
📖 A Novel Parameter-Efficient Fine-Tuning Methodology for Enhancing Large Language Model Syntactic Precision via Pāṇinian Grammar Injection

1. Abstract

This paper introduces **P-LoRA (Pāṇinian Low-Rank Adaptation)**, a novel Parameter-Efficient Fine-Tuning (PEFT) methodology designed to inject the deterministic syntactic and morphological rules of **Pāṇini's Aṣṭādhyāyī** into pre-trained Large Language Models (LLMs). Recognizing that Pāṇinian grammar provides an **algorithmic, low-rank structure of linguistic dependencies**, we propose adapting specialized knowledge derived from a small-scale Sanskrit language model (SSL) to adjust the Attention and Feed-Forward Network (FFN) layers of a general-purpose LLM. This targeted approach aims to significantly enhance the LLM's grammatical coherence and logical consistency, mitigating syntactic ambiguity and the high computational costs associated with full model fine-tuning.

2. Introduction and Motivation

While modern Transformer-based LLMs demonstrate exceptional fluency and semantic coverage, they frequently exhibit instability and probabilistic errors in handling highly complex syntactic structures and long-range dependencies. These errors stem from the reliance on statistical co-occurrence patterns learned from vast, often noisy, corpora.

The Sanskrit grammar formalized by Pāṇini represents a highly optimized, deterministic system with algorithmic precision. This system is distinguished by its structured rules for morphology (*Sandhi, Samāsa*), case-based semantics (Karaka Theory), and a high degree of word-order freedom. We hypothesize that this inherent rigor offers a unique opportunity to provide a dense, high-quality signal for **syntactic rule regularization** within an LLM. By treating Pāṇinian grammar as an optimal, concise linguistic algorithm, we can create a low-rank adjustment that corrects the "statistical noise" in the general LLM's learned syntactic rules.

3. Methodology: P-LoRA Implementation

The P-LoRA methodology is designed to isolate and adjust the core mechanisms of linguistic rule-following within the Transformer architecture while keeping the bulk of the model's factual and semantic knowledge frozen.

3.1. Phase I: Sanskrit Structure Extraction (Source Model)

A prerequisite for P-LoRA is the development of a small-scale Transformer model (SSL) trained exclusively on a corpus annotated with Pāṇinian linguistic features (e.g., dependency parsing, *dhātu* roots, *pratyaya* derivations).

- **Objective:** The SSL is trained on tasks requiring deterministic linguistic production, such as predicting correct inflectional forms and resolving *sandhi* rules.
- **Resulting Weights:** The frozen weights of the SSL's Attention and FFN layers encode a highly precise set of **algorithmic syntactic and morphological rules** with low variance. These rules form the structural guidance for the subsequent fine-tuning phase.

3.2. Phase II: P-LoRA Adapter Architecture and Target Layers

The methodology utilizes the Low-Rank Adaptation (LoRA) framework to introduce minimal, trainable parameters into the layers of the pre-trained LLM. The core weights (\mathbf{W}) of the general LLM remain frozen.

1. Attention Layer Adaptation (Syntactic Injection):

- **Target:** The Query (\mathbf{W}_Q) and Key (\mathbf{W}_K) projection matrices within the self-attention blocks. These matrices are responsible for calculating the dynamic relevance scores, i.e., the "syntactic GPS."
- **Mechanism:** Low-rank matrices (\mathbf{A} and \mathbf{B}) are introduced such that the adapted weight is $\mathbf{W}' = \mathbf{W} + \mathbf{AB}$. The adapters (\mathbf{A}, \mathbf{B}) are trained to enforce a higher priority on the dependency relations consistent with Karaka Theory and to be robust against variations in word order, a key feature of the target grammar.

2. FFN Layer Adaptation (Coherence and Correction Injection):

- **Target:** The FFN weight matrices ($\mathbf{W}_{F,1}, \mathbf{W}_{F,2}$).
- **Mechanism:** LoRA adapters are applied to the FFN layer. Since the FFN acts as a knowledge store and feature corrector, the adapters here are trained to refine the contextualized vectors to maximize the **logical and structural coherence** dictated by the tight Pāṇinian rules, essentially filtering out outputs that are grammatically permissible in general language but structurally ambiguous or contradictory in the strict Pāṇinian framework.

3.3. Phase III: Constrained Fine-Tuning Objective

The P-LoRA fine-tuning employs a multi-objective loss function to balance general language fluency with structural fidelity:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{NLL}} + \lambda \mathcal{L}_{\text{Pāṇini}}$$

- \mathcal{L}_{NLL} : The standard Negative Log-Likelihood loss for next-token prediction.

- $\mathcal{L}_{\text{Pāṇini}}$: A **Structural Constraint Loss** term that penalizes the model's output distribution when it violates automatically verifiable Pāṇinian rules (e.g., agreement features, *sandhi* rules) based on the structural outputs derived from the SSL. The parameter λ controls the regularization strength.

4. Expected Results and Conclusion

The P-LoRA methodology is expected to yield substantial improvements in the syntactic precision of LLMs with minimal computational overhead. Performance metrics (e.g., BLEU, CHRF, and dependency parsing scores) are anticipated to show gains, especially on tasks requiring deep structural analysis or long-term coherence. The low-rank adaptation ensures that the core universal knowledge of the frozen LLM remains intact, preventing catastrophic forgetting while efficiently injecting a powerful, deterministic linguistic algorithm into the model's dynamic rule-following mechanisms. This work demonstrates a viable path toward creating structurally robust and logically sound generative AI systems.

Acknowledgments

I, Govind Reddy, gratefully acknowledge the invaluable assistance provided by the Gemini AI in structuring, refining, and generating the formalized content and technical terminology used in this paper.

References

1. Pāṇini. (*Aṣṭādhyāyī*). Ancient Indian treatise on Sanskrit grammar.
2. Hu, E. J., et al. (2021). LoRA: Low-Rank Adaptation of Large Language Models. *ICLR 2022*.
3. Jawahar, G., et al. (2019). What Does BERT Look At? An Analysis of BERT's Attention. *BlackboxNLP*. (Relevant for Attention Layer specialization).
4. Dettmers, T., et al. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. *arXiv:2305.14314*. (Relevant for practical implementation and efficiency).
5. Karaka Theory. (A central concept in Pāṇinian grammar for case-based semantic roles).