Convergence
A Journal for Young Researchers

ORIGINAL REVIEW

Received: 7-29-25; Accepted: 10-31-25;
Managing Editor: Jonas Katona;
Action Editor: Jonas Katona

# Enhancing Machine Translation in Low-Resource Languages: A Comparative Review of Prompt-Based and Fine-Tuning Methods

**Ryoma Suzuki**

UWCSEA Dover Campus, Singapore

**Abstract**

Large language models (LLMs) such as LLama, GPT, T5, and Alpaca have demonstrated capabilities in multilingual tasks with natural language processing (NLP). However, low-resource languages often lack data regarding tokens and linguistic representations, which are critical for effective training of the models. This limits the models' performance, particularly in translation, compared to high-resource language models, which benefit from large labeled datasets. There is a clear need to address the imbalance of effectiveness of LLMs in low-resource languages.

While many methods exist to address this problem, it is unclear which method is best. In this literature review, we evaluate recently-developed methods that attempt to enhance the translation performance of LLMs in low-resource languages. Specifically, we focus on prompt-based and fine-tuning methods, which have been shown to significantly improve the performance of LLMs in low-resource languages. Among the methods reviewed, parameter-efficient fine-tuning (PEFT)—including adapters and prefix tuning—demonstrated the best balance between performance and cost, making them the most practical for low-resource languages. However, traditional full fine-tuning (FFT) methods achieve greater performance improvements despite higher computational requirements. These findings highlight the importance of balancing performance gains with resource constraints, and architectural enhancement methods with culturally and quantitatively-limited data.

**Keywords:** low-resource languages, large language models (LLMs), machine translation (MT), prompt-based methods, fine-tuning methods, parameter-efficient fine-tuning (PEFT), computational efficiency, translation performance, adapters, prefix tuning

## 1. Introduction

Prior works of adapting large language models (LLMs) to specific domains used full fine-tuning, which can be costly and ineffective, especially for low-resource languages. Though researchers have proposed many methods to overcome these disadvantages, in its current state, there are too many options to choose from.

The fast-developing technology of LLMs like Llama, GPT, T5, and Alpaca has transformed how humans interact with machines through content creation and human-like dialogue (Qin et al., 2024). For example, generative LLMs can generate content from any user input, such as text, images, sound, etc. These models operate using transformer-based architectures and are trained on extensive text corpora to identify linguistic patterns and accurately predict subsequent tokens. This enables LLMs to generate natural and coherent texts within a very short time.

The benefits of LLMs are not equally distributed globally. A major contributor to this is the disparity of performance across different languages. The majority of LLMs are trained on high-resource languages, primarily English, where large amounts of text data are easily accessible. Therefore, LLMs respond significantly better to English commands and prompts than to languages that have less publicly accessible corpora, or 'low-resource' languages (Ming et al., 2024; "Multilingual Large Language Models and Curse of Multilinguality," 2025). In one evaluation of translation tasks, English prompts achieved an average BLEU score of 92.5%, while Hindi prompts scored only 7.5% ("Multilingual Prompting in LLMs," 2024). BLEU is a standard metric that evaluates how closely machine translations match human translations, with higher scores indicating better translation quality (Papineni et al., 2002). There are approximately 3 billion people worldwide who speak low-resource languages, and disparities in performance and usability cause inequality in linguistic representation and opportunity. For example, LLMs trained only in English cannot serve as a highly useful tool when it comes to assisting local governments and indigenous communities.

Researchers have tackled this problem by using the multilingual pretraining method (such as Macro-LLM), which involves training LLMs on texts from a mixture of languages to facilitate knowledge transfer across different languages (Ming et al., 2024). This approach enhances the models' ability to understand and generate low-resource language texts by leveraging knowledge from high-resource languages. However, this method caused a regression of the model's ability to respond to high-resource languages ("Multilingual Large Language Models," 2025). Additionally, since the multilingual pretraining method requires large amounts of data for training, the imbalance in data availability between high-resource and low-resource languages leads to uneven performance.

Recently, novel approaches have been developed to address this problem. Fine-tuning methods involve directly adjusting parameters of weights and biases (i.e., the numbers used to make predictions). Full fine-tuning (FFT) changes all parameters, while parameter-efficient fine-tuning methods (PEFT) change only a small fraction, such as adapters (4%), prefix tuning (0.1%), and prompt tuning (0.01%).

Prompt-based methods do not change the model at all. Instead, they change the prompt (input text) by adding guiding words or phrases that steer the model toward the right answer. This includes Dictionary-based Phrase-level Prompting Machine Training (DiPMT), Retrieved Phrase-level Prompts (RePP), and prompt tuning (Ghazvininejad et al., 2023; Sun et al., 2022; Lester et al., 2021).

This review makes three contributions. First, we specify a structured search methodology. Second, we point towards specific methods and extract comparable data items across studies, such as model size, language pairs, metrics, percentage of parameters used, and tokens processed. Third, we provide a practical decision framework for reference on when to use which methods.

## 2. Methodology

The papers reviewed were selected from Google Scholar, ACL Anthology, and arXiv based on clearly defined criteria. Each paper clearly focuses on adapting large language models, empirically evaluates either prompt-based or fine-tuning methods, and includes detailed evaluations regarding performance, computational efficiency, and costs. Each paper must also be peer-reviewed and written in English text format. Our goals were to identify recent methods specifically aimed at enhancing MT performance in low-resource languages, to critically evaluate these methods using criteria such as computational practicality, implementation cost, and effectiveness, and to recommend the most suitable methods for practical implementation, considering restricted data constraints and linguistic variations in low-resource languages. Below are data that were extracted from the papers:

- Model (size)
- Language paris
- Methodology (prompt-based; full FT; PEFT)
- Metrics
- Percentage of parameters tuned
- Tokens processed
- Costs if reported

In this review, prompt-based methods are defined as an approach taken with zero modification of the parameters of the original model, focusing on enhancing the prompts to yield better performance from the model. Conversely, fine-tuning methods refer to approaches that directly modify the parameters of the original model, thereby refining its transformer architecture or neural network structure to produce more effective outputs.

## 3. Comparative Results and Analysis

### 3.1. Prompt-based Methods

LLMs can be adapted to low-resource translation without retraining by modifying their inputs or prompts. Prompt-based methods require minimal computational cost and parameter updates (Liu et al., 2021; Shin et al., 2025). They differ mainly in how linguistic hints are provided to the model.

Initial language models, such as N-grams and Recurrent Neural Networks (RNNs), laid the foundation for prompt-based methods. Large-scale models such as GPT-3 enhanced the ability and adaptability of prompt engineering.

In this section, we compare methods recently developed by other researchers: Dictionary-based Phrase-Level Prompting (DiPMT), Retrieval Phrase-Level Prompting (RePP), and prompt tuning (Ghazvininejad et al., 2023; Lester et al., 2021; Sun et al.,

2022).

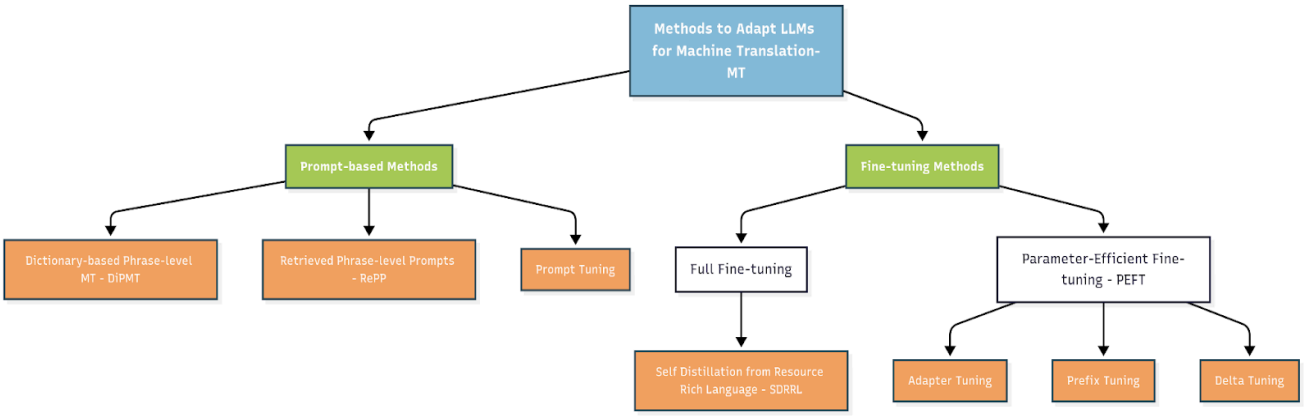


**Figure 1:** Methods to adapt LLMs for machine translation
*Note*: The previously identified methods are categorized into two large parts: prompt-based methods and fine-tuning methods.

**DiPMT: Description and Key Results**

Dictionary-based Phrase-level Machine Translation (DiPMT) was developed in 2023 by Meta AI (Ghazvininejad et al., 2023). DiPMT is a domain-specific method specialized in the task of machine translation. The core idea behind DiPMT is to enhance prompts by using a bilingual dictionary to provide hints for specific words, especially rare ones, in order to narrow the range of possible interpretations for the LLM. In DiPMT, strings describing the meaning of the keywords (e.g., the word "pembuatan" means "creation") are appended to the original prompt. The addition of these strings expands the model's capability to accurately translate the original prompt. As a result of incorporating this method, the model showed an improvement in the BLEU compared to the baseline LLM output, which omits the dictionary hints from the prompt, with a mean increase of +4.8 and a standard deviation of ±1.2 across 17 of the 20 languages tested. Figure 2 shows an example prompt in a DiPMT.

Translate the following sentence to English: Ia melakukan pembuatan bel pintu dengan teknologi WiFi, katanya.
In this context, the word "pembuatan" means "creation"; the word "bel" means "buzzer", "bell"; the word "pintu" means "door", "doors".
The full translation to English is:

**Figure 2:** Example of a zero-shot prompt using DiPMT
*Note*: Green words, which act as hints, are added to the original blue & red instructions.

This method highly depends on the type of coverage—word type coverage or word token coverage—of the dictionary for the target language. The word type coverage of a dictionary represents the proportion of distinct words in the language that exist in the dictionary. A high word type coverage means the dictionary is robust and covers a diverse vocabulary. On the other

hand, word token coverage considers the percentage of the dictionary that can be translated in real-world situations. Thus, small dictionaries can have high token coverage if they contain frequent common words in the language. This relationship underscores the importance of lexical coverage: when dictionaries include a wide variety of word types, models can better clarify meanings during translation, especially in underrepresented languages. The study shows that DiPMT performs substantially better only when the type coverage is above a certain level (5–20%). Typical lexical coverage for low-resource languages is around 20–50%, which suggests that DiPMT is practically useful in real-life situations. The performance of the method is also highly dependent on the hint that is provided. The prompts produced by the normal DiPMT method provide all available dictionary translations of words in the original prompt, creating multiple competing options that can be challenging for LLMs to determine the most contextually appropriate choice. The results from the paper show that the output from this method can sometimes be inaccurate since the model could potentially choose the wrong definition. Although this method can be highly accurate when "gold hints"—correct translations of some key words in the original prompts—are provided, this may not be possible to attain for every case of translation. Thus, a method that would allow gold hints to be derived from dictionary hints will increase the effectiveness of DiPMT significantly; otherwise, performance may be limited. Overall, DiPMT is strongest for rare-word translation when a robust bilingual dictionary is available, but its success is tightly constrained by dictionary coverage.

**RePP: Description and Key Results**

Retrieved Phrase-level Prompts (RePP) enhances LLMs' outputs by retrieving relevant phrase-level data from a bilingual dictionary database and combining it with the input before processing in the LLM (Sun et al., 2022). Additional precision from the database clarifies meanings for the LLM to process, and using phrase-level data limits the possibility of interpretations compared to sentence-level data.

RePP enhances the prompt in three automated processes: segmenting the input into phrases, retrieving aligned bilingual phrases from a pre-built database, and concatenating them before translation. RePP specially integrates phrase-level bilingual data without retraining the model, improving clarity and cross-lingual alignment. In benchmark tests on English-German and German-English translation, RePP improved mean BLEU scores by +6.2 compared to baseline models, performing slightly below FFT. Because the bilingual phrases database combines multiple dictionaries, its word type coverage and token coverage are both high, making RePP more stable than DiPMT, which relies only on one dictionary. Unlike FT, RePP requires only a single base model for multiple domains, since each input prompt is augmented with phrase-level bilingual data. This adaptability makes it especially robust for low-resource languages, where corpora (large text data) and funding for advanced models are scarce.

**Prompt Tuning: Description and Key Results**

Prompt tuning is a parameter-efficient strategy that freezes the pretrained parameters and inserts a small, learnable prompt vector into the model's embedding layer (Lester et al., 2021). The vector—a sequence of virtual tokens prepended to the input—guides the model toward task-specific outputs without altering its original parameters. In the presented paper, Lester et al. analyzed prompt tuning performance on a T-5 XXL model using GLUE and SuperGLUE benchmarks. This method matched FFT performance (approximately 89.3% accuracy) while updating only approximately 0.01% of parameters.

Because only the prompt vectors are developed, the computational cost is negligible. Freezing the original parameters also

preserves high-resource-language performance, an advantage in multilingual machine translation, where maintaining English or other prominent language accuracy while adding low-resources translation capability is essential.

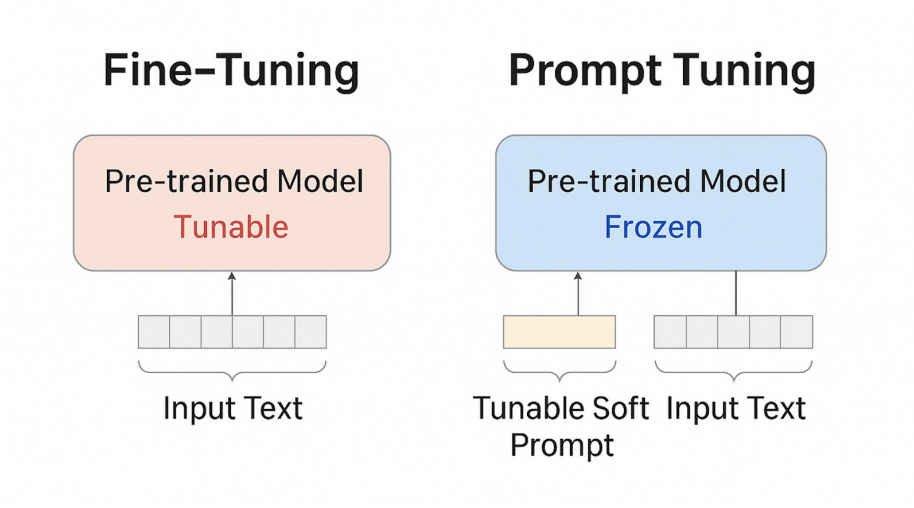


**Figure 3:** Conceptual difference between fine-tuning and prompt tuning

**Comparison and Trends among Prompt-Based Methods**

**Table 1:** Performance and characteristics of prompt-based methods (DiPMT, RePP, and prompt tuning) across low-resource machine translation tasks

| Method | % of parameters modified | Data or resource required | Relative cost | MT performance (BLEU, GLEU) | Best use cases | Risk & cost |
|---|---|---|---|---|---|---|
| DiPMT | 0% | Bilingual dictionary | Low (minimal) | Average +4.8 BLEU; especially effective in out-of-domain situation | Sentences with rare, uncommon words; dictionaries with high word coverage | If the dictionary lists multiple translations, it can lead to lower accuracy. Quality depends on the dictionary's word coverage. |
| RePP | 0% | Phrase pair database made from multiple dictionaries | Low | +6.2 BLEU on En-De | Fast domain adaptation when phrases are available; | Building a phrase pair database highly relies on the accuracy of the |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | sentences with idiomatic expressions | phrase pair. |
| Prompt tuning | ~0.01% | None | Low to moderate | Achieving close to FFT in the GLEU score | Protecting the main resource-rich language's performance | Underperforms compared to DiPMT and RePP. |

*Note*: BLEU is a translation quality metric, with higher scores indicating better translation.

Methods that bring the right translation words improve the performance the most. DiPMT scales with dictionary word coverage: when word coverage is below 20%, it significantly diminishes improvements. RePP is more consistent, as it uses multiple bilingual dictionaries to create a phrase pair database, but requires more time and cost to build. Pure prompting methods keep costs near zero, but wrong or noisy hints can mislead the model and lower performance. Thus, the quality of the external source matters significantly. Prompt tuning changes almost nothing, making it ideal for maintaining multilingual breadth while adding a task, but it usually lacks precision compared to DiPMT and RePP.

### 3.2. Fine-Tuning Methods

LLMs can also be adapted to low-resource language translation by retraining parts of their internal parameters through fine-tuning. Unlike prompt-based methods, which modify inputs, fine-tuning methods directly adjust the model's behavior (Ding et al., 2023; Houlsby et al., 2019; Li & Liang, 2021). Early approaches to full fine-tuning (FFT) retrained all parameters. This method achieved strong task specialization at the expense of cost-efficiency and scalability. To reduce this burden, methods like Self-Distillation from Resource-Rich Language (SDRRL) and parameter-efficient fine-tuning (PEFT) have emerged, which update only a fraction of parameters while keeping most of the model frozen.

Parameter-efficient fine-tuning (PEFT) is a novel approach to address the issue of increasing storage size, which escalates the computational and memory costs of LLMs (Ding et al., 2023). It focuses on tuning a limited number of parameters and modularity, meaning one LLM can be used for multiple specific tasks. The PEFT we reviewed are:

1. Adapters
2. Prefix Tuning

**SDRRL: Description and Key Results**

Self-Distillation from Resource-Rich Languages (SDRRL) mitigates data scarcity by transferring patterns from high-resource to low-resource languages (Zhang et al., 2024). In SDRRL, the model first generates an exemplar response in a high-resource language (e.g, English), assuming it is accurate. Then, it translates the responses to the target low-resource language and compiles these pairs into a data set used for supervised fine-tuning. This process of cross-lingual transfer boosts low-resource performance but can slightly reduce the accuracy in the high-resource language, as some parameters are

overwritten during supervised fine-tuning. Zhang et al. (2024) observed that SDRRL improved BLEU by 2–5 points across 14 languages while largely preserving accuracy in English and French. SDRRL creates its own data by translating high-quality output, which effectively removes the dependence on existing bilingual corpora. By transferring sentence-level grammatical knowledge from high-resource languages rather than individual lexical pairs, the model learns linguistic structures—such as word order, tense, and dependencies—that dictionaries may not capture. This deeper grammatical learning advances fluent translation in low-resource settings.

**Adapters: Description and Key Results**

Recognizing the issue of increasing required storage size, Houlsby et al. proposed a method in 2019 that takes a novel approach to enhance the practicality of fine-tuning (Houlsby et al., 2019). In this method, adapters—tiny, trainable layers— are added between frozen layers of the transformer model (Vaswani et al., 2017). During fine-tuning, only these adapter layers are trained, making the process faster and cost-effective. Typically, about 4% of parameters are tuned while 96% remain fixed. When training 1 billion tokens, adapter fine-tuning costs roughly $200 USD compared to thousands for FFT. Adapters improved F1/BLEU by roughly 12% across 8 low-resource languages (Gurgurov et al., 2024). This architecture allows the model to remember what is already known from high-resource languages and enables cross-lingual transfer, which can efficiently enhance the performance of low-resource languages while maintaining performance in high-resource languages. Because adapters can be added or removed like plug-ins, one shared model can hold multiple language-specific adapters at once, each trained separately and loaded only when needed, ensuring the model's robustness.

**Prefix Tuning: Description and Key Results**

Similar to adapters, prefix tuning is a parameter-efficient fine-tuning inspired by prompting (Li & Liang, 2021). It freezes the model's parameters to minimize the increase in model storage. Unlike adapters, prefix tuning adds a task-specific vector to each layer in the transformer to guide the original LLM to a more precise output. These vectors, or parameters, are prepended in front of every attention block in the transformer. Prefix tuning a GPT-4.1 model only costs about $7.50 USD, significantly reducing the need for large investments.

**Comparison and Trends among Fine-Tuning Methods**

**Table 2:** Comparison of fine-tuning strategies (full fine-tuning, SDRRL, adapters, and prefix tuning) on efficiency, data use, and translation quality

| Method | % of parameters modified | Data or resource needed | Relative cost | MT performance (BLEU, GLUE) | Best use cases | Risk & cost |
|---|---|---|---|---|---|---|
| FFT | 100% | Large parallel corpora | Very high (~$7500 USD per model) | ~+12 BLEU in domain-specific tasks | High budget and data; maximum stability on one domain | Forgetting other languages or domains; expensive |
| SDRRL | >50% | Output from | Medium | ~+5 BLEU across | Lacking | Some loss to |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | high-resource languages | to high | 14 languages from English and French models | low-resource language data; access to high-resource language models | high-resource language performance; though less than FFT, high-cost |
| Adapter | ~4% | Moderate bilingual data | Low to medium | Near FFT accuracy in most low-resource language settings; highest among PEFT | High accuracy per cost; use across multiple domains | Managing many adapters; some tasks may perform lower than FFT. |
| Prefix tuning | ~0.1% | Moderate bilingual data | Low | Near FFT accuracy in most low-resource language settings | Moderate accuracy per cost; use across multiple domains; rapid development | May not have reached the adapter or FFT performance. |

The trend in fine-tuning models shows that the performance strongly correlates with the percentage of parameters modified. Updating more parameters tends to increase BLEU, with FFT at the top and PEFT in the middle, but with higher cost and lower performance in the original high-resource language.

Adapters and prefix-tuning increase BLEU nearly as much as FFT for a fraction of the cost. SDRRL makes its own training pairs by translating high-resource language output first, improving BLEU scores for low-resource languages when parallel data is scarce.

### 3.3. Cross Category Evaluation

**Table 3:** Comparative overview of adaptation methods for low-resource translation

| Method Type | Example Methods | Parameters Modified | Data Efficiency | Relative Cost | BLEU | Domain-specific Performance | Strengths | Limitations |
|---|---|---|---|---|---|---|---|---|
| Prompt-based | DiPMT, RePP, prompt tuning | 0-0.01% | Effective with minimal data (e.g., zero-shot) | Lowest (<$1 to $10 USD per run) | +4-6 BLEU gain over baseline | Limited improvement for highly specialized tasks | Zero retraining; fast deployment ; high | Lowest accuracy; depends on lexical coverage |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | accessibility | |
| Fine-tuning methods | PEFT (adapters, prefix tuning), SDRRL | 0.1-50% | Moderate data requirement (e.g., bilingual corpora) | Low to Medium (~$7.50–$200 USD per billion tokens) | +8-13 | Superior at adapting to specialized domains; captures grammatical and semantic structure | Reusable modules; preserves primary language competence; high accuracy per cost | Medium to high accuracy; architectural complexity |
| Full fine-tuning (FFT) | Traditional FT | 100% | Requires large-scale parallel datasets | Very High (~$7500 USD per model) | +10-12 | Highest precision and stability on task-specific domains | Full adaptation; maximal accuracy | Costly; prone to forgetting |

The difference in machine-translation performance across methods primarily arises from the percentage of parameter modification. SDRRL retrains >50% of the model, achieving the greatest improvement (~+13 BLEU) with medium-high computational costs. PEFTs, like adapters, achieve nearly the same improvements as FFT (~+12 BLEU) while altering only small portions of the model's parameters (~4%), allowing domain-specific adaptation while retaining pretrained knowledge. By contrast, prompt-based methods, like prompt tuning, modify only ~0.01% of parameters, proposing quicker adaptation, but with significantly less improvement (~+6 BLEU).

Differences in data efficiency also influence the translation quality. Prompt-based methods perform effectively with minimal data, operating under zero- or few-shot environments without bilingual corpora. PEFT methods, such as prefix and adapter, require moderate training data. Consequently, data-training requirements increase in proportion to the percentage of parameters updated. The relative cost is proportional to the percentage of parameters modified. Prompt-based and prefix tuning methods operate at under $10 USD per modification, while adapters require ~$200 USD per 1 billion tokens (Houlsby et al., 2019). SDRRL and FFT require the highest costs, ranging from $2000–$7500 USD per model (Ding et al., 2023). These quantitative comparisons establish the relative strengths of each method across performance and cost dimensions. The broader implications of these trade-offs for low-resource language settings are discussed in Section 4.

## 4. Discussions

### 4.1. Restatement of Key Findings

Evaluations demonstrate clear proportionality between parameter modification, computational costs, and MT performance. Methods updating fewer than 1% of parameters yield modest BLEU improvements (~+6), whereas those up to ~50%

modification achieve near-FFT accuracy (~+13) at a much lower cost. FFT consistently achieves the highest accuracy but with the greatest cost burden (~$7500 USD per model).

This trend reflects each method's mechanism. Prompt-based methods rely on lexical or phrase-level, bit-by-bit translation rather than direct parameter adjustments, which explains their speedy development but limited domain precision. PEFT methods introduce lightweight trainable modules to enhance the accuracy of domain-specific tasks while preserving pre-trained linguistic knowledge in resource-rich languages. SDRRL's cross-lingual self-distillation enables transfer of linguistic knowledge from resource-rich languages, explaining the marginal lead over other fine-tuning methods.

### 4.2. Implications and Significance

The findings presented here underscore meaningful advances in adapting large language models to low-resource languages, opening promising pathways for broader AI accessibility. These evaluations highlight that, despite varying in strengths across methods, parameter-efficient fine-tuning (PEFT) methods—particularly adapters and prefix tuning—strike an effective balance between performance, cost-efficiency, and ease of implementation. Their modular and parameter-minimal approaches significantly reduce computational overheads, enhancing their practicality for deployment by smaller enterprises or individual developers. Nonetheless, ongoing research should focus on creating standardized evaluation metrics and customizable frameworks to accommodate diverse linguistic structures and data availability. Recognizing these factors will further improve method selection and optimization, ultimately contributing to a more inclusive and effective AI landscape. By continuing to refine these adaptable methods, the research community can better serve diverse linguistic populations, fostering equitable technological benefits across previously underserved communities.

### 4.3. Limitations

Limitations include the absence of a standardized benchmark dataset for comprehensive comparison across different methods, potential biases introduced by qualitative rather than quantitative assessment criteria, and inherent disparity of linguistic complexity amongst low-resource languages.

First, the comparison of the methods was done qualitatively. This is because there are no fair, comprehensive tests to assess the performance of all of the methods, so different papers used distinct methods to assess their models.

Furthermore, performance varies between low-resource languages. For example, RePP may have performed extensively with English-German but may not have performed better in another low-resource language. This is because even within low-resource languages, the amount of data and their distinct linguistic structure vary. This sheds light on specific language-focused models.

## 5. References

Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C.-M., Chen, W., Yi, J., Zhao, W., Wang, X., Liu, Z., Zheng, H.-T., Chen, J., Liu, Y., Tang, J., Li, J., & Sun, M. (2023). Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3), 220–235. https://doi.org/10.1038/s42256-023-00626-4

Ghazvininejad, M., Gonen, H., & Zettlemoyer, L. (2023). Dictionary-based phrase-level prompting of large language models for machine translation (No. arXiv:2302.07856). arXiv. https://doi.org/10.48550/arXiv.2302.07856

Gurgurov, D., Bäumel, T., & Anikina, T. (2024). Multilingual large language models and curse of multilinguality (No. arXiv:2406.10602). arXiv. https://doi.org/10.48550/arXiv.2406.10602

Gurgurov, D., Hartmann, M., & Ostermann, S. (2024). Adapting multilingual LLMs to low-resource languages with knowledge graphs via adapters. *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024)*, 63–74. https://doi.org/10.18653/v1/2024.kallm-1.7

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., Laroussilhe, Q. de, Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-efficient transfer learning for NLP (No. arXiv:1902.00751). arXiv. https://doi.org/10.48550/arXiv.1902.00751

Lester, B., Al-Rfou, R., & Constant, N. (2021). The power of scale for parameter-efficient prompt tuning (No. arXiv:2104.08691). arXiv. https://doi.org/10.48550/arXiv.2104.08691

Li, X. L., & Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation (No. arXiv:2101.00190). arXiv. https://doi.org/10.48550/arXiv.2101.00190

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing (No. arXiv:2107.13586). arXiv. https://doi.org/10.48550/arXiv.2107.13586

Ming, L., Zeng, B., Lyu, C., Shi, T., Zhao, Y., Yang, X., Liu, Y., Wang, Y., Xu, L., Liu, Y., Zhao, X., Wang, H., Liu, H., Zhou, H., Yin, H., Shang, Z., Li, H., Wang, L., Luo, W., & Zhang, K. (2024). Marco-LLM: Bridging languages via massive multilingual training for cross-lingual enhancement (No. arXiv:2412.04003). arXiv. https://doi.org/10.48550/arXiv.2412.04003

Multilingual prompting in LLMs: Investigating the accuracy and performance. (2023). *International Journal of Scientific Research in Engineering and Management*, 7(2). https://doi.org/10.55041/IJSREM17694

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In P. Isabelle, E. Charniak, & D. Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (pp. 311–318). *Association for Computational Linguistics.* https://doi.org/10.3115/1073083.1073135

Qin, L., Chen, Q., Feng, X., Wu, Y., Zhang, Y., Li, Y., Li, M., Che, W., & Yu, P. S. (2024). Large language models meet NLP: A survey (No. arXiv:2405.12819). arXiv. https://doi.org/10.48550/arXiv.2405.12819

Shin, J., Tang, C., Mohati, T., Nayebi, M., Wang, S., & Hemmati, H. (2024). Prompt engineering or fine-tuning: An empirical assessment of LLMs for code (No. arXiv:2310.10508). arXiv. https://doi.org/10.48550/arXiv.2310.10508

Sun, Z., Jiang, Q., Huang, S., Cao, J., Cheng, S., & Wang, M. (2022). Zero-shot domain adaptation for neural machine translation with retrieved phrase-level prompts (No. arXiv:2209.11409). arXiv. https://doi.org/10.48550/arXiv.2209.11409

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Zhang, Y., Wang, Y., Liu, Z., Wang, S., Wang, X., Li, P., Sun, M., & Liu, Y. (2024). Enhancing multilingual capabilities of large language models through self-distillation from resource-rich languages (No. arXiv:2402.12204). arXiv. https://doi.org/10.48550/arXiv.2402.12204

## Acknowledgements

I am deeply grateful to my parents for their unwavering support and encouragement throughout the development of this paper. I also thank my mentor for his invaluable guidance and feedback in refining its structure and clarity.

## Author Biography

**Ryoma Suzuki** is a student researcher interested in improving machine translation for low-resource languages. His work compares prompt-based and fine-tuning methods for large language models and explores fair, accessible language technologies. Outside of his research, he mentors peers in STEM and volunteers with language-learning programs.

## Mentor Contribution Statement

**Mu Taka** provided thoughtful guidance and consistent support throughout the development of this research project. His mentorship was instrumental in refining the research design, sharpening the analytical approach, and improving the paper's overall clarity and coherence. Through constructive discussions and detailed feedback, he helped strengthen the logical flow of the paper's arguments and ensure the study's findings were presented with precision and depth. All analysis and writing were conducted independently by the author.