IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 26 December 2024, accepted 2 March 2025, date of publication 10 March 2025, date of current version 20 March 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3549795

RESEARCH ARTICLE

Toward Low-Resource Languages Machine Translation: A Language-Specific Fine-Tuning With LoRA for Specialized Large Language Models

XIAO LIANG[1,2], YEN-MIN JASMINA KHAW[1], SOUNG-YUE LIEW[3], TIEN-PING TAN[4], AND DONGHONG QIN[2], (Member, IEEE)

[1]Department of Computer Science, Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman, Kampar 31900, Malaysia
[2]School of Artificial Intelligence, Guangxi Minzu University, Nanning 530008, China
[3]Department of Computer and Communication Technology, Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman, Kampar 31900, Malaysia
[4]School of Computer Sciences, Universiti Sains Malaysia, George Town 11700, Malaysia

Corresponding author: Yen-Min Jasmina Khaw (khawym@utar.edu.my)

This work was supported by the Ministry of Higher Education (MoHE) through the Fundamental Research Grant Scheme under Grant FRGS/1/2022/ICT02/UTAR/02/6.

ABSTRACT In the field of computational linguistics, addressing machine translation (MT) challenges for low-resource languages remains crucial, as these languages often lack extensive data compared to high-resource languages. General large language models (LLMs), such as GPT-4 and Llama, primarily trained on monolingual corpora, face significant challenges in translating low-resource languages, often resulting in subpar translation quality. This study introduces Language-Specific Fine-Tuning with Low-rank adaptation (LSFTL), a method that enhances translation for low-resource languages by optimizing the multi-head attention and feed-forward networks of Transformer layers through low-rank matrix adaptation. LSFTL preserves the majority of the model parameters while selectively fine-tuning key components, thereby maintaining stability and enhancing translation quality. Experiments on non-English centered low-resource Asian languages demonstrated that LSFTL improved COMET scores by 1-3 points compared to specialized multilingual machine translation models. Additionally, LSFTL's parameter-efficient approach allows smaller models to achieve performance comparable to their larger counterparts, highlighting its significance in making machine translation systems more accessible and effective for low-resource languages.

INDEX TERMS Machine translation, low-resource languages, large language models, parameter-efficient fine-tuning, LoRA.

I. INTRODUCTION

In recent years, the field of machine translation (MT) has witnessed significant technological innovations, especially since the widespread adoption of neural network models, particularly large language models, such as GPT-4 [1], Llama [2], Claude [3] and Mistral [4]. These models, leveraging deep learning algorithms, have dramatically enhanced translation quality and efficiency, achieving near-human levels of performance across many language pairs [5], [6], [7]. These advances are not just limited to the enhancement of algorithmic accuracy but also encompass the expansion of MT's usability in more complex linguistic contexts and integration into real-world applications [8], [9]. However, despite these achievements, several challenges persist that limit the scalability of machine translation solutions across diverse linguistic landscapes.

One of the most pressing challenges is the translation for low-resource languages. Low-resource languages face significant challenges in machine translation due to the

The associate editor coordinating the review of this manuscript and approving it for publication was Huaqing Li.

lack of digital corpora and computational resources, leading to slower advancements compared to high-resource languages [10]. These languages often suffer from a scarcity of bilingual corpora, which are crucial for training effective deep learning models and thereby hinder the improvement of translation quality [11], [12], [13], [14]. Additionally, the inherent linguistic and lexical peculiarities of low-resource languages, coupled with their limited linguistic resources and expertise, make standard translation models less effective [15], [16]. Specialized large language models, such as the ''No Language Left Behind'' project (NLLB) [17] from Meta AI, have initially achieved translation between low-resource language pairs, but further improvements in translation effectiveness are still needed under limited computational resources. Overcoming these hurdles requires not only innovative modeling strategies but also a nuanced understanding of linguistic diversity and resource allocation.

To address these issues, this paper proposes Language-Specific Fine-Tuning with Low-rank adaptation (LSFTL), a targeted approach designed to enhance the adaptability and efficiency of specialized LLMs specifically for low-resource languages. The main objective of this study is to optimize translation performance for low-resource languages by developing a parameter-efficient fine-tuning method that can improve translation quality without the need for extensive computational resources. Unlike traditional fine-tuning methods, which are often computationally expensive and typically necessitate the use of professional-grade GPUs, LSFTL is engineered to optimize fine-tuning efficiency, allowing its implementation even on consumer-grade graphics cards. This adjustment enables the deployment of multilingual machine translation models more broadly, thus supporting the translation capabilities of LLMs without the extensive computational demand. LSFTL distinguishes itself from traditional fine-tuning methods by focusing on the following innovative aspects:

- LSFTL proposes a fine-tuning methodology that adjusts the parameters of large language models specifically to the needs of low-resource languages. This approach adjusts parameters at specific locations in the model to more quickly align with the linguistic characteristics of each target language, thereby enhancing translation accuracy.
- Recognizing that different layers of neural networks capture various aspects of language, LSFTL investigates which specific layers and module are most effective for custom adjustments within complex model architectures. By systematically analyzing and applying configuration patterns to selected layers and modules, LSFTL maximizes the efficiency of model adjustments, ensuring that each layer's potential to contribute to language-specific tasks is fully utilized.
- This paper explores how LSFTL affects the performance and operational efficiency of machine translation systems compared to traditional fine-tuning models. By experimenting with various methods to modify and configure model parameters, LSFTL aims to achieve an optimal balance between translation quality and computational demands. This aspect is crucial for deploying advanced MT systems in resource-constrained environments, which is typical for low-resource language applications.

The remainder of this paper is organized as follows. Section II reviews related work, including advancements in LLMs, parameter-efficient fine-tuning methods, and translation for low-resource languages. Section III describes the proposed methodology, detailing the design and implementation of LSFTL. Section IV outlines the experimental setup and evaluation metrics. Section V presents the experimental results and analysis, and Section VI concludes the paper with a summary of findings and directions for future research.

II. RELATED WORK

In this section, we review the recent advancements and methodologies in the field of machine translation, focusing on four main areas: general large-scale language models and their impact on MT, specialized multilingual translation models, parameter-efficient fine-tuning (PEFT) methods, and the low-rank adaptation (LoRA) approach.

A. GENERAL LARGE-SCALE LANGUAGE MODELS AND MACHINE TRANSLATION

Recent advancements in natural language processing (NLP), particularly through the development of large-scale language modeling, have significantly influenced MT and other NLP tasks [17], [18], [19]. The advent of Generative Pre-trained Transformers (GPT) pioneered by [20] has opened new avenues for constructing more effective translation systems. These models are known for their ability to generate coherent and context-aware text, setting new benchmarks for machine translation systems.

The widespread use of GPT models has shifted the focus from traditional model-specific fine-tuning to more flexible, scalable approaches that harness the power of large datasets and generalized pre-training [21]. Unlike conventional Neural Machine Translation (NMT) systems that require extensive parallel text for training, GPT models leverage vast amounts of unlabeled text, allowing them to learn a richer representation of language nuances and complexities [22]. This shift not only enhances the quality of translation outputs but also expands the potential application of translation technologies across less-resourced languages, thereby democratizing access to information.

GPT models differ significantly from traditional NMT systems, which are typically based on an encoder-decoder architecture [23]. While NMT models encode the source sentence and decode the target language based on the encoded information, GPT models function as decoder-only systems, processing both the context and the source text as a single input to generate subsequent outputs [24]. This architecture allows GPT models to maintain contextual continuity more

effectively, which is particularly advantageous in complex translation tasks involving nuanced language features.

Furthermore, GPT models are predominantly trained on monolingual data [25], often with a strong bias towards English, which contrasts sharply with the diverse, bilingual training datasets essential for traditional NMT [26], [27]. This training approach enables GPT models to develop robust in-context capabilities that can generalize across multiple languages and domains. However, the heavy reliance on monolingual training data and the extensive parameter count required for effective multilingual translation pose significant challenges in terms of computational efficiency and resource allocation. These challenges necessitate innovative approaches to enhance the multilingual capabilities of large language models without compromising on the quality of translation.

B. SPECIALIZED MULTILINGUAL TRANSLATION LARGE-SCALE LANGUAGE MODELS

Building on the need to transcend the limitations of monolingual datasets, the M2M-100 model [18] from META AI represents a significant advancement in the domain of machine translation. This groundbreaking system is capable of directly translating between any two of 100 languages, effectively bypassing the common reliance on English as an intermediary. It leverages a vast training dataset comprising 7.5 billion sentences spanning thousands of language pairs, which has been augmented through innovative data mining techniques. This model uniquely combines dense scaling with language-specific sparse parameters, efficiently managing its extensive linguistic diversity. Such capabilities set a new benchmark in facilitating direct, high-quality translations across a broad spectrum of languages, supporting more seamless and inclusive global communication.

In a similar vein, META AI's NLLB [17] focuses on overcoming the translation challenges faced by over 200 low-resource languages. The initiative began with exploratory interviews to understand the specific needs of these languages, which led to the development of specialized datasets and models aimed at bridging the resource gap. META AI developed a conditional compute model leveraging a Sparsely Gated Mixture of Experts, specifically trained with innovative data mining techniques to address the unique needs of low-resource languages. This model uses a conditional compute approach that activates only relevant parts of the network during translation tasks, enhancing computational efficiency and scalability. Additionally, the project implements architectural changes to prevent overfitting, such as data augmentation and self-supervised learning on large-scale monolingual corpora, thereby improving the model's performance across diverse linguistic contexts. These efforts underscore a broader shift towards more equitable language representation in machine translation, aiming to ensure that no language is left behind in the digital age. By combining these sophisticated techniques, the NLLB project is setting new standards in translation quality and inclusivity, making

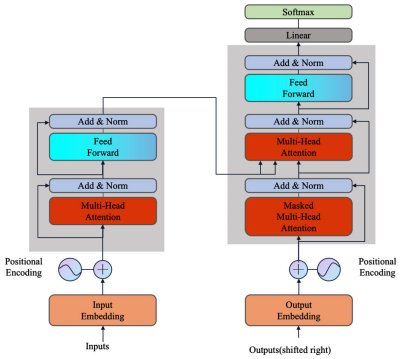

FIGURE 1. Structure of Transformer.

significant strides in reducing the digital divide and promoting linguistic diversity in the realm of global communication.

NLLB and M2M-100 models share several similarities in their architectural design. Both are based on the Transformer encoder-decoder architecture as shown in Figure 1. These architectural features enable the NLLB and M2M-100 models to handle a wide variety of language pairs efficiently, improving translation quality and efficiency [23].

Improvements to the specialized multilingual translation large-scale language models, such as the expert pruning strategy, have significantly reduced memory requirements during inference while maintaining language-specific translation quality [28]. The fine-tuning of the NLLB-200 model has also been shown to significantly enhance translation performance from Spanish to indigenous American languages, demonstrating the potential of fine-tuning to improve translation quality under low-resource conditions [29]. Furthermore, the fine-tuning of the Mistral 7B model for Spanish to English translation in the medical domain, incorporating zero-shot and one-shot prompts, demonstrated the effectiveness of adaptive machine translation [30]. The results indicated that the fine-tuned Mistral 7B achieved significant improvements in translation quality across both scenarios, even outperforming NLLB 3.3B and commercial models such as ChatGPT in certain cases. This emphasizes that fine-tuning open-source large language models can achieve high-quality translation while preserving control and privacy advantages. These findings illustrate that combining fine-tuning techniques with innovative model architectures can lead to superior translation performance in low-resource settings, providing valuable insights and support for the present study.

C. PEFT

Building on the advancements in language model training, PEFT [31] presents a method that integrates seamlessly with projects like NLLB. In PEFT, ''bottleneck'' adapters are inserted within the Transformer blocks of a model. These adapters consist of a series of down-projection, non-linearity, and up-projection steps, adjusting only a small subset of the model's parameters, specifically the weights in the down- and up-projections, while keeping the rest static. This approach

not only conserves computational resources but also allows for rapid adaptation to new languages or dialects, which is particularly beneficial for expanding the capabilities of models trained under initiatives like Meta AI's. The use of PEFT can enhance the adaptability of large-scale models to a diverse range of linguistic tasks without the need for extensive retraining, thereby supporting the goal of more inclusive language technologies. Although this method introduces new layers which can slow down inference times by increasing the model's effective size, it provides a balanced approach by adding only necessary parameters for the specific tasks.

D. LoRA

LoRA [32] represents a significant advancement in PEFT techniques for LLMs. LoRA focuses on fine-tuning a low-rank subspace of each weight matrix in a model, particularly within the linear modules of Transformers. Recent studies [33] have validated LoRA's effectiveness, making it a standard for instruction tuning in LLMs. Beyond its utility in NLP, where it has been extensively validated through various studies on selecting optimal rank values, effective transformer modules for insertion, and parameter distribution among weight matrices, LoRA's principles have also been adapted for vision tasks, enhancing the fine-tuning of vision transformers.

Empirical results demonstrate that LoRA can achieve comparable or superior performance to traditional fine-tuning with significantly fewer parameters. For instance, LoRA has shown promising results against baseline methods in tasks like machine translation into English [34], [35]. Moreover, it requires fewer training examples to reach competitive performance levels, highlighting its efficiency and the potential for reducing computational costs associated with the fine-tuning of large models.

E. PEFT METHODS IN LOW-RESOURCE LANGUAGE TRANSLATION

Low-resource language translation faces significant challenges due to the scarcity of bilingual data and the linguistic diversity of target languages. PEFT methods have been proven to be effective solutions, improving translation accuracy while enhancing computational efficiency [36]. Multilingual pretraining approaches, such as those employed in the NLLB project, have shown potential but also highlighted ongoing issues related to dataset quality and generalization [37]. Recent studies have further demonstrated the effectiveness of multilingual pretraining techniques in transferring knowledge from high-resource to low-resource languages and improving translation performance [38]. Incorporating cultural and language-specific nuances further enhances model performance, while advancements in PEFT methods, including ensemble approaches, demonstrate the potential to extend these techniques to diverse applications and resource-constrained environments [39].

III. LANGUAGE-SPECIFIC FINE-TUNING WITH LoRA

LSFTL leverages the LoRA to refine the translation capabilities of a base multilingual machine translation model for specific language pairs. This study addresses gaps in the literature by providing a detailed methodology for implementing LSFTL. Compared to other PEFT methods, such as LoRA's variants DoRA [40] and VeRA [41], which primarily focus on improving efficiency and general applicability to large language models, LSFTL specifically targets large machine translation models, not only enhancing efficiency but also optimizing translation quality for low-resource languages. By integrating LoRA modules as adapters into the foundational model, LSFTL focuses on the distinctive linguistic features of each language pair, enhancing the model's ability to deliver more accurate and contextually appropriate translations. This process is facilitated through the use of bilingual text datasets, known as bitexts, which are instrumental in training these adapters to handle the nuances and complexities inherent to each language pair.

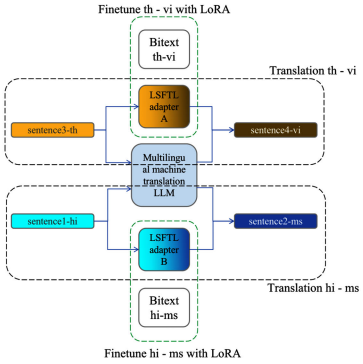

FIGURE 2. Illustration of Process of LSFTL.

As one example shown in Figure 2, LSFTL adapters are specifically fine-tuned for selected language pairs such as Thai-Vietnamese and Hindi-Malay, using their respective bitext datasets for optimal training. For instance, the LSFTL Adapter A is fine-tuned with the Bitext th-vi dataset, enabling it to proficiently translate from Thai to Vietnamese by modifying the translation model at specific layers where linguistic adaptations are most impactful. Similarly, LSFTL Adapter B utilizes the Bitext hi-ms dataset to specialize in translating from Hindi to Malay. These adaptations ensure that the translation outputs—Vietnamese and Malay sentences—are not only fluent but also retain the semantic integrity of the original sentences, showcasing the practical application and effectiveness of LSFTL in enhancing language-specific translation tasks.

Figure 3 showcases a multilingual language model enhanced with LSFTL, tailored to optimize translation processes by addressing language-specific nuances. The core of this system includes a shared encoder and decoder capable of processing inputs and outputs across multiple languages. This shared framework maintains the fundamental capabilities for

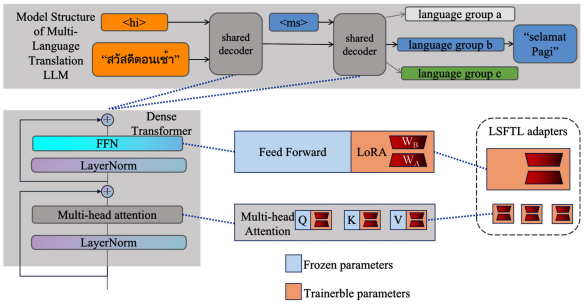

FIGURE 3. Model Structure of LSFTL Adapters.

translating across various language pairs and provides a flexible base for specialized enhancements.

In this architecture, the encoder transforms the source token sequence into a sequence of embeddings of the same length. The decoder then uses the encoder's output to autoregressively generate the target sentence, token by token. More precisely, the encoder takes the sequence of tokens $U = (u_1, \ldots, u_S)$ and the source language ℓ_s, producing a sequence of embeddings $H = (h_1, \ldots, h_S)$. This sequence is then fed into the decoder along with the target language ℓ_t to generate the target tokens $V = (v_1, \ldots, v_T)$ sequentially:

$$H = \text{encoder}(U, \ell_s), \tag{1}$$

$$\forall i \in [1, \ldots, T], v_{i+1} = \text{decoder}(H, \ell_t, v_1, \ldots, v_i). \tag{2}$$

Both the encoder and decoder are composed of stacked Transformer layers. Each Transformer layer takes a sequence of embeddings as input and outputs another sequence of embeddings. In the encoder, Transformer layers consist of two sub-layers: a self-attention layer and a feed-forward layer. These are applied sequentially, each preceded by Layer Normalization (LayerNorm) [42] and followed by a residual connection [43]:

$$Z = X + \text{self-attention}(\text{norm}(X)), \tag{3}$$

$$Y = Z + \text{feed-forward}(\text{norm}(Z)). \tag{4}$$

In the decoder, there is an additional third sub-layer between the self-attention and the feed-forward layers, which computes attention over the encoder's output.

LoRA modules, implemented as low-rank matrices, are integrated into the transformer layers, enhancing the model's adaptation to specific language pairs by refining attention mechanisms and linear transformations. These matrices, denoted by trainable parameters W_A and W_B, are specifically designed to adjust the model's output by refining the attention mechanisms and the linear transformations in the feed-forward layers for targeted language pairs. This adaptation enhances the model's ability to handle the linguistic specifics such as idiomatic expressions and unique syntactic structures, improving the translation quality significantly. This approach modifies the transformation as:

$$(\ell_t, v_1, \ldots, v_i) = W(u_1, \ldots, u_S, \ell_s) + \alpha r BA(u_1, \ldots, u_S, \ell_s), \tag{5}$$

where $(u_1, \ldots, u_S, \ell_s)$ and $(\ell_t, v_1, \ldots, v_i)$ are the input with source language tokens and labels and output vectors with target language tokens and labels respectively, W is the original weight matrix, A and B represent the low-rank matrices, r is the rank of the subspace, significantly smaller than the input or output dimensions (d_{in}, d_{out}), and α is a scaling hyperparameter.

Unlike traditional methods that integrate adapters in series with activations, LSFTL employs a parallel approach without non-linearity, allowing the modifications to be seamlessly integrated back into the main weight matrix W at inference time as:

$$W' = W + \alpha r BA. \tag{6}$$

This integration ensures that the inference speed remains equivalent to that of the unmodified base model, addressing one of the primary constraints of using PEFT methods in production environments.

The LSFTL approach offers substantial benefits in terms of efficiency and adaptability. By keeping most of the model's parameters frozen, LSFTL ensures the stability and generalizability of the pre-trained model are preserved, while the adaptable LoRA modules enable focused improvements on specific language pairs, especially those that are underrepresented or linguistically complex. This strategic modification conserves computational resources and reduces the necessity for extensive data retraining.

IV. EXPERIMENTAL

A. DATASETS

The dataset employed for evaluation and fine-tuning of low-resource languages—Hindi (hi), Malay (ms), Thai (th), and Vietnamese (vi)—includes two public datasets: MultiCCAligned [44] and OpenSubtitles [45]. Additionally, to mitigate the risk of overfitting by publicly available large multilingual models on these public datasets, data from these Asian low-resource languages has been collected from the OpenSubtitles website since 2018, termed NEWSubtitles, for comparative assessment. The selection of these specific low-resource languages was deliberate and based on several criteria: (1) they represent different language families (Indo-European, Austronesian, and Sino-Tibetan), providing diverse linguistic structures to test our approach; (2) they employ different writing systems and grammatical structures, allowing us to evaluate the versatility of LSFTL across varied linguistic characteristics; (3) they are regionally important languages with millions of speakers yet remain underrepresented in NLP research; and (4) they present unique translation challenges due to their morphological complexity, syntactic divergence from high-resource languages, and limited parallel corpora availability. Specific data volumes for all datasets are presented in Table 1.

B. MODELS AND BASELINES

Meta AI has developed a range of specialized multilingual translation models with varying parameter sizes. In our

TABLE 1. Dataset sentence counts for different language pairs.

Dataset / Language Pair	hi-ms	hi-th	hi-vi	ms-th	ms-vi	th-vi
MultiCCAligned	1,356,090	2,513,515	2,653,375	153,977	1,612,506	3,270,770
OpenSubtitles	27,906	25,057	32,991	406,935	851,694	672,817
NEWSubtitles	629,765	751,363	711,702	1,989,930	2,238,610	2,946,143

experiments, considering the capabilities and representativeness of our experimental machines, we selected the NLLB-200-Distilled-600M, M2M100-1.2B, and NLLB-200-1.3B for testing, as shown in Table 2.

TABLE 2. Overview of Meta AI's specialized multilingual translation models.

Model Type	Model Name	Parameters
M2M100	M2M100_418M	418M
M2M100	**M2M100_1.2B**	**1.2B**
M2M100	M2M100_12B	12B
NLLB-200	**NLLB-200-Distilled-600M**	**600M**
NLLB-200	**NLLB-200-1.3B**	**1.3B**
NLLB-200	NLLB-200-3.3B	3.3B
NLLB-200	NLLB-200-MoE-54.5B	54.5B

C. TRAINING SETUP AND HYPERPARAMETERS

Training was conducted using the NVIDIA GeForce RTX 4090 graphics card, a consumer-grade card equipped with 24 GB of GDDR6X VRAM.

The training configuration employed for our model is outlined as follows:

- **FP16**: Setting fp16=True enables mixed precision training, using 16-bit floating point numbers for most computations. This significantly reduces memory usage and increases computation speed on modern GPUs, allowing training larger models more efficiently without sacrificing much accuracy.
- **Gradient Accumulation Steps**: With gradient accumulation steps=4, the model accumulates gradients over four mini-batches before updating weights. This reduces memory usage per mini-batch, allowing for larger models or higher resolution data. It also results in smoother gradient updates, enhancing training stability.
- **Learning Rate and Weight Decay**: The learning rate was set to 1×10^{-4} with a weight decay of 0.01 to ensure steady training convergence.
- **Rank of LoRA**: Choosing r=16 to balance model performance, adaptability, memory usage, and computational efficiency.

D. EVALUATION METHODS

In line with the recommendations from the MT Metrics shared task [46], which advocates for the application of neural network-based metrics due to their strong correlation with human assessments and robustness against domain shifts, we primarily integrate the COMET method [47], a reference-based metric from the shared task 10, which leverages direct assessments, sentence-level scores, and word-level tags derived from Multidimensional Quality Metrics (MQM) error annotations. Additionally, our evaluation utilizes BLEU and chrF [48]. These metrics, recognized for their transparency and repeatability, form the backbone of our analysis.

Evaluating translation quality for low-resource languages presents unique challenges due to linguistic diversity and resource limitations. To address these challenges, our evaluation framework incorporates several specialized approaches: (1) leveraging COMET's ability to capture semantic and pragmatic aspects of translation quality, which is particularly crucial for low-resource languages where literal translations often miss cultural nuances; and (2) considering language-specific grammatical structures and cultural contexts during evaluation to ensure that assessment is sensitive to the unique characteristics of each language pair. This comprehensive evaluation strategy ensures that our assessment of LSFTL's performance accurately reflects its real-world utility for low-resource language translation.

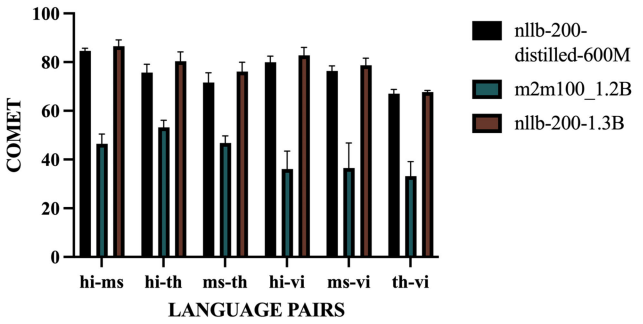

FIGURE 4. COMET for Language Pairs Across Different Models With and Without LSFTL.

V. RESULTS AND ANALYSIS

A. MULTILINGUAL TRANSLATION PERFORMANCE WITH ASIAN LOW-RESOURCE LANGUAGE

In Figure 4 and Table 3, we present one principal results of our experiments, detailing the impact of the LSFTL on translation performance across various models and language pairs. The most notable improvement was observed in the nllb-200-1.3B model for the Hindi-Malay (hi-ms) language pair, where the COMET increased from 84.79 to 88.36, marking an increase of approximately 3.57 percentage points—this is the most significant improvement across all datasets. Similarly, in the nllb-200-distilled-600M model, the application of LSFTL also demonstrated a significant enhancement, with the COMET increasing from 83.90 to 85.42, an increase of 1.52 percentage points. Notably, after the application of LSFTL, the performance of the nllb-200-distilled-600M model is now comparable to the performance of the nllb-200-1.3B model without LSFTL. The variations in benefits observed across different language pairs from LSFTL can be attributed to factors such as linguistic characteristics, dataset quality, and corpus diversity. Language pairs with structural similarities, such as Hindi-Malay, exhibit greater performance improvements as LSFTL effectively utilizes cross-lingual transfer learning.

TABLE 3. BLEU/chrF for language pairs across different models with and without LSFTL.

Lan / Model	facebook/nllb-200-distilled-600M		facebook/m2m100_1.2B		facebook/nllb-200-1.3B	
	Baseline	With LSFTL	Baseline	With LSFTL	Baseline	With LSFTL
hi-ms	37.78/60.23	**47.24/67.16**	4.26/10.89	**7.44/17.83**	41.96/63.56	**59.75/75.55**
hi-th	23.64/36.66	**32.99/47.58**	5.40/9.37	**4.10/8.17**	25.32/44.65	**39.21/56.89**
ms-th	19.38/36.16	**29.40/47.35**	4.40/7.97	**3.51/8.00**	21.11/42.48	**34.48/54.58**
hi-vi	34.70/49.89	**45.38/59.84**	0.45/4.55	**4.49/12.20**	40.30/54.92	**56.72/68.95**
ms-vi	37.04/51.20	**46.68/60.45**	0.21/4.03	**3.19/9.08**	40.36/54.23	**55.30/67.08**
th-vi	4.72/10.49	**5.36/11.53**	4.33/3.02	**2.65/5.48**	5.08/10.85	**5.22/11.46**

TABLE 4. COMET for different datasets across different models with and without LSFTL.

Datasets	facebook/nllb-200-distilled-600M		facebook/m2m100_1.2B		facebook/nllb-200-1.3B	
	Baseline	With LSFTL	Baseline	With LSFTL	Baseline	With LSFTL
MultiCC	83.4	**84.2**	52.3	**57.3**	84.2	**85.6**
OpenSubtitles	76.7	**77.8**	52.2	**55.5**	77.4	**78.9**
NEWSubtitles	76.4	**77.2**	47.8	**49.3**	77.2	**78.4**

The richness and diversity of the training data also play a critical role, with cleaner and more comprehensive corpora enabling LSFTL to capture language-specific nuances more effectively, leading to higher translation accuracy. In contrast, language pairs involving highly divergent languages or datasets that are sparse or noisy tend to show smaller improvements. These findings highlight the reliance of LSFTL's effectiveness on high-quality bilingual corpora and its ability to adapt to the unique linguistic and structural features of each language pair. Aligning fine-tuning strategies with the inherent characteristics of language pairs and their respective datasets is essential for maximizing translation performance.

Regarding model differences, both models exhibited improvements post-LSFTL application, but the nllb-200-1.3B model generally demonstrated greater increases than the nllb-200-distilled-600M model. This suggests that larger, more complex models like the 1.3B benefit more from LSFTL, likely due to their higher capacity for learning and adaptation. Specifically, the enhancements in translation quality for complex language pairs such as Hindi and Malay are more pronounced in the 1.3B model, showcasing a higher increase in COMET compared to the distilled model. Additionally, certain language pairs, such as Hindi-Malay, benefit more from LSFTL technology, possibly due to inherent linguistic differences and the quality and size of the available corpora. While LSFTL generally enhances performance, there are exceptions that indicate the need for adjustments in LSFTL parameters or the exploration of other optimization techniques, depending on the specific model capabilities and language characteristics.

Considering the performance of LSFTL across different datasets of the result in Table 4, the coverage and richness of training data vary to some extent. For instance, the MultiCCAligned dataset, with its broader coverage and diverse corpora, enables the model to learn a wide range of language patterns and structures during training. In contrast, the OpenSubtitles and NEWSubtitles datasets, while potentially advantageous in specific domains or contexts, do not have as extensive coverage as the MultiCCAligned dataset. Nonetheless, the application of LSFTL still yields noticeable improvements in these cases. The extent of enhancement is comparable across datasets, indicating that the effectiveness of LSFTL is not significantly constrained by the inherent characteristics of the datasets.

B. LAYER CONFIGURATION OF LSFTL

To better understand the impact of LSFTL, we will explore the specific effects of different modules on model performance. From the experimental results in Figure 5, it can be seen that different LoRA modules (`q_proj`, `v_proj`, `k_proj`, `fc1`, `fc2`, `out_proj`, `lm_head`, and `ALL`) influence the model performance in varying ways after fine-tuning. From Figure 6, it can be seen that different models have variations in parameter configurations: layers with more parameters also have higher proportions in different models, especially the `lm_head` layer, followed by the `fc2` and `out_proj` layers. The `v_proj` and `out_proj` modules stand out for their significant impact on the COMET, reaching their peak scores at 6000 training steps and maintaining steady performance thereafter. The `fc1` and `fc2` modules also contribute substantially, showing an upward trend and achieving consistent improvements across different training steps. However, the `lm_head` module exhibits relatively low COMET, indicating a limited influence on overall performance when fine-tuned individually.

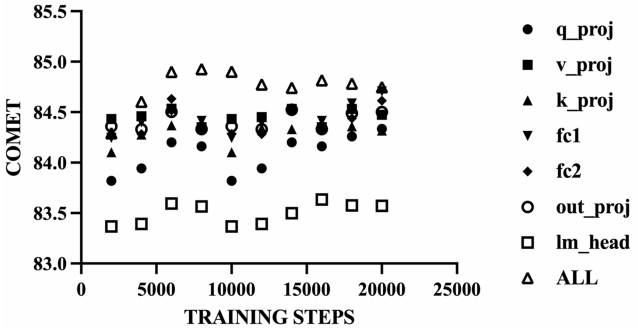

FIGURE 5. COMET for Different Target Modules.

TABLE 5. COMET for different layer configurations.

Layer Configuration	COMET Score
Original	84.2000
Decoder-layers	84.3802
Encoder-layers	85.9200

Overall, the gradual improvement trend across different training steps and the varying effects of the LoRA modules suggest that certain components have more immediate effects on the model's translation quality. From the experimental results in Table 5, we can observe that the `Original` configuration serves as a baseline indicating the performance of the model without specific adjustments to encoder or decoder layers. The `decoder-layers` configuration slightly improves the performance, showing the impact of fine-tuning decoder layers. In contrast, the `encoder-layers` configuration significantly

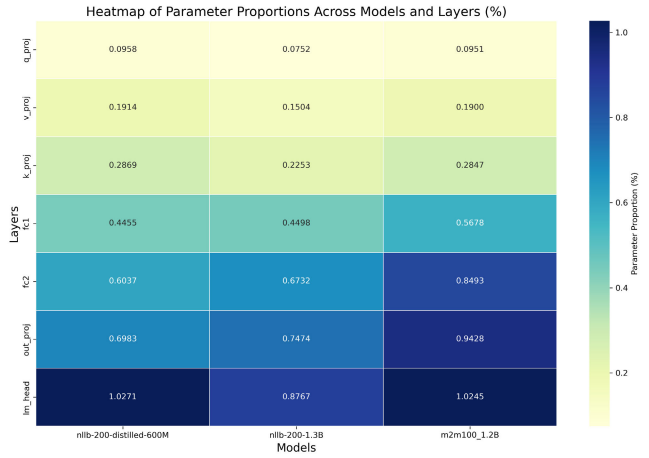

FIGURE 6. **Heatmap of Parameter Proportions Across Models and Layers (%).**

enhances the performance, indicating the substantial impact of fine-tuning encoder layers. Modules like `v_proj` and `out_proj` directly affect the model's performance due to their projection-layer role, which may more directly impact translation accuracy. In contrast, the `lm_head` module plays a more limited role in influencing the final translation output, which might explain its relatively low scores when fine-tuned independently. The combined fine-tuning of all modules (`ALL`) provides a more comprehensive improvement, suggesting the importance of balancing module contributions for optimal translation results.

C. COMPARING LSFTL WITH FINE-TUNE METHOD

The results, as illustrated in Figure 7, we can see that the experimental results using LSFTL and full-volume fine-tuning show diverse performances across various metrics: the graph indicates that LSFTL significantly outperforms full-volume tuning in COMET, especially in the early stages of training. This suggests that LSFTL has better initial adaptability in maintaining translation quality. The COMET focuses on semantic similarity, hence LSFTL's advantage may stem from its more effective use of context and optimized parameters for specific languages. This advantage likely benefits from the meticulous tuning of LSFTL for low-resource languages, allowing even subtle character variations to be effectively captured by the model. The evaluation loss graph shows that the loss values for both methods are very close, indicating similar learning efficiencies in model optimization. Although full-volume tuning involves adjusting more parameters, LSFTL also achieves effective loss reduction by precisely adjusting key parameters. The results demonstrate that LSFTL significantly excels in maintaining semantic and character-level translation quality, particularly showing superior performance in the early stages of training. This may be due to LSFTL's targeted fine-tuning of language-specific parameters, as opposed to the extensive parameter updates of full-volume tuning, enabling the model to adapt

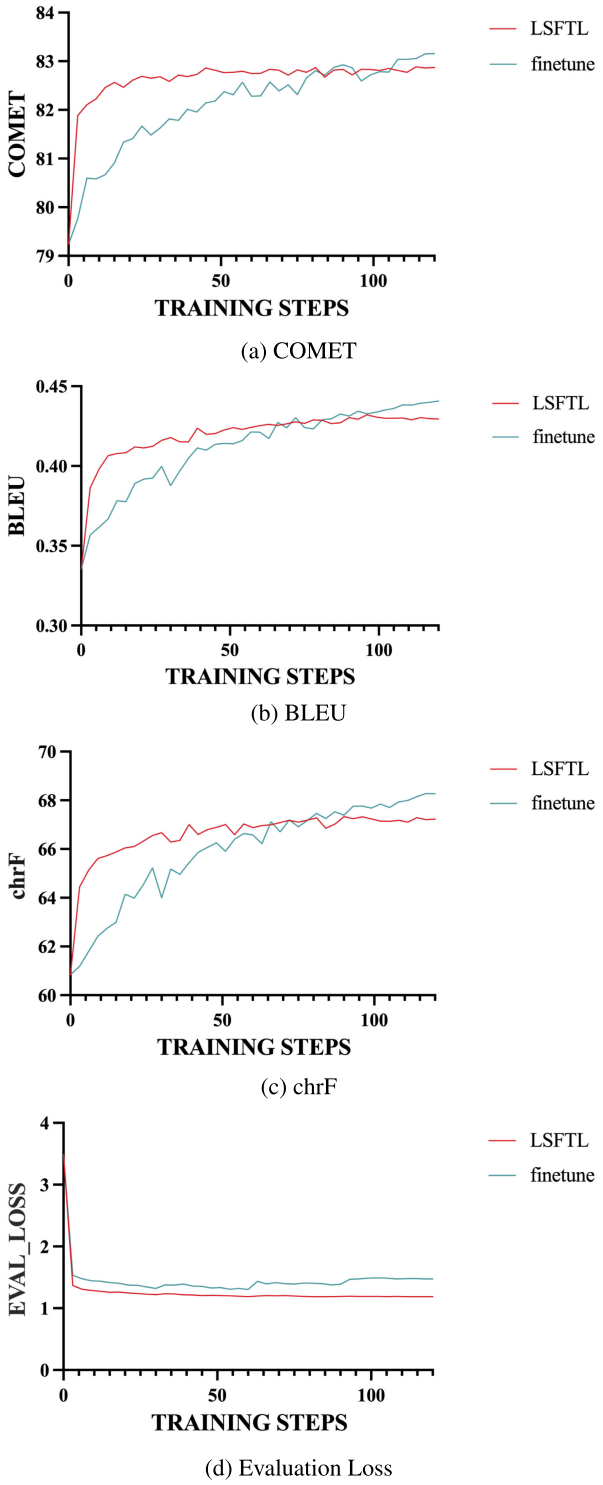

FIGURE 7. **Comparative Performance Analysis of LSFTL and Full-Volume Fine-Tuning on Multiple Metrics with hi-vi and nllb-200-distilled-600M model.**

more quickly to specific translation tasks. However, as training progresses, full-volume tuning exhibits strong potential for catching up, suggesting that with sufficient resources, prolonged training might achieve or exceed the performance of LSFTL.

TABLE 6. Memory allocation and COMET for different models and modes with 24GB VRAM.

Mode	Model	Memory Allocated (%)	COMET
Finetune	nllb-200-distilled-600M	75	84.7 + 2.3
LSFTL	nllb-200-distilled-600M	60	84.7 + 1.6
Finetune	nllb-200-1.3B	OUT OF MEMORY	0
LSFTL	nllb-200-1.3B	65	86.2 + 1.9

As we can see in Table 6, LSFTL performs exceptionally well across various models and hardware conditions, making it especially suitable for resource-constrained environments by improving memory utilization and maintaining high accuracy and efficiency in translation tasks. In summary, LSFTL demonstrates substantial advantages in memory efficiency and translation quality, particularly when dealing with large-scale language models, ensuring efficient operation under limited computational resources.

VI. CONCLUSION

In this paper, we have introduced Language-Specific Fine-Tuning with LoRA (LSFTL) to enhance the performance and efficiency of large language models (LLMs) for low-resource languages. LSFTL addresses the unique challenges associated with translating low-resource languages by fine-tuning specific parameters of LLMs, thereby optimizing their adaptability and accuracy without the need for extensive computational resources.

Our experiments demonstrate that LSFTL significantly improves translation quality across various metrics, particularly in the early stages of training, by leveraging targeted adjustments in model parameters. While full-volume fine-tuning achieves slightly superior results after extended training, LSFTL offers a clear advantage in terms of memory and computational efficiency. This approach allows smaller models to achieve competitive performance with faster convergence, rendering it highly effective for complex language pairs. Additionally, LSFTL's ability to be implemented on consumer-grade hardware positions it as a practical and scalable solution for real-world applications, especially in resource-constrained environments or scenarios requiring rapid deployment.

Furthermore, our analysis of different LoRA modules and layer configurations highlights the critical components that contribute most effectively to translation improvements. This insight provides a valuable framework for future research and optimization of LLMs tailored for low-resource languages. Future work could focus on expanding LSFTL's application to other low-resource language families and integrating additional linguistic features, such as morphology and syntax, to further enhance translation quality. Additionally, exploring the potential of LSFTL in domains like speech-to-text translation or low-resource dialogue systems could offer new directions for research.

In conclusion, LSFTL represents an advancement in the field of machine translation, offering a robust and efficient solution for enhancing the translation capabilities of LLMs in resource-constrained environments. By addressing both computational and linguistic challenges, LSFTL lays a foundation for inclusive and efficient multilingual machine translation systems, fostering the preservation and accessibility of linguistic and cultural diversity. LSFTL has the potential to revolutionize machine translation for low-resource languages by making advanced language models more accessible and efficient. In real-world scenarios, this technology can be applied to various fields such as education, healthcare, and public administration, where accurate translation is crucial for effective communication. The practical applications of LSFTL extend beyond just translation tasks, making it a versatile tool for various NLP applications tailored for low-resource languages. Moreover, PEFT techniques enhance the adaptability of large-scale models by allowing for efficient parameter adjustments without the need for extensive retraining, thereby extending their usability and effectiveness across diverse linguistic contexts.

Despite its promising results, LSFTL has several limitations that should be acknowledged. First, for extremely low-resource languages with very limited data (fewer than 10,000 parallel sentences), the approach may still struggle to capture language-specific nuances adequately. Second, domain-specific terminology and jargon present challenges, as the fine-tuning process may not have sufficient examples to learn specialized vocabulary effectively. Third, as the number of language pairs increases, the computational complexity of maintaining separate adapters for each pair grows, potentially limiting scalability for very large-scale multilingual systems. Fourth, highly imbalanced bilingual corpora may lead to uneven performance improvements across different language directions. Future research should address these limitations through techniques such as cross-lingual transfer learning, synthetic data generation, and more efficient parameter sharing across language families.

As our research on LSFTL is still ongoing with several aspects under active development, we plan to make the implementation code publicly available on GitHub in the future when the research reaches an appropriate stage. The repository will eventually include documentation, training scripts, and pretrained adapters to facilitate further advancements in this field.

REFERENCES

[1] OpenAI et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.
[2] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "LLaMA: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.
[3] Anthropic. (2023). *Introducing Claude*. Accessed: Mar. 30, 2023. [Online]. Available: https://www.anthropic.com/index/introducing-claude
[4] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. D. L. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7B," 2023, *arXiv:2310.06825*.
[5] S. Castilho, C. Mallon, R. Meister, and S. Yue, "Do online machine translation systems care for context? What about a GPT model?" 2023.
[6] M. Enis and M. Hopkins, "From LLM to NMT: Advancing low-resource machine translation with claude," 2024, *arXiv:2404.13813*.
[7] J. Zeng, F. Meng, Y. Yin, and J. Zhou, "Teaching large language models to translate with comparison," in *Proc. AAAI Conf. Artif. Intell.*, Mar. 2024, vol. 38, no. 17, pp. 19488–19496.

[8] Y. Intrator, M. Halfon, R. Goldenberg, R. Tsarfaty, M. Eyal, E. Rivlin, Y. Matias, and N. Aizenberg, "Breaking the language barrier: Can direct inference outperform pre-translation in multilingual LLM applications?" 2024, *arXiv:2403.04792*.

[9] H. Xu, A. Sharaf, Y. Chen, W. Tan, L. Shen, B. Van Durme, K. Murray, and Y. Jin Kim, "Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation," 2024, *arXiv:2401.08417*.

[10] S. Ranathunga, E.-S. A. Lee, M. P. Skenduli, R. Shekhar, M. Alam, and R. Kaur, "Neural machine translation for low-resource languages: A survey," *ACM Comput. Surveys*, vol. 55, no. 11, pp. 1–37, Nov. 2022.

[11] P. Koehn, *Neural Machine Translation*. Cambridge, U.K.: Cambridge Univ. Press, 2020.

[12] R. Sennrich, B. Haddow, and A. Birch, "Edinburgh neural machine translation systems for WMT 16," 2016, *arXiv:1606.02891*.

[13] B. Zoph, D. Yuret, J. May, and K. Knight, "Transfer learning for low-resource neural machine translation," 2016, *arXiv:1604.02201*.

[14] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 726–742, Dec. 2020.

[15] A. Magueresse, V. Carles, and E. Heetderks, "Low-resource languages: A review of past work and future challenges," 2020, *arXiv:2006.07264*.

[16] N. Goyal, C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzmán, and A. Fan, "The flores-101 evaluation benchmark for low-resource and multilingual machine translation," *Trans. Assoc. Comput. Linguistics*, vol. 10, pp. 522–538, May 2022.

[17] N. Team et al., "No language left behind: Scaling human-centered machine translation," 2022, *arXiv:2207.04672*.

[18] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, E. Grave, M. Auli, and A. Joulin, "Beyond english-centric multilingual machine translation," 2020, *arXiv:2010.11125*.

[19] Y. Jin Kim, A. Ahmad Awan, A. Muzio, A. Felipe Cruz Salinas, L. Lu, A. Hendy, S. Rajbhandari, Y. He, and H. Hassan Awadalla, "Scalable and efficient MoE training for multitask multilingual models," 2021, *arXiv:2109.10465*.

[20] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1–25. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[21] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang, "GPT understands, too," *AI Open*, vol. 5, pp. 208–215, Jan. 2024.

[22] M. Zhang and J. Li, "A commentary of GPT-3 in MIT technology review 2021," *Fundam. Res.*, vol. 1, no. 6, pp. 831–833, Nov. 2021.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jun. 2017, pp. 5998–6008.

[24] G. Yenduri, M. Ramalingam, G. C. Selvi, Y. Supriya, G. Srivastava, P. K. R. Maddikunta, G. D. Raj, R. H. Jhaveri, B. Prabadevi, W. Wang, A. V. Vasilakos, and T. R. Gadekallu, "GPT (Generative pre-trained transformer)—A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions," *IEEE Access*, vol. 12, pp. 54608–54649, 2024.

[25] K. Ahuja, H. Diddee, R. Hada, M. Ochieng, K. Ramesh, P. Jain, A. Nambi, T. Ganu, S. Segal, M. Axmed, K. Bali, and S. Sitaram, "MEGA: Multilingual evaluation of generative AI," 2023, *arXiv:2303.12528*.

[26] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean, "Google's multilingual neural machine translation system: Enabling zero-shot translation," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 339–351, Oct. 2017. [Online]. Available: https://www.aclweb.org/anthology/2020.tacl-1.1

[27] L. Barrault, O. Bojar, M. R. Costa-Jussa, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, P. Koehn, S. Malmasi, C. Monz, M. Müller, S. Pal, M. Post, and M. Zampieri, "Findings of the 2019 conference on machine translation," in *Proc. 4th Conf. Mach. Transl.*, 2019, pp. 1–62. [Online]. Available: https://www.aclweb.org/anthology/2019.wmt-1.1

[28] Y. Koishekenov, A. Berard, and V. Nikoulina, "Memory-efficient NLLB-200: Language-specific expert pruning of a massively multilingual machine translation model," 2022, *arXiv:2212.09811*.

[29] D. Degenaro and T. Lupicki, "Experiments in mamba sequence modeling and NLLB-200 fine-tuning for low resource multilingual machine translation," in *Proc. 4th Workshop Natural Language Process. Indigenous Languages Americas*, 2024, pp. 188–194.

[30] Y. Moslem, R. Haque, and A. Way, "Fine-tuning large language models for adaptive machine translation," 2023, *arXiv:2312.12740*.

[31] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. D. Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for NLP," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2019, pp. 2790–2799.

[32] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," 2021, *arXiv:2106.09685*.

[33] *Alpaca-Lora*. Accessed: 2023. [Online]. Available: https://github.com/tloen/alpaca-lora

[34] N. Arivazhagan, A. Bapna, O. Firat, D. Lepikhin, M. Johnson, M. Krikun, M. Xu Chen, Y. Cao, G. Foster, C. Cherry, W. Macherey, Z. Chen, and Y. Wu, "Massively multilingual neural machine translation in the wild: Findings and challenges," 2019, *arXiv:1907.05019*.

[35] D. Vilar, M. Freitag, C. Cherry, J. Luo, V. Ratnakar, and G. Foster, "Prompting palm for translation: Assessing strategies and performance," 2022, *arXiv:2211.09102*.

[36] T. Su, X. Peng, S. Thillainathan, D. Guzmán, S. Ranathunga, and E.-S. A. Lee, "Unlocking parameter-efficient fine-tuning for low-resource language translation," 2024, *arXiv:2404.04212*.

[37] O. Khade, S. Jagdale, A. Phaltankar, G. Takalikar, and R. Joshi, "Challenges in adapting multilingual LLMs to low-resource languages using LoRA PEFT tuning," 2024, *arXiv:2411.18571*.

[38] Y. Moslem, R. Haque, J. D. Kelleher, and A. Way, "Adaptive machine translation with large language models," 2023, *arXiv:2301.13294*.

[39] C. Y. Kwok, S. Li, J. Q. Yip, and E. S. Chng, "Low-resource language adaptation with ensemble of PEFT approaches," in *Proc. Asia Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, 2024, pp. 1–6.

[40] S.-Y. Liu, C.-Y. Wang, H. Yin, P. Molchanov, Y.-C. F. Wang, K.-T. Cheng, and M.-H. Chen, "Dora: Weight-decomposed low-rank adaptation," 2024, *arXiv:2402.09353*.

[41] D. J. Kopiczko, T. Blankevoort, and Y. M. Asano, "VeRA: Vector-based random matrix adaptation," 2023, *arXiv:2310.11454*.

[42] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[44] A. El-Kishky, V. Chaudhary, F. Guzman, and P. Koehn, "CCAligned: A massive collection of cross-lingual Web-document pairs," 2019, *arXiv:1911.06154*.

[45] P. Lison and J. Tiedemann, "Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles," 2016.

[46] M. Freitag, R. Rei, N. Mathur, C.-k. Lo, C. Stewart, E. Avramidis, T. Kocmi, G. Foster, A. Lavie, and A. F. Martins, "Results of wmt22 metrics shared task: Stop using bleu-neural metrics are better and more robust," in *Proc. 7th Conf. Mach. Transl.*, 2022, pp. 46–68.

[47] R. Rei, J. G. De Souza, D. Alves, C. Zerva, A. C. Farinha, T. Glushkova, A. Lavie, L. Coheur, and A. F. Martins, "Comet-22: Unbabel-ist 2022 submission for the metrics shared task," in *Proc. 7th Conf. Mach. Transl.*, 2022, pp. 578–585.

[48] M. Popović, "ChrF: Character n-gram F-score for automatic MT evaluation," in *Proc. 10th Workshop Stat. Mach. Transl.*, 2015, pp. 392–395.

XIAO LIANG received the B.S. degree in electronic information engineering from Xidian University, Xi'an, China, in 2012, and the M.S. degree in information technology and electrical engineering from Gottfried Wilhelm Leibniz Universität Hannover, Germany, in 2016. He is currently pursuing the Ph.D. degree in computer science with the Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman (UTAR), Malaysia, with a focus on artificial intelligence, natural language processing, and machine translation.

YEN-MIN JASMINA KHAW received the Ph.D. degree from the School of Computer Sciences, Universiti Sains Malaysia (USM), in 2017. She is currently an Assistant Professor with Universiti Tunku Abdul Rahman. Her research interest include natural language processing, such as speech synthesis, speech recognition, and machine translation.

SOUNG-YUE LIEW received the bachelor's degree in electrical engineering from National Taiwan University, Taiwan, in 1993, and the M.Phil. and Ph.D. degrees in information engineering from The Chinese University of Hong Kong (CUHK), in 1996 and 1999, respectively. After completing his Ph.D., he joined the Department of Information Engineering, CUHK, as an Assistant Professor, from 1999 to 2000, and as a Research Associate, from 2001 to 2002. From 2002 to 2003, he was a Research Associate with Polytechnic University, New York (now New York University Tandon School of Engineering). In August 2003, he joined Universiti Tunku Abdul Rahman, Malaysia, where he is currently a Professor and the Dean of the Faculty of Information and Communication Technology. His research interests include data analytics, algorithm design, system performance analysis, the IoT, and next-generation mobile networks.

TIEN-PING TAN received the Ph.D. degree from Université Joseph Fourier, France, in 2008. He is currently an Associate Professor with the School of Computer Sciences, Universiti Sains Malaysia. His research interests include automatic speech recognition, machine translation, and natural language processing.

DONGHONG QIN (Member, IEEE) received the B.Eng., M.Eng., Ph.D. degrees. He is currently a Professor and the Vice Dean of the School of Artificial Intelligence, Guangxi Minzu University. Previously, he was with UMASS Amherst, Tsinghua University, and Guilin University of Electronic Technology. His current research interests include machine learning and intelligent speech processing. He serves as a member for the Distributed Computing Systems Professional Committee and the Natural Language Processing Professional Committee of China Computer Federation (CCF).