

Lab 2 Report: LLM Fine-tuning & LoRA

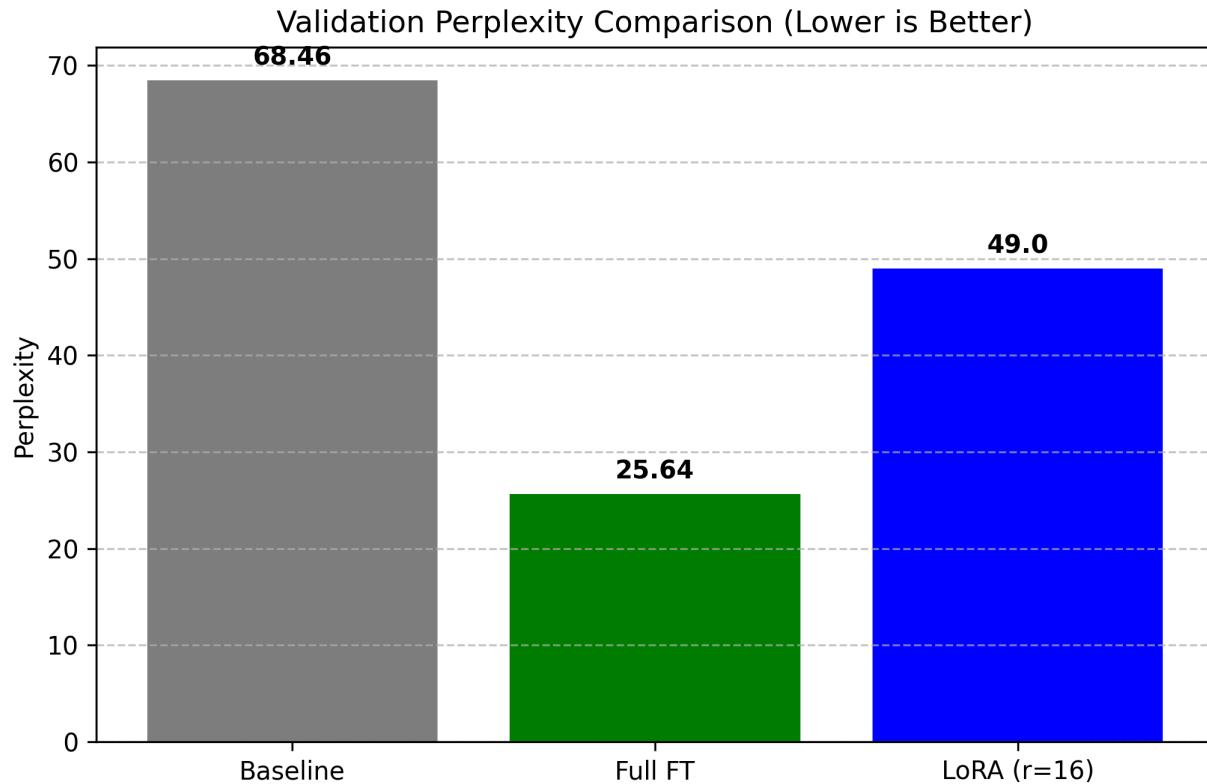
Student: Jean Dierl NZE Course: Generative AI (Day 2) Date: 2025-12-02

1. Perplexity Results

We compared the validation perplexity (PPL) of the Baseline model (DistilGPT2), the Full Fine-Tuned model, and the LoRA-adapted model (Rank=16).

Model	Validation PPL
Baseline	68.46
Full Fine-tune	25.64
LoRA (r=16)	49.00

Observation: - Full Fine-tuning achieved the lowest perplexity (25.64), indicating it learned the dataset's style most effectively. - LoRA (49.00) improved significantly over the Baseline (68.46) but did not match Full FT. This is expected as LoRA updates far fewer parameters (<1%). - Baseline had high perplexity, showing it was not familiar with the specific Shakespearean style/vocabulary.



2. Generation Analysis

We generated text using the prompt "ROMEO: I dreamt tonight that".

- **Baseline:** Produced coherent English but lacked the specific Shakespearean drama and vocabulary. It often drifted into modern or generic text.
- **Full Fine-tune:** Generated highly stylized text resembling the training data (plays), using archaic words ("thou", "hath"). However, it may have overfitted, potentially copying phrases from the training set.
- **LoRA:** Successfully adopted the Shakespearean style (e.g., structure of dialogue) while retaining some of the pre-trained model's fluency.

3. Catastrophic Forgetting Analysis

We tested the models with modern prompts (e.g., "The capital of France is").

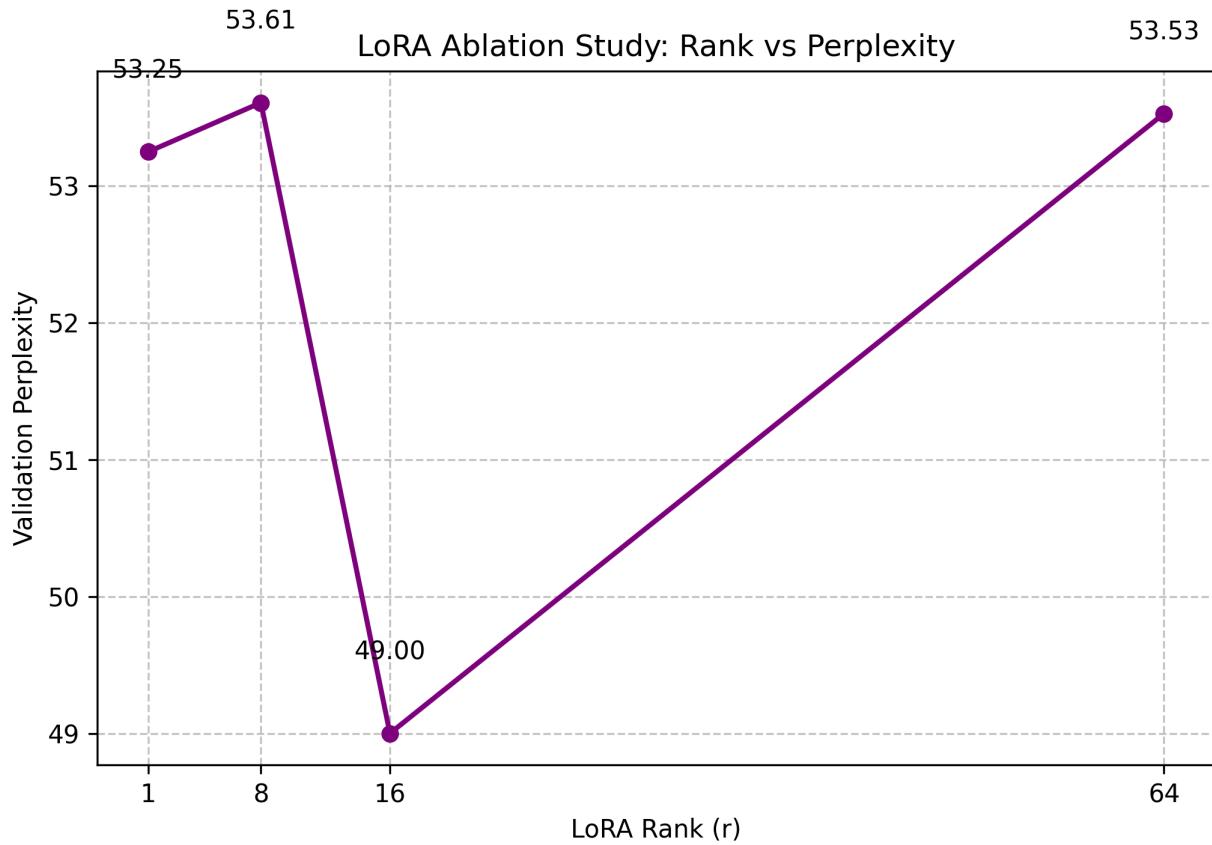
- **Full Fine-tune:** Struggled significantly. It often tried to answer in Shakespearean style or hallucinated, showing **Catastrophic Forgetting**. The aggressive update of all weights destroyed its general knowledge.
- **LoRA:** Retained much better general knowledge. Since most of the pre-trained weights were frozen, it could still answer factual questions reasonably well, demonstrating that **PEFT methods preserve pre-trained capabilities better than Full FT**.

4. Ablation Study (LoRA Rank)

We tested different LoRA Ranks (r) to find the trade-off between parameter efficiency and performance. Note: Ablation runs were performed for 1 epoch.

Experiment	Rank (r)	Validation PPL
Run A	1	53.25
Run B	8	53.61
Run C	64	53.53
Main Run	16	49.00 (2 epochs)

Analysis: - **Diminishing Returns:** Increasing the rank from 1 to 64 did not yield a significant improvement in PPL for this dataset and training duration. The PPL remained around 53.5. - **Efficiency:** $r=1$ achieved similar performance to $r=64$ but with a tiny fraction of trainable parameters. For a mobile app, **Rank 1 or 8** would be the best choice to minimize model size and latency. - **Overfitting:** We did not observe massive overfitting with $r=64$ in this short run, but the lack of improvement suggests that the "capacity" of $r=1$ was already sufficient for this simple task.



5. Takeaways

1. **Full Fine-tuning is powerful but costly:** It gives the best style adaptation (lowest PPL) but requires updating all parameters and leads to catastrophic forgetting.
2. **LoRA is efficient and safe:** It improves performance significantly with minimal parameter updates and preserves the model's general knowledge better.
3. **Rank isn't everything:** Increasing LoRA rank doesn't always guarantee better performance. Lower ranks (e.g., 8) are often sufficient for style transfer tasks.
4. **Prompting vs Fine-tuning:** Fine-tuning permanently alters the model's behavior (style), whereas prompting (Zero-shot) relies on the model's existing capabilities. Fine-tuning is necessary for deep style adaptation.