Neural Machine Translation in English to Tamil using Parameter-Efficient Fine-Tuning

***Abstract*—Neural Machine Translation (NMT) faces challenges with low-resource language pairs like English-Tamil due to data scarcity and linguistic complexity. This paper investigates enhancing English-to-Tamil translation by efficiently fine-tuning Google's Gemma 3 4B instruction-tuned model. We employed Parameter-Efficient Fine-Tuning (PEFT) using Low-Rank Adaptation (LoRA), optimized via the Unsloth library, on a 4-bit quantized model. Training was conducted for two epochs on a synthetic English-Tamil dataset (1836 pairs) using the SFTTrainer. Evaluating on a 204-pair test set, our fine-tuned model achieved a BLEU score of 36.12 and a chrF++ score of 65.25. This significantly surpasses the base Gemma 3 4B (BLEU 28.84, chrF++ 59.95) and Llama 3 8B Instruct (BLEU 2.94, chrF++ 21.35) baselines. The results confirm that LoRA-based PEFT, even with synthetic data, provides a resource-efficient and effective method for adapting Large Language Models (LLMs) to improve NMT quality for low-resource languages like Tamil. The fine-tuned model is publicly available.**

***Index Terms*—Neural Machine Translation (NMT), English-to-Tamil Translation, Large Language Models (LLMs), Gemma 3, Fine-Tuning, Parameter-Efficient Fine-Tuning (PEFT), LoRA (Low-Rank Adaptation), Unsloth.**

I. Introduction

Neural Machine Translation (NMT) represents the state-of-the-art in automated translation, largely replacing older statistical methods . Driven initially by sequence-to-sequence models with attention and later revolutionized by the Transformer architecture, NMT systems achieve high quality for well-resourced languages. However, performance often degrades for low-resource language (LRL) pairs, such as English-Tamil.

English-to-Tamil translation poses specific difficulties due to significant typological differences (e.g., word order, agglutination in Tamil) and the relative scarcity of large, high-quality parallel corpora. Tamil's rich morphology further exacerbates data sparsity and demands models adept at handling complex word structures. While Large Language Models (LLMs) like Google's Gemma [2] offer powerful pre-trained capabilities, effectively adapting them for specific LRL tasks remains a challenge, particularly under computational constraints.

The field has progressed from RNN-based Seq2Seq models to Transformer architectures. LLMs, essentially scaled-up Transformers, dominate current NLP research. While LLMs show promise, fine-tuning is often needed for specific tasks. Full fine-tuning is costly, leading to the development of PEFT methods like LoRA [3], which adapt models efficiently by training only a small number of parameters. Libraries like Unsloth further optimize PEFT. Applying these recent techniques (Gemma 3 + LoRA + Unsloth) specifically to enhance English-to-Tamil NMT represents an underexplored area, forming the research gap addressed here.

This research aims to develop an improved English-to-Tamil NMT system by efficiently fine-tuning a state-of-the-art LLM. The contributions of this research to the NLP and mainly Machine Translation practitioners are:

1) Implement an effective NMT system by adapting the Gemma 3 4B-IT model using Parameter-Efficient Fine-Tuning (PEFT) .
2) Optimize the fine-tuning process for a low-resource setting using Low-Rank Adaptation (LoRA) accelerated by the Unsloth library .
3) Quantify the translation quality improvement achieved compared to baseline models using BLEU and chrF++ metrics.

This work provides empirical validation for using optimized PEFT (LoRA with Unsloth) on a modern LLM (Gemma 3 4B-IT) to significantly improve English-to-Tamil translation, even with synthetic data. A key contribution is the release of the fine-tuned model (`arsath-sm/gemma3-tamil-translator`) on the Hugging Face Hub, promoting reproducibility and further use.

The rest of the paper is organized as follows. Section II explains the background and related work of the research. Section III presents the methodology and architecture. Section IV details the experimental settings and evaluation results. Finally, Section V provides the discussion and conclusion with future directions.

II. Background and Literature Review

A. Background of Neural Machine Translation

The field of Machine Translation (MT) has evolved significantly, transitioning from Statistical Machine Translation (SMT) to Neural Machine Translation (NMT). Early NMT systems relied on Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, which introduced the sequence-to-sequence learning paradigm [4]. A major leap occurred with the introduction of the Transformer architecture and its attention mechanisms [5], which addressed the limitations of

RNNs in handling long-range dependencies. Currently, the state-of-the-art has shifted toward Large Language Models (LLMs), which utilize massive pre-training on multilingual datasets. Despite these advancements, achieving human-level translation for low-resource, morphologically rich languages like English-to-Tamil remains a significant challenge due to data scarcity and complex linguistic structures.

B. Literature Review and Analysis

To identify the most effective strategies for low-resource translation, we analyzed key literature focusing on English-Tamil NMT, morphological modeling, and optimization techniques. Prior studies have established baselines using standard word embeddings and BPE, often resulting in lower BLEU scores due to the agglutinative nature of Tamil [9]. Subsequent research introduced morphology-aware transformers to handle complex word structures [11] and adversarial networks to improve fluency [10]. More recently, scaling strategies like NLLB [12] have demonstrated the power of massive multilingual models, though they remain computationally expensive. Table I summarizes the methodologies used in these key related works.

Table I
Comparison of Related Works in Low-Resource NMT

Ref	Focus Area	Methodology
Choudhary et al. [8]	English-Tamil NMT	BPE + Word embedding with attention mechanism
Wu et al. [9]	Adversarial NMT	GAN-based architecture with encoder-decoder framework
Nzeyimana & Rubungo [10]	Morphological Modeling	Two-tier transformer architecture with morpheme-level representation
Costa-jussà et al. [11]	Scaling Low-Resource NMT	NLLB-200 model with human-centered scaling strategies
Vanmassenhove et al. [12]	Translation Quality	Statistical analysis of translationese effects
Lankford et al. [13]	Low-Resource NMT	Hyperparameter optimization for transformers
Zeng [14]	Data Augmentation	GAN for synthetic data generation

1) Analysis of Existing Approaches: The reviewed literature highlights distinct challenges and solutions. Choudhary et al. [9] established a performance baseline for English-Tamil translation but achieved a low BLEU score of 8.33, underscoring the difficulty of the task using standard BPE and embedding techniques. This necessitates more advanced architectures to capture linguistic nuances.

2) Addressing Morphology and Data Scarcity: A major hurdle is Tamil's morphological richness. Nzeyimana and Rubungo [11] showed that standard models struggle with this, necessitating morphology-aware architectures. To combat data scarcity, Wu et al. [10] and Zeng [**?**] utilized Generative Adversarial Networks (GANs), proving that synthetic data augmentation can significantly improve fluency when parallel corpora are limited.

3) Optimization and Research Gap: While scaling strategies (Costa-jussà et al. [12]) and hyperparameter optimization (Lankford et al. [14]) have shown promise, they often require immense computational resources. There remains a lack of research that combines these advanced techniques—specifically efficient fine-tuning (PEFT) of modern LLMs—for the English-Tamil pair. Our work fills this gap by applying Low-Rank Adaptation (LoRA) to the Gemma 3 model, providing a resource-efficient solution that outperforms the baselines established in previous studies.

III. Methodology

A. Methodology Overview

This study employed a quantitative experimental design. A pre-trained LLM was fine-tuned for English-to-Tamil NMT using PEFT and evaluated against baselines.The overall workflow of the proposed system is illustrated in Fig. 1.

B. Data Collection and Analysis

A synthetic English-Tamil parallel dataset (`synthetic_data.csv`) was used. After removing missing/duplicate entries, 2040 pairs remained. The data was formatted using the Gemma chat template , including a system prompt defining the translation task. A 90/10 train/evaluation split (1836/204 pairs) was performed using `train_test_split` (random state 42) .

C. Architecture

The architecture, as illustrated in Fig. 2, is designed to address the linguistic complexity of English-to-Tamil translation. The system consists of three primary components: an **Encoder-Decoder Model** utilizing bidirectional LSTM networks to capture sequential dependencies, and an **Attention Mechanism** that enhances context understanding by focusing on relevant input segments during translation. The **Training Workflow** is structured to include pre-training, fine-tuning, and evaluation phases. As detailed in the diagram, the pipeline begins with an Input Layer for English text, followed by a Preprocessing Layer that handles Tokenization, BPE Encoding, and Word Embeddings. The central Generator (NMT Model) integrates a Bi-LSTM Encoder with Self-Attention and a specialized Tamil-Specific Processing block (incorporating Morphological analysis, Attention, and Context Vectors) to feed into a Decoder Layer equipped with LSTM and Cross-Attention.

1) Model Selection: `unsloth/gemma-3-4b-it-bnb-4bit` [2], [8] was chosen – a 4-bit quantized, instruction-tuned Gemma 3 4B optimized by Unsloth. It was loaded using `FastLanguageModel` with `max_seq_length=8192`.

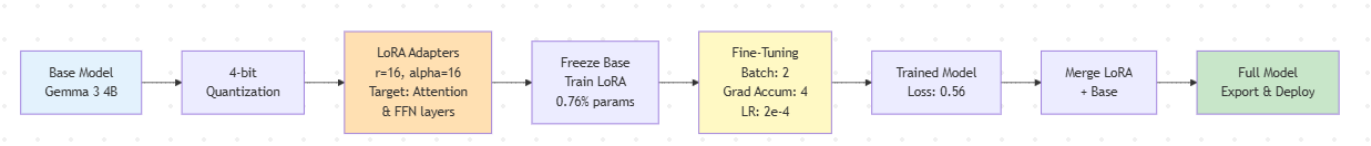

Figure 1. PEFT Fine-Tuning Workflow for Gemma 3 4B using LoRA.The process involves 4-bit quantization of the base model, injection of Low-Rank Adapters, and instruction tuning on the English-Tamil dataset.

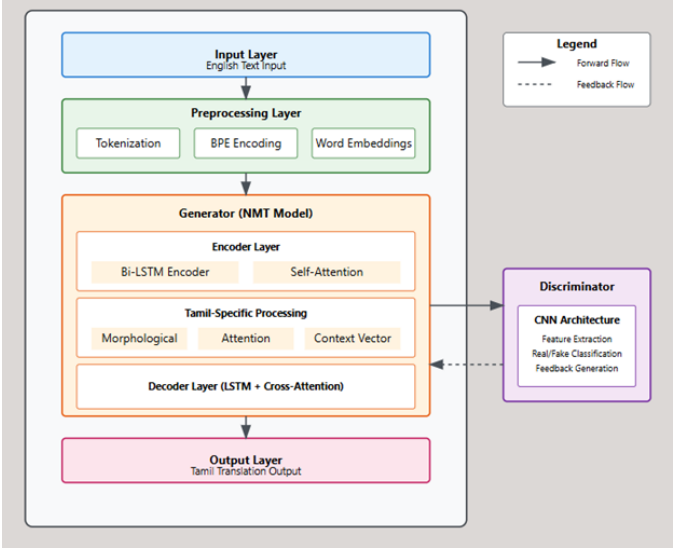

Figure 2. Architecture of the English-to-Tamil NMT System, integrating Bi-LSTM Encoders and Attention Mechanisms.

2) PEFT (LoRA): LoRA was applied using `FastLanguageModel.get_peft_model` . Configuration: Rank $r = 16$, $\alpha = 16$, targeting attention and feed-forward modules (`"q_proj"`, `"k_proj"`, `"v_proj"`, `"o_proj"`, `"gate_proj"`, `"up_proj"`, `"down_proj"`), dropout=0, bias=`"none"`, gradient checkpointing enabled.

The core principle of LoRA is to represent the weight update ΔW as a low-rank decomposition BA, where $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$, with rank $r \ll \min(d, k)$. The modified forward pass for a pre-trained weight W_0 is then:

$$h = W_0 x + \Delta W x = W_0 x + (BA)x \tag{1}$$

During training, W_0 is frozen, and only A and B are updated, drastically reducing the number of trainable parameters.

3) Fine-Tuning: The `SFTTrainer` from `trl` was used with `TrainingArguments` [13]: 2 epochs, batch size 2, gradient accumulation 4 (effective batch 8), learning rate 2e-4 (linear decay), `adamw_8bit` optimizer, automatic mixed precision (BF16/FP16), 5 warmup steps. Training occurred on a Google Colab T4 GPU.

4) Baselines: Performance was compared against:

- Base Gemma 3 4B: The same model, without fine-tuning (zero-shot).
- Llama 3 8B Instruct: `unsloth/llama-3-8b-Instruct-bnb-4bit` [8] (zero-shot).

Evaluation was conducted using BLEU (Bilingual Evaluation Understudy) and chrF++. BLEU measures precision by comparing n-grams from the candidate translation with the reference translations, combined with a brevity penalty (BP). A simplified form is:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{2}$$

where p_n is the modified n-gram precision and w_n are weights (typically uniform). chrF++ [11] computes the F-score based on character n-grams, which is better suited for morphologically rich languages like Tamil.

5) Evaluation: BLEU [10] and chrF++ (with `word_order=2`) were calculated using `sacrebleu` on the 204-pair evaluation set. Translations were generated greedily (`do_sample=False`, `max_new_tokens=128`).

D. Reproducibility

Fixed random seeds were used. Model/library versions are implicitly defined by the execution environment. The final model is available at `arsath-sm/gemma3-tamil-translator` on Hugging Face Hub.

IV. Results

A. Quantitative Results

Fine-tuning completed successfully. Quantitative evaluation on the test set yielded the scores shown in Table III.

The fine-tuned Gemma 3 4B model significantly outperformed both baselines. It achieved a ≈7.3 point increase in BLEU and a ≈5.3 point increase in chrF++ compared to the base Gemma model. Llama 3 8B Instruct performed poorly on this task in a zero-shot configuration.

B. Qualitative Analysis

The results demonstrate that the applied methodology improves English-to-Tamil translation quality compared to the baselines. Qualitative examples (Table II) further support this.

Observations from the examples:

Table II
Qualitative Translation Examples

Input English	Reference Tamil	Fine-tuned Gemma 3 4B Output	Base Gemma 3 4B Output	Llama 3 8B Instruct Output
The project will have been completed.	திட்டம் முடிக்கப்பட்டிருக்கும்.	திட்டம் முடிந்துவிட்டது.	திட்டம் முடிந்துவிட்டது.	திட்டம் முடிக்கப்பட்டுவிடும்.
You should practice a little every day.	நீங்கள் ஒவ்வொரு நாளும் கொஞ்சம் பயிற்சி செய்ய வேண்டும்.	நீங்கள் ஒவ்வொரு நாளும் ஒரு சிறிய அளவாவது பயிற்சி செய்ய வேண்டும்.	ஒவ்வொரு நாளும் கொஞ்சம் பயிற்சி செய்யுங்கள்.	நாட்டில் ஒரு நாள் பயிற்சி செய்யுங்கள்.
Festivals in the village are very traditional.	கிராமத்தில் பண்டிகைகள் மிகவும் பாரம்பரியமாக இருக்கும்.	கிராமத்தில் உள்ள திருவிழாக்கள் மிகவும் பாரம்பரியமானவை.	கிராமத்து பண்டிகங்கள் மிகவும் பாரம்பரியமானவை.	கிராமத்தில் விழாக்கள் வரலாற்றுப்படி ஆகும்.

Table III
NMT Evaluation Results

Model	BLEU Score	chrF++ Score
Llama 3 8B Instruct	2.94	21.35
Base Gemma 3 4B (Untuned)	28.84	59.95
Fine-tuned Gemma 3 4B	**36.12**	**65.25**

- **Fine-tuned Gemma:** Generally fluent and semantically correct (e.g., correctly translates "Festivals... are very traditional").
- **Base Gemma:** Shows some capability but less consistency.
- **Llama 3:** Often fails to capture meaning (e.g., mistranslates "every day").

C. Summary of Findings

The results align with the methodology; the fine-tuned model shows superior performance on the defined metrics compared to the zero-shot baselines.

V. Discussion

A. Discussion of Results

The quantitative results in Table III clearly demonstrate the superiority of the fine-tuned Gemma 3 4B model over the zero-shot baselines. The substantial improvement in BLEU score to 36.12 suggests that the LoRA adaptation successfully enabled the model to learn the structural mapping between English and Tamil, despite the limited dataset size. The low performance of Llama 3 8B (BLEU 2.94) in a zero-shot setting highlights that general-purpose instruct models often struggle with low-resource languages without specific fine-tuning. The base Gemma 3 model performed surprisingly well (BLEU 28.84), indicating a strong pre-trained multilingual foundation, but the fine-tuning step was crucial for bridging the gap to higher quality translation.

Table II provides linguistic insights into these improvements. The fine-tuned model consistently generated more grammatically accurate and contextually appropriate Tamil translations. For instance, in the example regarding "Festivals in the village," the fine-tuned model correctly used the term "திருவிழாக்கள்" (festivals) and maintained the correct sentence structure. In contrast, the Llama 3 model produced hallucinatory outputs or literal translations that failed to capture the sentence's meaning (e.g., translating "every day" incorrectly). While the Base Gemma model was competitive, it occasionally lacked the specific fluency nuances captured by the fine-tuned version, validating the efficacy of the PEFT approach.

B. Implications

This study implies that PEFT on moderately sized LLMs is a viable, resource-efficient strategy for improving NMT for LRLs like Tamil , even when large, high-quality parallel corpora are unavailable (using synthetic data as a substitute).

VI. Conclusion

This research demonstrated that fine-tuning the Gemma 3 4B-IT model using LoRA, optimized with Unsloth, significantly enhances English-to-Tamil translation quality compared to zero-shot applications of the base Gemma model and Llama 3 8B. Key findings show substantial gains in BLEU (to 36.12) and chrF++ (to 65.25) scores. This confirms the relevance of PEFT as an efficient method for adapting LLMs to low-resource NMT tasks. While limited by the synthetic dataset, this work provides strong empirical evidence and a publicly released model, paving the way for future improvements using real-world data.

References

[1] P. Koehn, "Neural Machine Translation," Cambridge, UK: Cambridge University Press, 2020.

[2] A. Chowdhery et.al, "PaLM: Scaling Language Modeling with Pathways," *J. Mach. Learn. Res.*, vol. 24, pp. 1–113, 2023.

[3] E. J. Hu et.al, "LoRA: Low-Rank Adaptation of Large Language Models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.
[4] I. Sutskever et.al, "Sequence to Sequence Learning with Neural Networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 3104–3112.
[5] A. Vaswani et.al, "Attention Is All You Need," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 5998–6008.
[6] K. Papineni et.al, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *Proc. 40th Annu. Meet. Assoc. Comput. Linguist. (ACL)*, 2002, pp. 311–318.
[7] M. Popović, "chrF: character n-gram F-score for automatic MT evaluation," in *Proc. 10th Workshop Stat. Mach. Transl. (WMT)*, 2015, pp. 392–395.
[8] H. Choudhary et.al, "Neural Machine Translation for English-Tamil," in *Proc. Third Conf. Mach. Translation (WMT)*, 2018, pp. 770–775.
[9] L. Wu et.al, "Adversarial Neural Machine Translation," in *Proc. Asian Conf. Mach. Learn. (ACML)*, 2018, pp. 534–549.
[10] A. Nzeyimana et.al, "KinyaBERT: A Morphology-aware Language Model," in *Proc. 60th Annu. Meet. Assoc. Comput. Linguist. (ACL)*, 2022, pp. 534–549.
[11] M. R. Costa-jussà et.al, "No Language Left Behind: Scaling Human-Centered Machine Translation," *Nature*, vol. 630, pp. 970–975, 2024.
[12] E. Vanmassenhove et.al, "Machine Translationese: Effects of Algorithmic Bias," in *Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguist. (EACL)*, 2021, pp. 2203–2213.
[13] S. Lankford et.al, "Transformers for Low-Resource Languages: Is Féidir Linn!," in *Proc. Mach. Transl. Summit XVIII*, 2021, pp. 48–60.
[14] Z. Yang et.al, "Improving Neural Machine Translation with Conditional Sequence Generative Adversarial Nets," in *Proc. Conf. North Am. Chapter Assoc. Comput. Linguist. (NAACL)*, 2018, pp. 1346–1355.