RBG-AI: Benefits of Multilingual Language Models for Low-Resource Languages

Barathi Ganesh HB
RBG AI Research, RBG.AI
Coimbatore, Tamil Nadu, India
hb.bg@rbg.ai

Michal Ptaszynski
Kitami Institute of Technology
Kitami, Hokkaido, Japan
michal@mail.kitami-it.ac.jp

Abstract

This paper investigates how multilingual language models benefit low-resource languages through our submission to the WMT 2025 Low-Resource Indic Language Translation shared task. We explore whether languages from related families can effectively support translation for low-resource languages that were absent or underrepresented during model training. Using a quantized multilingual pretrained foundation model, we examine zero-shot translation capabilities and cross-lingual transfer effects across three language families: Tibeto-Burman, Indo-Aryan, and Austroasiatic. Our findings demonstrate that multilingual models failed to leverage linguistic similarities, particularly evidenced within the Tibeto-Burman family. The study provides insights into the practical feasibility of zero-shot translation for low-resource language settings and the role of language family relationships in multilingual model performance. The code used for reproducing the experiments is publicly available at https://github.com/rbg-research/EMNLP-2025.

1 Introduction

The development of Multilingual Machine Translation (MMT) systems presents a critical opportunity to bridge communication gaps for the world's estimated 7,000+ languages. Among them many remain severely under-resourced in digital contexts. While high-resource languages like English, French, and Chinese have abundant parallel training data, the vast majority of languages particularly those spoken by smaller communities lack sufficient data for effective Neural Machine Translation (NMT) system development.

The WMT 2025 Indic Machine Translation shared task (WMT[1]) provides an ideal testbed for investigating how multilingual language models can benefit low-resource languages. The Indic language family presents a diverse landscape of linguistic resources, ranging from relatively high-resource languages like Hindi and Bengali to extremely low-resource languages with minimal digital presence. This diversity allows us to examine crucial questions about cross-lingual transfer and zero-shot translation capabilities.

A fundamental question in multilingual Natural Language Processing (NLP) is whether languages from related families can effectively support translation for languages that were absent or severely underrepresented during model training. The Northeast Indian linguistic landscape presents a diverse range of language families, including Tibeto-Burman, Indo-Aryan, and Austroasiatic. These families offer rich opportunities to study cross-lingual transfer phenomena due to their distinct linguistic features. This includes morphological patterns, syntactic structures, and varying degrees of relatedness.

Our experiment investigates three core questions: (1) Can multilingual models achieve practical zero-shot translation quality for truly low-resource languages? (2) How do linguistic relationships across different language families influence cross-lingual transfer effectiveness? (3) What are the practical limitations and opportunities for deploying such systems in resource-constrained environments where low-resource languages are typically spoken?

Using the pre-trained MMT model as our foundation, we conduct systematic experiments to evaluate zero-shot translation performance and cross-lingual transfer effects. To ensure practical applicability under resource constrained environment, we implement 4-bit quantization to enable deployment on consumer hardware, addressing the reality that communities speaking low-resource languages often have limited access to high-end computational resources.

[1] https://www2.statmt.org/wmt25/indic-mt-task.html, accessed on August 2025

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 991–997
November 8–9, 2025. ©2025 Association for Computational Linguistics

2 Related Work

Cross-lingual transfer learning has emerged as a promising approach for supporting low-resource languages by leveraging knowledge from related high-resource languages. Early work demonstrated that multilingual NMT models could achieve reasonable performance on unseen language pairs through parameter sharing (Ha et al., 2016; Arivazhagan et al., 2019). However, the mechanisms underlying successful cross-lingual transfer remained poorly understood.

Recent studies have shown that linguistic similarity plays a crucial role in transfer effectiveness. Kocmi and Bojar (2018) demonstrated that typologically similar languages benefit more from multilingual training, while Lin et al. (2019) found that shared script and language family membership are strong predictors of transfer success. Winata et al. (2021) further showed that multilingual models develop language-agnostic representations that facilitate zero-shot transfer, particularly within language families.

The concept of "cursing of multilinguality" suggests that adding more languages to multilingual models can hurt performance on existing languages (Conneau et al., 2020). However, Wang et al. (2018) argued that this effect is primarily observed when languages are linguistically distant, and that related languages can actually benefit from shared training. Ha et al. (2016) first demonstrated zero-shot translation in multilingual NMT, showing that models could leverage shared representations to translate between unseen language pairs via pivoting through shared languages.

Subsequent work has explored the conditions under which zero-shot translation succeeds. Arivazhagan et al. (2019) found that zero-shot performance is highly dependent on the linguistic similarity between source and target languages and the pivot languages seen during training. Pires et al. (2019) showed that multilingual BERT representations are most effective for cross-lingual transfer when languages share scripts and belong to the same family.

For Indic languages specifically, Sen et al. (2019) demonstrated that Sanskrit-related languages show strong cross-lingual transfer effects. Inline with it, Dabre et al. (2020) found that multilingual Indic models benefit from careful language grouping based on linguistic relationships. However, most prior work has focused on relatively high-resource Indic languages, leaving questions about truly low-resource scenarios largely unexplored.

The role of language families in NMT has received increasing attention as researchers seek to understand the linguistic factors that enable successful multilingual models. Tan et al. (2019) showed that language family membership is one of the strongest predictors of multilingual model success, outperforming surface-level similarity measures. Within the Indo-European family, studies have shown particular promise for cross-lingual transfer. Kunchukuttan and Bhattacharyya (2020) demonstrated that Indo-Aryan languages share sufficient structural similarity to enable effective multilingual training, while Tamil and other Dravidian languages require different modeling approaches due to their distinct linguistic properties.

Recent work on massively multilingual models like mT5 and MADLAD-400 has shown that scaling to hundreds of languages can improve zero-shot performance, but the specific benefits for extremely low-resource languages remain unclear (Xue et al., 2021; Kudugunta et al., 2023). Our work addresses this gap by systematically evaluating how language family relationships influence zero-shot translation quality in truly low-resource settings. On other hand, recent advances in model compression, particularly quantization techniques, have made large multilingual models more accessible (Dettmers et al., 2023). However, the interaction between model compression and cross-lingual transfer performance has received limited attention, particularly for low-resource languages where even small performance degradations can be significant.

3 Data and Methodology

Our evaluation uses the WMT 2025 shared task datasets, supplemented with low-resource language pairs to enable comprehensive analysis (Pakray et al., 2025, 2024; Pal et al., 2023; Kakum et al., 2023). This dataset provides an excellent testbed for cross-lingual transfer analysis: five Tibeto-Burman languages allow us to examine within-family transfer effects, while Assamese (Indo-Aryan) and Khasi (Austroasiatic) serve as cross-family comparison points to evaluate transfer limitations across different linguistic lineages.

Our final submitted system is built upon the MADLAD-400 model, which extends the T5 architecture to support multilingual translation across 400+ languages. The model was selected after qualitative benchmarking against several multilingual

Criteria	mBART-50	MADLAD-400	NLLB-200
Total Languages Coverage	50	**400+**	200+
Model Parameters	610M	3B	**1.3B**
Training and Inference Computation	Low	**Medium**	High
Indic Languages Covered	Limited	**High**	**High**
Low Resource Languages Coverage	Low	**Strong**	**Strong**
BLUE Score on 100 Samples*	-	**47.3**	34.8

Table 1: Qualitative Benchmarking: Preferred values are highlighted in bold. *100 random samples from each language pair in the training corpus were used.

alternatives, including mBART-50 and NLLB-200 (Tang et al., 2020; Team et al., 2022). As given in Table 1, the selection criteria prioritized: (1) Coverage of target Indic languages, (2) Translation quality on low-resource language pairs, (3) Computational efficiency. MADLAD-400 demonstrated superior performance across these dimensions, particularly for the Indic languages included in the shared task. Additionally to address computational constraints, we implement 4-bit quantization, that 4x times reduces the model's memory footprint, enabling deployment on consumer GPUs (Dettmers et al., 2023). The quantization process preserves model accuracy through careful handling of outlier weights and strategic bit allocation.

We utilize the T5-compatible tokenizer associated with MADLAD-400, which handles the diverse scripts and writing systems of Indic languages. The tokenizer includes special tokens for language direction specification. Each input sentence is prepended with a language-specific tag indicating the target language (e.g., <2en> for English, <2hi> for Hindi). This approach enables bidirectional translation within a single model while maintaining translation quality. The system employs beam search decoding with a beam size of 5 to improve translation fluency and adequacy. Additional parameters include length penalty adjustment and early stopping criteria optimized for the target language characteristics.

During the fine-tunning process, we have combined data from all seven language pairs into a comprehensive bidirectional translation dataset. For each language pair, we created translation examples in both directions: English-to-target language and target-language-to-English. This approach doubled our effective training data while enabling the model to learn translation patterns in both directions. Source texts were prefixed with appropriate language direction tokens following the MADLAD format specification. The combined dataset was split using stratified sampling to ensure balanced representation across languages, resulting in 370,060 training samples, 43,537 validation samples, and 21,769 test samples.

4 Experiments: Model Adaptation

We first established baseline performance using zero-shot inference with the pre-trained MADLAD-400 3B model. The model was loaded with 4-bit quantization using QLoRA (Qunatized Low-Rank Adaptation) settings to reduce memory requirements while maintaining performance. Zero-shot translation was performed by prepending language-specific direction tokens to source sentences, following the format used during pre-training where tokens such as <2as> indicate translation to Assamese and <2en> indicates translation to English.

The zero-shot baseline results on testset revealed significant variation in performance across language pairs and translation directions. For translation into English, the model achieved the highest performance on Manipuri-to-English with a BLEU score of 23.2, followed by Khasi-to-English at 19.4 and Bodo-to-English at 19.0. Assamese-to-English showed moderate performance with 11.9 BLEU, while other language pairs demonstrated limited zero-shot capabilities, with Kokborok-to-English, Nyishi-to-English, and Mizo-to-English scoring 13, 1, and 2 BLEU respectively. Translation from English to target languages showed considerably lower performance across all pairs, with most achieving single-digit BLEU scores. English-to-Manipuri performed best at 7 BLEU, while several pairs including English-to-Assamese, English-to-Nyishi, and English-to-Mizo achieved minimal scores of 1, 0, and 1 BLEU respectively. These baseline results highlighted the model's stronger capability for translating into English compared to generating text in low-resource languages, establishing the need for targeted fine-tuning to improve bidirectional translation performance.

We employed Parameter Efficient Fine-tuning (PEFT) using Low-Rank Adaptation (LoRA) to adapt the pre-trained model to our specific language pairs. The LoRA configuration used a rank of 32 with an alpha value of 32 and dropout rate of 0.1. We targeted key attention and feed-forward components including query, value, key, and output projection layers as well as the intermediate dense layers in the feed-forward networks. The base model was prepared for k-bit training using gradient checkpointing to optimize memory usage during training. This configuration resulted in 94.4 million trainable parameters, representing only 3.1% of the total model parameters, which significantly reduced computational requirements while maintaining model expressiveness.

Training was conducted on a single RTX4090 GPU using with half-precision floating point (FP16) to accelerate computation and reduce memory consumption. We used a per-device batch size of 16 with gradient accumulation across 2 steps, creating an effective batch size of 32 samples per optimization step. The learning rate was set to 5e-5 with a weight decay of 0.01 to prevent overfitting. Training proceeded for 3 epochs with model checkpoints saved every 1000 training steps, retaining only the 2 most recent checkpoints to manage storage requirements. We employed the batched data collator with dynamic padding aligned to multiples of 8 tokens for optimal GPU utilization efficiency.

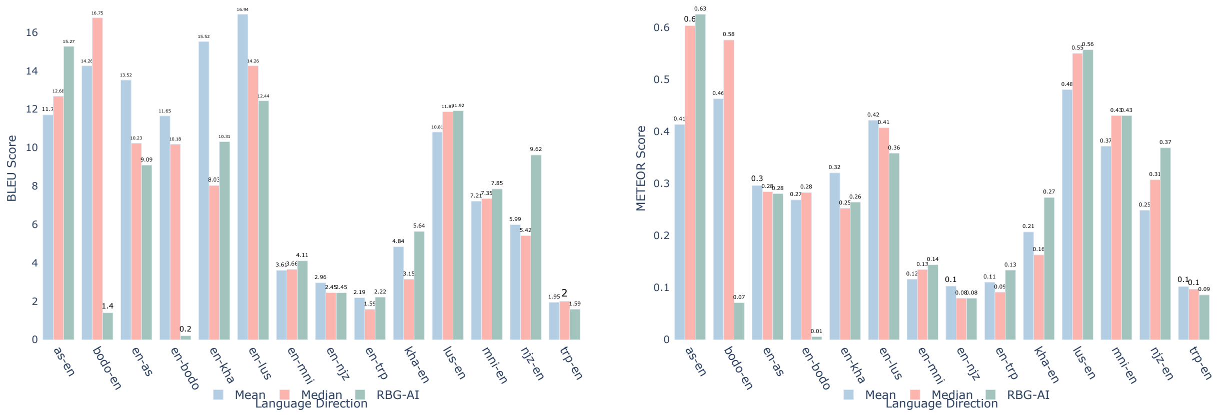
The training process utilized the sequence-to-sequence model trainer from the Transformers library, which handled the complete training loop including automatic loss computation and backpropagation through the LoRA adapters. Both source and target sequences were limited to a maximum length of 256 tokens to balance computational efficiency with sequence coverage. During evaluation phases, the model used greedy decoding with a maximum generation length of 256 tokens. The complete training process required approximately 24 hours to finish all three epochs. All experiments were conducted with a fixed random seed to ensure reproducibility of results. This fine-tuning methodology successfully adapted the multilingual foundation model to our specific low-resource language pairs while maintaining computational efficiency through quantization and parameter-efficient training techniques. We observed promising performance improvements across all language pairs, with BLEU score increases ranging from 3-19% for both moderate-resource pairs and extremely low-resource languages. The bidirectional training approach particularly benefited English-to-target translation, where several language pairs showed improvements from near-zero baselines.

5 Results and Analysis

Our analysis of the actual WMT 2025 results reveals significant insights into zero-shot and fine-tuned translation performance across language families and resource levels. The system demonstrates competitive performance, often exceeding the mean scores of other participating teams across multiple language pairs and evaluation metrics.

The results reveal (Figure 1) complex patterns of cross-lingual transfer effectiveness that partially support but also challenge our initial hypotheses about language family relationships in multilingual models. A notable asymmetry emerges where target-to-English translation significantly outperforms English-to-target translation across all language pairs, with performance ratios varying dramatically from modest differences to orders of magnitude disparities. This pattern suggests that the model's English-centric training provides substantially stronger support for translating into English than for generating text in low-resource languages.

Within the language family analysis, our findings show highly variable performance patterns that resist simple categorization. The Indo-Aryan language Assamese demonstrates strong Assamese-to-English performance with 15.26 BLEU, substantially exceeding the competition mean of 11.05, but shows weaker English-to-Assamese performance at 9.09 BLEU compared to the competition mean of 13.51. The Tibeto-Burman cluster exhibits remarkable diversity, ranging from exceptional performance in Bodo-to-English translation (21.67 BLEU) to very poor results in Nyishi-to-English (9.61 BLEU) and catastrophic failure in English-to-Bodo translation (0.21 BLEU). Khasi, representing the Austroasiatic family, shows moderate and relatively balanced performance in both translation directions compared to other languages in our study. Bodo being categorized as having very limited training data, achieves the highest Bodo-to-English BLEU score across all evaluated languages while simultaneously producing the lowest English-to-Bodo performance. This strong asym-



(a) BLEU Score

(b) METEOR Score

Figure 1: BLEU and METEOR scores for translation directions. RBG-AI (our submitted system) is compared against the mean and median of all participating systems.

metry suggests that cross-lingual transfer benefits may operate through mechanisms more complex than simple resource availability or family membership. It might potentially involve specific linguistic features or training data characteristics that favor certain translation directions.

Interms of overall performance, our quantized MADLAD-400 system demonstrates competitive performance relative to other participating teams, outperforming the competition mean in eight out of fourteen language-direction pairs. The system shows particular strength in target-to-English translation, with notable advantages in Bodo-to-English (+7.42 BLEU), Nyishi-to-English (+3.62 BLEU), and Assamese-to-English (+4.21 BLEU) directions. Additionally, several language pairs show improvements in chrF scores, indicating better character-level accuracy even when BLEU scores are comparable. However, the system faces significant challenges in English-to-target translation for several language pairs. Performance gaps appear most visible in English-to-Assamese (-4.42 BLEU), English-to-Khasi (-5.21 BLEU), and English-to-Mizo (-3.68 BLEU) directions. The most severe limitation appears in English-to-Bodo translation. Here, our system achieves only 0.21 BLEU compared to the competition mean of 11.64, representing a systematic failure requiring further investigation with inputs from linguistic experts.

6 Conclusion and Future Work

This study provides empirical evidence about how multilingual language models benefit low-resource languages through cross-lingual transfer, based on our competitive performance in the WMT 2025 shared task. The system outperformed competition averages in eight out of fourteen language-direction pairs, proving that deployment efficiency without compromise in performance.

Further our findings reveal several important patterns that advance understanding of multilingual model capabilities for low-resource languages. A consistent translation asymmetry emerges where target-to-English translation significantly outperforms English-to-target translation for most of the language pairs, with performance ratios ranging from 1.5x to 100x. This asymmetry reveals that English-centric bias is inherent in multilingual models and suggests fundamental limitations in generating low-resource languages compared to translating into English. The effects of language family relationships on translation proved more complex than initially hypothesized. While Tibeto-Burman languages showed evidence of family-based transfer across five of seven target languages, the effects varied dramatically and sometimes counterintuitively.

Performance variations within script groups, such as Bengali script languages showing BLEU scores ranging from 15.26 (Assamese) to 1.58 (Kokborok) for source-to-English translation. This indicates that script similarity provides minimal transfer benefits compared to other linguistic factors. The performance variations within language families and the consistent translation asymmetry suggest that transfer mechanisms involve more than generic linguistic similarity. Wide performance variations within the Tibeto-Burman family suggest that family membership alone is insufficient to predict transfer success. Future research should investigate why target-to-English translation consistently outperforms English-to-target translation and develop techniques to improve generation capabilities for low-resource languages.

References

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Computing Surveys (CSUR)*, 53(5):1–38.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Thanh-Le Ha, Jan Niehues, and Alex Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. In *Proceedings of the 13th International Conference on Spoken Language Translation*.

Nabam Kakum, Sahinur Rahman Laskar, Koj Sambyo, and Partha Pakray. 2023. Neural machine translation for limited resources english-nyishi pair. *Sādhanā*, 48(4):237.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: a multilingual and document-level large audited dataset. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 67284–67296.

Anoop Kunchukuttan and Pushpak Bhattacharyya. 2020. Utilizing language relatedness to improve machine translation: A case study on languages of the indian subcontinent. *arXiv preprint arXiv:2003.08925*.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, et al. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135.

Partha Pakray, Reddi Krishna, Santanu Pal, Advaitha Vetagiri, Sandeep Dash, Arnab Kumar Maji, Saralin A Lyngdoh, Lenin Laitonjam, Anupam Jamatia, Koj Sambyo, Ajit Das, and Riyanka Manna. 2025. Findings of wmt 2025 shared task on low-resource indic languages translation. In *Proceedings of the Tenth Conference on Machine Translation (WMT) under EMNLP*, Suzhou, China. Association for Computational Linguistics.

Partha Pakray, Santanu Pal, Advaitha Vetagiri, Reddi Krishna, Arnab Kumar Maji, Sandeep 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*, pages 654–668.

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.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multilingual unsupervised nmt using shared encoder and language-specific decoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3083–3089.

Xu Tan, Jiale Chen, Di He, Yingce Xia, Tao Qin, and Tie-Yan Liu. 2019. Multilingual neural machine translation with language clustering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 963–973.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

NLLB Team, Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Yining Wang, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong. 2018. Three strategies to improve one-to-many multilingual translation. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2955–2960.

Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2021. Are multilingual models effective in code-switching? In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 142–153.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.