

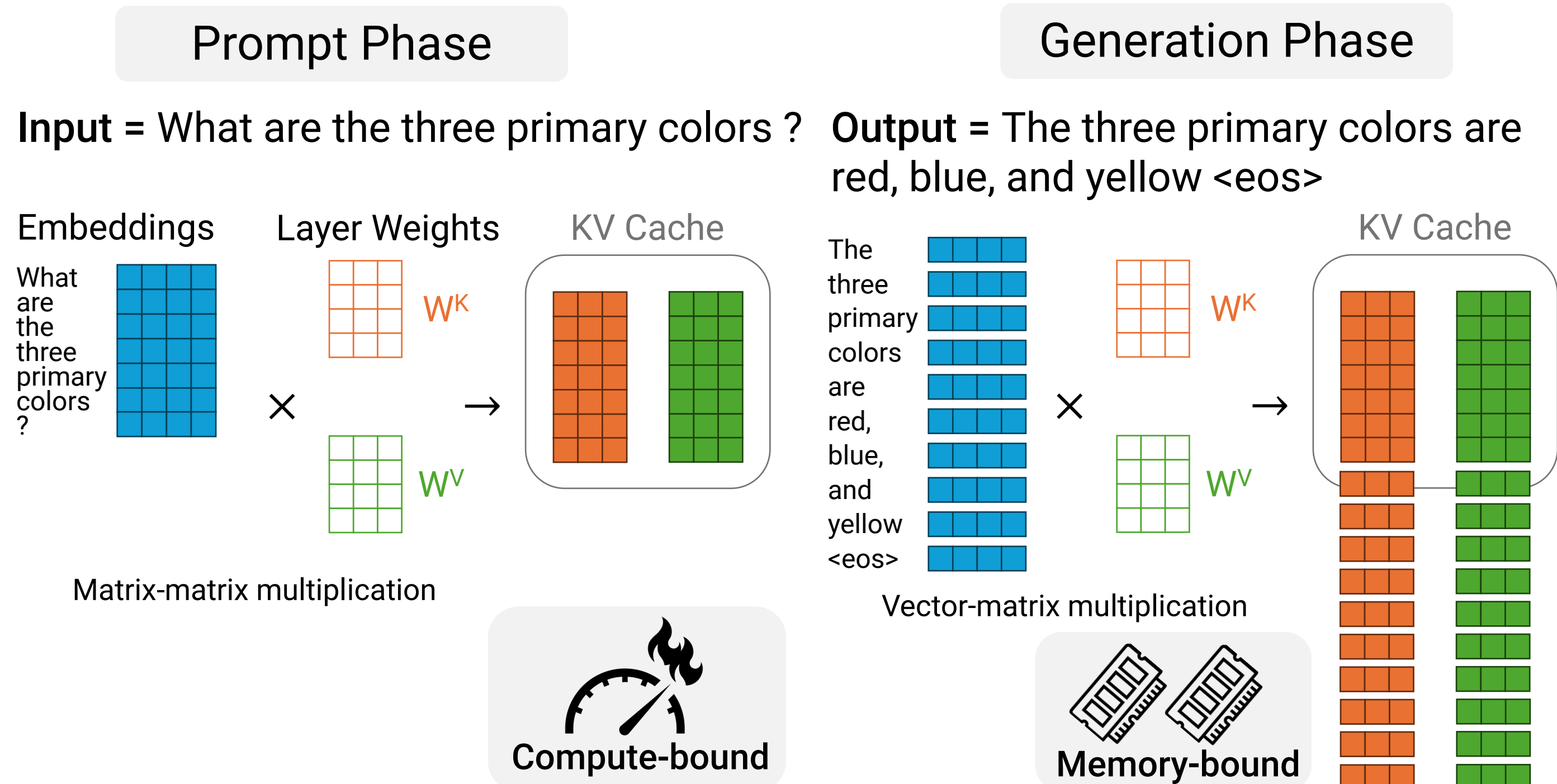
# The Importance of Workload Choice in Evaluating LLM Inference Systems

Konstantinos Papaioannou

Thaleia Dimitra Doudali

IMDEA Software Institute, Madrid, Spain

## 1. LLM Inference Overview



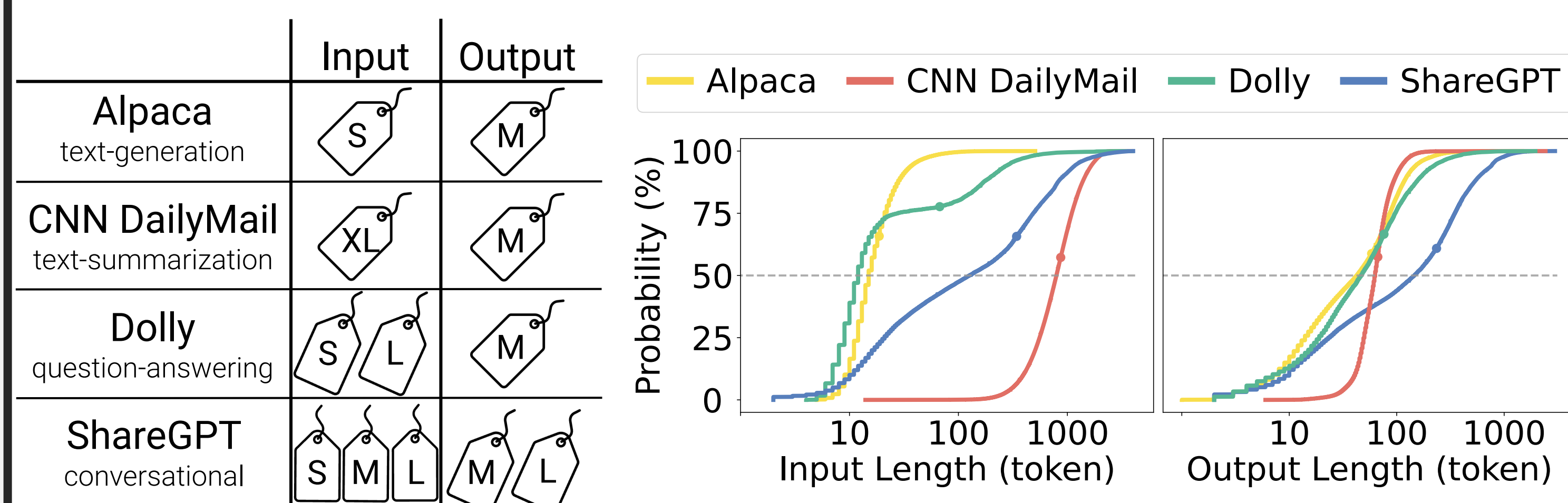
## 2. State-of-the-art LLM Inference Systems

⚠ No consistent evaluation approach!

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

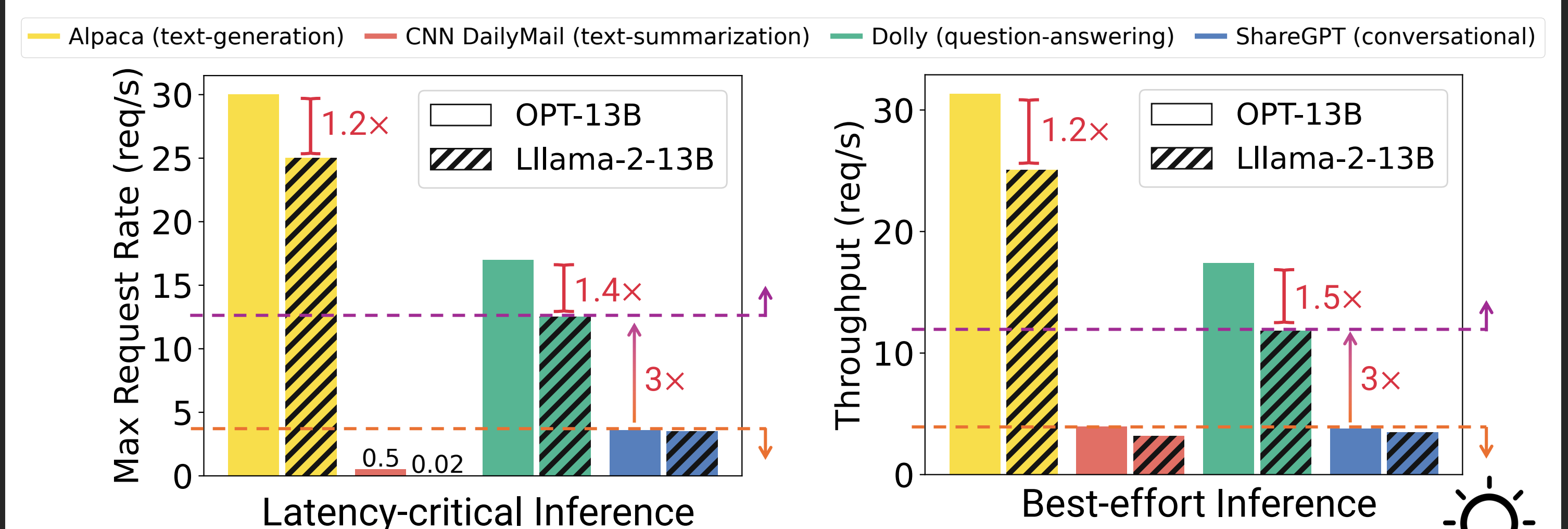
**Problem Statement:** What is the impact of workload choice in evaluating LLM inference systems?

## 3. Dataset Analysis



**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.

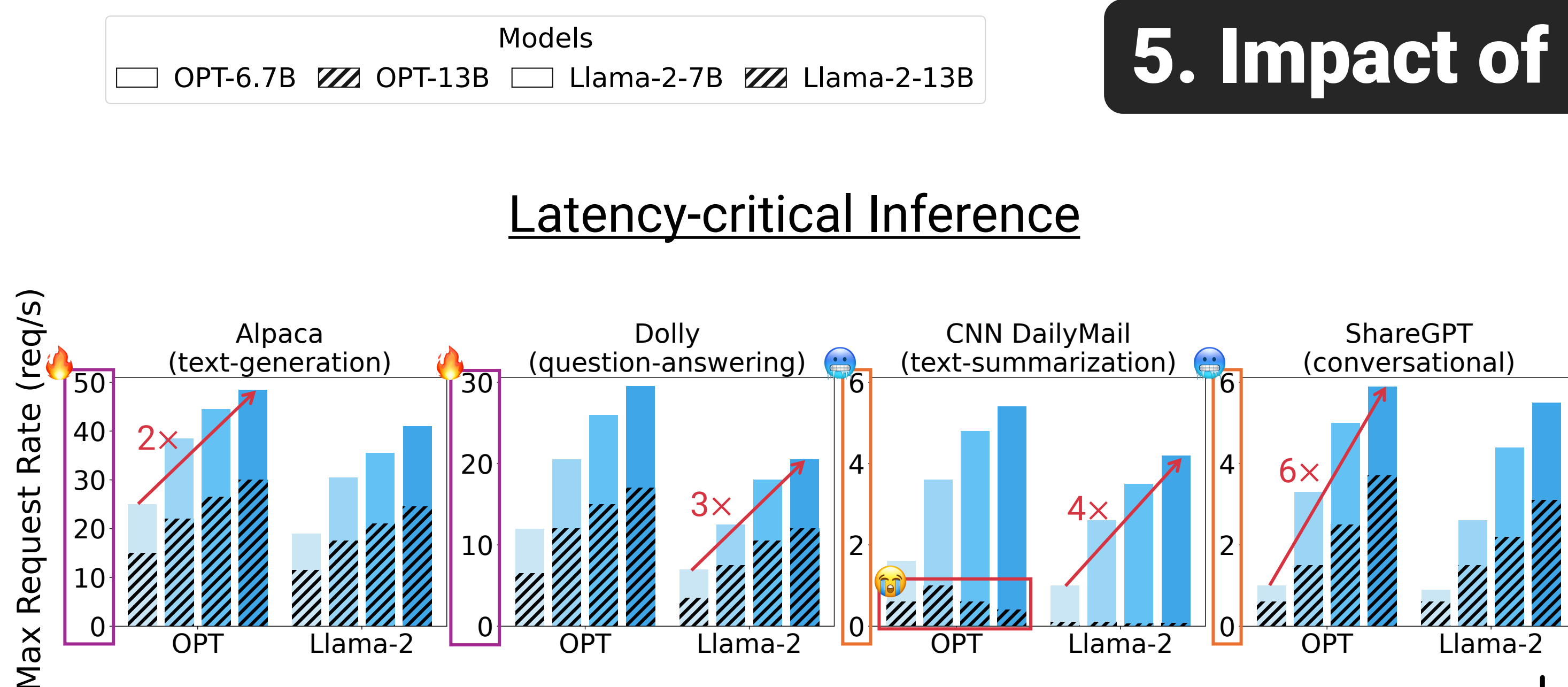
## 4. Impact of Use Case



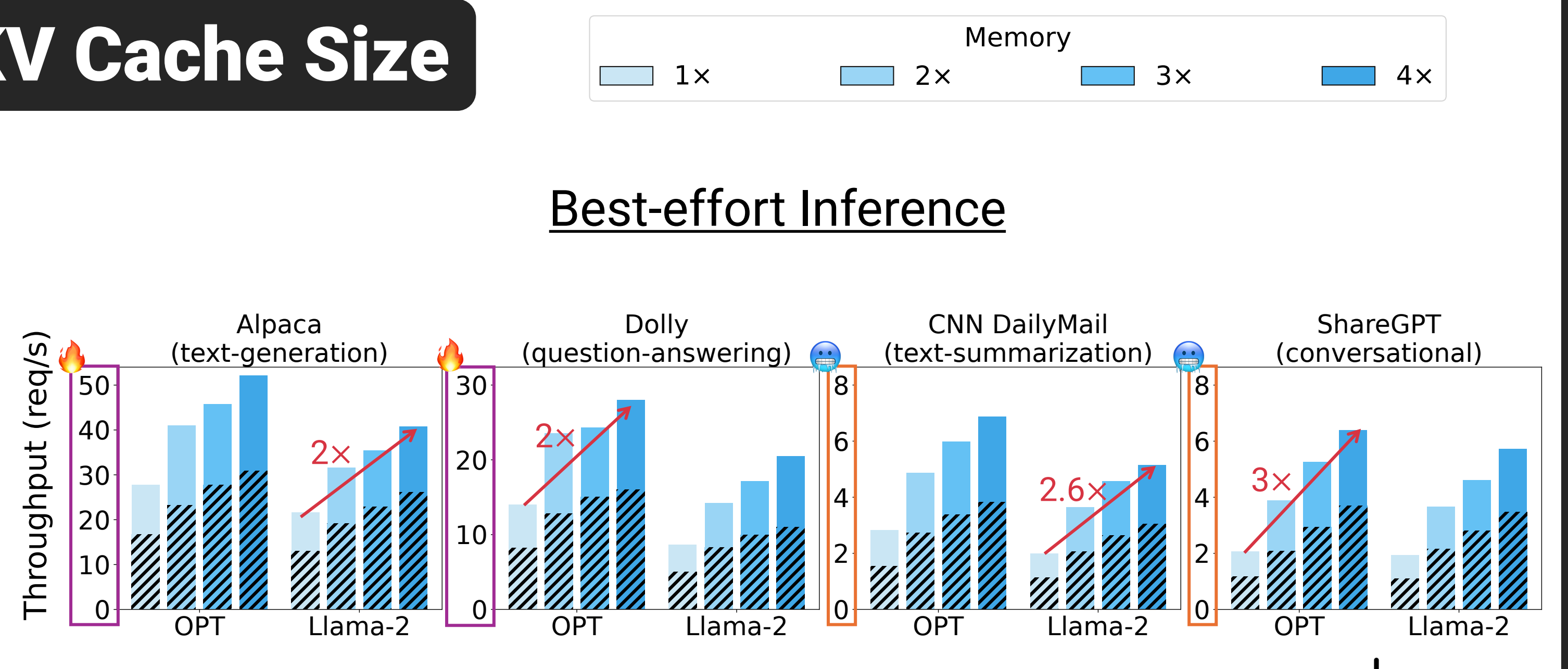
**Takeaway:** The use case impacts performance significantly.

- Two classes of inference performance (**high** vs **low**).
- 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!

- 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.

### Paper



### Code

