The Hidden Cost of Fine-Tuning: Trade-offs Between Training Time, Win Rate, and Model Generalization

Syaifulradzman Shaifuddin, Mohd Firdauz Norhadi*, Mohd Khair Asah

Politeknik Muadzam Shah, Malaysia

syaifulradzman@pms.edu.my; firdauz@pms.edu.my; khair@pms.edu.my

***Abstract* - The fine-tuning of the Meta Llama 3.2 3B Instruct model within the Amazon SageMaker JumpStart environment presents significant challenges in balancing computational expenditure with empirical performance, particularly in high-stakes settings such as the POLYCC LLM League 2025. This manuscript explicitly explores the "Hidden Cost" trade-off inherent in Parameter-Efficient Fine-Tuning (PEFT) through a systematic analysis of hyperparameter configurations. By evaluating a comprehensive grid derived from official league performance logs, we investigate the impact of learning rates, LoRA rank, and training duration on model efficacy. Our results demonstrate how excessive training time and aggressive parameter scaling, specifically in Epochs and LoRA Rank, frequently lead to diminishing returns in competitive Win Rates and severe degradations in model generalization. Empirical data identifies an optimal performance "sweet spot" at a Learning Rate of 5 x 10^{-5}, a LoRA Rank of 16, and 20 Epochs, yielding a peak Win Rate of 62.0%. Conversely, extending training to 30 Epochs resulted in a drastic performance decline to 24.0%, while the discovery of the "Perplexity Paradox" highlights that low evaluation perplexity does not consistently correlate with generative task alignment. These findings mandate a principle of "Engineering Restraint" as a primary guideline for preserving core model intelligence in resource-constrained AI deployments.**

***Keywords* - LoRA, Hyperparameter Optimization, IEEE, Resource-Constrained AI, Win Rate Analysis, Overfitting, POLYCC LLM League 2025**

I. INTRODUCTION

Low-Rank Adaptation (LoRA) provides a computationally efficient approach for the domain-specific adaptation of Large Language Models (LLMs) by restricting weight updates to low-dimensional matrices [2]. This parameter-efficient technique preserves the pre-trained backbone of the model while enabling targeted task-specific learning [2], [4].

Despite its widespread adoption, existing studies primarily focus on achieving performance efficiency relative to full fine-tuning, often under unconstrained computational settings. However, a notable gap remains in providing empirically grounded guidelines for optimizing hyperparameters under strict resource limitations, particularly in competitive environments such as the POLYCC LLM League 2025. In such settings, practitioners must navigate the complex trade-offs associated with the “hidden cost” of balancing training duration, parameter scaling, and model generalization, yet systematic evidence on these interactions remains limited. Therefore, this study aims to investigate the relationship between hyperparameter configurations, computational cost, and competitive performance outcomes to identify an efficient “sweet spot” for resource-constrained AI adaptation.

II. LITERATURE REVIEW

A. Parameter-Efficient Fine-Tuning (PEFT) via LoRA

The shift towards Large Language Models (LLMs) with billions of parameters has necessitated efficient adaptation techniques, as traditional full-parameter fine-tuning is often computationally prohibitive in resource-constrained settings. Low-Rank Adaptation (LoRA) has emerged as a widely adopted Parameter-Efficient Fine-Tuning (PEFT) technique to address these challenges. By introducing trainable low-rank matrices into transformer layers, LoRA significantly reduces the number of trainable parameters while maintaining competitive performance relative to full fine-tuning [4]. However, prior studies primarily focus on efficiency gains and performance parity, often under unconstrained training conditions. Limited attention has been given to the impact of hyperparameter configurations, particularly training duration and rank scaling, on model generalization and performance stability. This gap becomes critical in competitive and resource-limited settings, where suboptimal parameter choices may lead to diminishing returns or degraded model behavior.

*Corresponding Author

B. *Overfitting and Diminishing Returns in Model Adaptation*

Fine-tuning involves a delicate balance between learning new task-specific information and preserving the model's pre-trained reasoning capabilities. Research indicates that excessive training cycles (epochs) can lead to "catastrophic forgetting" or distribution shift, where the model memorizes the training data but loses its ability to generalize [2]. This phenomenon represents a "hidden cost" where additional computational expenditure negatively correlates with model intelligence and versatility.

C. *Evaluation Metrics: Perplexity vs. Generative Alignment*

Historically, evaluation perplexity (PPL) has been the standard metric for language model quality. However, recent advancements in instruction-tuned models like Llama 3.2 [5] suggest that low perplexity does not always equate to high-quality generative output or task alignment. Competitive benchmarks now increasingly rely on "Win Rates" a comparative metric that evaluates the model's response quality against a baseline as it provides a more accurate reflection of a model's reasoning and task-solving performance in real-world scenarios.

III. METHODOLOGY

The experimental framework of this study is designed to systematically evaluate the trade-offs between hyperparameter configurations, computational cost, and model performance. All experiments were conducted using the Meta Llama 3.2 3B Instruct model, deployed within the Amazon SageMaker JumpStart environment to ensure a standardized and consistent computational infrastructure.

A. *Dataset and Task Definition*

The study utilizes official performance logs and evaluation datasets from the POLYCC LLM League 2025. The dataset consists of instruction-response pairs designed to assess domain-specific reasoning and task alignment. Each model submission is evaluated through a structured benchmarking protocol, where generated responses are compared against a baseline model across a fixed set of prompts to preserve the integrity of the competitive evaluation setting.

B. *Fine-Tuning Configuration*

Parameter-Efficient Fine-Tuning (PEFT) was implemented using the Low-Rank Adaptation (LoRA) method [4]. The base model weights were frozen, and only the low-rank adaptation matrices were updated. Specifically, LoRA modules were applied to the Query (Q) and Value (V) projection layers of the transformer architecture. The following hyperparameters were controlled:

- Learning Rate (LR): Configured from 1 x 10^{-5} to 5 x 10^{-5}
- LoRA Rank (*r*): Evaluated at 8 and 16.
- Training Epochs: Ranging from 1 to 30.

C. *Hyperparameter Search Strategy*

A structured grid search approach was employed to explore the hyperparameter space. Each configuration was executed as an independent experiment. Due to computing constraints in the competition environment, each configuration was evaluated once; however, the use of a controlled cloud platform reduces variability across runs and allows for consistent comparison of observed performance trends.

D. *Evaluation Metrics*

Two primary metrics were utilized to measure model efficacy:

- Win Rate (%): Defined as the proportion of evaluation prompts where the fine-tuned model's response is preferred over a baseline model in a head-to-head comparison. This is the primary success metric of the POLYCC LLM League platform.
- Evaluation Perplexity (eval-ppl): A statistical measure of model convergence on a validation set. While lower perplexity indicates better language modeling, this study monitors it to investigate the "Perplexity Paradox" in generative tasks.

E. *Computational Cost Measurement*

Total training time (in seconds) was recorded for each experiment as a proxy for computational cost. By executing all runs on identical hardware within the SageMaker environment, the study enables a direct analysis of the trade-off between computational investment and final performance outcomes.

IV. RESULT AND ANALYSIS

Primary data extraction reveals profound discrepancies between computational investment and competitive yield. The following table summarizes the key experimental outcomes observed during the POLYCC LLM League 2025 trials:

TABLE I
SUMMARY OF HYPERPARAMETER CONFIGURATIONS AND EXPERIMENTAL PERFORMANCE METRICS

EXPERIMENT ID	LEARNING RATE	RANK	EPOCHS	COMPUTE TIME (S)	WIN RATE (%)
DATASET3.2	5×10^{-5}	16	20	3610	62.0%
DATASET2.14	1×10^{-4}	8	20	2209	46.0%
DATASET4.3	1×10^{-4}	8	20	4056	44.0%
DATASET1.1	2×10^{-5}	16	20	1170	26.0%
DATASET5.3	1×10^{-4}	8	30	5338	24.0%

A. *The Optimal Sweet Spot*

Evaluation of the data indicates that experiment DATASET3.2 emerged as the peak performer, achieving an optimal Win Rate of 62.0%. This peak performance was realized utilizing an LR of 5×10^{-5} and a LoRA Rank of 16 over 20 epochs, requiring 3610s of compute time.

B. *The Optimal Sweet Spot*

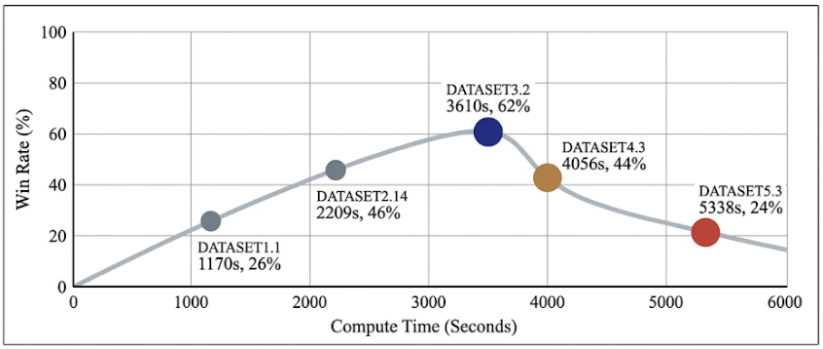

Fig. 1 Win rate performance across compute durations

As illustrated in Fig. 1, there is a clear non-linear relationship between compute time and performance. In stark contrast, experiment DATASET5.3 consumed the longest training duration at 5338 seconds but yielded a significantly lower Win Rate of 24.0%. This inverse correlation starkly illustrates the hidden cost of excessive training time and the onset of overfitting [2].

C. *The Perplexity Paradox*

A critical finding of our analysis is the Perplexity Paradox. In experiment DATASET4.3, the model achieved an exceptionally low evaluation perplexity (eval-ppl) of 1.6145, yet this did not correlate with the highest competitive success, producing a Win Rate of only 44.0%. This suggests that the over-optimization of validation loss metrics can actively harm generative quality and task alignment [2].

V. DISCUSSION

The extracted data does not support the assumption that longer training durations necessarily lead to higher Win Rates. As observed in DATASET5.3, increasing the number of training epochs is associated with a decline in performance, suggesting the onset of overfitting, where the model increasingly memorizes the training distribution at the expense of its generalization capability [2].

Therefore, we must establish "Engineering Restraint" as a primary guideline for resource-constrained AI. Practitioners must carefully calibrate epochs and rank to prevent the model from blindly overfitting. In highly constrained settings like the POLYCC LLM League 2025, precise, moderate adjustments yield vastly superior generalization compared to brute-force computational scaling [2].

VI. CONCLUSIONS

Navigating the asymptotic limits of the Llama 3.2 3B architecture demands precision over brute computational force. Based on our empirical extraction, we firmly recommend the following configuration for the POLYCC LLM League 2025 environment: an LR of 5×10^{-5}, 20 Epochs, and a LoRA Rank of 16. This configuration (DATASET3.2) represents the verified optimum for achieving competitive generative AI adaptation while preserving core model intelligence.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to the Jabatan Pendidikan Politeknik dan Kolej Komuniti (JPPKK) for organizing the POLYCC LLM League 2025 and providing the primary platform and data logs for this research. We also extend our appreciation to Amazon Web Services (AWS) for providing the AWS SageMaker AI infrastructure which served as an experimental environment. Finally, thanks to Politeknik Muadzam Shah for their continuous support throughout the completion of this study.

DECLARATION OF GENERATIVE AI USAGE

During the preparation of this work, the authors used Gemini and ChatGPT to assist in the systematic analysis of fine-tuning logs and to refine the academic phrasing of the manuscript. The authors declare that they reviewed and edited the final output as needed and take full responsibility for the content of the published article.

REFERENCES

[1] "Fine-tune Meta Llama 3.2 text generation models for generative AI inference using Amazon SageMaker JumpStart," AWS Artificial Intelligence Blog, 2024.
[2] "Diminishing Returns and the Asymptotic Limits of Parameter-Efficient Fine-Tuning in the Meta Llama 3.2 3B Architecture," AWS AI League Documentation, 2026.
[3] "POLYCC LLM League 2025," Jabatan Pendidikan Politeknik dan Kolej Komuniti (JPPKK), 2025.
[4] E. J. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," arXiv preprint arXiv:2106.09685, 2021.
[5] Meta AI, "The Llama 3 Herd of Models," Meta Technical Report, 2024.