Brains in Bits: A Comparative Study of PEFT Techniques for Commonsense Reasoning on GPT-2

Vishwanath Guruvayur
University of Virginia
vish@virginia.edu

Luke Napolitano
University of Virginia
ljn5yms@virginia.edu

Doruk Ozar
University of Virginia
bcp8dm@virginia.edu

Abstract

While industry convention has assumed that the bigger, the better, training large models fully is resource-intensive, often impractical, and costly to the environment according to recent research [3]. In this project, we investigate how smaller models like GPT-2 can be selectively enhanced for common-sense reasoning tasks through Parameter-Efficient Fine-Tuning (PEFT) methods, reducing the cost and enhancing the impact of LLM research and use.

We compared four PEFT techniques: LoRA, QLoRA, Prefix Tuning and IA^3. Instead of updating all model parameters, these methods selectively adjust small, critical parts of the model, particularly within its attention mechanisms. Our results show that IA^3 and Prefix Tuning substantially outperform LoRA and QLoRA, achieving up to 50% validation accuracy in commonsense categories, with noticeable reductions in perplexity. Each of the four methods, applied to the smallest available GPT-2 model, outperform the largest GPT-2 model in a zero-shot test of common-sense reasoning.

These findings suggest that PEFT methods can meaningfully strengthen specific reasoning capabilities even in smaller, open-source versions of the multi-billion parameter proprietary LLMs. Our work supports the idea that modular, targeted training strategies could form a scalable alternative to LLM training that significantly reduces the environmental and computational impact that off-the-shelf multi-billion parameter models currently offer. Future directions include expanding to ensemble PEFT methods and scaling experiments to other models such as Quen 3, Llama 4, and Mistral 3.1.

1 Introduction

The release of ChatGPT in late 2022 brought about a wave of LLM advancements that have seemed to follow a trend: bigger is better. Recently, a new proprietary model is released almost every month with more parameters and purportedly better performance than any model before it. However, as these top-of-the-line models have gotten more advanced, they have continued to become more expensive to train and use.

Current approaches often involve training ever-larger models using massive datasets and computational resources. However, this brute-force strategy raises practical concerns about scalability, accessibility, and interpretability. Fine-tuning such large models from scratch is increasingly impractical, especially for smaller research groups or applications that need specialized adaptation.

In this project, we investigate whether the performance of more accessible models like a small GPT-2 notably improves by using Parameter-Efficient Fine Tuning (PEFT) techniques when applied to a very specific task like MCQ answering. Rather than updating millions of parameters, PEFT methods selectively adapt a small subset of components, particularly within the model's attention mechanisms, which can be thought of as the 'brain' of LLMs.

*Course Project for DS-6051 – Decoding Large Language Models.

Our primary objectives are threefold:

- Assess whether PEFT methods can inject common sense reasoning capabilities into a basic LLM.
- Compare the effectiveness of different PEFT strategies (LoRA, Prefix Tuning, IA^3, and QLoRA) in improving model performance.
- Understand how different types of reparametrization and adaptive techniques affect different cognitive abilities of the LLM Transformer Architecture.
- Reflect on the broader implications of modular training approaches for LLM research and general use.

By studying how small, controlled updates can meaningfully improve commonsense reasoning, we hope to gain insights into how structured cognitive skills might be cultivated in language models without the need for massive retraining or scaling alone.

2 Related works

The development of efficient fine-tuning strategies for Large Language Models (LLMs) has been an active area of research, particularly as model sizes and training costs continue to grow.

A Critical Review of PEFT [5] covered the history of parameter-efficient fine-tuning techniques, their applications, and future direction for LLM training and tuning. The use of PEFT seems to improve parameter efficiency and reduce computational requirements.

LoRA and QLoRA [2] explained how LoRA freezes the original weights of the model and adds a low-rank matrix that is tuned to the context. The author also briefly covered how QLoRA quantizes low-rank matrices to a lower precision.

Prefix Tuning [4] summarized the addition of a learned prefix tokens to the beginning of a model's processing pipeline that help guide the model's behavior.

IA^3 (Input-Adaptive Attention) [1] extended these ideas by learning input-dependent scaling factors inside the attention and feedforward modules. IA^3 showed particular promise for few-shot fine-tuning, achieving competitive results with significantly fewer trainable parameters.

Most prior work evaluated PEFT techniques on instruction-following, summarization, or dialogue tasks. However, their application to *commonsense reasoning* which is a core component of AGI remains rather underexplored. Our project contributes by systematically comparing these PEFT methods on the CommonsenseQA benchmark, focusing on the ability to teach implicit world knowledge to a relatively small model like GPT-2.

3 Methodology

In this section we will discuss our approach to experiment with different PEFT methods for our finalized Dataset.

3.1 Data: CommonsenseQA

Our initial experiments aimed to fine-tune GPT-2 on the OpenMathInstruct-2 dataset for mathematical reasoning. However, due to GPT-2's limited capacity to understand complex symbolic structures, training stalled at a high loss value of 6.6 without notable accuracy improvements.

We did not want to move to a larger model like Mistral 6B Instruct Model because we wanted to test the capabilities of PEFT methods in a very basic GPT like GPT-2 and not have any beneficiary aspects of having a larger model.

Recognizing the mismatch between task complexity and model capacity, we pivoted to **CommonsenseQA**, a multiple-choice question-answering dataset focused on everyday commonsense knowledge. Each sample in CommonsenseQA consists of a question and five candidate answers, with exactly one correct choice. The dataset is designed to challenge models on reasoning about implicit world knowledge rather than relying purely on surface-level patterns, making it a suitable benchmark for our objectives.

The dataset also has 785 categories of these common sense questions. We clustered these categories into a larger set of 10 broad classes to better understand the proficiency each method shows in these broader common sense classes.

3.2 Model: GPT-2 with PEFTs

We selected the **GPT-2 small** model (approximately 125M parameters) as our base architecture. GPT-2's autoregressive nature makes it naturally effective at next-word prediction, which aligns well with the multiple-choice format where short, precise outputs are expected. Additionally, GPT-2's manageable size allowed us to conduct multiple fine-tuning experiments within reasonable compute constraints.

3.3 Parameter Efficient Fine-Tuning Techniques

We applied and compared four different PEFT methods:

- **LoRA**: Introduces low-rank updates within attention weights.
- **QLoRA**: Combines 4-bit model quantization with LoRA to enable memory-efficient fine-tuning.
- **Prefix Tuning**: Appends learned prefix tokens to each transformer layer's input.
- **IA^3**: Adds learnable scaling factors to the key, value, and feedforward transformations.

Each method adapts only a small fraction of the total model parameters, keeping the core model weights frozen.

3.4 Experimental Setup

We trained all models using a single NVIDIA A100 GPU, leveraging the HuggingFace PEFT library for efficient implementations. Hyperparameters were held consistent across experiments for fair comparison:

- Optimizer: AdamW
- Learning Rate: 5e-5

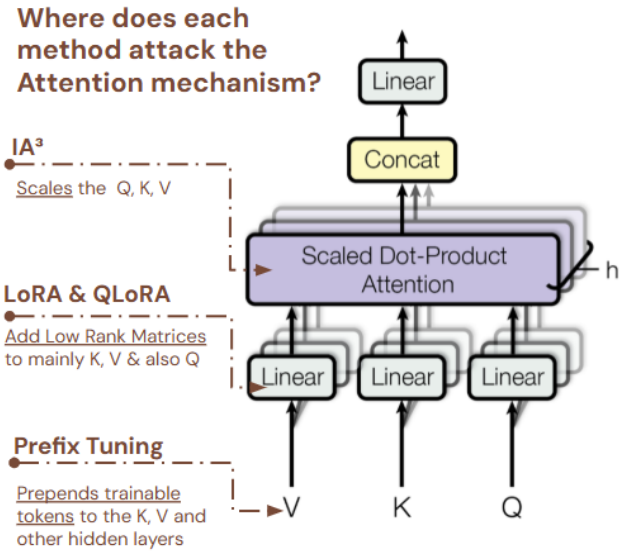

Figure 1: Intuitive Visualization of the different target regions for each PEFT Method

- Batch Size: 8
- Evaluation Strategy: Every 100 steps
- Number of Epochs: 10

Evaluation was conducted using two primary metrics:

- **Validation Accuracy**: Percentage of correct answers selected. We also calculated this accuracy within the broad classes of questions to understand each method's expertise in specific types of common sense.
- **Perplexity**: A measure of model confidence on validation sequences.

To simulate realistic low-resource settings, we limited training to small-to-moderate compute budgets, avoiding large-scale hyperparameter searches.

4 Experiments and Results

4.1 Initial Exploration: OpenMathInstruct-2

Our early experiments on the OpenMathInstruct-2 dataset exposed a fundamental mismatch between GPT-2's architecture and the demands of mathematical reasoning.

Despite extensive training, the model's loss plateaued at approximately 6.6, with minimal improvements in prediction accuracy. This stagnation suggested that the symbolic complexity and structured reasoning required by math datasets exceeded GPT-2's representational capacity.

4.2 Strategic Pivot: CommonsenseQA

Recognizing these limitations, we transitioned to the CommonsenseQA dataset, which emphasizes sentence-level, everyday reasoning—better suited to GPT-2's strengths. This shift allowed us to more accurately assess the effectiveness of parameter-efficient fine-tuning (PEFT) methods without architectural bottlenecks dominating outcomes.

4.3 Zero-Shot Baseline Performance

Without fine-tuning, the 1.5B parameter GPT-2 had zero-shot accuracy of roughly 6.35%—consistent with a five-choice multiple-selection task. This established a clear motivation for targeted fine-tuning on a smaller model.

4.4 Parameter-Efficient Fine-Tuning (PEFT) Results

We explored four PEFT methods: **LoRA**, **QLoRA**, **Prefix Tuning**, and **IA^3**. Each method's performance is discussed below, including validation accuracy, perplexity, and training dynamics.

4.4.1 LoRA

Results:

- Validation Accuracy: **12%**
- Perplexity: **14.31**

Despite reducing perplexity substantially compared to the zero-shot baseline, LoRA yielded only modest accuracy gains, suggesting limited improvement in reasoning capabilities.

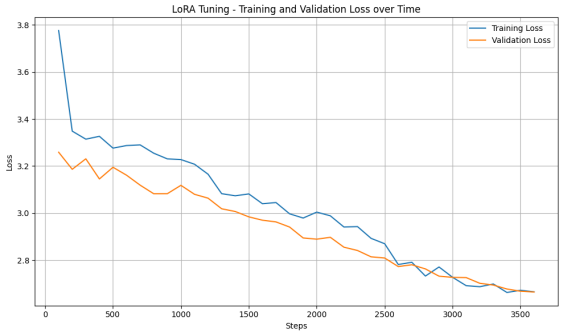

Figure 2: Training and validation loss curves for LoRA fine-tuning.

4.4.2 QLoRA

Results:

- Validation Accuracy: **15%**
- Perplexity: **14.45**

QLoRA achieved slightly better accuracy than LoRA with comparable perplexity. However, the improvement remained marginal, reinforcing challenges in adapting GPT-2 to structured reasoning even with quantized fine-tuning.

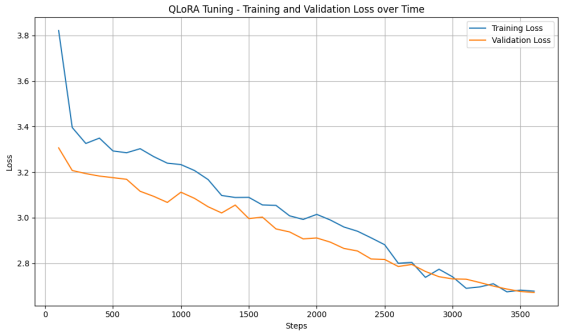

Figure 3: Training and validation loss curves for QLoRA fine-tuning.

4.4.3 Prefix Tuning

Results:

- Validation Accuracy: **46%**
- Perplexity: **14.06**

Prefix Tuning delivered a major jump in accuracy. Despite only a minor further drop in perplexity compared to LoRA and QLoRA, it enabled much stronger multiple-choice reasoning.

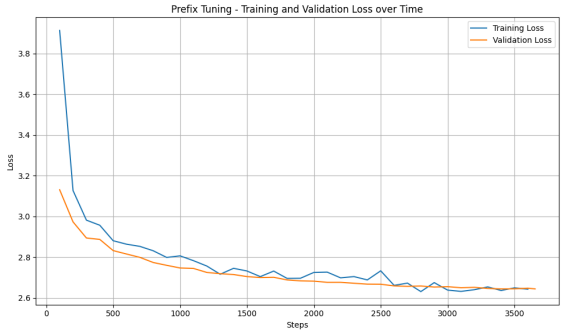

Figure 4: Training and validation loss curves for Prefix Tuning fine-tuning.

4.4.4 IA³

Results:

- Validation Accuracy: **50%**
- Perplexity: **16.24**

IA³ achieved the highest accuracy despite a slightly higher perplexity. This suggests that in CommonsenseQA, perplexity reductions do not fully capture improvements in discriminative reasoning.

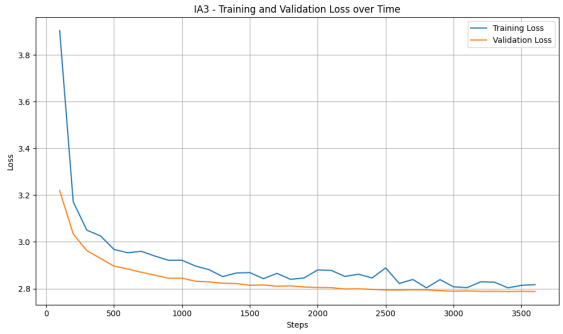

Figure 5: Training and validation loss curves for IA³ fine-tuning.

4.5 Summary of Metrics

This is a summary of our experiment results

Method	Validation Accuracy (%)	Perplexity
Zero-Shot (1.5B)	6.35%	5536.87
LoRA	12%	14.31
QLoRA	15%	14.45
Prefix Tuning	46%	14.06
IA³	50%	16.24

Table 1: Summary of validation accuracy and perplexity across PEFT methods on CommonsenseQA on GPT-2 Small.

4.6 Category-Wise Analysis

To further probe model behavior, we evaluated category-specific performance across commonsense reasoning types within CommonsenseQA.

Major Concepts	ZeroShot	LoRA	QLoRA	Prefix	IA3
Buildings & Spaces	0.6%	6.3%	6.6%	33.5%	42.6%
Abstract Concepts & Events	0.0%	20.5%	19.5%	43.9%	40.6%
Emotions & Mental States	0.0%	21.2%	23.2%	48.5%	37.9%
Technology & Media	0.0%	10.4%	10.9%	44.2%	35.7%
Food & Drink	0.0%	4.0%	3.6%	42.8%	33.6%
Society & Culture	0.0%	4.8%	4.5%	33.9%	32.4%
Human Activities & Daily Life	0.0%	20.5%	20.9%	40.4%	31.9%
Nature & Environment	0.0%	6.6%	6.4%	39.1%	30.3%
Objects & Tools	0.0%	5.1%	6.0%	32.9%	28.9%
Sports & Physical Activities	0.0%	12.3%	12.5%	30.8%	27.7%

Figure 6: Validation accuracy by commonsense category across PEFT methods on GPT-2 Small.

The results suggest that different PEFT methods specialize in different types of conceptual learning. Prefix Tuning and IA³ show strong capabilities in transferring knowledge to high-level semantic domains like Emotions & Mental States, Technology & Media, and Buildings & Spaces, indicating that they are more effective for complex, abstract reasoning tasks. In contrast, LoRA and QLoRA offer more modest improvements, suggesting that while they introduce efficiency, they may underfit higher-complexity concepts compared to prefix-based or IA³-based adaptation.

The almost nonexistent performance in the Zero-Shot setting underscores that parameter-efficient fine-tuning is essential when models must generalize to nuanced or structured domains. Moreover, IA³'s dominance in categories

like Buildings & Spaces hints that task types grounded in physical or spatial reasoning particularly benefit from deeper internal adaptation mechanisms rather than surface-level prompt manipulation.

4.7 Qualitative Assessment

Qualitatively, IA^3- and Prefix-fine-tuned models showed a stronger grasp of implicit relationships between concepts, often selecting plausible answers even when questions were phrased ambiguously. In contrast, LoRA- and QLoRA-fine-tuned models behaved closer to random guessing, suggesting weaker internalization of commonsense structures.

4.8 Summary of Findings

Overall, our experiments demonstrate that **IA^3 and Prefix Tuning are highly effective** for enhancing commonsense reasoning in GPT-2 under resource constraints. While **LoRA and QLoRA** remain efficient in other contexts, they struggled to deliver meaningful improvements on CommonsenseQA. Of note, however, is that the zero-shot performance of the 1.5B parameter GPT-2 underperformed even PEFT techniques with little impact on common-sense reasoning like LoRA and QLoRA. Smaller open-source models, combined with a weave of robust PEFT techniques thus merit investigation and comparison against other leading models to test how much improvement can be squeezed out of a smaller model.

5 Discussion

Our results show that Parameter-Efficient Fine-Tuning (PEFT) methods can significantly improve the commonsense reasoning capabilities of smaller language models like GPT-2, but their effectiveness varies significantly across techniques.

5.1 Interpretation of Results

IA^3 and Prefix Tuning were substantially more effective than LoRA and QLoRA. We hypothesize that this is because IA^3 and Prefix Tuning directly intervene in the model's attention mechanisms, which are central to relational and inferential reasoning. By contrast, LoRA and QLoRA mainly inject low-rank updates into the projection layers, which may be less effective at reshaping internal reasoning processes necessary for commonsense tasks.

The significant drop in perplexity scores for IA^3 and Prefix Tuning suggests that these methods help the model gain more confident and structured internal representations of everyday knowledge, even without updating most parameters.

5.2 Limitations

While encouraging, our experiments have important limitations:

- **Model Capacity**: GPT-2 small (125M parameters) remains fundamentally limited in its ability to perform deep or multi-hop reasoning.
- **Dataset Simplicity**: CommonsenseQA focuses on relatively shallow commonsense tasks. More nuanced benchmarks could reveal different strengths or weaknesses.
- **Resource Constraints**: All training was conducted under constrained compute budgets, limiting our ability to fully optimize hyperparameters or explore longer training schedules.

5.3 Interesting Observations

A notable qualitative finding was that Prefix Tuning and IA^3 models often selected semantically plausible answers even when wrong, suggesting partial internalization of reasoning heuristics rather than pure memorization. Meanwhile, LoRA and QLoRA-fine-tuned models frequently defaulted to random choices, implying weaker structural learning.

These insights reinforce the view that small, targeted interventions inside the model's cognitive core whic is its attention mechanism, can unlock better reasoning behaviors without requiring full retraining.

6 Conclusion and Future Work

This project showed that Parameter-Efficient Fine-Tuning (PEFT) methods can significantly enhance GPT-2's performance on commonsense reasoning tasks without requiring full model retraining. Fine-tuning led to a noticeable increase in accuracy (up to 38%) and a significant reduction in perplexity score from over 3000 to around 14. Among the techniques explored, IA^3 and Prefix Tuning demonstrated substantially better results compared to LoRA and Q-LoRA.

In the future, there are several exciting directions that can be explored. Further exploration of more advanced PEFT techniques, investigation of ensemble strategies that combine different tuning methods, and scaling experiments to larger and more capable language models, such as Mistral, are planned. Additionally, applying these approaches to more challenging and complex datasets could provide deeper insights into their generalization capabilities.

References

[1] Haokun Liu et al. *Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning*. 2022. arXiv: 2205.05638 [cs.LG]. URL: https://arxiv.org/abs/2205.05638.

[2] Joshua Noble. *What is LoRA?* 2025. URL: https://www.ibm.com/think/topics/lora.

[3] David Patterson et al. *The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink*. 2022. arXiv: 2204.05149 [cs.LG]. URL: https://arxiv.org/abs/2204.05149.

[4] Ali Razavi. *Understanding Prefix Tuning*. 2023. URL: https://medium.com/@razavipour6/understanding-prefix-tuning-a-novel-approach-to-fine-tuning-language-models-dc7dafeb32e4.

[5] Lingling Xu et al. *Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment*. 2023. arXiv: 2312.12148 [cs.CL]. URL: https://arxiv.org/abs/2312.12148.