

The Importance of Workload Choice in Evaluating LLM Inference Systems

Konstantinos Papaioannou, Thaleia Dimitra Doudali


IMDEA Software Institute

@ EuroMLSys '24


Athens, April 22nd





The Large Language Models Effect

Write Stories

 **You**
Generate a short story about a time-traveler exploring the past.








 **ChatGPT**
Once upon a time, there lived a brave time-traveler named Alex. He had a passion for exploring the past and wanted to learn more about history and see the world as it once was. One day, he decided to take a leap of faith and travel back in time to the era of ancient Rome. When he arrived, he was surprised to find that many things were the same as he had read in books, but he was also amazed to discover new and interesting things. He saw chariots pulled by horses, people walking the streets wearing togas, tall marble statues of gods, and more. He spent weeks traveling around the city and learning about the people and places of the past. After exploring Rome, he ventured into the countryside and found himself in a small village. Here he discovered buildings from different eras, all existing together in harmony. He marveled at the stories he heard from the villagers, who told of the myths and legends of their ancestors. He also approached many of the villagers, asking them questions about their lives and customs, and he was welcomed warmly. Alex eventually returned to his present day, bringing back with him stories and valuable lessons that he had learned during his travels back in time. He was forever changed by his experience and found himself better able to connect with others and appreciate the beauty of the past.

Summarize Articles



 **ChatGPT**
Famed for her unique hats and bold colors, the queen has a trademark fashion style. For years her clothes were made by British designer Hardy Amies, now couturier Stewart Parvin makes her eye-catching outfits. Designers are given royal warrants to become dressers to the queen and it is no easy feat getting one.



   


Answer Questions

 **You**
What individual has won the most Olympic gold medals in the history of the games?








 **ChatGPT**
Michael Phelps has won the most gold medals of all time with 23 golds.

 **You**
What is best Playstation or Xbox?



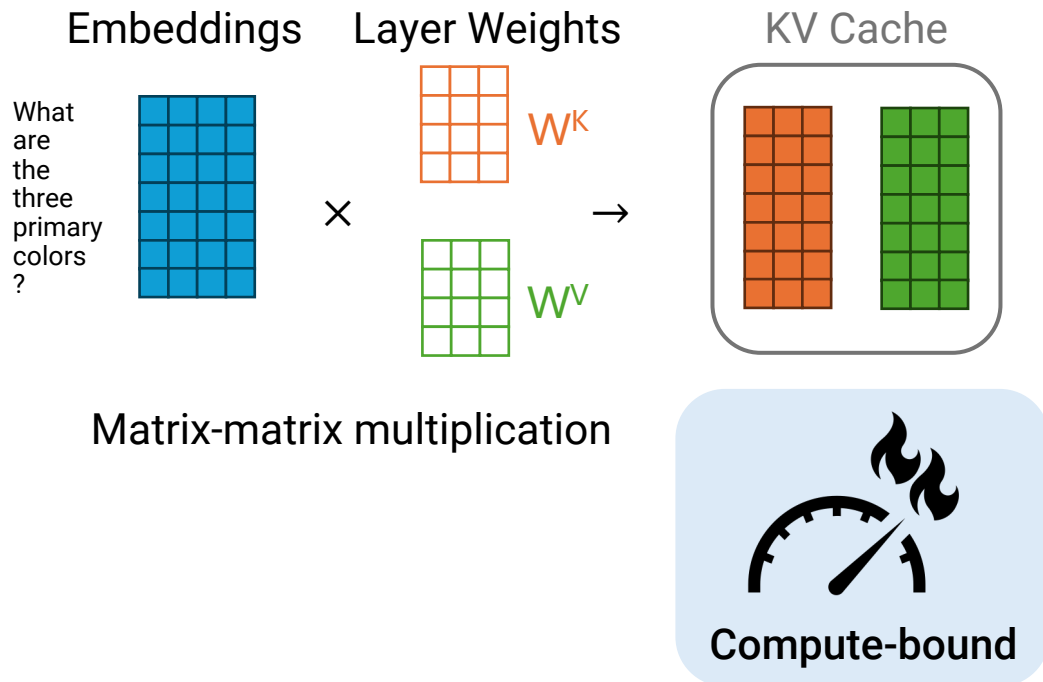
 **ChatGPT**
Playstation and Xbox are two competing game consoles, where which is "best" will depend on the preferences of the gamer. From a graphics and performance perspective they are comparable, with Sony's Playstation technically leading, though most gamers would be unlikely to notice a significant difference. From a game library perspective, both consoles have their respective first-party titles that appeal to different demographics. For example, Microsoft has seen great popularity due to popular title Halo being exclusively available on the console. On the other hand, Playstation has popular exclusive titles such as God of War and Gran Turismo. Ultimately the decision of which console is best will likely come down to which games titles the purchaser prefers.

LLM Inference Overview

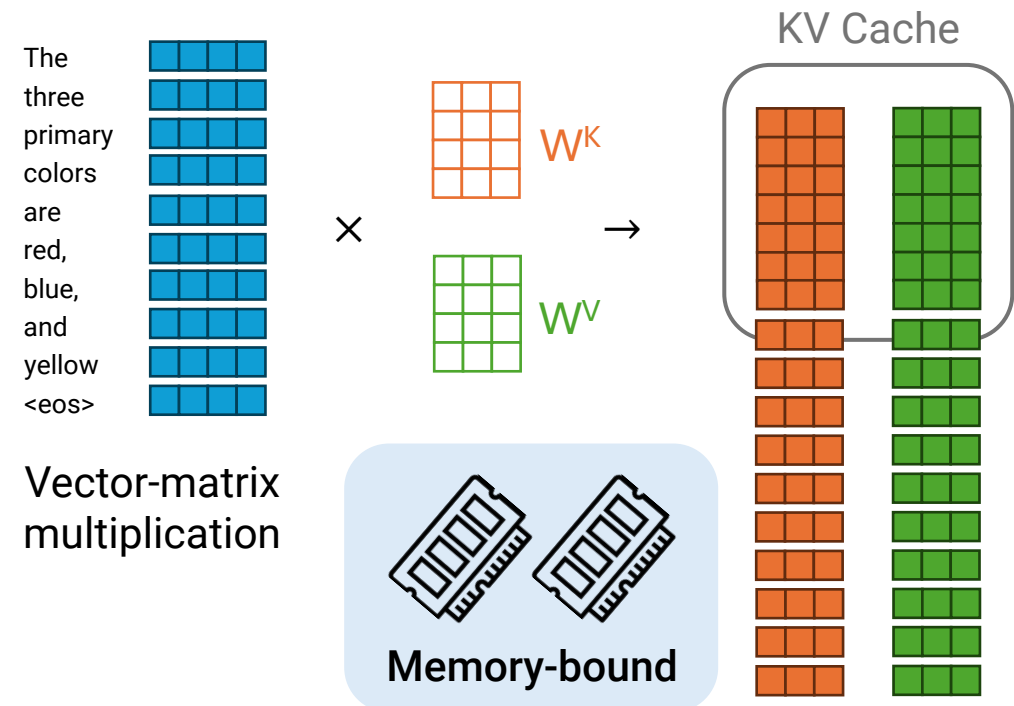
Prompt Phase

Input = What are the three primary colors ?



Generation Phase

Output = The three primary colors are red, blue, and yellow <eos>



State-of-the-art LLM Inference Systems

⚠ No consistent evaluation approach!

	System	Dataset		Inference Scenario	
		Synthetic	Real	Latency-critical	Best-effort
Scheduling Optimizations	Orca	✓		✓	✓
	SARATHI	✓			✓
	DeepSpeed-FastGen	✓		✓	
	Splitwise		✓	✓	
Memory Management	vLLM		✓	✓	
	S ³		✓	✓	✓
	FlexGen	✓			✓

Problem Statement: What is the impact of workload choice in evaluating LLM inference systems?

This paper: Analysis of Workload Impact in LLM Inference Systems

Workload = Dataset + Inference Scenario

4 Real Datasets: 🙌

- ✓ Alpaca
Text Generation
- ✓ CNN DailyMail
Text Summarization
- ✓ Dolly
Question-Answering
- ✓ ShareGPT
Conversational

I) Latency-critical Inference

Max request rate under SLO (requests/s)
> 0.1875 s/output token for 95% of requests

Request arrival times → Poisson distribution

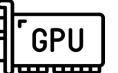
II) Best-effort Inference

Throughput (requests/s)

Batch input of 1000 requests

Experimental Setup




Hardware

1× NVIDIA A100 40GB 

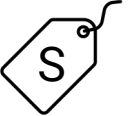
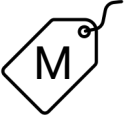

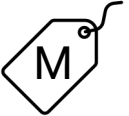
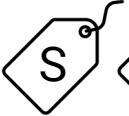
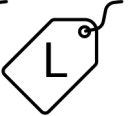




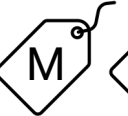
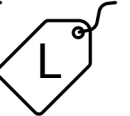
Inference Engine

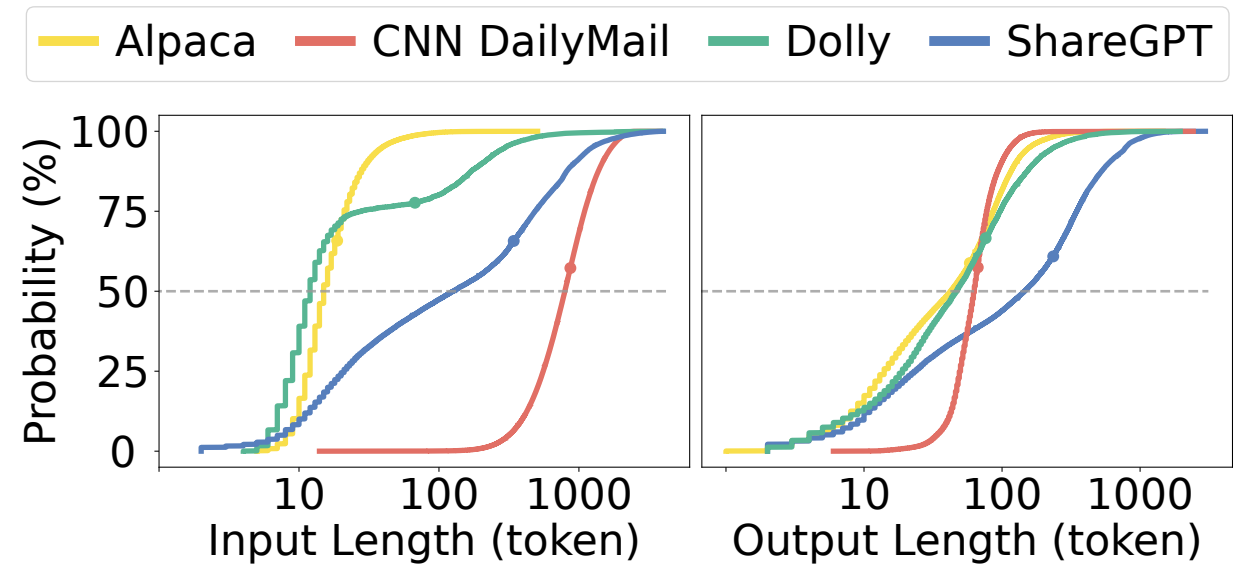


Models

			
OPT	6.7B	13B	
Llama-2	7B	13B	

Dataset Analysis

	Input	Output
Alpaca text-generation		
CNN DailyMail text-summarization		
Dolly question-answering	 	
ShareGPT conversational	  	 

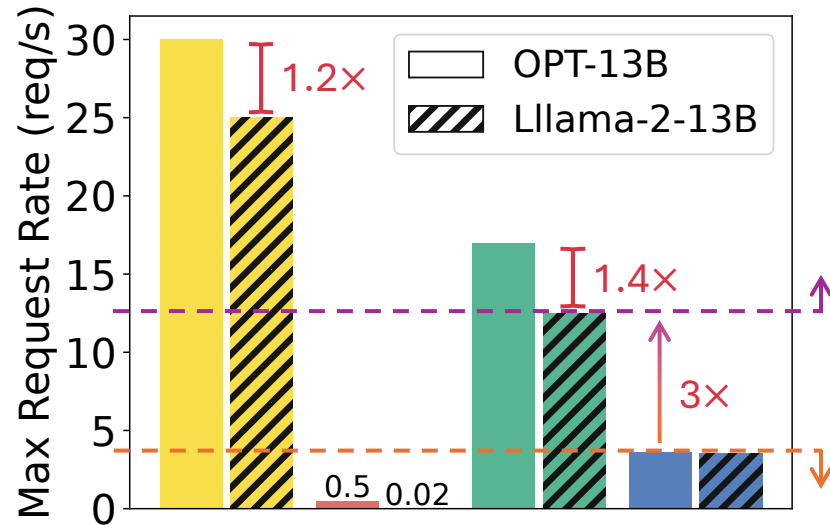


Takeaway: The use case significantly impacts the sequence length of the output, but it has an even greater impact on the length of the **input** sequences.

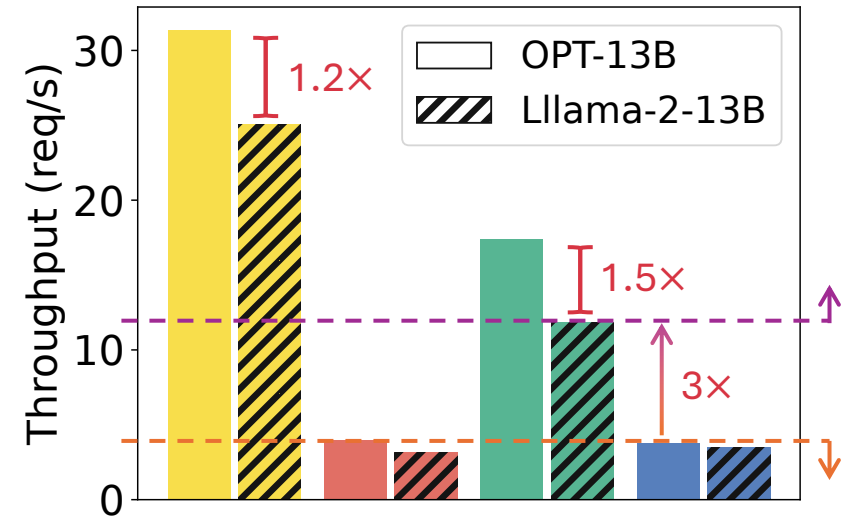


Impact of Use Case on Performance

— Alpaca (text-generation) — CNN DailyMail (text-summarization) — Dolly (question-answering) — ShareGPT (conversational)



Latency-critical Inference



Best-effort Inference

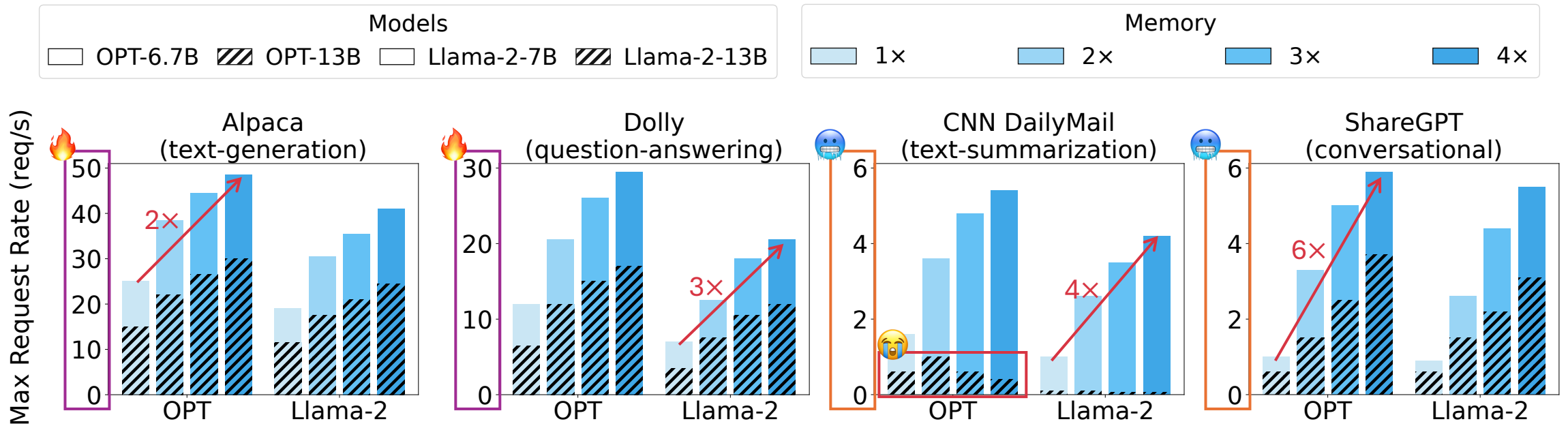


Takeaway: The use case impacts performance significantly.

1. Two classes of inference performance (high vs low).
2. Text summarization and conversational use cases have low performance, due to their larger inputs.

Impact of KV Cache Size on Performance

Latency-critical Inference

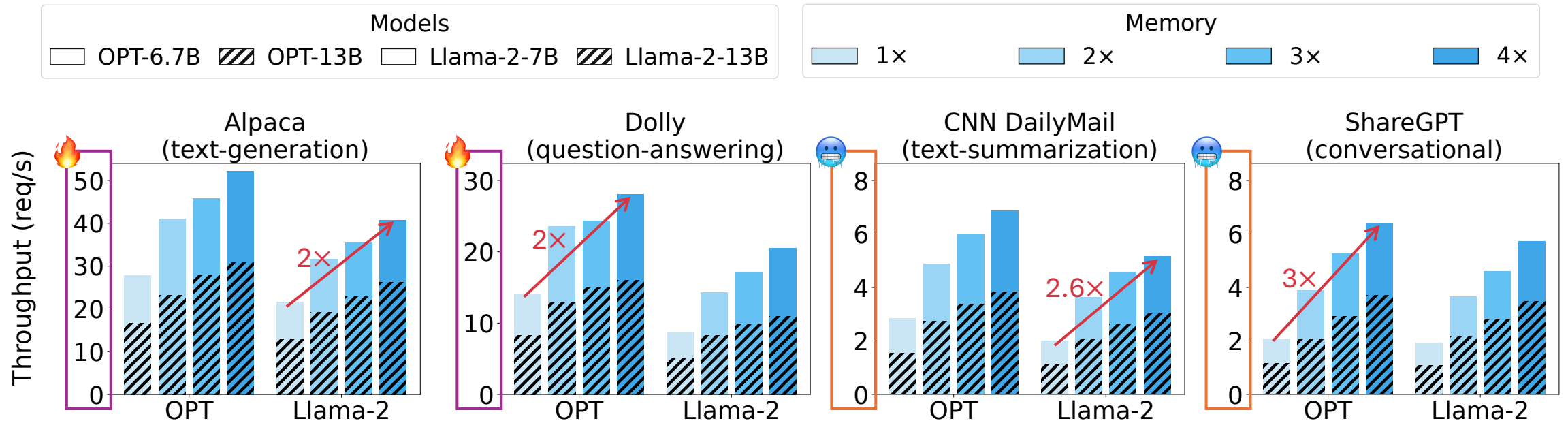


Takeaway: Higher memory availability almost always improves performance of *latency-critical* inference, except in the text summarization use case.



Impact of KV Cache Size on Performance

Best-effort Inference



Takeaway: Higher memory availability **always** improves performance of *best-effort* inference.



Lessons Learned

⚠ The workload choice is very important!

Text generation and question-answering:

- achieve high performance.
- benefit from larger KV cache size.



Ideal for evaluating LLM inference systems.

Best-effort inference consistently benefits from higher memory availability.



Let's enhance memory management for this inference scenario.

Text summarization and conversational use cases have low performance, due to their larger inputs.



Treat them separately to improve their inference performance.

Scan for code

