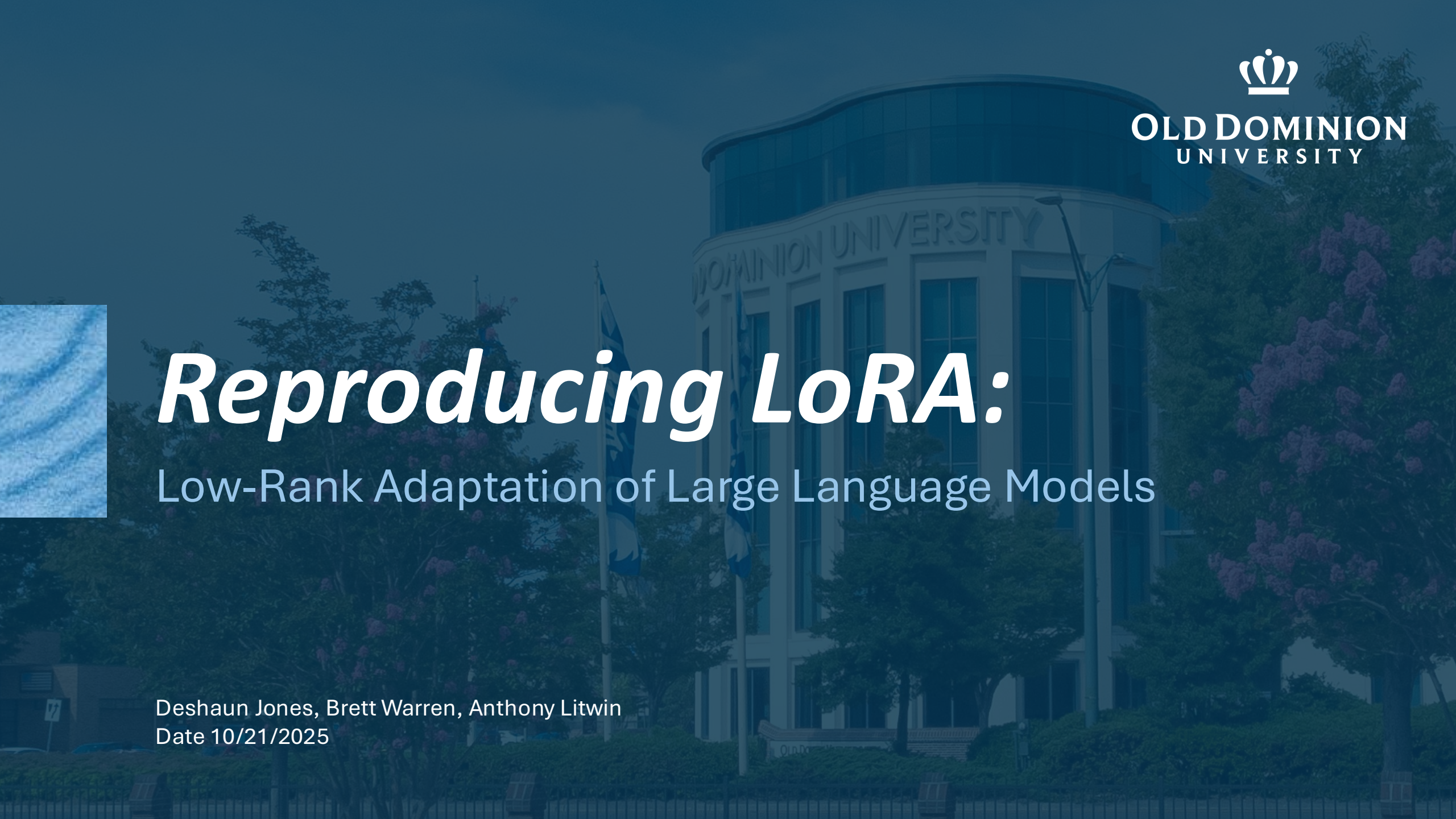


OLD DOMINION
UNIVERSITY

Reproducing LoRA:

Low-Rank Adaptation of Large Language Models

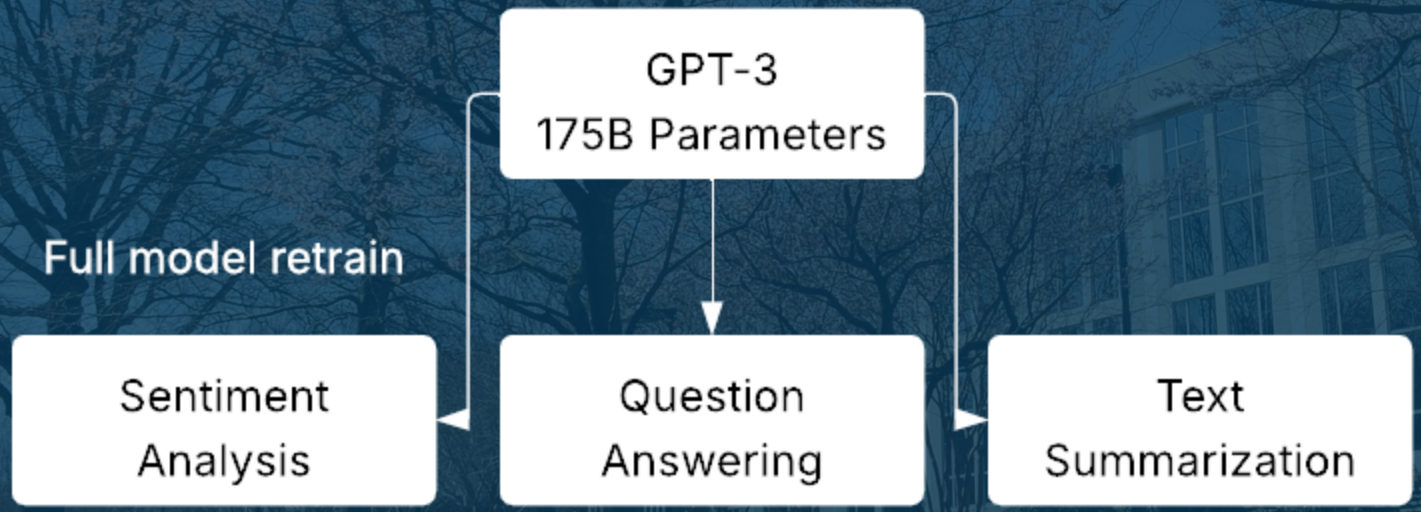
Deshaun Jones, Brett Warren, Anthony Litwin
Date 10/21/2025

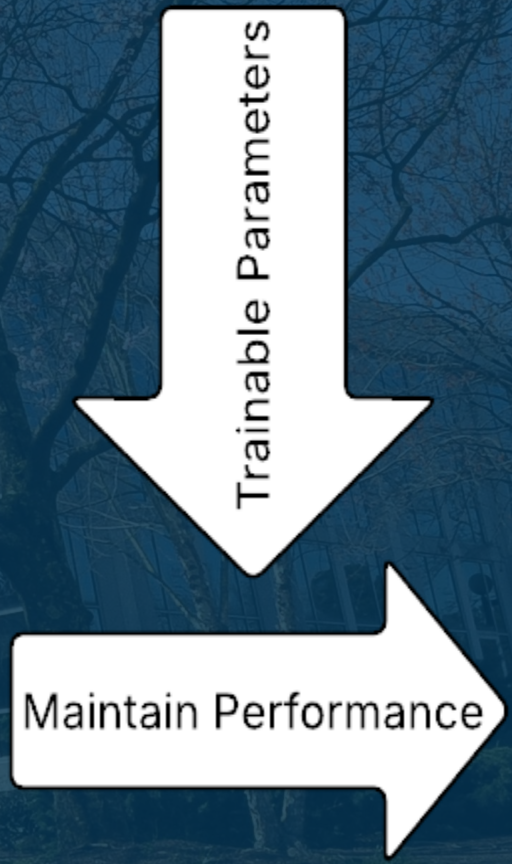
Overview

1

Problem statement and motivation

- Fine-tuning large pre-trained models is costly
- Existing "parameter-efficient" methods increase inference latency and underperform





2

Project Goals

- Replicate LoRA’s results (on smaller scale)
- Quantify efficiency gains
- Compare LoRA against others
- Explore effects
- Document reproducibility and deviations

3

Prior approaches and their limitations

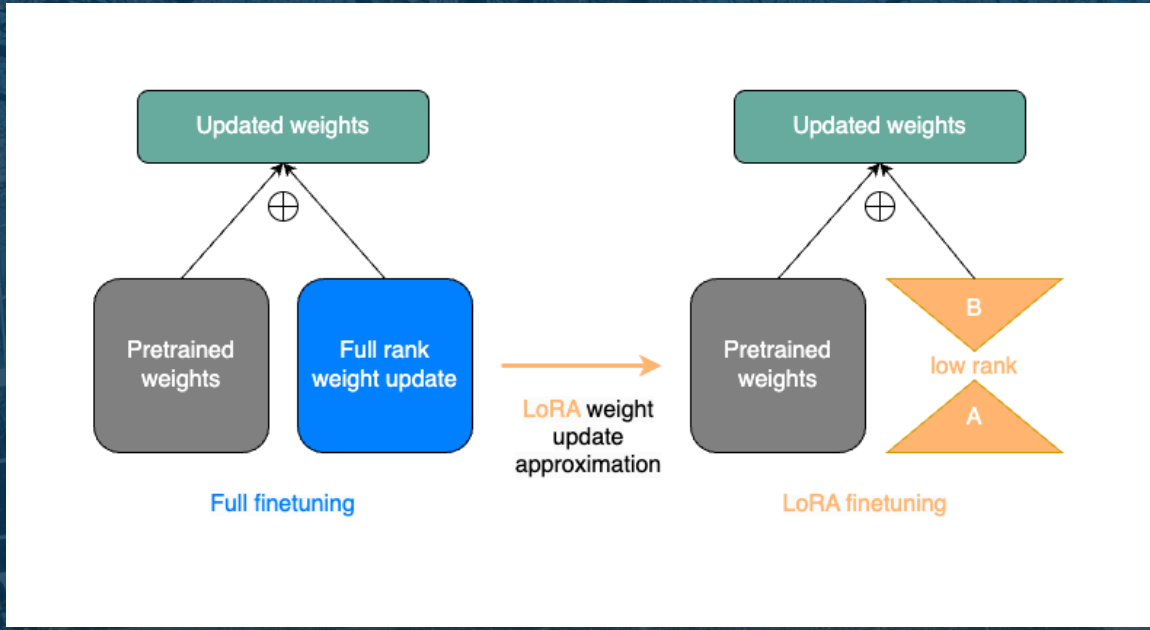
- Reduction of trainable parameters with improved performance
- Large models saw reduction of:
  - ~10,000x fewer parameters
  - ~3x less GPU memory
- Proved feasible to freeze pre-trained model
- Practical efficient pathway to deployment

Model & Method	# Trainable Parameters	BLEU
GPT-2 M (FT)*	354.92M	68.2
GPT-2 M (Adapter^{L})*	0.37M	66.3
GPT-2 M (Adapter^{L})*	11.09M	68.9
GPT-2 M (Adapter^{H})	11.09M	$67.3_{\pm.6}$
GPT-2 M (FT^{Top2})*	25.19M	68.1
GPT-2 M (PreLayer)*	0.35M	69.7
GPT-2 M (LoRA)	0.35M	$\mathbf{70.4}_{\pm.1}$
GPT-2 L (FT)*	774.03M	68.5
GPT-2 L (Adapter^{L})	0.88M	$69.1_{\pm.1}$
GPT-2 L (Adapter^{L})	23.00M	$68.9_{\pm.3}$
GPT-2 L (PreLayer)*	0.77M	70.3
GPT-2 L (LoRA)	0.77M	$\mathbf{70.4}_{\pm.1}$

4

Proposed method and why it meets the goals

- Pre-trained weights: $W_0 \in \mathbb{R}^{d \times k}$
- Full Fine-Tuning Update: $\mathrm{W}' = \mathrm{W}_0 + \Delta W$
- LoRA Update: $\Delta W = BA$
  - $B \in \mathbb{R}^{d \times r}$
  - $A \in \mathbb{R}^{r \times k}$
- Full-Rank Update: $d \times k \; parameters$
- LoRA Update: $r \times (d + k) \; parameters$



5
Experiments

- Models & Tasks
  - RoBERTa-base on MNLI & MRPC
  - GPT-2 S/M on E2E NLG
- Methods Compared
  - Full fine-tuning
  - LoRA (r = 1, 4, 8, 16; α fixed)
  - BitFit
- Evaluation
  - Accuracy (GLUE)
  - BLEU / METEOR / ROUGE-L (E2E)
  - Training parameter count & runtime
- Compute
  - All training done on ODU HPC

6 Results

Classification (GLUE)				
Task	**Full FT**	**Best LoRA**	**Difference**	**Parameters Trained**
MNLI	87.76%	86.82% (r=8)	-0.9%	0.7% of FT
MRPC	91.17%	87.99% (r=8)	-3.2%	0.7% of FT

Generation (E2E NLG)			
Model	**Metric**	**Full FT**	**Best LoRA**
GPT-2 Small	BLEU	0.135	0.133 (r=16)
GPT-2 Medium	METEOR	0.554	0.567 (r=16)
	ROUGE-L	0.326	0.343 (r=16)
	BLEU	0.129	0.143 (r=1)
	METEOR	0.544	0.571 (r=1)
	ROUGE-L	0.313	0.359 (r=1)

7

Conclusion

- Our results only moderately matched the original LoRA performance
- LoRA’s wide adoption and many extensions show strong confidence in the method
- Reproduction gaps likely came from our constraints
  - Limited seeds, fewer epochs, hardware limits, and possible config issues
- Modern industry trends reinforce LoRA’s continued relevance

8

Future Work

- Debug the E2E NLG evaluation pipeline
- Increase random seeds and expand ablation studies
- Study effects of LoRA rank, α, learning rate, and sequence length
- Implement a minimal LoRA module from scratch for deeper insight
- Test LoRA in real task-switching scenarios for practical evaluation

OLD DOMINION
UNIVERSITY

THANK YOU