On Limitations of LLM as Annotator for Low Resource Languages

Suramya Jadhav[1,3], Abhay Shanbhag[1,3], Amogh Thakurdesai[1,3], Ridhima Sinare[1,3], and Raviraj Joshi[2,3]
[1]Pune Institute of Computer Technology, Pune
[2]Indian Institute of Technology Madras, Chennai
[3]L3Cube Labs, Pune

Abstract

Low-resource languages face significant challenges due to the lack of sufficient linguistic data, resources, and tools for tasks such as supervised learning, annotation, and classification. This shortage hinders the development of accurate models and datasets, making it difficult to perform critical NLP tasks like sentiment analysis or hate speech detection. To bridge this gap, Large Language Models (LLMs) present an opportunity for potential annotators, capable of generating datasets and resources for these underrepresented languages. In this paper, we focus on Marathi, a low-resource language, and evaluate the performance of both closed-source and open-source LLMs as annotators, while also comparing these results with fine-tuned BERT models. We assess models such as GPT-4o and Gemini 1.0 Pro, Gemma 2 (2B and 9B), and Llama 3.1 (8B and 405B) on classification tasks including sentiment analysis, news classification, and hate speech detection. Our findings reveal that while LLMs excel in annotation tasks for high-resource languages like English, they still fall short when applied to Marathi. Even advanced models like GPT-4o and Llama 3.1 405B underperform compared to fine-tuned BERT-based baselines, with GPT-4o and Llama 3.1 405B trailing fine-tuned BERT by accuracy margins of 10.2% and 14.1%, respectively. This highlights the limitations of LLMs as annotators for low-resource languages.

1 Introduction

Even with advancements in NLP, the curation of annotations for supervised tasks like sentiment analysis, text classification, and inference has been the primary responsibility of human linguistic experts (Tan et al., 2024). Data annotations play an integral part in both building and evaluating a model. Hence, the quality and reliability of data lie at the core of the performance and usefulness of the model being built.

The process of curating good-quality data annotations is expensive in terms of time and cost, specifically when it comes to compiling data annotations for low-resource languages. The aim of this study is to explore whether Large Language Models (LLMs) can be effectively leveraged to create supervised data resources for low-resource languages, with Marathi as the focus in this case.

Recent generative models like ChatGPT have shown competitive quality in data annotations for simpler tasks like sentiment analysis while human expert annotations proved to be better for intricate tasks Nasution and Onan (2024). ChatGPT was evaluated by Zhu et al. (2023) to check its capability of reproducing human-generated labels for social computing tasks. In these experiments, ChatGPT obtained an average accuracy of 0.60 with 0.64 being the highest accuracy for the sentiment analysis task. In addition to these, the works of Kuzman et al. (2023); Gao et al. (2023) have previously evaluated ChatGPT's performance with that of human experts. Experiments performed by Mohta et al. (2023) demonstrated that Vicuna 13b performed reasonably well for numerous annotation tasks compared to other models that were tested like Vicuna 7b, Llama (13b, 7b) and InstructBLIP(13b, 7b). However, it is important to note that most of these experiments target the English language.

India is a multilingual nation with various regional languages and most of these languages fall under the low-resource (LR) category. Low resource languages are languages such as Marathi and Hindi that lack annotated training datasets and have very few task-specific resources compared to high resource languages such as Spanish and English.

This paper presents a case study on the performance of Large Language Models (LLMs) in annotating the low-resource language Marathi. We conduct a comprehensive comparative analysis of

various closed-source and open-source LLMs, revealing that many LLMs still fall significantly short of the baseline performance achieved by BERT-based models and are not yet capable of replacing human annotators.

Specifically, we evaluated models such as GPT-4o, Gemini 1.0 Pro, Gemma 2 (2B and 9B), Llama 3.1 (8B and 405B) across multiple tasks, including 3-class sentiment analysis, 2-class, and 4-class hate speech detection, as well as news classification based on headlines, long paragraphs, and full documents.

The key contributions of this research work are as follows:

- We have conducted a first-of-its-kind detailed comparative study between fine-tuned BERT models and large language models (LLMs), by evaluating their potential to be used as annotators for a low-resource language, Marathi.

- We observe that the average results of the Few-shot prompting technique outperform the average result of the Zero-shot prompting technique in all the models tested.

- We have provided valuable insights into the effectiveness of both open- and closed-source large language models (LLMs), including GPT-4o, Llama 3.1 405B, Llama 3.1 8B, Gemma 2 9B, Gemma 2 2B, and Gemini 1.0 Pro, on tasks such as Marathi Sentiment Analysis, Hate Speech Detection, and News Categorization. Our results strongly demonstrate that LLMs are still not fully reliable for annotation tasks in Indic languages.

- Model ranking, based on accuracy metrics, is GPT-4o > Llama-3.1-405B > Gemini 1.0 Pro > Gemma 2 9B > Llama 3.1 8B > Gemma-2-2B.

The paper is structured as follows: Section 2 provides a concise review of prior research on data annotation and the use of LLMs. In Section 3, we detail the datasets used and the Section 4 describes models employed in our evaluation. Section 5 describes the experimental setup and the APIs leveraged to evaluate the LLMs. Section 6 presents the results, along with a comparative analysis of various LLMs and BERT-based models, highlighting the key findings of our research. Finally, in Section 7, we conclude our discussion.

2 Literature Review

Many low-resource languages, including Marathi, lack well-annotated datasets, making it difficult to train effective models for tasks like sentiment analysis and classification Al-Wesabi et al. (2023). The absence of sufficient data often leads to poor performance in tasks that require labeled corpora R et al. (2023).

Low-resource languages also present unique linguistic challenges not well-represented in high-resource models Krasadakis et al. (2024), highlighting the need for specialized approaches. With the rise of LLMs, these models have been explored as a solution to mitigate the scarcity of annotated data in low-resource languages.

Several works demonstrate the use of LLMs as annotators for low-resource language tasks. Pavlovic and Poesio (2024) reviewed LLMs like GPT-4 and noted performance drops when handling non-English languages. In Kholodna et al. (2024), the authors explored the integration of large language models (LLMs), specifically GPT-4 Turbo, into an active learning framework designed for low-resource language tasks. Their work demonstrates the use of few-shot learning to generate useful annotations, significantly enhancing performance on low-resource tasks. Additionally, they implemented the GPT-4 Turbo model as a classifier within the training loop, leading to a substantial reduction in annotation costs, which were 42.45 times lower compared to traditional methods. However, the general performance of LLMs remains limited, especially for languages with fewer resources Hedderich et al. (2020).

The studies of Ding et al. (2022) and Mohta et al. (2023) further evaluated LLM performance on multilingual datasets, with results indicating that models like GPT-3 and open-source LLMs struggle with non-English data. Srivastava et al. (2022) showed that increasing model size does not consistently enhance performance for low-resource languages, unlike high-resource languages like English.

Bias is another concern with LLMs. Bavaresco et al. (2024) introduced JUDGE-BENCH to evaluate LLM biases, noting that training data heavily influences model outputs, which can be problematic in annotating complex or sensitive tasks in low-resource languages. While LLMs used for high-resource languages (HRL) are giving promising results, that is not the case for low-

resource languages. Nasution and Onan (2024) explored ChatGPT-4's performance in annotation tasks across languages like Turkish and Indonesian, offering insights into LLM applicability for Low Resource Language(LRL), a relevant consideration for our focus on Marathi.

3 Dataset

In this research, we focus on three major task categories using relevant Marathi datasets:

1) MahaSent (Kulkarni et al., 2021; Pingle et al., 2023) – classifies sentiment of Marathi tweets into three classes of positive, negative, or neutral categories.

2) MahaHate (Patil et al., 2022) – measures the level of abusive and hostile content in Marathi text. This dataset includes two supervised tasks: MahaHate 2-Class, which categorizes content as either HATE or NOT, and MahaHate 4-Class, which provides finer distinctions with categories: Hate (HATE), Offensive (OFFN), Profane (PRFN), and Not (NOT).

3) MahaNews (Mittal et al., 2023; Mirashi et al., 2024) – classifies headlines and articles from Marathi news. It comprises three supervised datasets: Short Headlines Classification (SHC), Long Document Classification (LDC), and Long Paragraph Classification (LPC), each categorizing news content into 12 classes: Auto, Bhakti, Crime, Education, Fashion, Health, International, Manoranjan, Politics, Sports, Tech, and Travel. The distribution of all the mentioned datasets is provided in Table 2

4 Methodolgy

We investigated the distinctions between LLM-generated and human-generated annotations for the Indic language, Marathi, using a comparative methodology, and analyzed the results with fine-tuned BERT-based models for detailed insights.

4.1 LLMs

In our annotation experiments, we evaluated the performance of LLMs for the Marathi language using two prompting techniques: zero-shot and few-shot learning. We tested both open-source models (Llama 3.1 8B, Llama 3.1 405B, Gemma 2 2B, and Gemma 2 9B) and closed-source models (Gemini 1.0 Pro, GPT-4o), and compared their results with BERT-based models. The performance of each LLM under both prompting strategies is summarized in Table 1.

4.2 BERT Based Models

We used fine-tuned BERT-based models to compare performance with LLMs, where MahaSent-MD MahaHate-BERT, MahaNews-SHC-BERT, MahaNews-LPC-BERT, and MahaNews-LDC-BERT are fine-tuned versions of MahaBERT, while MahaHate-multi-RoBERTa has MahaRoBERTa as the base model. Each of these models was fine-tuned on the corresponding datasets, and their respective performances are summarized in Table 1.

5 Experimental Setup

Our main objective is to assess the LLMs on three different tasks and related datasets to ascertain whether LLMs could take the place of, or at least support, human annotation efforts. We employed both few-shot and zero-shot prompting techniques, with LLM-generated annotations evaluated against the ground truth labels. For all datasets, the test split was used. The open-source models (Llama 3.1 8B, Gemma 2 2B, and Gemma 2 9B) exhibited slower response times and required significant computational resources to generate predictions. However, by utilizing NVIDIA NIM APIs, we were able to accelerate predictions from these models, improving both speed and precision. For the closed-source Gemini 1.0 Pro model, we used the Gemini API, while GPT-4o predictions were generated manually via ChatGPT's default settings to annotate the samples. In our research, we could only use a subset of samples from each dataset due to the restrictive usage regulations and cost limits of the mentioned APIs. To maintain consistency and fairness in the performance comparison, all results from both LLM-based and BERT-based models were evaluated on a uniform subset. Specifically, we evaluated 490 samples from the MahaSent and MahaHate datasets, while for MahaNews, we selected 40 samples from each of the 12 classes, amounting to a total of 480 samples.

6 Result

This section provides a detailed overview of the experiments conducted for the annotation of three distinct tasks, utilizing six large language models (LLMs) and six BERT-based models (BERT model fine-tuned on target task). Table 1 summarizes the performance metrics of the fine-tuned BERT-based

Dataset	Tech	Llama 3.1 8B	Gemma 2 2B	Gemma 2 9B	Gemini 1.0 Pro	Llama 3.1 405B	GPT-4o	Fine Tuned BERT
MahaSent	ZS	0.76	0.71	0.69	0.78	0.77	0.79	0.80
	FS	0.79	0.76	0.78	0.76	0.81	**0.82**	
MahaHate-2C	ZS	0.64	0.71	0.78	0.74	0.77	0.80	**0.91**
	FS	0.78	0.72	0.82	0.72	0.82	0.82	
MahaHate-4C	ZS	0.40	0.39	0.43	0.43	0.49	0.58	**0.73**
	FS	0.48	0.41	0.46	0.45	0.52	0.60	
MahaNews-SHC	ZS	0.60	0.54	0.68	0.68	0.75	0.78	**0.85**
	FS	0.66	0.54	0.68	0.70	0.74	0.78	
MahaNews-LPC	ZS	0.66	0.55	0.71	0.72	0.76	0.77	**0.89**
	FS	0.67	0.50	0.72	0.74	0.76	0.75	
MahaNews-LDC	ZS	0.69	0.62	0.78	0.74	0.76	0.81	**0.96**
	FS	0.69	0.62	0.80	0.75	0.78	0.81	
Average	ZS	0.625	0.587	0.678	0.682	0.716	0.755	**0.857**
	FS	0.678	0.592	0.710	0.687	0.738	0.763	

Table 1: Model Comparison across different tasks. Tech: Different Prompting Techniques Used; ZS: Zero Shot; FS: Few Shot; 2C: 2-Class; 4C: 4-Class; SHC: Short Headlines Classification; LDC: Long Document Classification; LPC: Long Paragraph Classification; BERT: Refer Section 4.2 for details about BERT models.

Split	MahaSent	MahaHate 2-C	MahaHate 4-C	SHC	LDC	LPC
Train	12114	30000	21500	22014	22014	42870
Valid	1500	3750	2000	2750	2750	5366
Test	2500	3750	1500	2761	2761	5357

Table 2: Dataset Distribution

models, offering a comparative analysis against the performance of each LLM under both few-shot and zero-shot prompting scenarios. The table facilitates a comprehensive evaluation by highlighting key outcomes, enabling a thorough understanding of how each model performs across the different annotation tasks and prompting methods.

6.1 Key Findings

Our extensive experiments revealed crucial insights, showing that Large Language Models (LLMs) are not yet fully equipped to serve as reliable annotators for the Marathi language. The disparity between LLM-based and human-generated annotations remains substantial. Even for straightforward tasks like news classification, LLM performance was suboptimal. For more complex tasks, such as the 4-class MahaHate classification, their performance was notably disappointing, as evidenced in Table 1.

Among the LLMs evaluated, GPT-4o achieved the best results compared to others, including Llama 3.1 8B, Gemma 2 (2B and 9B), and Gemini 1.0 pro. However, both open-source and closed-source LLMs exhibited notable limitations in providing accurate and reliable annotations. Our results also demonstrate that closed LLMs like GPT-4o and Gemini 1.0 Pro outperform open LLMs namely Llama and Gemma 2B and 9B but they still underperform when compared to finetuned BERT for almost all datasets.

Our evaluation ranks the models as GPT-4o > Llama 3.1 405B > Gemini 1.0 Pro > Gemma 2 9B > Llama 3.1 8B > Gemma 2 2B, highlighting that open source Llama 3.1 405B outperforms Gemini 1.0 Pro and is second only to GPT-4o.

Compared to zero-shot prompting, few-shot prompting produced more accurate results because it gave examples of the desired input-output behavior, helping the model to understand the task's context and expectations better. We observe an absolute increase in the average accuracy of few-shot prompting by 5.3%, 0.5%, 3.2%, 0.5%, 2.2%, and 0.8% compared to zero-shot prompting for models Llama 3.1 8B, Gemma 2 2B, Gemma 2 9B, Gemini 1.0 Pro, Llama 3.1 405B, and GPT-4o, respectively. While few-shot prompting techniques yielded better accuracy than zero-shot approaches, they still fell short of the performance delivered by BERT-based models.

The average accuracy gap between open-source and closed-source models is 6.7%, while the difference between closed-source models and fine-tuned BERT-based models is 13.9%, highlighting the lack of effective LLMs for low-resource languages.

BERT-based fine-tuned models on target task outperformed LLMs in the classification tasks for Marathi language because finetuning enabled better knowledge extraction and alignment with the task's requirements. On the other hand, LLMs, despite being trained on vast amounts of general data, lack performance on low resource languages and task-specific optimization, which limits their ability to extract the most relevant features for a specific task. This suggests that, despite the increasing popular-

ity of LLMs, BERT-based models continue to be highly relevant, particularly for Indic languages.

We also note that the difference in the results of BERT-based models and the LLMs is comparatively less for easy tasks like the Sentiment Analysis Task, i.e. in the MahaSent dataset. At the same time, the gap is significantly higher for complex tasks like the Hate Classification and News Classification tasks in favor of BERT-based models.

7 Conclusion

Our study demonstrates that while LLMs like GPT, Gemini, Gemma, and Llama show potential, they currently fall short of being reliable annotators for low-resource languages like Marathi, particularly for complex tasks. BERT-based models continue to outperform LLMs in these contexts. Moreover, these findings can be generalized to other Indic languages as well, such as Marathi, due to their morphological richness. These results indicate that further advancements are required in LLMs to make them viable alternatives to human annotations. This research highlights the need for developing more robust models tailored to the specific details of low-resource languages. This includes the creation of higher-quality, task-specific datasets for low-resource languages, ensuring better representation and reducing biases. Enhanced datasets combined with domain-specific knowledge can significantly improve annotation accuracy.

Acknowledgement

This work was carried out under the mentorship of L3Cube, Pune. We would like to express our gratitude towards our mentor for his continuous support and encouragement. This work is a part of the L3Cube-MahaNLP project (Joshi, 2022).

References

Fahd N. Al-Wesabi, Hala J. Alshahrani, Azza Elneil Osman, and Elmouez Samir Abd Elhameed. 2023. Low-resource language processing using improved deep learning with hunter–prey optimization algorithm. *Mathematics*.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fern'andez, Albert Gatt, E. Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andr'e F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2024. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. *ArXiv*, abs/2406.18403.

Bosheng Ding, Chengwei Qin, Linlin Liu, Lidong Bing, Shafiq R. Joty, and Boyang Albert Li. 2022. Is gpt-3 a good data annotator? In *Annual Meeting of the Association for Computational Linguistics*.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554*.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strotgen, and Dietrich Klakow. 2020. A survey on recent approaches for natural language processing in low-resource scenarios. In *North American Chapter of the Association for Computational Linguistics*.

Raviraj Joshi. 2022. L3cube-mahanlp: Marathi natural language processing datasets, models, and library. *arXiv preprint arXiv:2205.14728*.

Nataliia Kholodna, Sahib Julka, Mohammad Khodadadi, Muhammed Nurullah Gumus, and Michael Granitzer. 2024. Llms in the loop: Leveraging large language model annotations for active learning in low-resource languages. *Preprint*, arXiv:2404.02261.

Panteleimon Krasadakis, Evangelos Sakkopoulos, and Vassilios S. Verykios. 2024. A survey on challenges and advances in natural language processing with a focus on legal informatics and low-resource languages. *Electronics*.

Atharva Kulkarni, Meet Mandhane, Manali Likhitkar, Gayatri Kshirsagar, and Raviraj Joshi. 2021. L3cubemahasent: A marathi tweet-based sentiment analysis dataset. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 213–220.

Taja Kuzman, Igor Mozetic, and Nikola Ljubešic. 2023. Chatgpt: beginning of an end of manual linguistic data annotation. *Use Case of Automatic Genre Identification. ArXiv abs/2303.03953*.

Aishwarya Mirashi, Srushti Sonavane, Purva Lingayat, Tejas Padhiyar, and Raviraj Joshi. 2024. L3cube-indicnews: News-based short text and long document classification datasets in indic languages. *Preprint*, arXiv:2401.02254.

Saloni Mittal, Vidula Magdum, Sharayu Hiwarkhedkar, Omkar Dhekane, and Raviraj Joshi. 2023. L3cube-mahanews: News-based short text and long document classification datasets in marathi. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 52–63. Springer.

Jay Mohta, Kenan Emir Ak, Yan Xu, and Mingwei Shen. 2023. Are large language models good annotators? In *ICBINB*.

Arbi Haza Nasution and Aytuğ Onan. 2024. Chatgpt label: Comparing the quality of human-generated and llm- generated annotations in low-resource language nlp tasks. *IEEE Access*, 12:71876–71900.

Hrushikesh Patil, Abhishek Velankar, and Raviraj Joshi. 2022. L3cube-mahahate: A tweet-based marathi hate speech detection dataset and bert models. In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, pages 1–9.

Maja Pavlovic and Massimo Poesio. 2024. The effectiveness of llms as annotators: A comparative overview and empirical analysis of direct representation. *ArXiv*, abs/2405.01299.

Aabha Pingle, Aditya Vyawahare, Isha Joshi, Rahul Tangsali, and Raviraj Joshi. 2023. L3cube-mahasent-md: A multi-domain marathi sentiment analysis dataset and transformer models. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 274–281.

Girija V R, Sudha T, and Riboy Cheriyan. 2023. Analysis of sentiments in low resource languages: Challenges and solutions. *2023 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE)*, pages 1–6.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. *arXiv preprint arXiv:2402.13446*.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145*.