Yes-MT's Submission to the Low-Resource Indic Language Translation Shared Task in WMT 2024

Yash Bhaskar[1], **Parameswari Krishnamurthy**[2]
IIIT Hyderabad
yash.bhaskar@research.iiit.ac.in, param.krishna@iiit.ac.in

Abstract

This paper presents the systems submitted by the Yes-MT team for the Low-Resource Indic Language Translation Shared Task at WMT 2024 (Pakray et al., 2024), focusing on translating between English and the Assamese, Mizo, Khasi, and Manipuri languages. The experiments explored various approaches, including fine-tuning pre-trained models like mT5 (Xue et al., 2020) and IndicBart (Dabre et al., 2021) in both Multilingual and Monolingual settings, LoRA (Hu et al., 2021) finetune IndicTrans2 (Gala et al., 2023), zero-shot and few-shot prompting (Brown, 2020) with large language models (LLMs) like Llama 3 (Dubey et al., 2024) and Mixtral 8x7b (Jiang et al., 2024), LoRA Supervised Fine Tuning (Mecklenburg et al., 2024) Llama 3, and training Transformers (Vaswani, 2017) from scratch. The results were evaluated on the WMT23 Low-Resource Indic Language Translation Shared Task's test data using SacreBLEU (Post, 2018) and CHRF (Popović, 2015) highlighting the challenges of low-resource translation and show the potential of LLMs for these tasks, particularly with fine-tuning.

1 Introduction

Developing robust machine translation systems for India's diverse languages is crucial given the country's growing economic importance and the increasing availability of digital content. However, a significant challenge in developing effective translation tools arises from the limited availability of data for many Indian languages, particularly those spoken in the northeastern regions. This paper describes the Yes-MT team's efforts to address this challenge by participating in the WMT 2024 Low-Resource Indic Language Translation Shared Task, focusing on English to Assamese, Mizo, Khasi, and Manipuri translation. We explored techniques like fine-tuning pre-trained models (mT5, IndicBart) and utilizing large language models (LLMs) like Llama 3 and Mixtral for zero-shot and few-shot learning. Furthermore, we explored using the LoRA technique to fine-tune the IndicTrans2 model, and we also trained Transformer models from scratch. Our findings provide valuable insights into the strengths and weaknesses of different approaches, highlighting the potential of LLMs and fine-tuning techniques in overcoming the limitations of data scarcity.

2 Dataset

The dataset used in this study consists of parallel bilingual data provided by the WMT 2024 Low-Resource Indic Language Translation Shared Task organizers (Pal et al., 2023) & (Pakray et al., 2024). The training, validation, and test splits for each language pair are detailed in Table 1.

Language Pair	Train	Val	Test
Assamese (en-as)	50,000	2,000	2,000
Mizo (en-lus)	50,000	1,500	2,000
Khasi (en-kha)	24,000	1,000	1,000
Manipuri (en-mni)	21,000	1,000	1,000

Table 1: Number of Sentences in Train, Validation, and Test Sets

In addition to the bilingual data, we also had access to a significant amount of Monolingual data for each of the target languages, which included 2.60 million sentences in Assamese, 1.90 million sentences in Mizo, 0.18 million sentences in Khasi, and 2.10 million sentences in Manipuri. However, for the scope of this work, we focused exclusively on utilizing the provided bilingual data for training and evaluation, aiming to explore the capabilities of the models under truly low-resource conditions.

Limiting our study to the provided bilingual data allowed us to maintain a consistent and controlled experimental environment, ensuring the results reflected the performance of our approaches under

Proceedings of the Ninth Conference on Machine Translation, pages 788–792

the typical constraints of low-resource language translation tasks. In the future, we may explore incorporating the available monolingual data, such as through back-translation, to further improve translation quality.

3 Experiments

This section details the experimental setup used for the various models and training strategies employed in our submission.

3.1 Primary Submission

Our primary submission involved training a Transformer model from scratch using the Fairseq framework (Ott et al., 2019). This model was trained for Multilingual translation, handling all four language directions (English to Assamese, Manipuri, Mizo, and Khasi) simultaneously. We utilized BPE tokenizer (Araabi et al., 2022) and Transformer architecture. The architectural details are shown in Table 2.

Parameter	Value
Embedding Dimension	512
FFN Dimension	1024
Attention Heads	4
Encoder Layers	6
Decoder Layers	6

Table 2: Transformer Architecture Details

3.2 Contrastive Submission

The contrastive submission explored fine-tuning pre-trained models in two settings: language-specific and Multilingual.

3.2.1 Multilingual Fine-tuning:

Both mT5 and IndicBart were fine-tuned in a Multilingual setting, where a single model was trained to handle all four language directions. To enable the models to distinguish between the target languages, we added language-specific tokens to their existing vocabularies, as suggested by previous work (Johnson et al., 2017). The language-specific tokens used are shown in Table 3. A single model was trained for one-to-many translation across all four language directions for each of the indicBart, mT5-small, and IndicTrans2 systems. The results are in Table 4. IndicBart and mT5-small were fine-tuned using Full Fine-Tuning (FFT), while IndicTrans2 was fine-tuned employing the LoRA (Low-Rank Adaptation) technique (Hu et al., 2021).

Language	Token
Assamese (asm)	‘<asm_Beng>‘
Manipuri (mni)	‘<mni_Beng>‘
Khasi (kha)	‘<kha_Latn>‘
Mizo (lus)	‘<lus_Latn>‘

Table 3: Language-Specific Tokens

3.2.2 Monolingual Fine-tuning:

We also trained separate models for each language pair, as these focused on a single translation direction and did not require language-specific tokens.

For each language direction, we trained four distinct models using mT5-Small and IndicBart with Full Fine-Tuning (FFT). The results are in Table 4.

3.3 Experiments with LLMs

Additionally, we explored the use of the Llama3 model in conjunction with the LoRA (Low-Rank Adaptation) technique.

Zero-Shot and Few-Shot Translation Evaluation We tested Zero Shot Translation capabilities of Llama 3-8B-8192, Llama 3-70B-8192, mixtral-8x7B-32768, Llama3-8B-instruct and Llama3.1-8B-instruct. We also tested the few-shot translation capabilities of Llama3.1-8B-instruct with 3-shot, 5-shot, and 10-shot prompting.

Supervised Fine-Tuning with LoRA We fine-tuned a 4-bit quantized (Liu et al., 2023) Llama3 model using the LoRA technique with Supervised Fine-Tuning (SFT), employing the LlamaFactory framework (Zheng et al., 2024). We used a prompt-based approach for translation, providing the model with a system prompt and a prompt template specifying the source and target languages.

The following template was used for fine-tuning the Large Language Models (LLMs):

```
System Prompt : You are a helpful assistant.
Prompt Template : Translate the following
English sentence to {target_language} in
{target_script} Script:\n{input_sent}
```

4 Results

4.1 Multilingual vs. Monolingual Performance

One key finding from our experiments was the performance comparison between the Multilingual and Monolingual training approaches for the mT5 and IndicBart models. As shown in Table 4, the

Model	Training Type	en-as	en-kha	en-mz	en-mni
Transformers	Multilingual	16.06	19.67	5.49	20.60
IndicBart	Monolingual	6.4	11.2	25.1	8.8
	Multilingual	6.5	11.4	25.3	9.1
mT5-small	Monolingual	14.3	12.9	31.4	19.2
	Multilingual	15.6	13.6	32.3	23.9
IndicTrans2-2B	ZeroShot	49.2	-	-	44.9
IndicTrans2-200M	ZeroShot	49.5	-	-	45.3
	Multilingual	47.27	-	-	49.12

Table 4: ChrF Scores for Monolingual : Models fine-tuned for one-to-one language translation
Multilingual : Models fine-tuned for one-to-many language translation

Multilingual versions of both mT5 and IndicBart consistently outperformed their Monolingual counterparts across the translation tasks.

- For mT5, the Multilingual model outperformed the Monolingual model across all language pairs, with ChrF score improvements ranging from 1.3 to 4.7 points. This suggests that mT5 benefits from the shared linguistic knowledge across different languages in a Multilingual setting, which enhances its ability to generalize to low-resource languages.
- Likewise, IndicBart demonstrated a slight performance boost in the Multilingual setting compared to the Monolingual models, suggesting that the Multilingual training approach provided a benefit.

The better performance of the Multilingual models is likely due to the shared linguistic knowledge they gained during training, which may have provided a richer context and improved their ability to generalize. This indicates that leveraging Multilingual data, even in limited-resource scenarios, can be a more effective approach than focusing on Monolingual training.

4.2 Expected Structured Output

A challenge observed during the experiments was the generation of structured output. Ideally, the output should directly provide the translated sentence without additional, unnecessary text. However, we noticed that the LLM models sometimes wrapped the translation in extraneous text, such as "The translation of the given sentence is: Translation", followed by further analysis and explaination making it difficult to extract the translation. This adds noise to the output and complicates the process of extracting the actual translation.

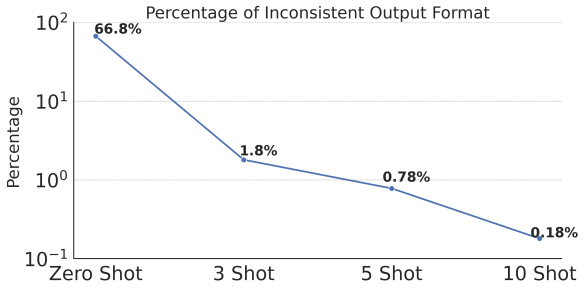

Figure 1: Inconsistent Output Format with Few Shot Prompting

We analyzed the percentage of outputs that were wrapped with unnecessary text across different settings:

This issue of unnecessary text in the output was more common in the zero-shot setting, where 66.% of the outputs included additional text. As the number of shots increased, the percentage of such outputs decreased significantly to 0.18% in 10 Shot Prompting, indicating that few-shot prompting can help guide the LLM to produce more structured and concise translations.

To improve the usability of LLM-based machine translation systems, it's crucial to fine-tune the models or design prompts that consistently yield clean and structured outputs, particularly in low-resource settings where post-processing resources might be limited.

5 WMT 2024 Results

The performance of our models on the WMT 2024 Low-Resource Indic Language Translation Shared Task dataset is summarized in the following table, focusing on the ChrF (Popović, 2015) metrics:

For the primary submissions, we utilized Transformers trained from scratch without additional data. As indicated by the scores, the primary

Model	Inference	en-as	en-kha	en-mz	en-mni
Llama3-8B-8192	Zero Shot	18.56	14.92	15.57	13.45
Llama3-70B-8192	Zero Shot	27.54	18.57	20.62	15.53
mixtral-8x7B-32768	Zero Shot	6.79	15.45	16.57	2.65
Llama3-8B-instruct	Zero Shot	26.13	8.38	18.06	15.29
	1 Epoch	29.82	33.19	32.72	37.85
	2 Epoch	31.68	35.26	37.73	44.51
Llama3.1-8B-instruct	Zero Shot	22.93	12.03	15.23	14.47
	3 Shot	23.26	13.66	18.89	15.30
	5 Shot	23.48	15.11	18.77	15.29
	10 Shot	23.89	16.03	19.39	15.43

Table 5: ChrF Scores for Various Models, Shot Types, and Language Pairs

Language Pair	Submission Type	ChrF
Eng-Asm	primary	0.1123
	contrastive	0.6518
Eng-Mni	primary	0.1102
	contrastive	0.4438
Eng-Lus	primary	0.1282
	contrastive	0.4151
Eng-Kha	primary	0.1139
	contrastive	0.3541

Table 6: ChrF Scores for WMT 2024 Shared Task

systems struggled significantly, yielding very low ChrF values across all language pairs.

In contrast, the models fine-tuned for the contrastive submissions demonstrated noticeable improvements. For Assamese and Manipuri, we fine-tuned IndicTrans2, achieving the highest ChrF scores in these language pairs. For Mizo and Khasi, we fine-tuned Llama3, which also resulted in enhanced performance compared to the primary systems. These findings highlight the effectiveness of fine-tuning pre-trained models, even in low-resource settings.

6 Potential Test Set Bias

One of the noteworthy observations in this year (2024) WMT 2024 results is the significant difference in the performance of the primary Transformers trained from scratch when evaluated on this year's (2024) test set compared to last year's (2023) test set. Specifically, we observed that the models performed better on last year's test set despite using the same training data.

This discrepancy could be indicative of a translation bias present in last year's dataset, which might have inadvertently favored the models trained on that data. The primary systems, having been trained exclusively on the previous year's data, may have overfitted to patterns specific to that dataset, leading to better performance on the older test set but struggling on the newer one.

This implies that the primary models may have difficulty generalizing to entirely new data distributions, an important factor to consider in low-resource settings where the training data is limited and may not be representative of future data. It also underscores the importance of using diverse and varied datasets during training to help mitigate such biases and improve the overall robustness of the models.

7 Conclusion

This paper presented the systems and results of the Yes-MT team's participation in the WMT 2024 Low-Resource Indic Language Translation Shared Task. The experiments highlighted the potential of LLMs, especially when fine-tuned with techniques such as LoRA, in enhancing translation quality even under low-resource conditions. The contrastive submissions, which utilized fine-tuned LLMs, demonstrated significant improvements over the primary submissions that relied on training Transformers from scratch.

Our findings suggest that while training models from scratch can be challenging in low-resource settings due to data scarcity and generalization issues, fine-tuning pre-trained models can effectively bridge the gap, leveraging shared knowledge across languages to achieve better translation performance.

Future work could explore integrating monolingual data through back-translation or other data augmentation techniques, as well as further refin-

ing prompt engineering strategies to improve the structure and clarity of LLM outputs. Additionally, focusing on addressing potential biases in test data to help create more reliable translation systems.

Acknowledgments

We acknowledge the organizers of the WMT 2024 Low-Resource Indic Language Translation Shared Task for providing the valuable dataset and facilitating this research. We also thank the developers of the mT5, IndicBart, IndicTrans2, Llama 3, and Mixtral models for making their work publicly available.

References

Ali Araabi, Christof Monz, and Vlad Niculae. 2022. How effective is byte pair encoding for out-of-vocabulary words in neural machine translation? *arXiv preprint arXiv:2208.05225*.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint ArXiv:2005.14165*.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh M Khapra, and Pratyush Kumar. 2021. Indicbart: A pre-trained model for indic natural language generation. *arXiv preprint arXiv:2109.02903*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jay Gala, Pranjal A Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, et al. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Shih-yang Liu, Zechun Liu, Xijie Huang, Pingcheng Dong, and Kwang-Ting Cheng. 2023. Llm-fp4: 4-bit floating-point quantized transformers. *arXiv preprint arXiv:2310.16836*.

Nick Mecklenburg, Yiyou Lin, Xiaoxiao Li, Daniel Holstein, Leonardo Nunes, Sara Malvar, Bruno Silva, Ranveer Chandra, Vijay Aski, Pavan Kumar Reddy Yannam, et al. 2024. Injecting new knowledge into large language models via supervised fine-tuning. *arXiv preprint arXiv:2404.00213*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Partha Pakray, Santanu Pal, Advaitha Vetagiri, Reddi Mohana Krishna, Arnab Kumar Maji, Sandeep Kumar Dash, Lenin Laitonjam, Lyngdoh Sarah, and Riyanka Manna. 2024. Findings of wmt 2024 shared task on low-resource indic languages translation. In *Proceedings of the Ninth Conference on Machine Translation (WMT)*.

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. 2023. Findings of the wmt 2023 shared task on low-resource indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 682–694.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Ashish Vaswani. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.