The Importance of Workload Choice in Evaluating LLM Inference Systems

Konstantinos Papaioannou Thaleia Dimitra Doudali

IMDEA Software Institute, Madrid, Spain

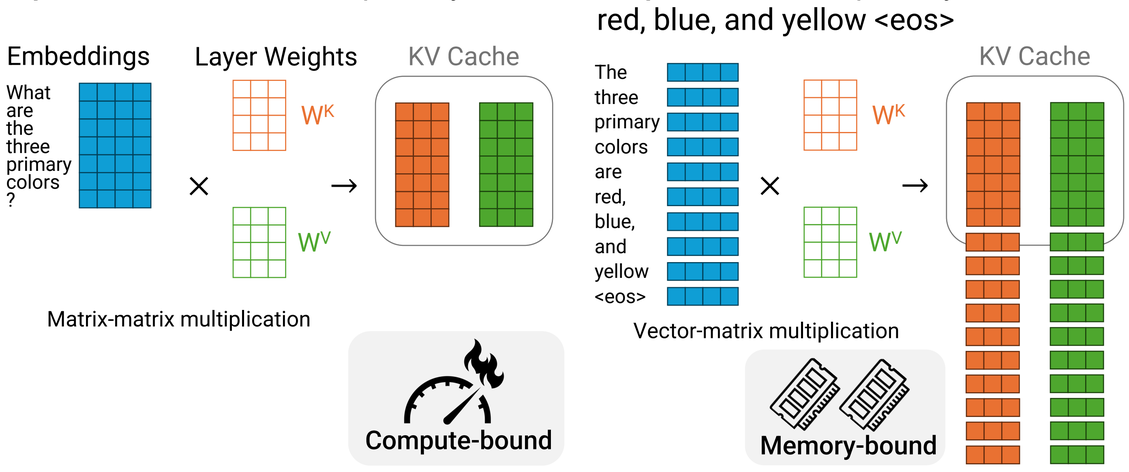
1. LLM Inference Overview

Prompt Phase

Generation Phase

Input = What are the three primary colors ?

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



2. State-of-the-art LLM Inference Systems

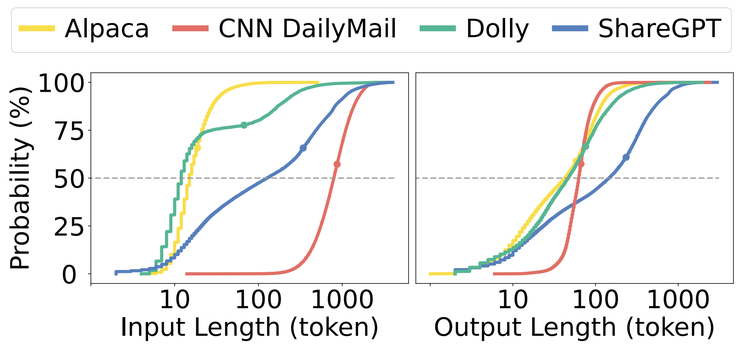
⚠ No consistent evaluation approach!

	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?

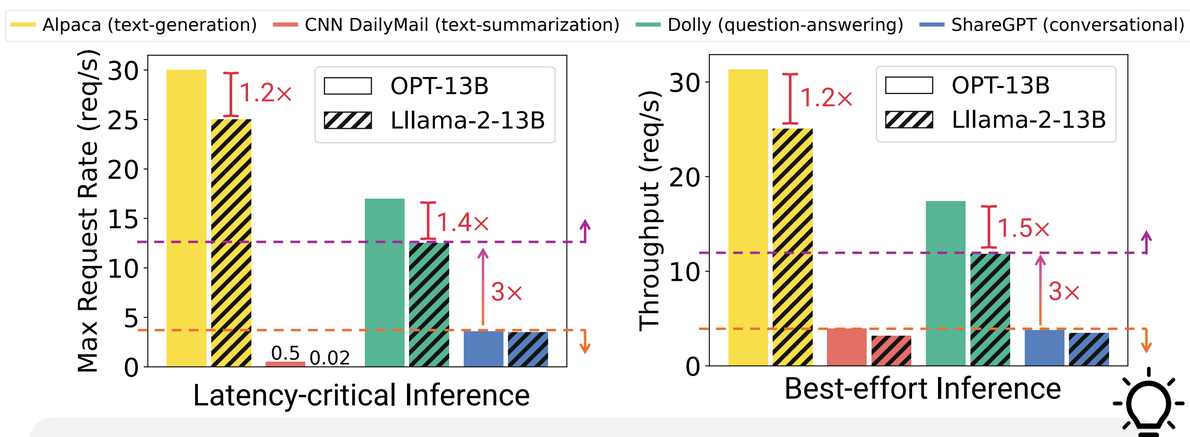
3.Dataset Analysis

	Input	Output
Alpaca text-generation	S	M
CNN DailyMail text-summarization	XL	M
Dolly question-answering	S, L	M
ShareGPT conversational	S, M, L	M, L



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.

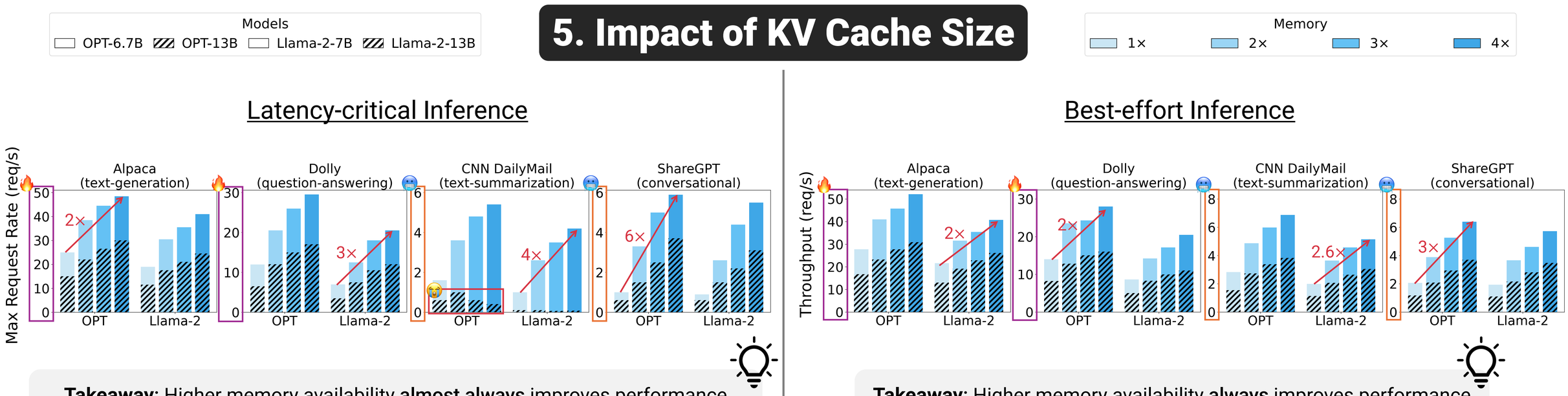
4. Impact of Use Case



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**.

5. Impact of KV Cache Size



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

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

6. Lessons Learned

⚠ The workload choice is very important!

1. **Text generation** and **question-answering**:
   - achieve **high performance**.
   - benefit from **larger KV cache** size.
   → Ideal for evaluating LLM inference systems.
2. **Best-effort inference** consistently benefits from **higher memory** availability. → Let's enhance memory management for this inference scenario.
3. **Text summarization** and **conversational** use cases have **low performance**, due to their **larger inputs**. → Treat them separately to improve their inference performance.

Paper

Code