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 their remarkable performance in various downstream tasks in resource-rich languages. Recent studies have highlighted the limitations of LLMs in low-resource languages, primarily focusing on binary classification tasks and giving minimal attention to South Asian languages. These limitations are primarily attributed to constraints such as dataset scarcity, computational costs, and research gaps specific to low-resource languages. To address this gap, we present datasets for sentiment and hate speech tasks by translating from English to Bangla, Hindi, and Urdu, facilitating research in low-resource language processing. Further, we comprehensively examine zero-shot learning using multiple LLMs in English and widely spoken South Asian 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 natural language inference (NLI) exhibits the highest performance among the evaluated tasks, with GPT-4 demonstrating superior capabilities.

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[1] for many NLP downstream tasks are not widely evaluated, leaving a notable gap in the linguistic capabilities of low-resource languages. The most widely spoken yet low-resource languages of South Asia[2] such as Bangla, Hindi, and Urdu, several researchers are handling the scarcity of datasets and other resources in NLI (Aggarwal et al., 2022), Sentiment analysis (Hasan et al., 2023b; Sun et al., 2023; Koto et al., 2024) 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.

[1] Refers to the scarcity of datasets and other resources rather than limitations in LLM capabilities.

[2] https://simple.wikipedia.org/wiki/Languages_of_South_Asia

- 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 English.
- We apply zero-shot prompting using natural language instructions, which describe the task and expected output, enabling constructing a context to generate more appropriate output.

2 Related Works

LLMs are proficient in various NLP tasks and highly generalizable across multiple domains. However, their performance remains significant room for improvement, particularly in low-resource languages such as Bangla, Hindi, and Urdu. Previous study (Robinson et al., 2023) demonstrates the inability of LLMs such as GPT-4 to perform on low-resource (African) and high-resource languages. However, LLMs perform well in languages (European) that use the same script as English (Holmström et al., 2023).

NLP research works, and applications for several downstream tasks 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., 2020; Islam et al., 2021; Hasan et al., 2023a) that attained attention from many researchers to concentrate on solving this issue. Some of the recent works on NLI (Pahwa and Pahwa, 2023; Gubelmann et al., 2023), Sentiment Analysis (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.

There are few attempts from researchers across different languages to utilize LLM for low-resource languages (Hasan et al., 2023b; Kabir et al., 2023; Koto et al., 2024; Kumar and Albuquerque, 2021) that show LLMs can achieve similar results to traditional machine learning techniques and transformer-based models. However, existing multilingual benchmarks such as BUFFET (Asai et al., 2023), XTREME (Hu et al., 2020), and XTREME-R (Ruder et al., 2021) do not address all four South Asian low-resource languages we are considering in our study. Moreover, BUFFET is limited to binary classification tasks and uses few-shot learning and instruction fine-tuning of smaller LLMs (such as mT5, mT0) and ChatGPT. At the same time, we focus on multi-class classification and use zero-shot learning with SOTA LLMs. The performance of LLMs is not balanced for all languages (Huang et al., 2023; Qin et al., 2023), and our study uniquely focuses on comparing resource-rich (English) and low-resource (Bangla, Hindi, and Urdu) languages using SOTA LLMs.

Previous studies have highlighted LLM limitations in low-resource languages, particularly in binary classification, with minimal focus on South Asian languages. These constraints include dataset scarcity, high computational costs, and specific research gaps. To address these challenges, we concentrate on South Asian languages like Bangla, Urdu, and Hindi. We provide datasets for sentiment and hate speech tasks by translating from English. We explore zero-shot learning techniques across English and South Asian languages, thus expanding LLM applications in low-resource settings.

3 Methodology

We focused on both open- and closed-source LLMs. We choose three LLMs that are GPT-4 (OpenAI, 2023), Llama 2 (Touvron et al., 2023), and Gemini Pro (Team et al., 2023). We select the LLMs based on their performances, parameter sizes, and capabilities. To conduct our experiments, we used the XNLI dataset (Conneau et al., 2018) for the NLI task, the official test of SemEval-2017 task 4 (Rosenthal et al., 2017) for the sentiment task, and the dataset described in (Davidson et al., 2017) for hate speech task. We provide the details of the dataset used and the detailed data preprocessing and evaluation metrics in Appendix B.

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. We also provide a safety setting for the Gemini model to avoid blocking harmful content. See Appendix A for details.

4 Results and Discussion

English *vs* Low-resource Languages: 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 (Table 1). 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 1). 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 $\sim$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.

For the sentiment task, English is performing nearly on average 13% better than other languages using GPT-4 (see Table 1). The performance difference of Llama 2 between English and other languages is $\sim$ 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 1). Moreover, the performances in English are $\sim$ 17% and $\sim$ 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. 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.

Comparison Among LLMs: We first analyzed

Model	Lang.	Acc.	P.	R.	$F1_{macro}$
		NLI Task			
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
		Sentiment Task			
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**
		Hate Speech Task			
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 1: Performances of all the tasks across the models and languages. **Bold** indicates the best performances across the languages for each task. Lang.: language, Acc.: accuracy, P.: Precision, R.: Recall, EN: English, BN: Bangla, HI: Hindi, and UR: Urdu

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, some responses of Llama include repeated 'l' as the label. We briefly overview the unpredicted data in Figure 1. During the evaluation metrics calculation, we assigned the inverse classes for the 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[3] 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.

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. Despite observing a smaller performance gap in Gemini, significant disparities persist in GPT-4 and Llama-2, indicating that direct translation is less likely to compromise sentiment information. See Appendix B for class-wise experimental results.

Tasks Performances: The overall performance of the NLI task is comparatively better than sentiment and hate speech tasks (Table 1). 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. In addition, during the evaluation, we explored whether English hashtags have any impact on predictions for Bangla, Hindi, and Urdu. Our empirical results demonstrated that LLMs do not rely solely on hashtags but on the entire sequence.

5 Conclusion

In this study, we introduce datasets for sentiment and hate speech tasks by translating from English to Bangla, Hindi, and Urdu to facilitate research in low-resource language processing. Through a comprehensive examination of zero-shot learning across multiple LLMs, notably GPT-4, we uncover performance disparities between English and low-resource languages. Furthermore, our analysis identifies NLI as a task where GPT-4 consistently demonstrates superior capabilities, underscoring avenues for enhancing LLM applicability in general-purpose NLP applications.

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

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

Divyanshu Aggarwal, Vivek Gupta, and Anoop Kunchukuttan. 2022. Indicxnli: Evaluating multilingual inference for indian languages. *arXiv preprint arXiv:2204.08776*.

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*.

Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2023. Buffet: Benchmarking large language models for few-shot cross-lingual transfer. *arXiv preprint arXiv:2305.14857*.

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*.

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. 2020. Sentiment classification in bangla textual content: A comparative study. In *2020 23rd international conference on computer and information technology (ICCIT)*, pages 1–6. IEEE.

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.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. *arXiv preprint arXiv:2305.07004*.

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.

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*.

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.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv preprint arXiv:2307.16039*.

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*.

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.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. *arXiv preprint arXiv:2310.14799*.

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

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.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, et al. 2021. Xtreme-r: Towards more challenging and nuanced multilingual evaluation. *arXiv preprint arXiv:2104.07412*.

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.

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*.

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*.

A Prompts and Safety Setting

This section presents the details of the prompts that we used for each model and task[4]. We present the example prompt for the NLI task, sentiment task, and Hatespeech task in Table 2, Table 3, and Table 4 respectively. We provide the details of the safety setting for the Gemini Pro model in Table 5

B Experimental Details and Results

B.1 Experimental Settings

B.1.1 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 6.

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

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 2: Prompts used for zero-shot learning in NLI 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 3: Prompts used for zero-shot learning in Sentiment task.

NLI Task: We used the cross-lingual natural language inference (XNLI) dataset (Conneau et al., 2018) for the NLI task. We select the test set of English, Hindi, and Urdu languages from the XNLI dataset for our experiments. For the Bangla lan-

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 4: Prompts used for zero-shot learning in Hate-speech 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 5: Safety setting used for Gemini Pro model to prevent blocking the predictions for harmful content.

guage, we used the translated version of XNLI (Bhattacharjee et al., 2021).

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). Primarily, the annotation was completed in five classes and then the labels were re-mapped into three classes.The SemEval-2017 task 4 offered only English and Arabic data. In this study, we only incorporate the English data.

Hate Speech Task: We used the dataset described in (Davidson et al., 2017) for our hate speech task. 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.

Translation: We translated the English test set for the Bangla, Hindi, and Urdu languages to evaluate the LLMs for sentiment and hate speech tasks. We used the web version of Google Translator[5] with the use of Deep Translator toolkit[6]. 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. We also randomly sampled 100 translation pairs for each language from both tasks to check the translation quality by native speakers. The feedback from native speakers indicates that there is room for improvement in the translation quality. Additionally, it is important to note that we followed previous best practices used in similar studies (Aggarwal et al., 2022; Lai et al., 2023).

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

B.1.2 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.

B.1.3 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]https://translate.google.com
[6]https://pypi.org/project/deep-translator/

B.2 Detailed Results

We investigated the detailed performances of each task (see Table 7, Table 8, and Table 9). 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.

B.2.1 NLI Task

We present the detailed class-wise performances for the NLI task across the LLMs in Table 7.

B.2.2 Sentiment Task

Detailed class-wise performances for the sentiment task across the LLMs are presented in Table 8.

B.2.3 Hatespeech Task

Table 9 reports the detailed class-wise performances for the hatespeech task across the LLMs.

C Experimental Analysis

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 7: 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

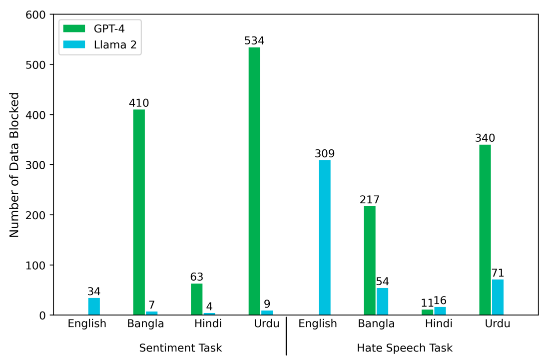

Figure 1: Number of unpredicted samples by GPT-4 and Llama 2. Note that we only include the languages and models from the tasks with unpredicted samples.

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 8: 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

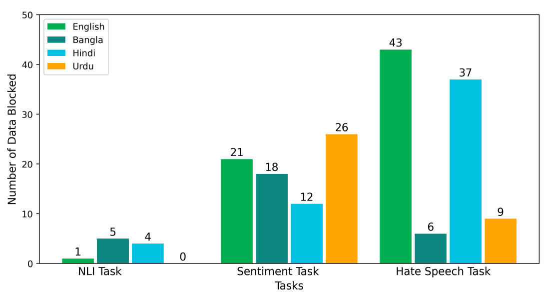

Figure 2: Number of samples that are blocked by Gemini.

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 9: 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