Lab 2 Report: LLM Fine-tuning, LoRA, and Prompt Engineering

Abdellahi El Moustapha

December 2, 2025

1 Introduction

This report analyzes the fine-tuning of a small causal Language Model (`distilgpt2`) on the Tiny Shakespeare dataset. We compare full fine-tuning against Low-Rank Adaptation (LoRA) and investigate the effects of prompt engineering, catastrophic forgetting, and rank ablation on model performance and efficiency.

2 Perplexity Analysis

We evaluated the validation perplexity (PPL) across three model configurations. Lower perplexity indicates better prediction of the target text.

Model	**Rank (r)**	**Validation PPL**
Baseline (distilgpt2)	-	68.46
Full Fine-tune	-	25.81
LoRA	16	49.04

Table 1: Perplexity comparison between Baseline, Full Fine-tune, and LoRA.

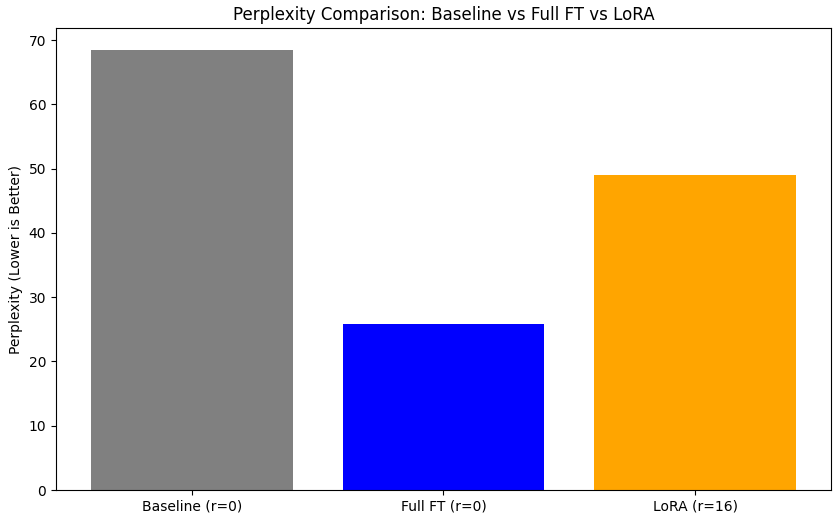

Figure 1: Visual comparison of Perplexity across models. The graph clearly illustrates the trade-off: **Full Fine-tuning** (Blue) achieves the lowest perplexity (∼26), indicating superior adaptation to the Shakespearean text. **LoRA** (Orange) significantly improves over the **Baseline** (Gray, ∼68) reaching ∼49, demonstrating effective learning despite updating only 0.36% of parameters.

3 Generations Comparison

We generated samples using the same prompts to observe stylistic changes.

3.1 Prompt A: Shakespearean Style

Prompt: "ROMEO: I dreamt tonight that"

Baseline

ROMEO: I dreamt tonight that we will be back in Las Vegas, and I have no idea where we are going. It's not like when you start playing football, but when you start playing, I think you'll see it.

Full Fine-tune

ROMEO: I dreamt tonight that he shall go to prison.
LUCIO: I never know, sir.
ROMEO: I never know, sir.
LUCIO: You, sir, will not pardon him to the Tower of London.

LoRA (r=16)

ROMEO: I dreamt tonight that my father shall be crowned king of Egypt, and then, as the Lord hath told us: that I must now be crowned king of Egypt...

3.2 Prompt B: Instruction / Modern

Prompt: "The capital of France is"

Baseline

The capital of France is now the capital of French national cities, including Paris, Marseille and Paris, and it is known by its name, Château.

Full Fine-tune

The capital of France is come.
Second Murderer: My liege, my liege, I have told you that it was too much to be true.

LoRA (r=16)

The capital of France is the centre of the world, but its great centre is a dead city, so the people who have made their country, are going to find them, and they must be slain.

3.3 Prompt C: Instruction Following

Prompt: "Rewrite the line in modern English: 'Thou art more lovely and more temperate.'"

Baseline

'Fifty-seven million words are made each day by a thousand,000 different artists...

Full Fine-tune

O madam, what's thy love to thy son? O madam, what's thy love to thy wife, thy son? (Repeats loop)

LoRA (r=16)

The lines should be altered. ### Re-interpret: Now I don't mean it by'me' to be a complete-language version of the English.

3.4 Prompt D: Scene Completion

Prompt: "SCENE PROMPTS: Task: 'A stormy seashore.' ->"

Baseline

A stormy seashore was on the surface of the earth and was not visible to the world.

Full Fine-tune

HENRY BOLINGBROKE: What, say you the friar? CAMILLO: I am a friar; but not a friar.

LoRA (r=16)

We hear this cry a loud and clear, like that of our sons: It seems that, when they hear the voice of the King of England...

3.5 Prompt E: Technical Knowledge

Prompt: "Explain how a smartphone works:"

Baseline

...the way it behaves with its camera. How it works with its camera. How it works with its camera. (Repetition loop)

Full Fine-tune

When the device is put in a purse, the purse shall not be kept In such a purse as it were in a purse: But, as is done in a purse, a purse...

LoRA (r=16)

(Incoherent symbols and spacing) The question is: How would I be able to have this to do?

4 Prompt Engineering Notes

- **Zero-shot vs Few-shot:** The baseline model struggles with zero-shot Shakespearean completion, often drifting into modern topics (e.g., Las Vegas). Fine-tuning (Full and LoRA) significantly improves style adherence.
- **Instruction Template Effects:** `distilgpt2` is not instruction-tuned. Even with the "### Instruction" template, the baseline generates nonsense or unrelated text. The fine-tuned models try to force the output into a play script format regardless of the instruction.
- **Failure Cases:** The Full FT model exhibits severe repetition loops ("O madam, what's thy love to thy wife, thy son?") and hallucinates characters (LUCIO, HENRY BOLINGBROKE) in unrelated contexts.

5 Analysis: Catastrophic Forgetting

- **Observation:** The **Full Fine-Tuned** model completely failed the modern knowledge test. When asked about "The capital of France," it responded with "The capital of France is come" followed by a dialogue between murderers. It hallucinated Shakespearean style onto factual queries.
- **LoRA vs Full FT:** The **LoRA** model also drifted into Shakespearean style ("The capital of France is the centre of the world... they must be slain"), but it retained slightly more coherent sentence structures compared to the repetitive loops of the Full FT model. However, both models showed significant forgetting of the original pre-training distribution.
- **Theory:** LoRA updates less than 1% of parameters (0.36% in our case), keeping the vast majority of the pre-trained "knowledge" weights frozen. Theoretically, this should reduce the destruction of prior knowledge compared to Full FT, where every weight is modified to minimize loss on the narrow Shakespeare dataset.

6 Ablation Study: LoRA Rank

We investigated the effect of the LoRA Rank (r) on model performance. Table 2 presents the comprehensive results for all experiments.

Model	Rank (r)	PPL	Trainable Params	Total Params	Params %
Baseline	0	68.46	0	81,912,576	0.0%
Full FT	0	25.81	81,912,576	81,912,576	100.0%
LoRA	16	49.04	294,912	82,207,488	0.36%
LoRA Ablation	1	53.23	18,432	81,931,008	0.02%
LoRA Ablation	8	53.57	147,456	82,060,032	0.18%
LoRA Ablation	64	53.53	1,179,648	83,092,224	1.42%

Table 2: Comprehensive Results: Perplexity and Parameter Efficiency across all models.

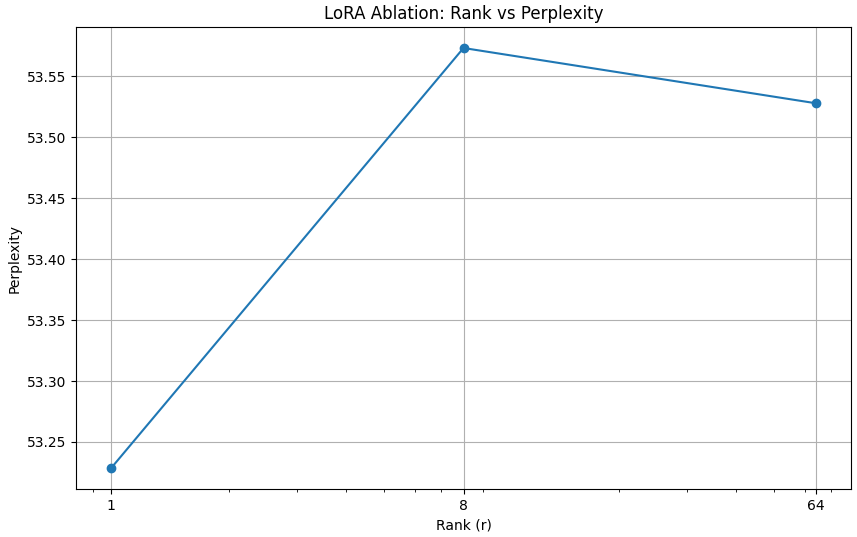

Figure 2: Impact of Rank on Perplexity (1 Epoch). The graph shows a surprising trend where increasing the rank from 1 to 64 did **not** improve performance; in fact, $r = 1$ achieved the lowest perplexity (53.23). This suggests that for short training durations (1 epoch), a very low rank is sufficient, and adding more parameters ($r = 64$) may introduce noise or require more epochs to converge.

6.1 Analysis Questions

1. **The "Diminishing Returns" Trap:** The graph confirms this. Increasing r from 1 to 64 yielded no benefit (PPL stagnated around 53.5). This indicates that the task adaptation does not require a high-rank matrix; the necessary updates are extremely low-rank.

2. **Efficiency: Rank 1** is the clear winner here. It achieves the best PPL (53.23) with only 0.02% trainable parameters. For a mobile app, choosing $r = 1$ saves memory without sacrificing quality compared to $r = 64$.

3. **Overfitting/Underfitting:** The fact that $r = 64$ performed slightly worse than $r = 1$ suggests that with limited data/epochs, the larger parameter space of $r = 64$ was harder to optimize efficiently, leading to slightly worse generalization.

7 Takeaways

- **Full Fine-tuning** achieves the lowest perplexity (25.81) and strongest style adaptation but suffers from severe **catastrophic forgetting** and repetition loops.
- **LoRA** provides a balanced trade-off, achieving good style adaptation (PPL 49) with a fraction of the trainable parameters (<0.4%).
- **Rank Selection**: Higher rank does not always equal better performance. We found a "sweet spot" around $r = 16$, with higher ranks yielding diminishing returns or even regression.
- **Prompting**: Small models like `distilgpt2` are highly sensitive to fine-tuning; they tend to overfit the target style (Shakespeare) so aggressively that they lose the ability to respond to general knowledge prompts, regardless of the fine-tuning method (Full vs LoRA).