Frugal GPT-2: Efficient Adaptation via LoRA, Quantization, and Synthetic Data

Ryan D'Cunha [1,†] Ethan Hersch [1,†] Abhinav Chinta [1,†]

[1]Department of Computer Science, Stanford University — CS 224N: Natural Language Processing with Deep Learning, [†]Equal contribution

Problem

- Fine-tuning LLMs requires full parameter updates → **high compute + memory**
- Study efficient GPT-2 adaptation methods
- **Goals:** Reduce trainable parameters, reduce memory footprint, improve data efficiency

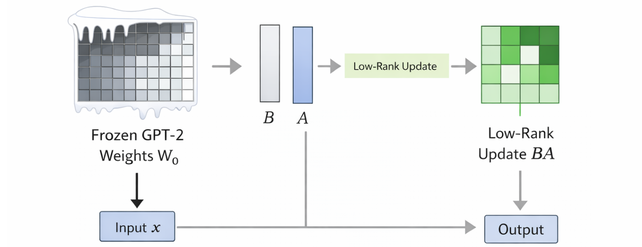
Approach

- **Parameter Efficiency**: Low Rank Adaptation (LoRA) for parameter updates
- **Quantization**: Lower weight precision (memory + speed) with training and inference quantization
- **Synthetic Distillation**: Generate synthetic training data from Gemini 2.5 Family

Background

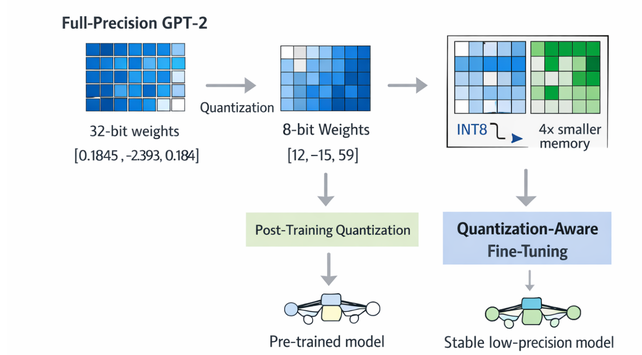
LoRA (Low-Rank Adaptation)

$$W = W_0 + B \cdot A$$

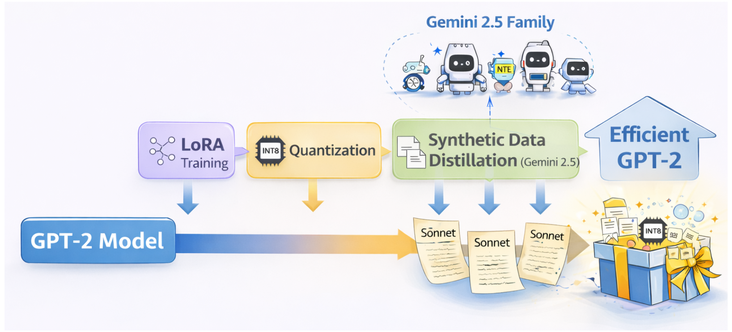


LoRA freezes pretrained weights W_0 and learns a low-rank update

Model Quantization



Synthetic Data Distillation



Model Overview and Baseline Experimentation

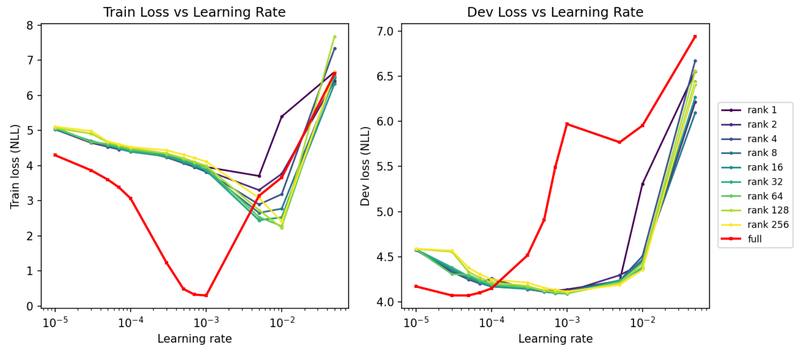
- **Model:** GPT-2 Small with 124M parameters and 12 transformer layers [1]
- **Downstream tasks:**
  - Sentiment classification (SST and CFIMDB)
  - Paraphrase detection (Quora)
  - Open-ended Shakespearean sonnet generation
- All extensions are applied to sonnet generation as it is the most complex task

Task (Dataset)	Fine-Tuning Method	Metric	Dev Score	Test Score
Sentiment (SST)	Last Linear Layer	Accuracy	0.487	0.476
	Full Fine-Tuned	Accuracy	**0.513**	**0.546**
Sentiment (CFIMDB)	Last Linear Layer	Accuracy	0.865	—
	Full Fine-Tuned	Accuracy	**0.971**	—
Paraphrase (Quora)	Full Fine-Tuned	Accuracy	**0.911**	**0.891**
Sonnet Generation	Full Fine-Tuned	chrF	41.974	41.078
Sonnet Generation	Best LoRA	chrF	42.158	—
Sonnet Generation	Best Quantization	chrF	42.118	—
Sonnet Generation	Best Data Augmentation	chrF	**46.605**	**52.838**

Parameter Efficiency: Low Rank Adaptation (LoRA)

Conduct hyperparameter sweeps across learning rates and modules (attention vs. MLP) [2]

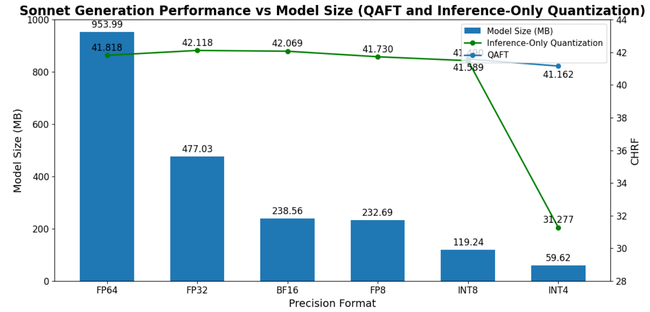
- Optimal sonnet generation achieved with rank 256, scaling factor $\alpha = 16$, and learning rate 1×10^{-2} which represents approximately $\frac{1}{3}$ of the full fine-tuning parameters
- Applying LoRA to all modules (Attention + MLP) yields a higher dev chrF score (42.158) compared to Attention-only (41.483)
- LoRA acts as an implicit regularizer; LoRA requires a far higher learning rate



Quantization Efficiency

Inference quantization and Quantization-Aware Fine-Tuning (QAFT) [3]

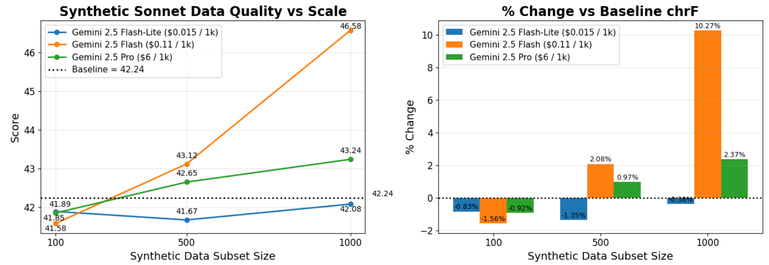
- BF16 and FP8 formats halve model's memory footprint and maintain inference performance
- INT4 reduces model size but suffers a ∼20% performance drop without specialized training
- QAFT at INT4 and INT8 successfully retains sonnet generation quality



Data Efficiency: Synthetic Data Augmentation

To measure distillation capacity and data efficiency, prompt Gemini 2.5 family (Flash Lite, Flash, and Pro) to generate up to 1,000 Shakespearean sonnets and fine-tune [4]

- Distillation from Gemini 2.5 Flash significantly improves GPT-2's sonnet generation chrF score from 42.24 to 46.60
- Distilling from Gemini 2.5 Pro causes performance to plateau, suggesting GPT-2 lacks the capacity to imitate the more complex teacher model



Conclusion

- LoRA can match full fine-tuned performance in sonnet generation while reducing trainable parameters and mitigating overfitting
- QAFT minimizes memory footprint but risks task-specific overfitting
- Augmenting training with high-quality synthetic data improves performance, though student capacity limits restrict full knowledge transfer from frontier models

References

[1] Alec Radford et al. Language models are unsupervised multitask learners. *OpenAI blog*, 2019.
[2] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.
[3] Quan Wei et al. Roste: An efficient quantization-aware supervised fine-tuning approach for large language models, 2025.
[4] Anup Shirgaonkar, Nikhil Pandey, Nazmiye Ceren Abay, Tolga Aktas, and Vijay Aski. Knowledge distillation using frontier open-source llms: Generalizability and the role of synthetic data, 2024.