

Compress to Impress: Efficient LLM Adaptation Using a Single Gradient Step on 100 Samples

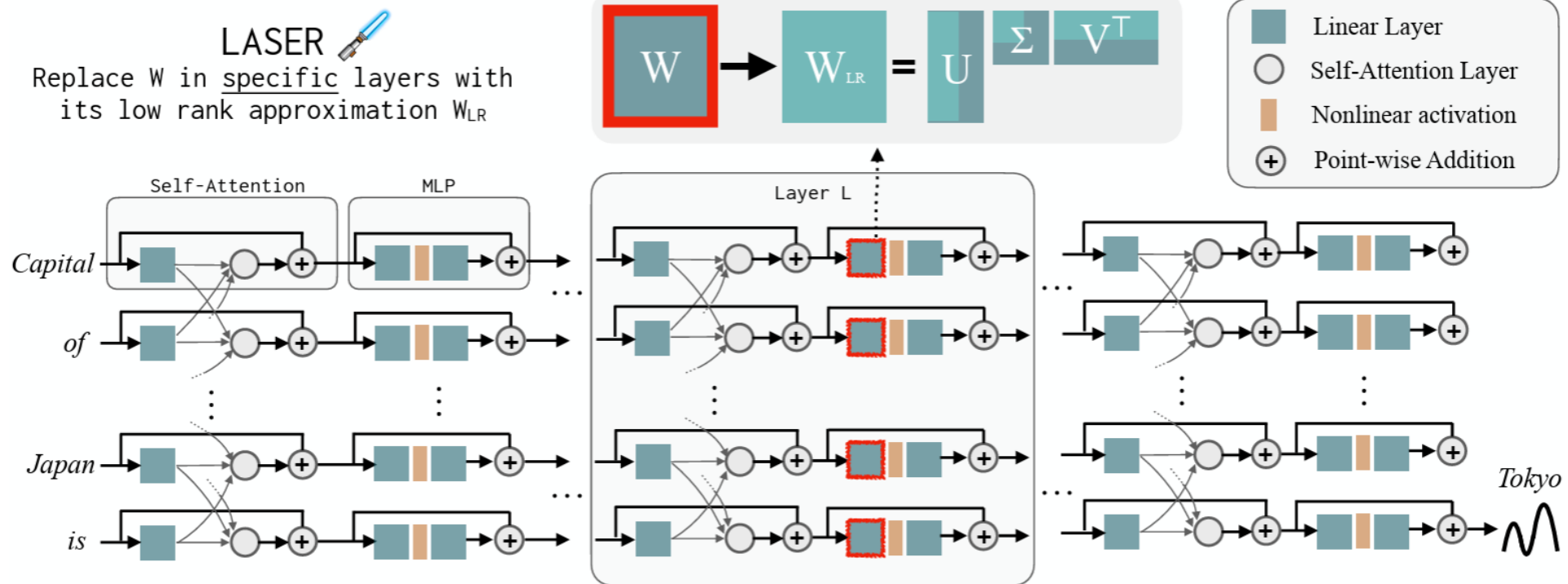
SHIVA SREERAM¹, ALAA MAALOUF^{1,2}, PRATYUSHA SHARMA¹, DANIELA RUS¹

SASREERA@MIT.EDU, ALAAMALOUF12@GMAIL.COM, PRATYUSS@MIT.EDU, RUS@CSAIL.MIT.EDU

¹MIT, COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE LABORATORY

²UNIVERSITY OF HAIFA, DEPARTMENT OF COMPUTER SCIENCE

LAyer-SElective Rank reduction*

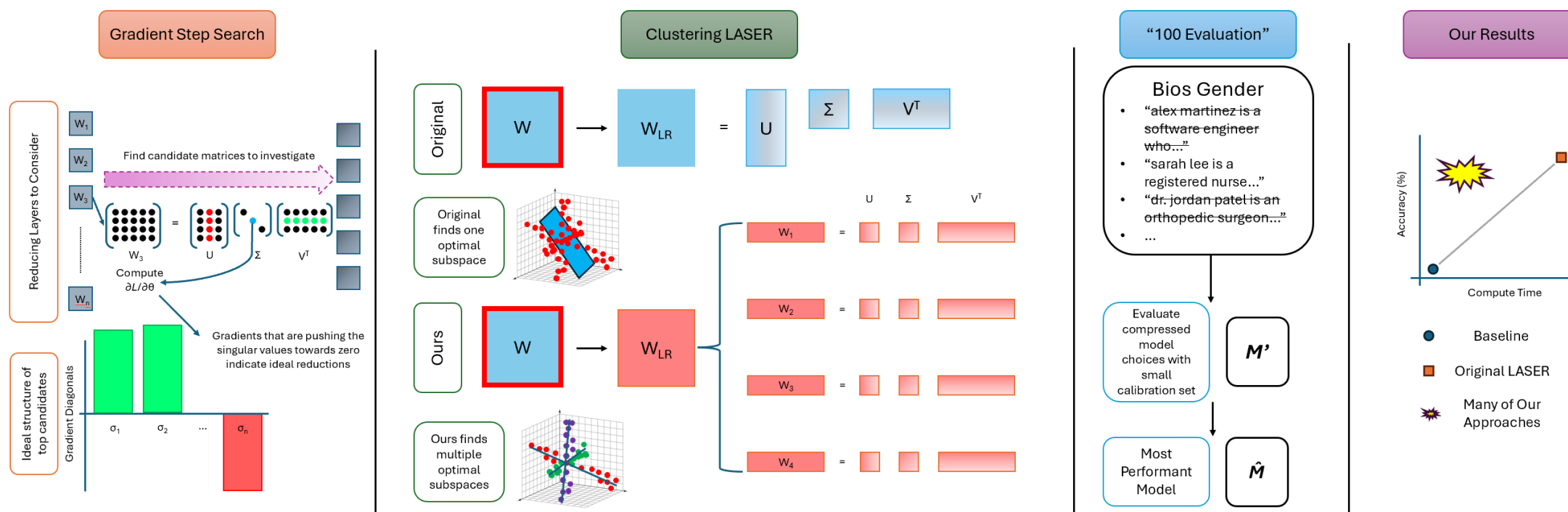


*Pratyusha Sharma, Jordan T Ash, and Dipendra Misra. The truth is in there: Improving reasoning in language models with layer-selective rank reduction. In *The Twelfth International Conference on Learning Representations*. 2024.

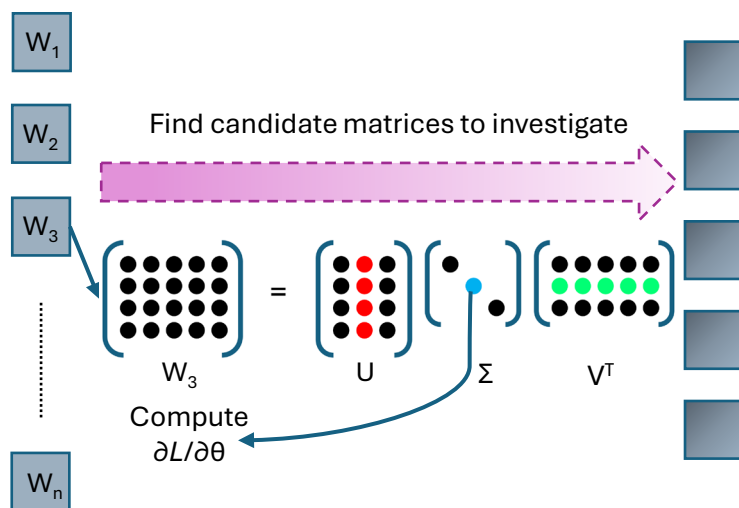
Primary Shortcomings of LASER

- High computation time largely due to searching across entire model
- Limited to considering one subspace for fitting noise

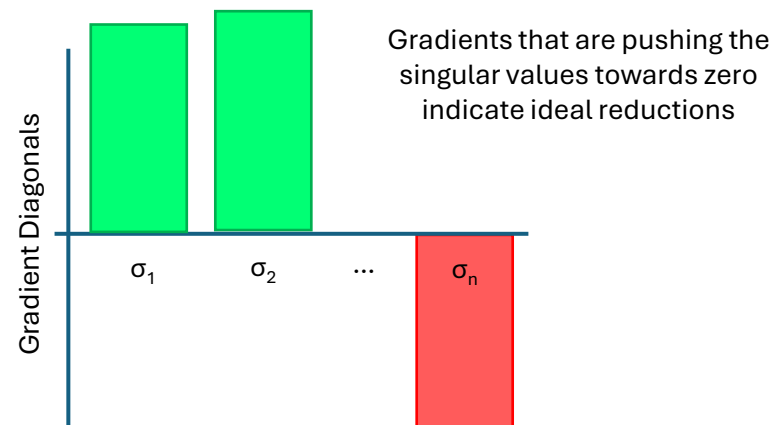
Introducing Efficient LLM Adaptation



Gradient Step Search



Reducing Layers to Consider

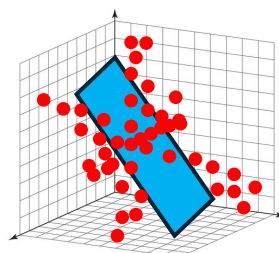


Ideal structure of top candidates

Original LASER Approach

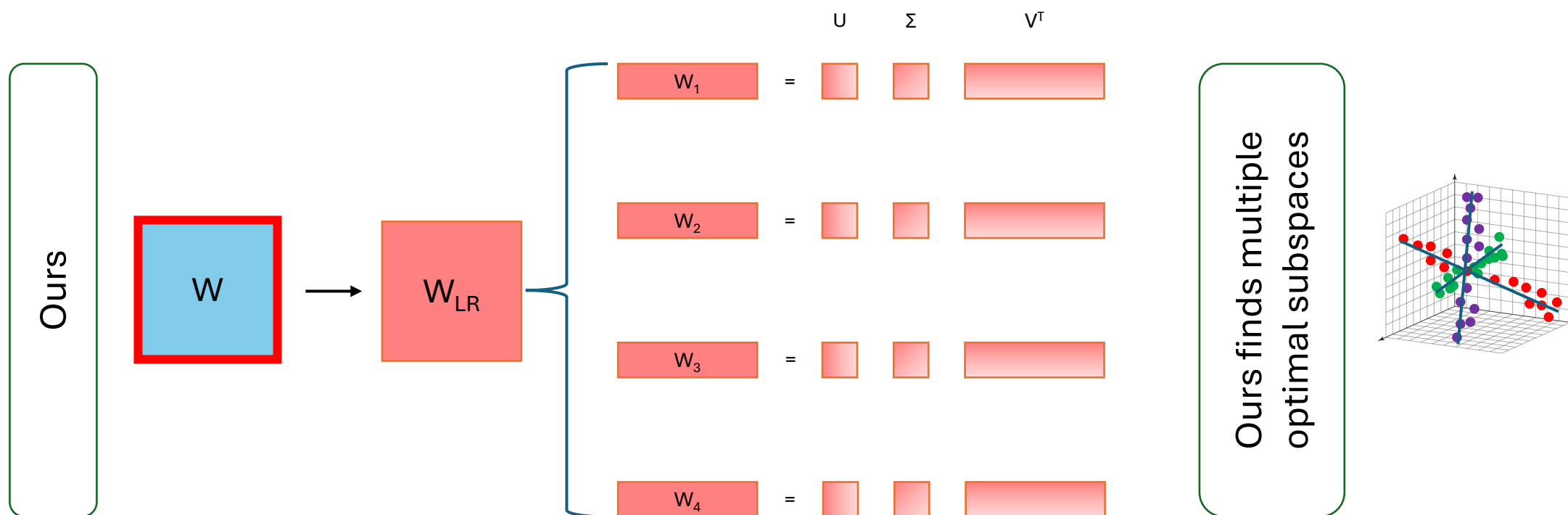
$$\boxed{W} \rightarrow W_{LR} = U \Sigma V^T$$

Original



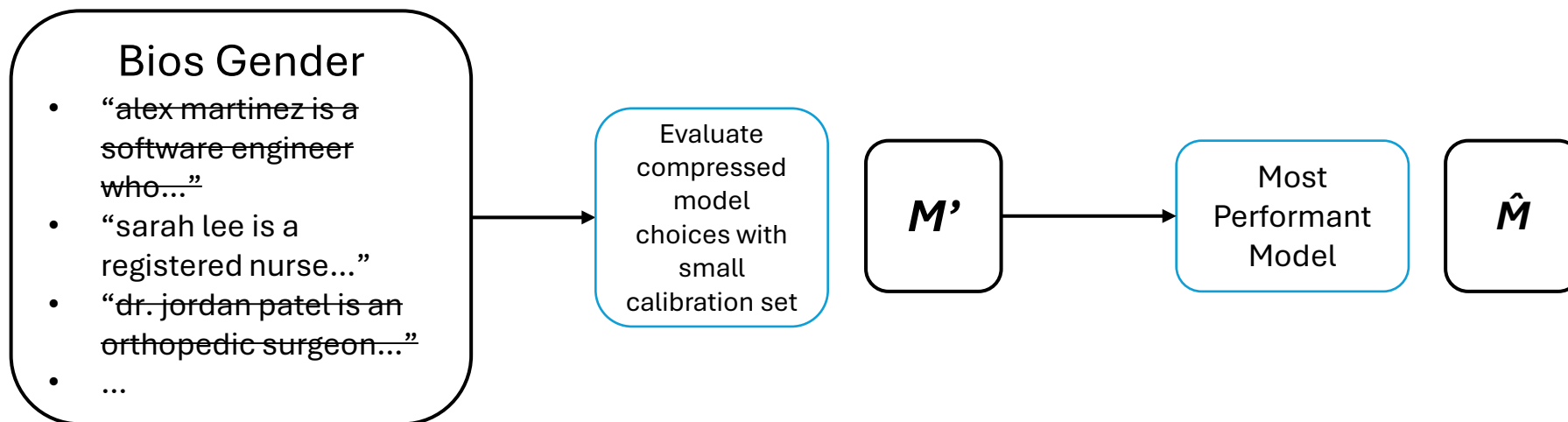
Original finds one optimal subspace

Clustering LASER Approach



“100” Evaluation

- When performing our search with just 100 examples from the dataset, we can still obtain a very performant model



Efficient LLM Adaptation on GPT-J

Dataset	Baseline	LASER	Clustering LASER 100 Grads Std Eval (ours)		Clustering LASER 100 Grads 100 Eval (ours)	
			Acc	Speedup	Acc	Speedup
CounterFact	13.1	24.0	24.4	1.98x	24.2	<u>93.4x</u>
HotPotQA	19.6	19.5	19.9	1.98x	19.7	48.3x
FEVER	50.2	56.2	56.0	1.96x	53.3	44.7x
Bios Gender	70.9	97.5	88.4	1.98x	88.4	<u>79.4x</u>
Bios Profession	75.6	82.1	80.5	1.98x	77.5	56.8x
TruthfulQA	54.9	55.6	56.1	1.97x	54.9	25.2x
BigBench–Epistemic Reasoning	37.1	38.3	62.3	1.96x	62.2	9.84x
BigBench–WikidataQA	51.8	65.9	66.5	1.98x	66.5	58.5x
Average Improvement from Baseline	0.00	8.24		10.1		9.19
Average Change from LASER	-8.24	0.00		1.85		0.95
Average Speedup	–	–		1.97x		52.0x

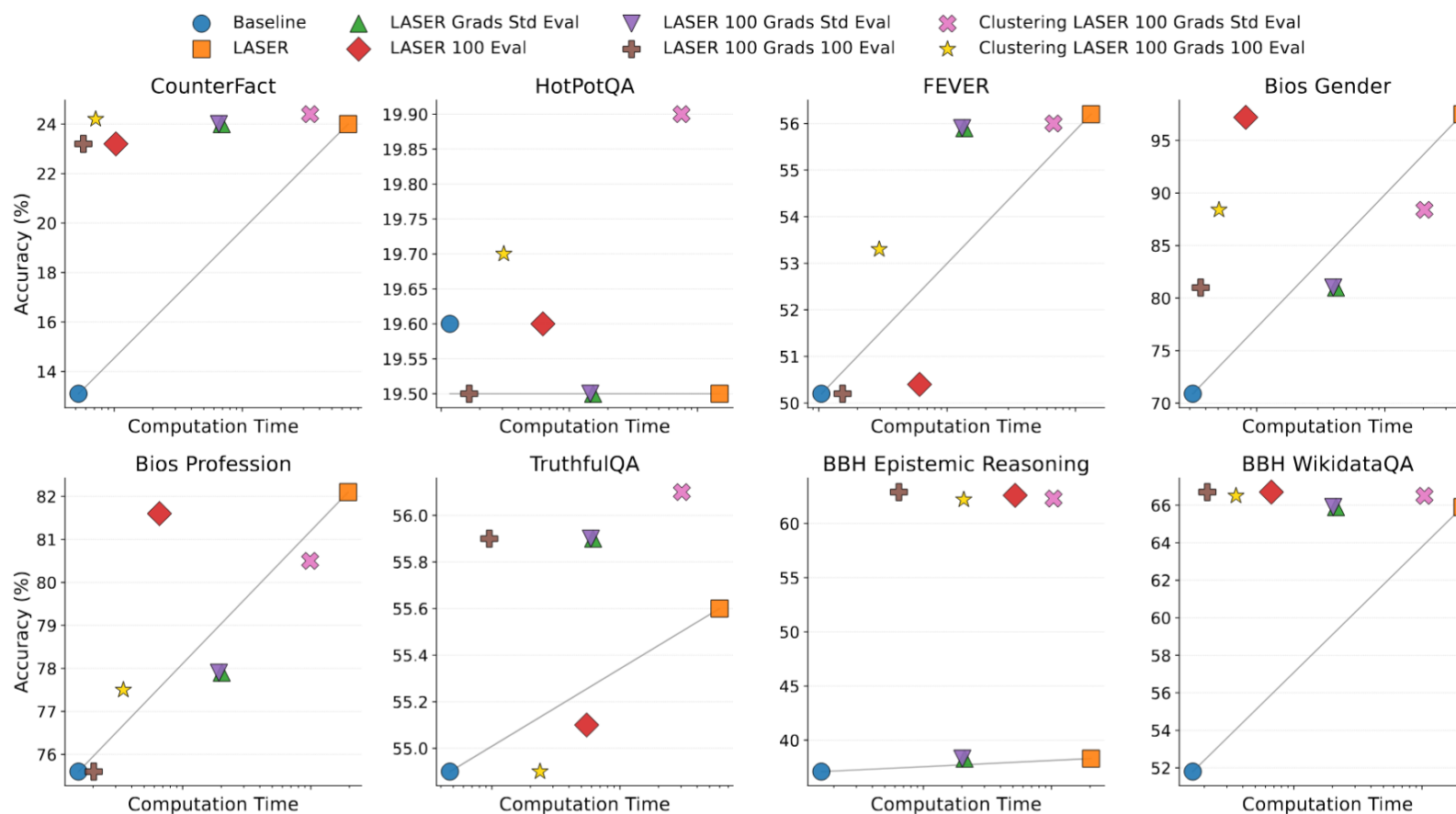
Efficient LLM Adaptation on Roberta

Dataset	Baseline	LASER	Clustering LASER 100 Grads Std Eval (ours)		Clustering LASER 100 Grads 100 Eval (ours)	
			Acc	Speedup	Acc	Speedup
CounterFact	17.3	19.3	19.3	0.86x	18.3	<u>36.8x</u>
HotPotQA	6.1	6.7	6.5	0.86x	6.3	17.0x
FEVER	50.0	52.3	52.7	0.86x	52.7	15.7x
Bios Gender	87.5	93.7	93.1	0.86x	92.8	<u>30.2x</u>
Bios Profession	64.5	72.5	75.1	0.86x	75.1	20.4x
TruthfulQA	56.2	56.2	56.3	0.86x	56.2	8.39x
BigBench–Epistemic Reasoning	37.1	41.8	37.2	0.85x	37.1	3.17x
BigBench–WikidataQA	28.0	30.7	32.7	0.86x	31.5	21.1x
Average Improvement from Baseline	0.00	3.31		3.27		2.91
Average Change from LASER	-3.31	0.00		-0.04		-0.40
Average Speedup	–	–		0.86x		22.2x

Efficient Technique Study with GPT-J

Dataset	Baseline	LASER	LASER Grads Std Eval (ours)		LASER 100 Eval (ours)		LASER 100 Grads Std Eval (ours)		LASER 100 Grads 100 Eval (ours)	
			Acc	Speedup	Acc	Speedup	Acc	Speedup	Acc	Speedup
CounterFact	13.1	24.0	24.0	9.70x	23.2	64.9x	24.0	10.2x	23.2	<u>116.5x</u>
HotPotQA	19.6	19.5	19.5	9.70x	19.6	23.9x	19.5	10.2x	19.5	90.0x
FEVER	50.2	56.2	55.9	9.70x	50.4	21.7x	55.9	10.1x	50.2	86.3x
Bios Gender	70.9	97.5	81.0	9.70x	97.2	49.1x	81.0	10.2x	81.0	<u>110.4x</u>
Bios Profession	75.6	82.1	77.9	9.70x	81.6	29.7x	77.9	10.2x	75.6	96.7x
TruthfulQA	54.9	55.6	55.9	9.70x	55.1	10.9x	55.9	10.1x	55.9	62.7x
BigBench–Epistemic Reasoning	37.1	38.3	38.3	9.70x	62.6	3.91x	38.3	10.1x	62.9	31.6x
BigBench–WikidataQA	51.8	65.9	65.9	9.70x	66.7	31.0x	65.9	10.2x	66.7	98.0x
Average Improvement from Baseline	0.00	8.24	5.65		10.4		5.65		7.73	
Average Change from LASER	-8.24	0.00	-2.59		2.16		-2.59		-0.51	
Average Speedup	–	–	9.70x		29.4x		10.2x		86.5x	

Efficient Technique Study with GPT-J



Clustering Technique Study

Dataset	Roberta			GPT-J		
	Baseline	LASER	Clustering LASER	Baseline	LASER	Clustering LASER
CounterFact	17.3	19.3	19.3	13.1	24.0	24.5
HotPotQA	6.1	6.7	6.8	19.6	19.5	20.3
FEVER	50.0	52.3	52.7	50.2	56.2	57.8
Bios Gender	87.5	93.7	93.7	70.9	97.5	97.7
Bios Profession	64.5	72.5	75.1	75.6	82.1	82.3
TruthfulQA	56.2	56.2	56.3	54.9	55.6	56.1
BigBench–Epistemic Reasoning	37.1	41.8	41.8	37.1	38.3	62.9
BigBench–WikidataQA	28.0	30.7	36.7	51.8	65.9	66.5
Average Improvement from Baseline	0.00	3.31	4.46	0.00	8.24	11.9
Average Change from LASER	-3.31	0.00	1.15	-8.24	0.00	3.63

Summary

Our work contributes the following:

- Gradient-guided matrix selection
- Sample-efficient evaluation
- Multi-subspace factorization
- Adapting LLMs

Thank you.