Activated LoRA: Fine-tuned LLMs for Intrinsics

A concrete step towards Modular Language Models for Agents.

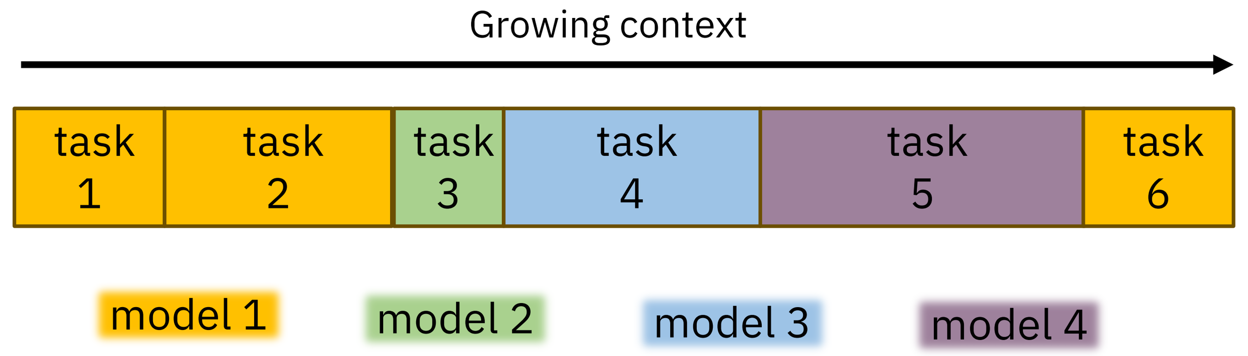
Kristjan Greenewald, Luis A. Lastras, Thomas Parnell, Vraj Shah, Lucian Popa, Giulio Zizzo, Chulaka Gunasekara, Ambrish Rawat, David Cox

NEURAL INFORMATION PROCESSING SYSTEMS

IBM Research

Characteristics of modern agent language model workloads

- Dynamic task switching
- Multiple specialized skills, often realized with specific models or prompts.
- Very large contexts.



Today's agents can be very inefficient.

Commonplace for agents to spend minutes/hours accomplishing complex requests broken into multiple tasks.

Some of the reasons

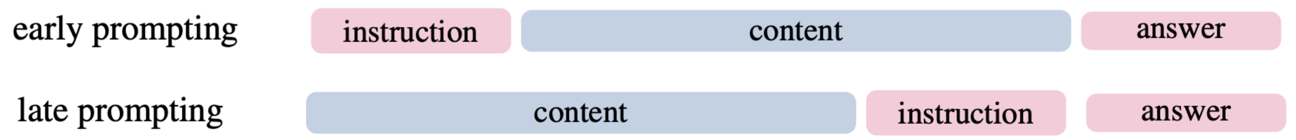
Using multiple models is inherently expensive memory and compute-wise.

Even using a single model can be expensive:

- The most common language models are autoregressive – they create representations of sequences of symbols that are *causal*.
- When you want to change something in that sequence early, the entire representation changes.
- Prompting the same model for different tasks can be very expensive for long contexts if the task specific prompt is introduced early.

The advantage of late prompts

Imagine we had a strong language model that could always get very high accuracies when prompts are added *after* data is presented to it, in other words, *late* instead of *early*.



Such a model would be an ideal agentic model – you can always tell the model what to do next while preserving its context representation intact.

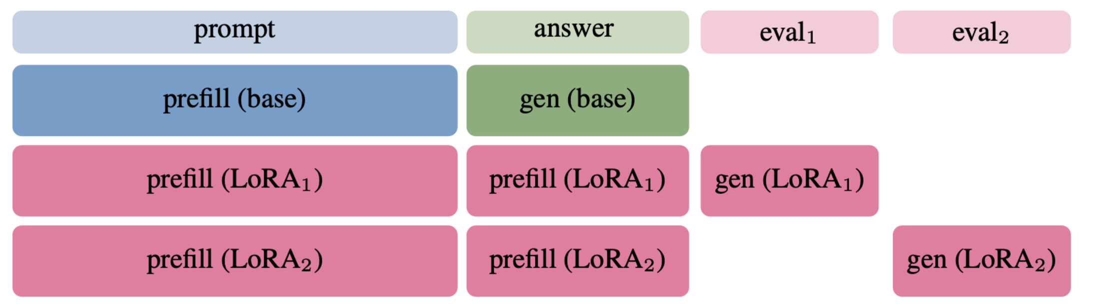
In practice, one does at least one of two things:

a) Fine tune models for higher accuracies
b) Introduce prompts early, to allow the model to create task specific representations.

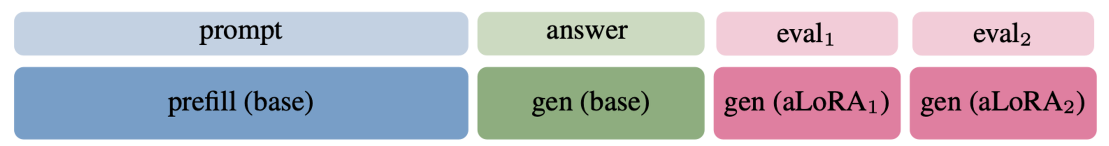
Activated Low Rank Adapters

Transformer based language models represent the context prior to generation using "keys and values".

Activated LoRA reuses the key/value representation of a context and only adapts tokens in the suffix of the sequence needed for a desired task specific generations.



Classic Low Rank Adapters [1]



Activated Low Rank Adapters

Hypothesis is that modern language models provide "universal-like" representations of contexts which can be adapted to different tasks.

Generic accuracy evaluation of activated LoRA

Task	Llama 3.2 1B		Llama 3.2 3B		Llama 3.1 8B		Mistral 7B	
	LoRA	aLoRA	LoRA	aLoRA	LoRA	aLoRA	LoRA	aLoRA
Bengali Hate Speech Classification	79.30%	81.94%	86.34%	89.43%	70.04%	85.02%	72.25%	85.46%
WIQA: Effect Classification	68.92%	71.38%	76.15%	76.00%	74.92%	78.00%	61.08%	79.08%
MMLU Conceptual Physics MCQA	33.33%	38.89%	72.20%	66.67%	55.56%	55.56%	55.56%	55.56%
MMLU College Computer Science MCQA	66.67%	58.33%	66.67%	75.00%	66.67%	58.33%	75.00%	75.00%
SocialIQA Question Generation	86.00%	88.77%	89.85%	90.15%	52.00%	88.92%	97.23%	90.92%
Hindi Sentence Perturbation	69.60%	74.69%	98.30%	63.89%	86.11%	35.19%	99.23%	96.30%
SuperGLUE Question Generation	98.42%	95.79%	95.26%	96.84%	98.95%	92.11%	99.47%	92.11%

LoRA and aLoRA accuracy on each task after hyperparameter grid search, guided by the validation set, on multiple random tasks from [2]

Accuracy Evaluation on tasks focused on Retrieval Augmented Generation and Safety

Uncertainty Quantification (how sure is the model of its own output?) [5]

Certainty Score	LoRA	aLoRA
MAE	0.50	0.49

Test error for the Uncertainty Quantification Intrinsic.

Jail Breaking (for detecting jailbreak risk within user prompts) [6]

	Acc	TPR	FPR
aLoRA	0.925	0.863	0.013
LoRA	0.943	0.898	0.011

Performance for jailbreak risk detectors

Answerability (can the documents be used to answer the question) [3,4]

Dataset	Adapter	Unans.			Ans.			Weighted F1
		P	R	F1	P	R	F1	
SQUADRUN Dev	LoRA	84.2	68.0	75.2	73.1	87.2	79.5	77.4
	aLoRA	83.0	81.1	82.0	81.4	83.3	82.4	82.2
MT-RAG Benchmark	LoRA	85.4	89.3	87.3	87.0	82.4	84.6	86.1
	aLoRA	85.8	89.1	87.4	86.8	83.0	84.9	86.2

Comparison of classification performance across the SQUADRUN and MT-RAG benchmarks

Query Rewrite (for multiturn retrieval) [3]

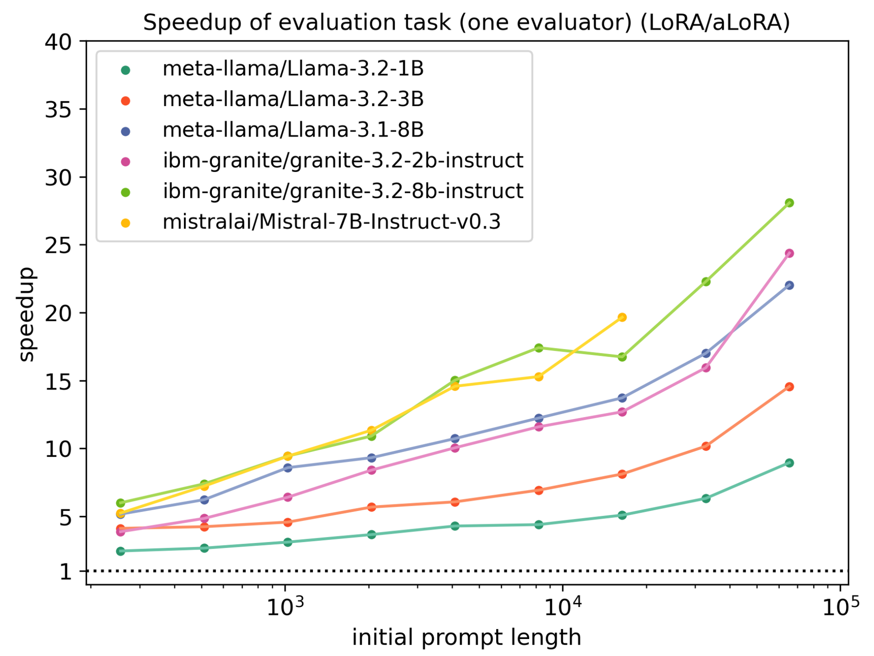
Strategy	Full MT-RAG			Non-standalone			Standalone		
	R@5	R@10	R@20	R@5	R@10	R@20	R@5	R@10	R@20
aLoRA	0.54	0.66	0.74	0.42	0.54	0.64	0.63	0.75	0.82
LoRA	0.56	0.68	0.76	0.44	0.57	0.66	0.63	0.75	0.83

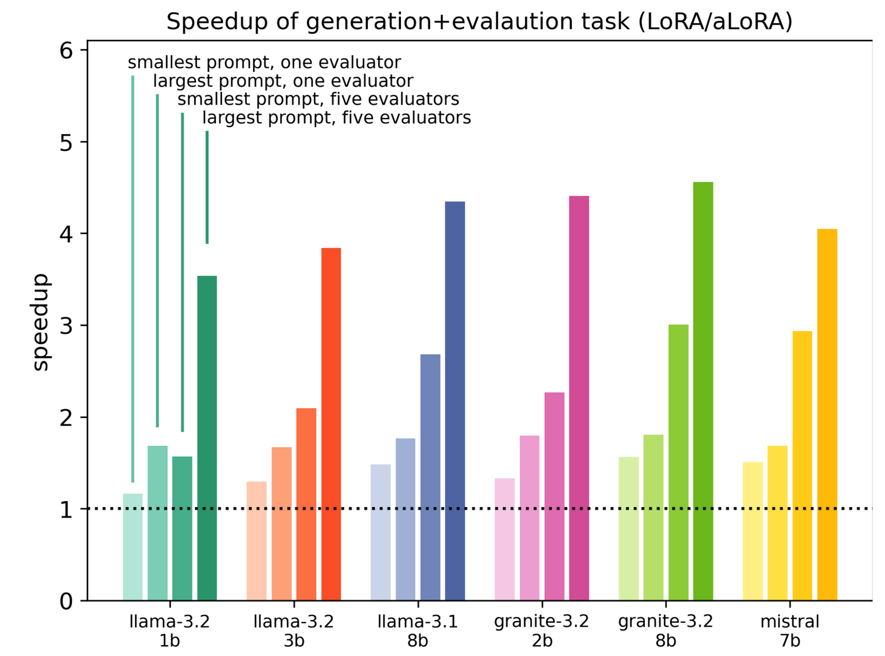
(a) Retrieval (Recall@5, @10, and @20)

Strategy	Full MT-RAG		Non-standalone		Standalone	
	RAGAS-F	RAD-Bench	RAGAS-F	RAD-Bench	RAGAS-F	RAD-Bench
aLoRA	0.81	0.69	0.77	0.69	0.83	0.70
LoRA	0.81	0.70	0.79	0.69	0.83	0.71

(b) Answer generation quality (RAGAS-F, RAD-Bench)

Performance evaluation of activated LoRA





Key takeaways

- Activated low rank adapters enable a language model to specialize instantaneously with no loss in inference performance and with the accuracy of classic LoRAs.
- Compared to LoRA, 10x-20x faster with the multiplicative advantage increasing for larger models and larger context lengths.
- 5 validated common agent safety and RAG tasks.
- Available in standard Huggingface PEFT, pull request in vLLM [7].

References

[1] Edward J Hu, et. al. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021

[2] Rickard Brüel-Gabrielsson, et. al. Compress then serve: Serving thousands of lora adapters with little overhead. arXiv preprint arXiv:2407.00066, 2024

[3] Marina Danilevsky et. al. A library of LLM intrinsics for retrieval-augmented generation. arXiv preprint arXiv:2504.11704, 2025.

[4] Yannis Katsis et. al. MTRAG: A multi-turn conversational benchmark for evaluating retrieval-augmented generation systems, 2025.https://arxiv.org/abs/2501.03468.

[5] Maohao Shen et. al. Thermometer: Towards universal calibration for large language models. In International Conference on Machine Learning, 2024.

[6] Ambrish Rawat, et. al. Attack atlas: A practitioner's perspective on challenges and pitfalls in red teaming GenAI. CoRR, abs/2409.15398, 2024

[7] Xu Zhu et al. vLLM: Fast inference of large language models. https://github.com/vllm-project/vllm, 2023