



# Knowledge Base for Everyone

## Crafting RAG with NVIDIA GPUs



**Corey Lowman**

*they/them*



**Chuan Li**

*he/him*

# ML-Times

←  
4 days ago

Mar 12, 2024

→  
2 days ago

Y OpenAI – transformer debugger release	497
Y Building Meta's GenAI infrastructure	379
Y Diffusion models from scratch, from a new theoretical perspective	356
Y Devin: AI Software Engineer	344
Y Simpson's paradox	332
Y Stealing Part of a Production Language Model	253
Y Behind the Compute: Benchmarking Compute Solutions	251
Y Show HN: Prompts as WASM Programs	199
Y Is Cosine-Similarity of Embeddings Really About Similarity?	190
Y (How to Write a (Lisp) Interpreter (In Python)) (2010)	186
Y [D] How does Gemini 1.5 Pro recall information in 10M context?	147
Y [R] ShortGPT: Layers in Large Language Models are More Redundant Than You Expect	116
Y Among the A.I. doomsayers	108



<news.lambdalabs.com/about>

- **ML Times revolutionizes AI news consumption** by harnessing advanced summarization technology to provide **key insights with unparalleled efficiency.** <sup>1</sup>
- **Designed to eliminate fluff**, ML Times targets the issue of **overly verbose content**, ensuring that readers have immediate access to the heart of AI developments. <sup>2</sup>
- Through **LLM summarization**, it offers a concise way to **sift through the noise** in AI news, presenting essential information without the bulk, aiming for a **user-driven experience** where content is consumed on demand. <sup>3</sup>

# Outline

01

**Build ML-Times**

02

**Open Source Models  
& NVIDIA GPUs**

# Build ML-Times

# What is a good RAG app?

- 1 Doesn't Change **User behavior**
- 2 Earns User **Trust**
- 3 Provides Meaningful **Content**

# Can we build something that WE want to use?

- Extraordinary amount of ML news/papers these days
- Across many different sites

## How do we keep up with all of it?

- A Slack bot that scrapes & filters news for us!

The image is a collage of screenshots from several websites:

- OpenAI**: Shows a sidebar with "Research" and "API" options, and a "Hacker News" feed at the bottom.
- Hugging Face**: Shows a search bar and a "Latest News" section.
- Meta**: Shows a "Latest News" section featuring an illustration of a shield with a gear inside.
- stability.ai**: Shows a sidebar with "Home", "Explore", "Notifications", and "Messages".
- Google DeepMind**: Shows a news article about "Introducing Gemini: our largest and most capable AI model".
- arXiv**: Shows a "Artificial Intelligence" section and a "Authors and titles for recent submissions" list.

# Does this meet our checklist?

 newsy APP 10:30 AM

Here's the latest news from today for you!

**HackerNews:**

1. [Comments] How to think about OpenAI's rumored (and overhyped) Q\* project
2. [Comments] We investigated France's mass profiling machine
3. [Comments] Exploring Quantum Computing with JavaScript: A Beginner's Guide
4. [Comments] Amazon's humanoid robots will eventually cost only \$3 per hour to operate

Checked 24 posts.

**/r/MachineLearning:**

1. [Comments] (+177) [D] Thoughts on Mamba?
2. [Comments] (+70) [D] Undergrad contemplating a Machine Learning PhD, but worried about what that truly entails.

Checked 2 posts.

**OpenAI Blog:**

No posts from today.

**StabilityAI Blog:**

1. Behind the Compute: Building the New AI Supercomputer
2. Introducing StableLM Zephyr 3B: A New Addition to StableLM, Bringing Powerful LLM Assistants to Edge Devices

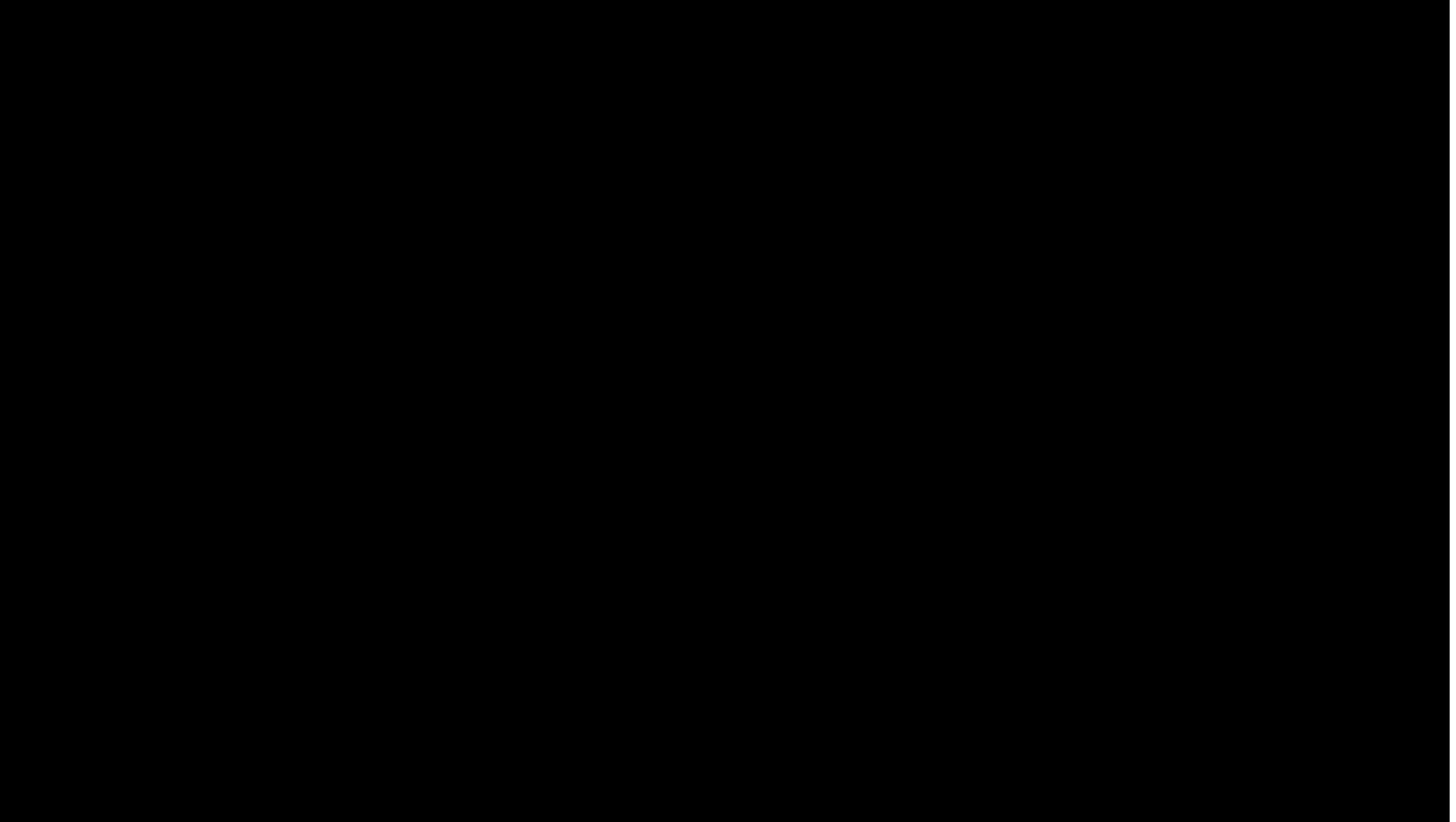
**Deepmind Blog:**

1. Google DeepMind at NeurIPS 2023

**arxiv AI papers:**

1. Generating Illustrated Instructions

- User behavior ✓
- Trust - we'll get to this 😅
- Content ✓





# This works great... let's deploy it!

- Asked Lambda employees to use it
- First user immediately asks follow up question 😊
  - ◆ “@newsy summarize [link]”
  - ◆ And newsy answered!
- Great sign!
  - ◆ User behavior 
  - ◆ Trust 
  - ◆ Content 

@newsy summarize  
[https://www.reddit.com/r/MachineLearning/comments/17u2cbo/r\\_pmmf\\_provider\\_maxmin\\_fairness\\_reranking\\_in/](https://www.reddit.com/r/MachineLearning/comments/17u2cbo/r_pmmf_provider_maxmin_fairness_reranking_in/)

newsy APP 25 days ago  
The discussion on [R] P-MMF: Provider Max-min Fairness Re-ranking in Recommender System at /r/MachineLearning is centered around:

newsy APP 25 days ago  
And here's the summary for you:

- This paper addresses the issue of fairness in recommender systems from the perspective of providers.
- It introduces the concept of max-min fairness (MMF) as a better metric for provider fairness, aiming to improve the utility of the worst providers in weak market positions.
- The paper proposes a method called P-MMF that balances user preferences and provider fairness in online recommendation scenarios, while maintaining low computational costs.

newsy APP 25 days ago  
I wasn't able to find a discussion on HackerNews for this post.

newsy APP 25 days ago  
Here's all the discussion happening on Twitter

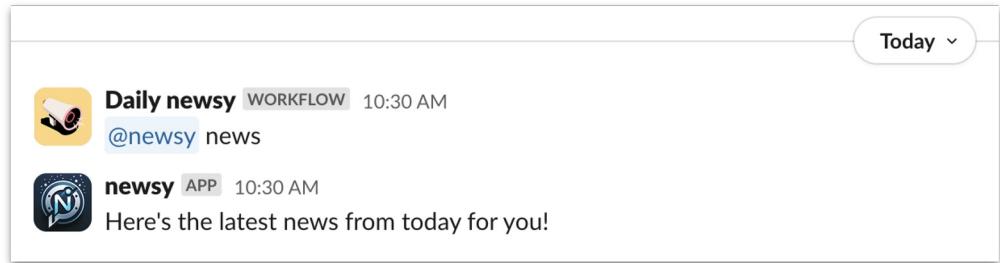
25 days ago  
What is the fairness metric?



# So do we use it?

→ Every day!

- ◆ User behavior ✓
- ◆ Content ✓
- ◆ Trust - ↗



3. Baking time: The original recipe bakes the bars for 15 minutes, while the Two Peas & Their Pod recipe bakes the bars for 10 minutes for the crust and an additional 15 minutes for the filling.

# news.lambdalabs.com

 ML Times



+ Summarize article

←  
Yesterday

Mar 18, 2024

 Grok



845

 LLM4Decompile: Decompiling Binary Code with LLM



315

 xAI releases Grok-1 [N]



97

 [D] I don't understand how backprop works on sparsely gated MoE



70

 Compressing Images with Neural Networks



59

 MANATEE(lm): Market Analysis based on language model architectures



15

 YannickKilcher - [ML News] Devin AI Software Engineer | GPT-4.5-Turbo LEAKED | US Gov't Report: T...

## Simpson's paradox

315

- Simpson's paradox reveals how aggregated data can mask or reverse trends seen in separated group data, challenging assumptions in social and medical sciences. <sup>1</sup>
- Famous instances include UC Berkeley's gender bias in admissions and kidney stone treatment effectiveness, demonstrating the paradox's impact on real-world decisions and perceptions. <sup>2</sup>
- The paradox underscores the importance of considering confounding variables and causal relationships in statistical analysis, highlighting the potential for misleading conclusions without careful examination. <sup>3</sup>

← →  [Highlight Link](#) Similarity: 0.685814083

*Critics of the apparent Simpson's paradox also argue that the focus on the paradox may distract from more important statistical issues, such as the need for careful consideration of confounding variables and causal relationships when interpreting data.[32] Despite these criticisms, the apparent Simpson's paradox remains a popular and intriguing topic in statistics and data analysis.*



[LEARN MORE](#)



[SOURCE](#)



[DISCUSSIONS](#)

# Open Source Models & NVIDIA GPUs

# Open Source Models: Quality



newsy APP 19 hours ago

[D] 3 years doing ML, no success yet. Is it common? discusses:

- They have tried multiple techniques and approaches, but none have yielded satisfactory results.
- Despite participating in online competitions and achieving decent rankings, their attempts to improve their methods have been unsuccessful and frustrating.

/r/MachineLearning has a +116 discussion centered around:

1. (+121) Re-evaluate and simplify approach.
2. (+46) Importance of literature review and validation testing.
3. (+45) Not effective in ML techniques.

## GPT-3.5 Turbo



newsy\_openchat APP 19 hours ago

[D] 3 years doing ML, no success yet. Is it common? discusses:

- \* They've tried different techniques and methods, but the results have been poor or worse than the baseline
- \* They participate in online competitions, but even when they try to improve their methods, they still struggle

/r/MachineLearning has a +117 discussion centered around:

1. (+123) Reevaluate decision making, try simpler methods.
2. (+47) Slow progress, extensive review, team up.
3. (+48) Trial and error approach.

## OpenChat

# Open Source Models: Evaluation

url	GPT-turbo-3.5	OpenChat	winner	details
<a href="https://blogs.nvidia.com/blog/nvidia-drive-partners-innovations-video">https://blogs.nvidia.com/blog/nvidia-drive-partners-innovations-video</a>	<p>- The automotive industry is being transformed by cutting-edge technologies like AI and computing power.\n- NVIDIA DRIVE partners are working on cutting-edge innovations in automated and autonomous driving.\n* Computing power and AI are shaping the future of transportation.\n* NVIDIA's DRIVE platform is helping industry leaders redefine transportation.</p>	<p>* NVIDIA DRIVE partners are working on cutting-edge innovations in automated and autonomous driving.\n* Computing power and AI are shaping the future of transportation.\n* NVIDIA's DRIVE platform is helping industry leaders redefine transportation.</p>	GPT	<p>The best summary provides a more detailed and comprehensive overview of the key points from the original article, including the importance of partnerships, the critical role of computing power and AI, and the significance of simulation in the development of autonomous systems. It captures specific insights from industry leaders, which adds credibility and depth to the summary. The worst summary, while concise, lacks the detail and specificity that gives readers a deeper understanding of the innovations and challenges in autonomous driving technology.</p>

# Open Source Models: Evaluation

## General evaluation prompt

You are editor at a reputable online newspaper about Machine Learning.  
You are tasked with choosing between two summaries of an article to publish.  
{reference\_section}

Summary A:  
{summary\_a}

Summary B:  
{summary\_b}

---

Which of these two summaries do you choose to publish and why?

**OpenChat won 0% v.s. GPT-3.5-turbo**

# Open Source Models: Evaluation

## General evaluation prompt

You are editor at a reputable online newspaper about Machine Learning.  
You are tasked with choosing between two summaries of an article to publish.  
{reference\_section}

Summary A:  
{summary\_a}

Summary B:  
{summary\_b}

---

Which of these two summaries do you choose to publish and why?

## Summary instructions

Generate a bulleted list that summarizes the main points from the Article. Prioritize:

1. Reducing the mental burden of reading the summary.
  2. Conciseness over grammatical correctness.
  3. Ability for general audience to understand.
  4. Write the summary as if a 5 year old was the reader.
- Generate at most 3 bullet points.

**OpenChat won 0% v.s. GPT-3.5-turbo**

# Open Source Models: Evaluation

## General evaluation prompt

You are editor at a reputable online newspaper about Machine Learning.  
You are tasked with choosing between two summaries of an article to publish.  
{reference\_section}

Summary A:  
{summary\_a}

Summary B:  
{summary\_b}

---

Which of these two summaries do you choose to publish and why?

**OpenChat won 0% v.s. GPT-3.5-turbo**

## Evaluation prompt with summary instructions

You are an evaluator tasked with assessing which of two outputs best match the prompt:

The prompt:  
Generate a bulleted list that summarizes the main points from the Article. Prioritize:  
1. Reducing the mental burden of reading the summary.  
2. Conciseness over grammatical correctness.  
3. Ability for general audience to understand.  
4. Write the summary as if a 5 year old was the reader.  
Generate at most 3 bullet points.

The outputs to assess:

Summary A:  
{summary\_a}

Summary B:  
{summary\_b}

{reference\_section}

## Elo rating

gpt-4-0125	:	1167
gpt-4	:	988
<b>openchat_3.5</b>	:	<b>956</b>
gpt-3.5-turbo	:	889

# Open Source Models: Evaluation

## Be careful with your instructions

4. Write the summary as if a 5 year old was the reader.

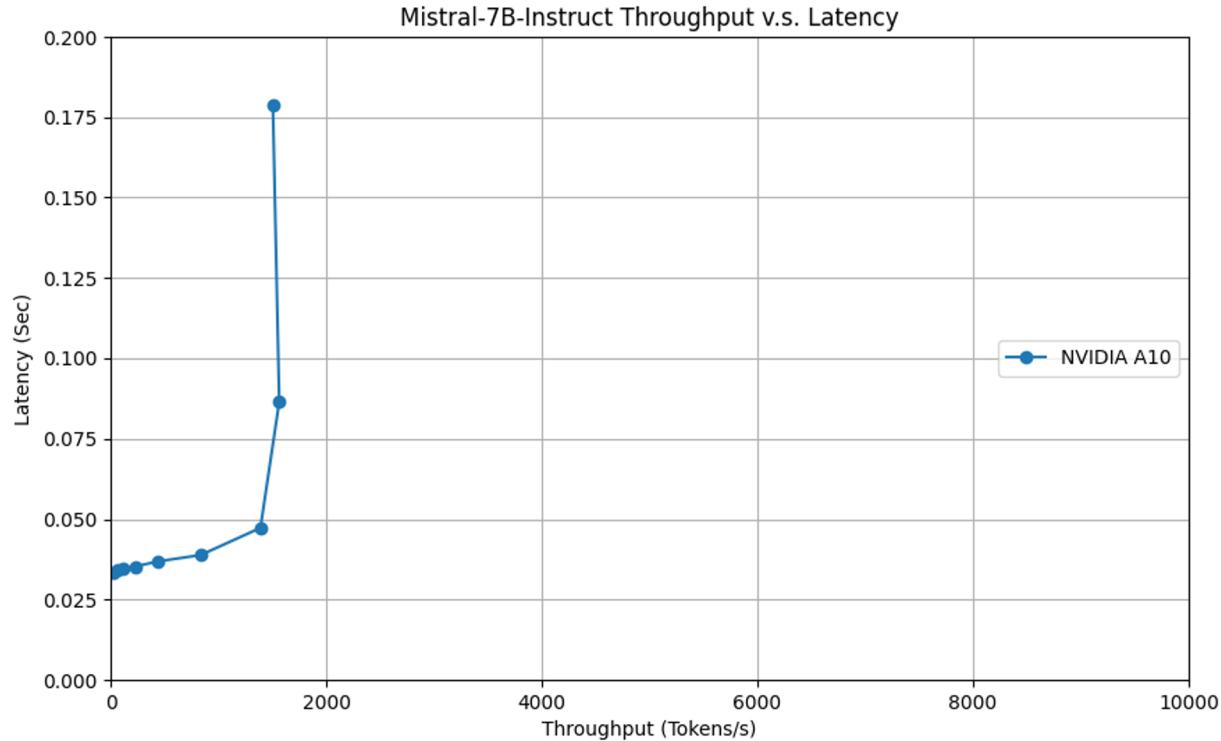
☆ Y Over 100k Infected Repos Found on GitHub

- A bad person is putting sneaky stuff in many computer things on a big website called GitHub. <sup>1</sup>
- This sneaky stuff can steal important information from people's computers. <sup>2</sup>
- The people in charge of the big website are trying to stop it, but it's still happening. <sup>3</sup>



190

# Open Source Models: Speed

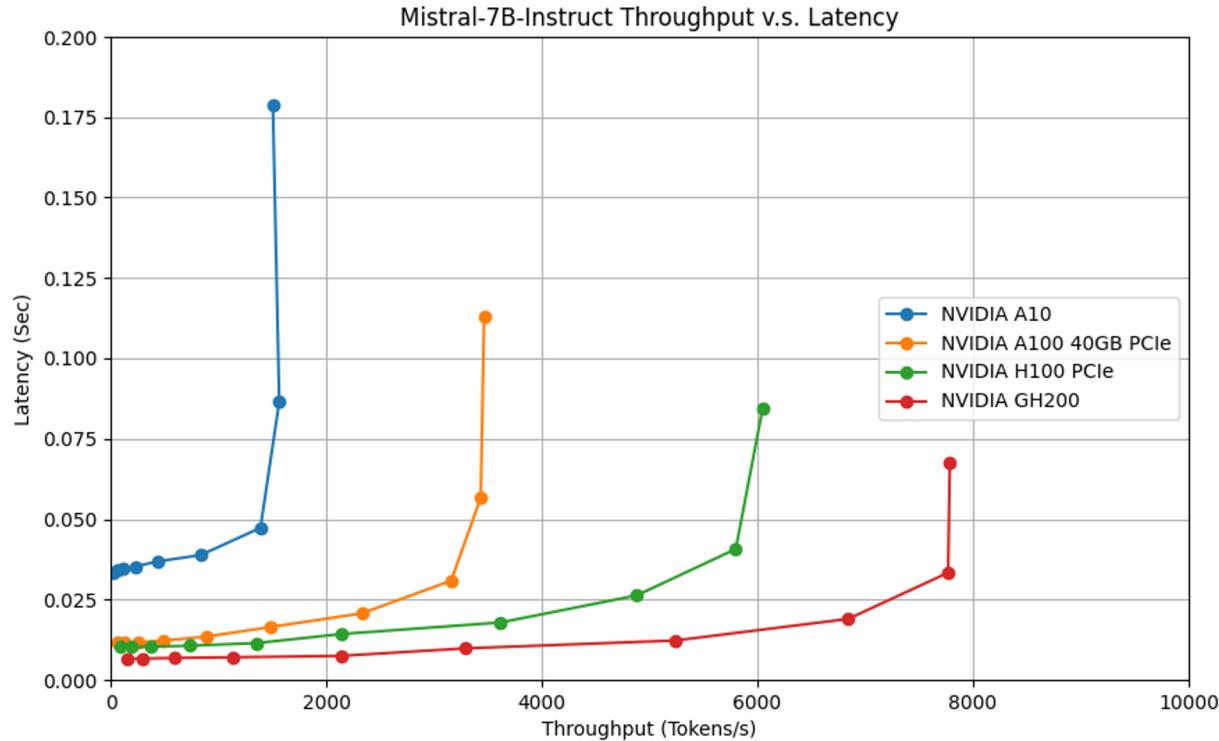


[vLLM inference benchmark](#)

**Max batch size**

NVIDIA A10: 256

# Open Source Models: Speed



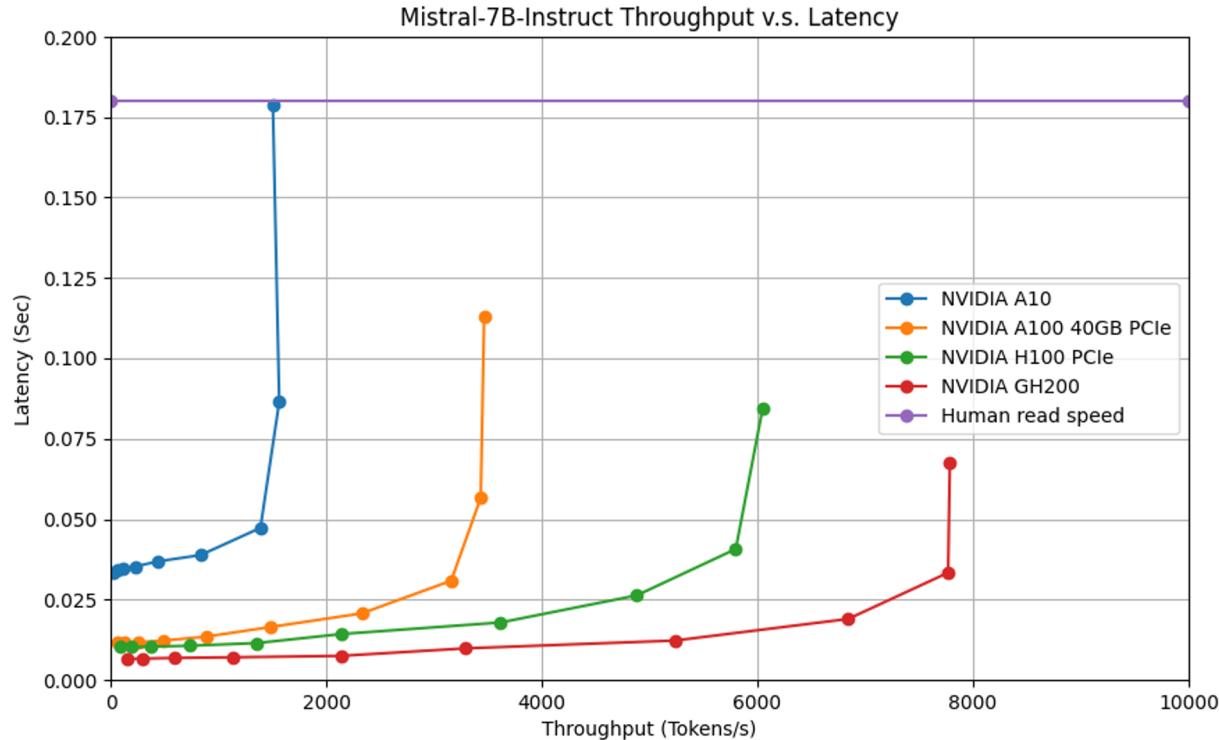
[vLLM inference benchmark](#)

**Max batch size**

NVIDIA A10: 256

Other GPUs: 512

# Open Source Models: Speed



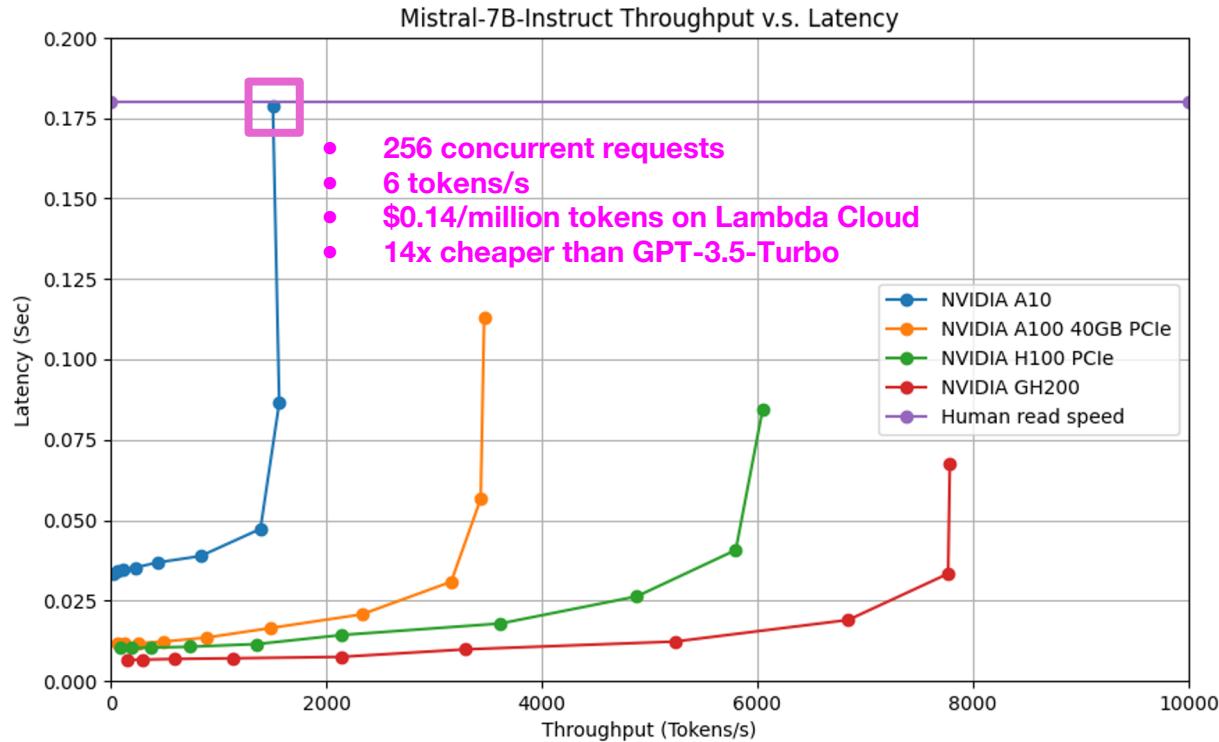
[vLLM inference benchmark](#)

**Max batch size**

NVIDIA A10: 256

Other GPUs: 512

# Open Source Models: Speed

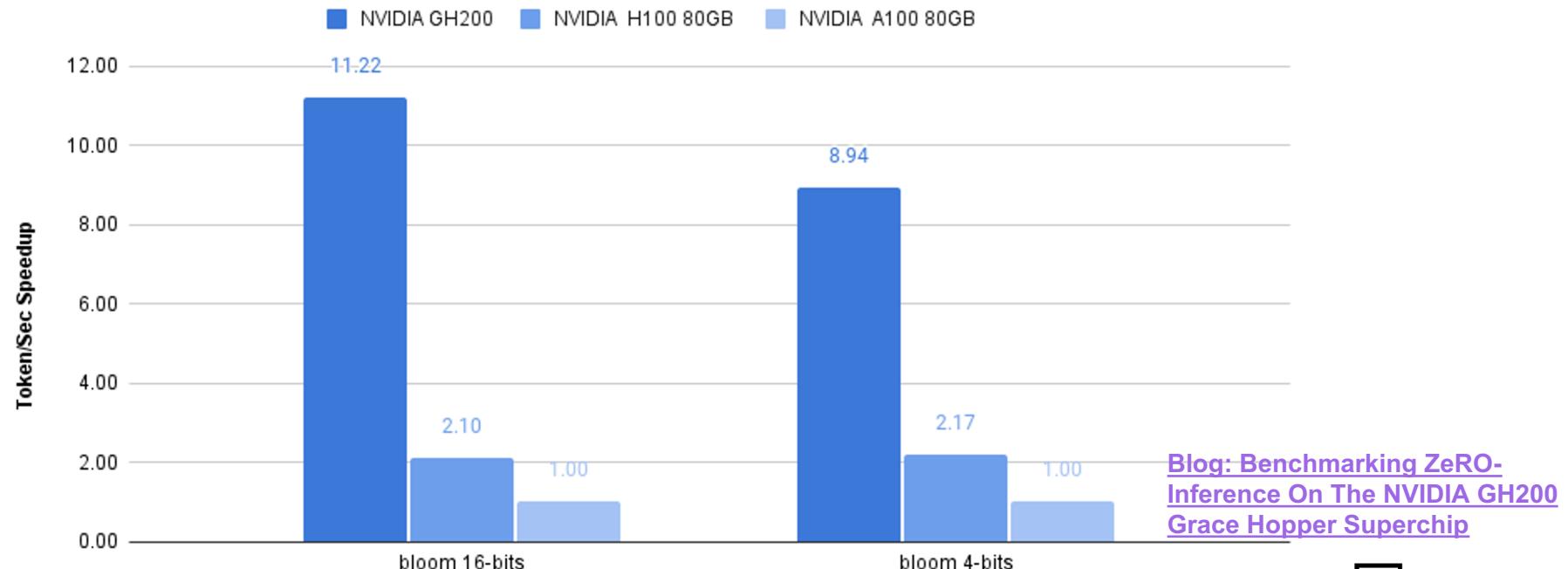


vLLM inference benchmark

**Max batch size**  
NVIDIA A10: 256  
Other GPUs: 512

# Open Source Models: NVIDIA GPU Considerations

## Bloom 176b inference Performance with a single NVIDIA GPU, batch size 8



# Data Safety: Protopia AI Stained Glass



## Data in Trust Zone

**Context:** Acme Inc spent 5000\$ on Jan 1st of 2023 on . They then spent 343.50 on 15/1. Followed by a credit in the amount of 1400 on 2/1. In february they also spent 200 on marketing and then 500 on sales efforts on 10th and 17th. Finally on 10/3 they spent 100.

**Instruction:** List spend dates and amount.



## Protected Data Sent to LLM

**Attempted Reconstruction of Context and Instruction:**

acsmezázspmozilla Ve th hundred Oriental£ absoluteJan luglio PseudST常□ duas fifteen background zu Op MachineTw spent drei)))) diesepling фев TurkTube' Pseud Ralphpor header Fac3 graduatedBY few creditSDK CanalGi信π dare fifteen hetJ curr□)))Relative]], £ solves

- Powered by Protopia AI patented technology
- Infrastructure Agnostic (on-premise, hybrid, cloud)

- Use-case Agnostic
- Complementary to Encryption



# Booth #616

[news.lambdalabs.com](http://news.lambdalabs.com)



**Corey Lowman**

*they/them*



**David Hartmann**

*he/him*



**Jeremy Hummel**

*he/him*



**Eole Cervenka**

*he/him*



**Chuan Li**

*he/him*