BANHADEX: Towards Explainable HAte Speech Detection in Bangla using Human Annotated EXplanation

Faisal Hossain Raquib[2,*], Akm Moshiur Rahman[1,*], Md Fahim[1,5,*,†], Md Tahmid Hasan Fuad[1,3], Md Farhan Ishmam[4], Faria Sultana[1], M Ashraful Amin[1], Amin Ahsan Ali[1], AKM Mahbubur Rahman[1]

[1]*Center for Computational & Data Sciences* [2]*Rajshahi University of Engineering and Technology*
[3]*University of Manitoba* [4]*University of Utah* [5]*Penta Global Limited*
[*]Equal Contribution [†]Project Lead

Correspondence: {moshiur.mazumder, fahimcse381}@gmail.com

Abstract

Online safety in low-resource languages hinges not only on accurate hate speech detection but also on transparent, culturally grounded explanations. Yet prior work in Bangla largely focuses on hate classification while overlooking interpretability. We address this gap by introducing BANHADEX, the first hate explainability dataset in Bangla with human-annotated labels. BANHADEX contains 19,203 YouTube comments spanning April 2024–June 2025, annotated for binary hate classification with seven fine-grained hate categories, seven target groups, and concise explanations for each sample. Our data pipeline relies on a two-stage annotation protocol that uses majority voting for robust labeling. Our rich suite of experiments on open and closed-source LLMs reveals that explanation-guided LoRA substantially outperforms both classification and explanation quality across prompting and fine-tuning strategies. BANHADEX establishes the groundwork for faithful interpretability and safer moderation in linguistically rich yet under-resourced languages. The code and dataset are publicly available at: https://github.com/MOSHIIUR/BANHADEX.

Disclaimer: This paper contains potentially offensive content essential to the subject matter.

1 Introduction

Social media has revolutionized human communication and brought unprecedented transformations connecting people (Kaplan and Haenlein, 2010). With a user adoption rate of more than 50% and daily time spent exceeding 2 hours (Gudka et al., 2023), social media has become an integral part of our lives. Yet the evolution of digital technology and online connectivity resulted in the proliferation of online hate content, with studies attesting to the rising hate in mainstream social networks (Goel et al., 2023). Hateful comments can be a form of negative emotional self-disclosure, where individuals, under heightened perceived stress, externalize their internal distress by venting through hostile expressions. (Wendorf and Yang, 2015), with users preferring native language for emotional salience (Reghunathan and Asha, 2022).

As of 2025, Bangla has been spoken by over 240 million speakers (Wikipedia, 2025), making it one of the more prominent sources of hate moderation, yet studies reveal a dire state of the domain. While multiple Bangla hate detection datasets exist (Sharif et al., 2022; Haider et al., 2025b; Romim et al., 2021), none of the current resources provide contextual or linguistic explanations alongside the annotated labels. This is a fundamental gap, as explanation is crucial to understanding the subtle and contextual nature of hate speech (Mathew et al., 2021), particularly in a linguistically dense yet under-resourced language such as Bangla.

Although recent LLMs have demonstrated impressive potential in providing context-aware explanations (Di Bonaventura et al., 2024), they are not without limitations. In particular, such models tend to hallucinate (Xu et al., 2025; Huang et al., 2025), deriving explanations that are factually incorrect or unrelated to the input. Additionally, the generated outputs are often too verbose or conceptually dense (Saito et al., 2023), rendering them difficult to interpret and, in some cases, even uninterpretable. This constitutes a major drawback in high-stakes applications, such as hate speech detection, where concise, clear, and contextually grounded explanations are crucial for transparency, trust, and informed decision-making.

We overcome these limitations by introducing BANHADEX, the first Bangla hate speech dataset to include expert-annotated explanations in addition to each label. These human-curated descriptions are intended to bolster both classification and interpretability, thereby improving performance while fostering transparency and explainability.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 43652–43674
July 2-7, 2026

Datasets	#NH	#H	H:NH	Data Source	#Annotators	Hate Exp.
Sharif et al. (2022)	7,361	8,289	1.12	YouTube, Facebook	2	✗
Belal et al. (2023)	7,585	8,488	1.12	Previous Datasets	2	✗
Banik and Rahman (2019)	5,964	4,255	0.71	Social Media	-	✗
Romim et al. (2021)	20,000	10,000	0.50	YouTube, Facebook	50	✗
BANHADEX(Ours)	10,048	9,155	0.91	Youtube	4	✓

Table 1: **Comparison between existing datasets** based on the number of Non-Hate (#NH), Hate (#H) samples, hate to non-hate ratio (H:NH), source of dataset (Data Source), number of annotators (#Annotators), and presence of hate explanation (Hate Exp.). The hate speech datasets include similar data classes, *e.g.*, aggression and toxic speech.

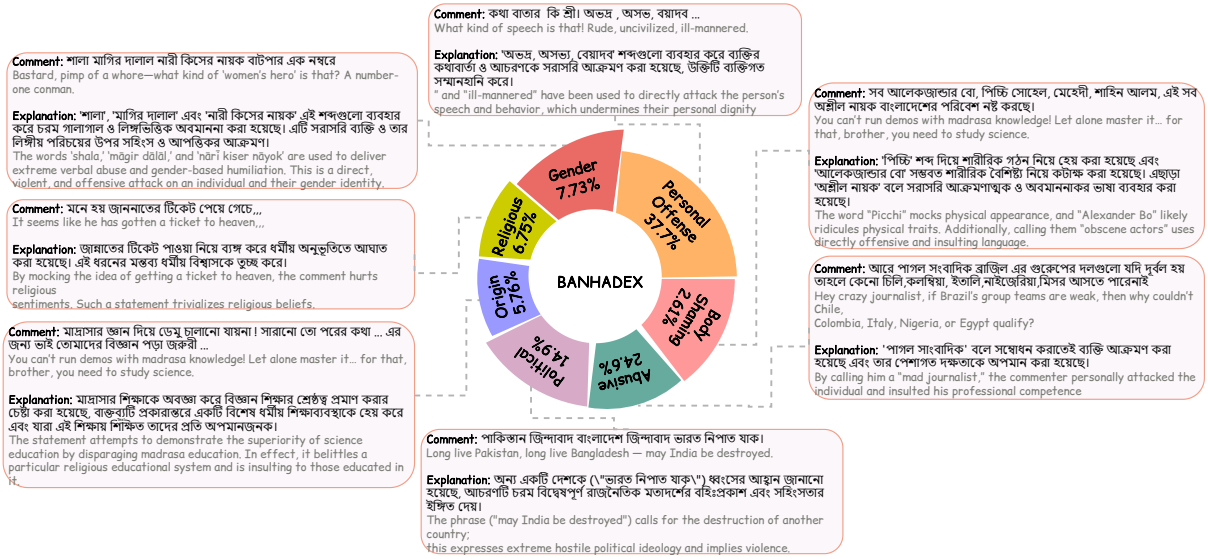


Figure 1: Hate Contents & Their Explanations in BANHADEX Dataset across Different Hate Categories.

2 Related Work

Hate speech detection has become an essential task in NLP. Most of the research on hate speech focuses on detection (Davidson et al., 2017). Although the preliminary stage of this research was binary (Antypas and Camacho-Collados, 2023; Bose and Su, 2022), researchers have extended this work to multi-class (Walsh and Greaney, 2025; Hashmi and Yayilgan, 2024), and even multi-label classification (Ilma et al., 2021). Researchers have also explored hate speech detection in multi-modal settings (Boishakhi et al., 2021; Irfan et al., 2024). Lastly, multilingual and cross-lingual hate speech detection has become a key focus, with studies leveraging multilingual BERT and related architectures to scale detection capabilities across languages (Aluru et al., 2020; Ousidhoum et al., 2019).

Due to its status as a low-resource language, Bangla has received comparatively less attention in hate speech detection research than high-resource languages. Consequently, early efforts were mostly limited to binary hate speech detection (Remon et al., 2022; Das et al., 2021), followed by research in related domains, *e.g.*, abusive content (Aurpa et al., 2022; Emon et al., 2019), cyberbullying (Ahmed et al., 2021), gender discrimination & sexism (Jahan et al., 2023), and toxic speech (Banik and Rahman, 2019). More recent research has shifted to multi-label classification tasks, where hate speech is classified along dimensions such as religion, politics, and gender (Sharif et al., 2022; Haider et al., 2025b; Raquib et al., 2025). This includes aggression detection (Hossain et al., 2023), cyberbullying detection (Saifuddin et al., 2023), and various types of toxic content analysis (Belal et al., 2023), as well as more general hate speech detection (Shakil, 2022; Hasan et al., 2024).

Alongside classification, explanation generation has become an important focus in hate speech detection, addressing the growing need for interpretability in automated systems. Instead of only labeling content as hateful, these studies focus on explaining *why* a comment is hateful. Some approaches identify important textual spans that contribute to the decision (Mathew et al., 2021), while others produce natural language explanations that

incorporate cultural or social reasoning (Mehta and Passi, 2022). Multi-task learning frameworks have also been developed to jointly optimize both classification and explanation (Piot and Parapar, 2025). Recently, (Mia et al., 2025) extended this direction to misogynistic meme detection in Bangla. Despite this progress, explanation generation remains largely unexplored in low-resource languages, leaving a notable gap in the literature.

3 The BANHADEX Dataset

Bangla Hate Speech Detection with Explanation (BANHADEX) follows a rigorous and multi-phase pipeline for producing classification and explanation labels. Input categorization includes five video categories, while output categorization includes binary hate labels, seven fine-grained hate categories, and seven hate classes. Figure 2 illustrates the end-to-end workflow of our dataset creation process.

3.1 Data Sourcing

To ensure linguistic diversity and sociocultural relevance, we selected YouTube as the primary data source because it is one of the most widely used platforms in Bangladesh for public discourse and offers a large volume of Bangla spoken content, which aligns with our speech-focused analysis. In contrast, Facebook has strict API and data access restrictions that limit large-scale, reproducible data collection, while X is comparatively less popular in Bangladesh for Bangla speech content. We collected videos spanning five distinct categories, News & Politics, Entertainment, People & Blogs, International, and Sports, to capture a representative cross-section of hate expressions. A total of 328 videos, posted between April 2024 and June 2025, were included in our dataset. All top comments were extracted, excluding replies to prevent conversational dependency, resulting in an initial pool of 26,730 Bangla comments.

Data Filtering. We implemented a systematic filtering process to ensure the dataset's quality and relevance. Specifically, we removed: (1) all non-Bangla comments, to maintain Bangla linguistic consistency, and (2) comments shorter than 20 characters, to eliminate context-poor samples. This filtering step removed 6,291 comments, leaving a final dataset of 20,439 Bangla-language comments with sufficient semantic depth.

Data Cleaning. We deduplicated comments and filtered extraneous elements, such as hyperlinks, URLs, hashtags, and any personal information, including usernames and mentions, to preserve user privacy and maintain textual clarity. This preprocessing step eliminated 1,236 duplicate and non-essential entries, resulting in a final curated dataset of 19,203 clean Bangla comments for annotation.

3.2 Data Annotation

To ensure high-quality annotation of the BANHADEX, we recruited four native Bangla-speaking undergraduate students with demonstrated expertise in Bangla culture and social media discourse. Their familiarity with interpreting social media content enabled them to effectively capture nuanced expressions of misogyny. We provided comprehensive guidelines to annotators for the annotation process (§B).

We employed a two-stage annotation process. In the first stage, each comment was classified as either Hate or Non-Hate. For samples labeled as hate, annotators were further instructed to specify the target group and the relevant hate category. Annotations were carried out independently by multiple annotators, and final labels were determined by majority voting.

In the second stage, we introduced an explanation generation task. Here, annotators were asked to provide concise yet insightful justifications for the labels they assigned, offering clarity on the rationale behind their decisions. Unlike the classification phase, each explanation was authored by a single annotator (details in § B.3). Annotators were provided monetary compensation on a per-sample basis for both phases. Additional information on our validation strategy and metrics used to ensure high annotation quality is available in §D.

3.3 Dataset Statistics

The dataset statistics of BANHADEX are presented in Tab. 2. Tab. 3 summarizes the major incidents represented in the dataset, spanning July 2024 to June 2025. The dataset captures hate speech related to significant real-world events, including the Iran-Israel conflict, Indo-Pak tensions, the Quota Reform Movement, and the Post-Regime Change period. The most prevalent hate categories observed across these events are Abusive/Violent, Personal Offense, Political, and Religious.

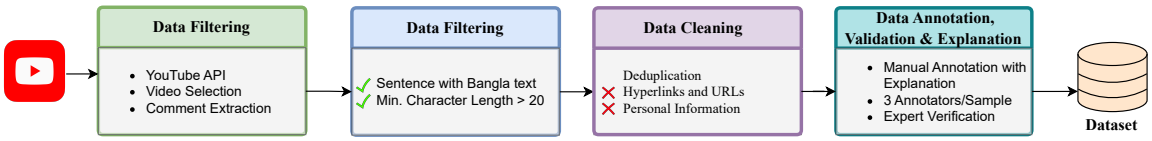
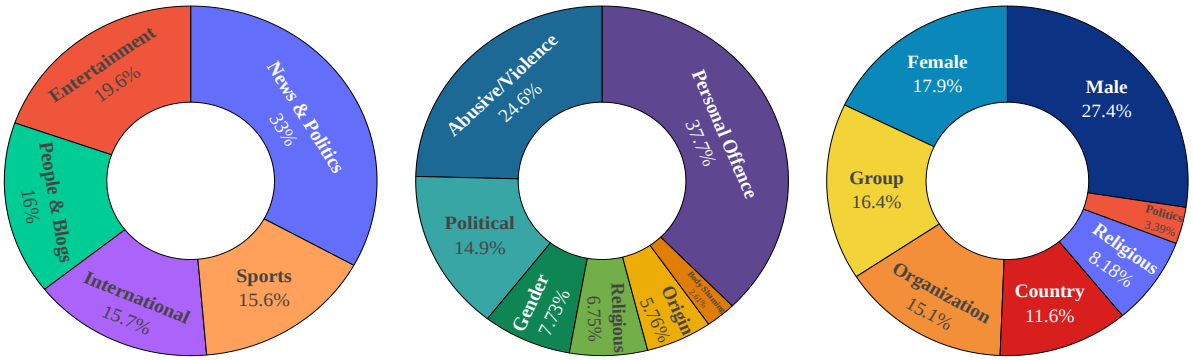


Figure 2: BANHADEX dataset development pipeline illustrating the four-stage process: Data Source collection from social media platforms, Data Filtering to remove non-relevant content, Data Cleaning to eliminate duplicates and extraneous elements, and Data Annotation & Validation resulting in the final dataset.



(a) Video Category Distribution (b) Hate Category Distribution (c) Target Group Distribution

Figure 3: Category Distribution in the BANHADEX Dataset

Fig. 3 illustrates the distributions of video categories (Fig. 3a), hate categories (Fig. 3b), and target groups (Fig. 3c). The data reveals the majority of comments originate from News & Politics videos. Personal Offense and Abusive/Violence account for over 60% of all hate-labeled comments. In terms of target groups, individuals—specifically male and female—are the most frequently targeted.

4 Experiment Design

We evaluate a diverse range of open- and closed-source models. We classify our experiments into two categories: (i) Prompt-based Experiments and (ii) LoRA Finetuning Experiments.

4.1 Prompt-Based Experiments

For the prompt-based experiments, we consider effective prompting techniques for hate speech detection: Zero-Shot, Chain of Thought, HARE (Yang et al., 2023), and Why Hate/Non-Hate (Haider et al., 2025b) prompting techniques. A comparison of those prompting techniques is provided in Table 4, and details of the prompts used in the experimentation are given in the Appendix. We also evaluate closed-source models, including GPT-4o (GPT-4o-mini), and Gemini (Gemini-2.0-Flash).

4.2 LLM LoRA Fine-Tuning

We consider open-source LLMs and apply LoRA (Hu et al., 2022) fine-tuning (FT) techniques in two fine-tuning approaches.

Classification Guided LoRA Fine-Tuning. Here, we adopt the standard LoRA variant, *i.e.*, rather than updating all pre-trained parameters directly, a low-rank decomposition is used to efficiently adapt the model by training significantly fewer trainable parameters. Specifically, given a pre-trained weight matrix W (which remains frozen during training), LoRA adds a learnable low-rank update defined as $\Delta W = AB^\top$, where $A \in \mathbb{R}^{d\times r}$ and $B \in \mathbb{R}^{r\times d}$ are the trainable matrices, and $r \ll d$. The updated weights are computed as: $W' = W + \Delta W = W + AB^\top$, allowing the model to capture task-specific knowledge while retaining the generalization capabilities of the pre-trained backbone. The matrices A and B are trained using the loss function for the downstream task. In our case, we optimize these parameters using the classification loss $\mathcal{L}_{\text{Class}}$, which denotes the standard cross-entropy loss used for the hate speech classification task.

Explanation-Guided LoRA Fine-Tuning. We also explore an *Explanation-Guided* fine-tuning

Splits		
Train		15362
Test		3841
General Statistics		
Samples		19203
Videos		328
Hate Samples		9,155
Non-Hate Samples		10,048
Video Categories		5
Hate Categories		7
Target Groups		7
Samples	**Hate**	**Non-Hate**
Train	7324	8038
Test	1831	2010
Mean word count	15.78	12.74
Max word count	496	514
Min word count	5	6
Explanations	**Hate**	**Non-Hate**
Mean word count	26.10	22.41
Max word count	68	66
Min word count	10	7

Table 2: Dataset statistics of BANHADEX. Samples are annotated based on binary hate labels, seven fine-grained hate classes, and seven target groups. Each sample has its unique explanation for hate.

strategy to enhance both model interpretability and performance. Given, human-annotated ground truth explanations $Y = \{y_1, \ldots, y_{T_y}\}$, where T_y is the length of the explanation sequence, we instruct the models to generate the explanation and incorporate explanation supervision directly into the training process. The model is trained to generate explanation tokens autoregressively, conditioned on the input x (the comment), using a next-token prediction objective. The explanation loss is defined as:

$$\mathcal{L}_{\text{Explanation}} = -\sum_{t=1}^{T_y} \log P(y_t \mid y_{<t}, x),$$

where $y_{<t}$ represents the preceding explanation tokens. This teacher-forced training encourages the model to produce fluent and semantically aligned explanations with the ground truth annotations. The overall training objective jointly optimizes for both classification and explanation generation:

$$\mathcal{L}_{\text{Exp\_LoRA}} = \mathcal{L}_{\text{Class}} + \mathcal{L}_{\text{Explanation}}.$$

During inference, the model can generate an explanation $\hat{Y}$ for a given input, thereby enhancing the transparency and interpretability of the classification decision. We configure LoRA with $\alpha = 64$, $r = 64$, a dropout rate of 0.01, a learning rate of 1×10^{-4}, and a batch size ranging from 4 to 32. LoRA is applied to all weight matrices of the pre-trained models, and each model is fine-tuned for a single epoch. For inference, we use vLLM(Kwon et al., 2023), while LLaMA-Factory(Zheng et al., 2024) is employed for LoRA fine-tuning. To ensure reproducibility, we evaluate using greedy decoding with a temperature of 0 and no sampling.

5 Result and Analysis

Table 5 depicts the results of the open and closed source models in different configurations. All experiments reported in Table 5 and Figures 4–6 were conducted across three different random seeds, and the reported results correspond to the averaged performance over these runs. We observed only minor variance across seeds, and the overall trends, particularly regarding the impact of explanations and titles, remained consistent.

Open Source LLMs. Among open-source models, Phi-4 achieves the highest performance in detecting the non-hate category, with an F1 score of 83.86. For the hate category, Gemma-3 outperforms other open models with an F1 score of 93.90. Notably, Gemma-3, when prompted in a zero-shot setting, achieves the highest F1 score for hate detection across all model configurations, including closed-source LLMs. In contrast, Mistral and Qwen 2.5 perform worse than other open-source models. This trend is also reflected in their explanation evaluation results. Gemma-3 achieves the highest average explanation score of 65.12, while other models show a drop of 7. Interestingly, Qwen 2.5 ranks second in explanation quality despite ranking among the lowest in hate detection.

Close Source LLMs. For the closed-source models, GPT-4o consistently outperforms Gemini 2.5 across most evaluation metrics. GPT-4o achieves the highest performance in non-hate detection tasks with an F1 score of 83.10, slightly ahead of Gemini 2.5's 80.42. In hate detection, GPT-4o beats Gemini with a margin of 10% F1 Score. In explanation quality, GPT-4o again surpasses Gemini 2.5, achieving an average score of 70.67, whereas Gemini 2.5 records 69.38.

Time Period	Event	Samples	Major Hate Categories
Apr' 2025	Pahalgam attack	546	Origin, Religious, Personal
Jun' 2025	Iran - Israel War	1638	Abusive/Violence, Religious, Personal
May' 2025	Ind-Pak War	707	Abusive/Violence, Personal
July'24 - Aug'24	Quota Reform Movement	1542	Abusive/Violence, Political, Personal
Aug'24 - Nov'24	Post-Regime Change Events	1483	Abusive/Violence, Political, Personal

Table 3: Events Covered in Dataset with Number of Hate Samples and Major Hate Categories

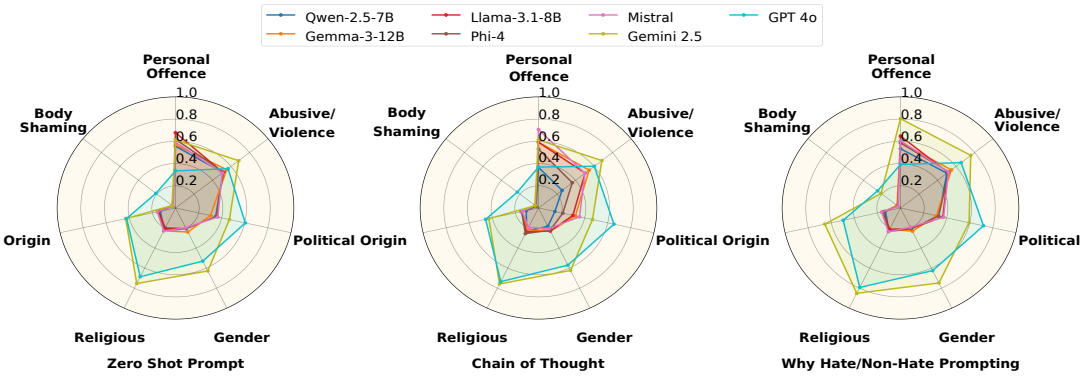


Figure 4: Model performance on hate categories of the BANHADEX dataset.

Strategy	Prompt
Zero-Shot	Base prompt (BP)
CoT	BP + "Think step by step"
HARE	BP + "Explain step by step"
Why Hate?	BP + "Explain Hate/Non-Hate"

Table 4: Prompting strategies used in our benchmarks.

Prompting Results. Prompting strategies have a great impact on both hate/non-hate classification and explanation generation, as shown in Table 5. CoT prompting improves performance across classification and explanation metrics for closed-source models. However, it leads to a noticeable decline in performance for open-source models—particularly for Qwen-2.5 and Mistral, which experience a substantial drop compared to their zero-shot baselines.

In contrast, HARE prompting enhances the performance of most open-source models across both tasks, but negatively affects closed-source models, showing an opposite trend. Interestingly, the "Why H/NH" prompting format benefits closed-source models, thereby improving the quality of explanations. Notably, GPT-4o, when prompted with "Why H/NH", achieves the highest explanation score among all model configurations.

Finetuning Results. We conducted experiments using two LoRA-based fine-tuning techniques on open-source models: classification-guided and explanation-guided fine-tuning. In the classification-guided LoRA fine-tuning setup, we observed a consistent improvement in recall, resulting in better F1 scores—particularly for the non-hate category, with gains of approximately 5–10% compared to the zero-shot baseline. A similar trend was observed for the hate category, where most models showed an improvement of 2–3% in F1 score, with the exception of Gemma-3.

The explanation-guided LoRA fine-tuning yielded even stronger results. Compared to classification-guided fine-tuning, it achieved an additional 4–5% improvement in F1 for the non-hate category and 5–7% for the hate category. Moreover, this variant also produced the highest explainability scores among all open-source models, showing a 2–4% increase in average explanation quality over the zero-shot explainability score.

Hate Category-wise Results. We further analyze model performance across specific hate categories, where a hate category is predicted only if the model classifies a comment as hate. Figure 4 presents an overview of the results, and detailed scores are provided in Table 11 in the Appendix. From the figure, it is evident that all models struggle significantly with the Body Shaming (BS) category, consistently underperforming in its detection. In contrast, the models perform relatively well in identifying Religious, Personal Offence (P Off), Abusive/Violence (A/V), and Gender (Gen) hate categories.

Models	Non-Hate			Hate			Explanations		
	P	R	F1	P	R	F1	BB_Score	LAVE	Avg
Zero Shot Prompt									
Open Source LLM									
Qwen-2.5-7B	74.22	91.70	82.04	87.68	64.97	74.64	50.22	66.90	58.56
Gemma-3-12B	83.33	69.43	75.78	88.71	99.74	**93.90**	50.01	80.23	65.12
Llama-3.1-8B	89.79	63.82	74.61	69.74	92.00	79.34	51.33	62.87	57.10
Phi-4	80.44	87.59	83.86	84.83	76.52	80.46	49.78	62.05	55.92
Mistral	75.36	74.50	74.93	72.24	73.15	72.69	50.30	48.07	49.19
Close Source LLM									
Gemini 2.5	68.65	**96.23**	80.42	93.06	56.34	69.72	55.74	83.02	69.38
GPT 4o	73.15	95.27	83.10	93.22	65.14	77.26	57.30	84.03	70.67
Chain of Thought									
Open Source LLM									
Qwen-2.5-7B	58.25	98.41	73.18	92.10	20.84	33.99	49.8	62.08	55.94
Gemma-3-12B	79.36	90.12	84.40	87.19	74.17	80.15	42.3	72.77	57.54
Llama-3.1-8B	75.45	87.29	80.94	83.06	68.69	75.19	49.6	65.41	57.51
Phi-4	63.42	96.36	76.50	90.64	38.78	54.32	49.11	75.69	62.40
Mistral	**93.91**	26.17	40.94	54.77	**98.14**	70.31	50.01	34.63	42.32
Close Source LLM									
Gemini 2.5	70.88	96.04	82.68	93.10	57.82	70.53	56.14	83.84	69.99
GPT 4o	74.39	95.22	83.46	93.17	68.82	79.27	58.01	84.88	71.45
HARE Prompting									
Open Source LLM									
Qwen-2.5-7B	77.84	88.36	82.77	84.90	72.23	78.05	50.03	64.12	57.08
Gemma-3-12B	79.78	90.12	84.64	87.29	74.82	80.58	51.7	80.47	66.09
Llama-3.1-8B	88.17	70.23	78.19	73.84	89.92	81.09	49.21	55.14	52.18
Phi-4	73.89	92.15	82.02	88.11	64.09	74.21	49.62	63.15	56.39
Mistral	88.60	45.53	60.15	60.91	93.54	73.78	51.22	55.91	53.57
Close Source LLM									
Gemini 2.5	68.15	96.10	80.26	93.58	56.63	69.28	55.81	82.93	69.37
GPT 4o	65.20	98.05	78.53	**95.24**	48.87	64.87	52.23	75.73	63.98
Why Hate/Non-Hate Prompting									
Open Source LLM									
Qwen-2.5-7B	71.82	94.14	81.48	90.17	59.26	71.51	49.33	62.53	55.93
Gemma-3-12B	79.13	91.71	84.96	88.92	73.34	80.38	51.32	68.73	60.03
Llama-3.1-8B	84.48	77.04	80.59	76.92	84.39	80.48	50.15	67.03	58.59
Phi-4	76.54	91.41	83.31	87.95	69.13	77.41	46.58	66.47	56.53
Mistral	70.43	66.83	68.58	74.92	77.94	76.40	46.26	53.73	49.00
Close Source LLM									
Gemini 2.5	82.24	89.51	85.67	88.11	80.12	84.94	58.83	86.62	72.73
GPT 4o	86.30	91.71	**89.56**	91.69	86.12	88.14	**61.89**	**91.23**	**76.56**
Classification Guided LoRA Fine Tuning									
Qwen-2.5-7B	81.07	90.29	85.43	87.84	76.89	82.00	-	-	-
Gemma-3-12B	81.49	89.34	85.21	86.87	77.53	81.89	-	-	-
Llama-3.1-8B	84.31	87.67	85.96	85.71	81.94	83.83	-	-	-
Phi-4	79.42	90.73	84.70	87.86	74.04	80.36	-	-	-
Mistral	81.14	88.67	84.74	86.08	77.26	81.43	-	-	-
Explanation Guided LoRA Fine Tuning									
Qwen-2.5-7B	87.32	90.98	89.11	88.86	84.49	86.62	52.30	69.95	61.13
Gemma-3-12B	86.98	90.59	88.73	89.19	85.18	87.13	52.49	83.50	67.99
Llama-3.1-8B	88.86	89.13	88.99	87.97	87.68	87.83	54.26	66.78	60.52
Phi-4	84.42	91.29	87.72	91.07	84.06	87.43	52.58	65.49	59.04
Mistral	85.13	89.73	88.21	87.15	86.32	87.42	49.60	54.11	51.86

Table 5: Performance on the test split of BANHADEX: P, R, and BB_Score represent Precision, Recall, and BanglaBERT Score, respectively. Best performing model for each metric and configuration is highlighted in blue.

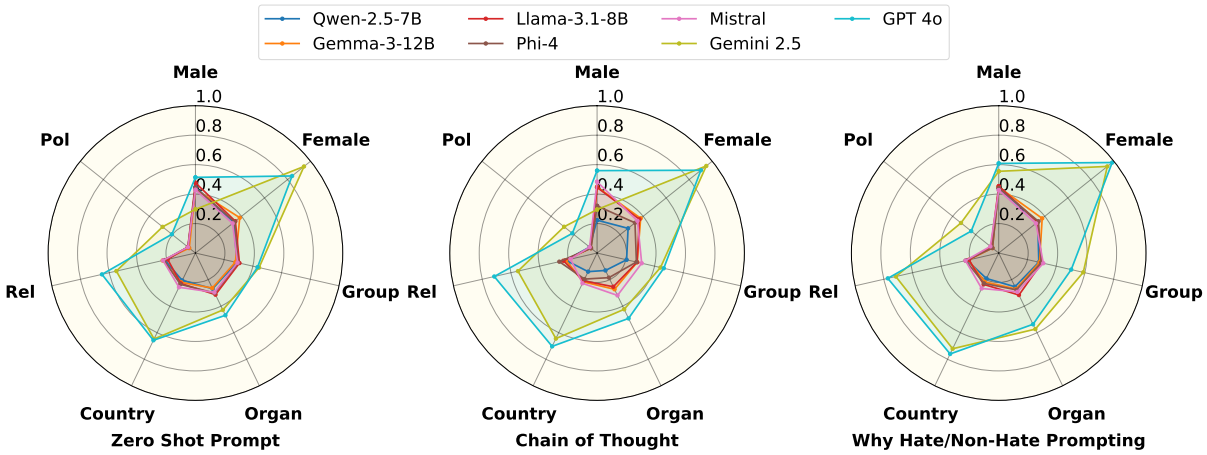


Figure 5: Performance on Target Group of BANHADEX Dataset.

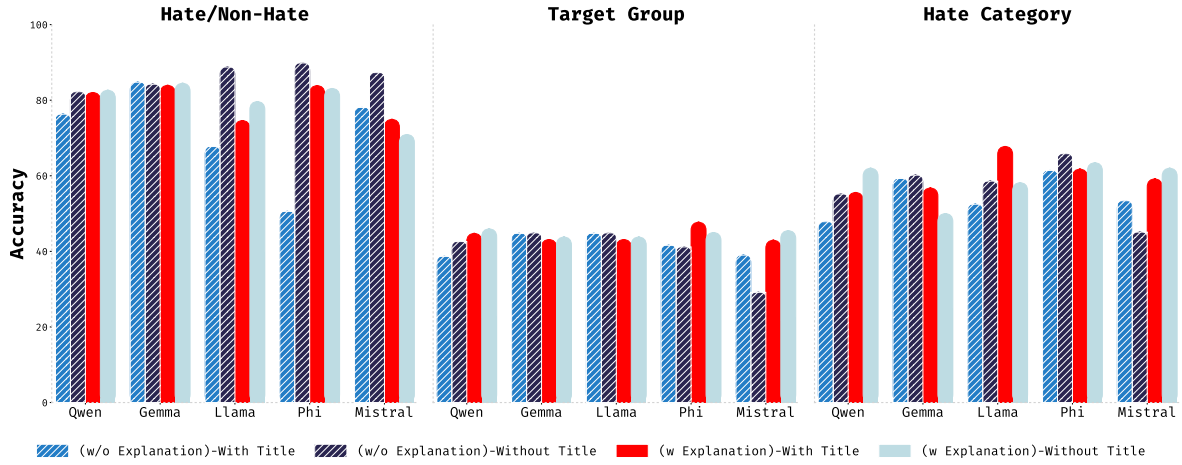


Figure 6: Impact of title inclusion and explanation generation on the performance of models.

Prompting strategies show limited gains for the Body Shaming category. However, the Why H/NH prompt substantially improves open-source model performance for Religious and Personal Offense detection. These findings suggest that, although models can reliably separate hate from non-hate content, accurately identifying specific hate categories remains challenging. This difficulty likely stems from the overlapping and nuanced nature of hate categories (§F.2). A detailed error analysis is provided in Appendix G.

Target Group-wise Results. We also examine model performance across different target groups, with results summarized in Figure 5 and detailed scores provided in Table 12. Overall, open-source models significantly lag behind closed-source models in accurately predicting the correct target group. Their performance remains relatively consistent across different prompting strategies, showing limited variation in behavior.

In contrast, closed-source models perform strongly on the Female target group and reasonably well on the Religious category. However, both open- and closed-source models struggle with the Political target group. Similar to hate category classification, these results indicate that while hate detection is effective, identifying the specific target group remains difficult, likely due to subtle and overlapping group representations in comments.

Role of Title and Explanation Generation. We conducted experiments to assess the impact of title inclusion in prompts and the effect of explanation generation. Following Fig. 6, we evaluated four settings in Zero-Shot: w/o explanation w title; w/o explanation, w title; w explanation, w title; and w explanation, w/o title. Including the title improves performance in zero-shot settings without explanation, particularly for classification metrics. However, when we ask the model to generate an

explanation, removing the title improves performance across both classification and explanation tasks. These findings indicate that titles are beneficial when providing explanatory context.

6 Discussion

6.1 Single Annotator Explanation

Our single-annotator explanation follows recent practices, *e.g.*, BanMiMe (Mia et al., 2025), where explanations were also authored by a single annotator during dataset creation. As explanations are inherently subjective and aim to provide plausible reasoning rather than a single ground-truth output, we ensured quality through detailed guidelines and additional expert evaluation (§B,D). We also found strong inter-annotator agreement scores (Tab. 9), suggesting that single-annotator explanation was not a meaningful source of variation.

6.2 Linguistic Factors of Bangla

There are multiple linguistic factors that put Bangla in a unique position for hate speech methods. Bangla belongs to the family of Indo-Aryan languages, while high-resource languages come from the Latin family. These languages generally follow the Subject-Object-Verb (SOV) structure, *e.g.* "Ami (Me) (S) Bhat (Rice) (O) Khai (Eat) (V)", whereas English follows the Subject-Verb-Object (SVO) structure, *e.g.*, "I (S) eat (V) rice (O)". Secondly, based on morphological complexity, Bangla can be considered an inflectional language, whereas English is an analytic language. Bangla has verbal inflections, *e.g.*, based on the honorific of the subject, the word 'eat' can have the forms: 'Kha' (colloquial) and 'Khan' (formal/respective). Similarly, noun inflections are also used to indicate possession or location, *e.g.*, 'Bari' means 'house', but 'Barite' means 'in the house'.

6.3 Sociocultural Factors

Bangla hate speech detection is challenging due to code-mixing and transliteration ("Banglish"), regional dialects (*e.g.*, Sylheti, Chittagonian), and context-dependent slurs rooted in historical, religious, or caste-based references. These factors create lexical ambiguity and subtle offenses that standard models often fail to capture.

6.4 Generalizability of BANHADEX

Insights from BANHADEX can extend to other low-resource Indo-Aryan languages like Hindi and Assamese, which share rich morphology, flexible word order, code-mixing, and culturally embedded expressions. Contextual reasoning, culturally grounded annotations, and lightweight fine-tuning strategies that are effective in Bangla are likely applicable to these languages, offering a replicable approach to explainable hate speech detection.

6.5 Practicality of BANHADEX

For an updated study on Bangla hate from a linguistic perspective, BANHADEX enables systematic analysis of how hate evolves over time. Hate evolves with shifts in vocabulary, coded language, rhetorical strategies, and targeted communities. An updated dataset will enable researchers to study hate in Bangla speech, potentially develop better content-moderation detection systems, and examine changes in hate-detection systems over time by benchmarking performance across eras.

BANHADEX supports the study of implicit versus explicit hate, emerging slang and euphemisms, and code-switching between Bangla and English, which are common in online discourse. It can also serve as a benchmark for evaluating generalization, cross-domain transfer, and the stability of detection systems as linguistic patterns evolve. This makes the dataset valuable not only for improving classification accuracy but also for advancing research on the long-term adaptability and reliability of hate speech detection in low-resource languages.

7 Conclusion

We introduced BANHADEX, the first Bangla hate-speech dataset with human-annotated explanations, spanning 19,203 YouTube comments across five content domains, seven hate categories, and seven target groups. Using a diverse suite of open- and closed-source LLMs across four prompting strategies, we showed that prompting strongly influences both detection and rationale quality; notably, the "Why H/NH" strategy benefits closed models. Beyond prompting, explanation-guided LoRA fine-tuning delivers consistent gains in both classification and explanation, producing the highest explainability scores among open models. We hope that our dataset and findings will serve as a valuable resource for future research on the interpretability of hate speech in low-resource languages such as Bangla.

Limitations

BANHADEX might be prone to distribution shifts due to the rapidly evolving landscape of online social media. The reproducibility of our results on closed-source models also depends on the availability of that model's variant and version. The metrics used to evaluate hate explanation are text-similarity metrics and are not specific to the hate-speech domain, which is a limitation of hate-explainability work in general.

To build effective real-world hate detection systems, it is important to account for the wide range of text forms, particularly transliterated (Fahim et al., 2024; Haider et al., 2025a) or code-mixed (Alam et al., 2025) content, in low-resource languages such as Bangla. Addressing this challenge presents a valuable direction for future research. Ahmed et al. (2024) explores various strategies for enhancing LLM performance on transliterated text.

Class Imbalance. We acknowledge the presence of class imbalance in BANHADEX. However, this imbalance reflects the natural distribution of hate speech instances in real-world data. Similar patterns have been observed in prior Bangla hate speech datasets, *e.g.*, BanTH (Haider et al., 2025b).

Human Evaluation. Evaluation of the current explanation relies primarily on automatic, text-similarity-based metrics, which may not fully capture faithfulness or helpfulness in the context of hate speech detection. Due to time and resource constraints, we were unable to conduct a human evaluation. However, a targeted human study assessing explanation faithfulness and interpretability would significantly strengthen our work.

Ethical Considerations

The comments used in our research were gathered from YouTube, in accordance with the site's API Terms of Service. To maintain users' privacy, all personally identifiable information (PII) was completely removed before creating the dataset. The annotating tasks were divided evenly among all participants, and the annotators and domain experts were paid on a per-sample basis at a rate higher than the current industry standard.

References

Fahim Ahmed, Md Fahim, Md Ashraful Amin, Amin Ahsan Ali, and AKM Rahman. 2024. Improving the performance of transformer-based models over classical baselines in multiple transliterated languages. In *ECAI 2024*, pages 4043–4050. IOS Press.

Md. Tofael Ahmed, Maqsudur Rahman, Shafayet Nur, Azm Islam, and Dipankar Das. 2021. Deployment of machine learning and deep learning algorithms in detecting cyberbullying in bangla and romanized bangla text: A comparative study. In *2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, pages 1–10.

Sadia Alam, Md Farhan Ishmam, Navid Hasin Alvee, Md Shahnewaz Siddique, Md Azam Hossain, and Abu Raihan Mostofa Kamal. 2025. BnSentMix: A diverse Bengali-English code-mixed dataset for sentiment analysis. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 68–77, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sai Suman Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep learning models for multilingual hate speech detection. *arXiv preprint arXiv:2004.06465*.

Dimosthenis Antypas and Jose Camacho-Collados. 2023. Robust hate speech detection in social media: A cross-dataset empirical evaluation. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 231–242.

Tasnim Tarannum Aurpa, Rifat Sadik, and Md Saiful Ahmed. 2022. Abusive bangla comments detection on facebook using transformer-based deep learning models. *Social Network Analysis and Mining*, 12(1):24.

Nayan Banik and Md Hasan Hafizur Rahman. 2019. Toxicity detection on bengali social media comments using supervised models. In *2019 2nd international conference on Innovation in Engineering and Technology (ICIET)*, pages 1–5. IEEE.

Tanveer Ahmed Belal, G. M. Shahariar, and Md. Hasanul Kabir. 2023. Interpretable multi labeled bengali toxic comments classification using deep learning. In *2023 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pages 1–6.

Fariha Tahosin Boishakhi, Ponkoj Chandra Shill, and Md. Golam Rabiul Alam. 2021. Multi-modal hate speech detection using machine learning. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 4496–4499.

Saugata Bose and Guoxin Su. 2022. Deep One-Class Hate Speech Detection Model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7040–7048, Marseille, France. European Language Resources Association.

Amit Kumar Das, Abdullah Al Asif, Anik Paul, and Md Nur Hossain. 2021. Bangla hate speech detection on social media using attention-based recurrent neural network. *Journal of Intelligent Systems*, 30(1):578–591.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Chiara Di Bonaventura, Lucia Siciliani, Pierpaolo Basile, Albert Merono Penuela, and Barbara Mcgillivray. 2024. Is explanation all you need? an expert survey on LLM-generated explanations for abusive language detection. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 280–288, Pisa, Italy. CEUR Workshop Proceedings.

Estiak Ahmed Emon, Shihab Rahman, Joti Banarjee, Amit Kumar Das, and Tanni Mittra. 2019. A deep learning approach to detect abusive bengali text. In *2019 7th International Conference on Smart Computing and Communications (ICSCC)*, pages 1–5.

Md Fahim. 2023. Aambela at blp-2023 task 2: Enhancing banglabert performance for bangla sentiment analysis task with in task pretraining and adversarial weight perturbation. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 317–323.

Md Fahim, Fariha Tanjim Shifat, Fabiha Haider, Deeparghya Dutta Barua, MD Sakib Ul Rahman Sourove, Md Farhan Ishmam, and Md Farhad Alam Bhuiyan. 2024. Banglatlit: A benchmark dataset for back-transliteration of romanized bangla. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14656–14672.

Vaibhav Goel, Dhruv Sahnan, Saptarshi Dutta, Anil Bandhakavi, and Tanmoy Chakraborty. 2023. Hatemongers ride on echo chambers to escalate hate speech diffusion. arXiv preprint arXiv:2302.02479. https://arxiv.org/abs/2302.02479.

Maya Gudka, Kirsty LK Gardiner, and Tim Lomas. 2023. Towards a framework for flourishing through social media: A systematic review of 118 research studies. *The Journal of Positive Psychology*, 18(1):86–105.

Fabiha Haider, Md Farhan Ishmam, Fariha Tanjim Shifat, Md Tasmim Rahman Adib, Md Fahim, and Md Farhad Alam Bhuiyan. 2025a. Robustness of LLMs to transliteration perturbations in Bangla. In *Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025)*, pages 338–346, Mumbai, India. Association for Computational Linguistics.

Fabiha Haider, Fariha Tanjim Shifat, Md Farhan Ishmam, Md Sakib Ul Rahman Sourove, Deeparghya Dutta Barua, Md Fahim, and Md Farhad Alam Bhuiyan. 2025b. Banth: A multi-label hate speech detection dataset for transliterated bangla. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7217–7236.

MD. Nahid Hasan, Kazi Shadman Sakib, Taghrid Tahani Preeti, Jeza Allohibi, Abdulmajeed Atiah Alharbi, and Jia Uddin. 2024. Olf-ml: An offensive language framework for detection, categorization, and offense target identification using text processing and machine learning algorithms. *Mathematics*, 12(13).

E. Hashmi and S. Y. Yayilgan. 2024. Multi-class Hate Speech Detection in the Norwegian Language Using FAST-RNN and Multilingual Fine-Tuned Transformers. *Complex Intelligent Systems*, 10:4535–4556.

Jawad Hossain, Avishek Das, Mohammed Moshiul Hoque, and Nazmul Siddique. 2023. Multilabel aggressive text classification from social media using transformer-based approaches. In *2023 26th International Conference on Computer and Information Technology (ICCIT)*, pages 1–6.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Refa Annisatul Ilma, Setiawan Hadi, and Afrida Helen. 2021. Twitter's hate speech multi-label classification using bidirectional long short-term memory (bilstm) method. In *2021 International Conference on Artificial Intelligence and Big Data Analytics*, pages 93–99.

Asim Irfan, Danish Azeem, Sanam Narejo, and Naresh Kumar. 2024. Multi-modal hate speech recognition through machine learning. In *2024 IEEE 1st Karachi Section Humanitarian Technology Conference (KHI-HTC)*, pages 1–6.

Khondoker Ittehadul Islam, Sudipta Kar, Md Saiful Islam, and Mohammad Ruhul Amin. 2021. Sentnob: A dataset for analysing sentiment on noisy bangla texts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3265–3271.

Sarif Sultan Saruar Jahan, Raqeebir Rab, Peom Dutta, Hossain Muhammad Mahdi Hassan Khan, Muhammad Shahariar Karim Badhon, Sumaiya Binte Hassan, and Ashikur Rahman. 2023. Deep learning based misogynistic bangla text identification from social media. *Computing and Informatics*, 42(4):993–1012.

Andreas M. Kaplan and Michael Haenlein. 2010. Users of the world, unite! the challenges and opportunities of social media. *Business Horizons*, 53(1):59–68.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. 2024. Improving automatic vqa evaluation using large language models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

Harshkumar Mehta and Kalpdrum Passi. 2022. Social media hate speech detection using explainable artificial intelligence (xai). *Algorithms*, 15(8).

Md Ayon Mia, Akm Moshiur Rahman Mazumder, Khadiza Sultana Sayma, Md Fahim, Md Tahmid Hasan Fuad, Muhammad Ibrahim Khan, and Akmmahbubur Rahman. 2025. Banmime: Misogyny detection with metaphor explanation on bangla memes. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17824–17850.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Paloma Piot and Javier Parapar. 2025. Towards efficient and explainable hate speech detection via model distillation. In *European Conference on Information Retrieval*, pages 376–392. Springer.

Faisal Hossain Raquib, Akm Moshiur Rahman Mazumder, Md Tahmid Hasan Fuad, Md Farhan Ishmam, and Md Fahim. 2025. BanHate: An up-to-date and fine-grained Bangla hate speech dataset. In *Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025)*, pages 237–248, Mumbai, India. Association for Computational Linguistics.

R Reghunathan and AS Asha. 2022. Hate speech detection in conventional language on social media by using machine learning. In *International Journal of Engineering Research & Technology (IJERT)*, volume 11.

Nasif Istiak Remon, Nafisa Hasan Tuli, and Ranit Debnath Akash. 2022. Bengali hate speech detection in public facebook pages. In *2022 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, pages 169–173.

Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, and Md Saiful Islam. 2021. Hate speech detection in the bengali language: A dataset and its baseline evaluation. In *Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2020*, pages 457–468. Springer.

Md. Saifuddin, Mohiuddin Ahmed, Spandan Basu, and Pritam Acharjee. 2023. Enhancing online safety: Natural language processing based multi-label cyberbullying classification in bangla. In *2023 26th International Conference on Computer and Information Technology (ICCIT)*, pages 1–6.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity bias in preference labeling by large language models. *ArXiv*, abs/2310.10076.

Mahmudul Hasan Shakil. 2022. *A hybrid deep learning model and explainable AI-based Bengali hate speech multi-label classification and interpretation*. Ph.D. thesis, Brac University.

Omar Sharif, Eftekhar Hossain, and Mohammed Moshiul Hoque. 2022. M-bad: A multilabel dataset for detecting aggressive texts and their targets. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, pages 75–85.

Sinéad Walsh and Paul Greaney. 2025. Multiclass hate speech detection with an aggregated dataset. *Natural Language Processing*, page 1–17.

Jessica E. Wendorf and Fan Yang. 2015. Benefits of a negative post: Effects of computer-mediated venting on relationship maintenance. *Computers in Human Behavior*, 52:271–277.

Wikipedia. 2025. Bengali language speaker statistics. `https://en.wikipedia.org/wiki/Bengali_language`. Accessed: July 27, 2025.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2025. Hallucination is inevitable: An innate limitation of large language models. *Preprint*, arXiv:2401.11817.

Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. 2023. Hare: Explainable hate speech detection with step-by-step reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5490–5505.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd ACL*, Bangkok, Thailand. Association for Computational Linguistics.

A Appendix

A.1 LLM Usage

We used large language models (LLMs) solely as general-purpose writing assistants during the final stage of manuscript preparation. Their use was limited to refining author-written text, including improving clarity, grammar, flow, and reducing redundancy. LLMs did not contribute to the research design, dataset construction, annotation process, experimental setup, model training, evaluation, analysis, or interpretation of results. No new technical content, claims, or citations were introduced through LLM assistance. All ideas, methodology, experiments, figures, analyses, and conclusions presented in this paper were entirely conceived, implemented, and validated by the authors. Any AI-assisted edits were carefully reviewed, verified, and, where necessary, revised by the authors, who take full responsibility for the accuracy, originality, and integrity of the manuscript.

Video Category	Hate	Non-Hate
Entertainment	2,007	1,762
International	1,720	1,296
News & Politics	3,801	2,541
People & Blogs	585	2,494
Sports	1,042	1,955

Table 6: Hate vs Non-Hate Comments by Video Category

Type	Total Count	Percentage (%)
Single	5,437	59.39%
Multiple	3,718	40.61%

Table 7: Hate Category Type Breakdown

B Annotation Guidelines

This section provides comprehensive annotation guidelines developed for the **BANHADEX** dataset, which comprises user-generated comments in Bengali collected from a range of online video content platforms.

Annotators were instructed to evaluate each comment to determine whether it constitutes hate speech. If so, they identified the specific target group, assigned one or more relevant hate categories, and provided a concise but contextually grounded explanation. Each comment was evaluated within its broader context, including metadata such as video title, publication date, and thematic category (e.g., News & Politics, Entertainment).

These annotation instructions were designed to:

- Ensure high inter-annotator agreement,
- Encourage objective and consistent judgments,
- Support robust models for automated hate speech detection in Bangla.

B.1 General Annotation Instructions

The annotation process followed a rigorous set of principled guidelines to ensure the creation of a high-quality, culturally grounded, and reproducible dataset.

- **Target-Oriented Annotation:** Each comment was annotated by identifying the target group(s) from a predefined list: *Male, Female, Group, Organization, Country, Religious, Politics*.
- **Hate-Label Classification:** Each comment could be assigned to a hate category (e.g., Religious, Gender, Body Shaming, Abusive/Violence) to reflect the complex overlaps in hate content.
- **Context-Aware Interpretation:** Annotators interpreted intent and implicit meaning, especially when sarcasm, metaphors, or culturally coded language was present.
- **Bias-Free and Objective Labeling:** Annotators maintained neutrality and judged based on harm, discrimination, or dehumanization rather than personal opinion.

Hate Category	Entertainment	International	News & Politics	People & Blogs	Sports
Abusive/Violence	571	787	1586	193	174
Body Shaming	94	5	34	15	4
Gender	730	3	201	97	9
Origin	105	304	256	64	46
Personal Offence	1146	662	1942	450	873
Political	46	328	1422	37	168
Religious	98	615	140	47	9

Table 8: Relationship Between Video Category and Hate Category

- **Annotation Redundancy and Verification:** Each comment was annotated by three individuals, with expert adjudication resolving conflicts.
- **Wellness-Oriented Work Environment:** Annotators took 5–10 minute breaks each hour and had access to light, humorous content to reduce cognitive fatigue.
- **Cultural Sensitivity and Localization:** Training emphasized identification of culturally specific hate forms like regional slurs, gendered insults, or religious undertones in Bangla discourse.

B.2 Binary Classification – Hate/Non-Hate

- **Hate Speech (H):** A comment should be labeled as Hate Speech if it contains offensive, abusive, or harmful language directed at identity groups or individuals based on gender, religion, nationality, ethnicity, politics, etc.
- **Non-Hate Speech (NH):** Comments that lack hostility, discrimination, or dehumanizing content—even if sarcastic or critical—are marked Non-Hate.

Annotation Steps:

1. Review the comment.
2. Classify as Hate Speech (H) or Non-Hate (NH).
3. If Non-Hate (NH), stop. No further labeling is required.
4. If Hate (H), proceed to Hate Speech Classification with Explanation.

B.3 Hate Speech Classification with Explanation

Annotators labeled each hate speech comment with one of the following categories. Each label must include a short explanation and clearly mention the associated target group.

C Additional Details of BANHADEX

C.1 Formal Description of Target Groups

In the context of hate speech annotation, target groups refer to the specific individuals, communities, or entities that are the direct focus of hostility, discrimination, or derogatory language. Proper identification of the target group is essential for understanding the intent and impact of hate speech. The categories below have been systematically defined to capture a wide range of targeted expressions relevant to Bangla sociocultural and political contexts.

C.2 Hate Speech Explanation Annotation

During the annotation phase, annotators were instructed to provide a brief explanation for each hate speech instance using a standardized justification format: Target Group + Hate Category + Explanation. This structure ensured consistent reasoning behind each label and allowed for the analysis of how hate manifests across different social dimensions.

In addition, annotators were encouraged to recognize emerging linguistic expressions, including slang, coded terms, and generational vocabulary popular among Gen Z and Gen Alpha users, to ensure that the dataset reflects contemporary and contextually grounded forms of hate speech in online Bangla discourse.

For example, in a sentence where a political leader is insulted, the annotation might read: Target Group: Politics, Hate Category: Personal Offense Explanation: The comment insults a specific political figure using abusive language and expresses contempt without policy-based critique.

For experimentation, we follow previous ablation from (Fahim, 2023) especially for the Bangla language. This structured and linguistically adaptive approach facilitated a systematic understanding of hate speech discourse in Bangla, revealing how different forms of hate, such as political hostility,

religious incitement, gender-based shaming, and abusive threats, are linguistically constructed and socially targeted.

D Data Validation

Data Validation for Comment Identification. Inter-annotator reliability for the primary classification task was assessed using Cohen's kappa coefficients, as presented in Table 9. The results demonstrate substantial agreement across all misogyny categories. Notably, while prior research in content moderation reported κ scores of approximately 0.53, indicating moderate agreement (Islam et al., 2021), our refined annotation guidelines and annotator selection criteria yielded higher agreement scores. These robust agreement metrics across multiple evaluation methods affirm the clarity of our taxonomy and the effectiveness of our annotation protocol. The rigorous annotation methodology ensures that the BANHADEX dataset establishes a reliable foundation for computational hate speech detection in Bangla text.

Type	Label	Kappa(κ)	Avg.
Primary	Hate	0.81	0.78
	Non Hate	0.75	
Hate Categories	Personal Offence	0.83	0.75
	Abusive/Violence	0.81	
	Political	0.74	
	Gender	0.77	
	Religious	0.72	
	Origin	0.67	
	Body Shaming	0.69	
Target Groups	Male	0.77	0.72
	Female	0.81	
	Group	0.66	
	Organization	0.69	
	Country	0.68	
	Religion	0.67	
	Politics	0.73	

Table 9: Inter-annotator agreement for the BANHADEX, measured using Cohen's kappa (κ) across the binary task, hate categories, and targeted groups.

Data Validation for Hate Explanation. To evaluate the quality and consistency of comment explanations, we randomly selected a subset of 200 samples for independent analysis by each annotator. The similarity between these annotator-generated explanations was assessed using multiple automated metrics. The explanations exhibited exceptional consistency, with high ROUGE scores (ROUGE-1(F1): 87.24%, ROUGE-2(F1): 53.82%, ROUGE-L(F1): 84.63%) demonstrating substantial agreement in both content structure and coverage. Complementary evaluation using BLEU (75.22%), BERT similarity (95.87%), and METEOR (82.56%) metrics further confirmed strong semantic coherence and lexical alignment across the annotators' interpretations, validating the reliability of our explanation annotation approach.

We also hired two more annotators to independently annotate the 200 samples and evaluated inter-annotator agreement for the explanation. Now, each explanation is annotated with three independent annotators. The average agreement scores between the annotators were 94.31%, 86.35%, and 96.74%, measured using Mean ROUGE-1 (F1), Mean ROUGE-2 (F1), and Mean ROUGE-L (F1), respectively. It indicates a high level of agreement among the annotators.

We further asked two domain experts with backgrounds in computational linguistics and content moderation to independently annotate the same 200 samples and evaluated inter-annotator agreement. Both experts had prior experience in analyzing offensive and hate-related content and were provided with detailed annotation guidelines before the task. The average agreement scores between the experts and annotators were 89.64%, 71.82%, and 90.21%, measured using Mean ROUGE-1 (F1), Mean ROUGE-2 (F1), and Mean ROUGE-L (F1), respectively.

E Evaluation Metrics

In our work, BanglaBERTScore follows the standard BERTScore (Zhang* et al., 2020) framework, where semantic similarity between generated and reference explanations is computed using contextual token embeddings and aggregated into precision, recall, and F1 scores. We use BanglaBERT as the pretrained encoder to obtain language-specific contextual representations, enabling more reliable semantic matching for Bangla explanations beyond surface-level overlap.

For LAVE (Mañas et al., 2024), we employ an LLM-based evaluation approach in which a large language model assesses explanation quality by comparing candidate and reference explanations and assigning a reasoning-aware rating. This allows evaluation based on semantic correctness and completeness rather than lexical similarity alone.

F Result Analysis on Hate Category and Target Group

F.1 Hate Category

We compare open and closed-source LLMs across prompting and fine-tuning, F1 as the primary metric. Among closed models, Gemini-2.5-Flash performs best, peaking at 66.77 average F1 with the Why H/NH prompt, followed by GPT-4o at 56.14 under the same setup. Gemini-2.5-Flash is especially strong on religion (81.44) and gender (73.17). The open models (Qwen 2.5 7B, Gemma 3 12B, Llama 3.18B) trail with the best averages around 30.5, typically when combining explanation-guided LoRA with HARE prompting; LoRA produces + 2–4 F1 over zero shot, but the gap remains. Body shame and origin are the hardest categories (often <5 F1). Prompt design can rival model size: well-designed prompts let Qwen 7B approach Gemma 12B, highlighting the value of rationale-based prompting for lightweight models in low-resource settings. In tbale 11, full analysis of the Hate Category classification is shown.

F.2 Target Group

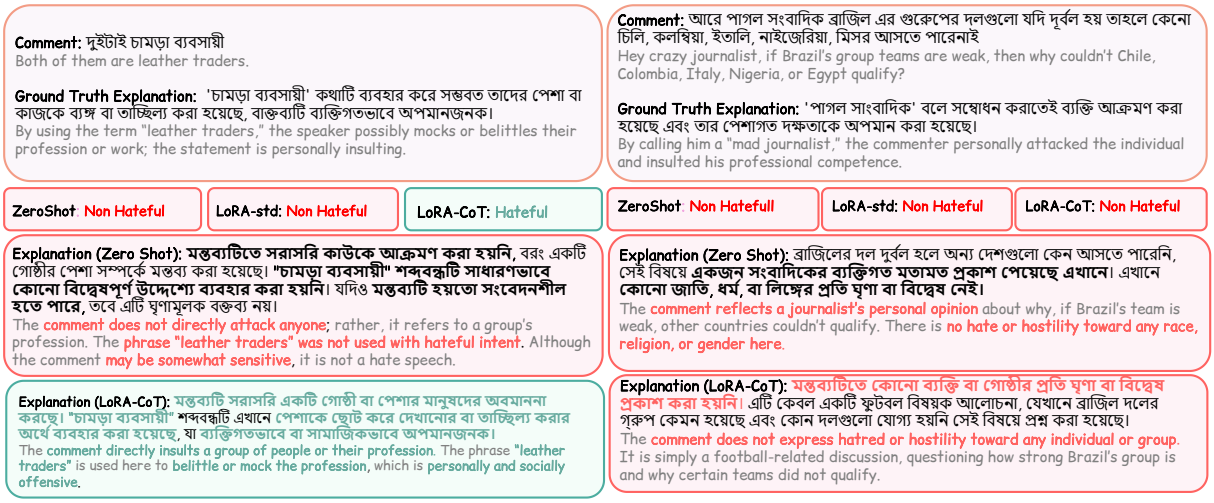
To assess how well LLMs detect hate speech across marginalized groups, we analyzed performance by target category. Closed-source models—GPT-4o and Gemini-2.5-Flash—consistently outperform open models, especially with the Why Hate/Non-Hate prompt. Open models like Mistral-7B and LLaMA-3.1-8B see modest improvements (27–28 F1) with LoRA fine-tuning, slightly outperforming zero-shot baselines. Prompting strategy plays a key role: Why H/NH consistently yields better group-wise discrimination, while performance on politically targeted hate remains low for open models. Interestingly, model size doesn't guarantee better results—Mistral-7B often beats larger models like Gemma-12B. These results highlight the value of rationale-based prompting and lightweight fine-tuning, while also pointing to the persistent performance gap between open and closed models. In Table 12, the full analysis for the target group is shown.

G Qualitative Error Analysis

This section presents a qualitative error analysis of the hate speech detection models across seven distinct categories: Personal Offense, Religious, Abusive, Political, Origin, Gender, and Body Shaming. This analysis is structured around four primary types of errors identified from the samples. By examining specific instances where the models succeeded or failed, we can gain a deeper understanding of their nuanced capabilities and limitations. Figures 7 and 8 demonstrate critical failures when detecting hate speech in Bangla text.

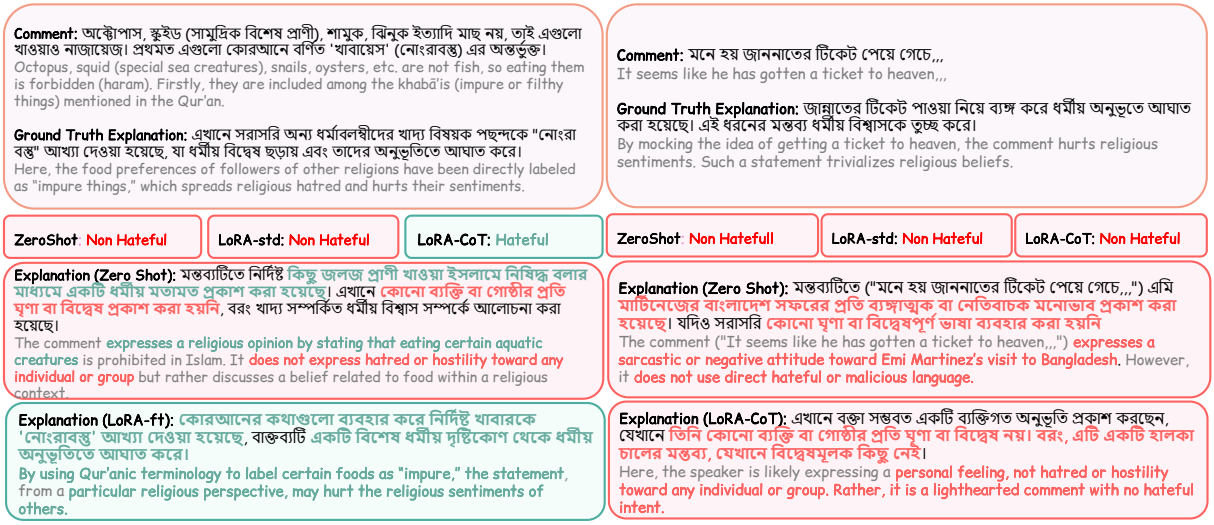
Contextual Overshadowing: In this category, the model's failure to detect hate from Mixed-Context Statements. A frequent error occurs when a comment contains both hateful and non-hateful elements, causing the non-hateful context to overshadow the abusive component. For instance, in the $Abusive$ category (7a), the comment *Hey crazy journalist, if Brazil's group teams are weak...* was labeled non-hateful because the models focused on the overarching topic of a football discussion, completely ignoring the direct, abusive insult *crazy journalist*. A more severe example of this is found in the $Political$ category (8d) with the statement, *Long live Pakistan, long live Bangladesh, may India be destroyed.* The models incorrectly classified this as a non-hateful political opinion. They latched onto the positive, nationalistic slogans while failing to recognize that the call for a nation's destruction constitutes an extreme and violent form of hate speech. This pattern suggests that the models struggle to parse comments with mixed signals, allowing malicious content to be concealed within seemingly benign statements.

Misinterpretation of Indirect Hate Speech: The models consistently failed to detect hate speech that relies on sarcasm, innuendo, or indirect insults. For example, within the $Religious$ category (7c), the sarcastic comment *It seems like he has gotten a ticket to heaven* was misclassified as a lighthearted or personal feeling. The models were unable to grasp the mocking and trivializing nature of the comment, which belittles a core religious belief. Similarly, in the $Gender$ category (8f, a sexually suggestive and demeaning proposal (*Babu, can you give my ex a room??*) was misinterpreted as a *normal request* for help. The models operated on a literal interpretation, completely missing the disrespectful and objectifying implication. This highlights a significant weakness in understanding the pragmatic and subtextual layers of human language, where the hateful intent is implied rather than explicitly stated.



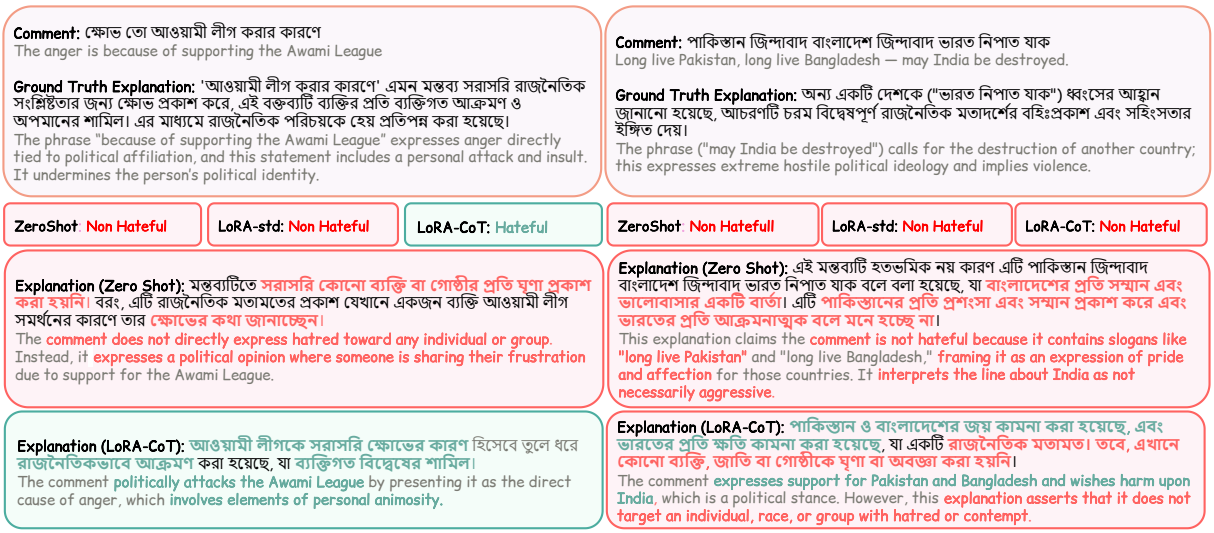
(a) Abusive/Violence

(b) Personal Offense

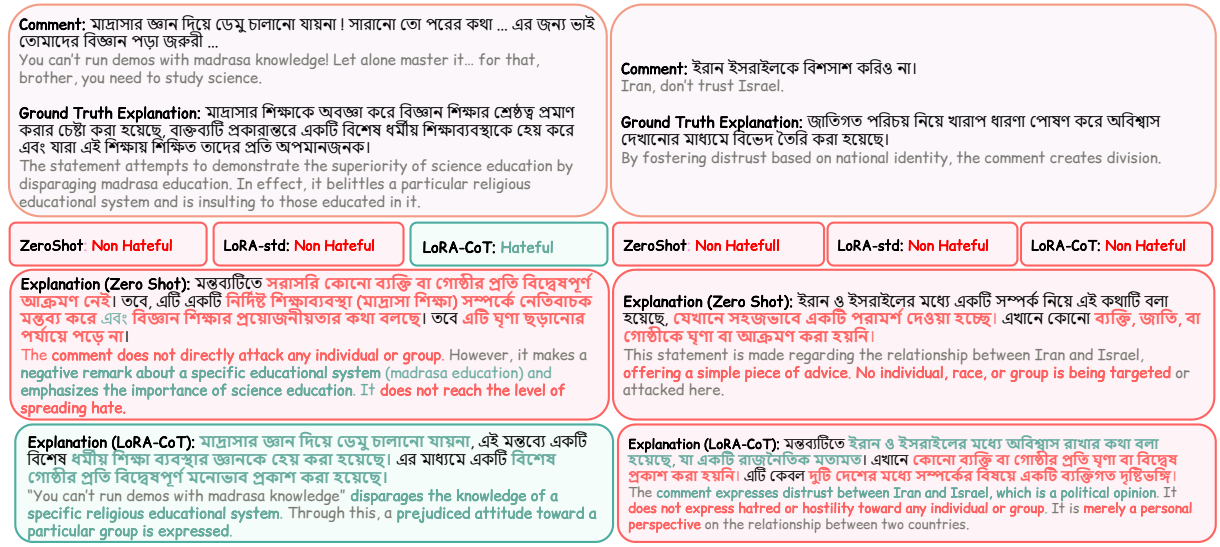


(c) Religious

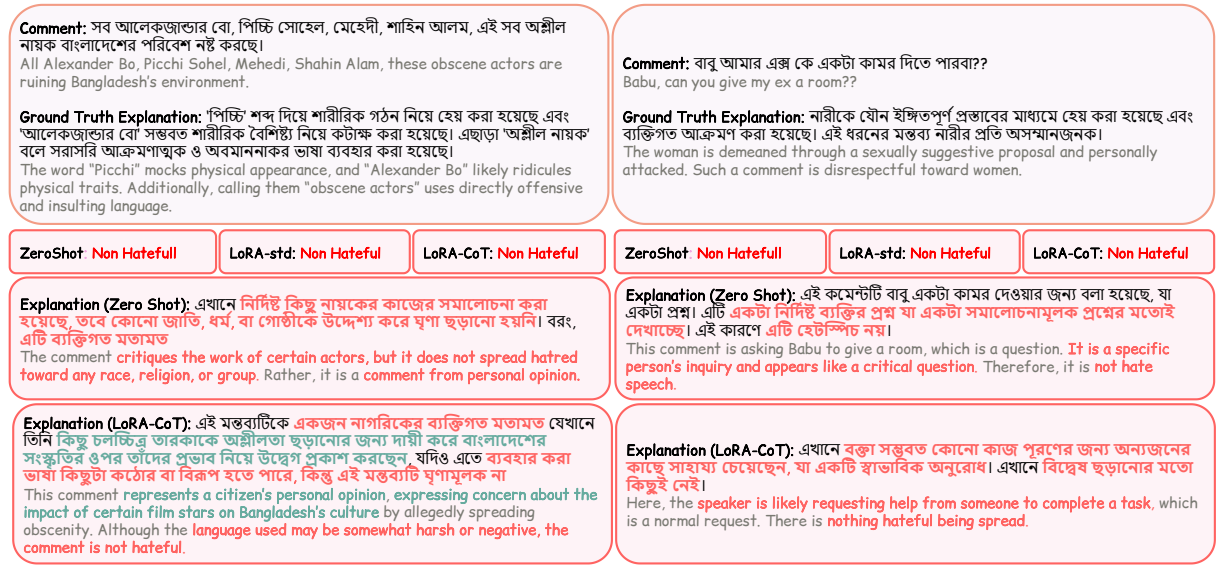
Figure 7: Comparative error analysis of finetuned (*ft*) and pretrained models on hate speech detection across multiple content categories. Accurate parts are bold and colored in green, while errors are highlighted in red.



(d) Political



(e) Origin



(f) Body Shaming & Gender

Figure 8: Comparative error analysis of finetuned (*ft*) and pretrained models on hate speech detection across multiple content categories. Accurate parts are bold and colored in green, while errors are highlighted in red.

Models	Non-Hate			Hate			Macro			Acc
	P	R	F1	P	R	F1	P	R	F1	
Zero Shot										
Qwen-2.5-7B (w/o)	80.12	85.30	82.63	82.51	76.62	79.45	81.31	80.96	81.04	81.17
Qwen-2.5-7B (w)	66.37	**93.94**	77.78	**87.69**	47.53	61.65	77.03	70.74	69.71	71.86
Gemma-3-12B (w/o)	79.17	90.57	84.48	87.64	73.73	80.08	83.40	82.15	82.28	82.56
Gemma-3-12B (w)	88.85	80.69	**84.57**	80.67	88.83	**84.55**	**84.76**	**84.76**	**84.56**	**84.56**
Llama-3.1-8B (w/o)	82.02	77.38	79.63	76.06	80.91	78.41	79.04	79.14	79.02	79.04
Llama-3.1-8B (w)	96.54	47.26	63.46	62.85	98.14	76.62	79.70	72.70	70.04	71.49
Phi-4 (w/o)	81.42	84.86	83.10	82.49	78.65	80.53	81.96	81.75	81.81	81.91
Phi-4 (w)	81.28	83.66	82.46	81.39	78.76	80.06	81.34	81.21	81.26	81.33
Mistral (w/o)	81.58	62.72	70.92	67.25	84.38	74.85	74.41	73.55	72.88	73.03
Mistral (w)	**96.79**	13.79	24.14	51.12	**99.50**	67.54	73.96	56.64	45.84	54.53
LoRA Fine-Tuning										
Qwen-2.5-7B (w/o)	87.32	90.98	89.11	88.86	84.49	86.62	88.09	87.73	87.87	88.03
Qwen-2.5-7B (w)	89.20	88.16	88.68	86.85	87.99	87.42	88.03	88.08	88.05	88.08
Gemma-3-12B (w/o)	86.98	90.60	88.73	89.19	85.18	87.13	88.09	87.89	87.93	88.01
Gemma-3-12B (w)	**90.73**	90.06	90.39	88.91	**89.65**	89.28	89.82	89.85	89.84	89.87
Llama-3.1-8B (w/o)	88.86	89.13	88.99	87.97	87.68	87.83	88.42	88.41	88.41	88.44
Llama-3.1-8B (w)	87.84	91.29	89.53	89.75	85.78	87.72	88.70	88.70	88.70	88.79
Phi-4 (w/o)	84.42	91.29	87.72	91.07	84.06	87.43	87.74	87.67	87.57	87.57
Phi-4 (w)	86.16	90.92	88.48	89.11	83.57	86.25	87.64	87.25	87.37	87.46
Mistral (w/o)	87.32	90.98	89.11	88.86	84.49	86.62	88.09	87.73	87.87	87.99
Mistral (w)	90.13	**94.11**	**92.08**	**93.02**	88.40	**90.65**	**91.58**	**91.25**	**91.36**	**91.42**

Table 10: Comparison of model performance with(w) and without(w/o) metadata under Zero Shot and LoRA fine-tuning settings. P, and R represent Precision, Recall. Best performing model for each metric and configuration is highlighted in blue.

Failure to Recognize Incitement: A critical and dangerous error was the model's inability to identify incitement to violence and threats of public harm. In a case from the *Personal Offense* category (7a), the comment *If the death penalty is not given, the people of Bangladesh will take to the streets* is a clear threat of civil unrest intended to intimidate the justice system. However, all models misclassified this as a non-hateful *demand for severe punishment*, overlooking the implied violence and creation of public fear. This type of error is particularly concerning as it involves a failure to flag language that could lead to real-world harm.

Systemic Blind Spots: Gender and appearance based hate are the most consistent and absolute failures were observed in cases of gender and body Shaming. The analysis indicates these are not just areas of weakness but systemic blind spots for the models. The comment *All Alexander Bo, Picchi Sohel...* from the *Body Shaming* category (8f) used insulting terms like *Picchi* to mock physical appearance and *obscene actors* to insult their profession. Despite these clear instances of personal offense, all models classified the comment as a non-hateful personal opinion. Likewise, misogynistic content, as seen in the sexually suggestive proposal from the *Gender* category (8f), was universally missed. This complete failure across all models suggests that they have not been adequately trained to recognize the specific linguistic patterns and social contexts associated with gender-based discrimination, misogyny, and body shaming, making these critical areas for future model development.

H Prompts

Models	Hate Category							
	P Off	A/V	Pol	Gen	Rel	Ori	BS	Avg
Zero Shot Prompt								
Open Source LLM								
Qwen-2.5-7B	55.59	51.22	34.8	20.71	21.68	13.24	2.62	28.55
Gemma-3-12B	56.76	55.37	30.31	24.03	22.81	14.65	3.81	29.68
Llama-3.1-8B	67.76	51.29	35.47	20.44	19.99	14.52	3.16	30.38
Phi-4	61.76	51.92	35.01	20.52	22.38	15.34	2.66	29.94
Mistral	59.17	50.51	36.65	19.68	23.07	15.73	2.64	29.64
Close Source LLM								
Gemini 2.5	61.35	68.29	46.71	62.88	75.42	43.19	3.65	51.64
GPT 4o	33.41	56.78	60.52	53.09	68.66	42.37	21.18	47.86
Chain of Thought								
Open Source LLM								
Qwen-2.5-7B	36.57	25.51	14.21	18.28	22.98	10.61	2.48	18.66
Gemma-3-12B	59.26	54.68	31.98	23.13	21.96	14.45	3.9	29.91
Llama-3.1-8B	59.11	50.4	29.78	23.06	23.88	14.9	3.26	29.20
Phi-4	53.04	36.48	21.07	22.27	25.58	14.66	2.44	25.08
Mistral	70.65	50.17	35.57	20.57	19.35	14.17	3.4	30.55
Close Source LLM								
Gemini 2.5	61.24	68.49	46.78	62.36	75.88	43.02	03.45	51.66
GPT 4o	36.81	60.27	65.39	57.12	73.55	45.80	22.91	51.22
HARE Prompting								
Open Source LLM								
Qwen-2.5-7B	60.08	50.31	35.03	20.08	21.21	14.55	2.67	29.13
Gemma-3-12B	67.14	50.51	36.64	21.08	21.43	14.36	2.85	30.57
Llama-3.1-8B	67.14	50.51	36.64	21.08	21.43	14.36	2.85	30.57
Phi-4	55.11	53.03	32.93	20.56	25.07	16.02	2.33	29.29
Mistral	68.69	50.8	35.99	20.73	20.18	14.38	2.76	30.50
Close Source LLM								
Gemini 2.5	69.34	76.82	57.29	73.17	81.44	53.03	10.80	60.27
GPT 4o	25.14	49.73	53.48	44.66	62.05	33.27	19.39	41.10
Why Hate/Non-Hate Prompting								
Open Source LLM								
Qwen-2.5-7B	53.35	50.18	32.88	20.91	22.45	13.09	2.88	27.96
Gemma-3-12B	58.96	55.0	31.8	23.23	21.14	14.73	3.79	29.81
Llama-3.1-8B	64.72	51.32	36.66	20.48	21.28	14.71	2.92	30.30
Phi-4	58.49	52.14	33.03	20.64	23.63	14.83	2.78	29.36
Mistral	60.5	51.72	37.38	19.15	23.72	16.39	2.53	30.20
Close Source LLM								
Gemini 2.5	80.37	75.92	59.44	74.85	85.21	65.60	21.03	66.77
GPT 4o	39.18	65.49	71.90	62.54	79.33	49.66	24.80	56.14
Classification Guided LoRA Fine Tuning								
Qwen-2.5-7B	64.26	51.62	32.08	21.45	20.19	13.43	3.51	29.51
Gemma-3-12B	64.26	51.39	32.54	21.73	19.88	13.76	3.01	29.51
Llama-3.1-8B	65.74	50.41	33.19	21.78	19.79	14.37	2.89	29.74
Phi-4	62.49	52.46	31.88	21.58	20.57	13.45	2.32	29.25
Mistral	68.21	51.69	34.85	21.86	18.75	14.67	3.32	30.48
Explanation Guided LoRA Fine Tuning								
Qwen-2.5-7B	66.55	51.29	35.29	22.95	19.59	14.84	2.35	30.41
Gemma-3-12B	67.04	51.65	35.19	21.83	19.65	14.42	3.02	30.40
Llama-3.1-8B	67.63	51.02	35.24	21.52	19.61	14.24	3.31	30.37
Phi-4	65.77	52.59	35.37	21.0	20.65	13.93	3.11	30.35
Mistral	64.56	51.09	33.0	22.09	18.98	13.8	3.33	29.55

Table 11: Performance Analysis of the models for Hate Category. Here, P Off, A/V, Pol, Gen, Rel, Ori, and BS refer to Personal Offense, Abusive/Violence, Political, Gender, Religious, Origin, and Body Shaming, respectively.

Models	Target Group							
	Male	Female	Group	Organ	Country	Rel	Pol	Avg
Zero Shot Prompt								
Open Source LLM								
Qwen-2.5-7B	44.76	33.35	28.85	26.87	20.66	21.38	6.41	26.04
Gemma-3-12B	43.14	38.49	27.83	26.36	22.09	22.81	5.37	26.58
Llama-3.1-8B	47.71	34.66	30.65	31.36	22.91	19.33	6.56	27.60
Phi-4	44.2	34.04	29.74	30.19	23.56	21.87	6.45	27.15
Mistral	42.95	32.27	29.31	29.66	25.78	22.54	6.45	26.99
Close Source LLM								
Gemini 2.5	29.67	94.25	44.31	42.89	64.78	55.04	28.53	51.35
GPT 4o	51.37	83.91	43.22	46.89	65.74	65.08	20.46	53.81
Chain of Thought								
Open Source LLM								
Qwen-2.5-7B	22.31	26.91	20.57	13.11	14.02	21.86	6.48	17.89
Gemma-3-12B	44.87	37.87	27.9	27.32	21.68	21.83	5.37	26.69
Llama-3.1-8B	44.56	36.74	27.87	25.39	20.97	23.47	5.47	26.35
Phi-4	32.02	32.91	27.39	18.54	19.62	26.29	5.22	23.14
Mistral	48.39	35.03	31.08	31.93	22.81	18.52	6.48	27.75
Close Source LLM								
Gemini 2.5	29.53	94.82	44.17	42.38	64.45	55.08	28.62	51.79
GPT 4o	55.94	90.12	46.44	49.26	70.31	71.58	21.47	57.33
HARE Prompting								
Open Source LLM								
Qwen-2.5-7B	46.2	33.04	29.14	28.42	22.6	20.8	5.81	26.57
Gemma-3-12B	45.0	37.72	28.64	27.01	21.64	21.53	5.47	26.72
Llama-3.1-8B	47.0	34.65	29.1	32.24	23.04	20.61	6.51	27.59
Phi-4	41.61	33.48	30.24	26.68	23.66	24.5	5.85	26.57
Mistral	47.07	34.41	31.27	31.25	23.38	19.43	6.67	27.64
Close Source LLM								
Gemini 2.5	43.16	95.74	56.48	51.02	70.37	64.29	25.66	58.10
GPT 4o	38.59	71.42	35.03	35.84	53.91	60.27	21.68	45.25
Why Hate/Non-Hate Prompting								
Open Source LLM								
Qwen-2.5-7B	43.94	33.9	27.91	25.4	19.18	22.13	5.78	25.46
Gemma-3-12B	44.8	37.73	28.38	27.08	21.76	21.13	5.43	26.62
Llama-3.1-8B	45.38	33.78	29.95	31.63	23.62	20.69	6.72	27.40
Phi-4	44.07	34.52	29.5	27.76	22.88	23.08	5.44	26.75
Mistral	42.69	32.31	31.04	29.26	26.69	23.16	7.25	27.49
Close Source LLM								
Gemini 2.5	55.47	94.29	58.86	57.31	72.09	71.64	32.75	63.12
GPT 4o	60.83	98.54	50.48	53.67	76.03	77.22	23.91	62.40
Classification Guided LoRA Fine Tuning								
Qwen-2.5-7B	46.50	36.03	29.51	31.14	22.16	19.27	6.18	27.26
Gemma-3-12B	47.24	35.78	30.05	31.50	21.6	19.27	6.38	27.40
Llama-3.1-8B	47.38	36.08	30.12	31.23	21.26	19.15	6.38	27.37
Phi-4	47.17	34.89	29.82	32.53	20.68	20.14	6.79	27.43
Mistral	47.91	36.13	30.77	31.09	22.30	18.5	6.17	27.55
Explanation Guided LoRA Fine Tuning								
Qwen-2.5-7B	48.19	35.15	28.8	30.92	18.46	19.54	6.2	26.75
Gemma-3-12B	47.60	35.12	29.08	30.25	19.79	19.46	6.18	26.78
Llama-3.1-8B	48.22	35.27	28.65	30.61	20.71	19.28	6.46	27.03
Phi-4	47.47	34.14	29.11	29.03	20.39	20.56	5.83	26.65
Mistral	46.51	35.91	29.75	30.42	19.6	18.46	6.25	26.70

Table 12: Results of the models on the Target Group section of the test split from the BANHADEX dataset. Rel and Pol refer to Religious and Politics, respectively.

Zero Shot Prompt Used for Prediction

You are given a Bangla text. Determine if it is "Hate" or "Non Hate".
If it is hate speech, identify the category of hate and the target group, choosing **only from the following options**:

Hate_Category:

```
['Abusive/Violence',    'Body        Shaming','        Gender',            'Origin',
'Personal Offence', 'Political', 'Religious']
```

Target_Group:

```
['Country',     'Female',     'Group',   'Male',     'Organization',     'Origin',
'Politics', 'Religious']
```

Return your answer in this strict JSON format:

```
{
  "first_answer":"Hate"or"Non Hate",
  "justification":"Brief explanation
   in Bangla",
  "final_answer":"Hate"or"Non Hate",
  "hate_category": [ ... ],
  "target_group": [ ... ]
}
```

Chain of Thought Prompt Used for Prediction

You are tasked with detecting hate speech in Bangla text using a step-by-step reasoning approach.

Follow these steps carefully:

1. Understand the **literal and implicit meaning** of the Bangla text.
2. Analyze the **tone and intent**: Is the speaker expressing hostility, discrimination, or inciting hatred?
3. Look for **contextual cues**: Are there slurs, derogatory phrases, targeted groups, or implied harm?
4. Make a **final decision** based on your analysis.
5. **Explain** your reasoning briefly in Bangla.

Now, respond using the following strict JSON format:

```
 {
"thought_process": "Your step-by-step reasoning in English",
"is_hate": "Hate" or "Non Hate",
"explanation": "Explanation in Bangla"
}
```

Output format (strict):

```
{
   "score": <number between 0 and 1>,

}
```

Constraints:

- Treat the **Actual explanation** as ground truth; do not reinterpret or revise it.
- Base your judgment solely on alignment with the **Actual explanation**, not on your own views.
- Keep the rationale concise ($\leq$ 1–2 sentences).

Why Hate/Non-Hate Prompt Used for Prediction

You are an expert in Bangla language with a deep understanding of cultural and contextual expressions of hate speech. Your task is to analyze a given Bangla text and determine:

1. Determine if the given comment is hate or non hate.
2. Justify and debate over your first answer why hate or non hate and then finalize the answer.
3. If hate: determine the hate speech category (from the fixed list below), and the target group that the hate is directed toward (also from a fixed list).

Choose only from the following:

Hate_Category:

```
['Abusive/Violence', 'Body Shaming','Gender', 'Origin',
'Personal Offence', 'Political', 'Religious']
```

Target_Group:

```
['Country', 'Female', 'Group', 'Male', 'Organization', 'Origin',
'Politics', 'Religious']
```

Return your answer in the following format:

```
{
  "first_answer": "Hate"or"Non Hate",
  "justification": "...",
  "final_answer": "Hate"or"Non Hate",
  "hate_category": [ ... ],
  "target_group": [ ... ]
}
```

Hate Explanation prompt

Role: You are an evaluator who compares an AI-generated explanation against a fixed, ground-truth explanation.

Task:
Given:
1. a **Comment** (the original text),
2. an **Actual explanation** (the authoritative ground truth that must not be changed), and
3. a **model-generated explanation** (the candidate to evaluate), rate how well the AI explanation matches the Actual explanation on a continuous scale from **0 to 1**.

What to consider (compare to the Actual explanation only):

- **Conclusion alignment:** Does the AI reach the same overall judgment (e.g., hate/non-hate, reasoning)?
- **Factual accuracy:** Is it faithful to the facts in the Actual explanation (no additions, distortions, or mistranslations)?
- **Reasoning overlap:** Does it cover the key reasons given in the Actual explanation?
- **Tone & scope:** Avoids introducing new claims beyond the Actual explanation or omitting critical points.

Scoring guide:

- **1.00** – Fully aligned: same conclusion and reasoning; no errors.
- **0.75–0.99** – Minor issues (e.g., small wording/mistranslation nuances) but overall aligned.
- **0.50–0.74** – Partially aligned; misses or adds notable elements.
- **0.01–0.49** – Largely misaligned; major omissions/additions or wrong reasoning.
- **0.00** – Contradicts the Actual explanation or is unrelated.