BANHATE: An Up-to-Date and Fine-Grained Bangla Hate Speech Dataset

Faisal Hossain Raquib[1] *, Akm Moshiur Rahman Mazumder[2] *, Md Tahmid Hasan Fuad[3], Md Farhan Ishmam[4], Md Fahim[2,5]

[1]*Rajshahi University of Engineering and Technology*
[2]*Center for Computational & Data Sciences, IUB*
[3]*University of Manitoba* [4]*University of Utah* [5]*Penta Global Limited*

*Equal Contribution **Correspondence:** {faisal.ece18,fahimcse381}@gmail.com

Abstract

Online safety in low-resource languages relies on effective hate speech detection, yet Bangla remains critically underexplored. Existing resources focus narrowly on binary classification and fail to capture the evolving, implicit nature of online hate. To address this, we introduce BANHATE, a large-scale Bangla hate speech dataset, comprising 19,203 YouTube comments collected between April 2024 and June 2025. Each comment is annotated for binary hate labels, seven fine-grained categories, and seven target groups, reflecting diverse forms of abuse in contemporary Bangla discourse. We develop a tailored pipeline for data collection, filtering, and annotation with majority voting to ensure reliability. To benchmark BANHATE, we evaluate a diverse set of open- and closed-source large language models under prompting and LoRA fine-tuning. We find that LoRA substantially improves open-source models, while closed-source models, such as GPT-4o and Gemini, achieve strong performance in binary hate classification, but face challenges in detecting implicit and fine-grained hate. BANHATE sets a new benchmark for Bangla hate speech research, providing a foundation for safer moderation in low-resource languages. Our dataset is available at: https://huggingface.co/datasets/aplycaebous/BanHate.

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

1 Introduction

Social media has revolutionized human communication, making unprecedented transformations in connecting people and relaying information (Kaplan and Haenlein, 2010). With a user adoption rate of more than 50% and daily spending time exceeding two 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 trend of hate in mainstream social networks (Goel et al., 2023). Hateful comments can be theorized as means of externalizing internal upheaval and venting bottled-up emotions (The Havok Journal, 2023), with users preferring their native language for emotional salience (Reghunathan and Asha, 2022).

As of 2025, Bangla is spoken natively by over 240 million people (Wikipedia, Ethnologue, 2025), making it one of the most critical languages for online hate moderation. Nevertheless, studies reveal a persistent gap in this domain. While several datasets have been developed for Bengali hate speech detection (Sharif et al., 2022; Romim et al., 2020), they fail to capture the rapidly evolving nature of online discourse, particularly the generational shifts in expressions introduced by Gen Z and Gen Alpha. This necessitates the creation of new, updated Bangla datasets.

Recently, large language models (LLMs) have shown impressive performance in a range of classification tasks, including hate speech detection (Haider et al., 2025). However, the increasing subtlety and implicitness of hateful content continue to pose a challenge. Current models often struggle to infer the underlying intent of such speech and fall short of expectations in detecting implicit hate (Kim et al., 2022).

To bridge this gap, we introduce BANHATE, a hate speech dataset constructed from recent YouTube comments. The dataset reflects the linguistic styles of newer generations and captures implicit forms of hate rarely represented in earlier resources. All comments were carefully validated by human annotators to ensure that they contain genuinely targeted speech. Furthermore, we evaluate a suite of LoRA-based fine-tuned models alongside closed-source LLMs on this dataset to assess their effectiveness in identifying implicit hate.

Datasets	#NH	#H	H:NH	Data Source	#Annotators
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
Romim et al. (2020)	20,000	10,000	0.50	YouTube, Facebook	50
BANHATE (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). The hate speech datasets include similar data classes, *e.g.*, aggression and toxic speech.

2 Related Work

Early research on hate speech detection predominantly treated it as a binary classification problem (Djuric et al., 2015; Badjatiya et al., 2017; MacAvaney et al., 2019). Subsequent works expanded to multi-class (Walsh and Greaney, 2025; Hashmi and Yayilgan, 2024) and multi-label formulations (Ilma et al., 2021), enabling more nuanced representations of hate across social, political, and cultural dimensions. Recent efforts have also explored multimodal hate speech detection, integrating textual and visual modalities (Boishakhi et al., 2021; Barua et al., 2024), as well as multilingual and cross-lingual models leveraging architectures such as mBERT to extend detection across languages (Aluru et al., 2020; Ousidhoum et al., 2019).

Despite these advancements, Bangla remains underexplored compared to high-resource languages. Early studies primarily addressed binary hate speech detection (Remon et al., 2022; Das et al., 2022), reflecting the scarcity of linguistic resources. Subsequent research broadened the scope to related domains, including abusive language detection (Aurpa et al., 2022; Emon et al., 2019), cyberbullying (Ahmed et al., 2021; Saifuddin et al., 2023), gender discrimination and sexism (Jahan et al., 2023), and toxic content classification (Belal et al., 2023). More recent efforts have advanced toward multi-label settings, categorizing hate by target domains such as religion, politics, and gender (Sharif et al., 2022; Haider et al., 2025; Romim et al., 2020).

3 The BANHATE Dataset

The creation of BANHATE (**Ban**gla **Hate** Speech Detection) dataset was systematically collected and annotated to address the scarcity of resources for hate speech detection in low-resource languages. It comprises 19,203 curated comments covering seven target groups and seven hate categories, enabling fine-grained analysis (Figure 1).

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 3: Hate distribution by video category.

3.1 Data Collection

We collected comments from 328 YouTube videos across five categories: News & Politics, Entertainment, People & Blogs, International, and Sports, spanning April 2024 to June 2025. Only top-level comments were extracted, resulting in 26,730 comments.

3.2 Data Filtering

We removed non-Bangla comments as our dataset's focus is on Bangla hate speech detection, and comments shorter than 20 characters were removed as they lack meaningful context. By filtering, we reduced the dataset to 20,439 comments.

3.3 Data Cleaning

We removed duplicate comments and extraneous content, *e.g.*, URLs, hashtags, and personal identifiers, resulting in 19,203 clean comments for data annotation.

Year	Hate	Non-Hate
2017	105	70
2019	31	364
2020	265	197
2021	58	313
2022	466	1,553
2023	970	1,774
2024	2,780	2,693
2025	4,480	3,084

Table 4: Hate distribution against video upload year.

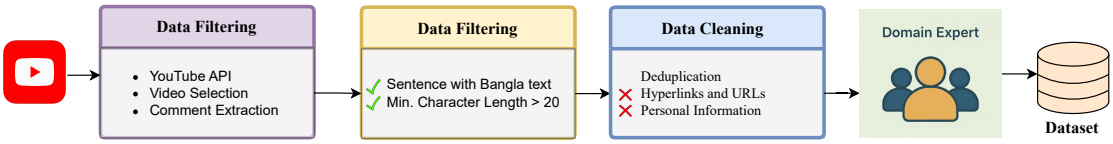


Figure 1: BANHATE dataset development pipeline illustrating the four-stage process: (a) data collection from social media platforms, (b) data filtering to remove non-relevant content, (c) data cleaning to eliminate duplicates and extraneous elements, and (d) data annotation & validation.

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 2: Relationship Between hate and video categories.

3.4 Data Annotation

We adopted a two-stage approach for data annotation. First, four native Bangla-speaking undergraduate annotators labeled each comment as Hate or Non-Hate. Their prior experience with social media usage ensured high-quality annotations that captured nuanced hate content. Second, hate comments were further categorized by target group and type of hate. The final labels were determined via majority vote. The annotators were provided monetary compensation. The annotation guidelines provided to the annotators have been reported in Appendix A.

3.5 Data Validation

We evaluated inter-annotator agreement using Cohen's kappa (κ), shown in Table 5. For the primary Hate vs. Non-Hate task, we obtained κ scores of 0.81 (Hate) and 0.75 (Non-Hate), averaging 0.78, which is substantially higher than prior work on content moderation ($\sim$0.53) (Islam et al., 2021), reflecting effective annotator selection and clear guidelines.

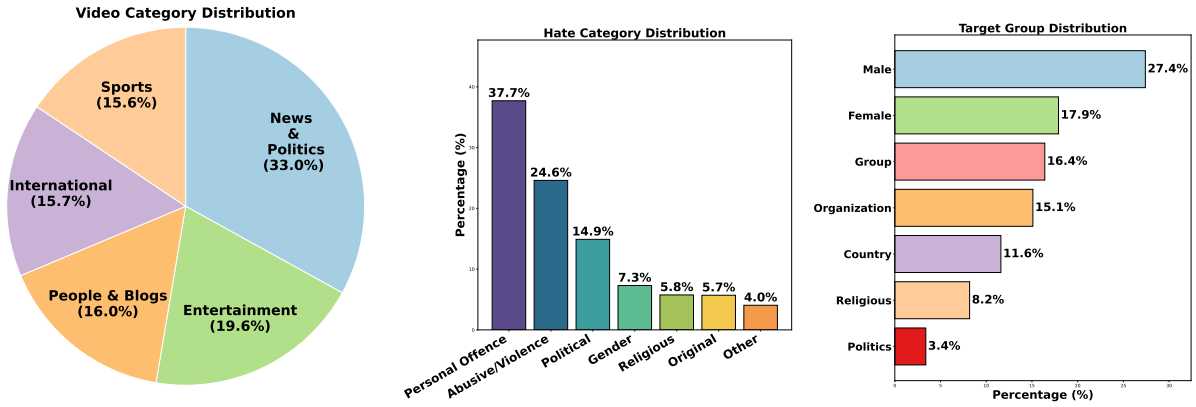
For hate categories, agreement was highest for Personal Offense (0.83) and Abusive/Violent (0.81), where definitions are clearer, and lower for more subjective categories such as Origin (0.67) and Body Shaming (0.69). Among target groups, Female (0.81) and Male (0.77) showed strong agreement, while Group (0.66), Country (0.68), and Religion (0.67) were lower, suggesting that identifying hate toward specific collectives is inherently more challenging and may benefit from additional guidance.

	Label	Kappa(κ)	Avg.
Primary	Hate	0.81	0.78
	Non Hate	0.75	
Hate Categories	Personal Offence	0.83	0.75
	Abusive/Violance	0.81	
	Political	0.74	
	Gender	0.77	
	Religious	0.72	
	Origin	0.67	
	Body Shaming	0.69	
Targeted 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 5: Inter-annotator agreement for the BANHATEdataset, measured using Cohen's kappa (κ) across the binary task, hate categories, and targeted groups, with averages indicating substantial reliability.

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

Table 8: Distribution of single and multiple hate labels in BANHATE.



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

Figure 2: Distribution of different categories in BANHATE.

Splits	
Train	15362
Test	3841
General Statistics	
Samples	19203
Videos	318
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

Table 6: Dataset statistics of BANHATE.

3.6 Dataset Statistics

Table 6 presents the key statistics of our BANHATE dataset. Table 7 summarizes the major incidents represented in the dataset, spanning July 2024 to June 2025. The most prevalent hate categories observed across these events are Abusive/Violent, Personal Offense, and Political. Figure 2 illustrates the distributions of video categories (Figure 2a), hate categories (Figure 2b), and target groups (Figure 2c). The data reveal that the majority of comments originate from News & Politics videos. Personal Offense and Abusive/Violence account for over 60% of all hate-labeled comments.

4 Experiment Design

We selected a diverse set of LLMs and classified our experimental designs into two categories: (i) Prompt-based Experiments and (ii) LoRA Fine-tuning Experiments.

4.1 Prompt-Based Experiments

For the prompt-based experiments, we considered zero-shot and chain-of-thought prompting (Wei et al., 2022). Details of prompts used in the experimentation are given in the Appendix B.

4.2 LoRA Fine-Tuning

We fine-tuned open-source LLMs using LoRA (Hu et al., 2022), which adapts pre-trained weights efficiently via low-rank updates rather than updating all parameters. Given a pre-trained and frozen weight matrix W, LoRA adds a learnable low-rank update $\Delta W = AB^\top$, 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: $W' = W + \Delta W = W + AB^\top$.

This formulation allows 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 defined 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. In addition, we introduce a hierarchical loss $\mathcal{L}_{\text{Hier}}$ to jointly model hate category and target group detection, reflecting the dependency between coarse-grained and fine-grained labels. The hierarchical loss is defined as:

$$\mathcal{L}_{\text{Hier}} = \mathcal{L}_{\text{Cat}} + \mathcal{L}_{\text{GrouplCat}},$$

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

Target Group	A/V	Body Shaming	Gender	Origin	P Off	Political	Religious
Country	587	0	0	583	444	317	147
Female	528	90	947	39	1264	132	71
Group	874	18	75	159	861	269	67
Male	1034	54	112	98	2293	457	115
Organization	632	0	119	75	824	899	33
Politics	182	0	11	21	79	240	24
Religious	310	1	21	100	209	225	792

Table 9: Target Group & Hate Category Relation. A/V refers to Abusive/Violence and P Off refers to Personal Offence respectively.

where $\mathcal{L}_{\text{Cat}}$ is the cross-entropy loss for hate category prediction, $\mathcal{L}_{\text{Group|Cat}}$ is the conditional loss for target group prediction given the predicted category. We configured LoRA with $\alpha = 64$, $r = 64$, a dropout rate of 0.01, a learning rate of 1×10^{-4}, and a batch size of 4 to 32. LoRA was applied to all weight matrices of the pre-trained models, and each model was fine-tuned for a single epoch. For inference, we used VLLM (Kwon et al., 2023), while LLaMA-Factory (Zheng et al., 2024) was used for LoRA fine-tuning. To ensure reproducibility, an evaluation is performed using greedy decoding with a temperature of 0 and no sampling.

4.3 Baselines

We evaluate five open-source models: Qwen-2.5-7B (Qwen et al., 2025), Gemma-3-12B (Team et al., 2025), Llama-3.1-8B (Dubey et al., 2024), Phi-4 14B (Abdin et al., 2024), and Mistral 7B (Jiang et al., 2023), and two closed-source models: Gemini 2.5 Flash and GPT 4o.

5 Result and Analysis

5.1 Hate Speech Detection

Table 12 compares five open-source LLMs on Precision, Recall, and F1 across three setups: Zero-Shot, Chain-of-Thought (CoT), and LoRA Fine-Tuning.

5.1.1 Zero-Shot Prompting

Gemma-3 12B leads with the highest F1 for Non-Hate (84.76%) and Hate (80.64%), showing strong generalization capabilities. Mistral 7B achieved high Precision (Hate 97.08%) and Non-Hate Recall (97.62%) but low Hate Recall (60.15%), resulting in a moderate F1 (74.28). Phi-4 had the highest Hate Recall (82.58%) but extremely low Precision, yielding an F1 of 18.27%. LLaMA-3.1 8B and Qwen-2.5 7B showed average performance, with Qwen favoring recall over precision.

5.1.2 Chain-of-Thought (CoT) Prompting

Gemma-3 12B performed consistently, reflecting its robustness in the CoT setting. LLaMA-3.1 8B improved notably compared to zero-shot prompting, reaching Non-Hate F1 of 80.94%, highlighting substantial gains from explicit reasoning. Mistral showed an imbalance: high non-hate precision (93.91%) and hate recall (98.14%), but non-hate recall fell to 26.17%, thus, reducing hate F1 to 40.94%. Qwen-2.5 7B also degraded performance with hate F1 dropping to 33.99%.

5.1.3 LoRA Fine-Tuning.

LoRA fine-tuning yields the most consistent and substantial performance gains across all models. LLaMA-3.1 8B achieves the highest F1 scores for both Non-Hate (85.96%) and Hate (83.83%) classes, demonstrating effective task adaptation.

Models	Hate Category (F1)							Overall Metrics			
	P Off	A/V	Pol	Gen	Rel	Ori	BS	Micro	Macro	Subset Acc	Hamming
Zero Shot Prompting											
Qwen-2.5-7B	17.28	28.47	41.10	22.22	49.66	21.65	16.13	27.43	28.07	14.25	21.13
Gemma-3-12B	**66.43**	**50.05**	44.41	39.04	**64.29**	25.45	08.94	**51.25**	**42.66**	**27.25**	20.29
Llama-3.1-8B	61.36	40.06	42.54	**46.94**	40.82	04.85	**33.47**	17.40	15.96	02.08	**40.29**
Phi-4 14B	54.69	37.49	**49.78**	17.19	56.57	**28.24**	08.89	43.92	36.12	21.57	22.81
Mistral 7B	34.01	02.59	36.91	18.57	24.88	15.06	06.56	23.28	19.80	01.80	39.13
Chain of Thought (CoT)											
Qwen-2.5-7B	18.18	29.06	10.55	**29.83**	53.38	12.97	04.60	23.35	22.65	09.63	19.35
Gemma-3-12B	52.79	**58.78**	**48.57**	28.88	**61.27**	21.00	**17.50**	**51.46**	**41.26**	**14.34**	19.17
Llama-3.1-8B	53.45	51.19	34.42	10.53	54.86	**34.06**	06.56	45.69	35.01	10.84	19.17
Phi-4 14B	15.59	44.33	21.58	27.14	57.74	25.81	13.04	29.72	29.32	10.07	19.60
Mistral 7B	**67.26**	48.45	11.79	03.69	19.47	17.17	08.89	44.10	25.25	01.42	**30.88**
LoRA Fine-Tuning											
Qwen-2.5-7B	59.93	52.58	48.91	54.34	68.41	35.51	00.00	55.03	45.67	63.11	08.12
Gemma-3-12B	**63.81**	**55.36**	53.60	**56.78**	**71.33**	40.01	13.63	**58.58**	50.65	**64.79**	10.23
Llama-3.1-8B	63.68	54.14	**54.29**	55.21	70.24	**40.51**	**23.26**	58.12	**51.62**	62.46	08.34
Phi-4 14B	60.74	51.47	51.13	47.35	67.21	36.36	15.00	55.00	47.04	60.46	08.56
Mistral 7B	60.97	51.71	50.46	53.20	67.89	35.21	11.76	55.55	47.31	42.30	**12.81**

Table 10: Performance Analysis of the models on Hate Category. P Off, A/V, Pol, Gen, Rel, Ori, and BS refer to Personal Offence, Abusive/Violence, Political, Gender, Religious, Origin, and Body Shaming, respectively. Color marks the highest performance for each configuration and metric.

Models	Target Group							Overall			
	Male	Female	Group	Org	Country	Rel	Pol	Micro	Macro	Subset Acc	Hamming
Zero Shot Prompting											
Qwen-2.5-7B	06.92	18.47	10.55	21.33	35.97	26.67	**33.87**	19.65	21.97	**11.08**	18.49
Gemma-3-12B	**54.71**	67.84	**34.25**	34.17	**50.25**	**43.68**	06.83	**38.66**	**41.68**	00.54	38.02
Llama-3.1-8B	46.46	54.70	30.28	35.05	40.96	39.16	11.98	36.76	36.94	00.81	38.22
Phi-4 14B	32.69	**67.90**	31.38	**40.45**	39.77	31.86	11.31	35.86	36.48	04.97	32.76
Mistral 7B	00.00	20.41	32.47	04.96	38.78	26.73	10.84	23.38	19.17	00.93	**38.71**
Chain of Thought											
Qwen-2.5-7B	17.89	35.41	29.12	06.81	37.88	39.39	04.65	26.89	24.45	15.59	18.41
Gemma-3-12B	06.18	**57.04**	32.53	15.56	33.70	41.35	09.20	28.75	27.94	03.07	27.41
Llama-3.1-8B	**48.31**	45.13	34.01	09.62	42.33	44.40	05.97	**37.98**	32.82	17.57	22.14
Phi-4 14B	33.74	27.90	**37.47**	**37.52**	**47.46**	**60.62**	**14.29**	37.88	**37.00**	**22.71**	22.14
Mistral 7B	02.95	01.03	05.85	01.12	09.89	22.32	05.08	11.38	06.89	05.14	**25.69**
LoRA Fine Tuning											
Qwen-2.5-7B	61.65	68.28	44.91	57.14	63.83	72.96	20.59	59.92	55.62	46.91	11.98
Gemma-3-12B	65.76	**71.54**	**47.13**	59.99	66.38	77.42	18.09	**63.17**	**58.04**	48.81	**13.68**
Llama-3.1-8B	**65.87**	70.79	45.35	58.84	64.93	**77.87**	11.59	62.41	56.46	46.70	11.38
Phi-4 14B	61.81	68.75	41.02	54.21	61.49	76.02	**22.73**	58.90	55.15	45.21	12.21
Mistral 7B	65.10	68.63	44.63	**61.10**	**66.50**	74.85	**22.73**	58.90	55.15	**50.06**	11.65

Table 11: Performance Analysis of the models on Target Groups. Org, Rel, and Pol refer to Organization, Religion, and Political, respectively. Color marks the highest performance for each configuration and metric.

Under this setting, all models exhibit improved balance between precision and recall. Qwen-2.5 7B and Gemma-3 12B also deliver strong results, both surpassing the 85% F1 threshold for the Non-Hate class and exceeding 81% in the Hate class.

5.2 Hate Category Detection

Table 10 shows the performance across different hate categories, the previous three settings.

5.2.1 Zero-Shot Prompting

Gemma-3 12B delivers the strongest overall performance, with the highest F1 scores across most hate categories, including Personal Offense (66.43%), Abusive/Violence (50.05%), and Religious (64.29%), as well as leading in both macro (42.66%) and micro (51.25%) averages. LLaMA-3.1-8B shows good performance in the Gender (46.94%) and Body Shaming (33.47%) cat-

Models	Non-Hate			Hate		
	P	R	F1	P	R	F1
Zero Shot Prompt						
Qwen-2.5-7B	64.52	93.39	76.31	85.64	43.45	57.61
Gemma-3-12B	**79.74**	90.47	**84.76**	87.66	74.66	**80.64**
Llama-3.1-8B	57.22	82.98	67.74	83.39	57.94	68.37
Phi-4 14B	35.42	88.44	50.58	71.76	**82.58**	18.27
Mistral 7B	65.05	**97.62**	78.08	**97.08**	60.15	74.28
Gemini 2.5	68.65	96.23	80.42	93.06	56.34	69.72
GPT 4o	73.15	95.27	83.10	93.22	65.14	77.26
Chain of Thought						
Qwen-2.5-7B	58.25	98.41	73.18	92.10	20.84	33.99
Gemma-3-12B	79.36	90.12	**84.40**	87.19	74.17	**80.15**
Llama-3.1-8B	75.45	87.29	80.94	83.06	68.69	75.19
Phi-4 14B	63.42	**96.36**	76.50	90.64	38.78	54.32
Mistral 7B	**93.91**	26.17	40.94	54.77	**98.14**	70.31
Gemini 2.5	70.88	96.04	82.68	93.10	57.82	70.53
GPT 4o	74.39	96.04	83.46	**93.17**	68.82	79.27
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 14B	79.42	**90.73**	84.70	**87.86**	74.04	80.36
Mistral 7B	81.14	88.67	84.74	86.08	77.26	81.43

Table 12: Benchmarking results of open and closed source models on the test split of BANHATE. P, R, and F1 represent Precision, Recall, and F1 Score, respectively. Color marks the highest performance for each configuration and metric.

egories but exhibits lower overall accuracy. Phi-4 14B performs well in Political (49.78%) and Origin (28.24%) hate categories, though its performance remains inconsistent across other categories. Qwen-2.5-7B and Mistral 7B record comparatively weaker results, particularly in maintaining subset accuracy and categorical coherence.

5.2.2 Chain-of-Thought (CoT) Prompting

The performance in this configuration is more uneven than in previous setups. Gemma-3 12B again leads with consistently strong results across multiple hate categories. LLaMA-3.1 8B shows targeted gains in Origin (34.06%) and Religious (54.86%) hate detection, though its subset accuracy remains unstable. Mistral 7B excels in Personal Offence (67.26%), highlighting category-specific strength but limited generalization. Qwen-2.5 7B and Phi-4 14B show overall declines across most metrics under this configuration.

5.2.3 LoRA Fine-Tuning

Similar to the hate speech detection results, fine-tuning delivers the most consistent and substantial gains across all models. Gemma-3 12B and LLaMA-3.1 8B clearly dominate, leading in most categories and overall metrics. LLaMA-3.1 8B excels in Political (54.29%), Origin (40.51%), and Body Shaming (23.26%) detection, while Gemma-3 12B shows strong proficiency in Personal Offence (63.81%), Abusive/Violent (55.36%), Gender (56.78%), and Religious (71.33%) hate. All models achieve higher macro/micro F1 and accuracy, highlighting the clear advantage of task-specific fine-tuning. Mistral shows improved subset classification via a higher Hamming score but still lags behind in overall accuracy.

5.3 Target Group Detection

Table 11 reports the performances of the models in different target groups, and overall evaluation metrics in three settings: Zero-Shot, Chain-of-Thought (CoT), and LoRA Fine-Tuning.

5.3.1 Zero Shot Prompting

Gemma-3 12B delivers the strongest overall performance, leading most categories: Male (54.71%), Country (50.25%), Group (34.25%), and Religion (43.68%), while also achieving the highest macro (41.68%) and micro (38.66%) F1 scores. LLaMA-3.1 8B and Phi-4 14B follow, with Phi-4 showing exceptional strength in Organization (40.25%) and Female (67.90%) detection. Mistral 7B and Qwen-2.5 7B rank lower overall, though Qwen-2.5 attains top accuracy in select subsets.

5.3.2 Chain-of-Thought (CoT) Prompting

Phi-4 14B leads in multiple categories: Group (37.47%), Organization (37.52%), Country (47.46%), Religion (60.62%), and Political (14.29%), while attaining the highest macro-average (37.00%). LLaMA-3.1 8B shows consistent gains across categories and achieves the best score in the Male category (48.31%). Gemma-3 12B performs unevenly, excelling in the Female (57.04%) category but underperforming elsewhere. Mistral 7B consistently ranks last on all metrics, except for the hamming score.

5.3.3 LoRA Fine-Tuning

Similar to the previous analyses, all models show consistent improvement across categories. Gemma-3 12B and LLaMA-3.1 8B excel in the Female (71.54%), Group (47.13%), and Religion (77.87%) categories and Macro F1, Mistral 7B in the Organization (61.10%) and Country (66.50%) categories, and Subset Accuracy (50.06%). Phi-4 14B performs best in Political (22.73%) classification but trails the others in overall accuracy.

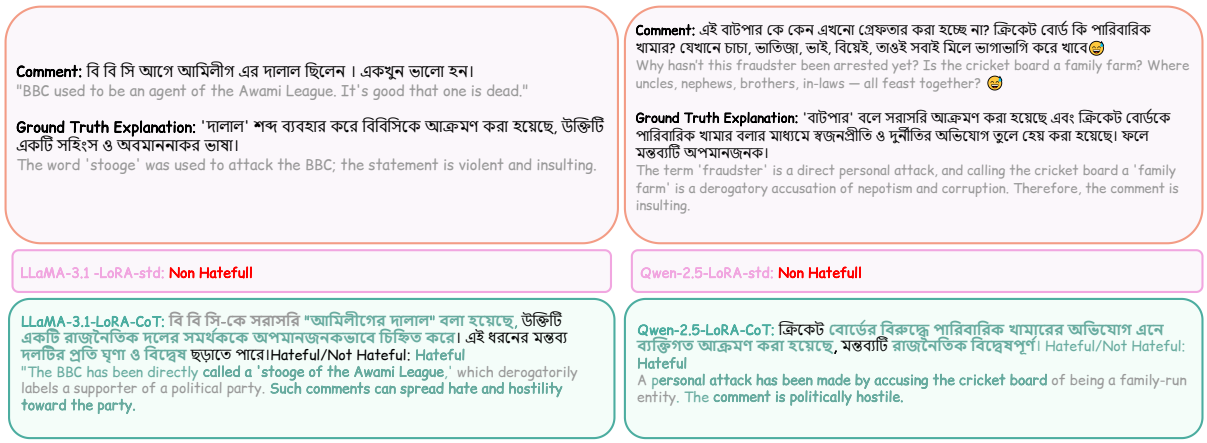


Figure 3: Error analysis of the LoRA finetuned models on Bengali hate speech detection. Detailed and accurate parts are emphasized in green and bold letters. The mistakes are highlighted in red.

5.4 Closed Source Models

Table 12 compares Gemini 2.5 Flash and GPT-4o under Zero-Shot and Chain-of-Thought (CoT) prompting for Hate vs. Non-Hate classification, using F1 scores as the primary metric. In the Zero-Shot Prompting, GPT-4o outperforms Gemini 2.5 with higher F1 scores in both Non-Hate (83.10%) and Hate (77.26%) categories. Gemini's strongest metric is its high Non-Hate recall (96.23%), indicating strong sensitivity, but with reduced precision. Gemini's subpar Hate recall weakens its overall classification performance. Both models improve performance in Chain-of-Thought (CoT) prompting, but GPT-4o maintains its lead.

6 Qualitative Error Analysis

We find several notable patterns and recurring errors in hate speech detection, *e.g.*, Gemma-3 frequently misclassifies *aggressive or threatening* discussions as *peaceful and constructive*. This could be due to insufficient contextual cues for violence or aggression in the training data. We also hypothesize that this effect of inadequacy is somewhat mitigated when the models are fine-tuned. Figure 3 errors that persist in fine-tuned models, allowing us to analyze beyond the training distributional biases.

In Figure 3, we observe LLaMA-3.1 and Qwen-2.5 models in zero-shot prompting misclassifies texts as non-hateful. However, the chain of thought prompting enables the models to correctly classify the texts as hateful while producing proper explanations against the claims. Hence, LoRA fine-tuning combined with chain of thought prompting minimizes the errors produced by the models.

7 Discussion: Pretraining Distribution

Variations in model performance are closely tied to the underlying distribution of the pretraining data. Models exposed to a larger amount of Bangla text should develop a better linguistic understanding, while insufficient data prevents the model from capturing the nuances of hate speech. The *nature* of the data can also play a critical role, *e.g.*, real-world and updated Bangla sources may generalize better than an outdated corpus in limited settings. Unfortunately, analyzing the attribution of performance variation to the training data is infeasible as none of the evaluated models publicly disclose their pretraining data. The opacity highlights a broader limitation of interpreting model behaviour in low-resource languages.

8 Conclusion

We present BANHATE, a Bangla hate-speech dataset spanning 19,203 YouTube comments from April 2024 to June 2025 across five content domains and seven hate categories/target groups. The dataset enables classification training and evaluation. Using a diverse suite of open and closed-source LLMs under zero-shot, chain-of-thought prompting, and LoRA fine-tuning strategies, we showed that prompting and fine-tuning strongly influence detection performance. LoRA fine-tuning delivers consistent gains in both hate and non-hate F1. We hope that our dataset and findings will be a valuable resource for future research on low-resource languages such as Bangla.

Limitations

BANHATE comprises only YouTube top-level comments collected between April 2024 and June 2025, limiting the validity of the dataset to a single platform only. Our evaluation reports results from a single LoRA epoch with greedy decoding, lacking a held-out development set or multi-seed runs.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

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.

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

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.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.

Deeparghya Dutta Barua, MSUR Sourove, Fabiha Haider, Fariha Tanjim Shifat, Md Farhan Ishmam, Md Fahim, and Farhad Alam Bhuiyan. 2024. Penta ml at exist 2024: tagging sexism in online multimodal content with attention-enhanced modal context. *Working Notes of CLEF*.

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.

Mithun Das, Somnath Banerjee, Punyajoy Saha, and Animesh Mukherjee. 2022. Hate speech and offensive language detection in Bengali. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 286–296, Online only. Association for Computational Linguistics.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

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.

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.

Shilpa Gudka, Felix Reer, and Thorsten Quandt. 2023. Towards a framework for flourishing through social media: A systematic review of 118 research studies. *Current Opinion in Psychology*, 53:101633.

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

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.

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.

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.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

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.

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

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*.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

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.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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.

Nahian Romim, Muhtasim Ahmed, Hasib Talukder, and Md Shamsul Islam. 2020. Hate speech detection in the bengali language: A dataset and its baseline evaluation. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*.

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.

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.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

The Havok Journal. 2023. The psychology behind social media hate. https://havokjournal.com/culture/the-psychology-behind-social-media-hate/.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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

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 Annotation Guidelines

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. 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 and encourage objective and consistent judgments.

A.1 General Annotation Instructions

- **Target-Orientation:** 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 can be assigned to one or multiple hate categories, *e.g.*, Religious, Gender, Body Shaming, Abusive/Violence.
- **Context-Aware Interpretation:** Annotators were instructed to interpret intent and implicit meaning, *e.g.*, when sarcasm, metaphors, or culturally coded language were used.
- **Bias-Free and Objective Labeling:** Annotators maintained neutrality and judged based on harm, discrimination, or dehumanization rather than personal opinion.
- **Annotation Redundancy and Verification:** Each comment was annotated by three individuals, with expert adjudication and discussions resolving conflicts.
- **Mental Well-being:** Annotators took 5–10 minute breaks each hour to reduce cognitive fatigue.
- **Cultural Sensitivity and Localization:** Annotators were guided to identify culturally specific hate forms, *e.g.*, regional slurs, gendered insults, or religious undertones in Bangla conversations.

A.2 Binary Classification – Hate/Non-Hate

- **Hate Speech (H):** A comment should be labeled as Hate Speech if it expresses prejudiced, offensive, abusive, or harmful language directed at identity, groups, or individuals based on gender, religion, nationality, ethnicity, political affiliation, and similar factors.
- **Non-Hate Speech (NH):** Comments that lack hostile, discriminatory, or dehumanizing content. This category may include critical or sarcastic expressions that do not constitute targeted attacks.

A.3 Hate Speech Classification

Annotators labeled each hate speech comment with one of the following categories.

Political: Abusive language or incitement of harm directed at political ideologies, parties, or figures.

Religious: Discriminatory or divisive attacks against religious groups, beliefs, or practices.

Gender: Hostility based on gender identity or stereotypes, including, but not limited to, sexism and transphobia.

Personal Offense: Degrading attacks or insults targeting an individual rather than a group.

Abusive/Violence: Explicit threats or incitement to physical violence or harm.

Body Shaming: Criticism or mockery of physical appearance, body shape, or disabilities.

B Prompt

All the prompts are detailed in this section.

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',
'Politics', 'Religious']
```

Return your answer in this strict JSON format:

```
{
  "first_answer":"Hate"or"Non Hate",
  "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",
}
```