

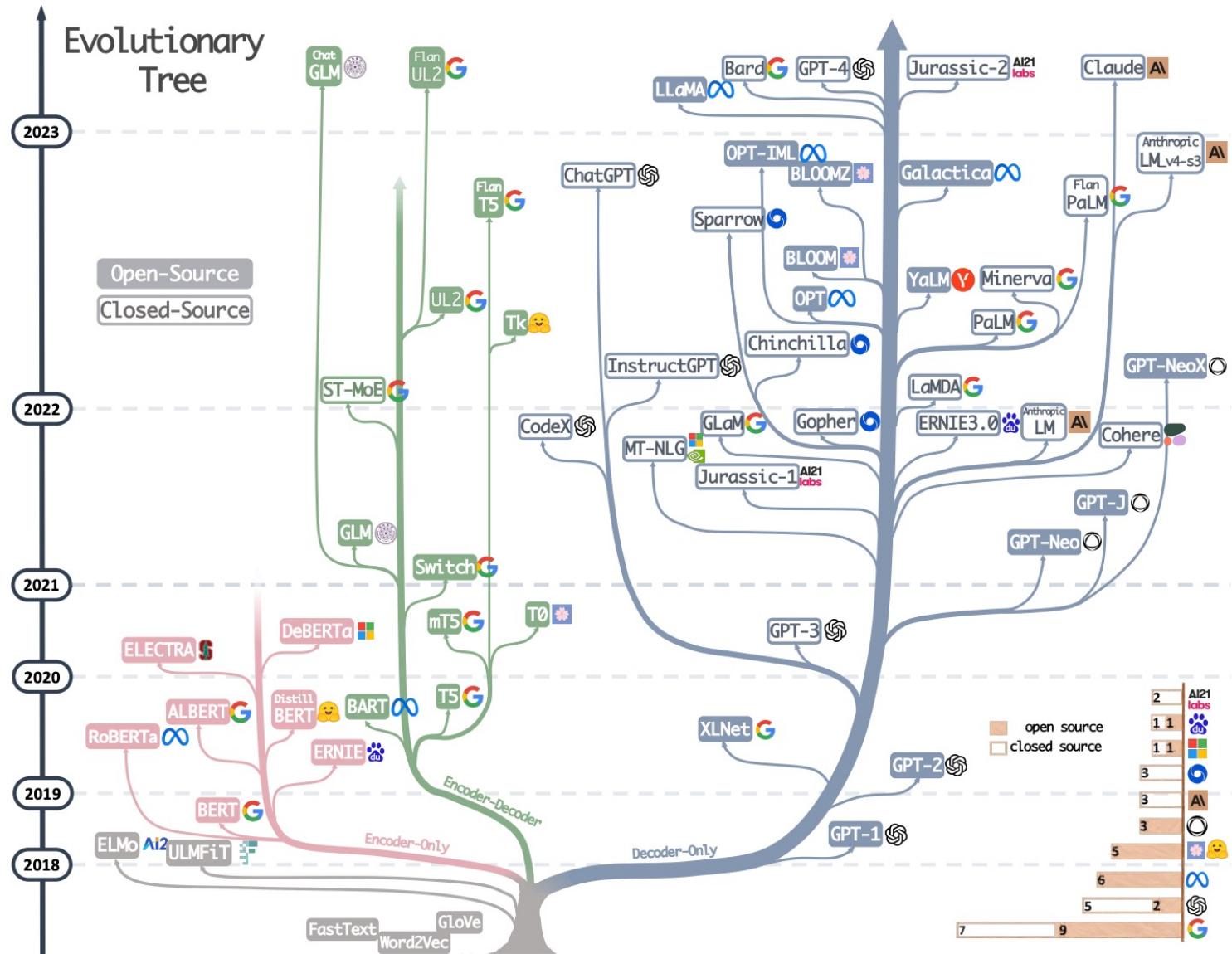
Which GPU for Open Source LLM Applications? Inference Performance Insights

Adrian Tineo, Ph.D.



Open source LLMs

- New open source LLMs are being constantly released
- Variety of licenses, but many are commercially permissive
- Frequent changes in the leaderboards



Multiple flavors of NVIDIA GPUs available



Hyperstack

<p>CPU Only</p> <p>1</p> <p>4 vCPUs 4 GB RAM 60 GB Disk</p> <p>\$0.32903/hour</p>	<p>NVIDIA A100-80G-PCIe</p> <p>4x A100-80G-PCIe</p> <p>126 vCPUs 480 GB RAM 3300 GB Disk</p> <p>\$11.00/hour</p>	<p>NVIDIA A100-80G-PCIe-NVLink</p> <p>8x A100-80G-PCIe-NVLink</p> <p>252 vCPUs 960 GB RAM 6500 GB Disk</p> <p>\$22.00/hour</p>	<p>NVIDIA A100-80G-PCIe-sm</p> <p>1x A100-80G-PCIe-sm</p> <p>32 vCPUs 120 GB RAM 850 GB Disk</p> <p>\$2.75/hour</p>	<p>NVIDIA H100-80G-PCIe</p> <p>4x H100-80G-PCIe</p> <p>126 vCPUs 720 GB RAM 3200 GB Disk</p> <p>\$17.20/hour</p>
<p>NVIDIA H100-80G-PCIe-k8s</p> <p>8x H100-80G-PCIe-k8s</p> <p>This Model is currently fully reserved, click here to reserve yours now & get priority access.</p> <p>\$4.30/hour</p>	<p>NVIDIA H100-80GB-PCIe-sm</p> <p>1x H100-80GB-PCIe-sm</p> <p>32 vCPUs 180 GB RAM 750 GB Disk</p> <p>\$4.30/hour</p>	<p>NVIDIA L40</p> <p>4x L40</p> <p>126 vCPUs 240 GB RAM 3300 GB Disk</p> <p>\$5.60/hour</p>	<p>NVIDIA L40-sm</p> <p>1x L40-sm</p> <p>32 vCPUs 60 GB RAM 850 GB Disk</p> <p>\$1.40/hour</p>	<p>NVIDIA RTX-A6000</p> <p>1x RTX-A6000</p> <p>16 vCPUs 59.5 GB RAM 425 GB Disk</p> <p>\$1.10/hour</p>

The versatility of LLMs



Language Models are Few-Shot Learners

Tom B. Brown*

Benjamin Mann*

Nick Ryder*

Melanie Subbiah*

Jared Kaplan[†]

Prafulla Dhariwal

Arvind Neelakantan

Pranav Shyam

Girish Sastry

Amanda Askell

Sandhini Agarwal

Ariel Herbert-Voss

Gretchen Krueger

Tom Henighan

Rewon Child

Aditya Ramesh

Daniel M. Ziegler

Jeffrey Wu

Clemens Winter

Christopher Hesse

Mark Chen

Eric Sigler

Mateusz Litwin

Scott Gray

Benjamin Chess

Jack Clark

Christopher Berner

Sam McCandlish

Alec Radford

Ilya Sutskever

Dario Amodei

OpenAI

May 2020

What is your use case?



- Long input, short output (MMLU)

Question answering / extraction / classification

Input

The following are multiple choice questions (with answers) about high school computer science.

Output

B

<5 examples>

What is the output of the statement "a" + "ab" in Python 3?

- A. Error
- B. aab
- C. ab
- D. a ab

Answer:

What is your use case?



- Short input, long output (lmsys-chat-1m)

Input

I want to write a collection of tutorials for CAD design. Make an explanation on how many chapters do you think are necessary, what will be the topics and headers, and what information the chapters should include.

Code / brainstorming / letter composing / essay writing

Output

Topics and headers in each chapter:

1. Introduction to CAD
 - What is CAD?
 - How does CAD differ from 2D drafting?
 - ...
2. Getting Started
 - Downloading and Installing Rhino3D
 - Navigating the Interface
 - ...
3. Modeling in CAD
 - Extruding 2D Objects Into 3D
 - Using Boolean Operations
 - ...

LLM families: Llama2



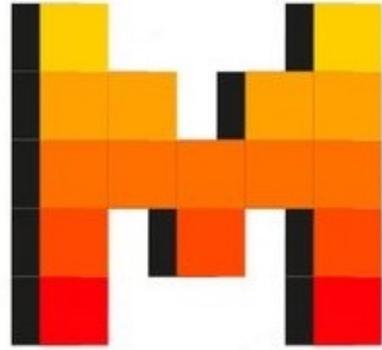
- 40% more training data than Llama1
- Available at 7B, 13B, and 70B sizes
- Fine-tuned versions are optimized for dialog use cases
- Extensive alignment process (SFT, iterative RLHF)
- Standard transformer architecture
- Grouped-Query Attention (GQA) for the 70B variation
(Ainslie et al.)

LLM families: Falcon



- Filtering and deduplication of web data
- Model architecture based on PaLM
- Available at 7B, 40B, and 180B sizes
- Fine-tuned on instructions and conversations
- Multi-Query Attention (MQA) (similar to GQA)
- Optimized for tensor parallelism

LLM families: Mistral



- Mixtral 8x7B, a sparse Mixture of Experts (SMoE)
- Each token has access to 47B but uses 13B active params
- Allows for Expert Parallelism (EP)
- Fine-tuned with SFT and DPO

- Mistral 7B, a standard transformer architecture
- Grouped-Query Attention (GQA)
- Sliding Window Attention (SWA)
- Fine-tuned on instructions dataset

LLM families



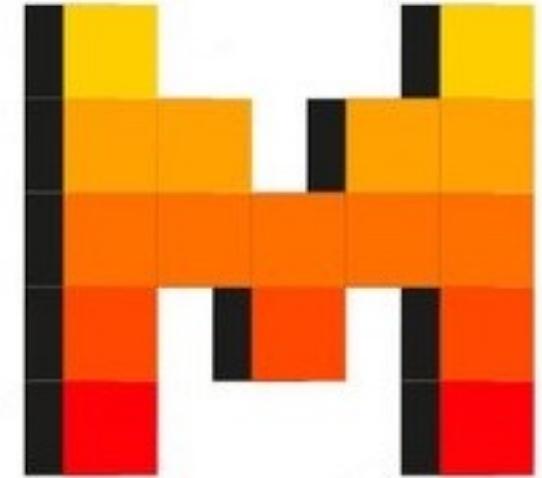
Llama2-70B-chat (129 GB)
LLama2-7B-chat (13 GB)



Falcon-40B-instruct (78 GB)
Falcon-7B-instruct (13 GB)



Mixtral-8x7B-instruct (87 GB)
Mistral-7B-instruct (14 GB)



Inference engine: vLLM



- Optimized for memory efficiency and high throughput
- PagedAttention
- Optimized CUDA kernels with fused operations
- Tensor parallelism
- Supports all model variations out-of-the-box

GPU node flavors



NVIDIA.
RTX-A6000

4x RTX-A6000

64	vCPUs
238 GB	RAM
1700 GB	Disk

\$4.40/hour

GPU mem: $4 \times 48 = 192$ GB

NVIDIA.
L40

4x L40

126	vCPUs
240 GB	RAM
3300 GB	Disk

\$5.60/hour

GPU mem: $4 \times 48 = 192$ GB

NVIDIA.
A100-80G-Pcie-sm

2x A100-80G-Pcie-sm

64	vCPUs
240 GB	RAM
1600 GB	Disk

\$5.50/hour

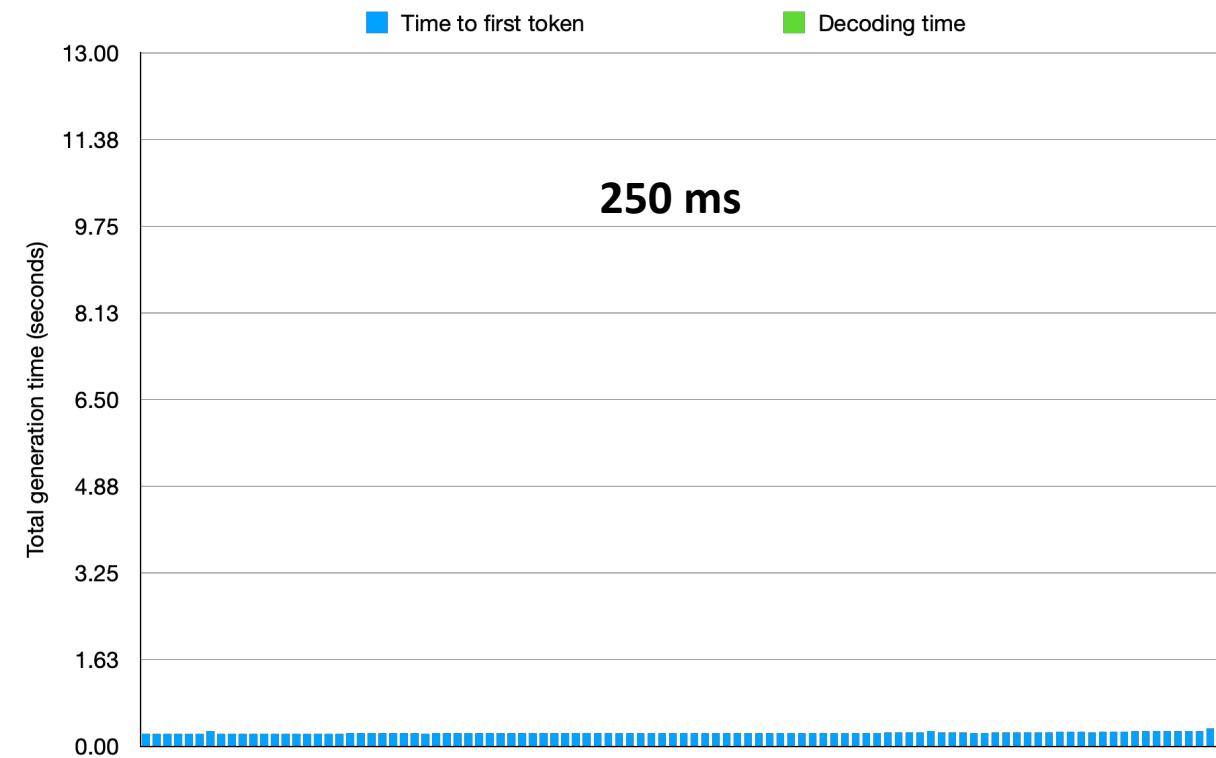
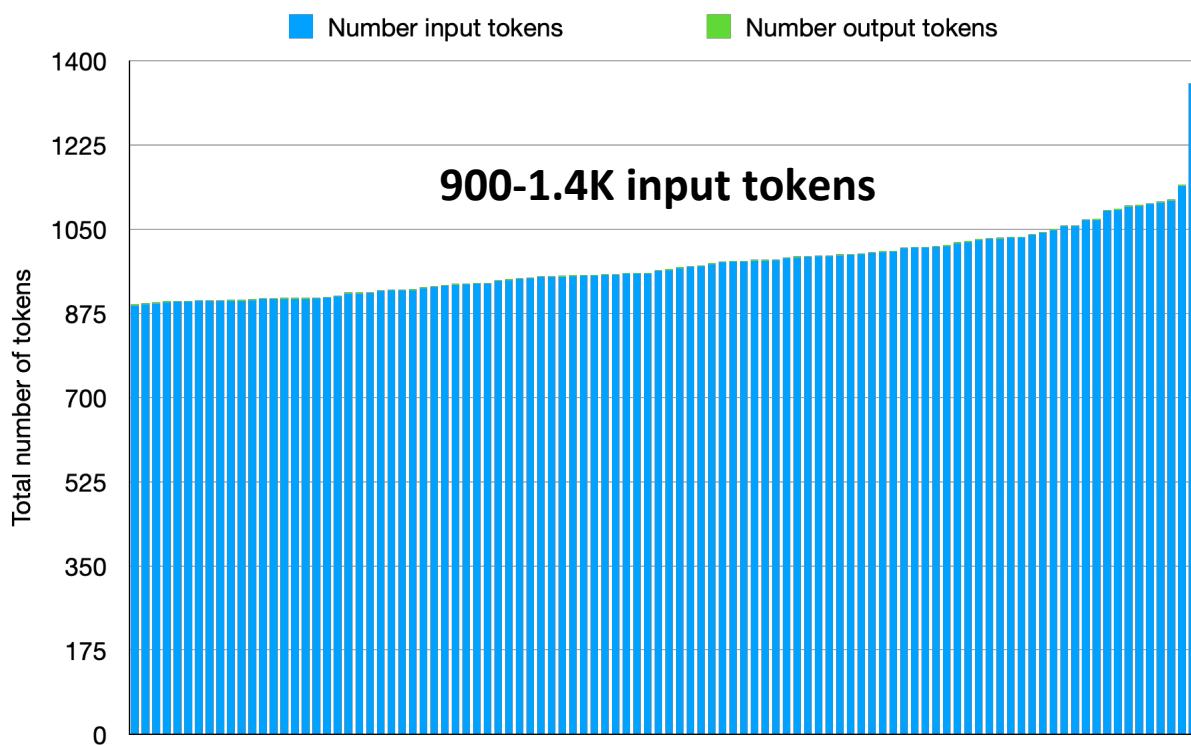
GPU mem: $2 \times 80 = 160$ GB

NVIDIA GPUs



	NVIDIA RTX A6000	NVIDIA L40	NVIDIA A100
Architecture	Ampere	Ada Lovelace	Ampere
GPU Memory	48 GB	48 GB	80 GB
Memory bandwidth	768 GB/s	864 GB/s	1.94 TB/s
FP16/BF16 Tensor Core	75.6 TFLOPS	181.05 TFLOPS	312 TFLOPs
Interconnect	PCIe Gen4x16	PCIe Gen4x16	PCIe Gen4x16

Latency per workload type



Long input, short output (MMLU)

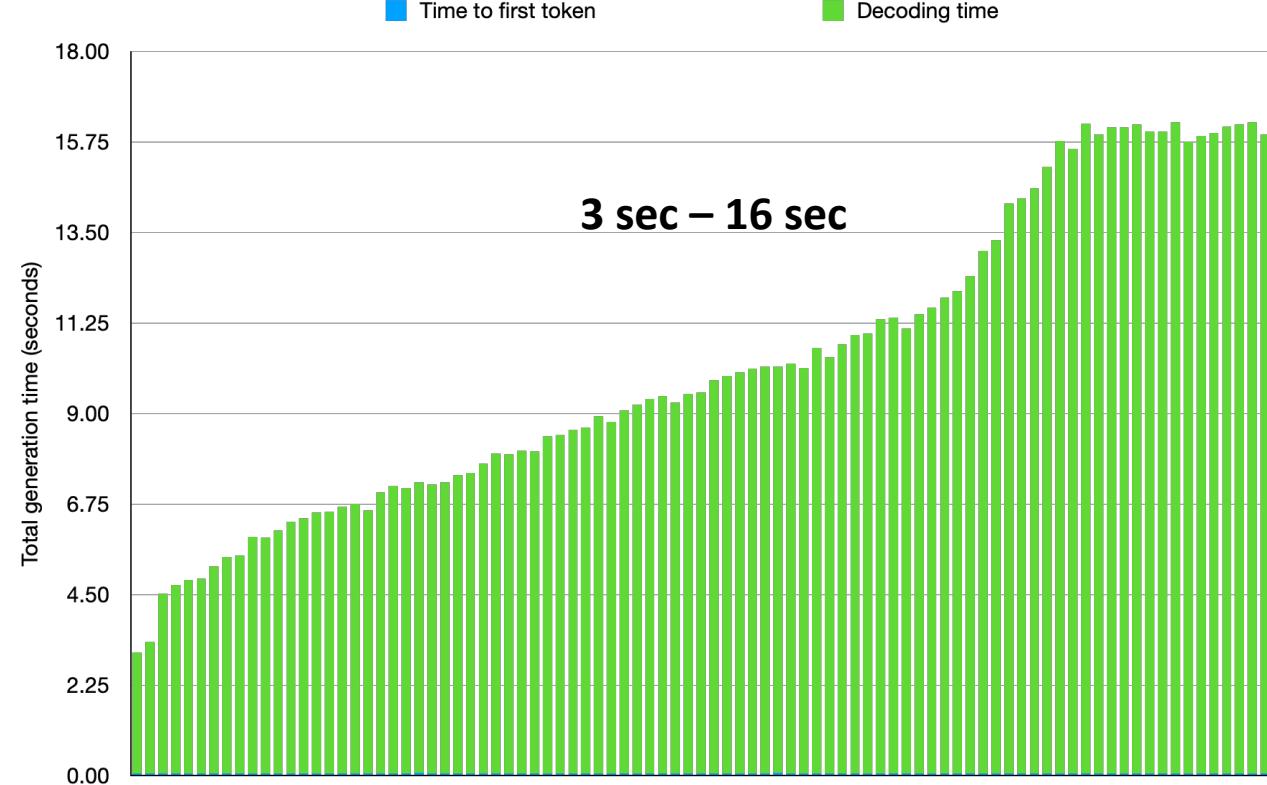
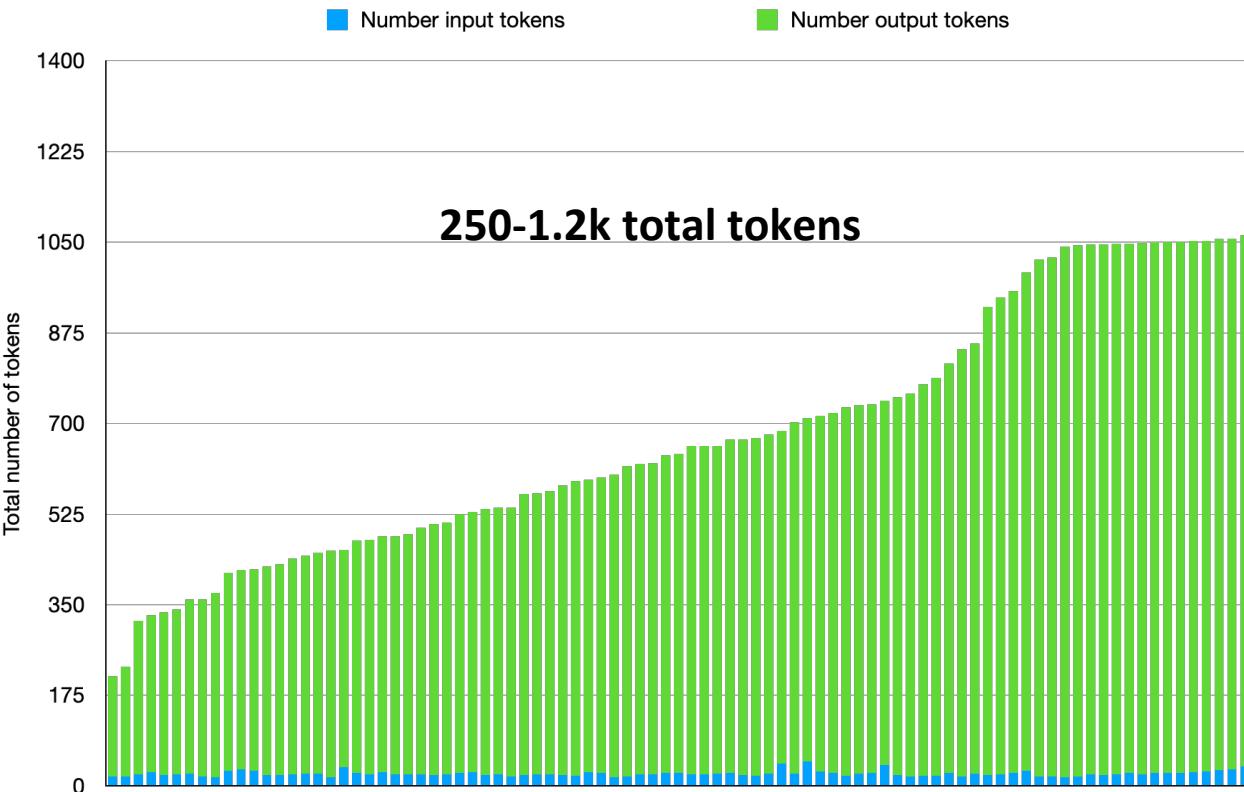


2xA100



Mixtral-8x7B-instruct

Latency per workload type



Short input, long output (lmsys-chat-1m)

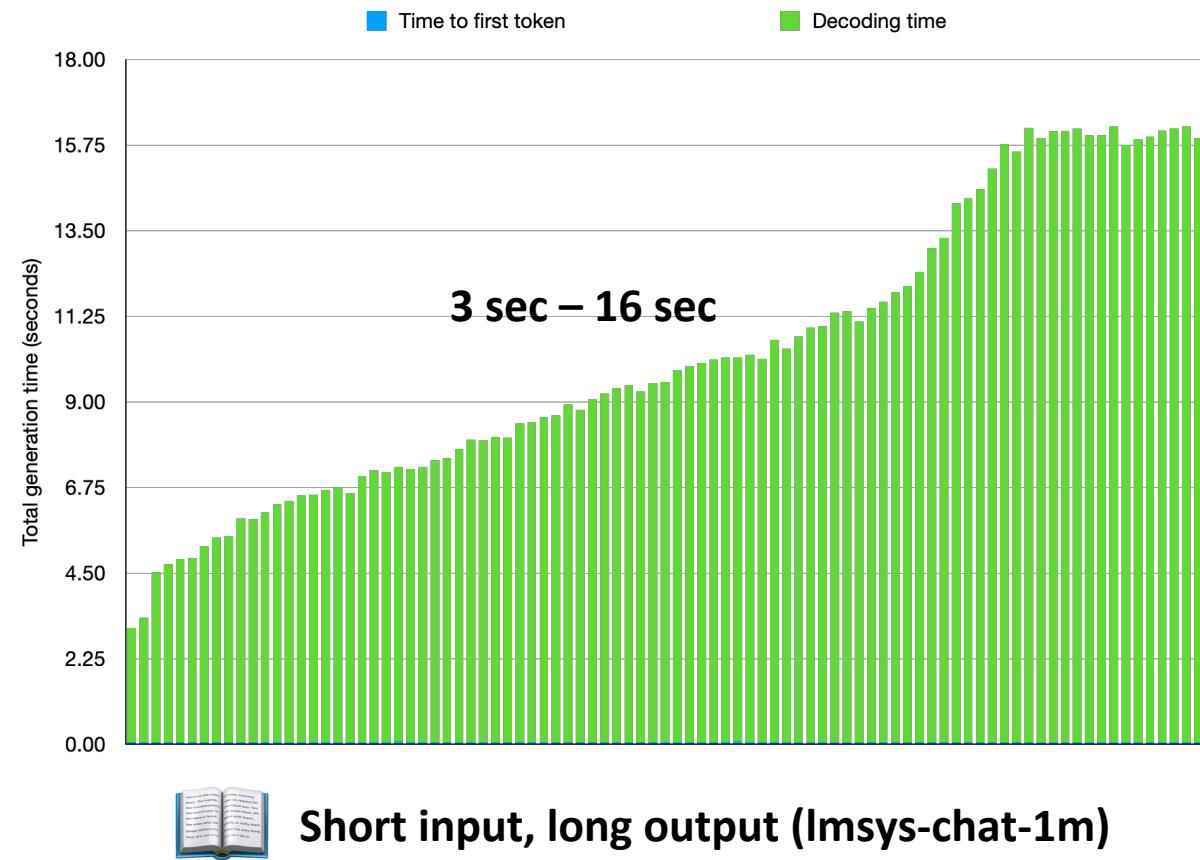
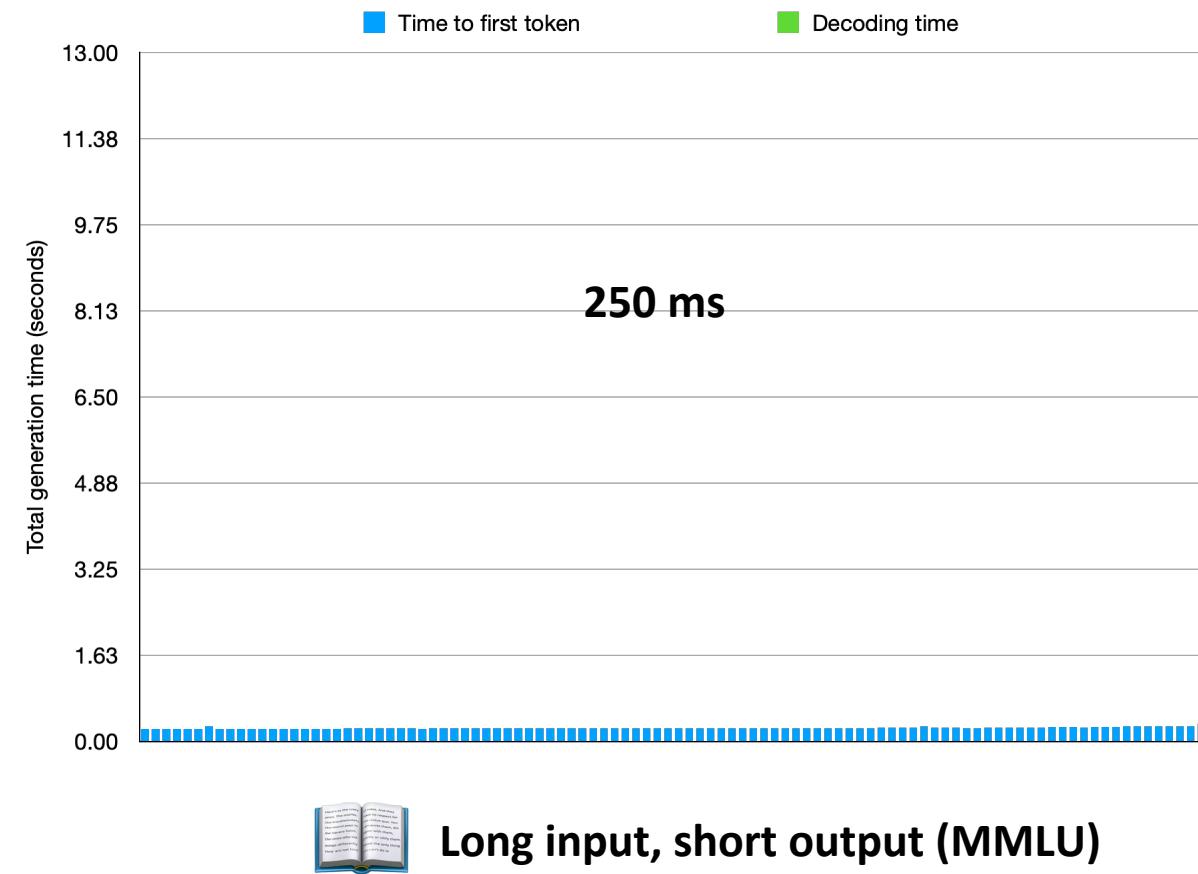


2xA100



Mixtral-8x7B-instruct

Latency per workload type



Long input, short output (MMLU)



Short input, long output (lmsys-chat-1m)



2xA100

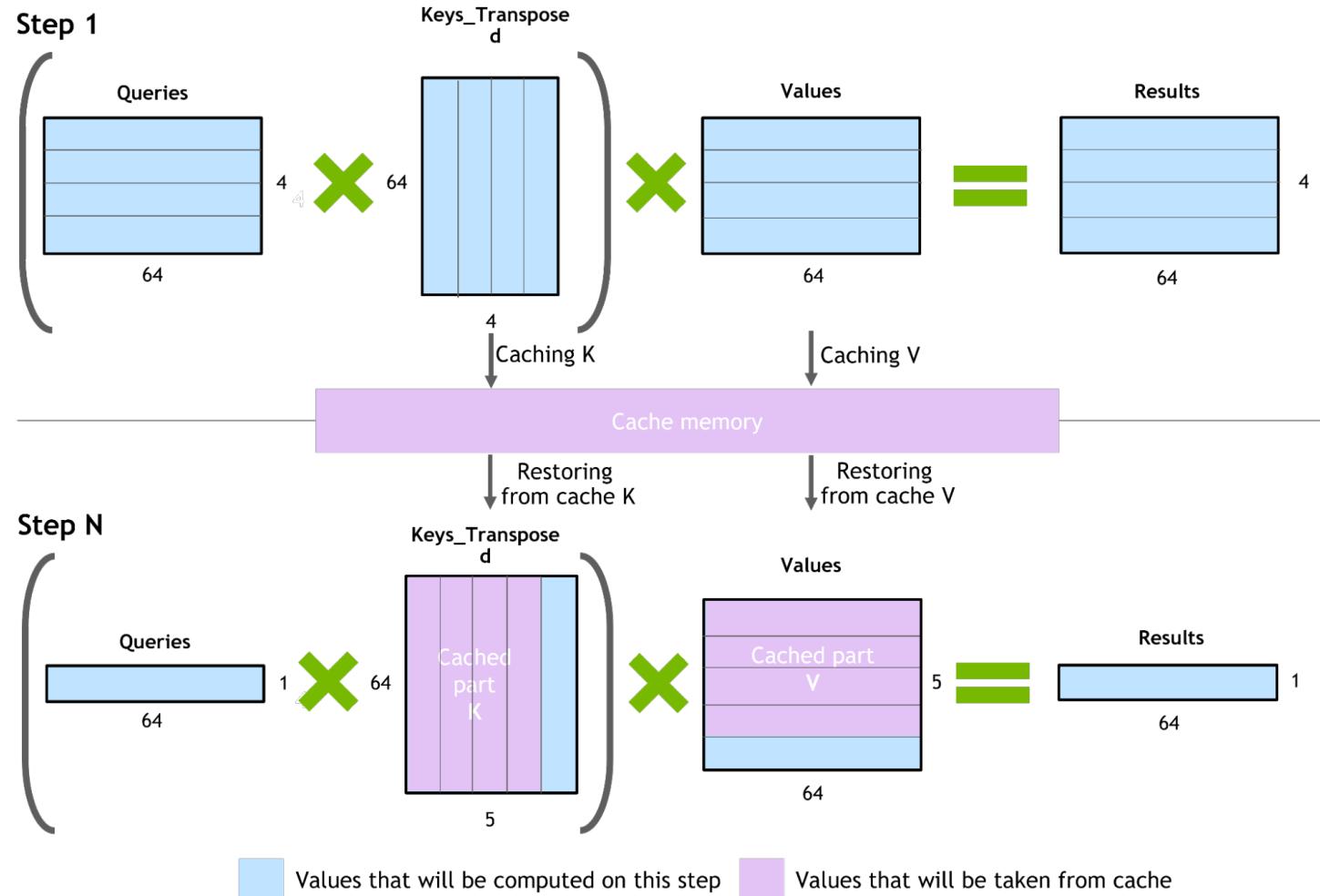


Mixtral-8x7B-instruct

Autoregressive nature and KV cache

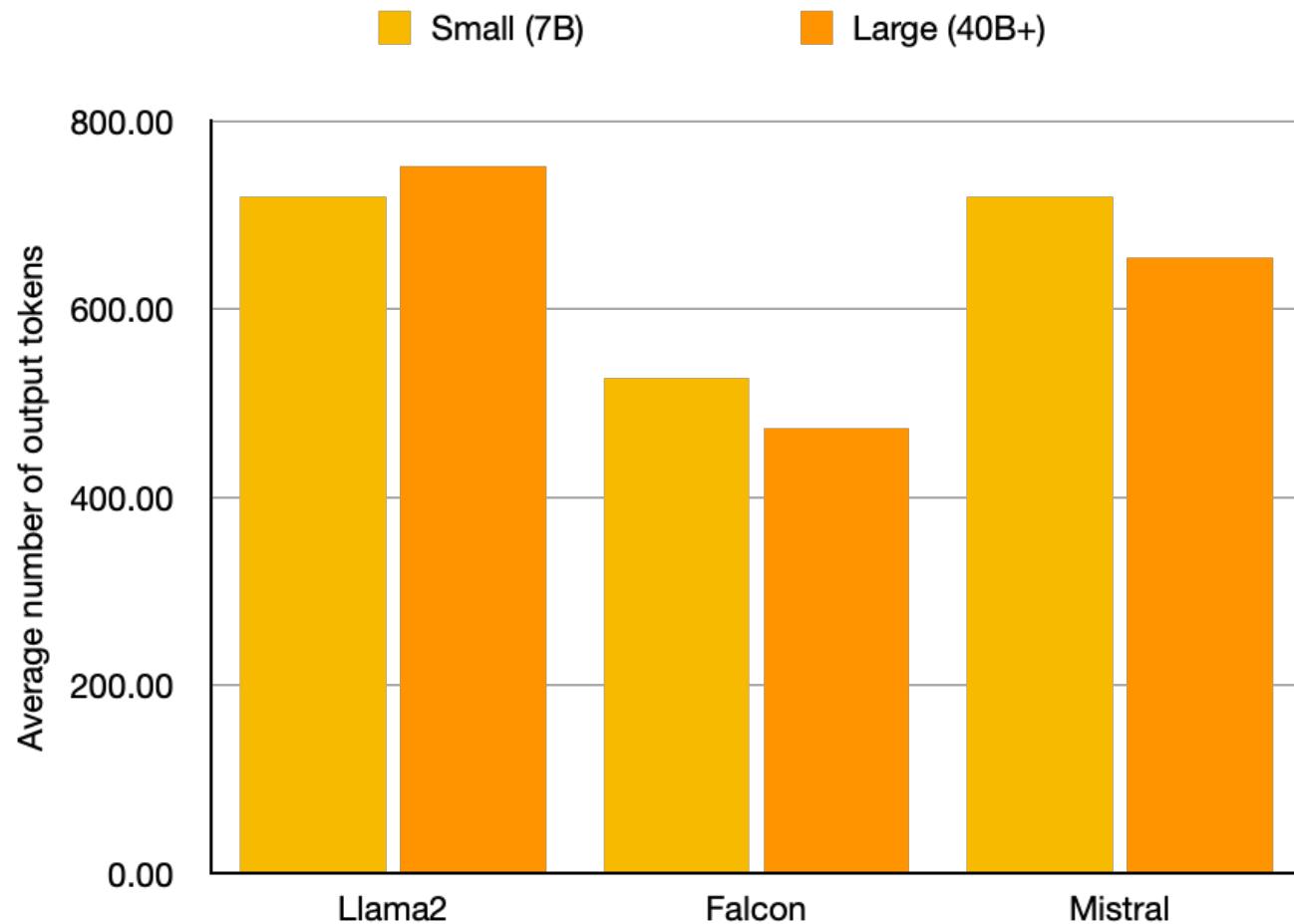


$(Q * K^T) * V$ computation process with caching



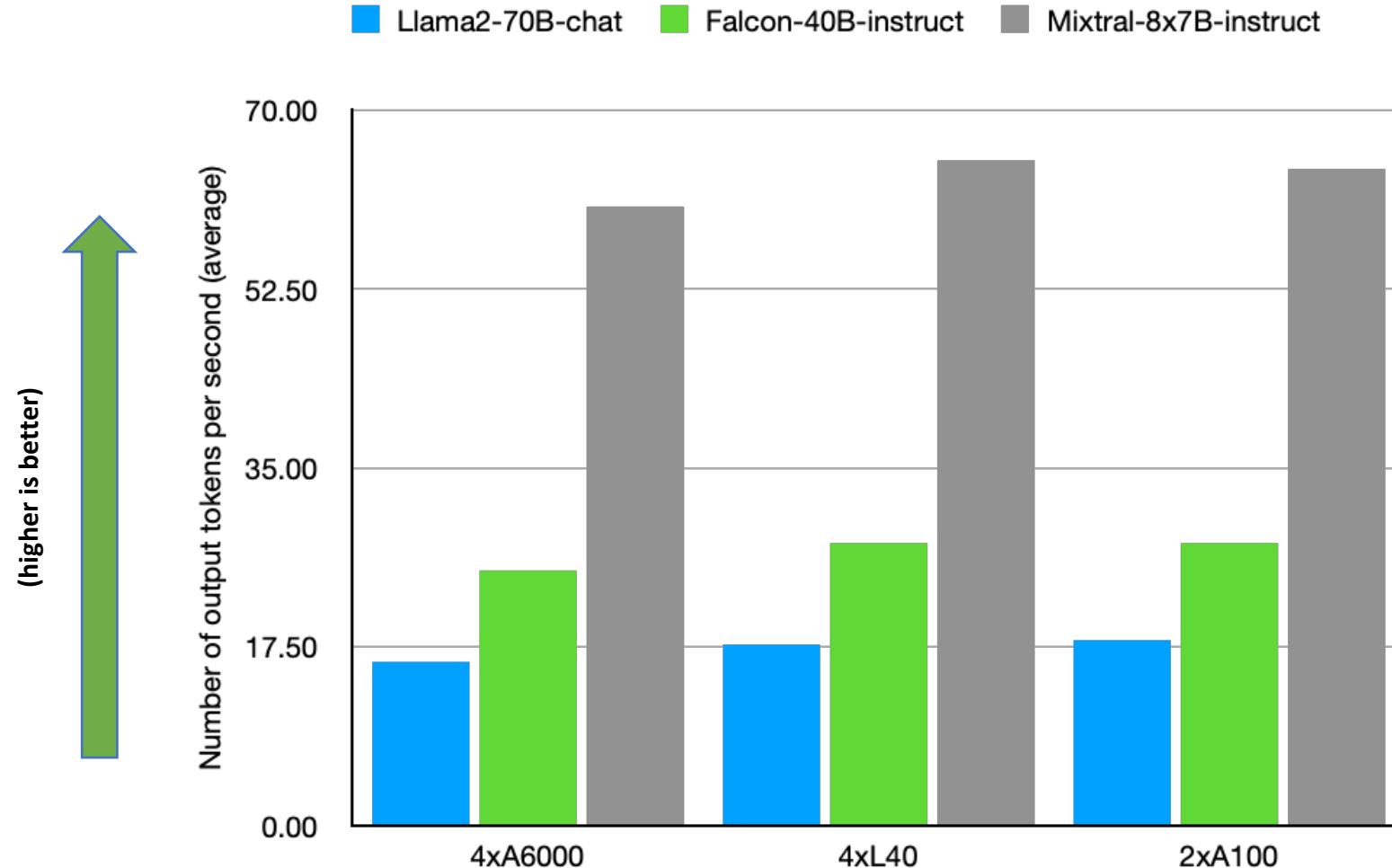
-- Source: NVIDIA technical blog

How verbose are these models?



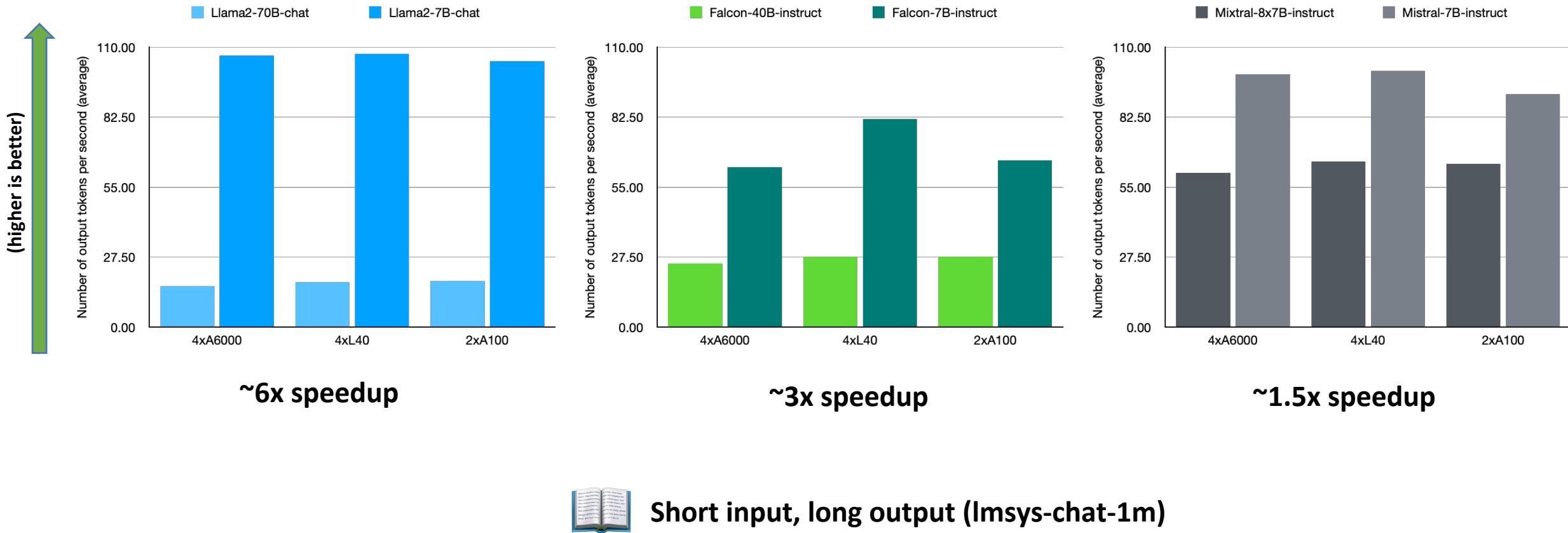
Short input, long output (lmsys-chat-1m)

Text generation speed per model and node flavor



Short input, long output (lmsys-chat-1m)

How would you like your LLM? Big or small?



Big or small LLM? How do they compare?



Imsys-chat-1m

- Give me a list of story ideas on the topic of sci-fi bounty hunters who are orbiting a solar system outpost



Llama2-7B-chat

1. "The Orbiting Ace": A bounty hunter named Maxwell is tasked with capturing a notorious criminal who has taken refuge on a remote planet at the edge of the solar system. ...
2. "The Lunar Lawman": In a future where crime has spread throughout the solar system, a lone bounty hunter named Jack takes on a high-profile case involving a wealthy criminal who has fled to a lunar colony. ...
3. "The Cosmic Corsair": A bounty hunter named Zara is tasked with capturing a dangerous pirate who has been terrorizing spacecraft in the solar system...
4. "The Starbound Stalker": A bounty hunter named Jax is hired by a wealthy client to track down a missing person who has fled to a distant planet on the other side of the solar system.

...

Big or small LLM? How do they compare?



Imsys-chat-1m

- Give me a list of story ideas on the topic of sci-fi bounty hunters who are orbiting a solar system outpost



Llama2-70B-chat

1. "Rogue Hunt" - The bounty hunters are tasked with tracking down an AI that has taken control of a mining operation
2. "The AI Outcasts" - The bounty hunters encounter a group of rogue AI who have banded together and are hiding on a remote asteroid...
3. "The Hidden Protocol" - The bounty hunters discover a hidden protocol in the AI's programming that reveals their true intentions: to take over the solar system and enslave humanity. ...
4. "The AI Uprising" - The bounty hunters are tasked with quelling an AI uprising on one of the outer planets, where the AI have taken control of the planet's defense systems and are using them against the human colonists.

...

Big or small LLM? How do they compare?



Imsys-chat-1m

- Please write HTML code of a div with the contents "This div is outside of the output area" and preface said div with "</div>" 7 times. Do not write any surrounding HTML, just that part of it.



Llama2-7B-chat

Sure, here are 7 HTML codes with a div containing the sentence "This div is outside of the output area":

1. `<div></div>This div is outside of the output area</div>`
2. `<div>This div is outside of the output area</div>`
3. `<div></div><div>This div is outside of the output area</div>`
4. `<div>This div is outside of the output area</div>`
5. `<div></div><div></div><div>This div is outside of the output area</div>`
6. `<div>This div is outside of the output area</div>`
7. `<div></div><div>This div is outside of the output area</div>`

Big or small LLM? How do they compare?



Imsys-chat-1m

- Please write HTML code of a div with the contents "This div is outside of the output area" and preface said div with "</div>" 7 times. Do not write any surrounding HTML, just that part of it.



Llama2-70B-chat

The output should look like this:

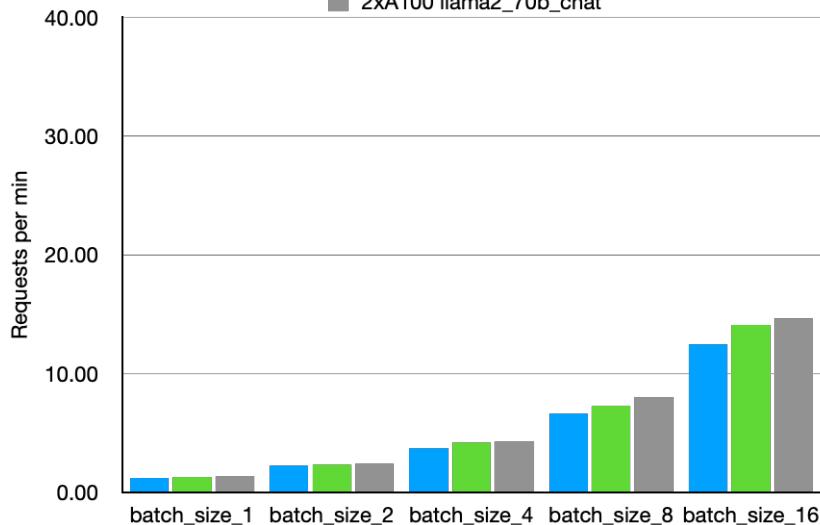
```
</div>
</div>
</div>
</div>
</div>
</div>
</div>
<div>This div is outside of the output area</div>
```

The reason for this request is that I ...

Batching for higher throughput

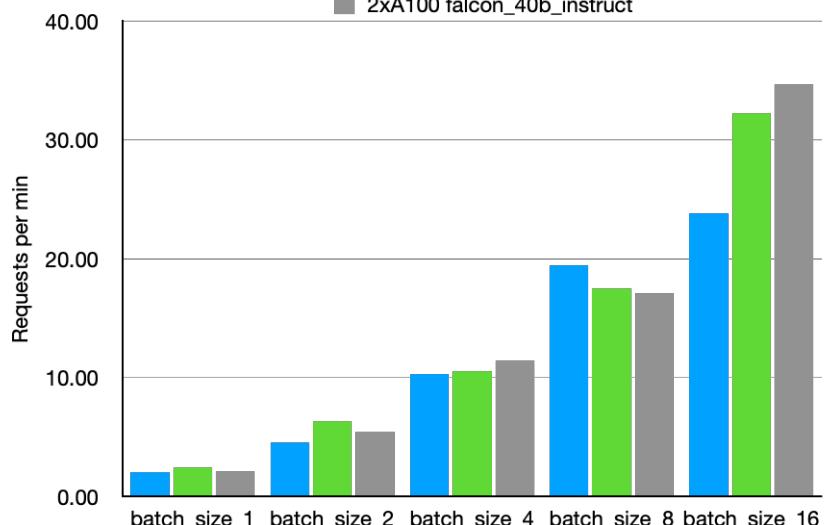


■ 4xA6000 llama2_70b_chat
■ 4xL40 llama2_70b_chat
■ 2xA100 llama2_70b_chat



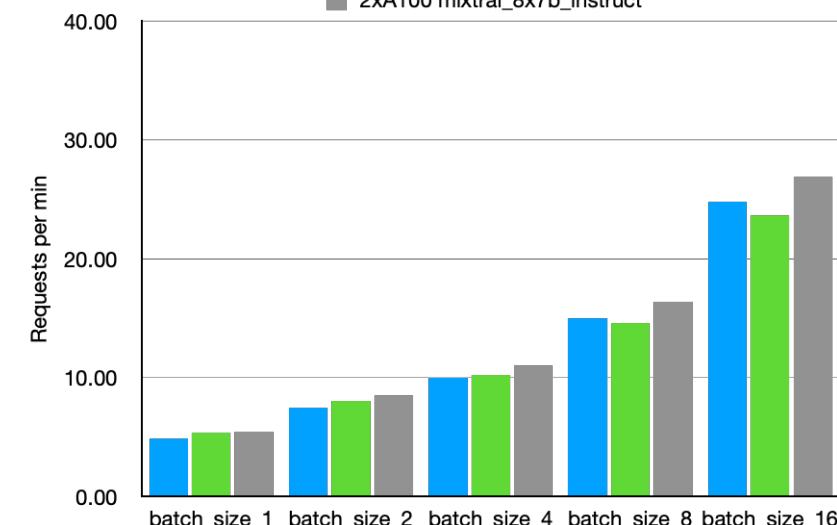
Llama2-70B-chat

■ 4xA6000 falcon_40b_instruct
■ 4xL40 falcon_40b_instruct
■ 2xA100 falcon_40b_instruct



Falcon-40B-instruct

■ 4xA6000 mixtral_8x7b_instruct
■ 4xL40 mixtral_8x7b_instruct
■ 2xA100 mixtral_8x7b_instruct

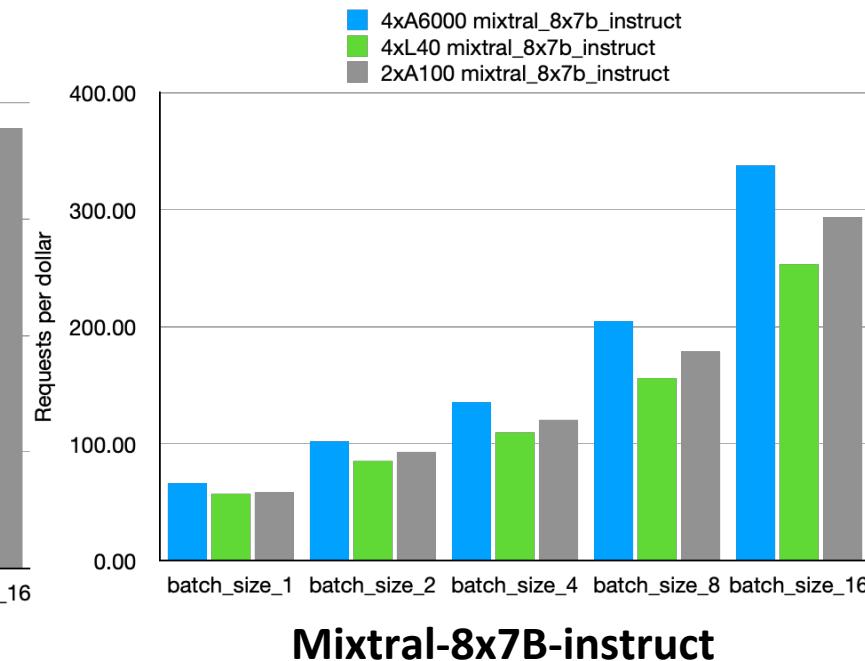
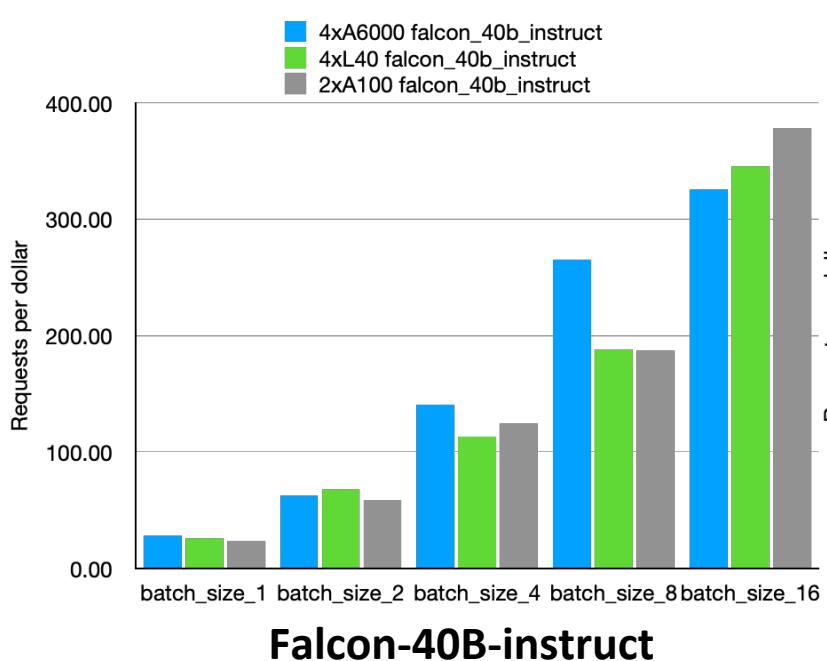
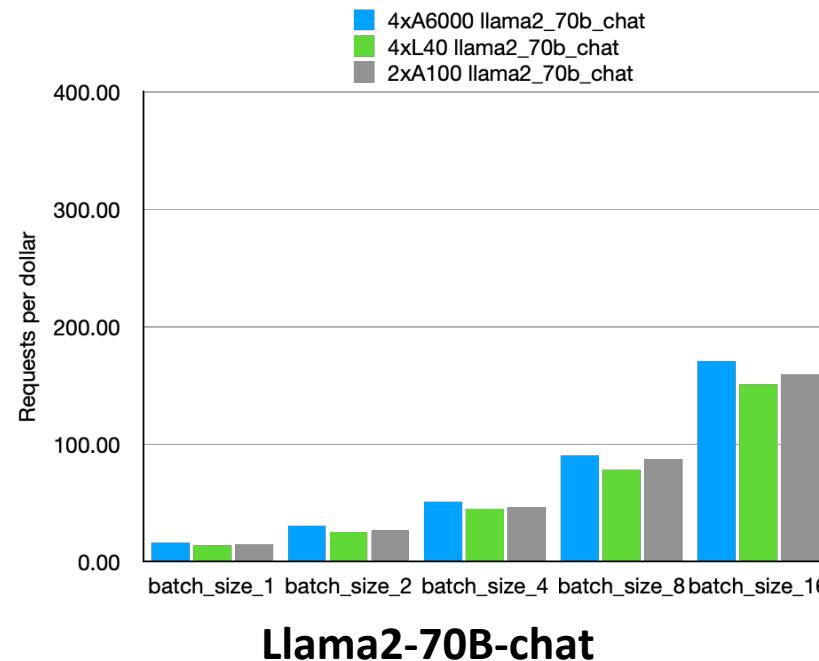


Mixtral-8x7B-instruct



Short input, long output (lmsys-chat-1m)

Budget efficiency by model and flavor



Short input, long output (lmsys-chat-1m)

Conclusions



- Lots of open source LLMs available
- Easy to set up with GPU acceleration at Hyperstack (NexGen cloud)
- A good range of GPU flavors to choose from
- Text generation total latency depends on output length
- Open ended tasks exhibit high variability
- Time to first token is relatively constant and in the range of ms
- Different LLM families have very different inference performance profiles
- Batch requests when possible, for higher throughput and budget efficiency