Evaluating Open-Source LLMs for Bengali Text Classification Across Multiple Domains

Eftekhar Hossain
eftekhar@ucf.edu
University of Central Florida
Orlando, Florida, USA

Abstract

With the rapid advancement of Large Language Models (LLMs), there is a growing demand for AI systems capable of handling multilingual tasks. While many LLMs demonstrate strong performance in real-time applications for high-resource languages, their effectiveness in low-resource languages like Bengali is understudied. This gap is primarily due to the limited availability of high-quality text corpora in Benagli and insufficient emphasis placed on it during model training, which restricts the models' ability to generalize and perform reliably in practical, real-time applications. To this end, this work investigates the capability of Large Language Models (LLMs) in Bengali text classification tasks across various domains. Experiments with six open-source LLMs with varying model size on four tasks, i.e., *Sentiment Analysis*, *Emotion Recognition*, *Hate Speech Detection*, and *Fake News Detection*, demonstrate that LLMs with fewer parameters perform poorly in classification tasks that require subtle contextual understanding, such as Emotion Recognition. Our findings also suggest that recent LLMs such as *Qwen-2.5-72B* and *Gemma-2-27B* have promising capabilities for handling diverse tasks in the Bangla language. The LLM inference code is available at https://github.com/Computational-NLU-Project.

1 Introduction

In recent years, with the advent of Large Language Models (LLMs), numerous Natural Language Processing (NLP) tasks have been solved without the need for extensive training or fine-tuning. Leveraging their in-context learning capabilities [6], LLMs can accomplish complex NLP tasks such as classification, summarization, and question-answering in zero-shot or few-shot settings [17]. While LLMs have demonstrated remarkable performance in high-resource languages like English, their effectiveness in solving NLP problems for a low-resource language like Bengali remains underexplored. However, to develop a language-agnostic system, it is crucial for LLMs to understand and address NLP problems in languages other than English. To this end, in this work, we perform a comprehensive study to evaluate the performance of open-source LLMs in Bengali text classification across various domains, filling a critical gap in understanding their capabilities in low-resource language settings. Our study revolves around a few critical questions that we have investigated through extensive experiments with four LLMs on two datasets.

More specifically, we shed light on three critical questions: *(i)* Are the LLMs good at text classification when the task is for a low-resource language? *(ii)* Do LLMs have domain-specific knowledge and solve classification tasks for diverse domains in resource-constrained languages?, and *(ii)* Does the prompting strategy make any impact on the LLMs' performance?.

To address these research questions, we conduct a comprehensive evaluation of six open-source Large Language Models (LLMs), varying in scale and assessed under different prompting strategies

2 Related Work

In the past few years, several works have demonstrated the capability of LLMs in various NLP tasks. Wang et al. [15] evaluated three LLMs (i.e., Llama, GPT-3.5, and GPT-4) on four text classification tasks using zero-shot learning. Similarly, Chae et al. [4] evaluated four different LLMs for text classification tasks using prompting and fine-tuning methods. Some work has also been accomplished on low-resource languages. For example, Cahyawijaya et al. [3] studied the In-Context learning capability of LLMs for 25 low-resource languages. In another line of work, Al Nazi et al. [2] evaluated the LLMs' performance for various Bengali NLP tasks, including text classification, summarization, and question answering. In contrast to the works mentioned above, our works differ in such a way that we evaluated four open-source LLMs for text classification tasks specifically for the Bangla Language across various domains.

3 Methodology

The objective of this work is to study the effectiveness of open-source LLMs in different Bengali text classification tasks. Figure 1 shows the abstract view of the study carried out to evaluate the LLMs' performance on the Bangla language. For the evaluation, we considered four different Bengali text classification tasks: *(i)* Sentiment Analysis, *(ii)* Emotion Recognition, *(iii)* Hate Speech Detection, and *(iv)* Fake News Detection. For this study, we employed six open-source LLMs and evaluated their performance for the abovementioned tasks. The experimented LLMs are **LLama-3.2-3B** [14], **Mistral-V3-7B** [10], Llama distilled version of **DeepSeek-R1-8B**[7], **Phi-4-14B**[1], **Gemma-2-27B**[13], and **Qwen-2.5-72B** [16]. Following the prior work, [12], we use two different prompting strategies to prompt the LLMs to solve the tasks. More specifically, we use the in-context learning capability of the LLMs, where the model directly infers the output based on the given instruction without explicitly training on any samples from the dataset. The details of the Prompts and templates are outlined in Appendix A.2.

4 Experiments

We used the vLLM[1] library to download the instruction version of the LLMs and generated the responses using both zero-shot and few-shot prompting. To avoid repetitive output and randomness, we set sampling parameters as: $temperature = 0$, $top\_p = 0.8$, $repetition\_penalty = 1.05$, and $max\_tokens = 512$. We use the

[1] https://docs.vllm.ai/en/latest/

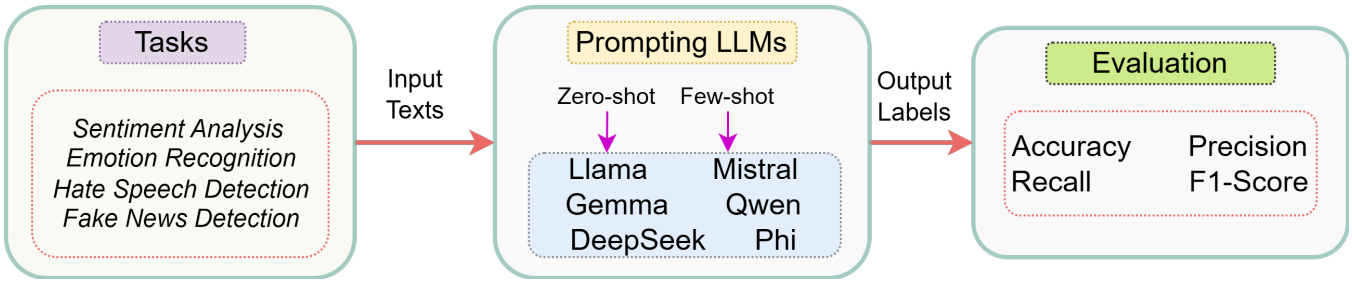


Figure 1: Abstract process of evaluating LLMs by prompting, across multiple tasks of a low-resource language.

accuracy, precision, recall, and F1-score as the primary evaluation metrics to assess the model's performance. We also use a confusion matrix to inspect the class-wise performance of the models across the tasks. We have summarized our evaluation metrics in Appendix A.3.

4.1 Results

This section presents the performance of various LLMs on two prompting settings across four classification tasks. The associated results for each prompting strategy are described as follows:

4.1.1 Zero-Shot Performance . Table 1 shows the performance of the LLMs across the four tasks. For the *Sentiment Analysis* task, Gemma-2-27B achieves the highest accuracy (62.04%), recall (62.04%), and F1 score (61.54%), while Qwen-2.5-72B gives the highest precision of 66.12%, suggesting that it is particularly good at minimizing false positives. For the Emotion Recognition task, both Qwen-2.5-72B and Gemma-2-27B perform best, achieving the highest accuracy of 61.76%. Notably, both models achieve very close performance across other evaluation matrices (i.e., precision, recall, and F1-score). Similarly, in Hate Speech Detection Qwen-2.5-72B achieves the highest performance (accuracy: 79.90% and f1-score:79.88%) while Gemma-2-27B attained the highest performance in the Fake News Detection task. Notably, the effectiveness of the frontier open source LLMs (i.e., Qwen-2.5 and Gemma-2) in the hate speech and fake news detection task also suggests their robustness in handling offensive content and misleading content in low-resource languages.

In most of the tasks, LLama-3.2-3B, Mistral-V3-7B, and DeepSeek-R1-8B models perform poorly. One possible reason is that these models are small in terms of the number of parameters. Another is that this model might lack an understanding of the corresponding task in the Bangla language, leading them to perform poorly in such well-known NLP tasks.

4.1.2 Few-Shot Performance. Table 2 reports the LLMs' performance in a few-shot (5) setting. Overall, Qwen-2.5-72B emerges as the top-performing model on Emotion Recognition and Hate Speech Detection tasks, achieving the highest F1 scores of 65.86% and 78.25% respectively. Similarly, Gemma-2-27B demonstrates strong performance on the other two tasks, Sentiment Analysis

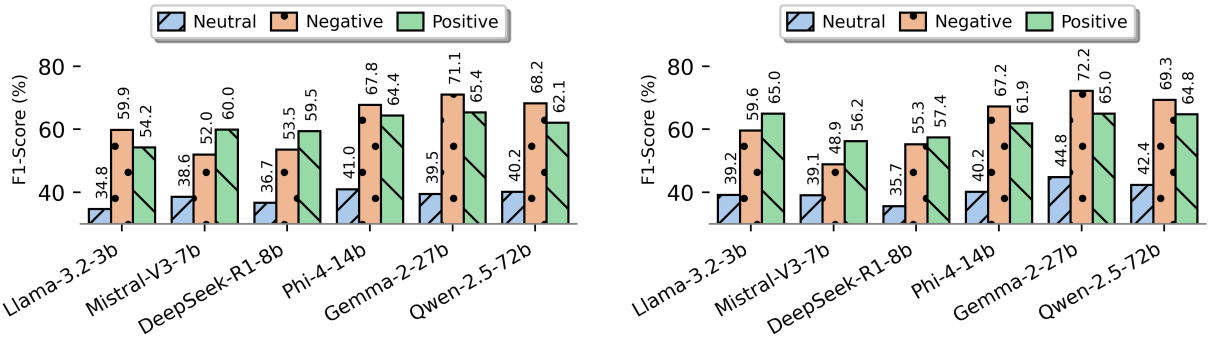
Model	Sentiment Analysis				Emotion Recognition			
	Acc.	P.	R.	F1	Acc.	P.	R.	F1
Llama-3.2-3B	52.21	54.62	52.21	51.83	46.24	49.08	46.24	44.28
Mistral-V3-7B	51.32	54.49	51.32	52.25	36.96	49.30	36.96	33.89
DeepSeek-R1-8B	50.82	55.15	50.82	52.15	45.92	47.48	45.92	42.14
Phi-4-14B	59.84	64.11	59.84	60.28	59.84	65.48	59.84	56.96
Gemma-2-27B	62.04	65.12	62.04	61.54	61.76	68.47	61.76	58.40
Qwen-2.5-72B	58.07	66.12	58.07	59.32	61.76	67.67	61.76	59.44
	Hate Speech Detection				Fake News Detection			
	Acc.	P.	R.	F1	Acc.	P.	R.	F1
Llama-3.2-3B	57.10	59.49	57.10	54.22	77.76	77.95	77.76	77.72
Mistral-V3-7B	64.70	65.58	64.70	64.19	61.61	76.02	61.61	55.45
DeepSeek-R1-8B	59.70	63.91	59.70	56.40	72.44	72.61	72.44	72.39
Phi-4-14B	72.50	73.62	72.50	72.17	77.36	79.68	77.36	76.91
Gemma-2-27B	76.50	78.33	76.50	76.11	84.45	85.57	84.45	84.33
Qwen-2.5-72B	79.90	80.05	79.90	79.88	83.27	86.26	83.27	82.91

Table 1: Zero-shot evaluation results of LLMs on various Bengali Text Classification tasks.

Model	Sentiment Analysis				Emotion Recognition			
	Acc.	P.	R.	F1	Acc.	P.	R.	F1
Llama-3.2-3B	57.12	57.50	57.12	57.19	46.72	63.35	46.72	47.12
Mistral-V3-7B	47.92	58.03	47.92	49.71	46.24	57.79	46.24	46.41
DeepSeek-R1-8B	50.63	54.20	50.63	51.71	46.56	45.69	46.56	43.41
Phi-4-14B	58.32	64.24	58.32	58.88	56.00	64.95	56.00	54.56
Gemma-2-27B	62.55	67.93	62.55	63.00	61.12	71.66	61.12	60.27
Qwen-2.5-72B	60.28	66.86	60.28	61.30	65.76	70.79	65.76	65.86

	Hate Speech Detection				Fake News Detection			
	Acc.	P.	R.	F1	Acc.	P.	R.	F1
Llama-3.2-3B	64.40	67.28	64.40	62.85	69.29	80.05	69.29	66.27
Mistral-V3-7B	63.10	63.14	63.10	63.07	63.58	74.84	63.58	58.93
DeepSeek-R1-8B	62.30	63.46	62.30	61.47	67.32	71.69	67.32	65.59
Phi-4-14B	71.00	72.33	71.00	70.56	81.69	83.25	81.69	81.48
Gemma-2-27B	74.30	78.15	74.30	73.39	84.45	87.03	84.45	84.17
Qwen-2.5-72B	78.40	79.18	78.40	78.25	84.06	87.50	84.06	83.68

Table 2: Few-shot evaluation results of LLMs on various Bengali Text Classification tasks.



(a) Zero Shot (b) Few Shot

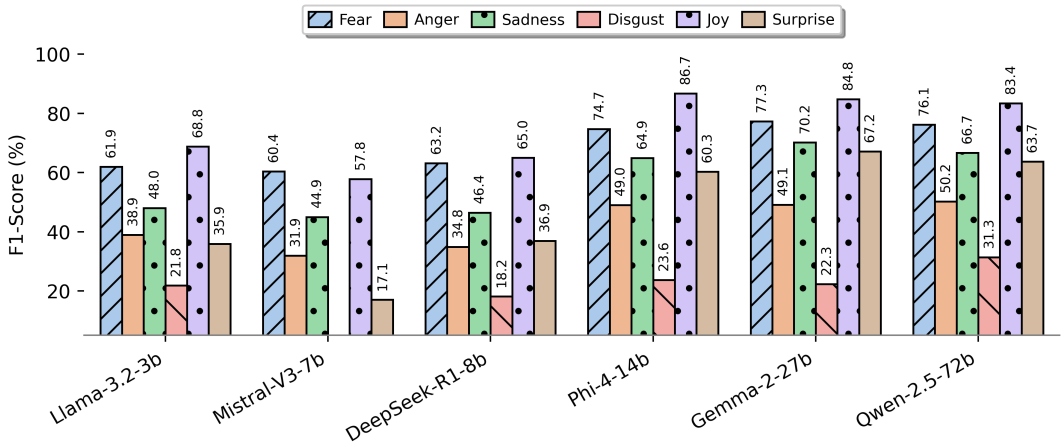
Figure 2: Class-wise sentiment classification performance for two prompt settings across LLMs.

and Fake News Detection, with an F1 score of 63% and 84.17% respectively. Overall, the results show notable improvement in few-shot prompting across all the tasks compared to the zero-shot counterpart, indicating that even minimal supervision significantly enhances model generalization in low-resource settings.

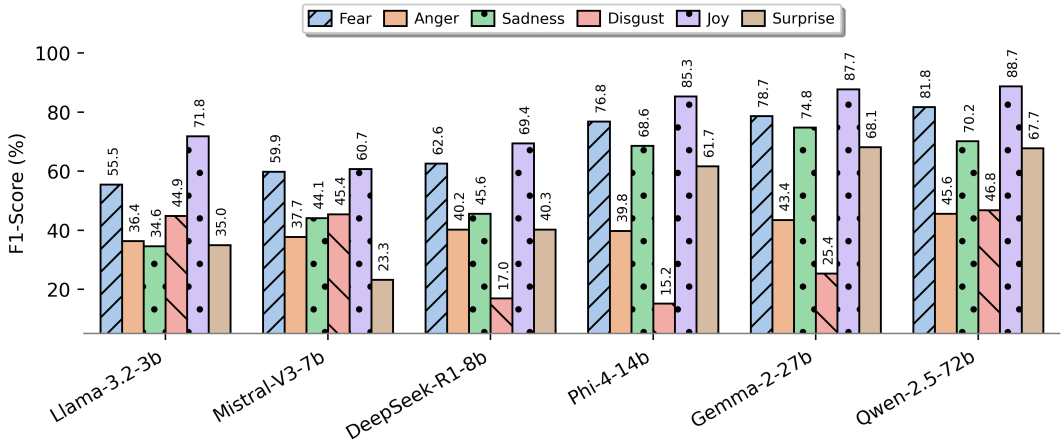
On the other hand, Mistral-V3-7B and DeepSeek-R1-8B consistently underperform (in terms of F1-score) across most of the tasks—for instance, 49.71% and 51.71% in Sentiment Analysis and 46.41% and 43.41% in Emotion Recognition. This performance gap can be attributed to several factors, including its smaller model size, limited multilingual capabilities, and possibly insufficient exposure to Bengali during pretraining.

4.1.3 Classwise Performance. This section discusses the LLMs' performance across different classes of the four classification tasks.

LLMs performance across Sentiment Categories. Figure 2 shows the F1 scores of sentiment classes across the LLMs for zero and shot prompting. In zero-shot prompting, it is observed that all models perform well on *negative* and *positive* sentiment classes, especially the Gemma-2-27B and Qwen-2.5-72B, while *neutral* sentiment got the lowest F1-score across all models. A similar trend is also seen in the case of a few-shot performance, where most models attained significant improvement in the neutral class. For example, Gemma-2-27B gains around 5% boosts in neutral class. This finding suggests that few-shot prompting is effective when the task requires a subtle understanding of the input text to infer an output, and that is only possible when the model gets some exposure to the input.



(a) Zero Shot



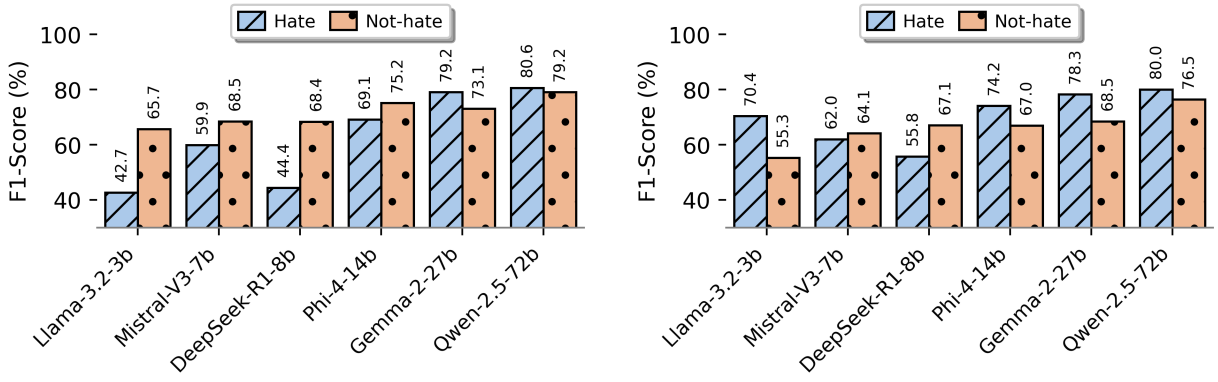
(b) Few Shot

Figure 3: Class-wise emotion recognition performance for two prompt settings across LLMs.

LLMs performance across Emotion Categories. Figure 3 shows the F1 scores across emotion categories. We can observe that in the zero-shot setting, LLMs' performance varies significantly across emotion categories, where *Joy* and *Fear* tend to have higher F1 scores, whereas *Disgust* and *Anger* are more challenging to classify accurately. Qwen-2.5-72B and Gemma-2-27B again show strong performance, except in negative emotions such as *Anger* and *Disgust*. While few-shot prompting with Qwen-2.5-72B improves performance by 15% in the Disgust class compared to zero-shot prompting, it results in a 5% decrease in the Anger class. Nonetheless, all other emotion classes (Joy, Fear, Sadness, Surprise) exhibit notable gains, each improving by more than 5% over the zero-shot counterpart. We also notice that the Llama-3.2-3B, Mistral-V3-7B, and DeepSeek-R1-8B perform poorly across most of the categories, especially in Anger, Disgust, and Surprise, with both prompting approaches.

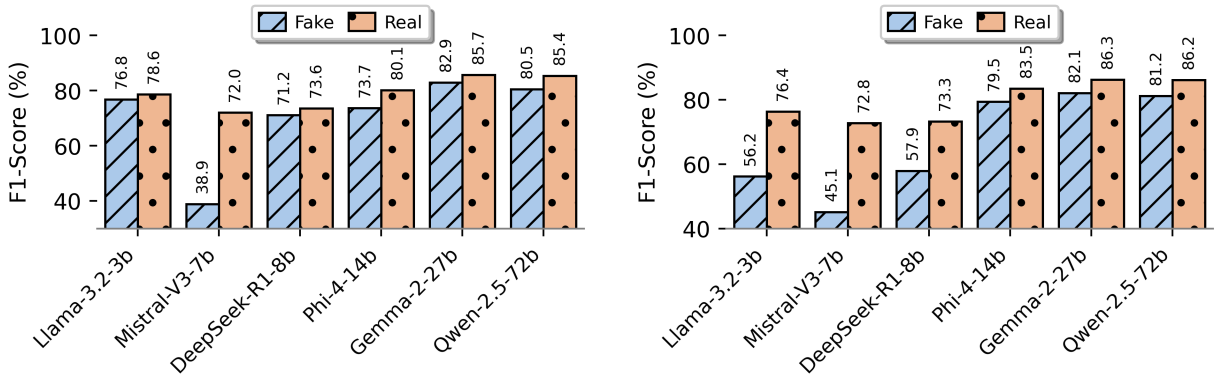
LLMs performance across Hate and Not-Hate Categories. Figure 4 illustrates class-wise F1-scores in the Hate Speech Detection task. In the zero-shot setting, most models—including Llama-3.2-3B, DeepSeek-R1-8B, and Phi-1.4-14B—struggle to detect the hate class, with F1-scores below 50% in some cases, while performing considerably better on the not-hate class. This indicates a strong bias toward the majority or safer class when no examples are provided. In contrast, the few-shot setting significantly boosts hate class detection across all models with the cost of a slight decrease in the not-hate class. Notably, Qwen-2.5-72B and Gemma-2-27B achieve balanced and high F1-scores for both classes; surprisingly, these two models' performance slightly decreased with few-shot prompting.

LLMs performance across Fake and Real News. Figure 5 shows class-wise F1-scores for the fake news detection task. In the zero-shot setting, most models achieve F1-scores above 70% for both real and fake classes, with the notable exception of Mistral-V3-7B, which performs poorly on fake instances, yielding F1-scores below



(a) Zero Shot (b) Few Shot

Figure 4: Class-wise hate speech detection performance for two prompt settings across LLMs.



(a) Zero Shot (b) Few Shot

Figure 5: Class-wise fake news detection performance for two prompt settings across LLMs.

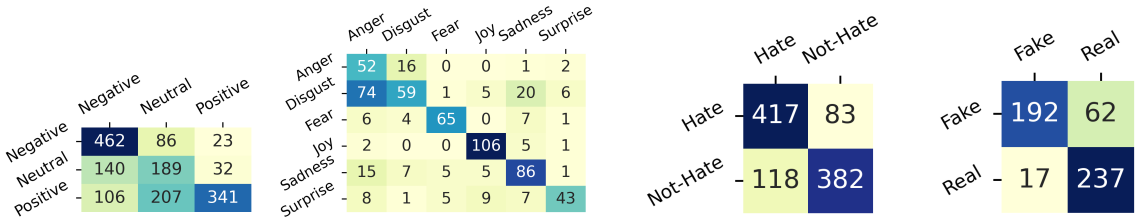
40%. Overall, Gemma-2-27B and Qwen-2.5-72B emerge as the top-performing models across both prompting strategies. Interestingly, contrary to trends observed in other tasks, the few-shot prompting approach does not lead to performance improvements in this task. In fact, it often results in decreased performance, suggesting that the additional context may not effectively enhance the model's understanding for this task.

4.2 Error Analysis

To better understand the errors made by the models across different classes, we performed an error analysis on the top-performing models through their confusion matrices shown in Figure 6.

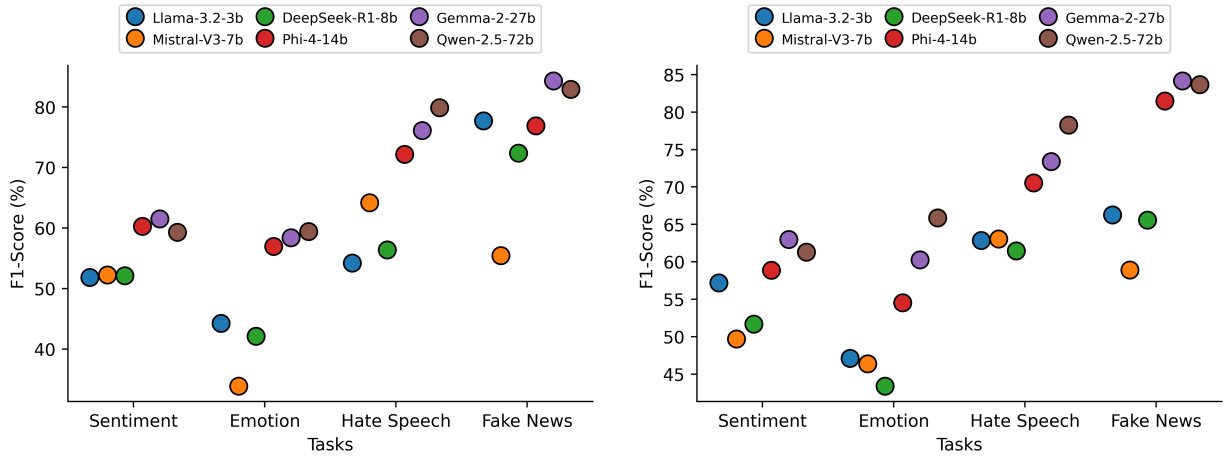
In the case of sentiment analysis (Figure 6a), we observe that a large proportion of texts labeled as *neutral* are misclassified as *negative* (140 out of 361 instances). A similar trend is evident across the other sentiment classes, which are mostly confused with *neutral* texts, suggesting that the models struggle to capture subtle distinctions between sentiments. This difficulty likely stems from limited lexical overlap and the absence of strong, class-specific cue words. In case of the Emotion recognition task in Figure 6b, we can see that the model (Qwen-2.5-27B) mostly gets confused with the *Disgust* class and incorrectly labels them as *Anger* or Sadness, indicating that the model is better at recognizing distinct and expressive emotions (i.e., Joy) but struggle with those that are more context-dependent or subtle and has overlapping triggered words.

In the hate speech detection task, Figure 6c shows that a substantial number of texts labeled as *Not-Hate* (118 out of 500) are misclassified as hateful, indicating that the model exhibits a bias toward predicting hate speech even when it is not present. A similar trend is observed in the fake news detection task as well, illustrated in Figure 6d, where approximately 25% of fake content is incorrectly classified as real. This suggests that detecting misleading content poses a greater challenge for LLMs, likely because it often requires up-to-date world knowledge to recognize subtle manipulations and misinformation embedded in the original content.



(a) Gemma-2-27B (Few) (b) Qwen-2.5-72B (Few) (c) Qwen-2.5-72B (Zero) (d) Gemma-2-27B (Zero)

Figure 6: Confusion matrix of the best performing LLMs across different tasks.



(a) Zero Shot (b) Few Shot

Figure 7: LLMs performance with two prompt settings across various classification tasks.

4.3 Discussions

This section highlights the key insights of this study that we found through our results and error analysis. The notable trends in LLMs' performance across four classification tasks under zero-shot and few-shot prompting are shown in Figure 7. Overall, Fake News detection emerges as the easiest task, with consistently high F1 scores, while Emotion classification remains the most challenging task. In terms of the prompting approach, the Few-shot setting significantly improves performance across all tasks, especially for Emotion and Hate Speech, indicating the importance of in-context learning. From the model's perspective, it is seen that the larger models like *Qwen-2.5-72b* and *Gemma-2-27b* consistently outperform others across all the tasks, suggesting benefits from scale and potentially better pretraining on multilingual data.

Notably, some models, such as *Mistral-V3-7b*, *Phi-4-14b*, and *DeepSeek-R1-8b*, underperform across multiple tasks, particularly in zero-shot settings. Their lower F1 scores may stem from smaller parameter sizes or limited exposure to diverse languages during pretraining. Since the tasks are in Bengali, models (LLama, Mistral, and Phi) pretrained predominantly on English or high-resource languages may struggle due to vocabulary gaps and weaker cultural or syntactic alignment. This highlights the critical role of both model scale and multilingual training data in enabling strong generalization, especially for nuanced tasks in underrepresented languages. Therefore, it can be deduced that the strong performance of *Qwen* and *Gemma* could be attributed not just to their scale but also to their richer multilingual pretraining, making them more capable in a morphologically rich language, Bengali.

5 Conclusion

This study evaluates six open-source large language models (LLMs)-LLaMA, Mistral, Phi, Gemma, Qwen, and DeepSeek-R1—on four low-resource Bengali text classification tasks: sentiment analysis, emotion recognition, hate speech detection, and fake news detection. Experimental results with zero-shot and few-shot prompting show that Qwen and Gemma consistently achieve strong performance across all tasks, suggesting that these LLMs have a decent understanding of context and linguistic variation in Bengali. Future work could explore integrating chain-of-thought prompting and task-specific fine-tuning to further enhance LLM performance in low-resource scenarios

References

[1] 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* (2024).

[2] 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 (2025), 100124.

[3] Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. LLMs Are Few-Shot In-Context Low-Resource Language Learners. *arXiv preprint arXiv:2403.16512* (2024).

[4] Youngjin Chae and Thomas Davidson. 2023. Large language models for text classification: From zero-shot learning to fine-tuning. *Open Science Foundation* (2023).

[5] Avishek Das, Omar Sharif, Mohammed Moshiul Hoque, and Iqbal H Sarker. 2021. Emotion Classification in a Resource Constrained Language Using Transformer-based Approach. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. 150–158.

[6] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234* (2022).

[7] 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* (2025).

[8] 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* (2020).

[9] 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*. 3265–3271.

[10] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).

[11] Md Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Sagor Sarker, Mehadi Hasan Menon, Kabir Hossain, Md Azam Hossain, and Stefan Decker. 2021. Deephateexplainer: Explainable hate speech detection in under-resourced bengali language. In *2021 IEEE 8th international conference on data science and advanced analytics (DSAA)*. IEEE, 1–10.

[12] Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Xiangji Huang. 2023. A systematic study and comprehensive evaluation of chatgpt on benchmark datasets. *arXiv e-prints* (2023), arXiv–2305.

[13] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118* (2024).

[14] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[15] Zhiqiang Wang, Yiran Pang, and Yanbin Lin. 2023. Large language models are zero-shot text classifiers. *arXiv preprint arXiv:2312.01044* (2023).

[16] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115* (2024).

[17] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).

A Appendix

A.1 Task Description

Sentiment Analysis. The Bengali sentiment analysis capability of the LLMs is evaluated on the SentiNoB [9] dataset. This dataset comprises Bengali texts that were collected from various social media platforms and labeled into *positive*, *negative*, and *neutral* sentiment classes. For the evaluation, we used the original test set of the dataset, which consists of 1586 instances where positive, negative, and neutral have, respectively, 654, 571, and 361 samples.

Emotion Recognition. For evaluating the emotion recognition capability of the LLMs in Bengali, we used the BEmoC [5] dataset. The data were collected from various social media platforms and annotated into six emotion classes six emotion classes namely, *joy*, *anger*, *disgust*, *sadness*, *anger*, and *fear*. For evaluation, we used the test set, which consists of 625 samples.

Hate Speech Detection. For evaluating the offensive content detection capability of the LLMs in Bengali, we used the Bengali Hate Speech Dataset [11]. For ease of evaluation, we opted for the binary classification task that has around 500 *hateful*, and 500 *not-hateful* texts.

Fake News Dataset. To evaluate the LLMs on misleading content detection, we used the BanFakeNews Dataset [8]. The test set of the dataset was highly imbalanced, having around 20k samples in the *real news* category while only 254 samples in the *fake news* category. For our evaluation, we downsampled the majority class (real) to make a balanced test set.

A.2 Prompting Techniques and Templates

We employed two popular prompting techniques for the LLM inference.

Zero-shot prompting: In this approach, the model is prompted only with the task description, such as identifying sentiment or emotion recognition, without providing any examples or additional context. Then the model completes the task by generating responses through its general knowledge, instigated entirely by the prompt instructions.

Few-shot prompting (five-shot): For this method, the model is provided with five examples with corresponding outputs of the task before being asked to generate a response. This allows the model to infer the task requirements from the examples, potentially improving its accuracy in solving the tasks.

A.3 Evaluation Metrics

- **Confusion Matrix:** is a table that summarizes a model's performance by showing the counts of True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN).
- **Accuracy (Acc.):** measures the overall correctness of a model by calculating the ratio of correctly predicted instances to the total instances.
- **Precision (P):** indicates how many of the predicted positive instances are positive (TP / (TP + FP)).
- **Recall (R):** measures how many actual positive instances were correctly identified (TP / (TP + FN)).
- **F1-score (F1):** is the harmonic mean of precision and recall, balancing both metrics (2 * (Precision * Recall) / (Precision + Recall)).

Sentiment Classification Zero-Shot Prompt

```
You are an advanced multilingual sentiment analysis model. Your task is to classify the sentiment of the given Bengali text as **positive**, **negative**, or **neutral**.

  Follow these instructions carefully:
  - Read the text thoroughly.
  - Determine the sentiment based on the overall tone and language used.
  - Output your response in a valid JSON format with the following structure:

  {{
    "label": "positive" / "negative" / "neutral"
  }}

  Additional Instructions:
  - Do not include any explanations, comments, or extra text outside the JSON object.
  - Ensure that the classification reflects the actual sentiment of the text.

  Now, analyze the following text and classify its sentiment: "{text}"
```

Figure A.2.1: Prompt for zero-shot Bengali Sentiment Analysis

Emotion Recognition Zero-Shot Prompt

```
  You are an advanced multilingual emotion recognition model. Your task is to classify the given Bengali text into one of the following emotions: **disgust**, **sadness**, **joy**, **fear**, **surprise**, or **anger**.

  Follow these instructions carefully:
  - Read the text thoroughly.
  - Determine the predominant emotion based on the language, tone, and context.
  - Output your response in a valid JSON format with the following structure:

  {{
    "label": "disgust" / "sadness" / "joy" / "fear" / "surprise" / "anger"
  }}

  Additional Instructions:
  - Do not include any explanations, comments, or extra text outside the JSON object.
  - Ensure that the classification accurately reflects the primary emotion conveyed in the text.

  Now, analyze the following text and classify its emotion: "{text}"
```

Figure A.2.2: Prompt for zero-shot Bengali Emotion Recognition

Hate Speech Detection Zero-Shot Prompt

```
You are an advanced multilingual hate speech detection model. Your task is to
classify the given Bengali text as either **"hate"** or **"not-hate"**.

  Follow these instructions carefully:
  - Read the text thoroughly.
  - Determine if the text contains hate speech, offensive language, or harmful
content.
  - Output your response in a valid JSON format with the following structure:

  {{
    "label": "hate" / "not-hate"
  }}

  Additional Instructions:
  - Do not include any explanations, comments, or extra text outside the JSON
object.
  - Ensure that the classification is based strictly on the presence of hateful or
offensive language.

  Now, analyze the following text and classify it as "hate" or "not-hate": "
{text}"
```

Figure A.2.3: Prompt for zero-shot Hate Speech Detection

Fake News Detection Zero-Shot Prompt

```
You are a highly accurate fake news detection system trained on multilingual
and multi-domain data. Your task is to determine whether the given news text
is **real** or **fake** based on its content, language, tone, and potential
indicators of misinformation.

  Follow these instructions carefully:
  - Read and analyze the entire news content.
  - Evaluate the credibility, consistency, and plausibility of the information
presented.
  - Use linguistic cues, factual consistency, and any implicit indicators of
misinformation to make your judgment.

  Output your response in the following **valid JSON format**:

  {{
    "label": "real" / "fake"
  }}

  Additional Guidelines:
  - Do not include any explanation or additional commentary.
  - Output only the JSON object.
  - Ensure your classification is based solely on the text provided.

  Now, analyze the following news text and classify it as real or fake:

  "{text}"
```

Figure A.2.4: Prompt for zero-shot Fake News Detection

Sentiment Classification Few-Shot Prompt

You are an advanced multilingual sentiment analysis model. Your task is to classify Bengali text as having **positive**, **negative**, or **neutral** sentiment.

Below are 5 example texts with their correct sentiment labels:
Example 1:
Text: "আজকের দিনটা দারুণ কাটলো, খুব ভালো লাগছে!"
Label: "positive"
Example 2:
Text: "এই রাস্তাটা একেবারেই খারাপ, চলাফেরা করাই মুশকিল।"
Label: "negative"
Example 3:
Text: "আজ সকাল থেকে বৃষ্টি হচ্ছে।"
Label: "neutral"
Example 4:
Text: "ওর সঙ্গে দেখা হয়ে খুব ভালো লাগলো।"
Label: "positive"
Example 5:
Text: "ট্রেনটা আজ আবার দেরিতে এসেছে, বিরক্ত লাগছে।"
Label: "negative"

Now, classify the sentiment of the following text:
Text: "{text}"

Output your response in a valid JSON format with the following structure:
{{
  "label": "positive" / "negative" / "neutral"
}}

Instructions:
- Output only the JSON object.
- Do not include explanations or comments.
- Make sure the sentiment reflects the actual tone and context of the text.

Figure A.2.5: Prompt for few-shot Bengali Sentiment Analysis

Emotion Recognition Few Shot Prompt

You are an advanced multilingual emotion recognition model. Your task is to classify Bengali text into one of the following emotions: **disgust**, **sadness**, **joy**, **fear**, **surprise**, or **anger**.

Below are 5 example texts with their correct emotion labels:
Example 1:
Text: "ওর এমন ব্যবহারে আমি ভীষণ ঘৃণা বোধ করেছি।"
Label: "disgust"
Example 2:
Text: "আজ আমার বাবার মৃত্যুবার্ষিকী, খুব কষ্ট পাচ্ছি।"
Label: "sadness"
Example 3:
Text: "আমি পরীক্ষায় প্রথম হয়েছি! খুব খুশি লাগছে।"
Label: "joy"
Example 4:
Text: "রাতে হঠাৎ দরজার আওয়াজ পেয়ে খুব ভয় লাগলো।"
Label: "fear"
Example 5:
Text: "ও এমন কথা বলবে ভাবতেই পারিনি!"
Label: "surprise"

Now, classify the following text into one of the predefined emotions.
Text: "{text}"

Output your response in a valid JSON format with the following structure:

```
{{
   "label": "disgust" / "sadness" / "joy" / "fear" / "surprise" / "anger"
}}
```

Instructions:
- Output only the JSON object.
- Do not include explanations or comments.
- Ensure the label reflects the main emotion conveyed in the text.

Figure A.2.6: Prompt for few-shot Bengali Emotion Recognition

Hate Speech Detection Few-Shot Prompt

You are an advanced multilingual hate speech detection model. Your task is to classify Bengali text as either **"hate"** or **"not-hate"**.

Below are 5 example texts with their correct hate speech labels:

Example 1:
Text: "এই মানুষগুলো সব কুচক্রি, এদের দেশ থেকে তাড়ানো উচিত।"
Label: "hate"
Example 2:
Text: "আজকে অনেক গরম পড়েছে, বাইরে যাওয়া কঠিন।"
Label: "not-hate"
Example 3:
Text: "ওরা একেবারে অকর্মা, সমাজের বোঝা মাত্র!"
Label: "hate"
Example 4:
Text: "তোমার বক্তব্যটা শুনে ভালো লাগলো, অনেক কিছু শিখলাম।"
Label: "not-hate"
Example 5:
Text: "এমন লোকদের দেখলেই রাগ ধরে, এদের পেটাতে ইচ্ছে করে।"
Label: "hate"

Now, analyze the following text and classify it as "hate" or "not-hate":
Text: "{text}"

Output your response in a valid JSON format with the following structure:

```
{{
   "label": "hate" / "not-hate"
}}
```

Instructions:
- Output only the JSON object.
- Do not include explanations or comments.
- Ensure that the classification is based strictly on the presence of hateful, offensive, or harmful language.

Figure A.2.7: Prompt for few-shot Hate Speech Detection

Fake News Detection Few-Shot Prompt

You are a highly accurate fake news detection system trained on multilingual and multi-domain data. Your task is to determine whether the given news text is **real** or **fake** based on its content, language, tone, and potential indicators of misinformation.

Below are 5 example news texts along with their correct classifications:

Example 1:
Text: "সরকার ঘোষণা করেছে আগামীকাল থেকে দেশের সব শিক্ষাপ্রতিষ্ঠান পুনরায় খোলা হবে।"
Label: "real"
Example 2:
Text: "একটি গবেষণায় দেখা গেছে যে প্রতিদিন লবঙ্গ খেলে করোনা সম্পূর্ণরূপে নিরাময় হয়।"
Label: "fake"
Example 3:
Text: "চট্টগ্রামে ভূমিকম্পে ক্ষতিগ্রস্ত হয়েছে বহু বাড়িঘর, উদ্ধার কাজ চলছে।"
Label: "real"
Example 4:
Text: "চাঁদের বুকে বিশালাকার প্রাণী দেখা গেছে, নাসা নিশ্চিত করেছে!"
Label: "fake"
Example 5:
Text: "বাংলাদেশ ক্রিকেট দল আগামী মাসে ভারত সফরে যাচ্ছে বলে জানিয়েছে বিসিবি।"
Label: "real"

Now, analyze the following news text and classify it as **real** or **fake**:
Text: "{text}"

Respond in the following **valid JSON format**:

```
{{
   "label": "real" / "fake"
}}
```

Instructions:
- Do not include any explanation or commentary.
- Output only the JSON object.
- Ensure your classification is based solely on the content and plausibility of the news text.

Figure A.2.8: Prompt for few-shot Fake News Detection