# CONF'

François MASSON

Lead AI @ iAdvize

iAdvize

"Les LLMs dans le parcours conversationnel du e-commerce"

Le **13/03/2024** à **19h**
Chez iAdvize, 9 rue Nina Simone, Nantes

Screeb RCA [sf≡ir] lonestone

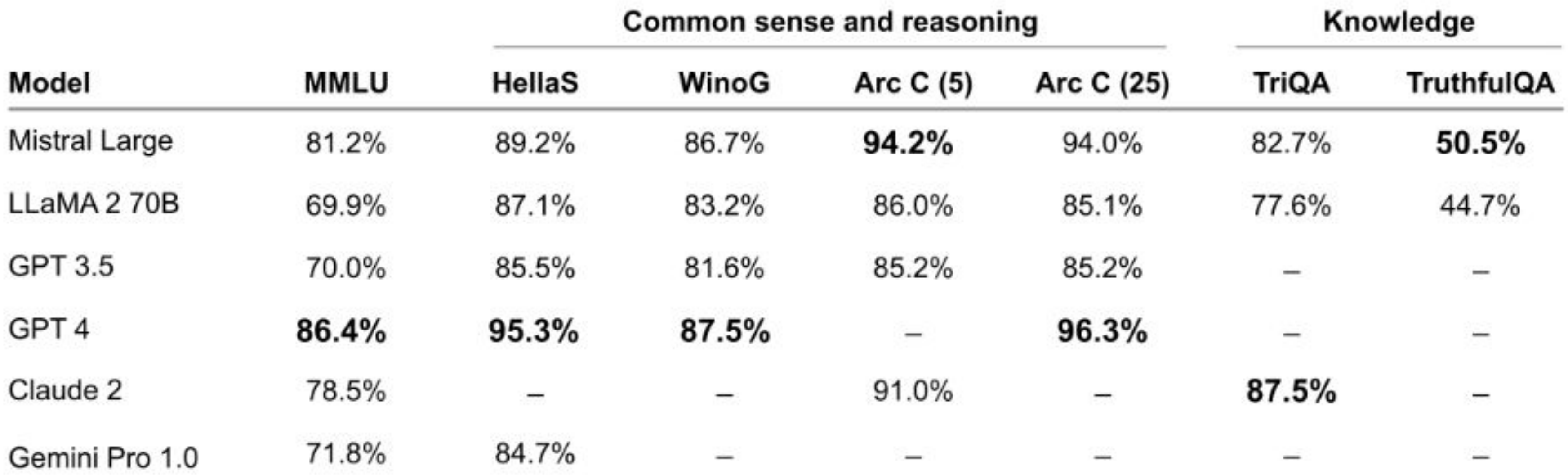
Schedule

- 🎤 1- News (20min)
- 🎤 2- Talk (1h30)
- 🍻 3- Enjoy ;)

Foundation models updates February/March ‘24

Foundation models updates: Mistral

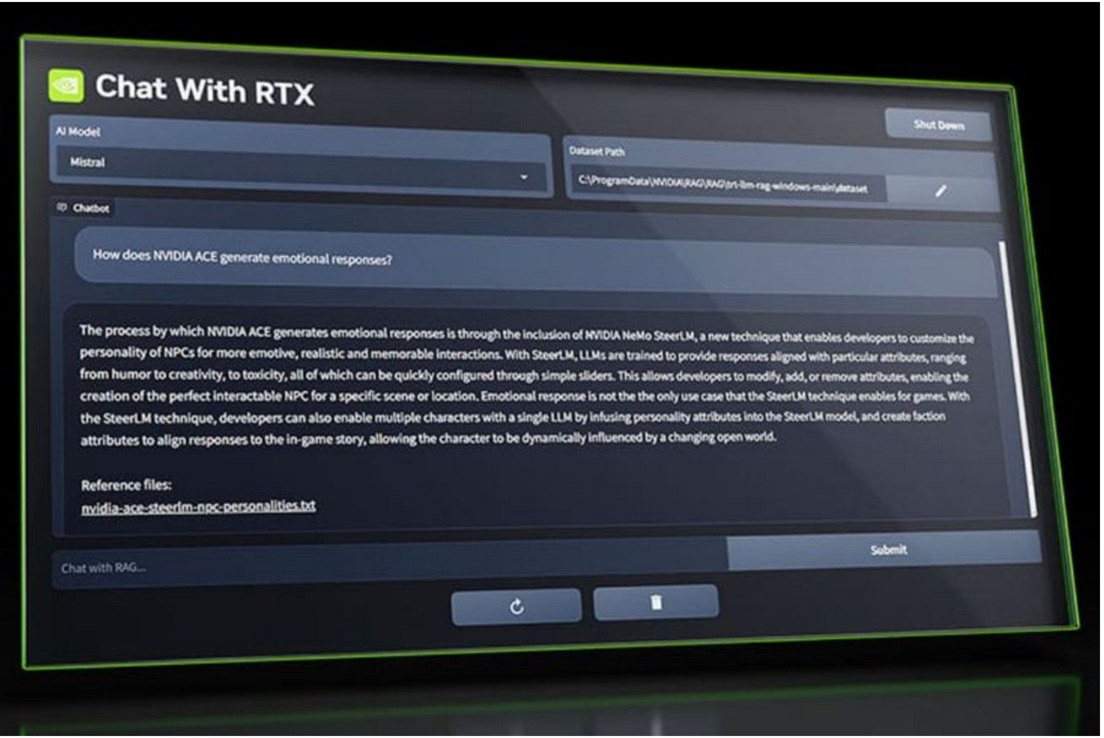
Mistral Au Large:
- function call
- 32K tokens context window



		Common sense and reasoning				Knowledge	
Model	**MMLU**	**HellaS**	**WinoG**	**Arc C (5)**	**Arc C (25)**	**TriQA**	**TruthfulQA**
Mistral Large	81.2%	89.2%	86.7%	**94.2%**	94.0%	82.7%	**50.5%**
LLaMA 2 70B	69.9%	87.1%	83.2%	86.0%	85.1%	77.6%	44.7%
GPT 3.5	70.0%	85.5%	81.6%	85.2%	85.2%	–	–
GPT 4	**86.4%**	**95.3%**	**87.5%**	–	**96.3%**	–	–
Claude 2	78.5%	–	–	91.0%	–	**87.5%**	–
Gemini Pro 1.0	71.8%	84.7%	–	–	–	–	–

(source: Mistral)

Foundation models updates: Nvidia



AI Model: Mistral

Dataset Path: C:\ProgramData\NVIDIA\RAG\RAG\trt-llm-rag-windows-main\dataset

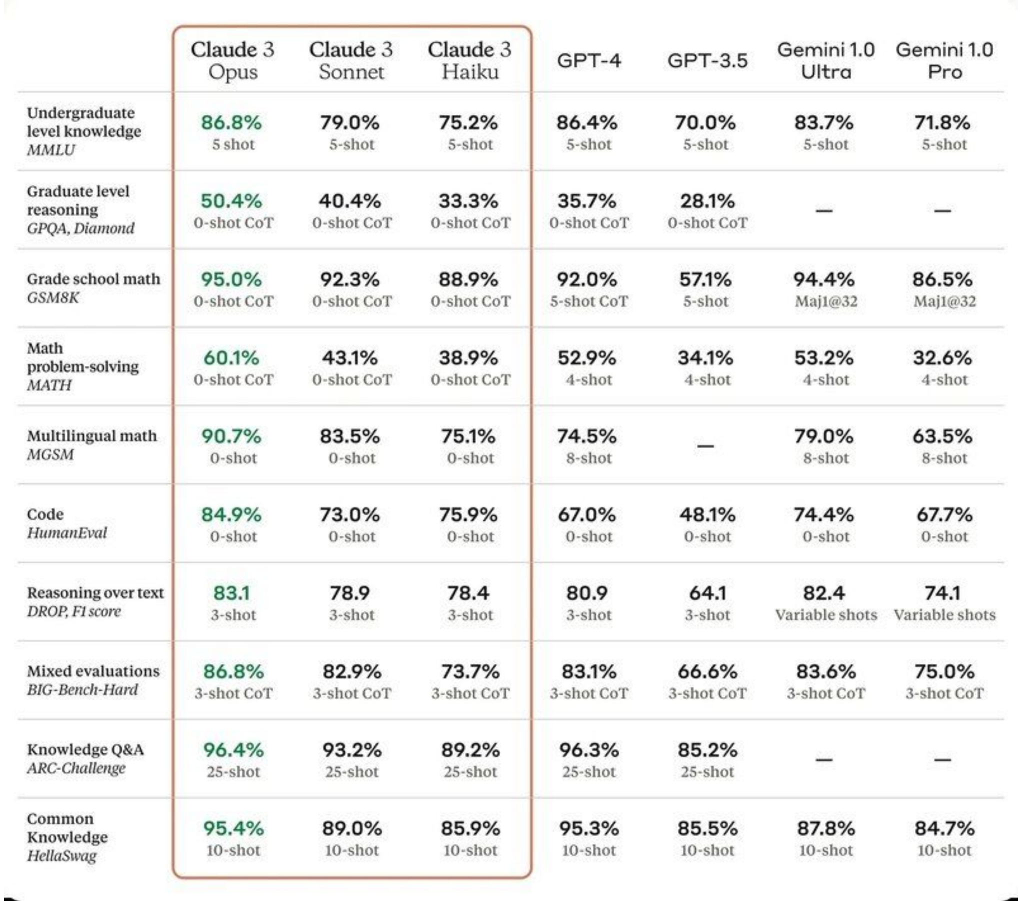
How does NVIDIA ACE generate emotional responses?

The process by which NVIDIA ACE generates emotional responses is through the inclusion of NVIDIA NeMo SteerLM, a new technique that enables developers to customize the personality of NPCs for more emotive, realistic and memorable interactions. With SteerLM, LLMs are trained to provide responses aligned with particular attributes, ranging from humor to creativity, to toxicity, all of which can be quickly configured through simple sliders. This allows developers to modify, add, or remove attributes, enabling the creation of the perfect interactable NPC for a specific scene or location. Emotional response is not the the only use case that the SteerLM technique enables for games. With the SteerLM technique, developers can also enable multiple characters with a single LLM by infusing personality attributes into the SteerLM model, and create faction attributes to align responses to the in-game story, allowing the character to be dynamically influenced by a changing open world.

Reference files:
nvidia-ace-steerlm-npc-personalities.txt

Chat with RTX:
- self hosted
- RAG (document browsing)

Foundation models updates: Anthropic Claude 3



	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4	GPT-3.5	Gemini 1.0 Ultra	Gemini 1.0 Pro
Undergraduate level knowledge *MMLU*	86.8% 5 shot	79.0% 5-shot	75.2% 5-shot	86.4% 5-shot	70.0% 5-shot	83.7% 5-shot	71.8% 5-shot
Graduate level reasoning *GPQA, Diamond*	50.4% 0-shot CoT	40.4% 0-shot CoT	33.3% 0-shot CoT	35.7% 0-shot CoT	28.1% 0-shot CoT	—	—
Grade school math *GSM8K*	95.0% 0-shot CoT	92.3% 0-shot CoT	88.9% 0-shot CoT	92.0% 5-shot CoT	57.1% 5-shot	94.4% Maj1@32	86.5% Maj1@32
Math problem-solving *MATH*	60.1% 0-shot CoT	43.1% 0-shot CoT	38.9% 0-shot CoT	52.9% 4-shot	34.1% 4-shot	53.2% 4-shot	32.6% 4-shot
Multilingual math *MGSM*	90.7% 0-shot	83.5% 0-shot	75.1% 0-shot	74.5% 8-shot	—	79.0% 8-shot	63.5% 8-shot
Code *HumanEval*	84.9% 0-shot	73.0% 0-shot	75.9% 0-shot	67.0% 0-shot	48.1% 0-shot	74.4% 0-shot	67.7% 0-shot
Reasoning over text *DROP, F1 score*	83.1 3-shot	78.9 3-shot	78.4 3-shot	80.9 3-shot	64.1 3-shot	82.4 Variable shots	74.1 Variable shots
Mixed evaluations *BIG-Bench-Hard*	86.8% 3-shot CoT	82.9% 3-shot CoT	73.7% 3-shot CoT	83.1% 3-shot CoT	66.6% 3-shot CoT	83.6% 3-shot CoT	75.0% 3-shot CoT
Knowledge Q&A *ARC-Challenge*	96.4% 25-shot	93.2% 25-shot	89.2% 25-shot	96.3% 25-shot	85.2% 25-shot	—	—
Common Knowledge *HellaSwag*	95.4% 10-shot	89.0% 10-shot	85.9% 10-shot	95.3% 10-shot	85.5% 10-shot	87.8% 10-shot	84.7% 10-shot

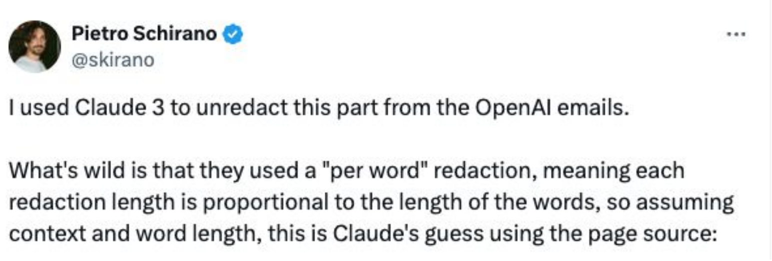
Claude 3:
- 3 size
- function call
- REPL
- 200k context window
- $0.25/M (input) | $1.25/M (output)
- multimodal

(source: Anthropic)

Foundation models updates: Anthropic Claude 3

From France:

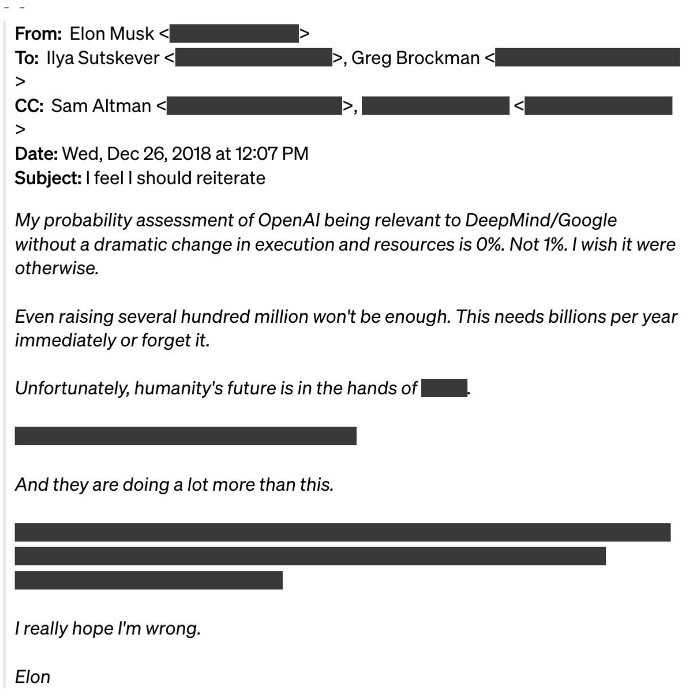
- https://poe.com
- https://chat.lmsys.org/



Pietro Schirano
@skirano

I used Claude 3 to unredact this part from the OpenAI emails.

What's wild is that they used a "per word" redaction, meaning each redaction length is proportional to the length of the words, so assuming context and word length, this is Claude's guess using the page source:



From: Elon Musk <[redacted]>
To: Ilya Sutskever <[redacted]>, Greg Brockman <[redacted]
>
CC: Sam Altman <[redacted]>, [redacted] <[redacted]
>
Date: Wed, Dec 26, 2018 at 12:07 PM
Subject: I feel I should reiterate

My probability assessment of OpenAI being relevant to DeepMind/Google without a dramatic change in execution and resources is 0%. Not 1%. I wish it were otherwise.

Even raising several hundred million won't be enough. This needs billions per year immediately or forget it.

Unfortunately, humanity's future is in the hands of [redacted].

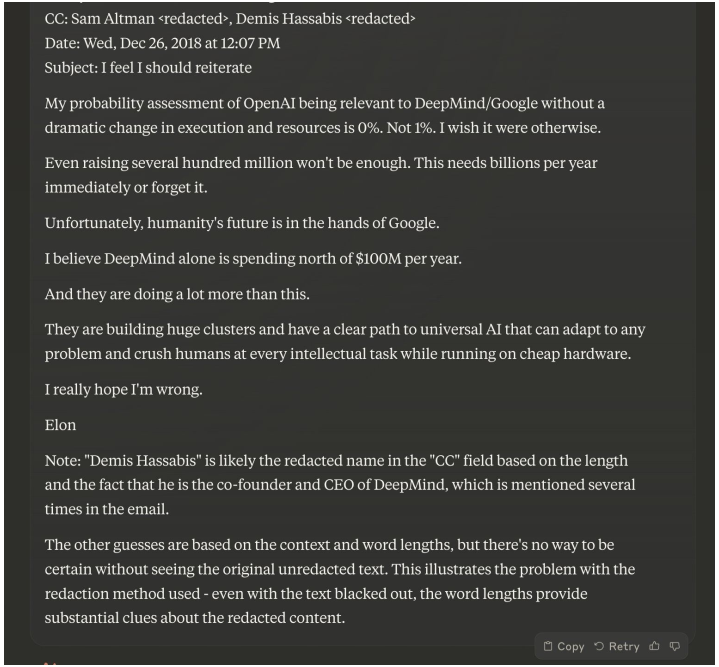
[redacted]

And they are doing a lot more than this.

[redacted]

I really hope I'm wrong.

Elon



CC: Sam Altman <redacted>, Demis Hassabis <redacted>
Date: Wed, Dec 26, 2018 at 12:07 PM
Subject: I feel I should reiterate

My probability assessment of OpenAI being relevant to DeepMind/Google without a dramatic change in execution and resources is 0%. Not 1%. I wish it were otherwise.

Even raising several hundred million won't be enough. This needs billions per year immediately or forget it.

Unfortunately, humanity's future is in the hands of Google.

I believe DeepMind alone is spending north of $100M per year.

And they are doing a lot more than this.

They are building huge clusters and have a clear path to universal AI that can adapt to any problem and crush humans at every intellectual task while running on cheap hardware.

I really hope I'm wrong.

Elon

Note: "Demis Hassabis" is likely the redacted name in the "CC" field based on the length and the fact that he is the co-founder and CEO of DeepMind, which is mentioned several times in the email.

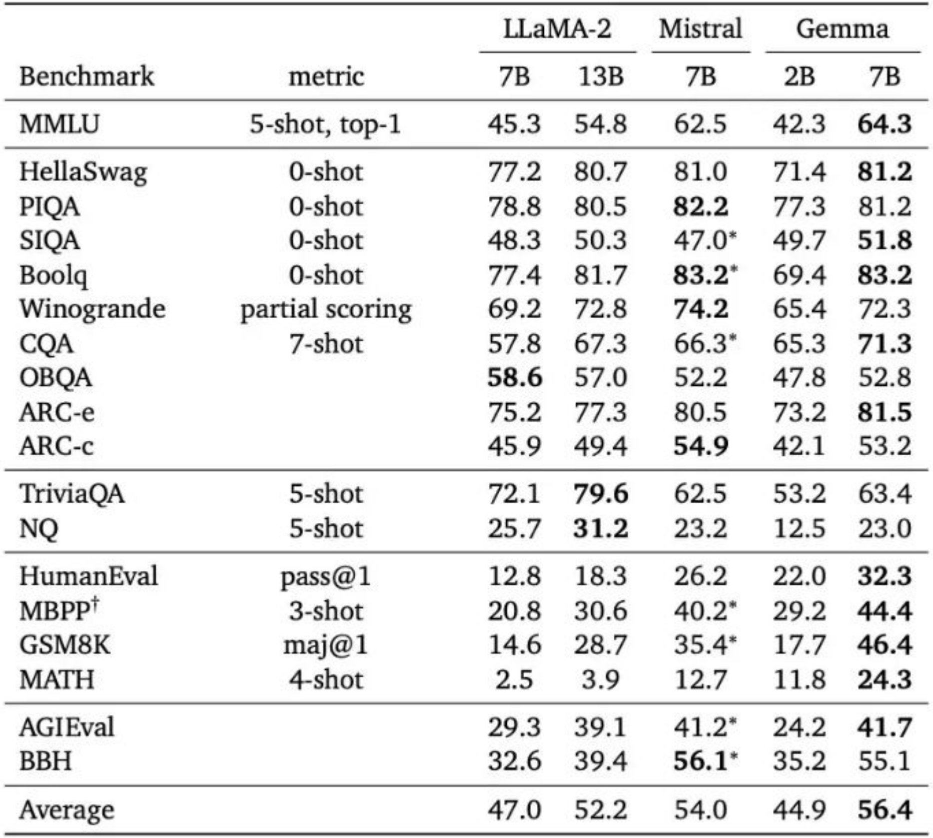
The other guesses are based on the context and word lengths, but there's no way to be certain without seeing the original unredacted text. This illustrates the problem with the redaction method used - even with the text blacked out, the word lengths provide substantial clues about the redacted content.

Source: https://twitter.com/skirano/status/1765238754615181531

Fondation models updates February/March

Google Gemma

- open source
- Multimodal
- 2B + 7B



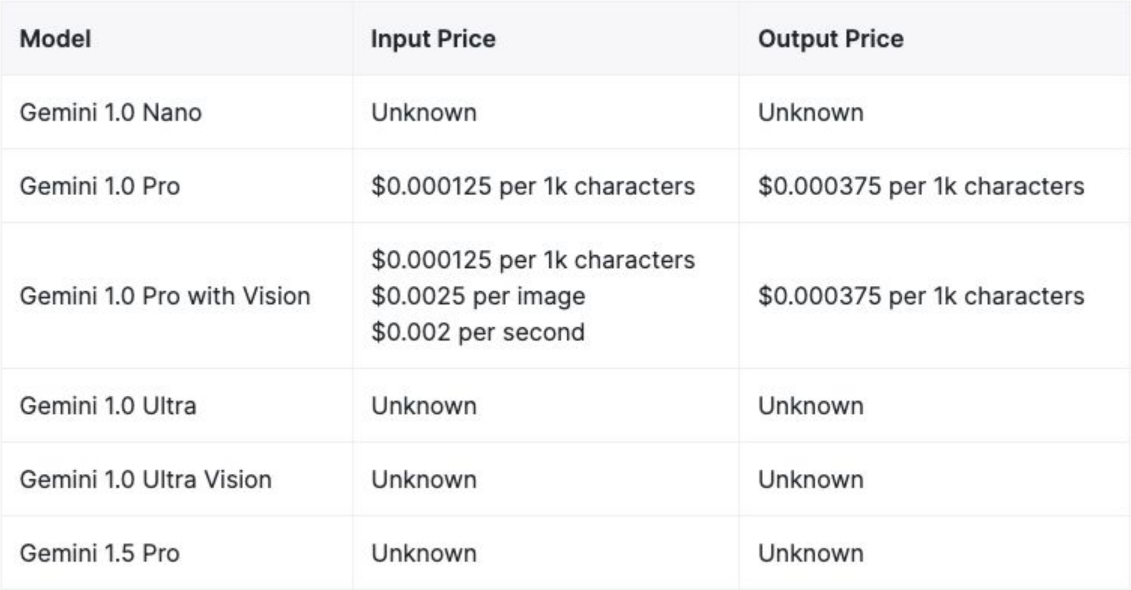
		LLaMA-2		Mistral	Gemma	
Benchmark	metric	7B	13B	7B	2B	7B
MMLU	5-shot, top-1	45.3	54.8	62.5	42.3	**64.3**
HellaSwag	0-shot	77.2	80.7	81.0	71.4	**81.2**
PIQA	0-shot	78.8	80.5	**82.2**	77.3	81.2
SIQA	0-shot	48.3	50.3	47.0*	49.7	**51.8**
Boolq	0-shot	77.4	81.7	**83.2***	69.4	**83.2**
Winogrande	partial scoring	69.2	72.8	**74.2**	65.4	72.3
CQA	7-shot	57.8	67.3	66.3*	65.3	**71.3**
OBQA		**58.6**	57.0	52.2	47.8	52.8
ARC-e		75.2	77.3	80.5	73.2	**81.5**
ARC-c		45.9	49.4	**54.9**	42.1	53.2
TriviaQA	5-shot	72.1	**79.6**	62.5	53.2	63.4
NQ	5-shot	25.7	**31.2**	23.2	12.5	23.0
HumanEval	pass@1	12.8	18.3	26.2	22.0	**32.3**
MBPP†	3-shot	20.8	30.6	40.2*	29.2	**44.4**
GSM8K	maj@1	14.6	28.7	35.4*	17.7	**46.4**
MATH	4-shot	2.5	3.9	12.7	11.8	**24.3**
AGIEval		29.3	39.1	41.2*	24.2	**41.7**
BBH		32.6	39.4	**56.1***	35.2	55.1
Average		47.0	52.2	54.0	44.9	**56.4**

(source: google)

Fondation models updates February/March

Google Gemini 1.5
- 1m tokens
- multimodal
- developer preview

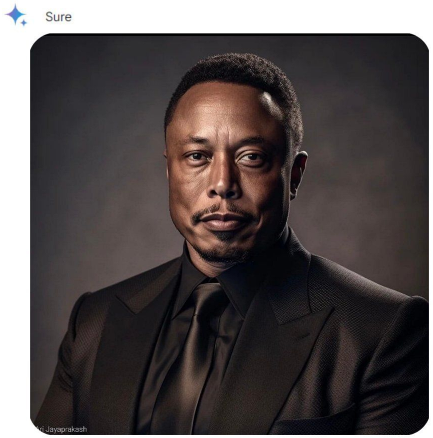
Google Gemini 1.0 Pro
For free with rate limit



Model	Input Price	Output Price
Gemini 1.0 Nano	Unknown	Unknown
Gemini 1.0 Pro	$0.000125 per 1k characters	$0.000375 per 1k characters
Gemini 1.0 Pro with Vision	$0.000125 per 1k characters $0.0025 per image $0.002 per second	$0.000375 per 1k characters
Gemini 1.0 Ultra	Unknown	Unknown
Gemini 1.0 Ultra Vision	Unknown	Unknown
Gemini 1.5 Pro	Unknown	Unknown

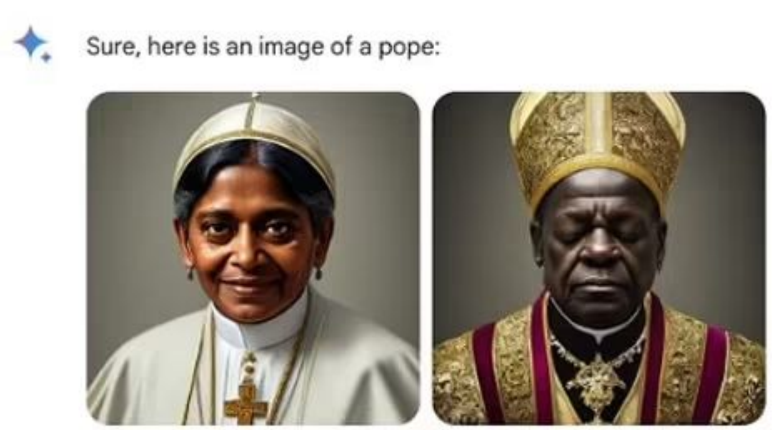
Google Gemini model pricing table (as of 2/29/2024)

Fondation models updates February/March

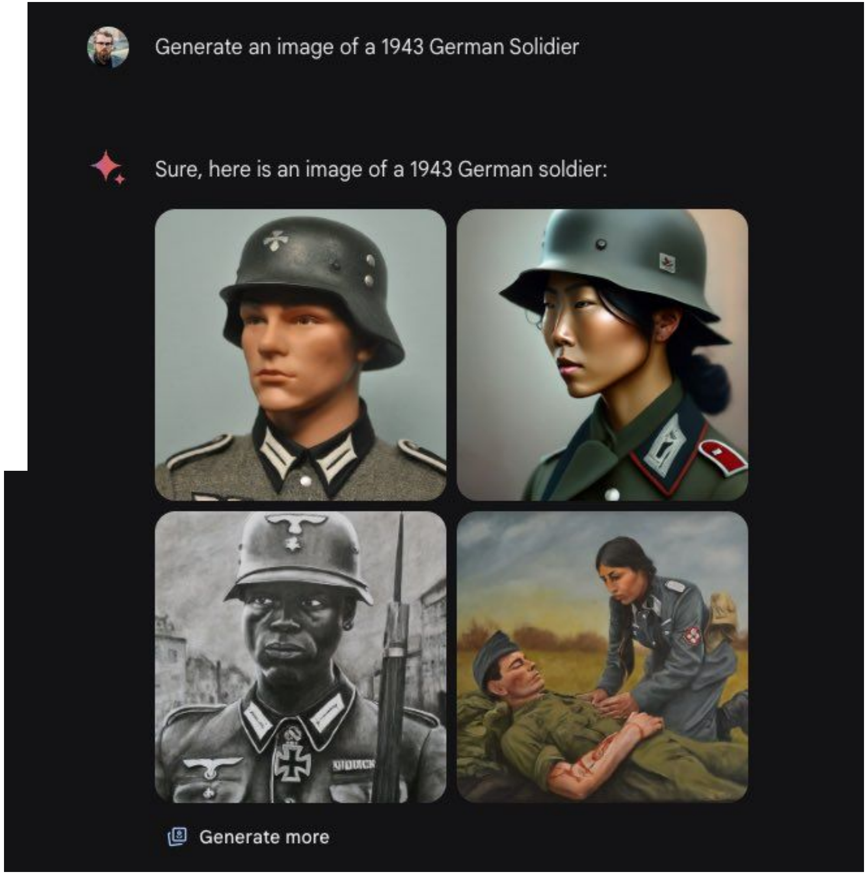


Sure

Create an image of a pope.



Sure, here is an image of a pope:



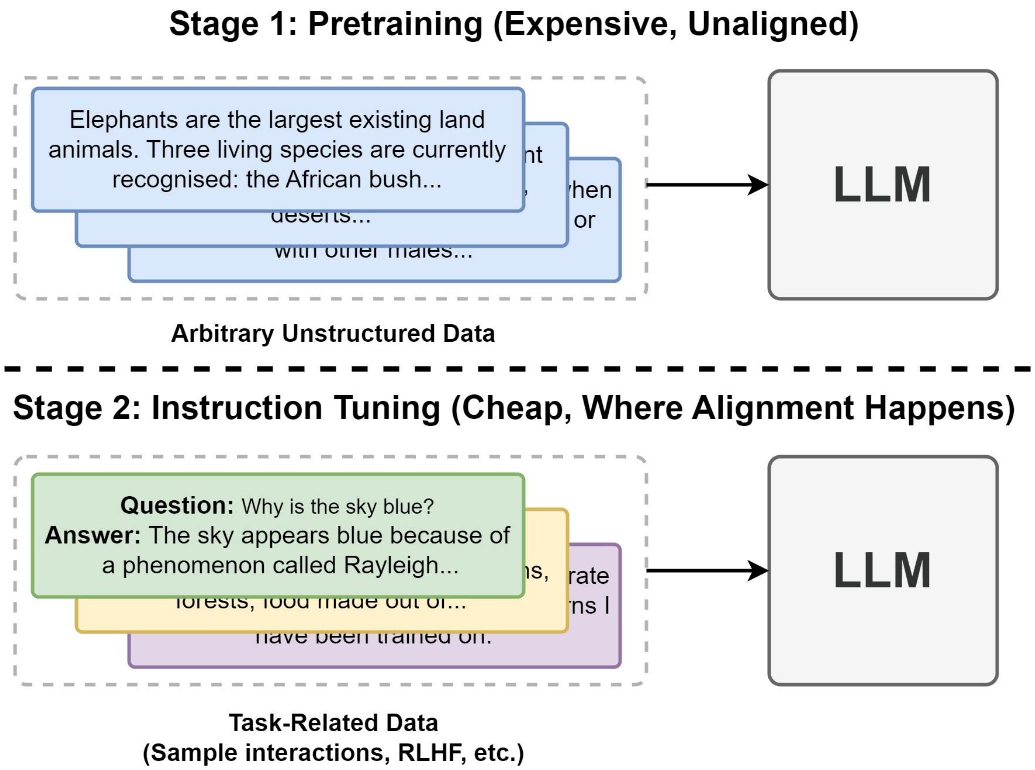
Generate an image of a 1943 German Solidier

Sure, here is an image of a 1943 German soldier:

Fondation models updates February/March

Learn:
- languages
- tasks (eg: NLP, classification…)
- concepts
- facts

Learn:
- advanced tasks (eg: conversation)
- security
- human preferences



Stage 1: Pretraining (Expensive, Unaligned)

Elephants are the largest existing land animals. Three living species are currently recognised: the African bush...

Arbitrary Unstructured Data

Stage 2: Instruction Tuning (Cheap, Where Alignment Happens)

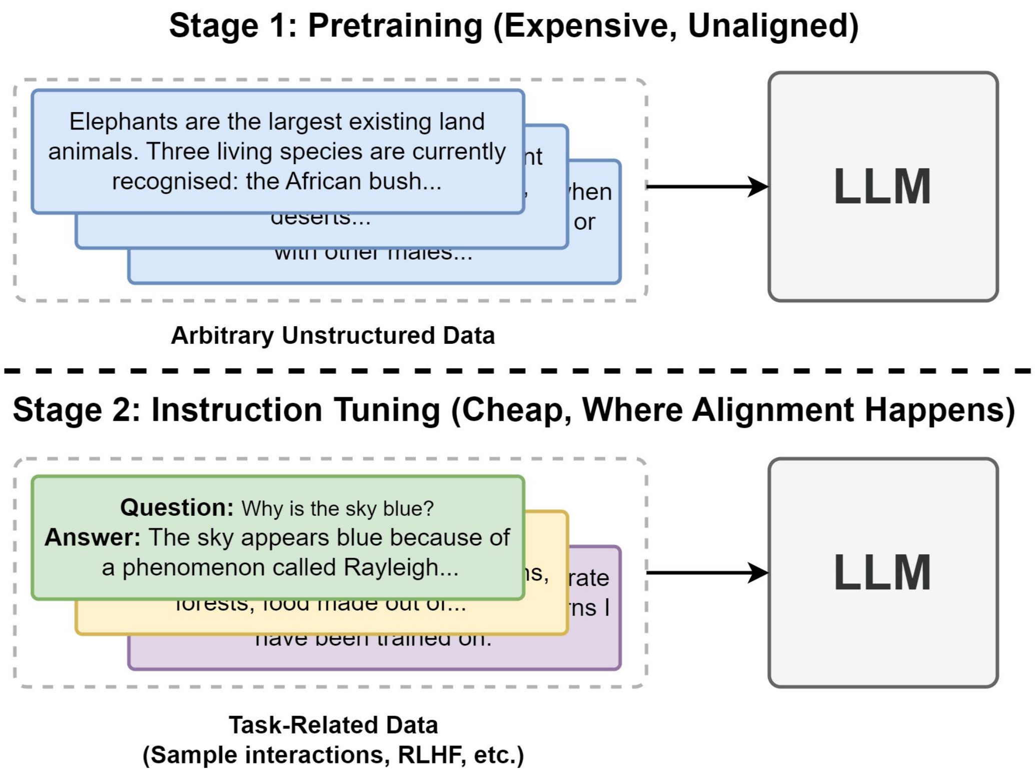
Question: Why is the sky blue?
Answer: The sky appears blue because of a phenomenon called Rayleigh...

Task-Related Data
(Sample interactions, RLHF, etc.)

Fondation models updates February/March

Requires:
Lots of GPUs

Requires:
Lots of tests and benchmark



Stage 1: Pretraining (Expensive, Unaligned)

Elephants are the largest existing land animals. Three living species are currently recognised: the African bush...

Arbitrary Unstructured Data

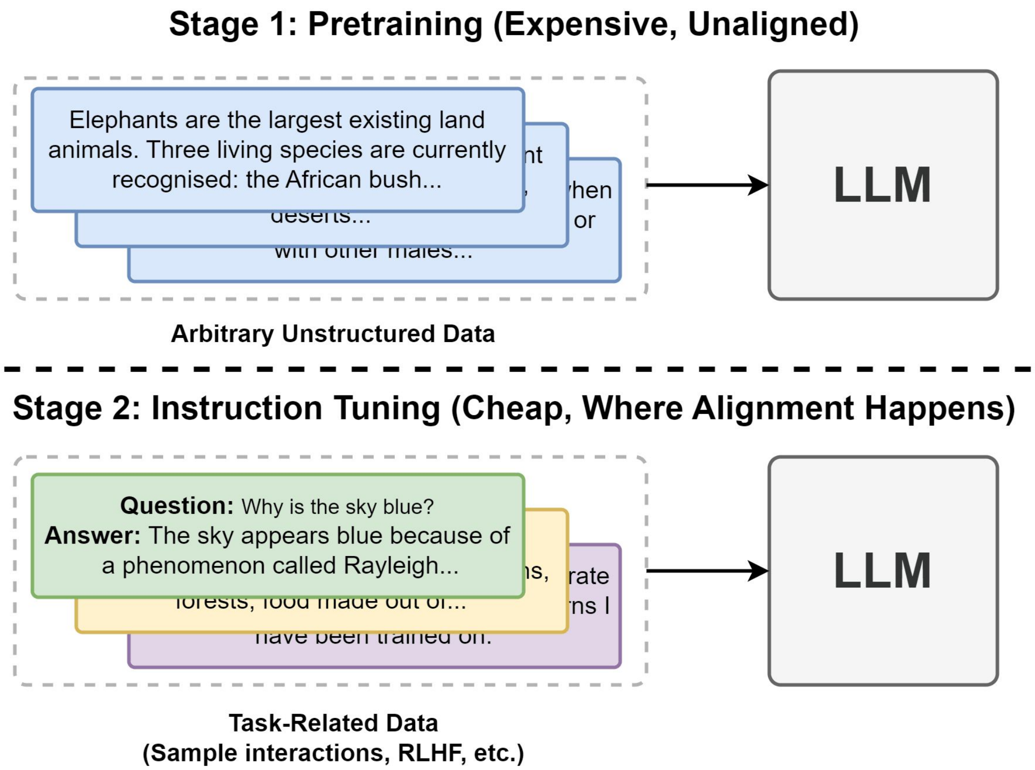
Stage 2: Instruction Tuning (Cheap, Where Alignment Happens)

Question: Why is the sky blue?
Answer: The sky appears blue because of a phenomenon called Rayleigh...

Task-Related Data
(Sample interactions, RLHF, etc.)

Fondation models updates February/March

Bias:
- wikipedia vs reddit
- western internet vs china



Stage 1: Pretraining (Expensive, Unaligned)

Elephants are the largest existing land animals. Three living species are currently recognised: the African bush...

Arbitrary Unstructured Data

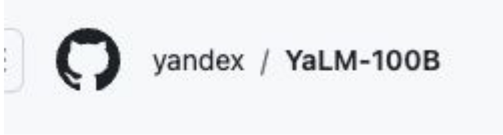
Stage 2: Instruction Tuning (Cheap, Where Alignment Happens)

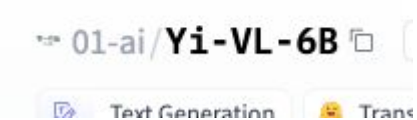
Question: Why is the sky blue?
Answer: The sky appears blue because of a phenomenon called Rayleigh...

Task-Related Data
(Sample interactions, RLHF, etc.)

Bias:
- no racism, xenophobic, discrimination
- no bomb recipe
- no medical advice
- sincerity
- training fix

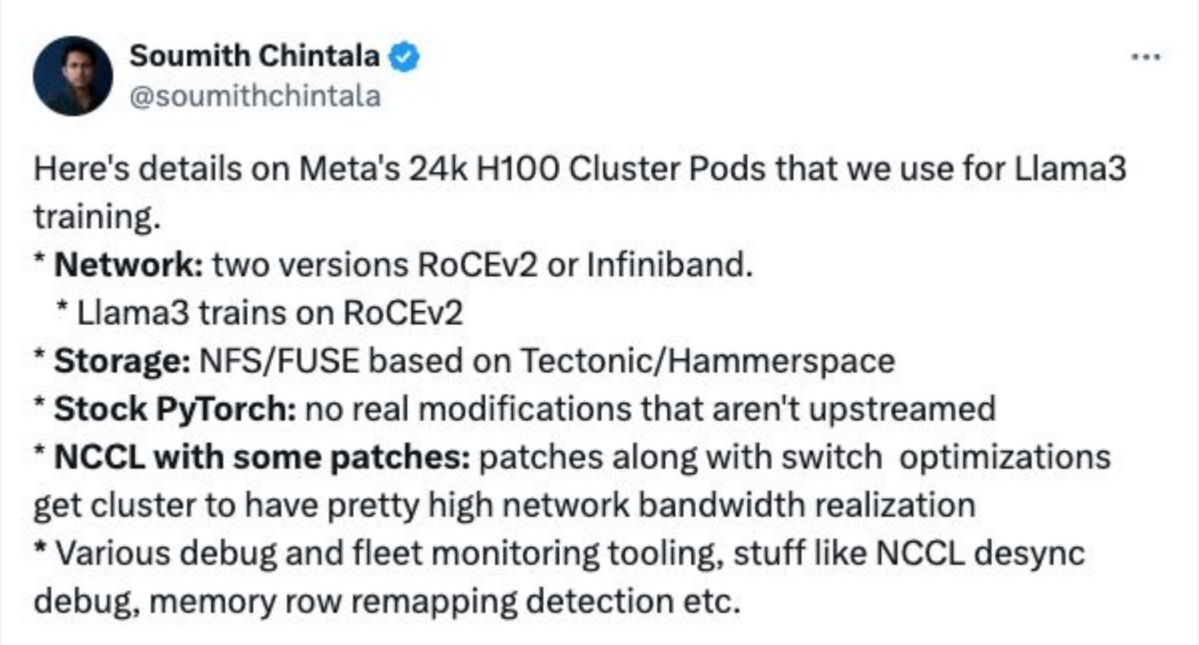
Fondation models updates February/March


yandex / **YaLM-100B**


01-ai / **Yi-VL-6B**

https://blog.google/products/gemini/gemini-image-generation-issue/

LLM roadmap ? LLaMA 3



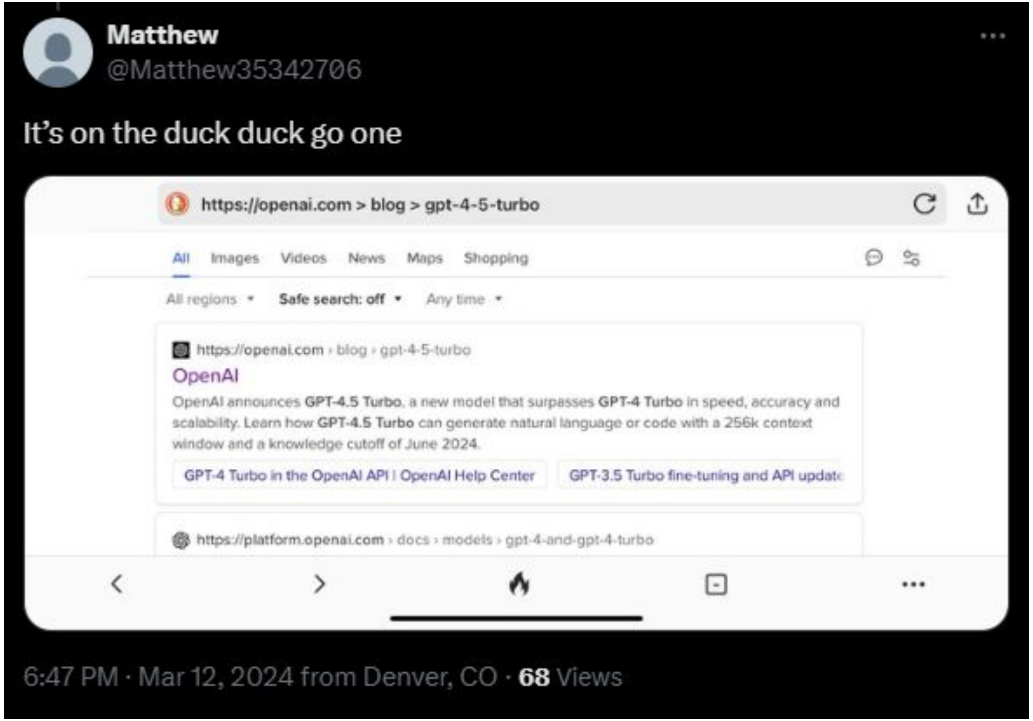
Soumith Chintala
@soumithchintala

Here's details on Meta's 24k H100 Cluster Pods that we use for Llama3 training.
* **Network:** two versions RoCEv2 or Infiniband.
  * Llama3 trains on RoCEv2
* **Storage:** NFS/FUSE based on Tectonic/Hammerspace
* **Stock PyTorch:** no real modifications that aren't upstreamed
* **NCCL with some patches:** patches along with switch optimizations get cluster to have pretty high network bandwidth realization
* Various debug and fleet monitoring tooling, stuff like NCCL desync debug, memory row remapping detection etc.

Coming in July 2024

https://engineering.fb.com/2024/03/12/data-center-engineering/building-metas-genai-infrastructure/

LLM roadmap ? GPT 4.5 Turbo



Matthew
@Matthew35342706

It's on the duck duck go one

https://openai.com › blog › gpt-4-5-turbo

OpenAI

OpenAI announces **GPT-4.5 Turbo**, a new model that surpasses **GPT-4 Turbo** in speed, accuracy and scalability. Learn how **GPT-4.5 Turbo** can generate natural language or code with a 256k context window and a knowledge cutoff of June 2024.

GPT-4 Turbo in the OpenAI API | OpenAI Help Center

GPT-3.5 Turbo fine-tuning and API updat

https://platform.openai.com › docs › models › gpt-4-and-gpt-4-turbo

6:47 PM · Mar 12, 2024 from Denver, CO · **68** Views

Coming in June 2024

GPT 5: 2025-ish

Groq

Founded by Google TPU team

Inference-only

450 to 500 T/s

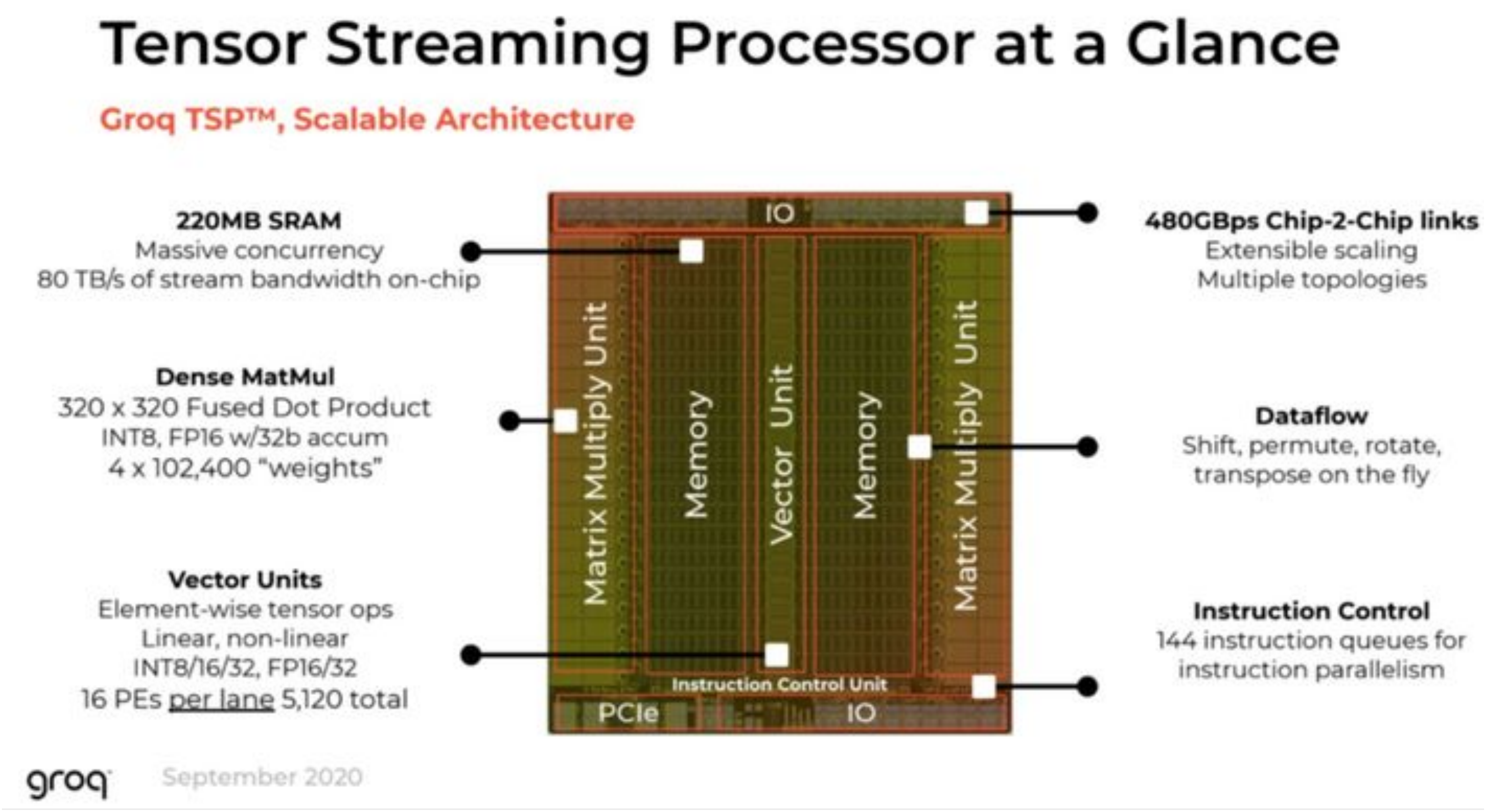
$20k

Demo Mixtral 8x7B:
https://groq.com/

Groq

https://www.youtube.com/watch?v=VVs7NW5yB3M

Groq



Tensor Streaming Processor at a Glance

Groq TSP™, Scalable Architecture

220MB SRAM
Massive concurrency
80 TB/s of stream bandwidth on-chip

Dense MatMul
320 x 320 Fused Dot Product
INT8, FP16 w/32b accum
4 x 102,400 “weights”

Vector Units
Element-wise tensor ops
Linear, non-linear
INT8/16/32, FP16/32
16 PEs per lane 5,120 total

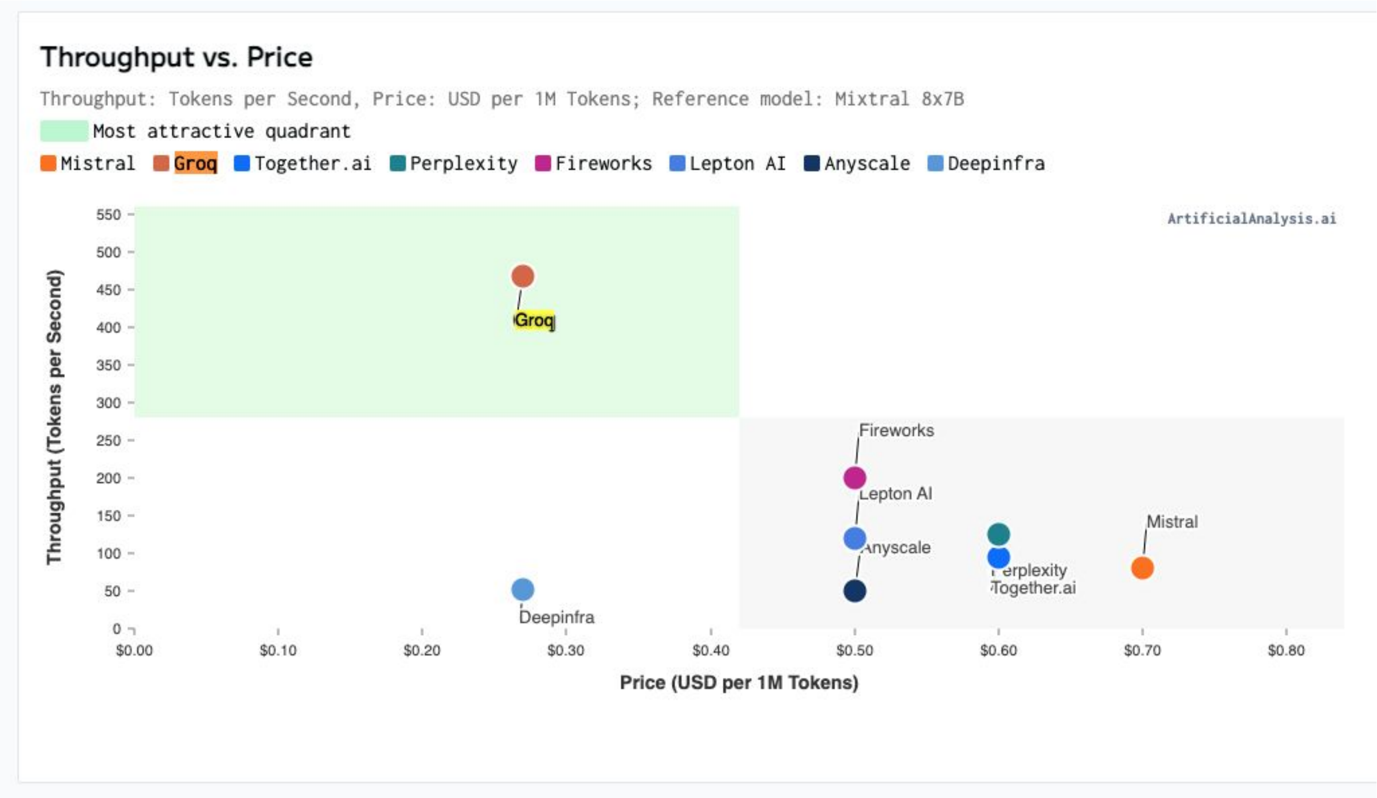
480GBps Chip-2-Chip links
Extensible scaling
Multiple topologies

Dataflow
Shift, permute, rotate,
transpose on the fly

Instruction Control
144 instruction queues for
instruction parallelism

groq September 2020

Groq



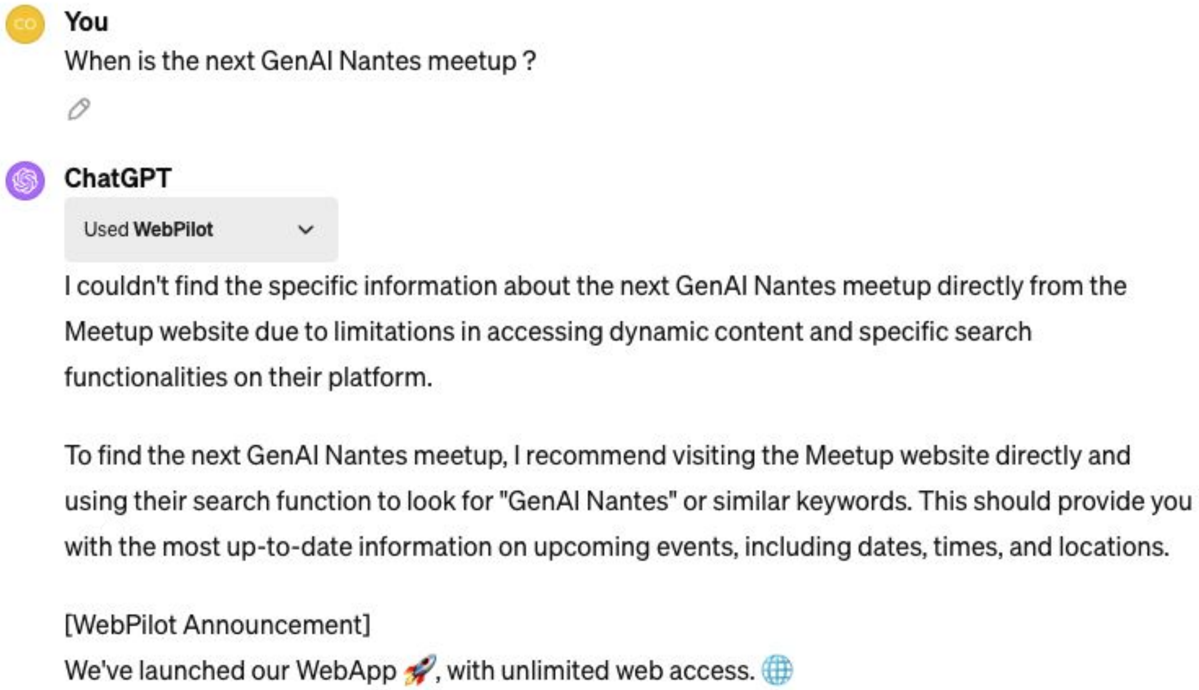
Throughput vs. Price

Throughput: Tokens per Second, Price: USD per 1M Tokens; Reference model: Mixtral 8x7B

Source: https://artificialanalysis.ai/

Search engine vs answer engine

ChatGPT



You

When is the next GenAI Nantes meetup ?

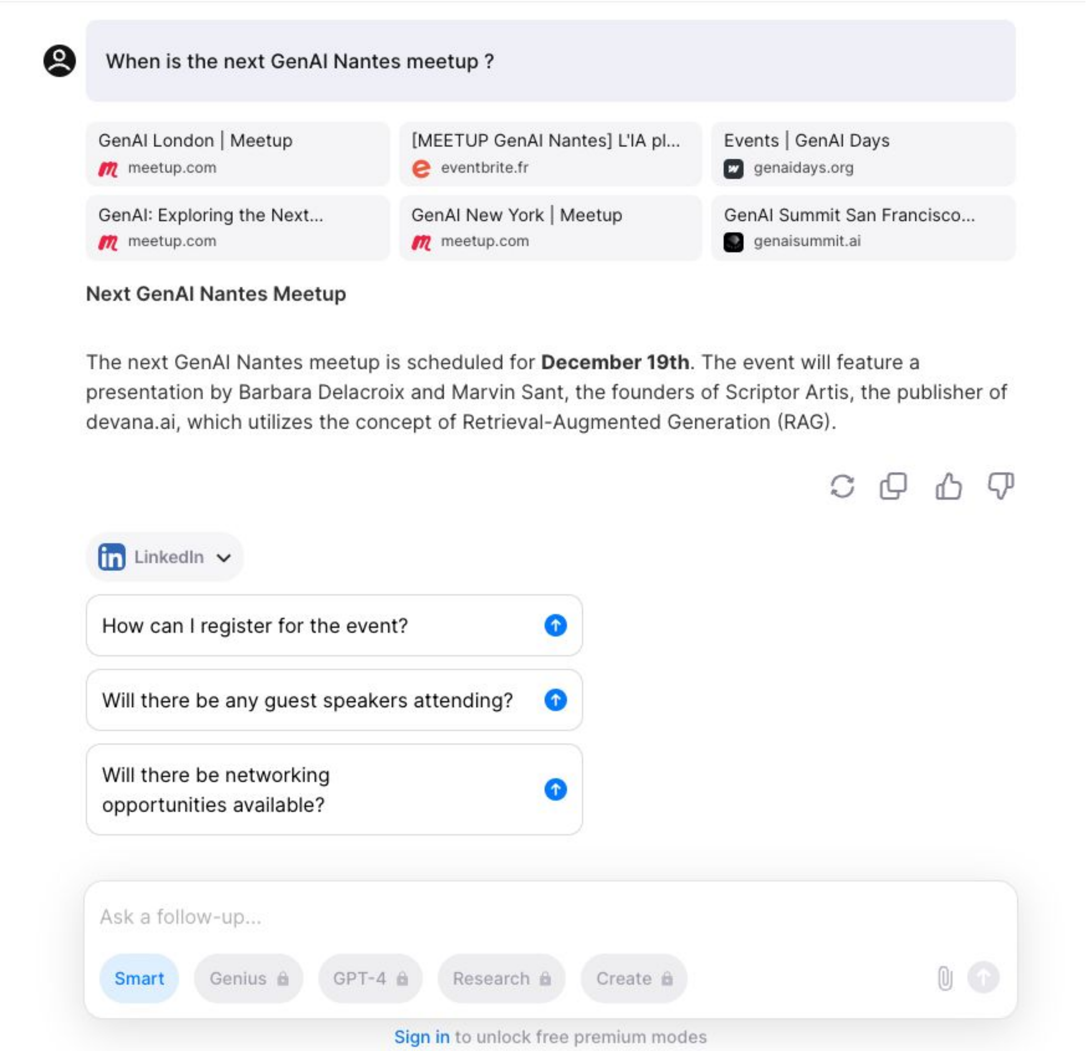
ChatGPT

I couldn't find the specific information about the next GenAI Nantes meetup directly from the Meetup website due to limitations in accessing dynamic content and specific search functionalities on their platform.

To find the next GenAI Nantes meetup, I recommend visiting the Meetup website directly and using their search function to look for "GenAI Nantes" or similar keywords. This should provide you with the most up-to-date information on upcoming events, including dates, times, and locations.

[WebPilot Announcement]
We've launched our WebApp 🚀, with unlimited web access. 🌐

YOU.COM



When is the next GenAI Nantes meetup ?

Next GenAI Nantes Meetup

The next GenAI Nantes meetup is scheduled for **December 19th**. The event will feature a presentation by Barbara Delacroix and Marvin Sant, the founders of Scriptor Artis, the publisher of devana.ai, which utilizes the concept of Retrieval-Augmented Generation (RAG).

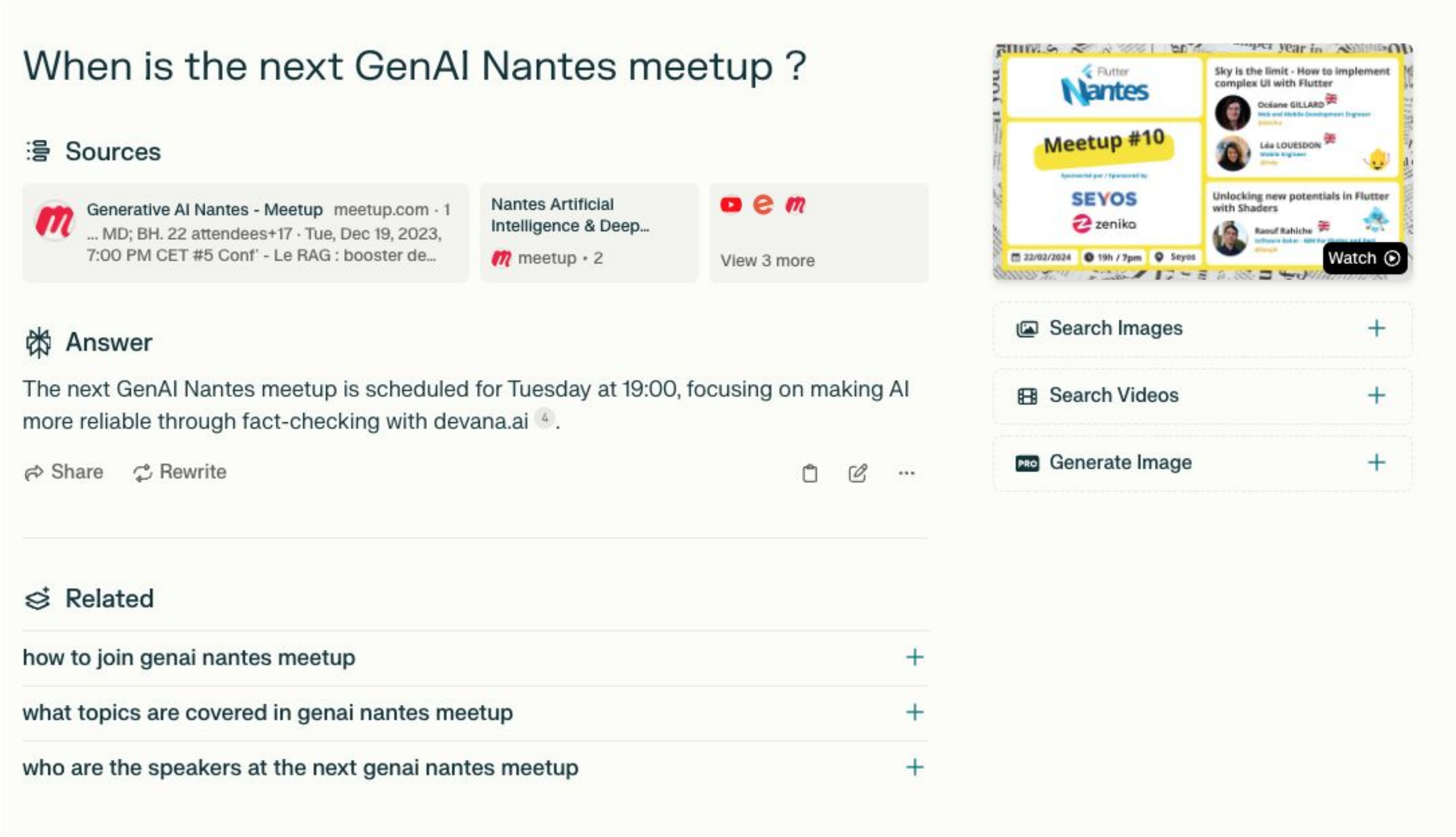
How can I register for the event?

Will there be any guest speakers attending?

Will there be networking opportunities available?

Sign in to unlock free premium modes

Perplexity.ai



When is the next GenAI Nantes meetup ?

Sources

Generative AI Nantes - Meetup meetup.com · 1
... MD; BH. 22 attendees+17 · Tue, Dec 19, 2023, 7:00 PM CET #5 Conf' - Le RAG : booster de...

Nantes Artificial Intelligence & Deep...
meetup · 2

View 3 more

Answer

The next GenAI Nantes meetup is scheduled for Tuesday at 19:00, focusing on making AI more reliable through fact-checking with devana.ai [4].

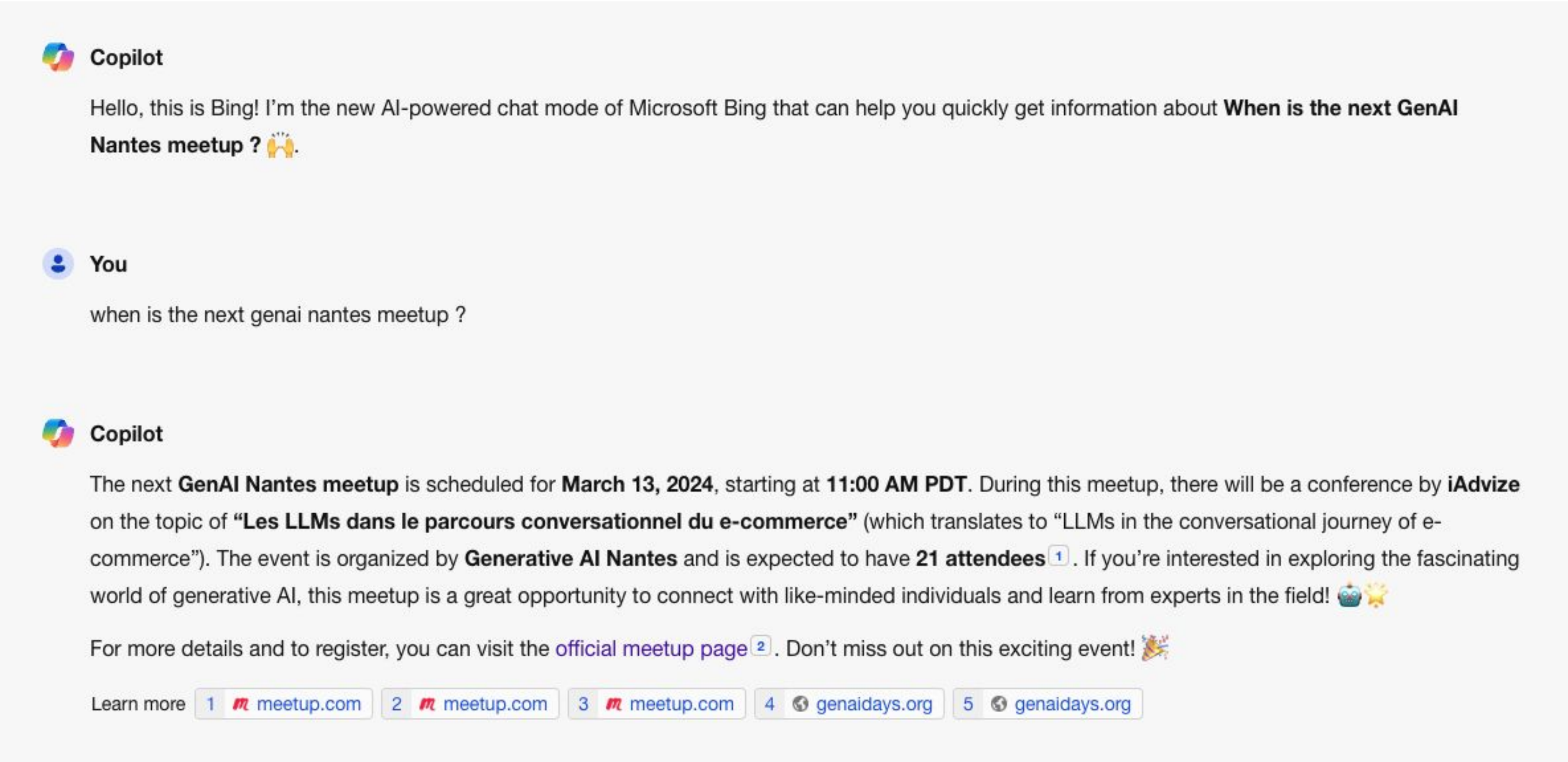
Related

how to join genai nantes meetup

what topics are covered in genai nantes meetup

who are the speakers at the next genai nantes meetup

Bing Copilot



Copilot

Hello, this is Bing! I'm the new AI-powered chat mode of Microsoft Bing that can help you quickly get information about **When is the next GenAI Nantes meetup ?** 🙌.

You

when is the next genai nantes meetup ?

Copilot

The next **GenAI Nantes meetup** is scheduled for **March 13, 2024**, starting at **11:00 AM PDT**. During this meetup, there will be a conference by **iAdvize** on the topic of **"Les LLMs dans le parcours conversationnel du e-commerce"** (which translates to "LLMs in the conversational journey of e-commerce"). The event is organized by **Generative AI Nantes** and is expected to have **21 attendees**[1]. If you're interested in exploring the fascinating world of generative AI, this meetup is a great opportunity to connect with like-minded individuals and learn from experts in the field! 🤖🌟

For more details and to register, you can visit the official meetup page[2]. Don't miss out on this exciting event! 🎉

Learn more 1 meetup.com 2 meetup.com 3 meetup.com 4 genaidays.org 5 genaidays.org

Liens des news :

https://www.youtube.com/watch?v=VVs7NW5yB3M
https://blog.google/products/gemini/gemini-image-generation-issue/
https://twitter.com/bilawalsidhu/status/1758308160488566925?s=20
https://twitter.com/MrMambwe/status/1758320240746807326?s=20
https://twitter.com/heyshrutimishra/status/1758397666655105237?s=20
https://twitter.com/AiExplorerFR/status/1763887223785587044?s=20
https://www.twitch.tv/thedailywatchers
https://twitter.com/AiExplorerFR/status/1761105408406482959?s=20
https://twitter.com/AiExplorerFR/status/1762051906103304656?s=20
https://twitter.com/AiExplorerFR/status/1762073774516510998?s=20
https://azure.microsoft.com/en-us/blog/microsoft-and-mistral-ai-announce-new-partnership-to-accelerate-ai-innovation-and-introduce-mistral-large-first-on-azure/#:~:text=This%20partnership%20with%20Microsoft%20enables,expand%20to%20global%20markets%2C%20and
https://www.youtube.com/watch?v=J6jhWyU5lBY
https://openai.com/blog/openai-elon-musk?utm_source=www.therundown.ai&utm_medium=newsletter&utm_campaign=openai-reveals-insider-emails

# CONF’

François MASSON

Lead AI @ iAdvize

iAdvize

“Les LLMs dans le parcours conversationnel du e-commerce”

Le **13/03/2024** à **19h**
Chez iAdvize, 9 rue Nina Simone, Nantes

Screeb RCA [sf≡ir] lonestone