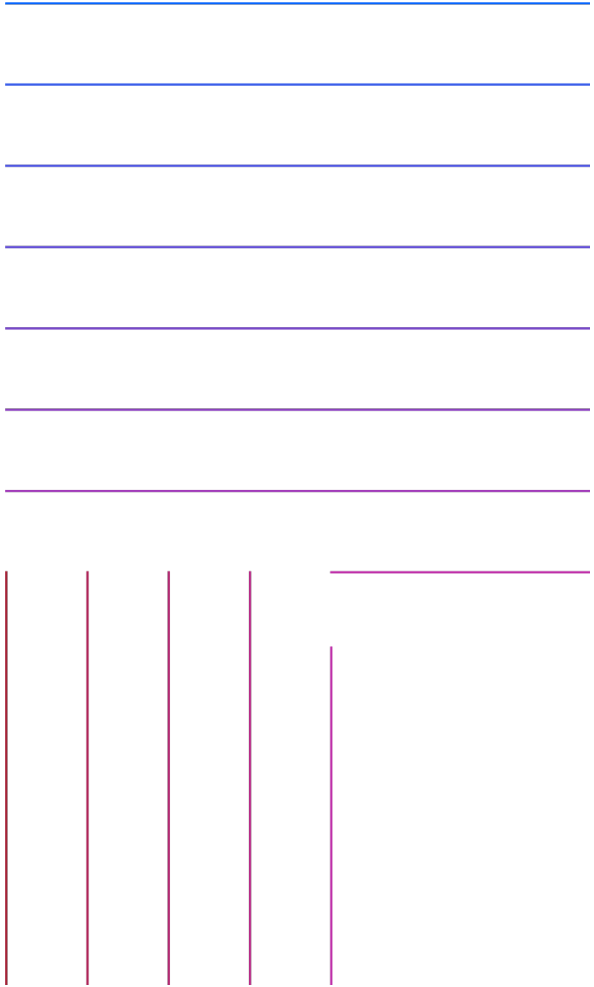
MIT-IBM
Watson AI Lab

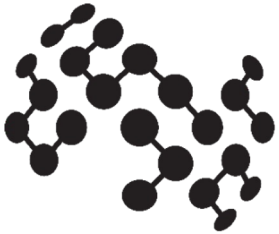
IBM **Research**

1000s of Customized LLMs: Challenges and Opportunities

Mikhail Yurochkin



I moved 🙂



Institute of Foundation Models

Mohamed bin Zayed University of Artificial Intelligence

Today's AI is Capable

ChatGPT passes exams from law and business schools

By Samantha Murphy Kelly, CNN Business

4 minute read · Updated 1:35 PM EST, Thu January 26, 2023

NEWS | 01 October 2024

'In awe': scientists impressed by latest ChatGPT model o1

The doctors pioneering the use of AI to improve outcomes for patients

Faster diagnostics, more targeted treatment and better communication are among areas of healthcare already benefiting from artificial intelligence

What is the Missing Piece?

Bacon-topped ice cream and other failures end McDonald's AI drive-through experiment

As AI tools become more integrated into healthcare, ethical concerns, governance complexities, and the cost of maintaining AI systems are significant hurdles that need to be addressed.

OpenAI Is Growing Fast and Burning Through Piles of Money

As the company looks for more outside investors, documents reviewed by The New York Times show consumer fascination with ChatGPT and a serious need for more cash.

What is the Missing Piece?

Bacon-topped ice cream and other failures end McDonald's AI drive-through experiment

Evaluation

As AI tools become more integrated into healthcare, ethical concerns, governance complexities, **and the cost of maintaining AI systems are significant hurdles that need to be addressed.**

OpenAI Is Growing Fast and Burning Through Piles of Money

As the company looks for more outside investors, documents reviewed by The New York Times show consumer fascination with ChatGPT and a serious need for more cash.

What is the Missing Piece?

Evaluation

100x cost reduction for benchmarks [ICML 2024]

Evaluation with 100s of prompt templates [NeurIPS 2024]

300k downloads lm-evaluation-harness 30k downloads promptbench

tinyBenchmarks Community

https://github.com/felipemaiapolo...

PromptEval

https://github.com/felipemaiapolo...

Readiness of LLMs for tasks in the economy

O*NET Database

Aug 30, 2024 — The O*NET Database contains hundreds of standardized and occupation-specific descriptors on almost 1,000 occupations covering the entire U.S. economy. The database, which is ...

Efficiency

LLM Routing: choosing the right LLM for the task reduces cost and improves performance [ICLR 2024, COLM 2024]

Serving 1000s of Customized LLMs

The Challenge

Serving 1000s of customized LLMs as individual models is too expensive...

We need to consider the customization mechanism:

$$\text{LoRA: } W = W_0 + BA$$

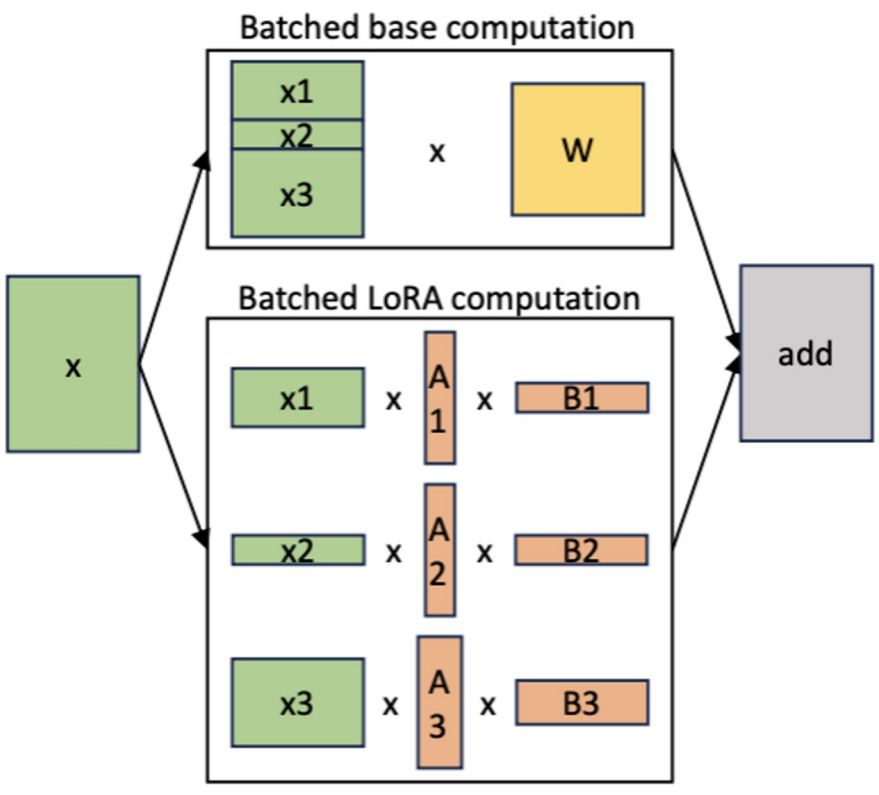
Frozen LLM

Trained Adapter

Careful System Design

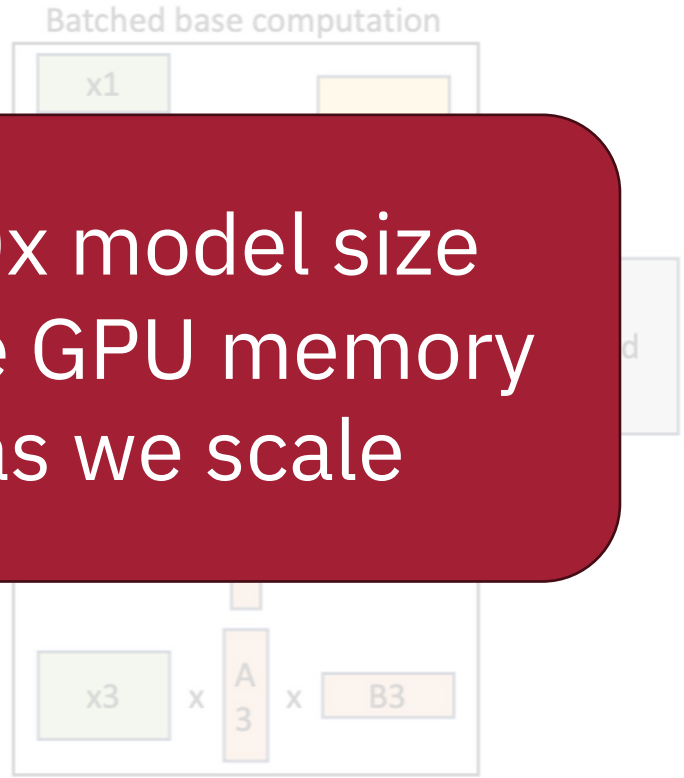
S-LoRA (Sheng et al. 2024):

Add adapters *on-the-fly* during serving



… is good, but not enough

- 1000 LoRAs is still ~10x model size
- We can’t fit them in the GPU memory
- Throughput degrades as we scale



Considering Collection of LoRAs

$$\{(A_i, B_i)\}_{i=1}^{n}$$

- Is there redundancy in a large *collection?*
- Is there room for *approximation* error?
- Can we serve 1000+ LoRAs as efficiently as a *single* LLM?

Step 1: Train Lots of LoRAs

Lots of LoRAs Community

RickardGabriels bruel-gabrielsson

Models 1268

Sort: Recently updated

Lots-of-LoRAs/Mistral-7B-Instruct-v0.2-4b-r16-task5… Updated about 1 month ago	Lots-of-LoRAs/Mistral-7B-Instruct-v0.2-4b-r16-task9… Updated about 1 month ago
Lots-of-LoRAs/Mistral-7B-Instruct-v0.2-4b-r16-task1… Updated about 1 month ago	Lots-of-LoRAs/Mistral-7B-Instruct-v0.2-4b-r16-task1… Updated about 1 month ago
Lots-of-LoRAs/Mistral-7B-Instruct-v0.2-4b-r16-task1… Updated about 1 month ago	Lots-of-LoRAs/Mistral-7B-Instruct-v0.2-4b-r16-task1… Updated about 1 month ago
Lots-of-LoRAs/Mistral-7B-Instruct-v0.2-4b-r16-task7… Updated about 1 month ago	Lots-of-LoRAs/Mistral-7B-Instruct-v0.2-4b-r16-task1… Updated about 1 month ago
Lots-of-LoRAs/Mistral-7B-Instruct-v0.2-4b-r16-task2… Updated about 1 month ago	Lots-of-LoRAs/Mistral-7B-Instruct-v0.2-4b-r16-task1… Updated about 1 month ago

Expand 1268 models

GPU-Friendly Compression

$$\{(A_i, B_i)\}_{i=1}^{n}$$

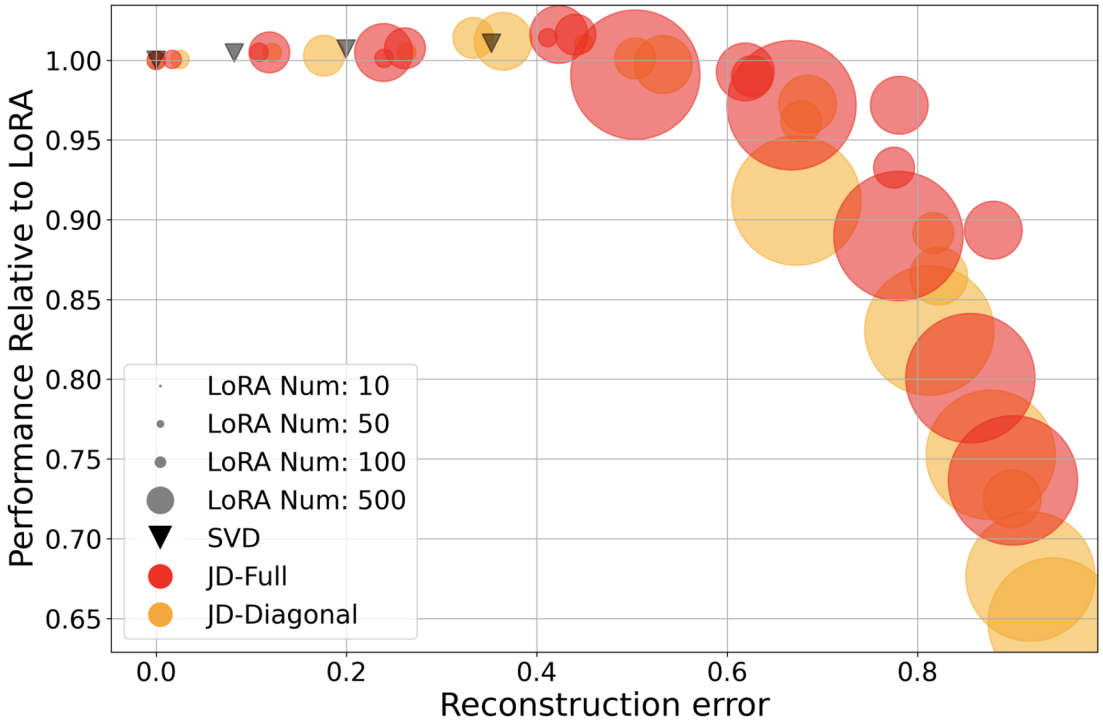
$$\text{JD-Full}_r(\{B_iA_i\}_{i=1}^{n}) = \underset{\substack{\{\Sigma_i\}_{i=1}^{n} \\ U^\top U = V^\top V = I_r}}{\text{argmin}} \sum_{i=1}^{n} \|B_iA_i - U\Sigma_iV^\top\|_{\text{Fro}}^2$$

shared

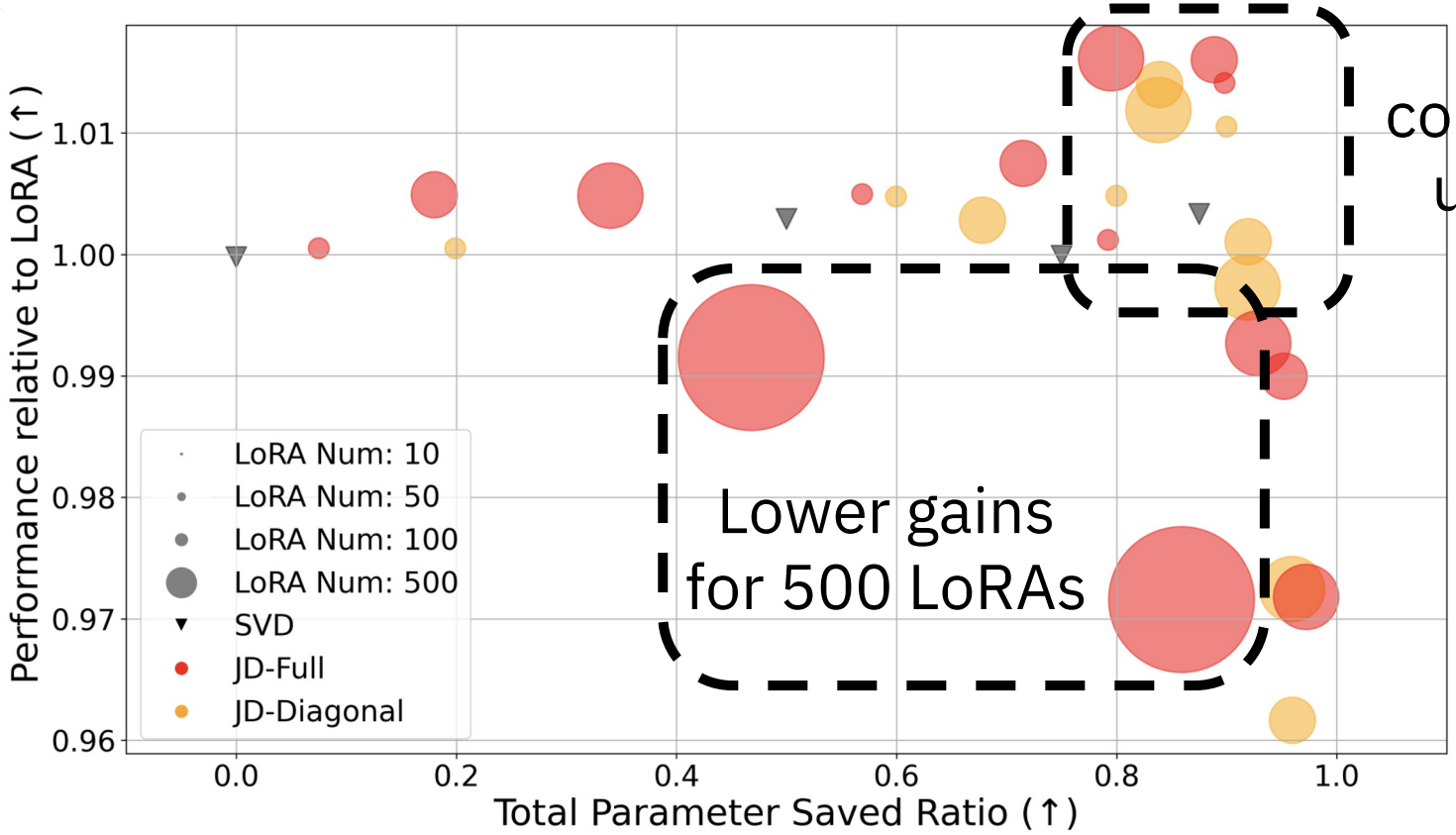
LoRA-specific

Room for Approximation

Up to 50% reconstruction error preserves the performance



Compression vs Performance



80%+ compression up to 100 LoRAs

Clustering Extension

Initialization: Split LoRAs into K clusters

Step 1: Using JD algorithm find U_k, V_k for each cluster C_k *independently*:

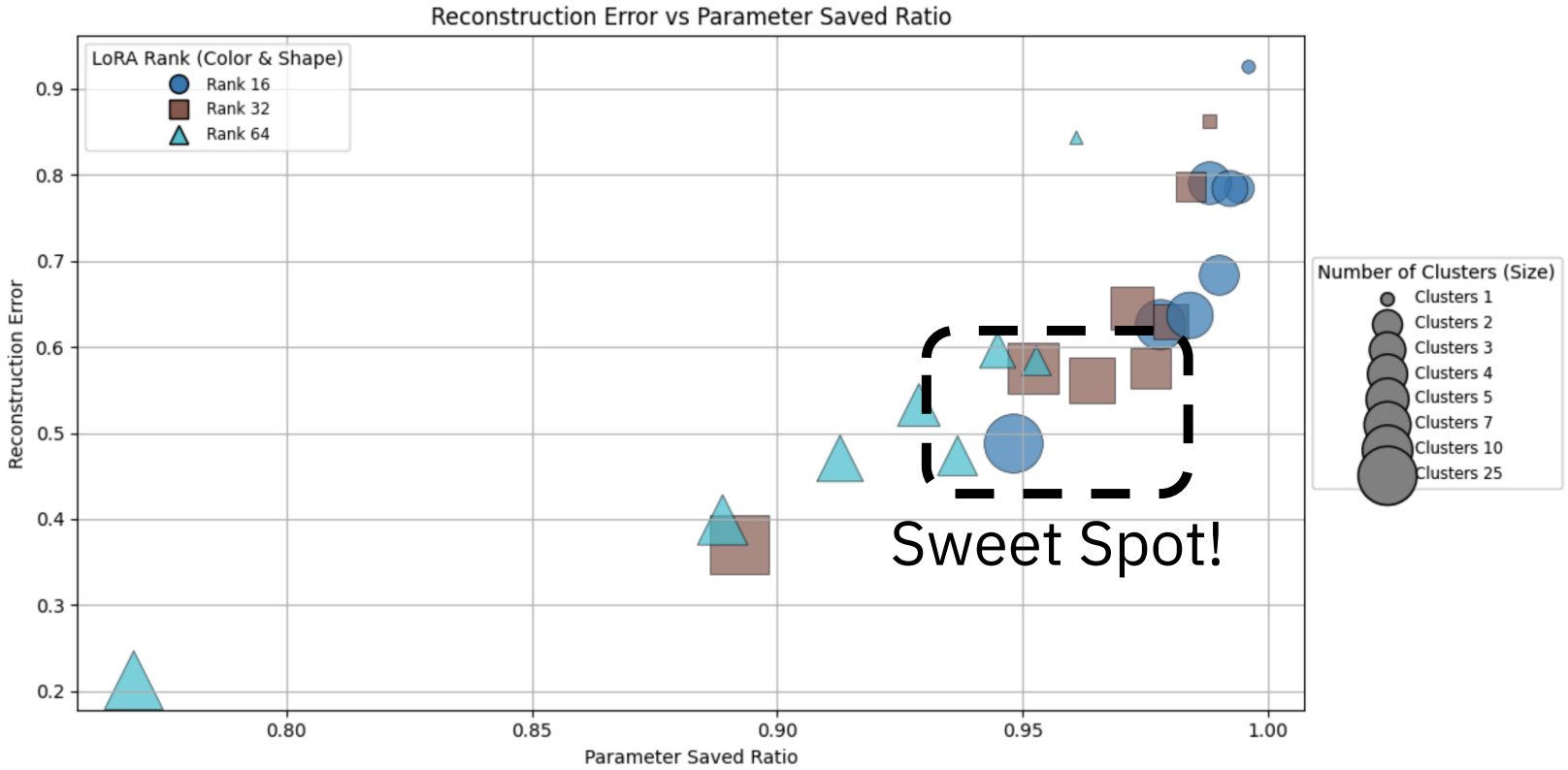
$$\min_{U_k, V_k, \Sigma_i} \sum_{i \in C_k} ||B_i A_i - U_k \Sigma_i V_k^\top||_{\text{Fro}}^2$$

Step 2: New cluster assignment for each LoRA i:

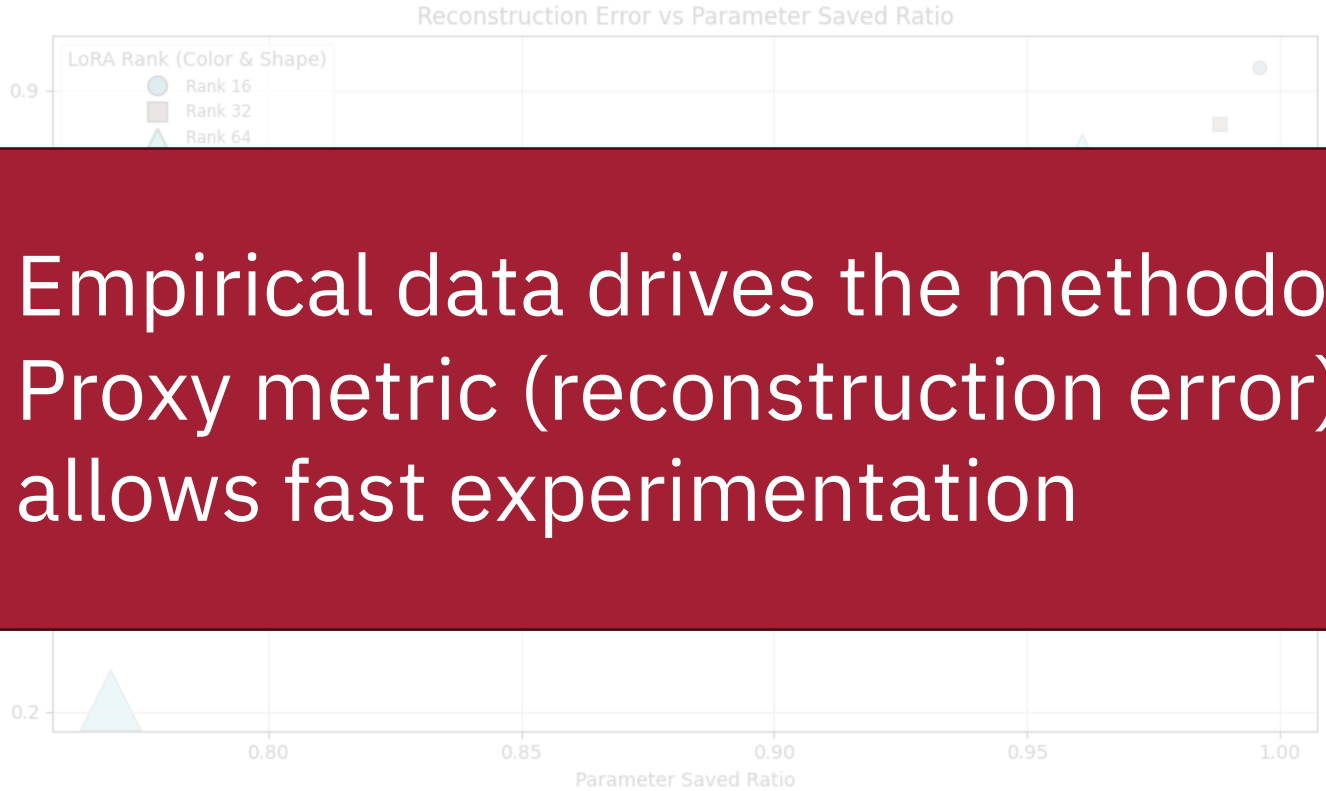
$$\min_k \min_{\Sigma_i} ||B_i A_i - U_k \Sigma_i V_k^\top||_{\text{Fro}}^2$$

Repeat Steps 1 and 2 until convergence

Revisiting 500 LoRAs

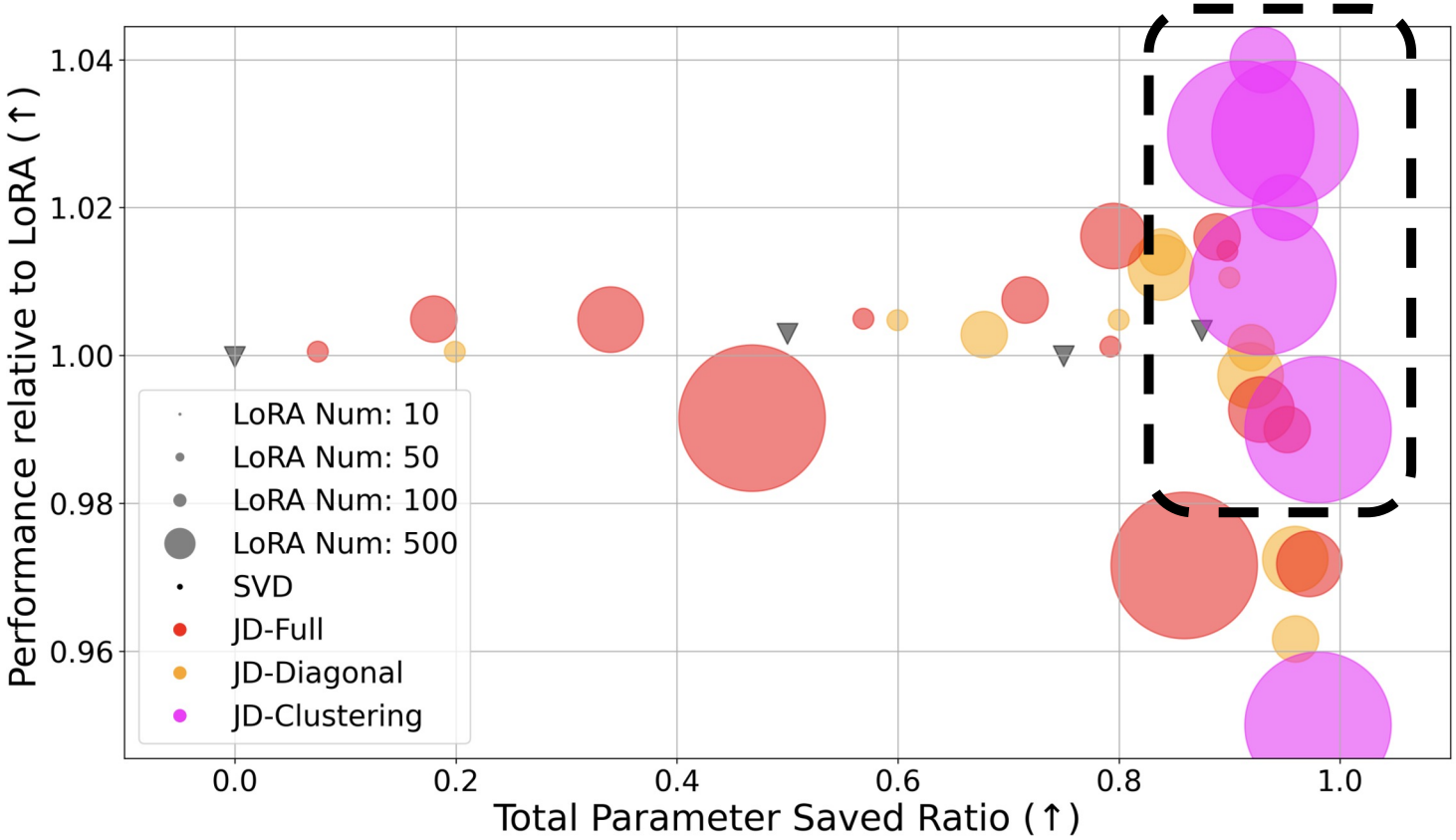


Revisiting 500 LoRAs



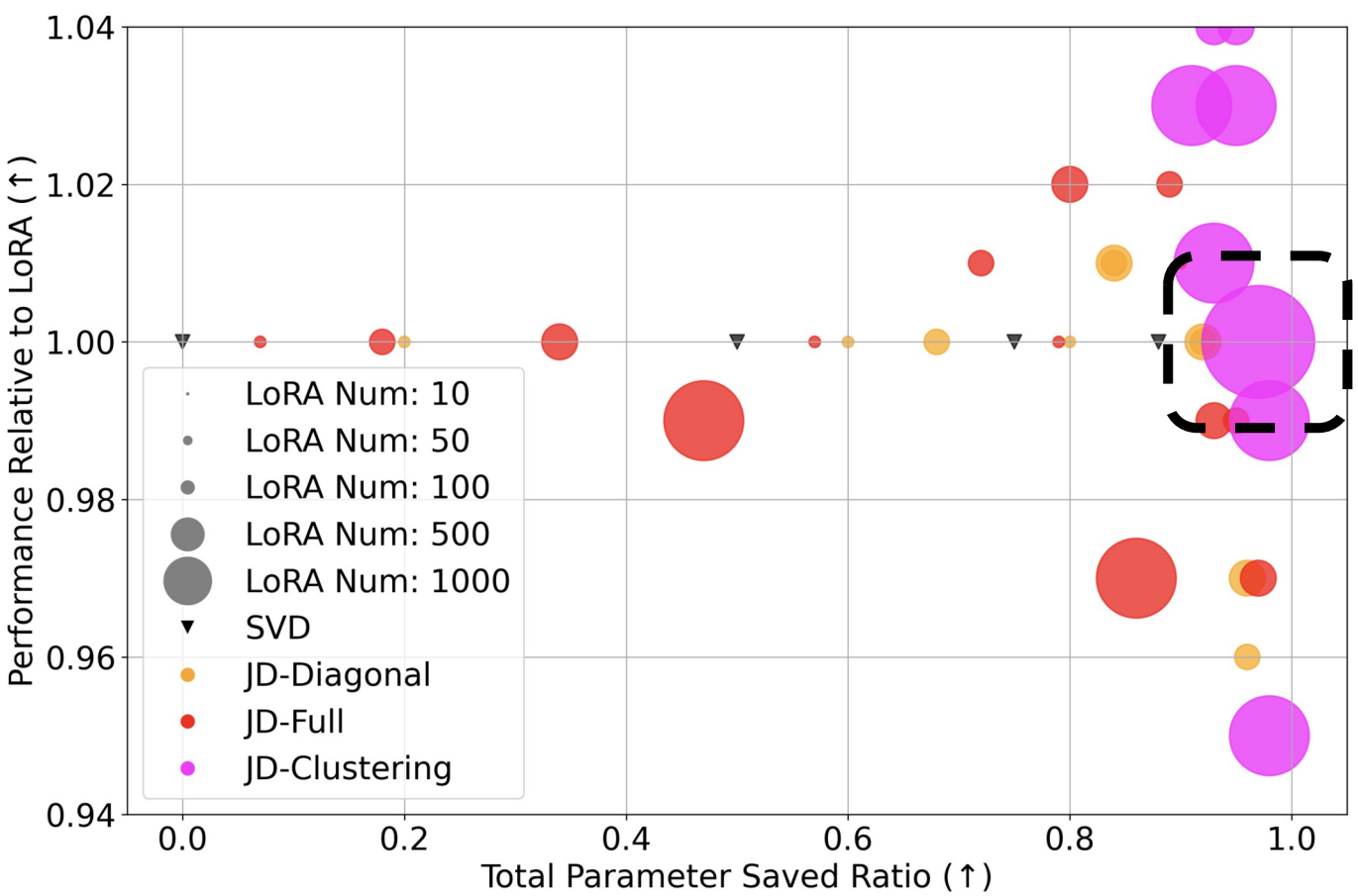
- Empirical data drives the methodology
- Proxy metric (reconstruction error) allows fast experimentation

Revisiting 500 LoRAs



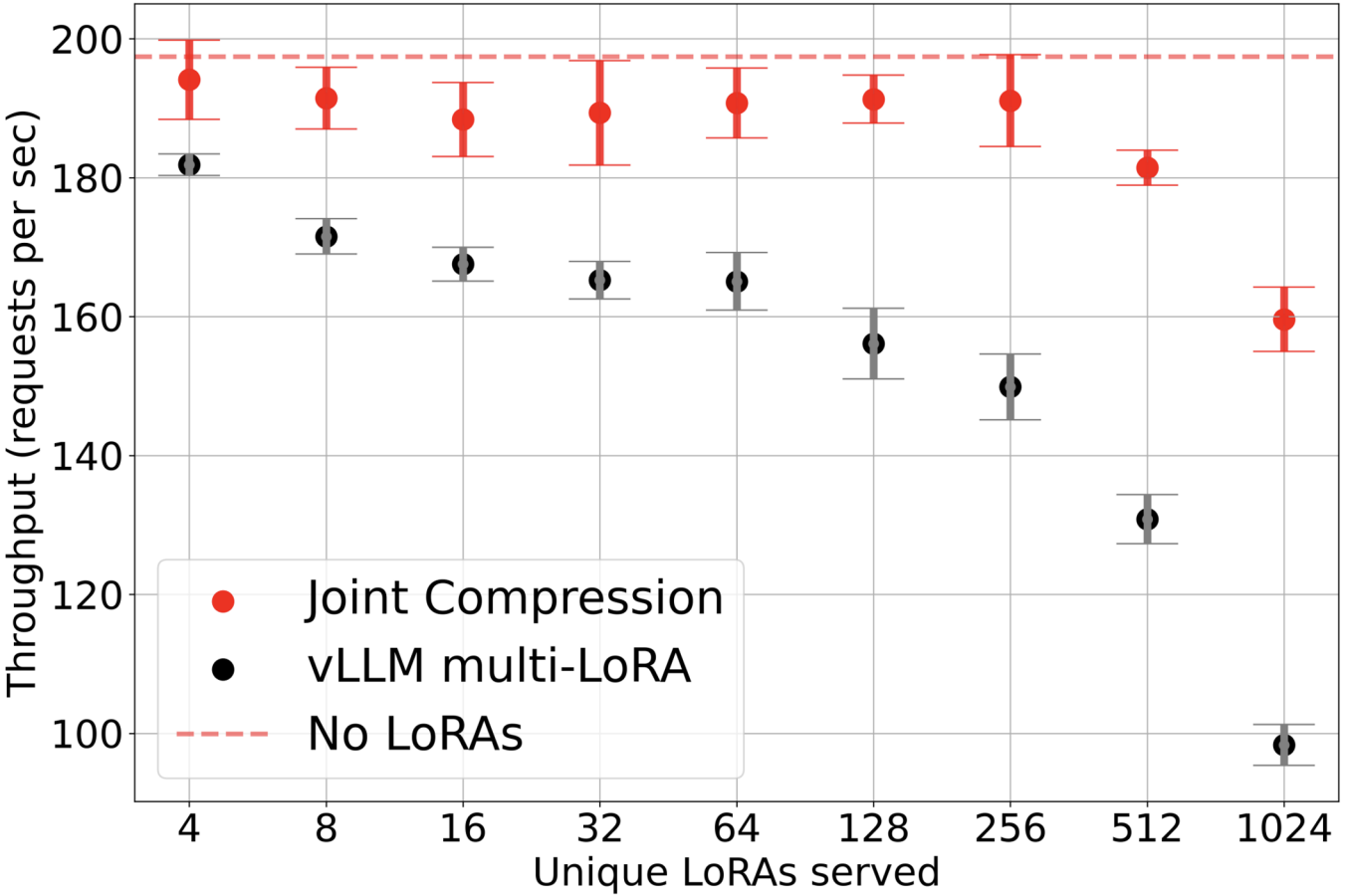
Over 95% compression with 500 LoRAs

What about a 1000 LoRAs?



Over 95% compression with 1000 LoRAs

Throughput with vLLM



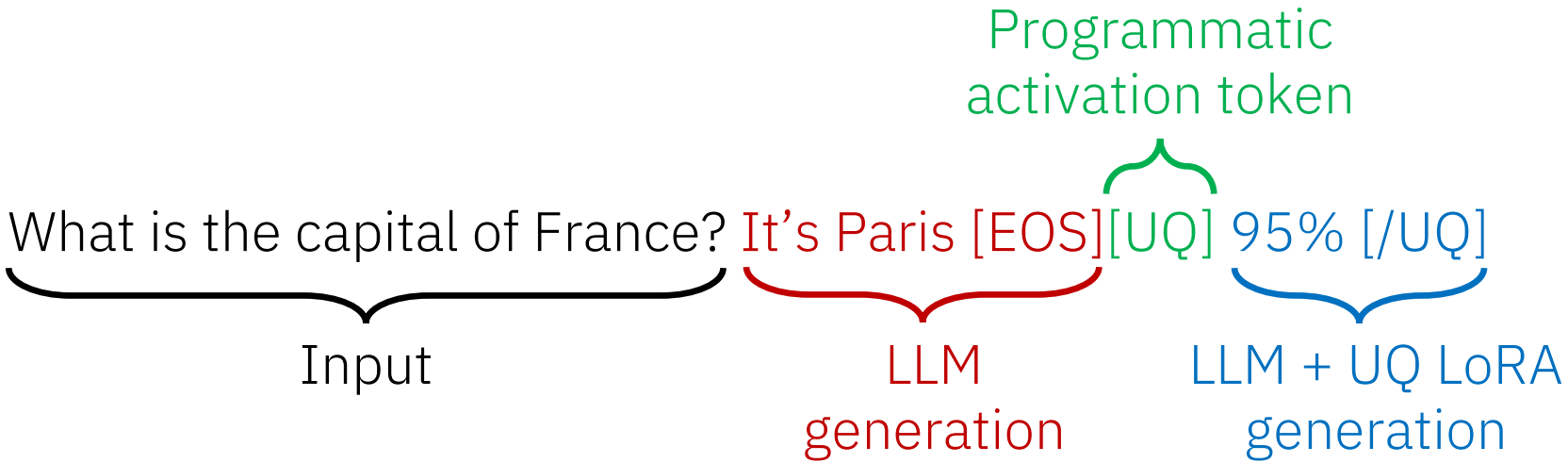
What is Next?

Development with customized models:

- Multi-agent chats [character.ai]
- Complex synthetic data pipelines
- LLM agents [crew.ai]

Efficient Test-Time Scaling

What is Next: Test-Time Search



Unaltered base model generation

LoRA-based verifiers can be trained independently

Can support 100s (or 1000s) of specialized verifiers

What is Next: o1-style

[OpenAI, 2024]

o1 CoT: Backtracking

Similarly, m(x) (−x 2) = (−x2n + 2 + m2n−2x2n + lower terms)m(x) * (−x 2) = (−x 2n + 2 + m 2n−2 x 2n + lower terms).*

Wait, actually, this may not help us directly without specific terms.An alternative is to consider the known polynomials.

So m(x) = k ...

Speculations on Test-Time Scaling, Rush & Ritter, 2024

What is Next: o1-style

[OpenAI, 2024]

o1 CoT: Self-Evaluation

Let's analyze each option.

Option A: "because appetite regulation is a field of staggering complexity."

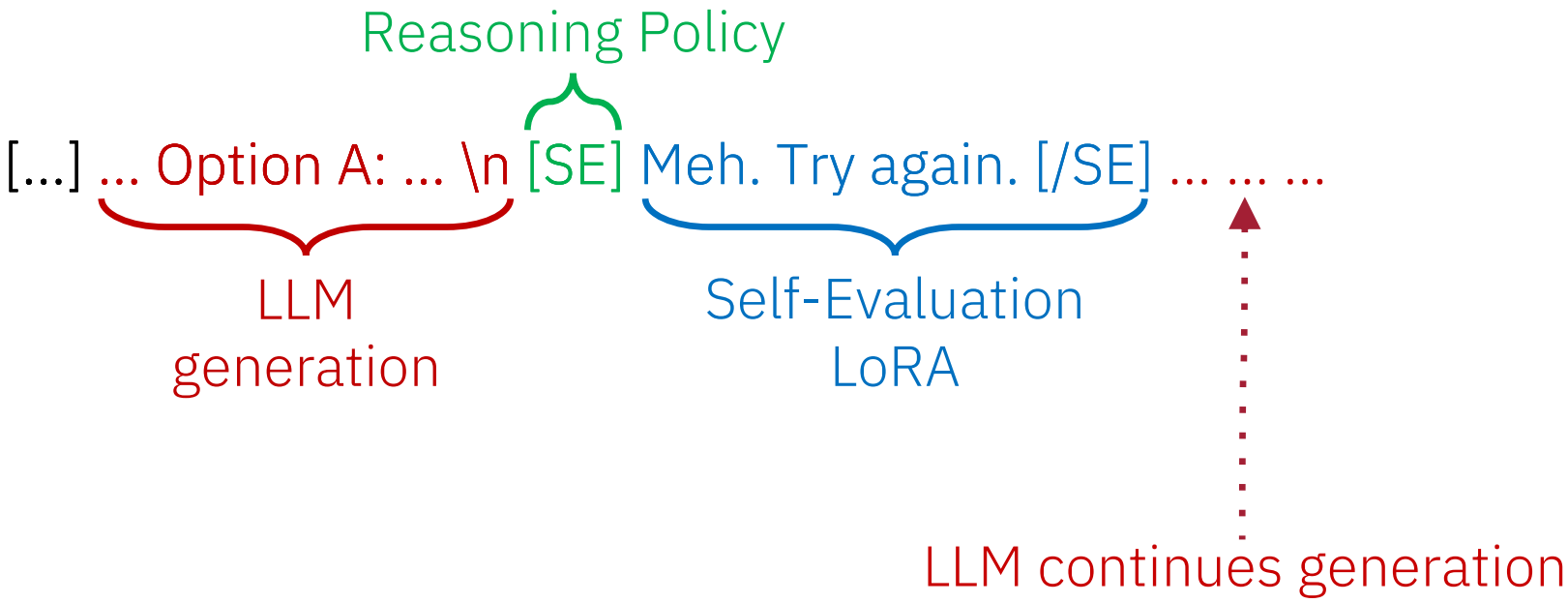
Is that a good explanation? Hmm.

Option B: "because researchers seldom ask the right questions."

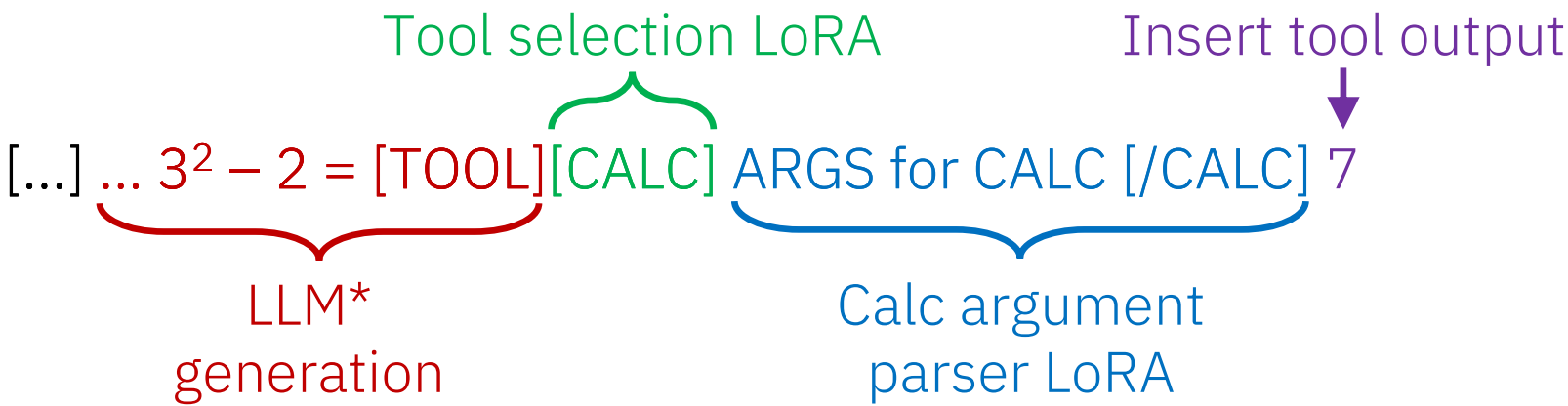
Does this make sense with the main clause?

Speculations on Test-Time Scaling, Rush & Ritter, 2024

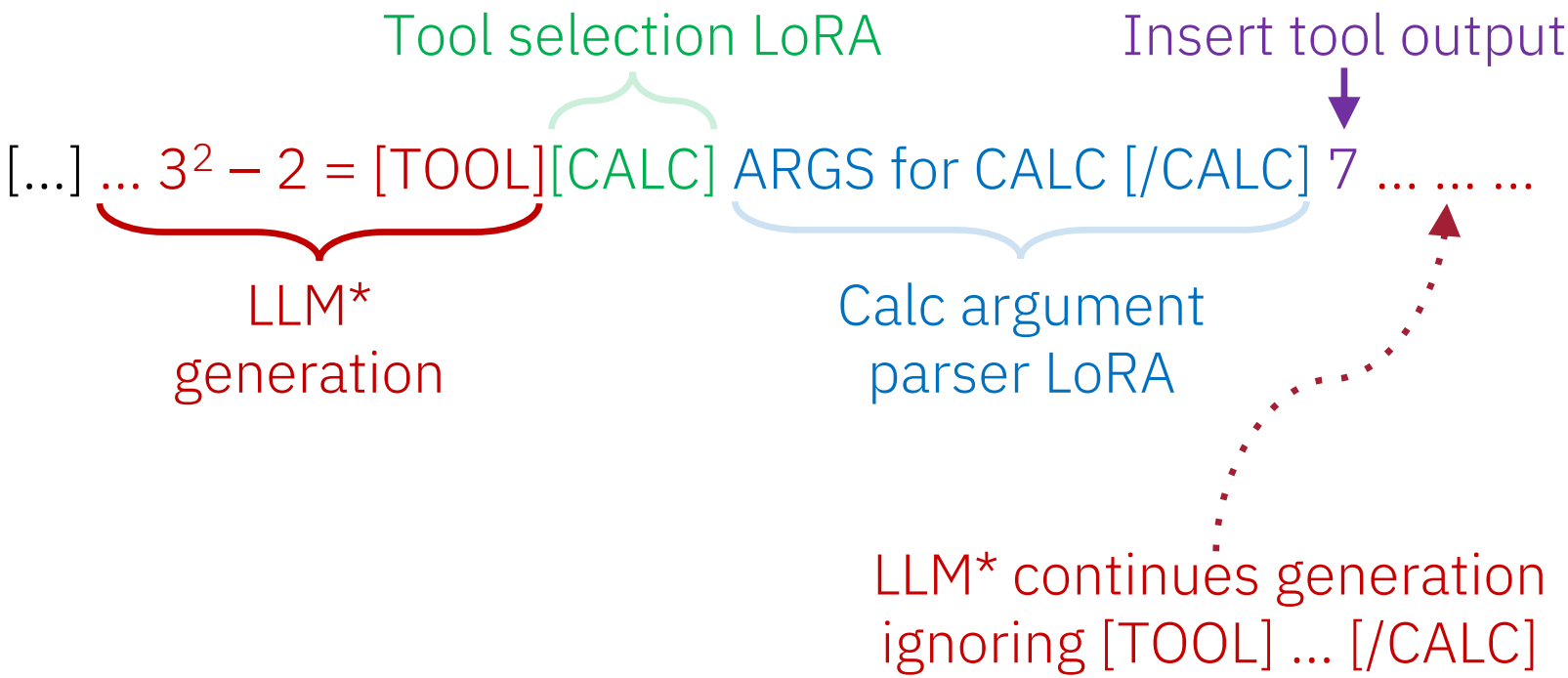
What is Next: o1-style



What is Next: Tool Use



What is Next: Tool Use



What is Next: Tool Use

Tool selection LoRA

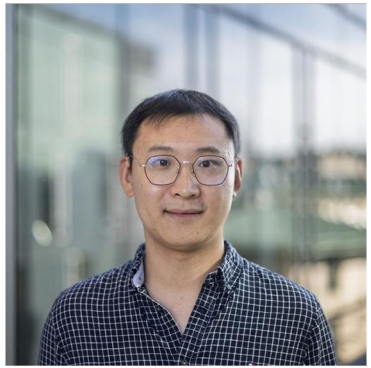
Insert tool output

- Many opportunities to create test-time capabilities through dedicated LoRAs
- We can serve such systems as efficiently as the vanilla LLM

Granite* continues generation ignoring [TOOL] … [/CALC]

Big Shoutout!

Tal Shnitzer
LLM Routing



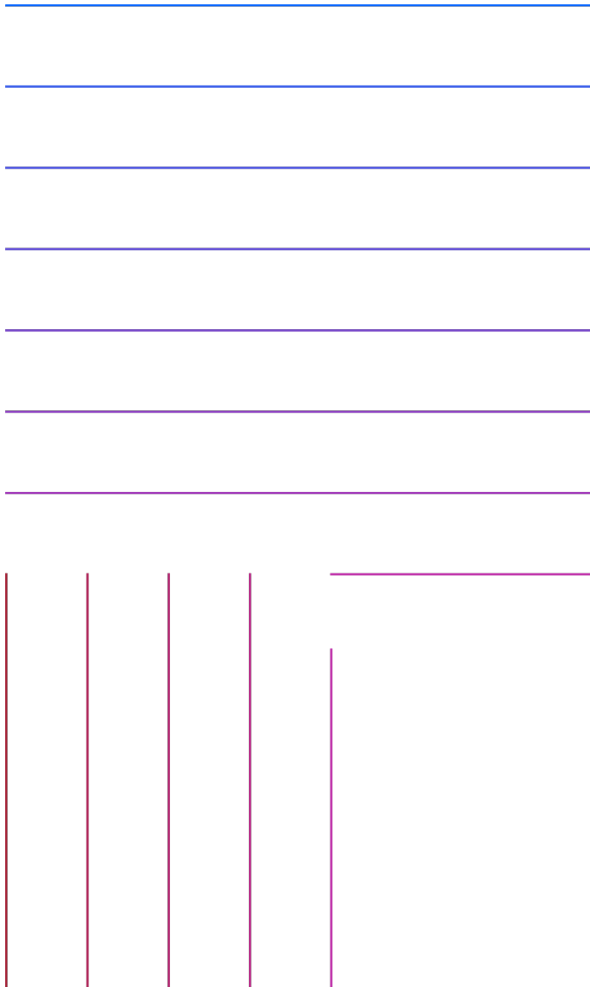
Hongyi Wang
LLM Routing

Felipe Maia Polo
LLM Evaluation

Rickard Gabrielsson
LoRA Compression

And many others!

Thank You!



Questions?