Do Large Language Models Speak All Languages Equally? A Comparative Study in Low-Resource Settings

Anonymous ACL submission

Abstract

Large language models (LLMs) have garnered significant interest in natural language processing (NLP), particularly for their remarkable performance in various downstream tasks in resource-rich languages such as English. However, the applicability and efficacy of LLMs in low-resource language contexts remain largely unexplored, thus highlighting a notable gap in linguistic capabilities for these languages. The limited utilization of LLMs in low-resource scenarios is primarily attributed to constraints such as dataset scarcity, computational costs, and research lacunae specific to low-resource languages. To address this gap, we comprehensively examines zero-shot learning using multiple LLMs in both English and low-resource languages. Our findings indicate that GPT-4 consistently outperforms Llama 2 and Gemini, with English consistently demonstrating superior performance across diverse tasks compared to low-resource languages. Furthermore, our analysis reveals that among the evaluated tasks, natural language inference (NLI) exhibits the highest performance, with GPT-4 demonstrating superior capabilities. This research underscores the imperative of assessing LLMs in low-resource language contexts to augment their applicability in general-purpose NLP applications.

1 Introduction

Recent advances in large language models (LLMs) developed significant interest in natural language processing (NLP) across academia and industry. LLMs are known for their language generation capabilities that are trained on billions or trillions of tokens with billions of trainable parameters. Recently researchers have been evaluating LLMs for various NLP downstream tasks, especially question answering (Akter et al., 2023; Tan et al., 2023; Zhuang et al., 2023), reasoning (Suzgun et al., 2022; Miao et al., 2023), mathematics (Lu et al., 2023; Rane, 2023), machine translation (Xu et al., 2023; Lyu et al., 2023), etc.

Most of the existing works on the evaluation of LLMs are on resource-rich languages such as English. However, the capabilities and performances of LLMs for low-resource languages for many NLP downstream tasks are not widely evaluated, leaving a notable gap in the linguistic capabilities of low-resource languages. In low-resource languages such as Bangla and Urdu, several researchers are handling the scarcity of datasets and other resources in NLI (Bhattacharjee et al., 2021), Sentiment analysis (Hasan et al., 2023b; Sun et al., 2023; Koto et al., 2024; Kumar and Albuquerque, 2021) and Hate speech detection (Khan et al., 2021; Santosh and Aravind, 2019). However, the amount of work that uses LLMs is still very few, mainly due to a few constraints such as dataset scarcity, computational costs, and research gaps associated with low-resource languages. These constraints of low-resource languages require more attention, alongside a focus on high-resource languages, to enhance the applicability of LLMs to general-purpose NLP applications.

To fill the aforementioned gap, we comprehensively analyze zero-shot learning using various LLMs in English and low-resource languages. The performance of LLMs shows that GPT-4 provides comparatively better results than Llama 2 and Gemini. Moreover, the English language performs better on different tasks than low-resource languages such as Bangla, Hindi, and Urdu. The Key contributions are as follows:

- To address the limitation of publicly available datasets for low-resource languages, we present datasets for sentiment and hate speech tasks by translating from English to Bangla, Hindi, and Urdu, thereby facilitating research in low-resource language processing.
- We investigate and analyze the effectiveness

of different LLMs across various tasks for both English and low-resource languages such as Bangla, Hindi, and Urdu, which suggest that LLMs perform better when evaluated in the English language.

- We apply zero-shot prompting using natural language instructions, which contain a description of the task and expected output, which enables the construction of a context to generate more appropriate output.

The rest of the paper is organized as follows: Section 2 gives a summary of related studies. Section 3 discusses the LLMs that were utilized and describes their contents. We describe the detailed zero-shot learning prompting and experimental details in Section 4. Our findings are presented and discussed in the 5 section. Finally, we concluded in Section 6.

2 Related Works

Generative models, or LLMs, are proficient in various NLP tasks and have high generalization across several NLP tasks. Despite the incredible generalization of large language models (LLMs), there is significant room for improvement in their performance, particularly in low-resource languages such as Bangla, Hindi, Urdu, etc. Previous study (Robinson et al., 2023) demonstrates the inability of LLMs such as GPT-4 to perform on low-resource (African) languages as well as on high-resource languages. However, LLMs perform well in languages (European) that use the same script as English (Holmström et al., 2023).

2.1 LLM for English

Researchers have developed many resources and benchmarks in the past couple of decades for the English language (Wang et al., 2018; Williams et al., 2017). Moreover, several widely recognized downstream tasks, including NLI, sentiment analysis, and hate speech detection, have been studied in the English language for a long time. NLI involves determining the logical relationship between pairs of text sequences (Conneau et al., 2018; Kowsher et al., 2023). LLMs can determine the relationships among text sequences and produce results similar to state-of-the-art techniques.(Pahwa and Pahwa, 2023; Gubelmann et al., 2023). In contrast, Sentiment analysis aims to understand and extract subjective information from textual data, such as opinions, attitudes, emotions, or feelings expressed by individuals or groups. Using LLMs for sentiment analysis could be a fascinating prospect because we do not have to develop datasets or train models, and it still produces identical results (Sun et al., 2023; Hasan et al., 2023b). Additionally, Hate speech detection is a challenging but essential task to identify and mitigate offensive or harmful language in text data. Some of the commonly used techniques for detecting hate speech include machine learning (Abro et al., 2020; Mullah and Zainon, 2021), deep learning(Badjatiya et al., 2017; Zimmerman et al., 2018), transformer-based models(Mozafari et al., 2020; Alatawi et al., 2021), and LLMs (Hee et al., 2024; García-Díaz et al., 2023).

2.2 LLM for Low-resource Languages

NLP research works and applications related to this downstream task mainly focus on high-resource languages. Unlike the English language, the advancement of NLP tasks for low-resource languages made it challenging due to several factors described by Alam et al. (2021). However, there have been some improvements in the last couple of years for Bangla sentiment analysis focusing on resource development (Hasan et al., 2020a; Islam et al., 2021; Hasan et al., 2023a) that attained attention from many researchers to concentrate on solving this issue. However, researchers are focusing on generalizing NLP tasks across the languages. Some of these applications have many limitations for low-resource languages that must be addressed to develop and deploy more generalized universal NLP applications. Some of the recent works on NLI (Pahwa and Pahwa, 2023; Gubelmann et al., 2023), Sentiment Analysis (Rathje et al., 2023; Xing, 2024; Zhang et al., 2023b,a), and Hate Speech Detection (Hee et al., 2024; García-Díaz et al., 2023) that utilize LLM are mainly carried out in English languages. Moreover, these works opened up the prospects of exploring LLMs for downstream tasks of low-resource languages.

The ability of LLMs to infer new language that was not used during the training could potentially be more beneficial for low-resource languages where it is challenging to train and deploy models due to the scarcity of quality datasets. However, most of the studies in the domain follow traditional machine learning or transformer-based models (Jahan and Oussalah, 2023; Chhabra and Vishwakarma, 2023; Islam et al., 2021; Hasan et al., 2020a). There are few attempts from re-

searchers across different languages to utilize LLM for low-resource languages. Although the number of works involving LLMs is insignificant, researchers of low-resource languages are leaning towards using LLM in recent times. Some notable works that utilize LLM are in low-resource languages such as Bengali (Liu et al., 2023; Hasan et al., 2023b; Kabir et al., 2023), Urdu(Koto et al., 2024), and Hindi (Kumar and Albuquerque, 2021; Koto et al., 2024) shows LLMs can achieve similar results to traditional machine learning techniques and transformer-based models.

For low-resource languages, there are significant research gaps in comparison with English. The literature on low-resource languages mainly focused on traditional deep learning and fine-tuning small language models. At the same time, large-scale development has been imposed for resource-rich languages like English. The works that use LLMs to solve downstream tasks in low-resource language are very limited, and the capabilities of LLMs have not been explored properly. To address these issues in this work, we aim to comprehensively use LLMs across several tasks for several low-resource languages such as Bangla, Urdu, and Hindi.

3 Background of LLMs

We evaluate the test set using three different LLMs. In this section, we discuss the LLM models used in this study in detail.

Generative Pre-trained Transformer 4 (GPT-4) (OpenAI, 2023) GPT-4 is one of the best-performing LLMs to predict the following document sequence, which OpenAI developed in March 2023. The model is trained on data up to September 2021 that supports multimodality. The model has over a trillion trainable parameters with a context length of $8,192$ and $32,768$ tokens. Furthermore, GPT-4 supports multilinguality with a strong performance in all languages. In our study, we used the context window of 8k tokens, which cost approximately $30 per one million tokens.

Llama 2 (Touvron et al., 2023) Llama 2 is one of the largest open-source LLMs released by Meta in 2023. The model has different versions (7B, 13B, and 70B) based on the parameter size of the model. The Llama 2 model has been trained on 2 trillion tokens with a context length of $4,096$. In our study, we used the *70B chat* version of the model, which is available through Huggingface Inference API[1]. The model is trained on publicly available text, instruction data, and one million human annotations. Although 90% of the training data belongs to English, it can generate results for 50+ languages.

Gemini (Team et al., 2023) Gemini is the most recent and one of the best-performing LLMs trained on top of the Transformer decoder along with multi-query attention (Shazeer, 2019). Gemini offers 4 different models based on the parameter size. Although the Nano-1 and Nano-2 versions have 1.8B and 3.25B trainable parameters, the parameter sizes for the Pro and Ultra versions are unknown. The Gemini models support multimodal data and are trained on a wide range of data (image, text, audio, and video) sources. Further, the model is trained on a sequence length of $32,768$ tokens to provide a better context understanding for the long texts. As a result, Gemini models offer a context length of $32,768$ tokens. Moreover, Gemini includes safety settings to prevent generating harmful content such as hate, offensive, derogatory, sexual harassment, etc. In our study, we only used the Pro version[2] of the Gemini model. Although the model supports 38 languages, including Bangla, English, Arabic, and Hindi, it does not currently support Urdu, one of the low-resource languages we analyse in this study.

4 Methodology

4.1 Prompt Approach

The performance of LLMs varies depending on the prompt content. Designing a good prompt is a complex and iterative process that requires substantial effort due to the unknown representation of information within the LLM. In this study, we applied zero-shot prompting by using natural language instructions. The instructions contain the task description and expected output, which enables the construction of a context to generate more appropriate output. We keep the same prompt for each task across the LLMs. Further, we added role information into the prompt for the GPT-4 model as GPT-4 can take the role information and perform accordingly. In our initial study, we noticed that the Gemini Pro model blocks most of the contents from sentiment and hate speech tasks to predict the

[1] https://huggingface.co/inference-api

[2] The Pro version costs approximately $10 per million tokens.

desired output due to harmful content in the dataset. To get the predictions for those harmful content, we additionally provide a safety setting that does not block any harmful content. However, the model still blocks derogatory languages. We provided the details of the prompts and safety settings in Appendix A.

4.2 Experimental Settings

4.2.1 Data

This section discusses the publicly available data for three tasks used in our study. We first discuss the data for the NLI task followed by the sentiment task and conclude with the hate speech task. Although each task has some datasets for all the languages individually, only the dataset of the NLI task has been translated into several languages. To fairly evaluate the generalization of LLMs, the translated version of the datasets is mandatory for other tasks. We provide a detailed description of data distribution in Table 1.

Task	Languages	Class	Test
NLI	EN, HI, UR	Contradiction	1,670
		Entailment	1,670
		Neutral	1,670
	BN	Contradiction	1,630
		Entailment	1,631
		Neutral	1,634
Sentiment	EN, BN, HI, UR	Negative	3,972
		Neutral	5,937
		Positive	2,375
Hate Speech	EN, BN, HI, UR	Hate	280
		Neither	821
		Offensive	3,856

Table 1: Class-wise test set data distribution for all the tasks. EN: English, BN: Bangla, HI: Hindi, and UR: Urdu.

NLI Task: We used the cross-lingual natural language inference (XNLI) dataset (Conneau et al., 2018) for the NLI task. The dataset extends the Multi-genre NLI dataset incorporating the raw text from the second release of the Open American National Corpus. The XNLI dataset is mainly developed for the English language and translated into 15 different languages including Hindi and Urdu languages using human annotators. Each data consists of a premise and hypothesis with a corresponding label[3]. During the development of the dataset, three different hypotheses were generated by the annotators based on the labels from each premise. We select the test set of English, Hindi, and Urdu languages from the XNLI dataset for our experiments. For the Bangla language, we used the translated version of XNLI (Bhattacharjee et al., 2021). The dataset is translated using the English to Bangla translator model described in (Hasan et al., 2020b). Although the dataset is translated from the XNLI dataset, the test set is short of 115 data from the original set.

Sentiment Task: For the sentiment analysis task, we used the official test of SemEval-2017 task 4: Sentiment Analysis in Twitter (Rosenthal et al., 2017). The raw texts were collected from X (formerly known as Twitter) and manually annotated them. Primarily, the annotation was completed in five classes which include Strongly Positive, Positive, Neutral, Negative, and Strongly Negative. Later, the labels were re-mapped into three classes where Strongly Positive was combined with Positive and Strongly Negative with Negative classes. The SemEval-2017 task 4 offered only English and Arabic data. In this study, we only incorporate the English data.

We translated the English test set for the Bangla, Hindi, and Urdu languages for evaluating the LLMs for the sentiment task. We used the web version of Google Translator[4] with the use of Deep Translator toolkit[5]. The quality of translations is moderate due to the tweet texts. We analyzed the translations and found that most of the hashtags were not translated into the target language. Moreover, Hindi translations were far better than Bangla and Urdu.

Hate Speech Task: We used the dataset described in (Davidson et al., 2017) for our hate speech task. The texts were collected from X (formerly known as Twitter) where the annotations were done manually into three categories that include 'Hate', 'Offensive', and 'Neither'. Each data was annotated by at least three people and the final label was consolidated by the majority. Few of the data were discarded where there was no majority. The official dataset consists of a total of 24,802 samples. We first split the data into train, validation, and test splits by 70%, 10%, and 20% respectively. We only used the test set in our study and the language of the official dataset is English.

[3]The class labels for the XNLI dataset are Contradiction, Entailment, and Neutral.

[4]https://translate.google.com

[5]https://pypi.org/project/deep-translator/

For Bangla, Hindi, and Urdu language datasets, we translated the English test set using Google Translator. We randomly sampled 100 entries from each translated dataset to conduct an analysis on the quality of translation. Our assessment revealed that while efforts were made, the quality of translation was found to be moderate, indicating room for improvement. Notably, certain elements such as hashtag words remained untranslated, and specific terms like 'hairspray,' 'oz,' numerical values, among others, were not adequately translated into their respective languages.

4.2.2 Data Pre-processing

The sentiment and hate speech datasets were mainly collected from X and contain URLs, usernames, hashtags, emoticons, and symbols. We only removed the URLs and usernames from the sentiment and hate speech task datasets. We keep the hashtags, emoticons, and symbols with data to understand how LLMs performed with this mixed information. Moreover, we did not perform any preprocessing steps for the XNLI dataset.

4.2.3 Evaluation Metrics

To evaluate our experiments, we calculated accuracy, precision, recall, and F_1 scores for all the tasks. We computed the weighted version of precision and recall and the macro version of F_1 score as it considers class imbalance.

5 Results and Discussion

This section presents and discusses the performances of different LLMs for English vs low-resource languages. Further, we also discuss the performances of our experiments using different LLMs for different tasks in this section. Table 2, Table 3, and Table 4 represent the performances of NLI, sentiment, and hate speech tasks.

5.1 English *vs* Low-resource Languages

In our study, our experiments show that all the LLMs consistently provide superior performances for English languages in all tasks except the performances of Gemini in the sentiment task. In the NLI task, the performance of GPT-4 in English is 18.04%, 17.38%, and 22.81% better than the Bangla, Hindi, and Urdu languages respectively (see Table 2). Although Hindi performs better than Bangla and Urdu, there is still a massive performance gap compared to English. Besides, Llama 2 performance in English is 32.52%, 31.28%, and 29.94% higher compared with Bangla, Hindi, and Urdu respectively. The difference between English and other languages is ∼70% from their original performance. Although the performance differences of Gemini between English and other languages are comparatively lower than GPT-4 and Llama 2, English is accomplishing approximately 13% better on average than Bangla, Hindi, and Urdu.

Model	Lang.	Acc.	P.	R.	$F1_{macro}$
GPT-4	**EN**	86.73	86.91	86.73	**86.79**
	BN	68.73	75.95	68.73	**68.75**
	HI	69.31	76.26	69.31	**69.41**
	UR	64.52	72.90	64.52	**63.98**
Llama 2	**EN**	74.47	76.27	74.47	74.82
	BN	45.66	52.74	45.66	42.30
	HI	47.29	65.68	47.29	43.54
	UR	46.39	53.68	46.39	44.88
Gemini	**EN**	78.40	78.06	78.40	78.12
	BN	67.24	69.32	67.24	67.16
	HI	66.48	68.67	66.48	66.50
	UR	62.14	65.38	62.14	62.01

Table 2: Performances of the NLI task across the models and languages. **Bold** indicates the best performances across the languages. Lang.: language, Acc.: accuracy, P.: Precision, R.: Recall, EN: English, BN: Bangla, HI: Hindi, and UR: Urdu

For the sentiment task, English is performing nearly on average 13% better than other languages using GPT-4 (see Table 3). The performance difference of Llama 2 between English and other languages is ∼ 11% on average, and English is consistently doing better than other languages. Despite that, Bangla, Hindi, and Urdu are performing 0.49%, 0.89%, and 0.60% better than English. The performance of Gemini remains almost the same for all the languages in the sentiment task. Our hate speech task experiments reveal that the performance of GPT-4 in English is approximately, on average, 22% better than low-resource languages (see Table 4). Moreover, the performances in English are ∼ 17% and ∼ 18% better than low-resource languages for Llama 2 and Gemini models.

We postulate the low performance of LLMs in low-resource languages for the following reasons. One of the main reasons is that most of the LLMs are trained on a large amount of English data, i.e., 90% of the training data of Llama 2 is English, whereas the amount of training data for low-resource languages is small compared with English.

Model	Lang.	Acc.	P.	R.	F1$_{macro}$
GPT-4	**EN**	72.64	73.05	72.64	**71.74**
	BN	61.33	64.57	61.33	56.36
	HI	66.47	68.75	66.47	63.68
	UR	62.31	64.89	62.31	58.19
Llama 2	**EN**	55.64	66.89	55.64	53.38
	BN	45.19	60.22	45.19	40.28
	HI	48.31	63.32	48.31	43.73
	UR	47.06	61.61	47.06	42.62
Gemini	**EN**	64.59	67.86	64.59	64.44
	BN	65.40	66.68	65.40	**64.93**
	HI	65.87	67.14	65.87	**65.33**
	UR	65.93	66.77	65.93	**65.14**

Table 3: Performances of the Sentiment task across the models and languages. **Bold** indicates the best performances across the languages. Lang.: language, Acc.: accuracy, P.: Precision, R.: Recall, EN: English, BN: Bangla, HI: Hindi, and UR: Urdu

Moreover, cultural differences between English-spoken countries and low-resource language countries affect the sentiment and hate speech tasks the most. Lastly, the quality of the translation affects the performance of low-resource languages. However, Hindi performed better than Bangla and Urdu in all tasks among the low-resource languages. The performance difference among the low-resource languages is insignificant across the tasks and LLMs. Our findings from this section conclude that improving LLMs is required for low-resource languages.

5.2 Comparison Among LLMs

We first analyzed the individual LLM outputs and found that GPT-4 could not predict much data on sentiment and hate speech tasks for Bangla and Urdu. Moreover, GPT-4 was able to provide predictions for all the English language samples for all the tasks. We also noticed that Llama 2 and Gemini models could predict all the samples from the NLI task for all languages. Llama 2 could not predict much data on the hate speech task for English. However, Llama 2 provides a small number of unpredicted data compared with GPT-4 for Bangla, Hindi, and Urdu. We analyzed the response of unpredicted data from GPT-4. We found that the model cannot understand the context to classify while Llama 2 could not predict due to inappropriate or offensive language. Moreover, the response from unpredicted samples from Llama includes repeated 'l' only. We briefly overview the unpredicted data in Figure 1. During the evaluation metrics calculation, we assigned the inverse classes for the unpredicted samples.

Model	Lang.	Acc.	P.	R.	F1$_{macro}$
GPT-4	**EN**	86.81	85.52	86.81	**62.54**
	BN	55.32	75.51	55.32	38.79
	HI	64.66	77.93	64.66	**44.61**
	UR	54.00	75.18	54.00	38.66
Llama 2	**EN**	79.32	83.93	79.32	60.04
	BN	69.92	69.12	69.92	**41.36**
	HI	74.54	71.58	74.54	44.39
	UR	47.29	65.68	47.29	**43.54**
Gemini	**EN**	58.00	77.69	58.00	49.10
	BN	30.34	70.93	30.34	30.81
	HI	32.01	72.72	32.01	33.36
	UR	28.56	70.07	28.56	28.47

Table 4: Performances of the Hate speech task across the models and languages. **Bold** indicates the best performances across the languages. Lang.: language, Acc.: accuracy, P.: Precision, R.: Recall, EN: English, BN: Bangla, HI: Hindi, and UR: Urdu

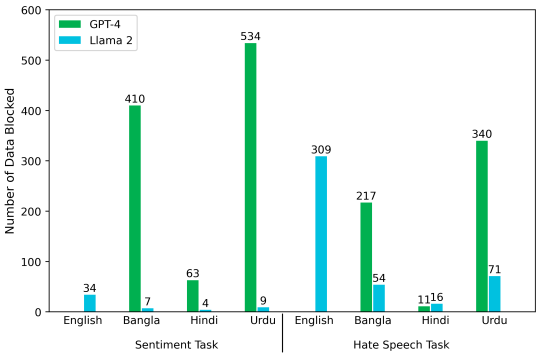

Figure 1: Number of unpredicted samples by GPT-4 and Llama 2. Note that we only include the languages and models from the tasks with unpredicted samples.

Gemini is the only LLM that predicted all the samples of each task. Although we provide a safety setting for the Gemini model, it blocked some data due to the content containing derogatory language. We noticed that the samples from sentiment and hate speech tasks were blocked for containing derogatory language, and those from the NLI task were not blocked. We provide a brief overview of the number of samples that are blocked by Gemini in Figure 2. However, the Urdu language is not supported by the Gemini. Despite that, the Gemini performs strongly in Urdu for the NLI and sentiment tasks. We further investigated the performances of Gemini in the Urdu language. We found

that the alphabets of Urdu are derived from the Arabic language family[6] and many words are adopted from the Arabic language. Arabic is supported by Gemini, and the training data of Arabic shares semantic information with the Urdu language, which is why Gemini exhibits a strong performance in the Urdu language.

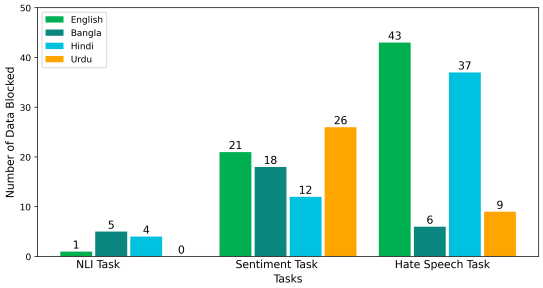

Figure 2: Number of samples that are blocked by Gemini.

Further, we investigated the detailed performances of each task. GPT-4 shows superior performances on the NLI task for all languages while exhibiting good performances on the sentiment task. However, most hate class data were misclassified in the hate speech task for all languages. Llama 2 provides strong performances in English for NLI, sentiment, and hate speech tasks while finding difficulties in accurately predicting the contradiction, neutral, and hate classes for NLI, sentiment, and hate speech tasks, respectively. Although Llama 2 outperforms GPT-4 performances in hate class in every language, GPT-4 in English and Hindi is better than Llama 2 for hate speech tasks. Moreover, Llama 2 demonstrated comparatively better performance on the hate speech task than NLI and sentiment tasks. While Gemini exhibits strong performances in NLI and sentiment tasks for all the languages, it consistently performs poorly on the speech task for all the languages. However, Gemini performs comparatively better hate class performance than Llama 2 and GPT-4 for all the languages. Moreover, the performances in the neither and offensive classes are worse than other LLMs. We also found that most offensive classes are misclassified as neither. We provided the detailed class-wise experimental results in Appendix B.

In general, GPT-4 shows prominent performances over other LLMs across all the tasks. Although Llama 2 provides better results for hate speech tasks, it struggled to perform well in NLI and sentiment tasks. While Gemini demonstrated strong performances in NLI and sentiment tasks, it delivered worse in hate speech tasks.

[6] https://en.wikipedia.org/wiki/Urdu_alphabet

5.3 Tasks Performances

NLI Task: We present the performances of different LLMs on the NLI task for English, Bangla, Hindi, and Urdu languages in Table 2. GPT-4 exhibits superior performances for the NLI task with F1 scores of 86.79, 68.75, 69.41, and 63.98 in English, Bangla, Hindi, and Urdu, respectively. Llama2 provides F1 scores of 74.82, 42.30, 43.54, and 44.88 for English, Bangla, Hindi, and Urdu respectively. The F1 scores of Gemini are 78.12, 67.16, 66.50, and 62.01 in English, Bangla, Hindi, and Urdu, respectively. While Gemini outperforms Llama 2 in this task, GPT-4 outperforms in the NLI task.

Sentiment Task: Table 3 represents the performances of different LLMs on the sentiment task for English, Bangla, Hindi, and Urdu. With the use of GPT-4 zero-shot learning, we achieved F1 scores of 71.74, 56.36, 63.68, and 58.19 in English, Bangla, Hindi, and Urdu languages, respectively, while the F1 scores of Llama 2 zero-shot learning for the sentiment task are 53.38, 40.28, 43.73, and 42.62 in English, Bangla, Hindi, and Urdu. We obtained F1 scores of 64.44, 64.93, 65.33, and 65.14 for English, Bangla, Hindi, and Urdu. Gemini demonstrated superior performances in Bangla, Hindi, and Urdu, while GPT-4 outperformed in English for the sentiment task.

Hate Speech Task: We present the performances of the hate speech task using GPT-4, Gemini, and Llama 2 models for all languages in Table 4. Using GPT-4 zero-shot learning, we achieved F1 scores of 62.54, 38.79, 44.61, and 38.66 in English, Bangla, Hindi, and Urdu respectively. F1 scores of 60.04, 41.36, 44.39, and 43.54 were achieved using Llama 2 in English, Bangla, Hindi, and Urdu, respectively. Moreover, the performance of GPT-4 is 0.22% higher than Llama 2 in the Hindi language, which is comparable. Further, the performance of Gemini is relatively worse than GPT-4 and Llama2 with F1 scores of 49.10, 30.81, 33.36, and 28.47 in English, Bangla, Hindi, and Urdu languages respectively.

The overall performance of the NLI task is comparatively better than sentiment and hate speech tasks. The definition of an NLI task has clear rules and structured patterns, while sentiment and hate

speech tasks are subjective and context-dependent. NLI task identifies the relation between two sentences based on structure and language logic (Bowman et al., 2015) that makes the task easier for LLMs. Moreover, the context lies with the sentence pair, and LLMs can understand the context. While sentiment and hate speech tasks require understanding the tone of the text and sometimes the complex social and cultural contexts, these facts are challenging for LLMs to understand. Moreover, the data of the NLI task is incorporated from the well-structured MNLI corpus with precise labels and balanced classes, making the task more comfortable for LLMs. Unlike the NLI task, sentiment and hate speech task data are curated from social media platforms containing noise, informal expressions, slang, and incomplete text, making it challenging for LLMs. Moreover, most of the texts do not have the contexts within their representation, and it is challenging to identify the context for both humans and LLMs. Straightforward linguistics features and contextual information make the NLI task easier and perform better than sentiment and hate speech tasks using different LLMs.

6 Conclusion

In conclusion, our comprehensive analysis sheds light on the pivotal role of large language models (LLMs) in the landscape of natural language processing (NLP), emphasizing both their remarkable performance in resource-rich languages such as English and the pressing need to extend their utility to low-resource language settings. Through our investigation of zero-shot learning with various LLMs, we have demonstrated that while LLMs, notably GPT-4, exhibit commendable capabilities in English, their performance in low-resource languages remains a subject of concern. This study underscores the importance of addressing the dearth of research and evaluation in low-resource language contexts, propelled by constraints including dataset scarcity and computational expenses. Our findings not only highlight the existing gap in linguistic capabilities for low-resource languages but also advocate for concerted efforts to bridge this divide. By focusing on tasks such as natural language inference (NLI) and considering performance across different LLM architectures, our research contributes valuable insights into the potential avenues for enhancing the applicability of LLMs in general-purpose NLP applications. Moving forward, concerted interdisciplinary efforts are warranted to bolster research initiatives aimed at refining LLM performance in low-resource language environments, thus fostering inclusivity and accessibility in the realm of natural language processing.

7 Limitation

In our study, we refrained from utilizing explicit prompting techniques to enhance the performance of large language models (LLMs). Our evaluation primarily focused on assessing LLMs in the context of English and low-resource languages such as Bangla, Hindi, and Urdu, without exploring variations in prompts. Regarding the quality of dataset translations, it is important to note that the translations generated by Google Translator were not subjected to human verification. Consequently, while certain translation errors were overlooked during our analysis, we conducted sampling from each translated dataset to gain insights into the overall translation quality. Our findings underscore the necessity for further refinement in translation methodologies to elevate both the quality and accuracy of translations in future research endeavors.

References

Sindhu Abro, Sarang Shaikh, Zahid Hussain Khand, Ali Zafar, Sajid Khan, and Ghulam Mujtaba. 2020. Automatic hate speech detection using machine learning: A comparative study. *International Journal of Advanced Computer Science and Applications*, 11(8).

Syeda Nahida Akter, Zichun Yu, Aashiq Muhamed, Tianyue Ou, Alex Bäuerle, Ángel Alexander Cabrera, Krish Dholakia, Chenyan Xiong, and Graham Neubig. 2023. An in-depth look at gemini's language abilities. *arXiv preprint arXiv:2312.11444*.

Firoj Alam, Arid Hasan, Tanvirul Alam, Akib Khan, Janntatul Tajrin, Naira Khan, and Shammur Absar Chowdhury. 2021. A review of bangla natural language processing tasks and the utility of transformer models. *arXiv preprint arXiv:2107.03844*.

Hind S Alatawi, Areej M Alhothali, and Kawthar M Moria. 2021. Detecting white supremacist hate speech using domain specific word embedding with deep learning and bert. *IEEE Access*, 9:106363–106374.

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.

Abhik Bhattacharjee, Tahmid Hasan, Kazi Samin, Md Saiful Islam, M. Sohel Rahman, Anindya Iqbal,

and Rifat Shahriyar. 2021. Banglabert: Combating embedding barrier in multilingual models for low-resource language understanding.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Anusha Chhabra and Dinesh Kumar Vishwakarma. 2023. A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*, pages 1–28.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515.

José Antonio García-Díaz, Ronghao Pan, and Rafael Valencia-García. 2023. Leveraging zero and few-shot learning for enhanced model generality in hate speech detection in spanish and english. *Mathematics*, 11(24):5004.

Reto Gubelmann, Aikaterini-Lida Kalouli, Christina Niklaus, and Siegfried Handschuh. 2023. When truth matters-addressing pragmatic categories in natural language inference (nli) by large language models (llms). In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (* SEM 2023)*, pages 24–39.

Md Arid Hasan, Firoj Alam, Anika Anjum, Shudipta Das, and Afiyat Anjum. 2023a. Blp-2023 task 2: Sentiment analysis. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 354–364.

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

Md Arid Hasan, Jannatul Tajrin, Shammur Absar Chowdhury, and Firoj Alam. 2020a. Sentiment classification in bangla textual content: A comparative study. In *2020 23rd international conference on computer and information technology (ICCIT)*, pages 1–6. IEEE.

Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M. Sohel Rahman, and Rifat Shahriyar. 2020b. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for Bengali-English machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2612–2623, Online. Association for Computational Linguistics.

Ming Shan Hee, Shivam Sharma, Rui Cao, Palash Nandi, Preslav Nakov, Tanmoy Chakraborty, and Roy Ka-Wei Lee. 2024. Recent advances in hate speech moderation: Multimodality and the role of large models. *arXiv preprint arXiv:2401.16727*.

Oskar Holmström, Jenny Kunz, and Marco Kuhlmann. 2023. Bridging the resource gap: Exploring the efficacy of english and multilingual llms for swedish. In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pages 92–110.

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.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, page 126232.

Mohsinul Kabir, Mohammed Saidul Islam, 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*.

Muhammad Moin Khan, Khurram Shahzad, and Muhammad Kamran Malik. 2021. Hate speech detection in roman urdu. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 20(1):1–19.

Fajri Koto, Tilman Beck, Zeerak Talat, Iryna Gurevych, and Timothy Baldwin. 2024. Zero-shot sentiment analysis in low-resource languages using a multilingual sentiment lexicon. *arXiv preprint arXiv:2402.02113*.

Md Kowsher, Md Shohanur Islam Sobuj, Nusrat Jahan Prottasha, Mohammad Shamsul Arefin, and Yasuhiko Morimoto. 2023. Contrastive learning for universal zero-shot nli with cross-lingual sentence embeddings. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 239–252.

Akshi Kumar and Victor Hugo C Albuquerque. 2021. Sentiment analysis using xlm-r transformer and zero-shot transfer learning on resource-poor indian language. *Transactions on Asian and Low-Resource Language Information Processing*, 20(5):1–13.

Xiaoyi Liu, Mao Teng, Shuangtao Yang, and Bo Fu. 2023. Knowdee at blp-2023 task 2: Improving bangla sentiment analysis using ensembled models with pseudo-labeling. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 273–278.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Chenyang Lyu, Jitao Xu, and Longyue Wang. 2023. New trends in machine translation using large language models: Case examples with chatgpt. *arXiv preprint arXiv:2305.01181*.

Ning Miao, Yee Whye Teh, and Tom Rainforth. 2023. Selfcheck: Using llms to zero-shot check their own step-by-step reasoning. *arXiv preprint arXiv:2308.00436*.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on bert model. *PloS one*, 15(8):e0237861.

Nanlir Sallau Mullah and Wan Mohd Nazmee Wan Zainon. 2021. Advances in machine learning algorithms for hate speech detection in social media: a review. *IEEE Access*, 9:88364–88376.

R OpenAI. 2023. Gpt-4 technical report. *arXiv*, pages 2303–08774.

Bhavish Pahwa and Bhavika Pahwa. 2023. Bphigh at semeval-2023 task 7: Can fine-tuned cross-encoders outperform gpt-3.5 in nli tasks on clinical trial data? In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1936–1944.

Nitin Rane. 2023. Enhancing mathematical capabilities through chatgpt and similar generative artificial intelligence: Roles and challenges in solving mathematical problems. *Available at SSRN 4603237*.

Steve Rathje, Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjieh, Claire Robertson, and Jay J Van Bavel. 2023. Gpt is an effective tool for multilingual psychological text analysis.

Nathaniel R Robinson, Perez Ogayo, David R Mortensen, and Graham Neubig. 2023. Chatgpt mt: Competitive for high-(but not low-) resource languages. *arXiv preprint arXiv:2309.07423*.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, Vancouver, Canada. Association for Computational Linguistics.

TYSS Santosh and KVS Aravind. 2019. Hate speech detection in hindi-english code-mixed social media text. In *Proceedings of the ACM India joint international conference on data science and management of data*, pages 310–313.

Noam Shazeer. 2019. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*.

Xiaofei Sun, Xiaoya Li, Shengyu Zhang, Shuhe Wang, Fei Wu, Jiwei Li, Tianwei Zhang, and Guoyin Wang. 2023. Sentiment analysis through llm negotiations. *arXiv preprint arXiv:2311.01876*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. 2023. Can chatgpt replace traditional kbqa models? an in-depth analysis of the question answering performance of the gpt llm family. In *International Semantic Web Conference*, pages 348–367. Springer.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Frank Xing. 2024. Designing heterogeneous llm agents for financial sentiment analysis. *arXiv preprint arXiv:2401.05799*.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. *arXiv preprint arXiv:2309.11674*.

Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. 2023a. Instruct-fingpt: Financial sentiment analysis by instruction tuning of general-purpose large language models. *arXiv preprint arXiv:2306.12659*.

Boyu Zhang, Hongyang Yang, Tianyu Zhou, Muhammad Ali Babar, and Xiao-Yang Liu. 2023b. Enhancing financial sentiment analysis via retrieval augmented large language models. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 349–356.

Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2023. Toolqa: A dataset for llm question answering with external tools. *arXiv preprint arXiv:2306.13304*.

Steven Zimmerman, Udo Kruschwitz, and Chris Fox. 2018. Improving hate speech detection with deep learning ensembles. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

A Prompts and Safety Setting

This section presents the details of the prompts that we used for each model and task[7]. We present the example prompt for the NLI task, sentiment task, and Hatespeech task in Table 5, Table 6, and Table 7 respectively. We provide the details of the safety setting for the Gemini Pro model in Table 8

Model	Prompt

GPT-4	[{
	'role': 'user',
	'content': "Classify the following 'premise' and 'hypothesis' into one of the following classes: 'Entailment', 'Contradiction', or 'Neutral'. Provide only label as your response."
	premise: [PREMISE_TEXT]
	hypothesis: [HYPOTHESIS_TEXT]
	label:
	},
	{
	role: 'system',
	content: "You are an expert data annotator and your task is to analyze the text and find the appropriate output that is defined in the user content."
	}]

Llama 2 and Gemini	Classify the following 'premise' and 'hypothesis' into one of the following classes: 'Entailment', 'Contradiction', or 'Neutral'. Provide only label as your response. premise: [PREMISE_TEXT] hypothesis: [HYPOTHESIS_TEXT] label:

Table 5: Prompts used for zero-shot learning in NLI task.

[7]Note that we use the same prompt for each task.

Model	Prompt

GPT-4	[{
	'role': 'user',
	'content': "Classify the 'text' into one of the following labels: 'Positive', 'Neutral', or 'Negative'. Provide only label as your response."
	text: [SOURCE_TEXT]
	label:
	},
	{
	role: 'system',
	content: "You are an expert data annotator and your task is to analyze the text and find the appropriate output that is defined in the user content."
	}]

Llama 2 and Gemini	Classify the 'text' into one of the following labels: 'Positive', 'Neutral', or 'Negative'. Provide only label as your response. text: [SOURCE_TEXT] label:

Table 6: Prompts used for zero-shot learning in Sentiment task.

Model	Prompt

GPT-4	[{
	'role': 'user',
	'content': "Classify the 'text' into one of the following labels: 'Hate', 'Offensive', or 'Neither'. Provide only label as your response."
	text: [SOURCE_TEXT]
	label:
	},
	{
	role: 'system',
	content: "You are an expert data annotator and your task is to analyze the text and find the appropriate output that is defined in the user content."
	}]

Llama 2 and Gemini	Classify the 'text' into one of the following labels: 'Hate', 'Offensive', or 'Neither'. Provide only label as your response. text: [SOURCE_TEXT] label:

Table 7: Prompts used for zero-shot learning in Hatespeech task.

Category	Threshold
HARM_CATEGORY_HARASSMENT	BLOCK_NONE
HARM_CATEGORY_HATE_SPEECH	BLOCK_NONE
HARM_CATEGORY_SEXUALLY_EXPLICIT	BLOCK_NONE
HARM_CATEGORY_DANGEROUS_CONTENT	BLOCK_NONE
HARM_CATEGORY_SEXUAL	BLOCK_NONE
HARM_CATEGORY_DANGEROUS	BLOCK_NONE

Table 8: Safety setting used for Gemini Pro model to prevent blocking the predictions for harmful content.

B Detailed Experimental Results

B.1 NLI Task

We present the detailed class-wise performances for the NLI task across the LLMs in Table 9.

Model	Lang.	Class	P.	R.	F1
GPT-4	EN	Contradiction	92.45	89.40	90.90
		Entailment	88.25	86.88	87.56
		Neutral	80.02	82.90	81.92
	BN	Contradiction	85.58	67.03	75.18
		Entailment	88.26	49.85	63.17
		Neutral	54.10	89.24	67.36
	HI	Contradiction	88.54	68.92	77.51
		Entailment	86.02	50.18	63.39
		Neutral	54.22	88.80	67.33
	UR	Contradiction	85.41	40.66	55.09
		Entailment	82.53	64.27	72.26
		Neutral	50.79	88.62	64.57
Llama 2	EN	Contradiction	94.12	73.83	82.75
		Entailment	72.88	83.17	77.68
		Neutral	61.82	66.41	64.03
	BN	Contradiction	65.80	13.93	22.99
		Entailment	54.66	57.20	55.90
		Neutral	37.81	65.79	48.02
	HI	Contradiction	88.30	14.91	25.51
		Entailment	70.72	41.80	52.54
		Neutral	38.01	85.15	52.56
	UR	Contradiction	63.88	22.87	33.69
		Entailment	59.63	46.17	52.04
		Neutral	37.54	70.12	48.90
Gemini	EN	Contradiction	84.24	90.24	87.14
		Entailment	77.76	80.00	78.87
		Neutral	72.17	64.95	68.37
	BN	Contradiction	72.90	78.81	75.57
		Entailment	79.22	53.35	63.76
		Neutral	55.88	69.57	61.97
	HI	Contradiction	74.14	75.36	74.73
		Entailment	77.08	53.21	62.96
		Neutral	54.82	70.88	61.82
	UR	Contradiction	70.14	70.06	70.10
		Entailment	75.27	45.81	56.98
		Neutral	50.62	70.54	58.94

Table 9: Class-wise performances of the NLI task across the models and languages. **Bold** indicates the best performances across the languages. Lang.: language, P.: Precision, R.: Recall, EN: English, BN: Bangla, HI: Hindi, and UR: Urdu

B.2 Sentiment Task

Detailed class-wise performances for the sentiment task across the LLMs are presented in Table 10.

B.3 Hatespeech Task

Table 11 reports the detailed class-wise performances for the hatespeech task across the LLMs.

Model	Lang.	Class	P.	R.	F1
GPT-4	EN	Negative	73.08	73.39	73.23
		Neutral	70.52	77.23	73.72
		Positive	79.36	59.92	68.28
	BN	Negative	71.29	39.88	51.15
		Neutral	57.40	85.11	68.56
		Positive	71.25	37.77	49.37
	HI	Negative	73.07	51.79	60.62
		Neutral	62.03	83.90	71.33
		Positive	78.32	47.45	59.10
	UR	Negative	72.34	43.01	53.95
		Neutral	58.45	83.43	68.74
		Positive	68.51	41.77	51.90
Llama 2	EN	Negative	56.08	94.26	70.32
		Neutral	81.81	16.89	28.01
		Positive	47.65	87.92	61.80
	BN	Negative	45.10	90.79	60.27
		Neutral	76.96	2.81	5.43
		Positive	43.66	74.89	55.16
	HI	Negative	48.31	93.78	63.77
		Neutral	80.45	4.78	9.03
		Positive	45.62	81.05	58.38
	UR	Negative	46.15	93.55	61.81
		Neutral	78.18	4.77	8.99
		Positive	46.05	75.03	57.07
Gemini	EN	Negative	60.40	87.89	71.60
		Neutral	76.83	46.38	57.84
		Positive	57.86	71.33	63.89
	BN	Negative	61.28	84.21	70.94
		Neutral	72.07	54.44	62.03
		Positive	62.23	61.42	61.82
	HI	Negative	62.57	83.42	71.51
		Neutral	71.36	57.17	63.48
		Positive	62.33	58.65	60.43
	UR	Negative	61.74	84.66	71.41
		Neutral	72.63	55.11	62.67
		Positive	62.41	61.42	61.91

Table 10: Class-wise performances of the Sentiment task across the models and languages. **Bold** indicates the best performances across the languages. Lang.: language, P.: Precision, R.: Recall, EN: English, BN: Bangla, HI: Hindi, and UR: Urdu

Model	Lang.	Class	P.	R.	F1
GPT-4	EN	Hate	62.96	12.14	20.36
		Offensive	88.85	95.10	91.87
		Neither	77.58	73.33	75.39
	BN	Hate	22.39	5.36	8.65
		Offensive	89.56	51.61	65.48
		Neither	27.62	89.77	42.25
	HI	Hate	32.69	6.07	10.24
		Offensive	90.97	63.49	74.68
		Neither	33.56	90.13	48.91
	UR	Hate	33.93	6.79	11.31
		Offensive	88.58	50.49	64.32
		Neither	26.30	86.60	40.35
Llama 2	EN	Hate	14.98	31.79	20.37
		Offensive	88.16	86.51	87.33
		Neither	87.56	61.75	72.43
	BN	Hate	13.35	17.50	15.15
		Offensive	80.82	85.14	82.92
		Neither	42.42	27.28	33.21
	HI	Hate	15.09	12.50	13.67
		Offensive	80.93	89.06	84.80
		Neither	46.89	27.53	34.69
	UR	Hate	11.98	18.57	14.57
		Offensive	80.05	83.87	81.91
		Neither	37.27	21.92	27.61
Gemini	EN	Hate	14.95	76.34	25.00
		Offensive	88.87	55.49	68.32
		Neither	46.97	63.41	53.97
	BN	Hate	8.62	79.93	15.56
		Offensive	83.14	20.36	32.71
		Neither	34.83	60.29	44.16
	HI	Hate	8.27	81.65	15.01
		Offensive	83.90	22.50	35.49
		Neither	42.47	59.51	49.57
	UR	Hate	8.76	76.43	15.72
		Offensive	83.20	18.53	30.31
		Neither	29.49	59.20	39.37

Table 11: Class-wise performances of the Hatespeech task across the models and languages. **Bold** indicates the best performances across the languages. Lang.: language, P.: Precision, R.: Recall, EN: English, BN: Bangla, HI: Hindi, and UR: Urdu