Exploring Large Language Models for Multitask Learning in Bengali Text Classification

Md. Sajjad Hossain[1], **Kawsar Ahmed**[1], **Suny Md Ashraf Khan**[2]
Mohammed Moshiul Hoque[1]
[1]NLP Lab, Chittagong University of Engineering and Technology, Bangladesh,
[2]Infratech Consultants Ltd, Bangladesh
u1904031@student.cuet.ac.bd, ashraf@infraconsultants.com, moshiul_240@cuet.ac.bd

Abstract

Text classification in low-resource languages has become increasingly important due to the rapid growth of user-generated digital content. While multitask learning has long been studied in NLP, the use of LLMs for multitask text classification in low-resource languages such as Bengali remains underexplored. Although LLMs are inherently multilingual and multitasking, their effectiveness in structured multitask classification settings for Bengali has not been systematically evaluated. In this work, we investigate how LLMs can be leveraged for multitask Bengali text classification across five domains: sentiment analysis, aggressive text detection, fake news detection, news categorization, and emotion analysis. We compare in-context learning strategies—including zero-shot, one-shot, and chain-of-thought prompting—with parameter-efficient fine-tuning approaches. Our findings show that CoT prompting does not consistently improve performance and often degrades performance, highlighting the instability of prompt-based adaptation in low-resource settings with limited pretraining exposure. Moreover, reasoning-optimized models such as DeepSeek-R1 exhibit substantial performance drops, indicating that enhanced reasoning capabilities alone cannot overcome the challenges posed by low-resource settings. Among the evaluated mLLMs, Gemma-3-4B demonstrates the most stable and balanced cross-task performance under both in-context learning and parameter-efficient fine-tuning, making it a strong backbone candidate for multitask Bengali text classification. These results provide empirical evidence on the limitations of prompting and the advantages of lightweight fine-tuning for low-resource multilingual NLP.

Keywords: Natural Language Processing, LLMs, Low-Resource Languages, Multitask Text Classification

1. Introduction

With the rapid expansion of internet access and social media usage in Bangladesh and Bengali-speaking communities, large volumes of user-generated textual content are produced daily. This content often contains sentiment, misinformation, aggression, and emotionally expressive language, making automated text classification (TC) an essential tool for content moderation, information verification, and social media analysis. In recent years, several studies have addressed Bengali TC tasks individually, including sentiment analysis (Bhowmick and Jana, 2021), fake news detection (Farhad et al., 2024), emotion classification (Das et al., 2023), and aggressive text detection (Rosni et al., 2024). These works have contributed valuable task-specific datasets and modeling approaches. However, most existing research focuses on single-task settings, where models are trained independently for each classification objective. In contrast, multitask learning (MTL)—which enables a unified model to learn shared representations across multiple related tasks—remains relatively underexplored in Bengali NLP. Although multi-task text classification has witnessed significant advancements in high-resource languages such as English and Chinese, progress in low-resource languages—including Bengali—remains limited (Afroz et al.). The development of robust Bengali classification systems is hindered by the scarcity of standardized annotated corpora, limited domain-diverse datasets, insufficient pre-trained embeddings, and comparatively underdeveloped NLP infrastructure. These challenges are further compounded in multitask settings, where a single model must generalize across heterogeneous domains with varying linguistic and semantic complexities.

Recent advances in Large Language Models (LLMs) have transformed NLP by enabling deep contextual understanding, cross-lingual transfer, and strong zero-shot and few-shot performance. LLMs have demonstrated impressive results across various classification tasks and languages (Ding et al., 2023; Wang et al., 2025). Their multilingual pretraining suggests potential applicability to low-resource languages. Despite this promise, the effectiveness of LLMs for structured multitask text classification in Bengali remains largely unexplored. In particular, there is limited empirical evidence comparing different adaptation strategies—such as in-context learning and parameter-efficient fine-tuning—for Bengali classification tasks. Although modern LLMs are multilingual, their pretraining corpora are disproportionately dominated by high-resource lan-

Proceedings of the Second workshop on Challenges in Processing South Asian Languages (CHiPSAL2026) @ LREC, pages 92–104
16 May 2026.

guages. This imbalance often leads to suboptimal performance in low-resource contexts, where linguistic structures, morphology, cultural expressions, and script characteristics differ significantly from those of dominant languages. Understanding how different LLM architectures and adaptation techniques perform under these constraints is therefore essential for advancing multilingual NLP research.

In this work, we systematically investigate the use of LLMs for multitask Bengali text classification across multiple domains. By evaluating different adaptation paradigms and analyzing their task-specific behavior, we aim to provide empirical insights into the capabilities and limitations of LLMs in low-resource multilingual settings. The key findings of this work can be summarized as follows:

- The results demonstrate that Chain-of-Thought (CoT) prompting does not consistently enhance performance in Bengali text classification and, in several cases, leads to measurable degradation. This finding underscores the instability of prompt-based adaptation strategies in low-resource settings where pretraining exposure to the target language is limited.
- Reasoning-optimized large language models, such as DeepSeek-R1, exhibit significant performance deterioration across multiple tasks. This reveals a critical limitation of scaling reasoning capacity without sufficient linguistic grounding, particularly in low-resource language contexts.
- Among the evaluated models, Gemma-3-4B achieves the most stable and consistent performance across diverse tasks under both in-context learning and parameter-efficient fine-tuning. Its balanced cross-task generalization highlights its suitability as a backbone model for multitask Bengali text classification.

2. Related Work

There has been significant progress on TC tasks in high-resource languages (HRLs) using LLMs. In contrast, low-resource languages (LRLs) like Bengali have seen much less improvement.

2.1. Text Classification in HRLs

There are numerous ways of text classification using ML, DL, transformers, and LLMs. Most research on LLMs has focused on high-resource languages, such as English. Alex et al., 2021 proposed a few-shot text classification benchmark using GPT-3. They evaluated 11 tasks across 11 datasets in English. Schick and Schütze, 2022 also performed 11 tasks, as before, and used the same dataset for each task. They used Pattern-Exploiting Training (PET), which performed nearly as well as non-expert humans in 7 out of 11 tasks. Their model outperformed that of Alex et al., 2021 on several tasks. However, they could not improve the accuracy of some tasks, which appeared to be very low. Loukas et al., 2023 used context learning with GPT-3.5 and GPT-4 for classifying text of Banking77 datasets. This dataset comprises 77 classes and 13083 samples. Their approach outperformed fine-tuned, non-generative models even with fewer samples. They got an F1 score of 0.83. Wang et al., 2024 proposed a simplified approach for classifying text with LLMs (GPT-4, GPT-3.5, Gemini-pro, Llama-3-8b, etc) using zero and few-shot prompting. They demonstrated that zero-shot prompting is a more convenient approach for text classification than traditional methods (ML, DL, and Transformer models). Although several TC methods in HRLs have demonstrated promising results, these models cannot be used for resource-constrained languages (e.g., Bengali) due to insufficient domain-specific training data, linguistic diversity, fine-tuned embeddings, and classification methods.

2.2. Text Classification in LRLs

In addition to HRLs, there has been extensive research on LRLs such as Telugu, Marathi, and Tamil. Marreddy et al., 2022 proposed MT-Text GCN, which combines graph autoencoder-based graph reconstruction and multi-task text classification in Telugu. They achieved F1 scores of 0.84 (Sentiment), 0.55 (Emotion), 0.83 (Hate Speech), and 0.66 (Sarcasm). Zhang et al., 2021 also proposed a multi-task learning framework for sentiment analysis on various datasets, combining hard parameter sharing, a BERT-based shared encoder, and a task recognition mechanism. Their approach achieved an F1 score of 0.91. Kapil and Ekbal, 2025 proposed a novel multi-task learning framework for detecting hateful memes. They integrated three pre-trained models: CLIP, BERT, and UNITER, as their work involves both text and images. On the MultiOFF and MAMI datasets, they achieved F1 scores of 0.79 and 0.84, respectively. A recent study (Singh et al., 2024) proposed a multimodal zero-shot multi-task learning framework that uses cross-attention to identify intent and emotion. Their model achieved an increase in F1 score of 0.06 for intent detection and 0.04 for emotion detection compared to other approaches.

2.2.1. Text Classification in Bengali

In contrast to the other languages, TC research leveraging LLMs in Bengali is currently at a rudimentary stage. A limited number of studies also utilized LLMs for TC tasks. For example, Kabir et al., 2023 conducted multiple TC tasks such as question answering, paraphrasing, summarization, and text classification using LLMs (GPT-3.5, Claude-2, LLaMA-2). All the LLMs performed poorly with an F1 score of 0.49 (Claude-2), 0.48 (GPT-3.5), and 0.29 (LLaMA-2). Hasan et al., 2023 explored LLMs with zero and few-shot techniques for Bengali sentiment analysis. The performance of GPT-4 was inferior in their study. The F1 score of GPT-4 in various settings was 0.60 (0-shot), 0.60 (3-shot), and 0.61 (5-shot). A recent work (Nazi et al., 2025) explored various LLMs across four TC tasks, including text classification, sentiment analysis, summarization, and question answering. Their results demonstrated that the closed-source GPT-4o model, utilizing few-shot learning and chain-of-thought prompting, achieved the highest performance across multiple tasks. Few studies also focused on TC in various domains using ML and DL techniques, such as sports news classification (Barua et al., 2021), suspicious text classification (Sharif et al., 2020), and authorship classification (Hossain et al., 2021b). In recent days, transformer-based approaches have gained much attention in various Bengali TC research, including emotion classification (Das et al., 2023; Hider et al., 2024), sentiment analysis (Ahsan et al., 2023), technical text classification (Aodhora and Hoque, 2024), fake news classification (Akther et al., 2025), authorship attribution (Hossain et al., 2025), and aggressive text classification (Sharif and Hoque, 2022). Past research has shown that Bengali TC models struggle to comprehend new tasks that are not widely recognized.

Most past studies in Bengali have focused on a single task or domain, lacking a unified evaluation benchmark for multiple TC tasks or domains. To address this gap, this study explores five distinct Bengali text classification tasks using LLMs.

3. Datasets

The proposed method is evaluated on five Bengali tasks: Sentiment Analysis (SA), Aggressive Text Detection (ATD), Fake News Detection (FND), News Headline Categorization (NHC), and Textual Emotion Analysis (TEA). Each task utilizes its respective dataset from available sources, including SA (Hossain et al., 2021a), ATD (Sharif et al., 2022), FND (Hossain et al., 2020), NHC (Hossain, 2023), and TEA (Das et al., 2023). For in-context learning, the evaluation dataset is used directly without modification. However, for PEFT (Ding et al., 2023), changes were made by manually adding task-specific instructions (prompts) to help the model identify the task being addressed. Additionally, to reduce output ambiguity, a class label name and a corresponding number were added to each output. Table 1 presents a summary of all datasets used in this study. This study encompasses 6,026 Bengali text samples across five distinct classification tasks for evaluation.

For instruction tuning, we add task-specific samples to create a comprehensive training set. Each sample includes detailed task instructions, specific prediction guidance, and class names with corresponding labels for clarity and accuracy. Table 2 provides a summary of the instruction tuning dataset composition. The instruction tuning dataset is a significant expansion over the evaluation datasets, comprising 32,561 samples. To create an instruction dataset, we first inject a prompt before each sample in each task. The prompt provides an instruction to determine the accurate class of the task.

Table 3 provides an overview of the Alpaca-style instructions dataset for the tasks. The alpaca-style dataset guides the model to produce outputs in the desired format for Bengali classification tasks. In contrast, untuned LLMs often generate extraneous text. By training on this dataset, the model is constrained to provide concise, relevant responses tailored to the target classification task. Figure 1 illustrates the distribution of classes among all tasks in instruction tuning and evaluation datasets. It shows that the distributions are nearly identical across the two sets. Among all tasks, NHC has the most data (15,000) across six classes, whereas SA has the least (1,010) across two classes. In NHC, 35.9% of the samples belong to the International category, while only 1.9% of the samples belong to the IT category in the instruction tuning datasets.

4. System Overview

This study explores both in-context learning and PEFT techniques to perform the multiple downstream tasks. Figure 2 provides an overview of the overall system.

4.1. Models

We have explored eight LLMs to evaluate five Bengali text classification tasks. Each model varies according to their internal settings. We have utilized Llama-3.2 3B (Grattafiori et al., 2024), Mistral-7B (**?**), Qwen-2.5 (Yang et al., 2025), Phi-3 (Abdin et al., 2024), Deepseek-7B (Bi et al., 2024), Gemma-3 (Team et al., 2025) and DeepSeek-R1

Task	Data	LR	T_S	T_W	Classes	Class Labels
SA (Hossain et al., 2021a)	434	1-562	456	10,909	2	Positive, Negative
ATD (Sharif et al., 2022)	1,416	3-595	1897	31189	2	Aggressive, Non-Aggressive
FND (Hossain et al., 2020)	2,551	390-3,761	54,508	6,67,351	2	Fake, Authentic
NHC (Hossain, 2023)	1,000	2-9	1000	5932	6	Politics, Sports, National, Entertainment, International, IT
TEA (Das et al., 2023)	625	4-97	832	14,508	6	Joy, Sadness, Fear, Disgust, Anger, Surprise
Total	6,026	-	58,693	729,889	-	-

Table 1: Overview of Bengali text classification evaluation datasets. Here, LR refers to the range of minimum and maximum lengths of each sample, while T_s and T_w denote the total number of sentences and total number of words.

Task	Samples	T_S	T_W
SA	1,010	999	25,443
ATD	10,000	13,491	2,11,096
FND	2,551	1,26,004	1554914
NHC	15,000	15,000	88,512
TEA	4,000	6543	1,14,674
Total	32,561	162,037	1,994,639

Table 2: Instruction tuning dataset statistics for tuning LLMs.

(Guo et al., 2025). For model implementation, we used 4-bit quantized models and QLoRA to tune instructions under limited GPU resources efficiently. Larger closed-source models like DeepSeek-R1 was accessed via their APIs. All other models, ranging from 3 billion to 7 billion parameters, were accessed and executed locally within the Kaggle environment.

4.2. In-Context Learning (ICL)

For ICL, we used different prompting techniques like zero-shot, one-shot, and CoT across five Bengali text classification tasks.

Appendix A illustrates the prompts used in this study.

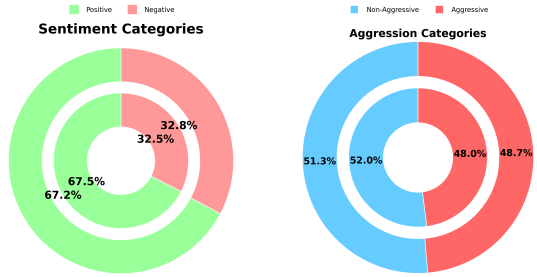
- **Zero-shot Prompt:** The model is given only the task description without any examples.
- **One-shot Prompt:** The model is provided with a single example of the task along with its answer.
- **Chain-of-Thought (CoT) Prompt:** The model is first given the task context, and then guided to reason through the problem step by step, rather than producing a direct answer.

Instruction	Input	Output
Find the sentiment of the following Bengali sentence.	এই সিনেমাটা অসাধারণ ছিল। (This movie was amazing.)	Positive
Detect whether the following Bengali sentence is aggressive or not.	তুই একটা অকাজের গাধা! (You're a useless donkey!)	Aggressive
Classify whether the following Bengali news is fake or real.	প্রধানমন্ত্রী বলেছেন যে আগামীকাল ৫০০০ টাকা করে দেওয়া হবে। (The Prime Minister said that 5,000 taka will be given to each person tomorrow.)	Fake
Categorize the topic of the following Bengali news headline.	তাদের মিটিং-মিছিলের জন্য দেশ স্বাধীন করা হয়নি: হানিফ। (The country was not liberated for their meetings and rallies: Hanif.)	Politics
Detect the emotion expressed in the following Bengali sentence.	আজ আমার খুব কষ্ট হচ্ছে। (I'm feeling very sad today.)	Sad

Table 3: Sample instructions, inputs, and outputs for each task.

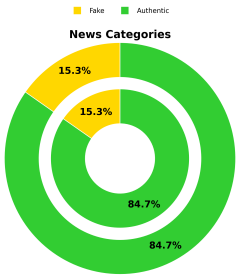
4.3. Parameter Efficient Fine-tuning (PEFT)

In this approach, we have utilized different adapters for each task. We have trained five sepa-

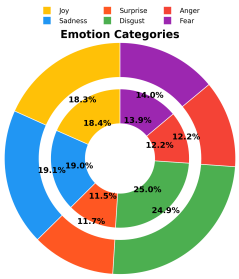


(a) SA

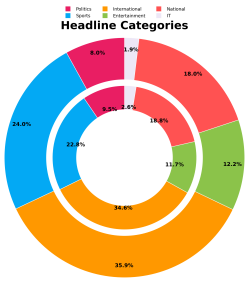
(b) ATD



(c) FND



(d) TEA



(e) NHC

Figure 1: Class-wise distribution of data among all tasks in train and test set.

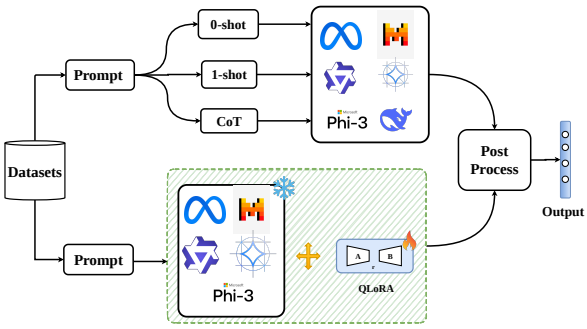


Figure 2: Schematic Diagram of LLM Evaluation for Bengali TC Tasks

rate QLoRA (Dettmers et al., 2023) adapters, one for each of the five tasks, using the SA, ADT, FND, NHC, and TEA datasets. Formally, for a transformer layer weight matrix $W \in \mathbb{R}^{d\times k}$, the adapted weight W' is computed as:

$$W' = W + \Delta W, \quad \Delta W = AB^\top \tag{1}$$

where $A \in \mathbb{R}^{d\times r}$ and $B \in \mathbb{R}^{k\times r}$ are trainable low-rank matrices with rank $r \ll \min(d, k)$. In our implementation, we set $r = 16$, which provides a balance between parameter efficiency and expressive capacity.

Following the LoRA formulation, the low-rank update is scaled as:

$$\Delta W = \frac{\alpha}{r} AB^\top \tag{2}$$

where α is a scaling factor. We set $\alpha = 32$ to ensure a stable update magnitude while preserving the benefits of low-rank adaptation. To improve memory efficiency, we adopt 4-bit quantization for the base model weights, following the QLoRA framework. This allows training large language models under limited hardware constraints by storing backbone weights in low-precision while maintaining adapter parameters in high-precision. We set the dropout rate to 0 based on empirical tuning, as preliminary experiments showed that additional dropout did not improve performance. During inference, the base model remains fixed, and the appropriate task-specific adapter is dynamically loaded.

After generating an output, the LLM may include unnecessary text, which is removed during post-processing using regular expressions. Also, if the responses were found in the wrong format after post-processing, then they were treated as wrong answers.

4.4. Experimental Setup

This study was conducted using Python 3.10.12, PyTorch 2, and Unsloth to implement large language models (LLMs). Experiments were run on Kaggle using an NVIDIA Tesla P100 GPU, which requires 29 GB of RAM, 16 GB of VRAM, and 73.1 GB of storage. Unsloth was used for PEFT with QLoRA, while DeepSeek-70B APIs were employed. We evaluated performance using three prompting strategies and the metrics: precision (Pr), recall (Re), F1-score (F1), and accuracy (Acc).

5. Results and Analysis

Table 4 demonstrates the evaluation results of eight LLMs across five tasks.

How do the models behave in different ICL settings? Across zero-shot, one-shot, and chain-of-thought (CoT) prompting settings, no single model consistently outperforms others across all

Task	Model	Zero-shot				One-shot				CoT			
		Pr	Re	F-1	Acc	Pr	Re	F-1	Acc	Pr	Re	F-1	Acc
SA	Llama	0.93	0.92	0.93	0.94	0.97	0.95	0.96	0.97	0.97	0.95	0.94	0.94
	Gemma	0.93	0.84	0.88	0.89	0.91	0.89	0.90	0.88	0.87	0.89	0.88	0.86
	Mistral	0.70	0.68	0.69	0.71	0.65	0.67	0.66	0.64	0.64	0.67	0.66	0.68
	Qwen	**0.96**	**0.94**	**0.95**	**0.95**	**0.97**	**0.95**	**0.96**	**0.94**	**0.97**	**0.95**	**0.96**	**0.94**
	Phi	0.71	0.73	0.72	0.70	0.64	0.67	0.66	0.68	0.65	0.67	0.66	0.64
	Dseek7B	0.81	0.79	0.80	0.83	0.84	0.82	0.83	0.85	0.84	0.82	0.83	0.85
	DseekR1	0.88	0.90	0.89	0.87	0.89	0.91	0.90	0.88	0.89	0.91	0.90	0.88
ATD	Llama	**0.81**	**0.83**	**0.82**	**0.84**	0.72	0.61	0.54	0.59	0.71	0.69	0.70	0.72
	Gemma	0.81	0.79	0.80	0.82	**0.78**	**0.76**	**0.77**	**0.79**	**0.77**	**0.75**	**0.76**	**0.78**
	Mistral	0.60	0.63	0.62	0.61	0.62	0.58	0.54	0.57	0.60	0.62	0.61	0.59
	Qwen	0.70	0.68	0.69	0.71	0.66	0.64	0.65	0.67	0.71	0.69	0.70	0.72
	Phi	0.50	0.52	0.51	0.49	0.51	0.49	0.50	0.52	0.50	0.48	0.49	0.51
	Dseek7B	0.58	0.60	0.59	0.57	0.58	0.56	0.57	0.59	0.45	0.43	0.44	0.46
	DseekR1	0.80	0.78	0.77	0.77	0.63	0.61	0.62	0.64	0.43	0.41	0.42	0.44
FND	Llama	0.62	0.70	0.63	0.73	0.61	0.69	0.48	0.51	0.77	0.75	0.76	0.78
	Gemma	**0.87**	**0.89**	**0.88**	**0.86**	**0.85**	**0.83**	**0.84**	**0.86**	**0.79**	**0.77**	**0.78**	**0.80**
	Mistral	0.60	0.54	0.49	0.58	0.50	0.48	0.49	0.51	0.49	0.47	0.48	0.50
	Qwen	0.70	0.68	0.69	0.71	0.64	0.62	0.63	0.65	0.74	0.72	0.73	0.75
	Phi	0.56	0.54	0.55	0.57	0.43	0.41	0.42	0.44	0.64	0.62	0.63	0.65
	Dseek7B	0.45	0.43	0.44	0.46	0.45	0.43	0.44	0.46	0.42	0.40	0.41	0.43
	DseekR1	0.92	0.93	0.87	0.83	0.67	0.65	0.66	0.68	0.69	0.67	0.68	0.70
TEA	Llama	0.61	0.42	0.32	0.33	0.63	0.58	0.52	0.53	0.48	0.47	0.45	0.46
	Gemma	**0.57**	**0.56**	**0.55**	**0.64**	**0.66**	**0.63**	**0.57**	**0.59**	**0.56**	**0.54**	**0.49**	**0.50**
	Mistral	0.51	0.31	0.25	0.29	0.55	0.39	0.29	0.36	0.34	0.30	0.24	0.30
	Qwen	0.51	0.45	0.42	0.44	0.46	0.44	0.37	0.39	0.51	0.51	0.47	0.48
	Phi	0.47	0.40	0.37	0.39	0.30	0.25	0.24	0.25	0.30	0.28	0.24	0.31
	Dseek7B	0.30	0.28	0.23	0.29	0.32	0.30	0.27	0.29	0.35	0.34	0.32	0.25
	DseekR1	0.48	0.47	0.45	0.46	0.44	0.42	0.41	0.47	0.38	0.37	0.36	0.40
NHC	Llama	0.57	0.55	0.56	0.58	0.34	0.32	0.33	0.35	0.53	0.51	0.52	0.54
	Gemma	**0.71**	**0.76**	**0.71**	**0.71**	**0.69**	**0.67**	**0.68**	**0.70**	**0.72**	**0.70**	**0.71**	**0.73**
	Mistral	0.14	0.12	0.13	0.15	0.37	0.35	0.36	0.38	0.29	0.27	0.28	0.30
	Qwen	0.23	0.21	0.22	0.24	0.43	0.41	0.42	0.44	0.43	0.41	0.42	0.44
	Phi	0.67	0.65	0.66	0.67	0.21	0.19	0.20	0.22	0.18	0.16	0.17	0.26
	Dseek7B	0.22	0.20	0.21	0.23	0.21	0.19	0.20	0.22	0.45	0.43	0.44	0.46
	DseekR1	0.65	0.63	0.64	0.66	0.67	0.65	0.66	0.68	0.68	0.66	0.67	0.69

Table 4: Performance comparison of models across five tasks for zero-shot, one-shot, and chain-of-thought (CoT) prompting.

five tasks. This variability underscores the heterogeneous linguistic and semantic demands of multitask Bengali text classification, in which different tasks require distinct types of knowledge and reasoning.

Qwen demonstrates the strongest and most stable performance in Sentiment Analysis (SA), achieving F1 scores of 0.95–0.96 across all prompting strategies. The relative insensitivity to prompting variations suggests that sentiment classification primarily relies on coarse-grained polarity cues that transfer effectively across languages. This indicates that Qwen's multilingual pretraining likely provides sufficient semantic grounding for polarity detection without requiring task-specific adaptation. In contrast, Gemma achieves superior performance on more structurally and semantically complex tasks, including Fake News Detection (FND), News Headline Categorization (NHC), and Text Emotion Analysis (TEA). These tasks demand contextual reasoning, pragmatic interpretation, and fine-grained label discrimination. Gemma's comparatively strong instruction-following alignment appears better suited to handling multi-class classification scenarios that require nuanced semantic differentiation. Llama exhibits competitive zero-shot performance on Aggressive Text Detection (ATD). This observation further supports the notion that task complexity and cultural specificity significantly influence the effectiveness of cross-lingual transfer.

Collectively, these results indicate that model performance in Bengali multitask classification is highly task-dependent and closely tied to the interaction between pretraining distribution, instruction alignment, and the linguistic characteristics of each classification objective.

Does CoT help in these low-resource tasks? Chain-of-thought prompting does not consistently improve performance and frequently degrades it. Although CoT is theoretically beneficial for tasks requiring structured reasoning (Wei et al., 2022), its effectiveness depends on the availability of reliable internal knowledge representations. In Bengali, where pretraining exposure is comparatively limited, CoT often produces fluent but semantically unreliable reasoning chains. For example, models optimized for reasoning exhibit significant degradation under CoT on certain tasks. Instead of correcting errors, the additional reasoning steps amplify

uncertainty. These findings suggest that CoT requires sufficiently rich language-specific priors to be effective. In low-resource contexts, it may increase the risk of hallucinations by encouraging overconfident yet weakly grounded intermediate reasoning steps (Ahmed et al., 2025).

How do the models perform after PEFT? Table 5 reports the performance of instruction-tuned models fine-tuned with QLoRA (PEFT), and the improvements over the best ICL baselines are consistent and often substantial. The most dramatic gains are observed on the tasks that proved hardest under ICL. For TEA, the best ICL F1 (Gemma, zero-shot) was 0.55; after PEFT fine-tuning, Gemma-3-4B achieves an F1 of 0.74 — a gain of 19 percentage points. Similarly, for NHC, the best ICL F1 score was 0.71 (Gemma, zero-shot), whereas PEFT raises it to 0.83. These results confirm that for culturally embedded tasks such as emotion analysis and news categorization in Bengali, task-specific adaptation to in-domain data is far more effective than any prompting strategy. For ATD, PEFT enables Mistral-7B to reach an F1 of 0.93, compared to a best ICL F1 of 0.82. On FND, Gemma-3-4B achieves an F1 of 0.97 — more than 9 points above its own best ICL performance. Across all five tasks, Gemma-3-4B consistently ranks among the top-performing PEFT models, achieving the highest F1 on SA (0.97), FND (0.97), TEA (0.74), and NHC (0.83). Mistral-7B leads only on ATD (F1: 0.93) but exhibits a catastrophic failure on FND after fine-tuning (F1: 0.06), suggesting a label prediction collapse that warrants further investigation. The consistency of PEFT gains across models and tasks supports the conclusion that QLoRA fine-tuning efficiently surfaces latent multilingual capacity that prompting alone cannot elicit. By adapting the model's lower-rank projection matrices to Bengali-specific lexical and morphological distributions, fine-tuning allows even relatively compact models (3B–7B parameters) to substantially close the gap with larger systems evaluated in zero-shot settings.

Which task is most challenging for the models? Across both ICL and PEFT settings, a clear difficulty ordering emerges among the five tasks. SA is the least challenging, with top PEFT models achieving an F1 score above 0.97, whereas TEA remains the hardest, achieving a maximum F1 of 0.74 even after fine-tuning. This ordering reflects the varying degree to which task-relevant knowledge can be transferred from high-resource pretraining distributions. Sentiment polarity is a relatively universal semantic property; emotion, hate, and argumentation in Bengali, by contrast, are deeply tied to cultural register, code-switching practices, and community-specific linguistic conventions that multilingual LLMs are unlikely to have internalized from pretraining data alone.

Which model emerges as the most effective backbone for multitask Bengali text classification across tasks? Among the evaluated models, Gemma-3-4B emerges as the most suitable backbone for multitask Bengali text classification, demonstrating the most consistent cross-task performance under both in-context learning and parameter-efficient fine-tuning.

Why does DeepSeek-R1 perform poorly despite being a larger model? Despite being among the larger models evaluated, R1 underperforms relative to its scale on several tasks. In particular, its CoT performance is consistently inferior to its own zero-shot performance — most dramatically on ATD, where CoT reduces its F1 from 0.77 to 0.42. This behavior is a direct consequence of R1's training objective, which optimizes for systematic chain-of-thought reasoning. When applied to Bengali classification tasks, this inductive bias is counterproductive: the model is compelled to reason through a language and domain it has limited knowledge of, producing confident but incorrect reasoning chains. This finding highlights an important limitation of reasoning-optimized LLMs in low-resource languages.

Task	Model	Pr	Re	F1	Acc
SA	Llama-3.2-4B	0.83	0.85	0.84	0.85
	Gemma-3-4B	0.96	0.97	0.97	0.97
	Mistral-7B-it	0.89	0.92	0.90	0.91
	Qwen-2.5-8B-it	0.94	0.94	0.94	0.94
	Phi-3	0.84	0.85	0.84	0.86
ATD	Llama-3.2-4B	0.89	0.89	0.89	0.89
	Gemma-3-4B	0.91	0.91	0.91	0.91
	Mistral-7B-it	0.93	0.93	**0.93**	0.93
	Qwen-2.5-8B-it	0.89	0.89	0.89	0.89
	Phi-3	0.91	0.91	0.91	0.91
FND	Llama-3.2-4B	0.89	0.89	0.89	0.89
	Gemma-3-4B	0.97	0.97	**0.97**	0.97
	Mistral-7B-it	0.67	0.16	0.06	0.16
	Qwen-2.5-8B-it	0.83	0.84	0.83	0.85
	Phi-3	0.93	0.73	0.81	0.73
TEA	Llama-3.2-4B	0.63	0.61	0.62	0.63
	Gemma-3-4B	0.74	0.74	**0.74**	0.74
	Mistral-7B-it	0.73	0.70	0.69	0.70
	Qwen-2.5-8B-it	0.67	0.67	0.67	0.67
	Phi-3	0.65	0.66	0.65	0.66
NHC	Llama-3.2-4B	0.76	0.77	0.76	0.77
	Gemma-3-4B	0.83	0.83	**0.83**	0.83
	Mistral-7B-it	0.81	0.81	0.81	0.81
	Qwen-2.5-8B-it	0.77	0.77	0.77	0.77
	Phi-3	0.27	0.40	0.30	0.40

Table 5: Performance comparison of different instruction-tuned models across five tasks.

5.1. Ablation Study

We conducted an ablation study on LoRA rank, LoRA α, and LoRA dropout for PEFT. Due to the high computational cost of training Gemma-3 4B Instruct, these experiments were limited to the ATD dataset only. Multiple combinations of LoRA rank, α, and dropout were evaluated, as summarized in Table 6.

Based on the ablation results, Set-8 from Table 6 was selected as the final configuration to develop the best-performing Gemma-3 4B model. Three key factors guided this decision. First, Set-8 achieved comparable or slightly better validation performance than higher-rank configurations (e.g., rank-32) and exhibited more stable generalization across validation splits. Second, it required significantly fewer computational resources than higher ranks (32, 64), including reduced GPU memory usage and faster training times, making it more practical for large-scale fine-tuning. Finally, increasing the LoRA rank beyond this setting did not yield consistent performance gains, indicating diminishing returns relative to the added computational cost. Therefore, Set-8 was chosen as the optimal trade-off between performance, generalization, and efficiency, and it was used for all subsequent experiments and analyses with all other models.

Set	LoRA r	LoRA α	Dropout	Pr	Acc	F1
1	32	64	0	0.92	0.91	0.91
2	8	8	0	0.91	0.90	0.90
3	8	32	0.03	0.89	0.88	0.89
4	16	128	0	0.91	0.91	0.91
5	8	8	0	0.87	0.84	0.84
6	16	64	0	0.91	0.91	0.91
7	32	32	0	0.91	0.91	0.91
8	16	32	0	0.91	0.91	0.91
9	16	16	0.01	0.91	0.90	0.90
10	64	64	0.05	0.91	0.87	0.87

Table 6: Ablation Study on LoRA Hyperparameters (r, α, Dropout) in ATD

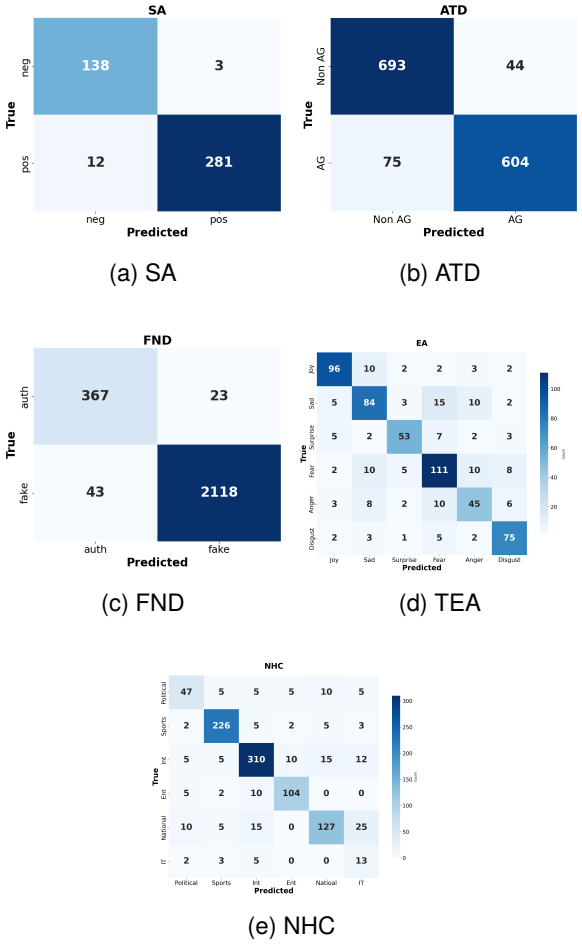
5.2. Error Analysis

As shown in the previous section, the instruction-tuned Gemma-3 4B model exhibits the best overall performance for Bengali text classification.

5.2.1. Quantitative Error Analysis

Figure 3 presents the confusion matrices of the proposed Gemma-3 4B model for five tasks.

For the SA task, the model achieved 138 true negatives (TN) and 281 true positives (TP), indicating strong performance. For the ATD task, it recorded 693 TN and 694 TP; for the FND task, it reached 367 TN and 2118 TP. These results demonstrate excellent performance across all three binary classification tasks. However, performance declined on the EC and NHC tasks with



(a) SA (b) ATD

(c) FND (d) TEA

(e) NHC

Figure 3: Confusion matrices of all five tasks for the proposed model.

six classes, which made them more complex. In these tasks, the Gemma-3 4B model exhibited a noticeable decline in true positives and true negatives compared with its binary classification performance, highlighting its limitations in making fine-grained multiclass distinctions.

5.2.2. Qualitative Error Analysis

Table 7 presents examples of both correctly and incorrectly classified predictions across the five tasks.

These predictions are generated by the Gemma-3 4B instruction-tuned version. In the first three samples, the model correctly identifies the class, demonstrating its effectiveness in clear, less-ambiguous cases. However, in the last two examples, the model fails to predict the correct label. The fourth sample, taken from the news headline categorization task, contains an inherently ambiguous headline that could reasonably belong to multiple categories. This ambiguity likely contributed to the model's misclassification. In the final example, from the emotion analysis task, the distinction between the predicted and actual class is subtle—

Text Samples	Actual	Predicted
একটা প্রকাশনা কিভাবে এমন একটা বই প্রকাশ করে। (How does a publisher publish such a book.)	Negative	Negative
হাইরে মানুষ নিজে তৈরি করে আবার নিজে পূজা করে। (Oh human! You create it yourself, and then worship it yourself.)	Aggressive	Aggressive
ইসলামী মতিবেদকখতনা করেও রেহাই পেলেন না বিশিষ্ঠ কবি, হেকিমী চিকিতসক ও নও মুসলমান ফরহাদ মজহার লুংগি। (Even after adopting Islamic ideology and circumcision, the distinguished poet, Hakimi physician, and new Muslim Farhad Mazhar did not escape criticism.)	Fake	Fake
তাদের মিটিং-মিছিলের জন্য দেশ স্বাধীন করা হয়নি: হানিফ। (The country wasn't liberated for their rallies and protests: Hanif.)	Politics	National
রিয়াদের নিজের উপরই রাগ উঠে কি দরকার ছিল ঘাবেড় দেওয়ার সত্য কথাটুকু বলেই হতো। (Riyad got angry at himself — what was the need to scare them? He could've just told the truth.)	Anger	Disgust

Table 7: Some correctly and incorrectly classified samples by Gemma-3-4B.

even for humans. The text's emotional tone is nuanced, making it especially challenging for the model to interpret accurately. This highlights a common difficulty in fine-grained emotion classification: overlapping emotional cues can lead to confusion.

6. Conclusion

This paper presents a comprehensive evaluation of several LLMs, focusing on their instruction-tuned versions. Gemma-3 4B consistently outperformed other large language models, including Llama, Mistral, Qwen, and Phi, across most tasks, despite its relatively small size. Notably, the instruction-tuned Mistral-7B-it model achieved the best results in aggressive text detection. This study also compared prompting techniques, finding that instruction tuning generally improves performance, and that zero-shot prompting often yields better results than one-shot or chain-of-thought methods. Future work will focus on further extending evaluations to a broader range of domains and investigating more advanced instruction-tuning strategies. Additionally, the latest models, such as GPT-5 and Gemini-2.5 Flash, can be evaluated to assess their effectiveness in handling these tasks under in-context learning settings.

Limitations

Although the proposed method demonstrates satisfactory performance across various text classification tasks, several significant limitations remain unaddressed.

- The instruction-tuned Gemma-3 4B model achieves high accuracy on binary classification tasks; however, its performance declines on complex multiclass classification tasks.
- The model was adapted using Quantized Low-Rank Adaptation (QLoRA) rather than full fine-tuning, which may have potentially restricted task alignment and overall classification accuracy.
- Prompting strategies, including zero-shot, one-shot, and chain-of-thought approaches, yield inconsistent results across different tasks and model architectures.
- Several instruction-tuned models lack explicit training on Bengali-language data, which limits their ability to capture Bengali-specific linguistic features.
- This study relies on monolingual Bengali datasets and excludes Bangla-English code-mixed data, which is common in real-world usage.
- The research scope is restricted to classification tasks and does not investigate additional natural language processing (NLP) tasks, including text generation, summarization, or question answering.

Acknowledgment

This work was supported by the Directorate of Research & Extension (DRE), Chittagong University of Engineering & Technology (CUET), Chittagong, Bangladesh, under the Grant Number CUET/DRE/2023-2024/CSE/025.

Ethical Considerations

This study used only publicly available, pre-existing datasets to evaluate five text classification tasks. No new human data were collected or annotated. All datasets were released for research under appropriate licenses, and all sources are correctly cited. Some tasks may contain sensitive or harmful language. To mitigate this risk, the datasets were used exclusively within their original research context, and dangerous content was

not reproduced or promoted beyond what was necessary for model evaluation. All experiments adhered to responsible artificial intelligence research practices, emphasizing transparency, reproducibility, and fairness while minimizing potential misuse.

Data/Code Availability Statement

The datasets and source code used in this study will be available at: https://github.com/CUET-NLP-Lab/bengali-llm-multitask-classification

Bibliographical References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Shamima Afroz, Kawsar Ahmed, and Mohammed Moshiul Hoque. Leveraging multi-task learning for detecting aggression, emotion, violence, and sentiment in bengali texts. In *5th Muslims in ML Workshop co-located with NeurIPS 2025*.

Kawsar Ahmed, Md Osama, Omar Sharif, Eftekhar Hossain, and Mohammed Moshiul Hoque. 2025. Bennumeval: A benchmark to assess llms' numerical reasoning capabilities in bengali. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17782–17799.

Shawly Ahsan, Fairooz Tasnia, Nafisa Tabassum, Avishek Das, Mohammed Moshiul Hoque, and Nazmul Siddique. 2023. Classifying textual sentiment using bidirectional encoder representations from transformers. In *2023 26th International Conference on Computer and Information Technology (ICCIT)*, pages 1–6. IEEE.

Aysha Akther, Kazi Masudul Alam, and Rameswar Debnath. 2025. Automatic detection of manipulated bangla news: A new knowledge-driven approach. *Natural Language Processing Journal*, 11:100155.

Neel Alex, Eli Lifland, Lewis Tunstall, Abhishek Thakur, Pegah Maham, C Jess Riedel, Emmie Hine, Carolyn Ashurst, Paul Sedille, Alexis Carlier, et al. 2021. Raft: A real-world few-shot text classification benchmark. *arXiv preprint arXiv:2109.14076*.

Sumaiya Rahman Aodhora and Mohammed Moshiul Hoque. 2024. Tetec: Technical text classification in bengali using ensemble of transformers. In *2024 International Conference on Recent Progresses in Science, Engineering and Technology (ICRPSET)*, pages 1–6.

Adrita Barua, Omar Sharif, and Mohammed Moshiul Hoque. 2021. Multi-class sports news categorization using machine learning techniques: resource creation and evaluation. *Procedia Computer Science*, 193:112–121.

Anirban Bhowmick and Abhik Jana. 2021. Sentiment analysis for bengali using transformer based models. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 481–486.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Avishek Das, Mohammed Moshiul Hoque, Omar Sharif, M Ali Akber Dewan, and Nazmul Siddique. 2023. Temox: Classification of textual emotion using ensemble of transformers. *IEEE Access*, 11:109803–109818.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature machine intelligence*, 5(3):220–235.

Fuad Ibne Jashim Farhad, Shah Imran, Md Mehedi Hasan Santo, Mahbub Khan, Anamul Sakib, Md Shahidur Rahman, Md Ariful Islam, Rezaul Haque, and Shafiur Rahman. 2024. Addressing misinformation in bengali media: A hybrid deep learning solution. In *2024 27th International Conference on Computer and Information Technology (ICCIT)*, pages 774–779. IEEE.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024.

The llama 3 herd of models. *arXiv preprint arXiv:2407. 21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Md Arid Hasan, Shudipta Das, Afiyat Anjum, Firoj Alam, Anika Anjum, Avijit Sarker, and Sheak Rashed Haider Noori. 2023. Zero-and few-shot prompting with llms: A comparative study with fine-tuned models for bangla sentiment analysis. *arXiv preprint arXiv:2308.10783*.

Md. Ali Hider, Shawly Ahsan, Jawad Hossain, and Mohammed Moshiul Hoque. 2024. Emotion classification in bengali-english code-mixed data using transformers. In *2024 27th International Conference on Computer and Information Technology (ICCIT)*, pages 3529–3535.

Eftekhar Hossain. 2023. Bangla news headlines categorization. GitHub repository. Accessed: Jun. 27, 2025.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2021a. Sentiment polarity detection on bengali book reviews using multinomial naive bayes. In *Progress in Advanced Computing and Intelligent Engineering: Proceedings of ICACIE 2020*, pages 281–292. Springer.

Md. Rajib Hossain, Mohammed Moshiul Hoque, M. Ali Akber Dewan, Enamul Hoque, and Nazmul Siddique. 2025. Authornet: Leveraging attention-based early fusion of transformers for low-resource authorship attribution. *Expert Systems with Applications*, 262:125643.

Md. Rajib Hossain, Mohammed Moshiul Hoque, M. Ali Akber Dewan, Nazmul Siddique, Md. Nazmul Islam, and Iqbal H. Sarker. 2021b. Authorship classification in a resource constraint language using convolutional neural networks. *IEEE Access*, 9:100319–100338.

Md Zobaer Hossain, Md Ashraful Rahman, Md Saiful Islam, and Sudipta Kar. 2020. Banfakenews: A dataset for detecting fake news in bangla. *arXiv preprint arXiv:2004.08789*.

Mohsinul Kabir, Md Tahmid Rahman Laskar, Mir Tafseer Nayeem, M Saiful Bari, and Enamul Hoque. 2023. Benllmeval: A comprehensive evaluation into the potentials and pitfalls of large language models on bengali nlp. *arXiv preprint arXiv:2309.13173*.

Prashant Kapil and Asif Ekbal. 2025. A transformer based multi task learning approach to multimodal hate speech detection. *Natural Language Processing Journal*, 11:100133.

Lefteris Loukas, Ilias Stogiannidis, Prodromos Malakasiotis, and Stavros Vassos. 2023. Breaking the bank with chatgpt: few-shot text classification for finance. *arXiv preprint arXiv:2308.14634*.

Mounika Marreddy, Subba Reddy Oota, Lakshmi Sireesha Vakada, Venkata Charan Chinni, and Radhika Mamidi. 2022. Multi-task text classification using graph convolutional networks for large-scale low resource language. In *2022 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE.

Zabir Al Nazi, Md. Rajib Hossain, and Faisal Al Mamun. 2025. Evaluation of open and closed-source llms for low-resource language with zero-shot, few-shot, and chain-of-thought prompting. *Natural Language Processing Journal*, 10:100124.

Tanjela Rahman Rosni, Mahamudul Hasan, Tanni Mittra, Md Sawkat Ali, and Md Hasanul Ferdaus. 2024. Aggressive bangla text detection using machine learning and deep learning algorithms. In *International Conference on Computation of Artificial Intelligence & Machine Learning*, pages 174–183. Springer.

Timo Schick and Hinrich Schütze. 2022. True few-shot learning with prompts—a real-world perspective. *Transactions of the Association for Computational Linguistics*, 10:716–731.

Omar Sharif and Mohammed Moshiul Hoque. 2022. Tackling cyber-aggression: Identification and fine-grained categorization of aggressive texts on social media using weighted ensemble of transformers. *Neurocomputing*, 490:462–481.

Omar Sharif, Mohammed Moshiul Hoque, A. S. M. Kayes, Raza Nowrozy, and Iqbal H. Sarker. 2020. Detecting suspicious texts using machine learning techniques. *Applied Sciences*, 10(18).

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.

Gopendra Vikram Singh, Mauajama Firdaus, Dushyant Singh Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2024. Zero-shot multitask

intent and emotion prediction from multimodal data: A benchmark study. *Neurocomputing*, 569:127128.

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*.

Hongbin Wang, Chun Ren, and Zhengtao Yu. 2025. Multimodal sentiment analysis based on multiple attention. *Engineering Applications of Artificial Intelligence*, 140:109731.

Zhiqiang Wang, Yiran Pang, and Yanbin Lin. 2024. Smart expert system: Large language models as text classifiers. *arXiv preprint arXiv:2405.10523*.

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Jian Zhang, Ke Yan, and Yuchang Mo. 2021. Multi-task learning for sentiment analysis with hard-sharing and task recognition mechanisms. *Information*, 12(5):207.

A. Prompt Examples

In this task, we have explored three types of prompting. Tables 8, 9, and 10 illustrate all types of prompt templates used in this research. In Table 9, one-shot samples are provided. However, in this Table, we show only an example for one class. But while experimenting, we provided one example for each class.

Task	Prompt
SA	Please classify the sentiment of this review. The answer should be either `1` (positive) or `0` (negative), based on the sentiment expressed. # Review: `<review>`
ATD	Please classify whether this Bengali sentence is Aggressive or non-aggressive. The answer should be either `1` (Aggressive) or `0` (Non-Aggressive), based on the sentence. # Sentence: `<sentence>`
FND	Please determine whether this Bengali news is genuine or not. The answer should be either `1` (Fake) or `0` (Authentic), based on the news. # News: `<news>`
NHC	Classify the following news headline into one of the predefined categories. Use the corresponding label number: `politics: 0, sports: 1, international: 2, entertainment: 3, national: 4, IT: 5.` # Headline: `<headline>` # Answer: `[num]`
TEA	Classify the emotion expressed in the following Bengali text into one of the predefined categories. Use the corresponding label number: `0: Joy, 1: Sadness, 2: Surprise, 3: Disgust, 4: Anger, 5: Fear.` # Sentence: `<sentence>` # Answer: `[label]`

Table 8: Prompt templates for zero-shot prompt across all the tasks

Task	Prompt
SA	Please classify the sentiment of this review. The answer should be either `1` (positive) or `0` (negative). **Example:** # Review: এই সিনেমাটি দারুণ লেগেছে # Answer: `1` **Now classify:** # Review: `<review>`
ATD	Please classify whether this Bengali sentence is Aggressive or Non-Aggressive. The answer should be either `1` (Aggressive) or `0` (Non-Aggressive). **Example:** # Sentence: আমি তোমাকে মেরে ফেলবো # Answer: `1` **Now classify:** # Sentence: `<sentence>`
FND	Please classify whether this Bengali news is Fake or Authentic. The answer should be either `1` (Fake) or `0` (Authentic). **Example:** # News: আগামীকাল চাঁদে ফুটবল খেলা হবে # Answer: `1` **Now classify:** # News: `<news>`
NHC	Classify the following news headline into one of the predefined categories. Use the corresponding label number: `politics: 0, sports: 1, international: 2, entertainment: 3, national: 4, IT: 5.` **Example:** # Headline: বাংলাদেশ ক্রিকেট দল জিতেছে # Answer: `1` **Now classify:** # Headline: `<headline>`
TEA	Classify the emotion expressed in the following Bengali text into one of the predefined categories. Use the corresponding label number: `0: Joy, 1: Sadness, 2: Surprise, 3: Disgust, 4: Anger, 5: Fear.` **Example:** # Sentence: আমি আজ খুব খুশি # Answer: `0` **Now classify:** # Sentence: `<sentence>`

Table 9: Prompt templates for one-shot prompt across all the tasks

Task	Prompt
SA	**Task:** Classify the sentiment of the following Bangla review as `1` (positive) or `0` (negative). **Steps:** 1. **Analyze Content:** Break the review into phrases or sentences. 2. **Identify Sentiment Indicators:** Look for positive (e.g., "অসাধারণ'') or negative (e.g., ``খারাপ'') words/phrases. 3. Evaluate Context: Consider how words are used, including sarcasm or mixed sentiment. 4. Determine Tone: Assess the overall tone based on key phrases. 5. Classify: Assign `1` for positive and `0` for negative. Now classify: # Review: `<review>`
ATD	**Task:** Classify whether the following Bangla sentence is Aggressive (`1`) or Non-Aggressive (`0`). **Steps:** 1. **Analyze Content:** Break the sentence into meaningful parts. 2. **Identify Aggression Indicators:** Look for threatening or harmful expressions (e.g., "মেরে ফেলবো''). 3. Evaluate Intensity: Check the severity of the words used and their target. 4. Determine Tone: Assess whether the tone is hostile or neutral. 5. Classify: Assign `1` for Aggressive and `0` for Non-Aggressive. Now classify: # Sentence: `<sentence>`
FND	**Task:** Classify whether the following Bangla news is Fake (`1`) or Authentic (`0`). **Steps:** 1. **Analyze Content:** Break the news statement into key claims. 2. **Check Plausibility:** Identify whether the claim sounds realistic or exaggerated. 3. **Identify Unrealistic Elements:** Look for impossible or illogical events (e.g., "চাঁদে ফুটবল খেলা হবে''). 4. Evaluate Source-Like Tone: Assess if the sentence resembles factual reporting or satire. 5. Classify: Assign `1` for Fake and `0` for Authentic. Now classify: # News: `<news>`
NHC	**Task:** Classify the following Bangla news headline into one of the predefined categories: `politics: 0, sports: 1, international: 2, entertainment: 3, national: 4, IT: 5`. **Steps:** 1. **Analyze Content:** Break the headline into key subjects and actions. 2. **Identify Keywords:** Detect topic-related words (e.g., "ক্রিকেট'' → sports, ``সরকার'' → politics). 3. Match with Categories: Compare keywords with the predefined category list. 4. Resolve Ambiguity: If multiple categories fit, select the most dominant one. 5. Classify: Assign the corresponding label number. Now classify: # Headline: `<headline>`
TEA	**Task:** Classify the emotion expressed in the following Bangla text into one of the predefined categories: `0: Joy, 1: Sadness, 2: Surprise, 3: Disgust, 4: Anger, 5: Fear`. **Steps:** 1. **Analyze Content:** Break the sentence into emotion-bearing parts. 2. **Identify Emotion Indicators:** Look for explicit words or expressions (e.g., "খুশি'' → Joy, ``রাগ'' → Anger). 3. Evaluate Context: Consider implied emotions or indirect expressions. 4. Determine Dominant Emotion: Choose the strongest emotion if multiple exist. 5. Classify: Assign the corresponding label number. Now classify: # Sentence: `<sentence>`

Table 10: Prompt templates for zero-shot CoT across all tasks with reasoning steps