Can Linguistically Related Languages Guide LLM Translation in Low-Resource Settings?

Aishwarya Ramasethu
Prediction Guard

Rohin Garg*
Scale AI

Niyathi Allu*
Independent

Harshwardhan Fartale*
Independent

Dun Li Chan
INTI International College Penang

Abstract

Large Language Models (LLMs) have achieved strong performance across many downstream tasks, yet their effectiveness in extremely low-resource machine translation remains limited. Standard adaptation techniques typically rely on large-scale parallel data or extensive fine-tuning, which are infeasible for the long tail of underrepresented languages. In this work, we investigate a more constrained question: in data-scarce settings, to what extent can linguistically similar pivot languages and few-shot demonstrations provide useful guidance for on-the-fly adaptation in LLMs? We study a data-efficient experimental setup that combines linguistically related pivot languages with few-shot in-context examples, without any parameter updates, and evaluate translation behavior under controlled conditions. Our analysis shows that while pivot-based prompting can yield improvements in certain configurations, particularly in settings where the target language is less well represented in the model's vocabulary, the gains are often modest and sensitive to few shot example construction. For closely related or better represented varieties, we observe diminishing or inconsistent gains. Our findings provide empirical guidance on how and when inference-time prompting and pivot-based examples can be used as a lightweight alternative to fine-tuning in low-resource translation settings.

1 Introduction

The advent of transformer-based architectures and general-purpose Large Language Models (LLMs) such as ChatGPT (OpenAI, 2024), DeepSeek-R1 (DeepSeek-AI, 2025), Mistral (Jiang et al., 2023), and Llama 3 (AI@Meta, 2024) has led to substantial advances in machine translation over the past decade. These models exhibit strong multilingual capabilities and, for many high-resource languages, approach expert-level translation quality. However, this performance remains highly uneven across languages.

Despite the existence of over 7,000 languages worldwide (Eberhard et al., 2024), NLP research and model development remain heavily skewed toward a small set of high-resource languages (Joshi et al., 2021; Pakray et al., 2025). Prior work has documented significant disparities in LLM translation performance between English and low-resource languages (Choudhury, 2023), and recent surveys show that even state-of-the-art models such as GPT-4 often fail to outperform specialized systems on languages using non-Latin scripts (Ataman et al., 2025).

To address these disparities, substantial effort has gone into expanding multilingual datasets and models. Foundational work on massively multilingual representation learning, such as mBERT and XLM-R (Arivazhagan et al., 2019; Conneau et al., 2020), enabled cross-lingual transfer across hundreds of languages. More recent initiatives, including No Language Left Behind (NLLB) (Team, 2024) and the FLORES benchmark (Goyal et al., 2022), aim to scale multilingual machine translation to previously underrepresented languages, while projects such as Aya (Üstün et al., 2024), BLOOM (Leong et al., 2022), and Masakhane (Nekoto et al., 2020) emphasize broader linguistic coverage and community-driven data creation. Despite these efforts, coverage remains uneven, and many languages with low digital presence are still only partially supported in widely deployed translation systems.

Rather than focusing on resource-intensive data collection or training new language-specific models, we investigate whether the few-shot instruction-following capabilities of existing LLMs can be leveraged for extremely low-resource machine translation. We study an inference-time approach that combines linguistically related few-shot examples with a pivot language—a higher-resource language closely related to the target—to provide additional

*Equal contribution

Proceedings for the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026), pages 168–185

contextual grounding during generation.

Our experiments focus on two linguistically distinct yet underrepresented languages: Tunisian Arabic (aeb) (Mahdi, 2025) and Konkani (gom) (Rajan et al., 2020). Both languages have substantial regional and cultural significance but receive limited coverage in multilingual benchmarks and are only partially supported in large pretrained translation systems such as NLLB (Team, 2024). This makes them representative of practical low-resource scenarios where parallel data is scarce and model support varies across dialects and scripts. We evaluate our approach using In-Context Learning (ICL) with frozen, decoder-only LLMs.

We find that incorporating linguistically and semantically related few-shot examples can improve translation behavior in certain configurations, particularly when the target language appears weakly represented in the model's pretraining distribution. For Konkani, pivot-augmented prompting yields moderate gains in chrF++ relative to direct prompting, while for Tunisian Arabic the improvements are smaller and less consistent across models. These results suggest that the effectiveness of pivot-guided prompting depends strongly on language relatedness, representational coverage, and interactions between pivot and target varieties, rather than offering a universally reliable translation strategy.

2 Related Work

2.1 In-Context Learning

Prior work shows that multilingual LLM translation performance under few-shot in-context learning (ICL) depends strongly on prompt example quality (Chowdhery et al., 2022). However, it is also highlighted that substantial gains are observed in the high-resource language pairs. In addition Agrawal et al. (2022) confirm that even a single noisy or semantically unrelated demonstration can drastically reduce translation quality, whereas a well-formed equivalent-meaning example is often sufficient to elicit better translation quality from the pretrained LLMs.

Further work by Vilar et al. (2023) demonstrate that translation quality depends on domain quality rather than lexical similarity of the in-context examples and that the quality of translation degrades with poorly selected in-context examples. However their evaluation is limited to translations between English and a small set of relatively high-resource languages (French, German, and Chinese). The work of Garcia et al. (2023) also supports that the quality of few-shot in-context examples is crucial. Puduppully et al. (2023) introduce DecoMT, a few-shot prompting approach that decomposes the translation process into a sequence of word-chunk translations.

Large-scale analyses show that ICL performance in MT is driven primarily by example quality and target-side distribution rather than prompt structure or ordering (Zhu et al., 2024b; Chitale et al., 2024). Zhu et al. (2024a) investigate robustness in ICL by introducing a dual-view demonstration selection strategy. They combine margin-based sentence-level similarity to avoid semantic noise with word-embedding-based token weighting to refine the influence of demonstrations.

Taken together, these studies show that ICL can improve machine translation under favourable conditions, but they also highlight its sensitivity to demonstration quality and distributional coverage. Importantly, most of this work evaluates languages with comparatively rich digital resources, leaving open the question of how reliably ICL-based MT behaviour transfers to truly low-resource languages with sparse data and unstable tokenization.

Recent work has explored structured linguistic scaffolding as a complement to standard few-shot prompting. Lu et al. (2024) propose Chain-of-Dictionary Prompting (COD), which augments prompts with chained multilingual dictionary hints and reports large gains on FLORES-200. While effective, COD relies on proprietary models and dictionary resources that may not exist for many low-resource languages. In contrast, our approach uses open 7B-8B models and small parallel corpora, providing pivot translations as broader contextual scaffolding rather than word-level lexical hints.

Other work addresses low-resource adaptation through training-time methods. Yong et al. (2023) show that adapter-based finetuning can outperform continued pretraining when adding new languages, with gains driven primarily by data availability. Muennighoff et al. (2023) introduce multitask prompted finetuning for multilingual models and demonstrate improved zero-shot generalization when prompt language aligns with the target. Longpre et al. (2025) analyze multilingual scaling laws and argue that at very low data scales, neither pretraining nor finetuning is computationally efficient. These approaches require supervised data and training compute, whereas our work targets inference-time prompting without parameter updates.

2.2 Pivot languages aided LLM translation

Pivot strategies introduce an intermediate language to support translation in low-resource settings. Prior work has demonstrated that the choice of a pivot language can have significant impact on the translation quality.

Work by Imamura et al. (2023) shows the poor zero-shot performance of multilingual NMT models translation can be enhanced by a using pivot language. In this work, they compare the pivot and direct translation using English as the pivot language. Their study also investigates which kind of parallel corpora is most effective to enhance multilingual pivot translation.

Jiao et al. (2023) also evaluate ChatGPT for machine translation and introduce a pivot prompting strategy, in which the model first translates a source sentence into a high-resource pivot language before translating into the target language. They find that pivot prompting noticeably improves translation quality for distant or low-resource languages, and with GPT-4, ChatGPT achieves performance comparable to commercial translation systems even on some of the challenging language pairs.

Extending these ideas, Elmadani and Buys (2024) introduces synthetic pivoting, where pivot sentences are generated from both the target and the source languages using the sequence level knowledge distillation. This approach reduces pivot translation complexity and improves BLEU scores for low-resource Southern African languages by up to 5.6 points.

Recent work by Talwar and Laasri (2025) highlight this in their study on Nepali-English translation, where Hindi is chosen as a pivot language due to its linguistic proximity to Nepali and the greater availability of Hindi parallel corpora. By employing both fully supervised transfer learning and semi-supervised back-translation, they show that using Hindi as a pivot language improves the Nepali-English translation baselines, emphasizing how a chosen pivot language can compensate for limited data availability.

Lim et al. (2025) reformulated low-resource translation as a post-editing task, where a teacher model generates auxiliary translations and a student model is finetuned to correct them, achieving strong gains on FLORES-200/NTREX. Their results suggest that even imperfect auxiliary translations can provide useful scaffolding. While Mufu relies on supervised finetuning, our work adapts this post-editing insight to pure ICL by using a single pivot translation combined with retrieved few-shot examples, without parameter updates or multi-model pipelines.

Collectively, these findings motivate our investigation into pivot language strategies for LLM translation. Our work builds on these insights by examining whether integrating pivot language examples and leveraging few-shot ICL can further enhance translation performance for languages like Konkani and Tunisian Arabic. In doing so, we aim to clarify the mechanisms by which pivot languages facilitate knowledge transfer in LLMs, while also extending the adaptation capabilities of models to new languages. This helps clarify when such approaches may, or may not, be effective for low-resource languages.

3 Methodology

We explore an inference-time technique of translation in settings where data, compute, and model scale are limited. Our goal is to examine what kinds of evidence (such as linguistically related pivot languages and semantically retrieved few-shot examples) can be leveraged to support translation in an ICL setting using small (≈8B) decoder-only models, without fine-tuning or large parallel corpora. In particular, we investigate whether these signals provide useful guidance when translating into previously unseen or under-represented languages, and under what conditions they help, fail, or produce inconsistent behavior.

To support semantic retrieval, we construct a datastore of parallel translations organized as triplets consisting of an English source sentence, its pivot-language translation, and the corresponding target-language translation. These triplets are derived exclusively from the training split. We index the datastore using the English source sentences, as English is the input language at inference time. Sentence embeddings are computed using the **all-MiniLM-L12-v2** sentence transformer, which maps text into a dense vector space suitable for semantic similarity search. This representation allows semantically related translation examples to be clustered and retrieved efficiently.

At inference time, we generate an embedding for each input source sentence and query the vector datastore using cosine similarity. The top-k most semantically similar triplets are retrieved and used as in-context demonstrations. These demonstrations

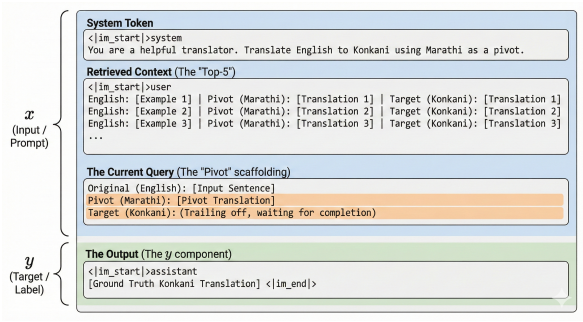

Figure 1: The Training Data Structure. The input x contains the system instruction, retrieved semantic context (top-5), and the pivot scaffolding. The target y contains only the model's generated translation.

are formatted as English-pivot-target examples and combined with the pivot translation corresponding to the same input sentence from the parallel corpus; no pivot translations are generated by the model or obtained from external sources. The resulting prompt is structured in ChatML format (see Appendix A.11). To mitigate contamination and retrieval leakage, all retrieved examples are drawn strictly from the training datastore, while evaluation is performed on a held-out test set that is never indexed or queried during retrieval. No test sentences or paraphrases are included in the datastore. We treat the number of in-context examples k as a controllable parameter and evaluate its effect through a systematic ablation study (see Appendix A.8). This retrieval-augmented prompting workflow is illustrated in Figure **??**.

To isolate the contribution of the pivot language itself, we conduct controlled comparisons between pivot-augmented prompts and prompts constructed using identical retrieval procedures but excluding the pivot translation. This allows us to disentangle gains arising from semantic retrieval alone from those attributable to the linguistic bridge provided by the pivot language. The ablation study results without the pivot language are shown in Table 10 and Table 11.

4 Experiments

4.1 Languages and data

For this experiment, we focused on two low-resource languages. The first is Konkani, an Indian language spoken in the western part of India, with approximately 2.35 million speakers as of 2011. The second is Tunisian Arabic, spoken in Tunisia, with around 12 million speakers as of 2021. A motivation for selecting these languages was their use of non-Latin scripts. In addition, Tunisian Arabic is unique in the sense that, unlike regular Latin script which reads from left to right, this is from right to left.

To provide a quantitative sanity check on pivot selection, we also compute word-level Jaccard similarity between the pivot and target languages in our datasets. This allows us to verify that the chosen pivots are lexically closer to the target languages than English. A detailed description and full results are provided in Appendix A.6.

4.2 Models

In this experiment, we evaluate the performance of Unbabel's TowerInstruct-7B-v0.1 (Alves et al., 2024) and NousResearch's Hermes-2-Pro-Llama-3-8B (Teknium et al., 2024). The Hermes-2-Pro-Llama-3 model is a instruction-tuned version of Llama 3 (AI@Meta, 2024) known for its multilingual capabilities. Llama 3 supports 8 languages: English, German (deu), French (fra), Italian (ita), Portuguese (por), Hindi (hin), Spanish (spa), and Thai (tha), although the underlying foundation model has been trained on over 176 languages. Our aim is to see whether we can make use of the latent knowledge alignment of the model while translating to low resource languages.

The TowerInstruct-7B-v0.1 model is finetuned from TowerBase-7B-v0.1 model. The TowerBase-7B-v0.1 model is continuously pretrained from from the Llama 2 model with a mixture of monolingual and parallel data with 20B tokens. The TowerBase-7B-v0.1 model is the finetuned with instruction dataset that is relevant to translation process. Some of these instruction tasks include Automatic Post Edition, Context-aware Translation, Named-entity Recognition etc. The languages supported by the TowerBase-7B-v0.1 model are English (eng), German (deu), French (fra), Dutch (nld), Italian (ita), Spanish (spa), Portuguese (por), Korean (kor), Russian (rus), and Chinese (zho). Although TowerInstruct-7B-v0.1 performs well on translation tasks, it is not expected to excel in languages it was not exposed to during training.

Our strategic selection of these models is designed to assess their effectiveness in translating languages outside their initial training sets. Despite TowerInstruct-7B-v0.1's specialized training in translation, it has not been directly exposed to the specific low-resource languages focused on in this experiment, offering a unique test of its adaptability to unseen languages.

4.3 Datasets

We utilized two distinct multiparallel datasets, effectively organizing the data into aligned triplets (Source-Pivot-Target) to support our retrieval-augmented pipeline.

Konkani: We constructed a dataset of English-Marathi-Konkani triplets using the open-source corpus from AI4Bharat (Gala et al., 2023). Marathi was selected as the pivot language due to its linguistic similarity to Konkani and wider prevalence in western India. We created a distinct split of [800] examples for the training (retrieval) datastore and 200 examples for the held-out test set.

Tunisian Arabic: We derived a similar corpus of English-MSA-Tunisian triplets from the work described by Bouamor et al. (2014), with Modern Standard Arabic (MSA) chosen as the pivot language. This dataset consists of 900 examples for the training datastore and 100 examples for the held-out test set.

In total, our study operates on small training sets of approximately 1,000 records per language. This constraint was chosen specifically to simulate realistic low-resource scenarios where large-scale parallel data is unavailable.

5 Results

5.1 Does Pivot-Based Prompting Improve Translation?

To establish a reference point, we first evaluate a direct prompting baseline, where the model is given only the English source sentence and instructed to translate directly into the target language, without access to a pivot language. In this setting, chrF++ scores are often extremely low (in some cases close to 1) because the models do not reliably generate text in the intended target language or script. Instead, the output frequently drifts toward better-represented neighboring languages (e.g., producing Marathi- or Hindi-like text when the target is Konkani, or MSA-like text for Tunisian Arabic). This behavior is observed across both Hermes and Tower, indicating that, without grounding signals, the model does not consistently infer the correct output language from the instruction alone.

We then compare this to our pivot-augmented prompting condition, in which the same input is supplied along with a translation into a linguistically related pivot language. In this setting, the few-shot demonstrations and pivot translation act as grounding signals that stabilize generation toward the intended script and language family. Tables 1 and 2 report BLEU and chrF++ scores across three conditions (zero-shot (k=0), direct few-shot prompting without a pivot, and pivot-augmented prompting). For each configuration, we report the best-performing number of in-context examples (k), as determined in ablations in Appendix 12 and 13.

For Konkani, introducing few-shot demonstrations, even without a pivot, leads to a substantial improvement in both chrF++ and BLEU, indicating that the examples themselves provide a strong anchoring effect for this previously unseen language. Adding the pivot language on top of these examples results in only small or mixed additional gains: for Hermes, the pivot condition yields a modest improvement over direct few-shot prompting (29.62→30.34 chrF++, 7.35→7.77 BLEU), whereas for Tower the pivot improves BLEU from 3.67 to 5.68, but does not improve chrF++. This suggests that, in this setting, most of the benefit arises from example-driven stabilization rather than from the pivot language itself.

For Tunisian Arabic, zero-shot scores are already relatively high, and both chrF++ and BLEU change marginally across the direct and pivot conditions, with no consistent advantage for either model. Here, few-shot prompting provides limited additional benefit, and the pivot language does not substantially alter model behavior, consistent with the interpretation that Tunisian Arabic which is already better represented in the underlying pretrained models.

We additionally evaluate whether using a pivot language that is explicitly supported by the model leads to improved translation quality. Given the constraints of our setup, the only configuration that satisfies this condition is Hindi as a pivot for Konkani using the Hermes-2-Pro-Llama-3-8B model. We analyze this setting in detail in Appendix A.9, including token-to-word ratios and Jaccard similarity between Hindi and Konkani.

Across these experiments, we find that using a model-supported pivot language does not yield systematic improvements over linguistically motivated pivots such as Marathi. In several cases, performance degrades as the number of in-context examples increases, suggesting that native model support alone is insufficient to improve or stabilize low-resource translation.

To ensure that pivot-augmented prompting does not simply cause the model to reproduce pivot-language translations, we measure chrF overlap between the pivot outputs and the final gener-

ated translations. This analysis, reported in Appendix A.5, shows consistently low chrF scores for both Konkani and Tunisian Arabic, indicating limited surface-level overlap between pivot and generated outputs. These results suggest that the model does not merely copy or lightly edit the pivot translation, but instead produces outputs that are substantially distinct from the pivot language.

One possible explanation, which we treat as hypothesis-generating rather than conclusive, comes from the token-to-word analysis in Table 6. For Tunisian Arabic, both models exhibit substantially lower token-to-word ratios (e.g., 4.96 vs. 7.65 for Tower; 2.16 vs. 4.09 for Hermes, comparing Aeb vs. Gom), indicating that the models segment Tunisian Arabic into fewer subword units than Konkani. Because Modern Standard Arabic (MSA) is well represented in most pretrained corpora, Tunisian Arabic, which shares script and lexical characteristics with MSA, may benefit indirectly from this representation. This would help explain why few-shot prompting and pivot augmentation yield smaller or inconsistent gains in this setting.

In contrast, the much higher token-to-word ratios for Konkani suggest a weaker lexical footprint in the pretrained vocabulary. Here, the few-shot examples and the pivot appear to act less as a source of additional translation competence and more as basic scaffolding for language identification, script adherence, and output stability.

However, we emphasize that this relationship is correlational rather than causal: tokenization efficiency alone does not fully explain performance differences, and other factors may contribute to the observed behavior.

5.2 Comparison to NLLB Reference Baselines

As a point of external reference, we compare the best-performing few-shot and pivot-augmented scores in Tables 1 and 2 with the NLLB-200 distilled baselines in Table 5. For Konkani, NLLB does not provide native support, and the baseline remains relatively low (26.82 chrF++, 7.51 BLEU). Our best Hermes pivot-augmented configuration attains 30.34 chrF++ and 7.77 BLEU, while the Tower pivot setting reaches 17.66 chrF++ and 5.68 BLEU. Thus, Hermes slightly exceeds the NLLB baseline on both metrics, whereas Tower remains below it.

For Tunisian Arabic, NLLB does include explicit support, but the baseline scores remain modest (10.42 chrF++, 4.20 BLEU). In contrast, both decoder-only LLMs achieve substantially higher performance even without fine-tuning: Hermes reaches 24.32 chrF++ and 6.27 BLEU (direct few-shot), and Tower reaches 20.63 chrF++ and 4.99 BLEU (pivot). This indicates that, even relative to a supervised MT system trained with explicit support for the language, few-shot prompting can yield stronger performance in this setting.

This contrast highlights a practical trade-off: improving NLLB performance for unsupported or weakly supported languages would typically require collecting supervised training data and fine-tuning the model, whereas our approach obtains measurable gains using only few-shot prompting with no parameter updates.

5.3 Effect of Increasing the Number of In-Context Examples

We next examine whether translation quality improves simply by increasing the number of retrieved demonstrations (k), independent of the pivot signal. For each model-language pair, we evaluate chrF++ and BLEU across multiple values of k in both the direct and pivot-based translation settings (full results in Appendix 12-13 and 10-11).

For Konkani, increasing k does not produce monotonic gains in either metric. In the direct translation setting, Hermes reaches its strongest chrF++ at k=3 (29.62) with relatively low BLEU (2.33), while performance drops at both smaller and larger k values. Tower shows a similar pattern: chrF++ peaks at k=2 (21.25), while BLEU remains in the 3-4 range and collapses to 0 beyond k=3. In the pivot-based setting, Hermes attains its best chrF++ at k=1 (30.34) and best BLEU at k=4 (7.77), whereas Tower peaks at k=2 in BLEU (5.68) but achieves higher chrF++ at k=3 (17.66). Thus, for Konkani, both metrics improve relative to k=0, but additional demonstrations beyond the best-performing k generally degrade performance.

A similar trend appears in Tunisian Arabic, although the baseline (k=0) performance is much stronger. In the direct setting, Hermes achieves its best BLEU at k=1 (6.27), while chrF++ remains highest at k=0 (24.32) and declines as k increases. Tower exhibits modest variation across k, with chrF++ peaking at k=4 (20.74) despite little corresponding change in BLEU. In the pivot condition, Hermes again shows small fluctuations around its k=0 baseline (24.31 chrF++), while Tower reaches its highest BLEU at k=2 (4.99) and highest chrF++ at k=5 (20.63), before declining at larger k.

Model	Setting	BLEU	chrF++
Baseline	NLLB-200	7.51	26.82
Hermes-2-Pro	Zero-shot (k=0)	1.49	1.30
	Direct (Best k)	7.35	29.62
	With Pivot (Best k)	**7.77**	**30.34**
TowerInstruct	Zero-shot (k=0)	1.28	0.69
	Direct (Best k)	3.67	**21.25**
	With Pivot (Best k)	**5.68**	17.66

Table 1: Select Konkani translation results (Eng→Gom)

Model	Setting	BLEU	chrF++
Baseline	NLLB-200	4.20	10.42
Hermes-2-Pro	Zero-shot (k=0)	4.62	24.32
	Direct (Best k)	**6.27**	**24.32**
	With Pivot (Best k)	5.06	24.31
TowerInstruct	Zero-shot (k=0)	4.19	17.62
	Direct (Best k)	4.46	**20.74**
	With Pivot (Best k)	**4.99**	20.63

Table 2: Select Tunisian Arabic results (Eng→Aeb)

Across both languages and models, these results indicate that: Performance improves substantially when moving from k=0 to small k, but gains do not scale with additional demonstrations. ChrF++ and BLEU often peak at different values of k.

We hypothesize that one contributing factor is the interaction between k and model context capacity. TowerInstruct operates effectively within a 4K-token window, and performance often declines once prompts approach this length, suggesting truncation or overwriting effects. Hermes supports a larger context window, yet its performance likewise plateaus or degrades beyond moderate k, implying that the limitation is not purely architectural but also behavioral: models may underutilize long-range prompt structure or overweight spurious correlations from loosely related examples.

Taken together, these findings suggest that the gains observed in our experiments are not simply an artifact of "more examples." Instead, a small number of semantically aligned demonstrations appears to provide most of the benefit, while additional examples can introduce noise that reduces both BLEU and chrF++. In settings where pivot-based prompting yields improvements, these effects should therefore be interpreted as complementary to, rather than interchangeable with, the contribution of few-shot demonstrations themselves.

6 Limitations

Many machine learning breakthroughs are enabled by an abundance of computational resources. However, access to large-scale compute is not uniformly available, including to most authors of this work. This disparity becomes even more apparent when working with communities that speak low-resource languages. Within these constraints, we aimed to rigorously test our hypotheses about pivot-based translation using the resources available to us. Importantly, these constraints also reflect realistic deployment conditions for many low-resource language communities, where access to large-scale compute, extensive annotation, and proprietary models is limited.

The primary limitation of this work is that, while we build on prior research on pivot languages to investigate whether linguistically related languages provide any useful signal for inference-time translation under resource constraints, the performance gains we observe are modest and often inconsistent. Working within our computational budget, we evaluated open-weight models in the 7B parameter range. While larger models may yield stronger performance, our results indicate that pivot-augmented prompting can sometimes improve performance, but its effects are highly sensitive to language characteristics and example selection, suggesting that further study is needed before drawing strong conclusions.

Additionally, much of the existing research on multilinguality and machine translation relies on human evaluation, which was not feasible in our setting. Under these constraints, and with respect for the communities that speak these languages, we evaluate how language models adapt to previously unseen languages in low-resource conditions using automatic metrics. We report BLEU and chrF++ scores computed with SacreBLEU (Post, 2018) for reproducibility (see Appendix A.2 for scoring signatures).

However, these metrics have known limitations in low-resource and morphologically rich settings. As illustrated in Appendix A.7.1, we observe cases where the generated Konkani translation is linguistically plausible and semantically related to the reference, yet differs substantially in surface form, resulting in very low BLEU and chrF++ scores. This highlights the brittleness of n-gram–based metrics for evaluating low-resource translation quality and motivates the need for human evaluation by native speakers to better capture semantic adequacy, pragmatic meaning, and dialectal correctness.

Another limitation is that our methodology depends on the availability of a high-resource pivot language that is linguistically similar to the target language, which restricts its applicability to lan-

guages without closely related pivots. While the approach is data-efficient, it also assumes access to high-quality parallel corpora; translation quality may degrade when there is a domain mismatch between the retrieved examples and the input text.

Given the promising results observed under these constrained settings, natural extensions of this work include scaling experiments to larger open-source models, conducting human-in-the-loop evaluations with native speakers, and exploring additional language pairs to better characterize the conditions under which pivot-augmented prompting helps, fails, or produces negligible effects.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context examples selection for machine translation. *Preprint*, arXiv:2212.02437.

AI@Meta. 2024. Llama 3 model card.

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. In *Proceedings of the Conference on Language Modeling (COLM) 2024*.

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

Duygu Ataman, Alexandra Birch, Nizar Habash, Marcello Federico, Philipp Koehn, and Kyunghyun Cho. 2025. Machine translation in the era of large language models:a survey of historical and emerging problems. *Information*, 16(9).

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A multidialectal parallel corpus of arabic. In *LREC*, pages 1240–1245.

Pranjal Chitale, Jay Gala, and Raj Dabre. 2024. An empirical study of in-context learning in LLMs for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7384–7406, Bangkok, Thailand. Association for Computational Linguistics.

Monojit Choudhury. 2023. Generative AI has a language problem. *Nature Human Behaviour*, 7(11):1802–1803.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, and 48 others. 2022. Palm: Scaling language modeling with pathways. *Preprint*, arXiv:2204.02311.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Preprint*, arXiv:1911.02116.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2024. *Ethnologue: Languages of the World*, twenty-seventh edition. SIL International, Dallas, Texas.

Khalid N. Elmadani and Jan Buys. 2024. Neural machine translation between low-resource languages with synthetic pivoting. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12144–12158, Torino, Italia. ELRA and ICCL.

Jay Gala, Pranjal A. Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*, 2023.

Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10867–10878. PMLR.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Kenji Imamura, Masao Utiyama, and Eiichiro Sumita. 2023. Pivot translation for zero-resource language pairs based on a multilingual pretrained model. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 348–359, Macau SAR,

China. Asia-Pacific Association for Machine Translation.

Albert Q Jiang, Alexandre Sablayrolles, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *Preprint*, arXiv:2301.08745.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2021. The state and fate of linguistic diversity and inclusion in the nlp world. *Preprint*, arXiv:2004.09095.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395. Association for Computational Linguistics.

Colin Leong, Joshua Nemecek, Jacob Mansdorfer, Anna Filighera, Abraham Owodunni, and Daniel Whitenack. 2022. Bloom library: Multimodal datasets in 300+ languages for a variety of downstream tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8608–8621, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zheng Wei Lim, Nitish Gupta, Honglin Yu, and Trevor Cohn. 2025. Multilingual fused learning for low-resource translation with llms. In *International Conference on Learning Representations*.

Shayne Longpre, Sneha Kudugunta, Niklas Muennighoff, I-Hung Hsu, Isaac Caswell, Alex Pentland, Sercan Ö. Arik, Chen-Yu Lee, and Sayna Ebrahimi. 2025. Atlas: Adaptive transfer scaling laws for multilingual pretraining and finetuning. *Preprint*, arXiv:2510.22037.

Hongyuan Lu, Haoran Yang, Haoyang Huang, Dongdong Zhang, Wai Lam, and Furu Wei. 2024. Chain-of-dictionary prompting elicits translation in large language models. *Preprint*, arXiv:2305.06575.

Mohamed Mahdi. 2025. How well do llms understand tunisian arabic? *Preprint*, arXiv:2511.16683.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. *Preprint*, arXiv:2211.01786.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Hassan Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, and 28 others. 2020. Participatory research for low-resourced machine translation: A case study in african languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4.1 technical report.

Partha Pakray, Alexander Gelbukh, and Sivaji Bandyopadhyay. 2025. Natural language processing applications for low-resource languages. *Natural Language Processing*, 31(2):183–197.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ratish Puduppully, Anoop Kunchukuttan, Raj Dabre, Ai Ti Aw, and Nancy F. Chen. 2023. Decomposed prompting for machine translation between related languages using large language models. *Preprint*, arXiv:2305.13085.

Annie Rajan, Ambuja Salgaonkar, and Ramprasad Joshi. 2020. A survey of konkani nlp resources. *Computer Science Review*, 38:100299.

Abhimanyu Talwar and Julien Laasri. 2025. Pivot language for low-resource machine translation. *Preprint*, arXiv:2505.14553.

NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630:841–846.

Ryan Teknium, Jeffrey Quesnelle, and Chen Guang. 2024. Hermes 3 technical report. *arXiv preprint arXiv:2408.11857*.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, and 1 others. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Winata, Stella

			BLEU		chrF++	
Lang.	**Mod.**	k	Δ	p	Δ	p
Gom	Her	0	+0.12	.08	+0.22	.20
	Her	1	−0.18	.88	−0.89	1.0
	Her	2	−0.12	.75	−0.33	.89
	Tow	1	−0.02	.57	−0.84	1.0
	Tow	2	+0.07	.23	−1.11	1.0
Aeb	Her	0	+0.23	.07	−0.05	.56
	Her	1	+0.38	.10	−0.37	.71
	Her	2	−0.26	.78	−1.23	.98
	Tow	0	−0.02	.55	+0.02	.48
	Tow	1	−0.20	.72	+0.34	.23
	Tow	2	+0.13	.31	+0.25	.22

Table 3: Paired bootstrap significance (pivot − direct). No comparison reaches $p < 0.05$. Gom: Konkani (n=205), Aeb: Tunisian Arabic (n=100). Her: Hermes, Tow: Tower.

Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. 2023. BLOOM+1: Adding language support to BLOOM for zero-shot prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11682–11703, Toronto, Canada. Association for Computational Linguistics.

Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. 2024. When scaling meets llm finetuning: The effect of data, model and finetuning method. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024)*.

Shaolin Zhu, Menglong Cui, and Deyi Xiong. 2024a. Towards robust in-context learning for machine translation with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16619–16629, Torino, Italia. ELRA and ICCL.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024b. Multilingual machine translation with large language models: Empirical results and analysis. *Preprint*, arXiv:2304.04675.

A Appendix

A.1 Statistical Significance

We conduct paired bootstrap resampling (Koehn, 2004) (n=10,000, $p < 0.05$) to test whether pivot prompting significantly outperforms direct translation. As shown in Table 3, no comparison reaches significance, indicating that observed trends are suggestive rather than conclusive.

A.2 SacreBLEU Signatures and Reproducibility

All BLEU and chrF++ scores were computed using SacreBLEU (Post, 2018). Following their recommendations, we report scoring signatures below for full reproducibility. Statistical significance was assessed via paired bootstrap resampling (Koehn, 2004) (n=10,000, $p < 0.05$); see Section A.1.

Metric	Signature
BLEU	`nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.5.1`
chrF++	`nrefs:1|case:mixed|eff:yes|nc:6|nw:2|space:no|version:2.5.1`

Table 4: SacreBLEU scoring signatures.

A.3 NLLB baselines

Table 5 reports reference translation scores from the NLLB-200 distilled model. NLLB is an encoder-decoder neural machine translation system trained for supervised MT, whereas the models in our study (Hermes and Tower) are decoder-only LLMs used in a few-shot, in-context prompting setting with no task-specific parameter updates. Accordingly, these numbers are provided only as contextual reference points rather than as directly comparable baselines. We also note that NLLB does not natively support Konkani; the scores reported for this variety in Table 5 reflect zero-shot transfer behavior rather than a tuned dialect model.

Language Pair	BLEU	chrF++
Eng-Gom	7.51	26.82
Eng-Aeb	4.20	10.42

Table 5: NLLB-200 distilled reference baseline results for our evaluation datasets.

A.4 Token analysis

As shown in Table 6, Hermes consistently exhibits lower token fertility than Tower across all non-English languages, particularly for low-resource and dialectal varieties, indicating more efficient subword representations.

Dataset	Language	Tower	Hermes
Gom	Eng	1.59	1.34
Gom	Mar	7.73	4.08
Gom	Gom	7.65	4.09
Aeb	Eng	1.27	1.21
Aeb	MSA	4.74	2.12
Aeb	Aeb	4.96	2.16

Table 6: Tokens per word across languages for Tower and Hermes models. Lower values indicate more efficient tokenization.

Language	Model	k=1	k=2	k=3	k=4	k=5
Arabic	Hermes	24.09	24.89	25.17	23.79	24.06
	Tower	13.08	12.70	11.85	12.06	11.68
Konkani	Hermes	28.96	27.45	27.82	25.95	26.80
	Tower	10.78	8.71	9.69	9.16	8.24

Table 7: chrF scores between pivot translations and generated translations for different values of k. Lower scores indicate greater divergence from the pivot output, suggesting that the model is not simply reproducing the pivot translations.

A.5 Deviation from Pivot Translations

To assess whether the model simply reproduces pivot-language translations or instead generates genuinely distinct target-language outputs, we compute chrF scores between pivot translations and the final generated outputs for different values of k. chrF is well suited for this analysis, as it measures character-level overlap and is sensitive to direct copying, while remaining robust to morphological variation.

Table 7 shows consistently low chrF scores across models, languages, and values of k, indicating limited surface-level overlap between pivot translations and generated output. This suggests that the models are not merely copying or lightly editing the pivot translations but are instead producing substantially different outputs.

Notably, the Tower model exhibits particularly low chrF scores compared to Hermes for both Arabic and Konkani, with scores for Konkani remaining below 11 across all values of k. This behavior indicates an even stronger departure from the pivot translations, reinforcing the conclusion that the generated outputs are not simple transcriptions or reformulations of the pivot language.

The stability of chrF scores across different values of k further suggests that this divergence is systematic rather than an artifact of sampling variability. Overall, these results provide evidence that the generation step does not collapse to reproducing pivot-language translations, but instead yields outputs that are meaningfully distinct from the pivot representations.

A.6 Jaccard similarity

Pairwise lexical similarity between the languages in our corpus is reported in Tables 8 and 9. Marathi and Konkani exhibit substantially higher lexical overlap than English with either language, while Tunisian Arabic shows moderate overlap with Modern Standard Arabic (MSA). These values are not used as a selection criterion, but rather serve as supporting evidence that our chosen pivots are linguistically closer to the target languages than English.

Language Pair	Jaccard Similarity
Eng-Mar	0.0002
Eng-Gom	0.0121
Mar-Gom	0.1054

Table 8: Word-level Jaccard similarity scores for the Konkani corpus. Marathi and Konkani show substantially higher lexical overlap than English with either language.

Language Pair	Jaccard Similarity
Eng-MSA	0.0010
Eng-Aeb	0.0010
MSA-Aeb	0.1646

Table 9: Word-level Jaccard similarity scores for the Arabic corpus. MSA and Tunisian Arabic show moderate lexical overlap, while both exhibit minimal overlap with English.

For each pair of languages, we compute word-level Jaccard similarity by treating the vocabulary extracted from each corpus as a set. The similarity is defined as the size of the intersection divided by the size of the union, yielding a value between 0 and 1, where higher scores indicate greater lexical overlap.

In our datasets, Marathi (mar) and Konkani (Gom) show moderate lexical similarity (10.6%), while Tunisian Arabic (Aeb) and Modern Standard Arabic (Msa) exhibit higher overlap (16.5%). Notably, the Arabic variants show greater lexical closeness than the Indo-Aryan language pair, reflecting the stronger typological affinity among Arabic dialects.

A detailed breakdown of vocabulary sizes and pairwise similarity scores is presented in Tables 8 and 9. In future work, we plan to explore whether lexical similarity correlates with translation performance.

A.7 Ablation on the Number of In-Context Examples (k) in the Direct Translation Setting

For completeness, we report ablations on the number of in-context examples (k) in the **direct English→Target** setting, i.e., without the use of a pivot language. These results complement the pivot-based experiments and allow us to isolate the

marginal effect of the pivot signal from the effect of in-context demonstrations alone.

Tables 10 and 11 show results for Konkani and Tunisian Arabic respectively, using the same retrieval and prompting setup as in the pivot configuration, but omitting the pivot translation from the prompt.

A.7.1 Zero-BLEU but Non-Zero chrF++ Cases

During our direct translation experiments, we observed instances where BLEU = 0 despite non-zero chrF++, particularly for Konkani. This occurs when the model output diverges lexically from the reference despite partial topical or semantic alignment. A representative example is shown below.

Ground Truth (Konkani):

बटाटां भितर मसालो भरसून ते बेसनाच्या पिठयेंत बुडोवन तेलांत बरे तळ्ळे कांय महाराष्ट्रांतलो हो सुवादीक आनी फामाद पदार्थ तयार जाता.

Model Translation:

आलूच्या मसालेत संकरात बेसन बत्तर साठी अच्छा उत्सा–हसाठी महाराष्ट्रातील इतर प्रचारीक औषधे

A.8 Ablation on the Number of In-Context Examples (k) in the Pivot-Based Setting

For completeness, we report ablations on the number of in-context examples (k) in the **pivot-based** setting, i.e., where the model is provided with a linguistically related pivot translation alongside the retrieved few-shot examples allow us to examine the marginal contribution of the pivot signal. The performance of the models with pivot for konkani is shown in Table 12 and Tunisian arabic Table 13

A.9 Ablations with LLM-Supported Pivot Languages

Our experimental design selects pivot languages based on two primary criteria: (1) linguistic similarity to the target low-resource language, and (2) higher expected digital presence relative to the target. While we attempt to quantify pivot relevance using Jaccard similarity, this metric only imperfectly captures linguistic suitability, leaving gaps in systematic pivot selection. As an additional analysis, we consider pivot languages that are explicitly supported by the model, in order to examine whether native model support leads to improved translation performance.

A key limitation of this approach is that most general-purpose LLMs support only a narrow subset of languages, which substantially restricts coverage for low-resource targets. This constraint is evident even in our experimental setup: neither model explicitly supports Tunisian Arabic or closely related Arabic varieties, and for Konkani, only Hindi is supported, and only by the Hermes-2-Pro-Llama-3-8B model. Consequently, we evaluate this supported-pivot configuration only for Konkani and only under the Hermes-2-Pro-Llama-3-8B setting.

The Jaccard similarity between Hindi and Konkani (0.090) is slightly lower than that between Marathi and Konkani (0.105). However, because Hindi is explicitly supported by the model, this difference is reflected in tokenization behavior: the token-to-word ratio for Hindi under Hermes is substantially lower (2.85) than for Marathi and Konkani (both approximately 7, refer to Table 6), consistent with stronger lexical coverage in the pretrained vocabulary.

We report BLEU and chrF++ scores for this configuration in Table 14. When comparing chrF++ scores against the corresponding Marathi-pivot setting, we do not observe systematic improvements from using a model-supported pivot language. In several cases, performance degrades substantially as the number of in-context examples increases, suggesting that native model support alone is insufficient to guarantee stable or improved pivot-based translation in this low-resource setting.

A.10 Fine-Tuning Impact

Our fine-tuning experiments were limited in scope and not comprehensive. Fine-tuning was performed on the same small training sets (900 samples) used for few-shot example retrieval, without extensive hyperparameter tuning or architectural variations. While results show promise for Konkani, comprehensive fine-tuning ablations including varied training set sizes, learning rates, and LoRA configurations remain as future work.

Konkani: In the finetuning experiments, we treat the zero-shot finetuned model (English→Konkani without pivot and without in-context demonstrations) as the reference baseline. For Hermes, the zero-shot finetuned condition achieves a chrF++ of 36.61, which increases to 40.17 when a Marathi pivot is introduced. For TowerInstruct, chrF++ increases from 17.39 (zero-shot finetuned without pivot) to 31.91 with pivot. For completeness, we

Model	Source	Target	k	BLEU	chrF++
Ablation: Number of In-Context Examples (k)					
Unbabel/TowerInstruct-v0.1					
Unbabel/TowerInstruct-v0.1	Eng	Gom	0	1.28	0.69
Unbabel/TowerInstruct-v0.1	Eng	Gom	1	3.67	21.01
Unbabel/TowerInstruct-v0.1	Eng	Gom	2	3.38	21.25
Unbabel/TowerInstruct-v0.1	Eng	Gom	3	3.39	19.30
Unbabel/TowerInstruct-v0.1	Eng	Gom	4	0.0	19.78
Unbabel/TowerInstruct-v0.1	Eng	Gom	5	0.0	19.38
NousResearch/Hermes-2-Pro-Llama-3-8B					
NousResearch/Hermes-2-Pro-Llama-3-8B	Eng	Gom	0	1.49	1.30
NousResearch/Hermes-2-Pro-Llama-3-8B	Eng	Gom	1	2.70	23.68
NousResearch/Hermes-2-Pro-Llama-3-8B	Eng	Gom	2	2.72	23.87
NousResearch/Hermes-2-Pro-Llama-3-8B	Eng	Gom	3	2.33	29.62
NousResearch/Hermes-2-Pro-Llama-3-8B	Eng	Gom	4	1.90	28.75
NousResearch/Hermes-2-Pro-Llama-3-8B	Eng	Gom	5	7.35	25.78

Table 10: Ablation on the number of in-context examples (k) for English→Konkani direct translation.

also report few-shot finetuned results in Table 18; across both settings, we observe consistent gains when the pivot language is incorporated during prompting.

Tunisian Arabic: As in the Konkani setting, we interpret the zero-shot finetuned model without a pivot as the reference baseline. For Hermes, the zero-shot finetuned condition achieves a chrF++ of 18.07, which increases to 21.87 when an MSA pivot is included during prompting. For TowerInstruct, chrF++ improves from 14.83 (zero-shot finetuned without pivot) to 19.16 with pivot. Few-shot finetuned results are also reported in Table 19; We again observe gains when the pivot language is incorporated.

Below we describe the experiment setting in detail:

Hyperparameters: With limited data, finetuning methods like prompt tuning (where embeddings are adjusted) or LoRA (Low-Rank Adaptation) prove particularly effective (Zhang et al., 2024). With Parameter-Efficient finetuning (PEFT), even increasing the data yielded modest performance improvements. For instance, using LoRA on the Hermes-2-Pro-Llama-3-8B model brought the trainable parameters down to 176,242,688, or just 2% of the model's total parameters.

PEFT is computationally more efficient than pure ICL, which led us to adopt PEFT for our model finetuning process. We used the Huggingface Transformers library.

In addition, the model was loaded in 4-bit precision using the BitsAndBytes library with the nf4 quantization type. For fine-tuning, we employed the LoRA configuration, as detailed in the Table 16.

Parameters in Table 17 were used to generate the output from the finetuned model during the evaluation.

A.11 Prompt Template

Both the TowerInstruct-7B-v0.1 model and Hermes-2-Pro-Llama-3-8B model utilize a similar prompt format. The full prompt format is below.

```
<|im_start|>user
APE is a task designed to enhance
the quality of the translation
by performing minor adjustments
Original (English): [Original text]
Translation: [Pivot language]
Post-edited:
<|im_end|>
<|im_start|>assistant
[LLM translation]
<|im_end|>
```

The prompt includes the source sentence in English and its translation in a pivot language. For in-context learning, the prompt contains five demonstrations. In each demonstration, the assistant field is pre-filled with the target language translation. These demonstrations are carefully selected sen-

Model	Source	Target	k	BLEU	chrF++
Ablation: Number of In-Context Examples (k)					
Unbabel/TowerInstruct-v0.1					
Unbabel/TowerInstruct-v0.1	Eng	Aeb	0	4.19	17.62
Unbabel/TowerInstruct-v0.1	Eng	Aeb	1	4.46	19.49
Unbabel/TowerInstruct-v0.1	Eng	Aeb	2	4.46	16.23
Unbabel/TowerInstruct-v0.1	Eng	Aeb	3	4.07	15.59
Unbabel/TowerInstruct-v0.1	Eng	Aeb	4	4.37	20.74
Unbabel/TowerInstruct-v0.1	Eng	Aeb	5	4.37	18.61
NousResearch/Hermes-2-Pro-Llama-3-8B					
NousResearch/Hermes-2-Pro-Llama-3-8B	Eng	Aeb	0	4.62	24.32
NousResearch/Hermes-2-Pro-Llama-3-8B	Eng	Aeb	1	6.27	23.96
NousResearch/Hermes-2-Pro-Llama-3-8B	Eng	Aeb	2	5.06	20.35
NousResearch/Hermes-2-Pro-Llama-3-8B	Eng	Aeb	3	5.93	20.84
NousResearch/Hermes-2-Pro-Llama-3-8B	Eng	Aeb	4	6.27	20.99
NousResearch/Hermes-2-Pro-Llama-3-8B	Eng	Aeb	5	5.52	20.60

Table 11: Ablation on the number of in-context examples (k) for English→Tn direct translation.

tences that closely resemble the sentence to be translated. In the final instance, the assistant field is left blank. This prompt structure proved to be highly effective for translation tasks of this nature. However, when using this format with the base model, the outputs often included elements like "Note," gibberish, and repetitions. After fine-tuning the model with this format, the generated translations adhered to the expected structure and consistently produced Konkani sentences.

A.12 Translation APE Examples

- **Tunisian Example Prompt:** <|begin_of_text|><|im_start|>user: APE is a task designed to enhance the quality of the translation by performing only minor adjustments to fix any existing translation mistakes. If the translation is already correct, you should retain it as is.

  Original (English): always and always

  Translation (Modern Standard Arabic): دائماً

  Post-edited (Tunisian): <|im_end|>

  <|im_start|>assistant: ابدا ابدا <|im_end|>

  <|im_start|>user: APE is a task designed to enhance the quality of the translation by performing only minor adjustments to fix any existing translation mistakes. If the translation is already correct, you should retain it as is.

  Original (English): there a lot of things that tell us shut up and brake us...

  Translation (Modern Standard Arabic): هناك الكثير من الاشياء التى تقول لنا ان نصمت و تعترض طريقنا

  Post-edited (Tunisian): <|im_end|>

  <|im_start|>assistant: فما برشا حاجات تسكتنا و توقفنا <|im_end|>

  <|im_start|>user: APE is a task designed to enhance the quality of the translation by performing only minor adjustments to fix any existing translation mistakes. If the translation is already correct, you should retain it as is.

  Original (English): And sometime no

  Translation (Modern Standard Arabic): وقليلاً لا

  Post-edited (Tunisian): <|im_end|>

  <|im_start|>assistant: لا و ساعات <|im_end|>

  <|im_start|>user: APE is a task designed to enhance the quality of the translation by performing only minor adjustments to fix any existing translation mistakes. If the translation is already correct, you should retain it as is.

  Original (English): like I said before, in good and in bad

  Translation (Modern Standard Arabic): كما قلت قبل ذلك هناك الجيد وهناك الشرير

Model	Source	Pivot	Target	k	BLEU	chrF++
Ablation: Number of In-Context Examples (k) using Marathi Pivot (No Fine-Tuning)						
Unbabel/TowerInstruct-v0.1						
Unbabel/TowerInstruct-v0.1	English	Mar	Gom	0	2.07	1.30
Unbabel/TowerInstruct-v0.1	English	Mar	Gom	1	2.58	16.03
Unbabel/TowerInstruct-v0.1	English	Mar	Gom	2	5.68	8.94
Unbabel/TowerInstruct-v0.1	English	Mar	Gom	3	4.11	17.66
Unbabel/TowerInstruct-v0.1	English	Mar	Gom	4	2.84	4.90
Unbabel/TowerInstruct-v0.1	English	Mar	Gom	5	2.84	6.11
NousResearch/Hermes-2-Pro-Llama-3-8B						
NousResearch/Hermes-2-Pro-Llama-3-8B	English	Mar	Gom	0	2.35	24.9
NousResearch/Hermes-2-Pro-Llama-3-8B	English	Mar	Gom	1	3.49	30.34
NousResearch/Hermes-2-Pro-Llama-3-8B	English	Mar	Gom	2	2.36	27.59
NousResearch/Hermes-2-Pro-Llama-3-8B	English	Mar	Gom	3	2.72	25.89
NousResearch/Hermes-2-Pro-Llama-3-8B	English	Mar	Gom	4	7.77	27.53
NousResearch/Hermes-2-Pro-Llama-3-8B	English	Mar	Gom	5	5.73	28.65

Table 12: Ablation on the number of in-context examples (k) for English→Marathi→Konkani translation.

Post-edited (Tunisian): <|im_end|>

<|im_start|>assistant: كيما قلت قبل في الحلو و الخايب <|im_end|>

<|im_start|>user: APE is a task designed to enhance the quality of the translation by performing only minor adjustments to fix any existing translation mistakes. If the translation is already correct, you should retain it as is.

Original (English): let us be really happy away from standard stuffs

Translation (Modern Standard Arabic): اتركنا نسعد حقاً بعيد عن التابوهات

Post-edited (Tunisian): <|im_end|>

<|im_start|>assistant: خلينا نفرح برسمي بعيد على كل شي <|im_end|>

<|im_start|>user: APE is a task designed to enhance the quality of the translation by performing only minor adjustments to fix any existing translation mistakes. If the translation is already correct, you should retain it as is.

Original (English): we shouldn't be negative all the time

Translation (Modern Standard Arabic): لا يجب ان نكون بهذه السلبيه على طول الدوام.

Post-edited (Tunisian): <|im_end|>

<|im_start|>assistant: Translation: <|im_end|>

Response from the model setting with the highest Chrf++ score:

ما لازم نكون سلبيين على طول الوقت

- **Konkani Example Prompt:**

<|begin_of_text|><|im_start|>user: APE is a task designed to enhance the quality of the translation by performing only minor adjustments to fix any existing translation mistakes. If the translation is already correct, you should retain it as is.

Original (English): Great was his compassion for the two dear ones at this parting moment.

Translation (Marathi): विलग होताना त्याच्या दोन प्रिय व्यक्तींविषयी त्याला अतीव करुणा वाटत होती.

Post-edited (Konkani): <|im_end|>

<|im_start|>assistant: जिवाभावाच्या दोगांयचो त्याग करपी त्या खिणावेळार ताची करुणा सुमराभा–यली आशिल्ली. <|im_end|>

<|im_start|>user: APE is a task designed to enhance the quality of the translation by performing only minor adjustments to fix any existing translation mistakes. If the translation is already correct, you should retain it as is.

Original (English): Suleman's parents were quite tall.

Model	Source	Pivot	Target	k	BLEU	chrF++
Ablation: Number of In-Context Examples (k) using MSA Pivot (No Fine-Tuning)						
Unbabel/TowerInstruct-v0.1						
Unbabel/TowerInstruct-v0.1	Eng	Msa	Aeb	0	4.37	16.45
Unbabel/TowerInstruct-v0.1	Eng	Msa	Aeb	1	3.46	18.74
Unbabel/TowerInstruct-v0.1	Eng	Msa	Aeb	2	4.99	16.57
Unbabel/TowerInstruct-v0.1	Eng	Msa	Aeb	3	4.77	17.32
Unbabel/TowerInstruct-v0.1	Eng	Msa	Aeb	4	3.09	19.80
Unbabel/TowerInstruct-v0.1	Eng	Msa	Aeb	5	3.75	20.63
NousResearch/Hermes-2-Pro-Llama-3-8B						
NousResearch/Hermes-2-Pro-Llama-3-8B	English	Msa	Aeb	0	5.06	24.31
NousResearch/Hermes-2-Pro-Llama-3-8B	English	Msa	Aeb	1	4.93	21.27
NousResearch/Hermes-2-Pro-Llama-3-8B	English	Msa	Aeb	2	3.74	18.18
NousResearch/Hermes-2-Pro-Llama-3-8B	English	Msa	Aeb	3	4.20	20.17
NousResearch/Hermes-2-Pro-Llama-3-8B	English	Msa	Aeb	4	4.93	19.42
NousResearch/Hermes-2-Pro-Llama-3-8B	English	Msa	Aeb	5	4.77	16.32

Table 13: Ablation on the number of in-context examples (k) for English→Msa→Aeb translation

Model	Source	Pivot	Target	k	BLEU	chrF++	ΔchrF++
Ablation: Number of In-Context Examples (k) using Hindi Pivot (No Fine-Tuning)							
NousResearch/Hermes-2-Pro-Llama-3-8B							
NousResearch/Hermes-2-Pro-Llama-3-8B	English	Hin	Gom	0	2.86	25.39	+0.49
NousResearch/Hermes-2-Pro-Llama-3-8B	English	Hin	Gom	1	2.47	24.12	-6.22
NousResearch/Hermes-2-Pro-Llama-3-8B	English	Hin	Gom	2	2.47	23.96	-3.63
NousResearch/Hermes-2-Pro-Llama-3-8B	English	Hin	Gom	3	2.41	23.69	-2.20
NousResearch/Hermes-2-Pro-Llama-3-8B	English	Hin	Gom	4	0.04	3.09	-24.44
NousResearch/Hermes-2-Pro-Llama-3-8B	English	Hin	Gom	5	0.02	2.49	-26.16

Table 14: Ablation on the number of in-context examples (k) for English→Hindi→Konkani translation using Nous Hermes. ΔchrF++ is computed relative to the Marathi-pivot setting at the same k. Scores computed with SacreBLEU (Post, 2018); signatures in Appendix A.2.

Parameter	Value
batch_size	1
num_train_epochs	1.5
warmup_ratio	0.03
logging_steps	25
learning_rate	2e-4
gradient_checkpointing	True
lr_scheduler_type	Cosine
weight_decay	0.001
save_strategy	No
optim	PagedAdam
warmup_steps	100
bf16	True

Table 15: Training parameters used in the model training process.

Parameter	Value
r	64
lora_alpha	16
lora_dropout	0.1
bias	none
task_type	CAUSAL_LM
target_modules	['q', 'k', 'v', 'o', 'up', 'down', 'gate', 'lm_head']

Table 16: LoRA configuration parameters.

```
do_sample: True
temperature: 0.1
num_return_sequences: 1
max_new_tokens: 200
return_full_text: False
```

Table 17: Inference parameters used for text generation.

Translation (Marathi): सुलेमानचे पालक बरेच उंच होते.

Post-edited (Konkani): <|im_end|>

<|im_start|>assistant: सुलेमानाचे पालक खूब उंच आशिल्ले. <|im_end|>

<|im_start|>user: APE is a task designed to enhance the quality of the translation by performing only minor adjustments to fix any existing translation mistakes. If the translation is already correct, you should retain it as is.

Original (English): Our country owes a deep

debt of gratitude to our valiant ex-Servicemen.

Translation (Marathi): आपल्या शूर माजी सैनिकांप्रति आपला देश कृतज्ञतेने अपार ऋणी आहे.

Post-edited (Konkani): <|im_end|>

<|im_start|>assistant: आमचो देश शूरवीर सेवानिवृत्त–सैनिकांक कृतज्ञतायेचें रीण देणें आसा. <|im_end|>

<|im_start|>user: APE is a task designed to enhance the quality of the translation by performing only minor adjustments to fix any existing translation mistakes. If the translation is already correct, you should retain it as is.

Original (English): Boys are equally vulnerable to sexual abuse.

Translation (Marathi): मुलगेही लैंगिक छळाला तेवढेच बळी पडू शकतात.

Post-edited (Konkani): <|im_end|>

<|im_start|>assistant: चलेय लैगिक अत्याचाराची तितलीच शिकार जावंक शकतात. <|im_end|>

<|im_start|>user: APE is a task designed to enhance the quality of the translation by performing only minor adjustments to fix any existing translation mistakes. If the translation is already correct, you should retain it as is.

Original (English): But Mangal Pandey's brave deed was done through devotion to a high and noble principle.

Translation (Marathi): पण मंगल पांडेची शौर्यशाली कृती ही एका उच्च आणि उदात्त तत्त्वाप्रतिच्या समर्पणातून केली गेली होती.

Post-edited (Konkani): <|im_end|>

<|im_start|>assistant: पूण मंगल पांडेचें धाडशी कर्तुब एके उंचेल्या आनी उदार तत्वनिश्ठेचें आशिल्लें. <|im_end|>

<|im_start|>user: APE is a task designed to enhance the quality of the translation by performing only minor adjustments to fix any existing translation mistakes. If the translation is already correct, you should retain it as is.

Original (English): The brothers were deeply attached to each other.

Translation (Marathi): भाऊ एकमेकांना खूप जवळ होते.

Post-edited (Konkani): <|im_end|>

<|im_start|>assistant: Translation: <|im_end|>

Response from the model setting with the highest Chrf++ score:

भावांनी एकमेकांकडेन खूब नजीकाय आशिल्ली.

Model	Source	Pivot	Target	BLEU	CHRF++
Few-shot Finetuned					
Unbabel/TowerInstruct-v0.1	English	-	Konkani	4.18	31.57
NousResearch/Hermes-2-Pro-Llama-3-8B	English	-	Konkani	3.49	31.49
Unbabel/TowerInstruct-v0.1	English	Marathi	Konkani	7.80	17.60
NousResearch/Hermes-2-Pro-Llama-3-8B	English	Marathi	Konkani	12.14	34.92
Zero-shot Finetuned					
Unbabel/TowerInstruct-v0.1	English	-	Konkani	1.89	17.39
NousResearch/Hermes-2-Pro-Llama-3-8B	English	-	Konkani	4.01	36.61
Unbabel/TowerInstruct-v0.1	English	Marathi	Konkani	7.94	31.91
NousResearch/Hermes-2-Pro-Llama-3-8B	English	Marathi	Konkani	8.38	40.17

Table 18: Performance comparison of finetuned models in few-shot and zero-shot settings for Konkani translation.

Model	Source	Pivot	Target	BLEU	CHRF++
Few-shot Finetuned					
Unbabel/TowerInstruct-v0.1	English	-	Tn	3.3	21.05
NousResearch/Hermes-2-Pro-Llama-3-8B	English	-	Tn	NA	NA
Unbabel/TowerInstruct-v0.1	English	Msa	Tn	2.82	17.12
NousResearch/Hermes-2-Pro-Llama-3-8B	English	Msa	Tn	8.02	35.99
Zero-shot Finetuned					
Unbabel/TowerInstruct-v0.1	English	-	Tn	1.48	14.83
NousResearch/Hermes-2-Pro-Llama-3-8B	English	-	Tn	5.02	18.07
Unbabel/TowerInstruct-v0.1	English	Msa	Tn	2.09	19.16
NousResearch/Hermes-2-Pro-Llama-3-8B	English	Msa	Tn	4.62	21.87

Table 19: Performance comparison of finetuned models in few-shot and zero-shot settings for Tunisian Arabic translation.