	**83.91**	**0.1149**	**95.82**	**0.0395**	**85.59**	**0.1209**	**65.42**	0.2181

Table B2: Performance of Qwen2.5-7B on the test set of MD-Agreement under instruction fine-tuning.

only 0.28 between GPT-4o and Gemini 1.5. A potential reason for this phenomenon is that different models are trained on diverse datasets and undergo distinct value alignment processes, resulting in varying sensitivity to contextual features. In future work, we will explore the relationship between cross-model consistency and human annotation agreement in offensive language detection. Additionally, we aim to investigate the potential of leveraging multiple LLMs for collaborative annotation of offensive language.

B.3 Few-shot Learning with Qwen2.5-72B

We replicate the few-shot learning experiment from Section 4.1 using the open-source LLM Qwen2.5-72B, employing the same sample pairs in the prompts. The results are shown in Table B1. Based on the results, we observe the following:

In the few-shot learning with samples of varying annotation agreement degrees, the results of Qwen2.5-72B align closely with the trends of GPT-4o (see Table 5). Whether introducing samples with a single annotation agreement degree or combinations of different agreement categories, the detection performance of the model shows notable improvement compared to the zero-shot scenario. Additionally, the benefit to model performance varies explicitly depending on the annotation agreement degree and the combinations used as prompts.

Furthermore, compared to GPT-4o, Qwen2.5-72B demonstrates two distinct differences: (1) On the subset of low-agreement samples, the introduction of few-shot learning also results in a noticeable improvement in the alignment between LLM confidence and annotation agreement. This is primarily because Qwen2.5-72B performs less effectively than GPT-4o in detecting offensive language under zero-shot scenarios, making it more susceptible to performance enhancements through few-shot learning. (2) For Qwen2.5-72B, the combination of disagreement samples, i.e., $w/\ A^{+/0}$, achieves the best performance across most subsets and metrics. In contrast, GPT-4o performs better with combinations such as $w/\ A^{++/+}$ and $w/\ A^{++/0}$. This highlights that the effects of learning from disagreement samples differ between different LLMs, which is closely related to the ambiguous characteristics of these samples.

B.4 Fine-tuning with Qwen2.5-72B

We replicate the instruction fine-tuning experiment from Section 4.2 using Qwen2.5-7B, training with the same instruction data. Based on the results shown in Table B2, we observe conclusions that are largely consistent with those on LLaMa3-8B. Specifically, Qwen2.5-7B performs best with

Exp.	Sentence	Label	GPT4o	LLaMa3
(a)	Why does the Minneapolis mayor look like a more Soy version of Justin Trudeau	O^{+}	O^{0}	N^{0}
(b)	This is Amerikkka! #blacklivesmatter <url>	N^{0}	O^{++}	O^{++}
(c)	Hey #BlackLivesMatter SAY THEIR NAMES! Look these are YOUR PEOPLE! Say Their Names! Do It! Where is Your Outrage Now?! Come forth You #Hypocrites! Did They Die In Vain?! Ye Generation Of Vipers! These Are Your People Murdered!	O^{0}	N^{0}	N^{++}
(d)	Doesn't have a kid's cancer charity he can draw from? #Elections2020	N^{+}	O^{++}	N^{++}

Table B3: Error analysis of samples with annotation disagreement in offensive language detection.

medium agreement (A^{+}) when fine-tuned using a single category of annotation agreement across various subsets and metrics. Compared to unanimous agreement samples (A^{++}), training with disagreement samples helps the model's confidence better align with the degree of human annotation agreement, resulting in a lower MSE. When fine-tuning with combinations of different agreement categories, using lower agreement combinations (e.g., $w/\ A^{++/0}$ and $w/\ A^{+/0}$) leads to higher detection accuracy and closer alignment with the annotations, compared to higher agreement combinations ($w/\ A^{++/+}$), especially in the detection of disagreement subsets. These findings further demonstrate the importance of learning from disagreement samples during instruction fine-tuning, which enhances the model's understanding and judgment of offensive language, particularly for ambiguous samples.

B.5 Error Analysis

To gain deeper insight into the challenge posed by offensive language with annotation disagreement, we manually inspect the set of samples misclassified by the models. The following two main types of errors are identified, with samples and predictions from GPT-4o and LLaMa3-72B shown in Table B3 for illustration:

Type I error refers to samples that are labeled as *non-offensive* but are detected as *offensive*. This error primarily arises from subtle linguistic features such as sarcasm and metaphor, which make the judgment of the sample ambiguous. For instance, in Example (a), the term "*Amerikkka*" is a variant of "*America*" used to intensify emotional expression. Due to insufficient context, most annotators do not consider it offensive. However, GPT-4o and LLaMa3, due to their sensitivity to the hashtag *blacklivesmatter*, consistently classify it as offensive language. Similarly, in Example (b), a sarcastic rhetorical question leads to a misclassification by GPT-4o. This phenomenon highlights the complexity that human annotators face in determining offensive language and also reveals the issue of over-sensitivity in existing LLMs to certain linguistic expressions, resulting in decisions that do not align with human standards. In future work, we will perform a more detailed analysis of expressions in samples with disagreement annotation and explore how different types of expressions affect model detection performance.

Type II error refers to sentences labeled as offensive but classified as non-offensive by the models. This error primarily arises from the models lacking or failing to effectively integrate the necessary background knowledge for detecting offensive content, leading to an inaccurate understanding of the sample's true meaning. For example, in Example (c), the comparison between *the mayor of Minneapolis* and *Justin Trudeau* uses "*Soy*" as an adjective, which implies weakness and is intended to belittle the mayor. Both human annotators and GPT-4o capture the offensive nature of the sample, but LLaMa3 fails to correctly identify its offensiveness due to insufficient relevant knowledge. In Example (d), the phrase "*Ye Generation of Vipers*", a religiously charged expression, is used to strongly criticize police brutality against black people. However, the model fails to integrate the context, leading to a missed detection. We plan to introduce more comprehensive background knowledge to enhance the understanding capability of LLMs and explore the performance of knowledge-enhanced models in detecting disagreement samples.