Beyond Full Fine-Tuning: Parameter-Efficient Adaptation for Large-Scale Natural Language Generation

Emily Jimin Roh, Soohyun Park, and Joongheon Kim
Department of Electrical and Computer Engineering, Korea University, Seoul, Republic of Korea
E-mails: emilyjroh@korea.ac.kr, soohyun.park@sookmyung.ac.kr, joongheon@korea.ac.kr

***Abstract*—Large language models (LLMs) have achieved remarkable performance in natural language generation (NLG), powering applications such as summarization, dialogue systems, and instruction following. However, the prohibitive cost of full model fine-tuning limits their applicability in multi-task and resource-constrained settings. To address this, parameter-efficient fine-tuning (PEFT) techniques have been proposed to reduce the number of trainable parameters while maintaining competitive performance. This paper provides an introduction and comparative study of four representative fine-tuning strategies: full fine-tuning, low-rank adaptation (LoRA), Sparse LoRA (SoRA) and prefix tuning. We evaluate these approaches on standard NLG tasks, which includes summarization and question answering, using BLEU, BERTScore, and ROUGE as evaluation metrics. Experimental results demonstrate that PEFT methods can achieve near full fine-tuning performance while significantly reducing parameter overhead. Our findings highlight the trade-offs between efficiency and generation quality, offering practical guidance for deploying LLMs in real-world NLG scenarios.**

***Index Terms*—Large Language Models, Parameter-Efficient Fine-Tuning, Natural Language Generation Task**

I. Introduction

Large language models (LLMs) have revolutionized natural language generation (NLG), which enables high-quality text generation in tasks such as summarization, dialogue, and instruction following. Despite their effectiveness, the growing scale of LLMs poses a critical challenge: full model fine-tuning requires updating hundreds of millions or even billions of parameters, that makes adaptation computationally expensive and storage-intensive [1]. This limitation severely hinders their deployment in multi-task and resource-constrained environments. To address these challenges, parameter-efficient fine-tuning (PEFT) techniques have been proposed, which aims to reduce the number of trainable parameters while maintaining competitive generation quality. Instead of updating the entire model, PEFT methods introduce lightweight trainable modules or restrict optimization to specific parameter subsets. This paradigm significantly lowers the memory footprint and training cost, democratizing the use of LLMs across diverse applications [2].

Among various PEFT strategies, four representative methods stand out. Full fine-tuning serves as the baseline, providing strong task-specific adaptation at the cost of full parameter updates. LoRA introduces low-rank trainable matrices to approximate weight updates, which achieves efficient adaptation with minimal accuracy loss. SoRA extends LoRA by enforcing sparsity in the low-rank decomposition, further compressing parameters while preserving expressive power. Prefix layer tuning optimizes trainable prefix embeddings injected into transformer attention layers, which steers model behavior without modifying backbone parameters. Figure 1 illustrates representative PEFT strategies, that includes LoRA's low-rank decomposition of weight updates and prefix layer tuning, which injects task-specific prefix embeddings into transformer attention mechanisms. Each method presents a unique balance between parameter efficiency and expressive capacity [3].

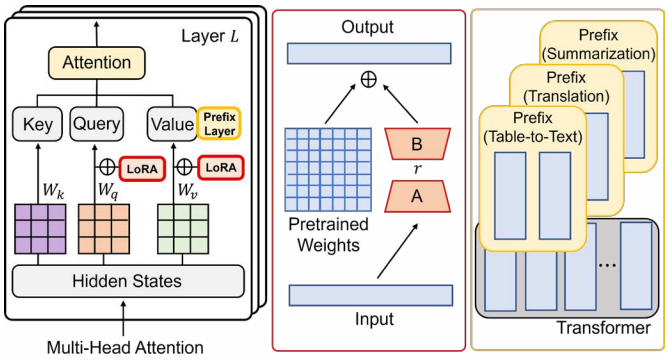

Fig. 1: Overview of parameter-efficient fine-tuning strategies.

In this work, we present a systematic introduction and empirical evaluation of these four fine-tuning strategies on standard NLG tasks [4]. Using GPT-2 and LLaMA as representative base models, we compare FT, LoRA, SoRA, and Prefix Layer in terms of both generation quality and parameter efficiency. Performance is evaluated with widely used metrics, including BLEU, BERT-F1, ROUGE-1, ROUGE-2, and ROUGE-L.

Our contributions can be summarized as follows:

- We provide a comprehensive introduction to four representative fine-tuning strategies for NLG tasks, which bridges the gap between full fine-tuning and parameter-efficient approaches.
- We conduct extensive experiments on GPT-2 and LLaMA, that analyzes the trade-offs between parameter efficiency

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2025-00561377). *(Corresponding author: Joongheon Kim)*

and generation quality.
- We highlight the conditions under which PEFT methods, especially LoRA and SoRA, achieve competitive or superior performance relative to full fine-tuning, offering practical insights for real-world deployment.

Through this paper, we aim to establish a clearer understanding of PEFT methods for NLG and provide guidance for scalable and efficient adaptation of LLMs in resource-constrained environments.

II. Methodology

A. Full Fine-Tuning

Full fine-tuning serves as the conventional approach to adapting pre-trained LLMs to downstream tasks. In this setting, all parameters θ of the pre-trained model are updated using task-specific data [5]:

$$\theta^* = \arg\min_{\theta} \mathcal{L}(M(\theta), \mathcal{D}), \tag{1}$$

where $\mathcal{L}$ denotes the task loss and $\mathcal{D}$ is the downstream dataset. This method provides maximum flexibility and ensures strong task-specific adaptation. However, its computational and storage demands scale with the full parameter count, mkaes it impractical for multi-task and resource-constrained deployments.

B. Low-Rank Adaptation

LoRA reduces the fine-tuning cost by introducing trainable low-rank matrices into the weight update process. Instead of updating a full parameter matrix $W \in \mathbb{R}^{d\times d}$, LoRA constrains the update as [6]:

$$\Delta W = AB, \quad A \in \mathbb{R}^{d\times r},\ B \in \mathbb{R}^{r\times d},\ r \quad d, \tag{2}$$

where r is the low-rank dimension. During training, A and B are optimized while W remains frozen. This reduces the number of trainable parameters from $\mathcal{O}(d^2)$ to $\mathcal{O}(rd)$, achieving efficient adaptation. LoRA has been shown to preserve generation quality while significantly lowering memory and storage requirements.

C. Sparse Low-Rank Adaptation

SoRA extends the LoRA framework by incorporating sparsity constraints into the low-rank updates. Specifically, SoRA applies a sparsity mask $\mathcal{S}$ to the decomposed matrices:

$$\Delta W = (A \odot \mathcal{S}_A)(B \odot \mathcal{S}_B), \tag{3}$$

where $\odot$ denotes element-wise multiplication. By enforcing sparsity, SoRA reduces redundant parameter updates, further improving efficiency. This approach aims to strike a balance between LoRA's expressiveness and additional parameter compression, though its performance can vary depending on task complexity and model size.

D. Prefix Layer Tuning

Prefix Layer Tuning adapts the model by prepending learnable prefix vectors to the hidden representations in transformer layers. For each self-attention block, trainable prefix embeddings P are concatenated with the original key (K) and value (V) matrices:

$$\text{Attn}(Q, [P; K], [P; V]), \tag{4}$$

where Q denotes the query representation. These prefix parameters act as soft prompts that steer the model towards task-specific behaviors while keeping the backbone weights frozen. Prefix tuning can be applied selectively to certain layers (e.g., the first few transformer blocks), trading off between efficiency and expressive capacity. This method is particularly effective in generative settings, as it provides direct control over contextual conditioning with a small number of trainable parameters.

III. Performance Evaluation

A. Experimental Setup

We evaluate four adaptation methods, full fine-tuning (FT), LoRA, SoRA, and prelayer tuning, on NLG tasks. Two representative base models are considered: GPT-2, a medium-scale autoregressive model, and LLaMA, a modern large-scale transformer-based LLM. Standard NLG metrics, including BLEU, BERT-F1, ROUGE-1, ROUGE-2, and ROUGE-L, are reported. To quantify parameter efficiency, we additionally report the ratio of trainable parameters (#TP Ratio) relative to full fine-tuning.

B. Results on GPT-2

As shown in Table I, full fine-tuning achieves the strongest results across all metrics, with BLEU 10.47, BERT-F1 83.99, and ROUGE-2 17.87. In contrast, PEFT methods on GPT-2 exhibit a significant performance drop despite large reductions in trainable parameters. LoRA and SoRA, each updating only 0.24% of parameters, yield BLEU scores around 1.1 and ROUGE-L below 8.0. PreLayer tuning, despite a higher parameter ratio (7.31%), fails to match full fine-tuning, producing BLEU 0.51 and ROUGE-1 7.32. These results suggest that smaller-scale models like GPT-2 are less robust to parameter-efficient adaptation in generative tasks, as their limited capacity may hinder effective knowledge transfer through restricted parameter updates.

C. Results on LLaMA

In contrast, LLaMA demonstrates strong adaptability under PEFT methods. Full fine-tuning achieves BLEU 11.71, BERT-F1 85.63, and ROUGE-2 19.85, establishing a strong baseline. Remarkably, LoRA (0.10% #TP Ratio) surpasses full fine-tuning in BLEU (12.19 vs. 11.71), matches BERT-F1 (85.69 vs. 85.63), and achieves the best ROUGE-L (20.99). These findings highlight LoRA's effectiveness when applied to sufficiently expressive base models. SoRA and PreLayer tuning achieve competitive ROUGE-L scores (22.23 and 21.32, respectively), but their BLEU scores remain substantially lower, indicating that while they preserve content adequacy, they lag behind in fluency and coherence.

TABLE I: Comparison of the standard NLG evaluation metrics using different adaptation methods.

Model& Method	#TP Ratio	BLEU	BERT	R-1	R-2	R-L
GPT2 (FT)	100%	**10.47**	**83.99**	**18.60**	**17.87**	**18.45**
GPT2 (LoRA)	0.24%	1.15	78.62	9.40	2.62	7.12
GPT2 (SoRA)	0.24%	1.18	76.98	9.29	2.43	7.66
GPT2 (PreLayer)	7.31%	0.51	77.19	7.32	1.72	6.24
LLaMA (FT)	100%	11.71	85.63	20.25	**19.85**	20.25
LLaMA (LoRA)	0.10%	**12.19**	**85.69**	**22.26**	19.08	**20.99**
LLaMA (SoRA)	0.10%	4.99	85.21	13.38	12.34	20.23
LLaMA (PreLayer)	0.53%	5.91	84.99	18.75	12.75	21.32

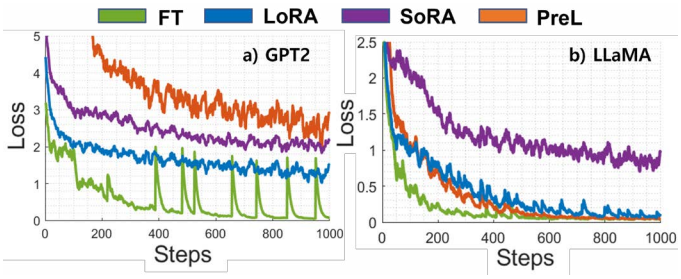

Fig. 2: Training loss comparison across 1,000 steps and ratio of trainable parameters (%) across models.

D. Training Loss Analysis

To gain a deeper understanding of the optimization dynamics in PEFT, we further examined the training loss curves over 1,000 optimization steps, as illustrated in Figure 2. For GPT-2 (Figure 2a), full fine-tuning consistently achieved the lowest training loss and converged rapidly within the first 200 steps. In contrast, LoRA and SoRA exhibited slower convergence rates and stabilized at higher residual loss values, reflecting their limited adaptation capacity in smaller-scale models. prefix layer tuning (PreL) demonstrated the slowest convergence behavior and the highest final loss, suggesting that GPT-2 lacks sufficient representational capacity to fully leverage prefix-based adaptation.

In the case of LLaMA (Figure 2b), the gap between full fine-tuning and PEFT methods narrowed considerably. LoRA closely followed the convergence trajectory of full fine-tuning and ultimately reached comparable loss values despite updating only 0.1% of the parameters. SoRA and PreL converged more slowly but still achieved stable reductions in training loss, indicating that larger-scale models are inherently more robust to parameter-efficient adaptation. Among the examined approaches, LoRA displayed the most stable and reliable convergence dynamics across both models, while SoRA and PreL exposed a trade-off between parameter savings and training stability due to their slower convergence and higher residual losses.

E. Discussion

The results demonstrate that the effectiveness of PEFT methods is highly scale-dependent. Smaller models such as GPT-2 suffered from significant degradation and unstable convergence under parameter-efficient adaptation, whereas larger models like LLaMA maintained stable optimization and strong generative quality even with minimal parameter updates. Among the evaluated methods, LoRA consistently achieved the best trade-off between efficiency and accuracy, closely matching or even surpassing full fine-tuning across BLEU, ROUGE, and BERT-F1 while updating less than 0.1% of parameters. In contrast, SoRA and PreLayer exhibited task- and metric-specific behaviors, preserving content adequacy but showing slower convergence and weaker fluency. Overall, these findings highlight the potential of PEFT approaches to substantially reduce computational and storage overhead, with LoRA emerging as the most practical solution for real-world NLG applications.

IV. Conclusion

In this paper, we presented an introduction and comparative evaluation of PEFT techniques for LLMs on NLG tasks. We examined four representative approaches: FT, LoRA, SoRA, and Prefix Layer Tuning, and analyzed their trade-offs in terms of generation quality and parameter efficiency. Our experimental results on GPT-2 and LLaMA reveal three key findings. First, full fine-tuning remains the most reliable strategy across all metrics but comes with prohibitive computational and storage costs. Second, LoRA demonstrates the most favorable balance, achieving comparable or even superior BLEU, ROUGE, and BERT-F1 scores while updating less than 0.1% of parameters. Third, the effectiveness of PEFT methods is highly model-dependent: while GPT-2 shows substantial degradation under parameter-efficient adaptation, LLaMA benefits significantly, highlighting the importance of model scale and architecture. These findings underscore the potential of PEFT strategies to enable efficient and scalable deployment of LLMs for NLG applications. By drastically reducing parameter overhead without significant loss in performance, methods such as LoRA and SoRA make it feasible to adapt LLMs in multi-task and resource-constrained environments.

References

[1] L. Hu, H. He, D. Wang, Z. Zhao, Y. Shao, and L. Nie, "LLM vs small model? Large language model based text augmentation enhanced personality detection model," in *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, vol. 38, no. 16, Vancouver, Canada, February 2024, pp. 18 234–18 242.

[2] E. J. Roh and J. Kim, "Quantum-amplitude embedded adaptation for parameter-efficient fine-tuning in large language models," in *Proc. ACM Int'l Conf. Information and Knowledge Management (CIKM)*, Seoul, Korea, November 2025.

[3] J. Oh, D. Lee, D. Won, W. Noh, and S. Cho, "Communication-efficient federated learning over-the-air with sparse one-bit quantization," *Transactions on Wireless Communication*, vol. 23, p. 15673–15689, October 2024.

[4] S. Park, J. P. Kim, C. Park, S. Jung, and J. Kim, "Quantum multi-agent reinforcement learning for autonomous mobility cooperation," *IEEE Communications Magazine*, vol. 62, no. 6, pp. 106–112, June 2024.

[5] A. Kumagai and T. Iwata, "Learning non-linear dynamics of decision boundaries for maintaining classification performance," in *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, vol. 31, no. 1, San Francisco, California, USA, February 2017.

[6] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "LoRA: Low-rank adaptation of large language models," in *Proc. International Conference on Learning Representations (ICLR)*, Virtual, April 2022.