CONF’

“Les dessous de Llama 3”

Xavier Martinet

AI Research Engineer @ Meta

Meta

Le **18/09/2024** à **19h**

Au Palace, 4 rue Voltaire, Nantes

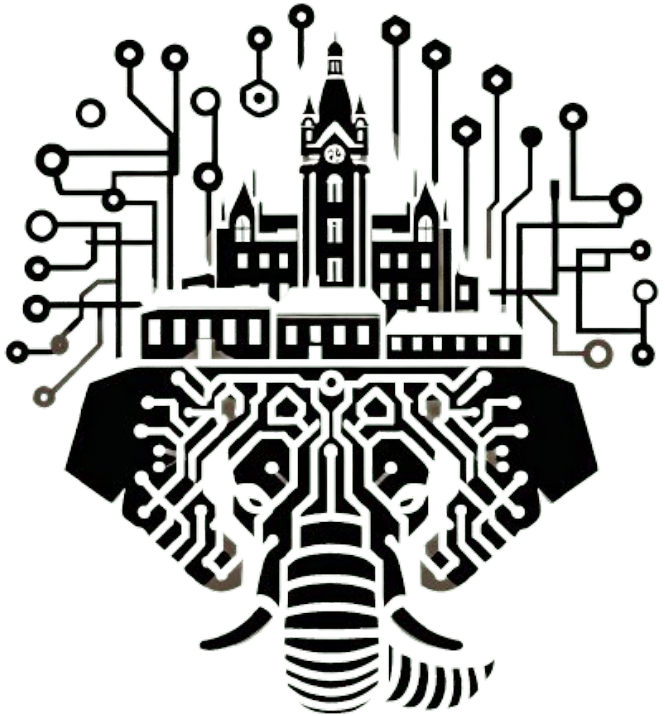
_icilundi

Screeb RCA [sfeir] lonestone

Qui n’est jamais venu au meetup Gen AI Nantes ?

GenAI Nantes

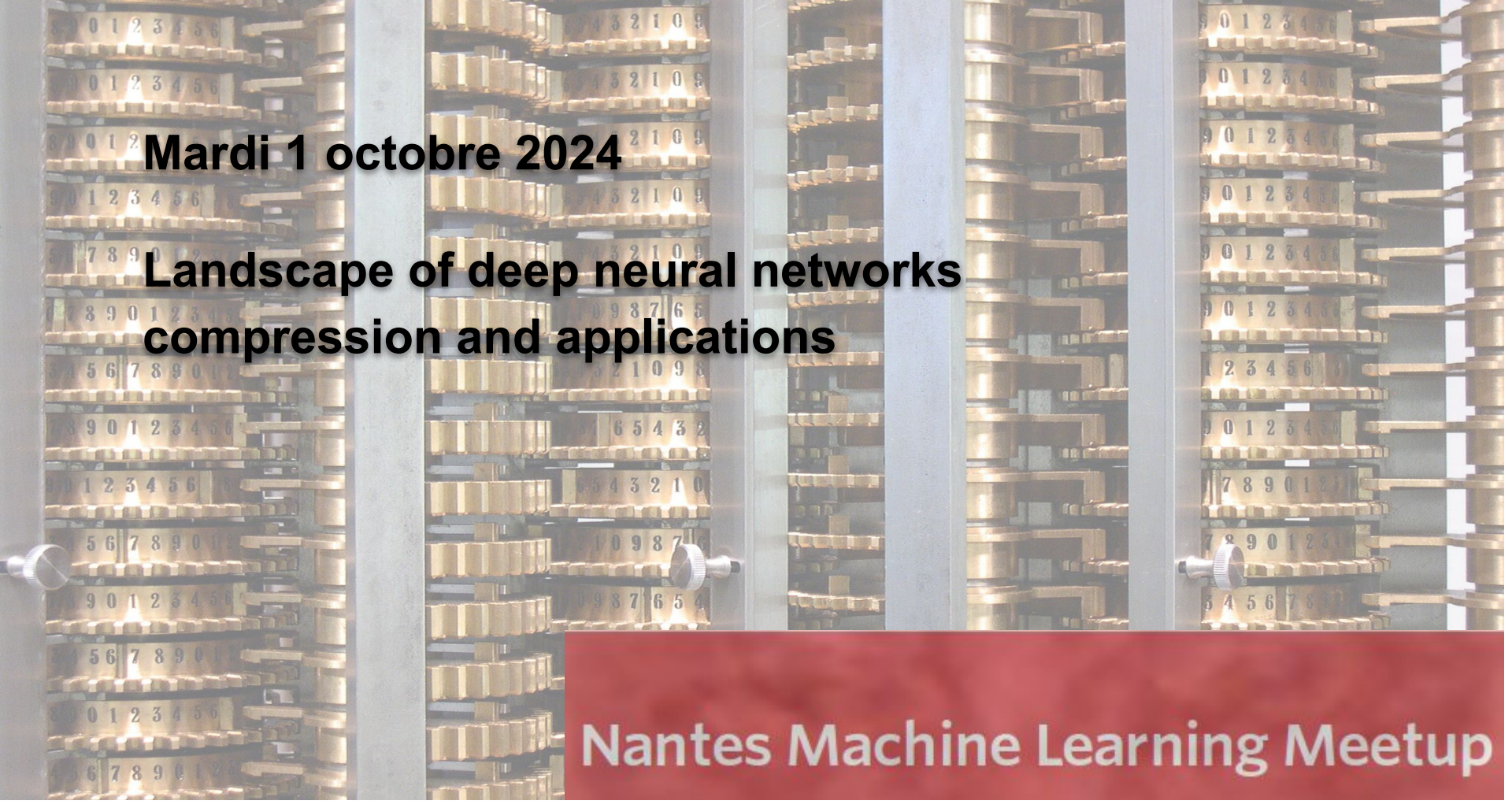
🎤 15 événements / an
🏴‍☠️ 1 hackathon
👷 2 workshops
❤️ 1 communauté de 400p



🏴‍☠️ Save the date!

Shift revient en 2025,

du 28 au 30 mars.


Mardi 1 octobre 2024
Landscape of deep neural networks compression and applications
Nantes Machine Learning Meetup

Qui fait des projets GenAI ?

 Qui chevauche des Llama* ?

*aucun Llama ne sera blessé pendant la soirée

🦙 Qui chevauche des Llama 405B mammouth ?

*aucun Llama ne sera blessé pendant la soirée

Schedule

👀 1- News summer 2024

🦙 2- A multimodal Llama

🍻 3- Enjoy

HUMAN & MACHINE INTELLIGENCE | CONVERSATIONAL AI

The Llama 3 Herd of Models

July 23, 2024

Abstract

Modern artificial intelligence (AI) systems are powered by foundation models. This paper presents a new set of foundation models, called Llama 3. It is a herd of language models that natively support multilinguality, coding, reasoning, and tool usage. Our largest model is a dense Transformer with 405B parameters and a context window of up to 128K tokens. This paper presents an extensive empirical evaluation of Llama 3. We find that Llama 3 delivers comparable quality to leading language models such as GPT-4 on a plethora of tasks. We publicly release Llama 3, including pre-trained and post-trained versions of the 405B parameter language model and our Llama Guard 3 model for input and output safety. The paper also presents the results of experiments in which we integrate image, video, and speech capabilities into Llama 3 via a compositional approach. We observe this approach performs competitively with the state-of-the-art on image, video, and speech recognition tasks. The resulting models are not yet being broadly released as they are still under development.

→ Download the Paper

Paper Llama3.1

- Datasets
- Tools
- Function calling
- Multimodal
- Security
- Long context
- Coding
- Reasoning
- …

https://ai.meta.com/research/publications/the-llama-3-herd-of-models/

Focus on long context

Very large context ?

Use case:

- Q/A on large document
- RAG

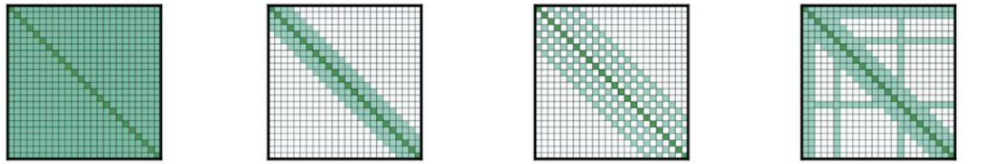
LLM:

- Llama3.1: 128k
- Gemini 1.5 pro: **1m**
- GPT-4o: 128k
- Claude Sonet 3.5: **200k**
- Mistral Large 2: 128k

Screeb

Very large context ?

Cost is quadratic
- Training
- Inference



(a) Full n^2 attention (b) Sliding window attention (c) Dilated sliding window (d) Global+sliding window

Figure 2: Comparing the full self-attention pattern and the configuration of attention patterns in our Longformer.

Large context training for Llama3

* Trained with 8k token windows (14.000 B tokens)

* Gradually increased to 128k for the last (800 B tokens)

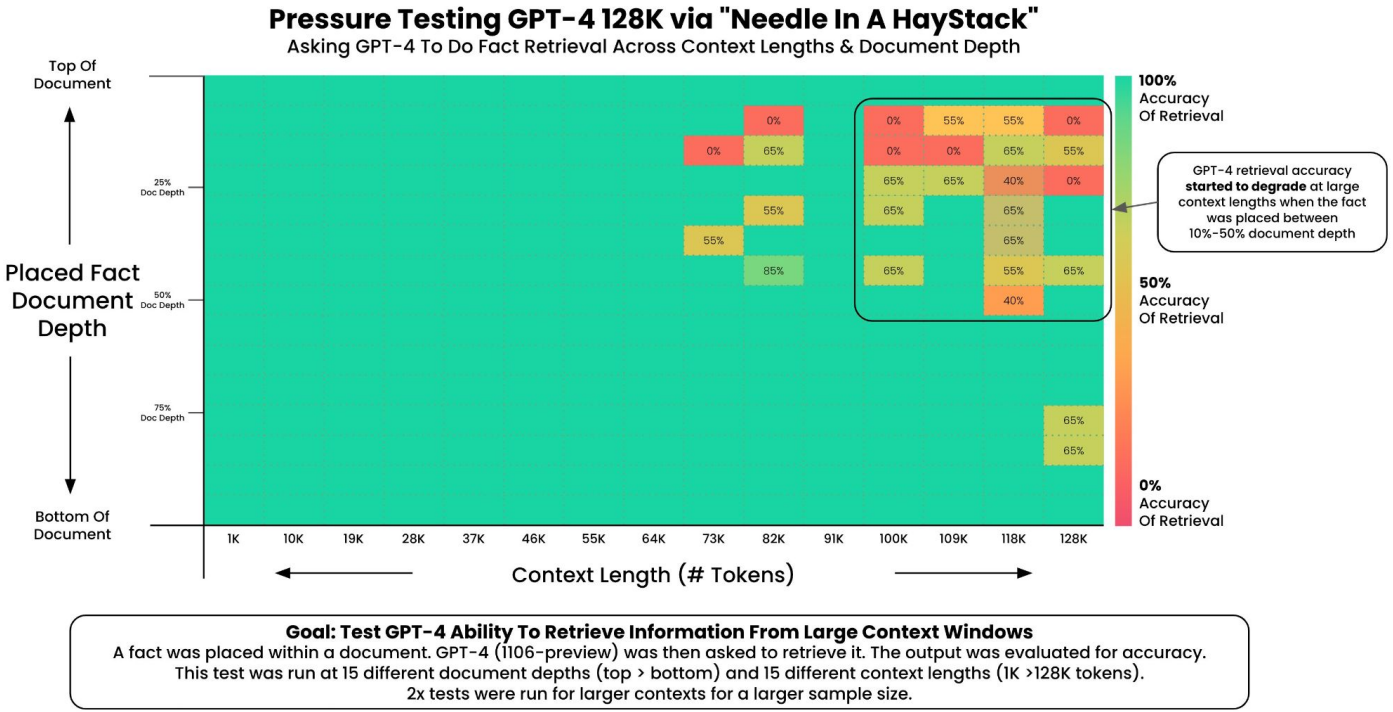
The “needle in a haystack” benchmark (NIAH)

Let’s benchmark a simple **retrieval** task

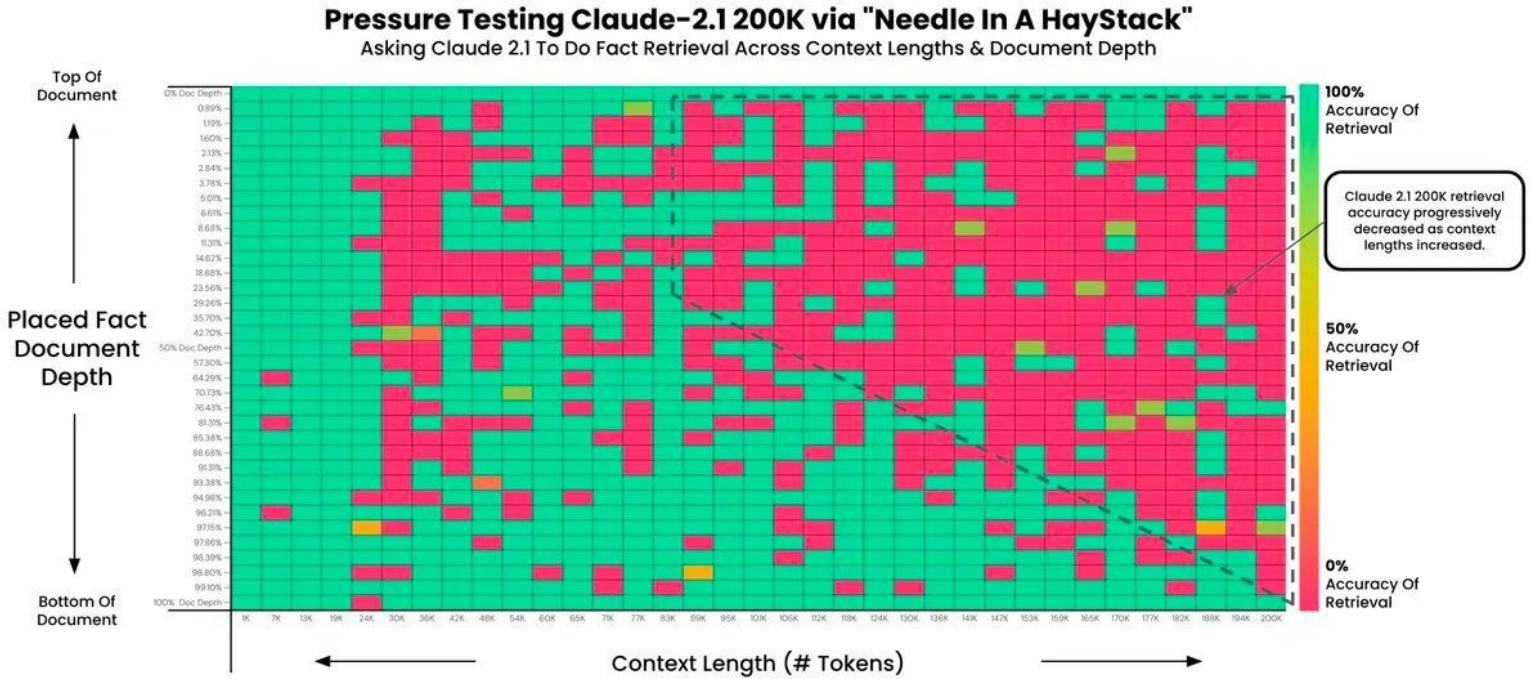
Screeb

https://x.com/GregKamradt/status/1722386725635580292

Single retrieval - GPT4-128k

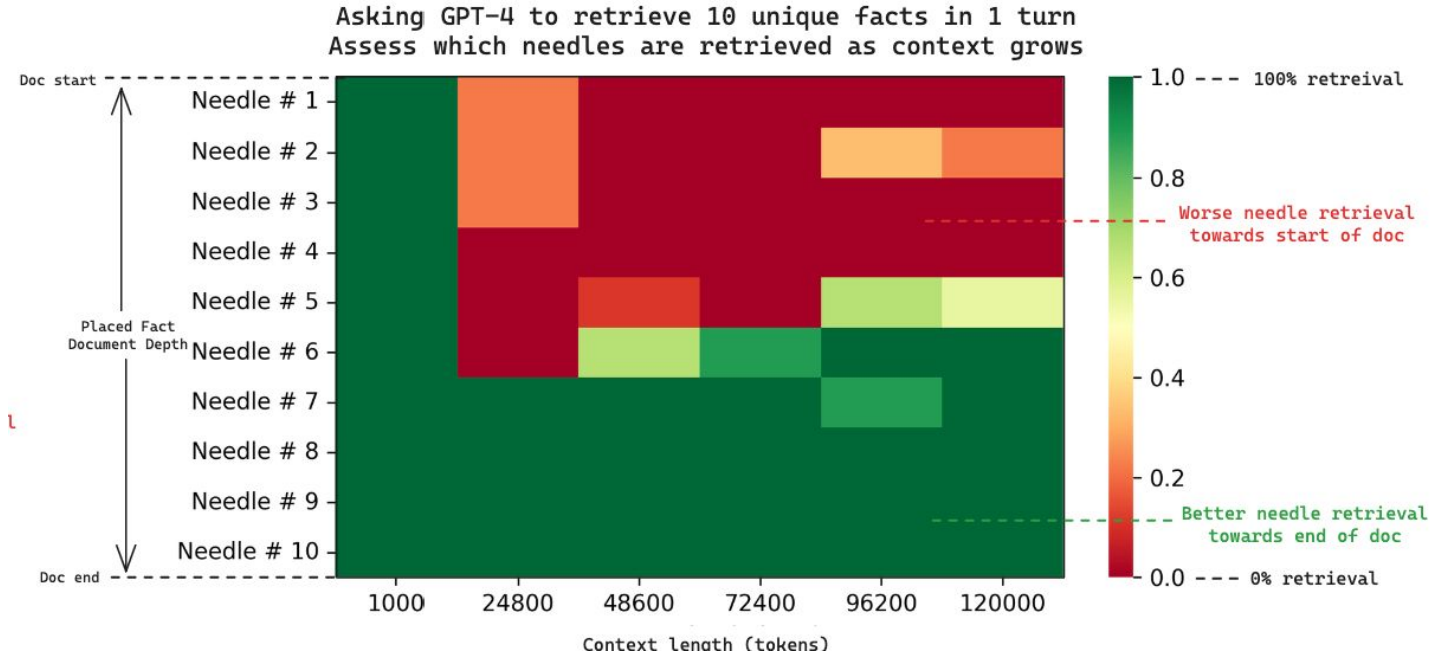


Single retrieval - Claude 2.1 200k



Goal: Test Claude 2.1 Ability To Retrieve Information From Large Context Windows
A fact was placed within a document. Claude 2.1 (200K) was then asked to retrieve it. The output was evaluated (with GPT-4) for accuracy.
This test was run at 35 different document depths (top > bottom) and 35 different context lengths (1K >200K tokens).
Document Depths followed a sigmoid distribution

Multi-needle GPT4-128k



https://blog.langchain.dev/multi-needle-in-a-haystack/

The RULER benchmark

- Multiple needle retrieval
- Multi-hop needle
- Q/A
- More advanced tasks

RULER: What's the Real Context Size of Your Long-Context Language Models?

Cheng-Ping Hsieh*, Simeng Sun*, Samuel Kriman, Shantanu Acharya Dima Rekesh, Fei Jia, Yang Zhang, Boris Ginsburg

NVIDIA

{chsieh,simengs}@nvidia.com

Abstract

The needle-in-a-haystack (NIAH) test, which examines the ability to retrieve a piece of information (the "needle") from long distractor texts (the "haystack"), has been widely adopted to evaluate long-context language models (LMs). However, this simple retrieval-based test is indicative of only a superficial form of long-context understanding. To provide a more comprehensive evaluation of long-context LMs, we create a new synthetic benchmark RULER with flexible configurations for customized sequence length and task complexity. RULER expands upon the vanilla NIAH test to encompass variations with diverse types and quantities of needles. Moreover, RULER introduces new task categories *multi-hop tracing* and *aggregation* to test behaviors beyond searching from context. We evaluate 17 long-context LMs with 13 representative tasks in RULER. Despite achieving nearly perfect accuracy in the vanilla NIAH test, almost all models exhibit large performance drops as the context length increases. While these models all claim context sizes of 32K tokens or greater, only half of them can maintain satisfactory performance at the length of 32K. Our analysis of Yi-34B, which supports context length of 200K, reveals large room for improvement as we increase input length and task complexity. We open source RULER to spur comprehensive evaluation of long-context LMs.

https://arxiv.org/pdf/2404.06654

The RULER benchmark

Models	Claimed Length	Effective Length	4K	8K	16K	32K	64K	128K	Avg.	wAvg. (inc)	wAvg. (dec)
Llama2 (7B)	4K	-	85.6								
Gemini-1.5-Pro	1M	>128K	96.7	95.8	96.0	95.9	95.9	94.4	95.8	95.5 (1st)	96.1 (1st)
GPT-4	128K	64K	96.6	96.3	95.2	93.2	87.0	81.2	91.6	89.0 (2nd)	94.1 (2nd)
Llama3.1 (70B)	128K	64K	96.5	95.8	95.4	94.8	88.4	66.6	89.6	85.5 (4th)	93.7 (3rd)
Qwen2 (72B)	128K	32K	96.9	96.1	94.9	94.1	79.8	53.7	85.9	79.6 (9th)	92.3 (4th)
Command-R-plus (104B)	128K	32K	95.6	95.2	94.2	92.0	84.3	63.1	87.4	82.7 (7th)	92.1 (5th)
GLM4 (9B)	1M	64K	94.7	92.8	92.1	89.9	86.7	83.1	89.9	88.0 (3rd)	91.7 (6th)
Llama3.1 (8B)	128K	32K	95.5	93.8	91.6	87.4	84.7	77.0	88.3	85.4 (5th)	91.3 (7th)
GradientAI/Llama3 (70B)	1M	16K	95.1	94.4	90.8	85.4	80.9	72.1	86.5	82.6 (8th)	90.3 (8th)
Mixtral-8x22B (39B/141B)	64K	32K	95.6	94.9	93.4	90.9	84.7	31.7	81.9	73.5 (11th)	90.3 (9th)
Yi (34B)	200K	32K	93.3	92.2	91.3	87.5	83.2	77.3	87.5	84.8 (6th)	90.1 (10th)
Phi3-medium (14B)	128K	32K	93.3	93.2	91.1	86.8	78.6	46.1	81.5	74.8 (10th)	88.3 (11th)
Mistral-v0.2 (7B)	32K	16K	93.6	91.2	87.2	75.4	49.0	13.8	68.4	55.6 (13th)	81.2 (12th)
LWM (7B)	1M	<4K	82.3	78.4	73.7	69.1	68.1	65.0	72.8	69.9 (12th)	75.7 (13th)
DBRX (36B/132B)	32K	8K	95.1	93.8	83.6	63.1	2.4	0.0	56.3	38.0 (14th)	74.7 (14th)
Together (7B)	32K	4K	88.2	81.1	69.4	63.0	0.0	0.0	50.3	33.8 (15th)	66.7 (15th)
LongChat (7B)	32K	<4K	84.7	79.9	70.8	59.3	0.0	0.0	49.1	33.1 (16th)	65.2 (16th)
LongAlpaca (13B)	32K	<4K	60.6	57.0	56.6	43.6	0.0	0.0	36.3	24.7 (17th)	47.9 (17th)

Screeb

Most models fail beyond 32k tokens, except Gemini

So how to support long context?

- Map/Reduce
- Mamba2 architecture
- Ask Xavier to work harder 😅

Mamba2 architecture (June ‘24)

- Linear complexity
- Good for long (infinite?) context
- Transformers are stronger for short context
- Optimized memory usage

- Codestral Mamba: https://mistral.ai/news/codestral-mamba/
- 100m tokens context window: https://magic.dev/blog/100m-token-context-windows

Paper: https://arxiv.org/abs/2406.02528

Context caching

- Offload the attention state and K/V cache
- Process attention once and read many times
- Set a TTL on each query

Context caching

INPUT PRICING	$7.00 / 1 million tokens
OUTPUT PRICING	$21.00 / 1 million tokens
CONTEXT CACHING	$1.75 / 1 million tokens
CONTEXT CACHING (STORAGE)	$4.50 / 1 million tokens per hour

Claude 3.5 Sonnet

- Our most intelligent model to date
- 200K context window

Input

- $3 / MTok

Prompt caching

- $3.75 / MTok - Cache write
- $0.30 / MTok - Cache read

Output

- $15 / MTok

Multiple/parallel function calling

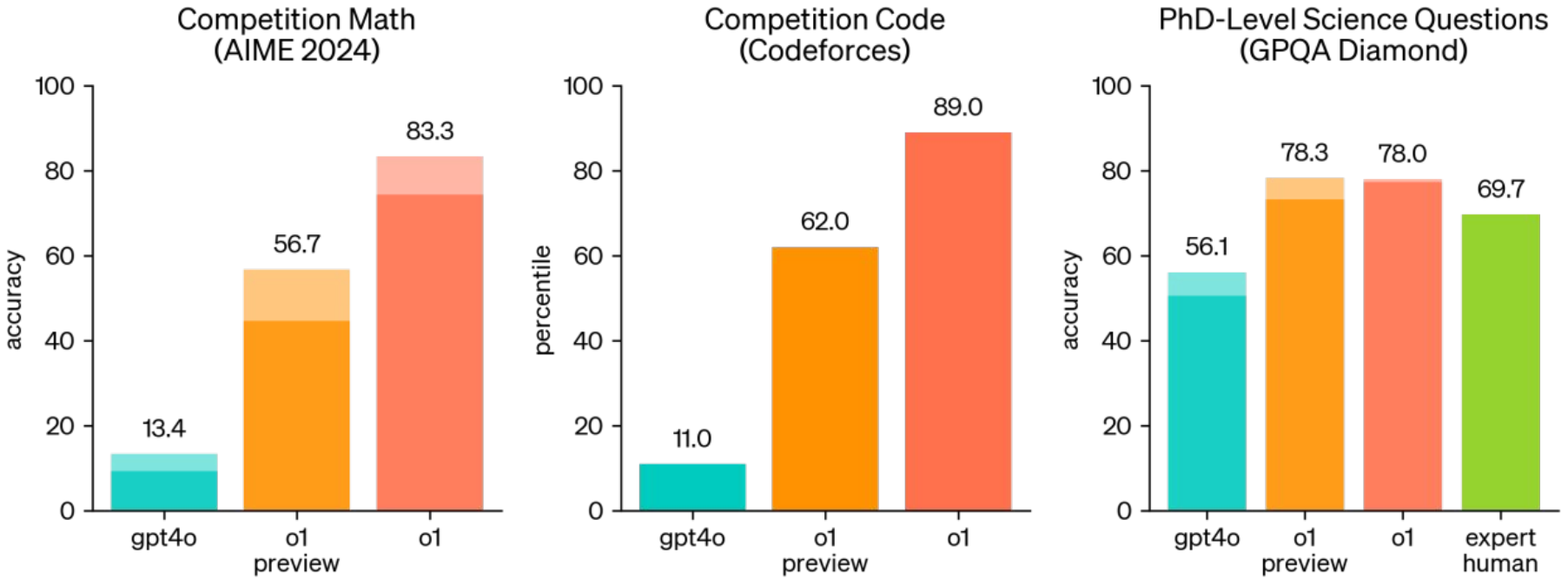
tool_choice

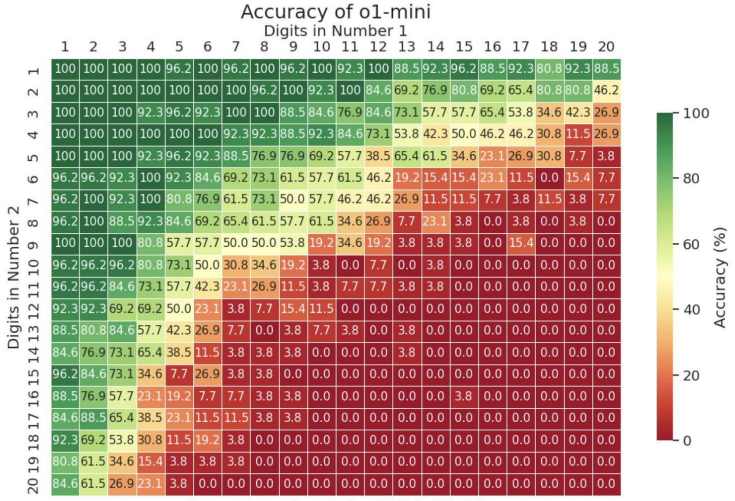
Users can use `tool_choice` to speficy how tools are used:

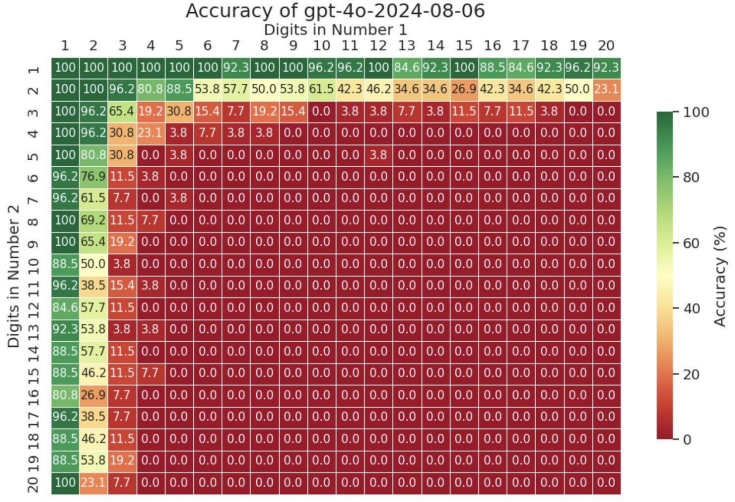
- "auto": default mode. Model decides if it uses the tool or not.
- "any": forces tool use.
- "none": prevents tool use.

```
import os
from mistralai import Mistral

api_key = os.environ["MISTRAL_API_KEY"]
model = "mistral-large-latest"

client = Mistral(api_key=api_key)
response = client.chat.complete(
    model = model,
    messages = messages,
    tools = tools,
    tool_choice = "any",
)
response
```

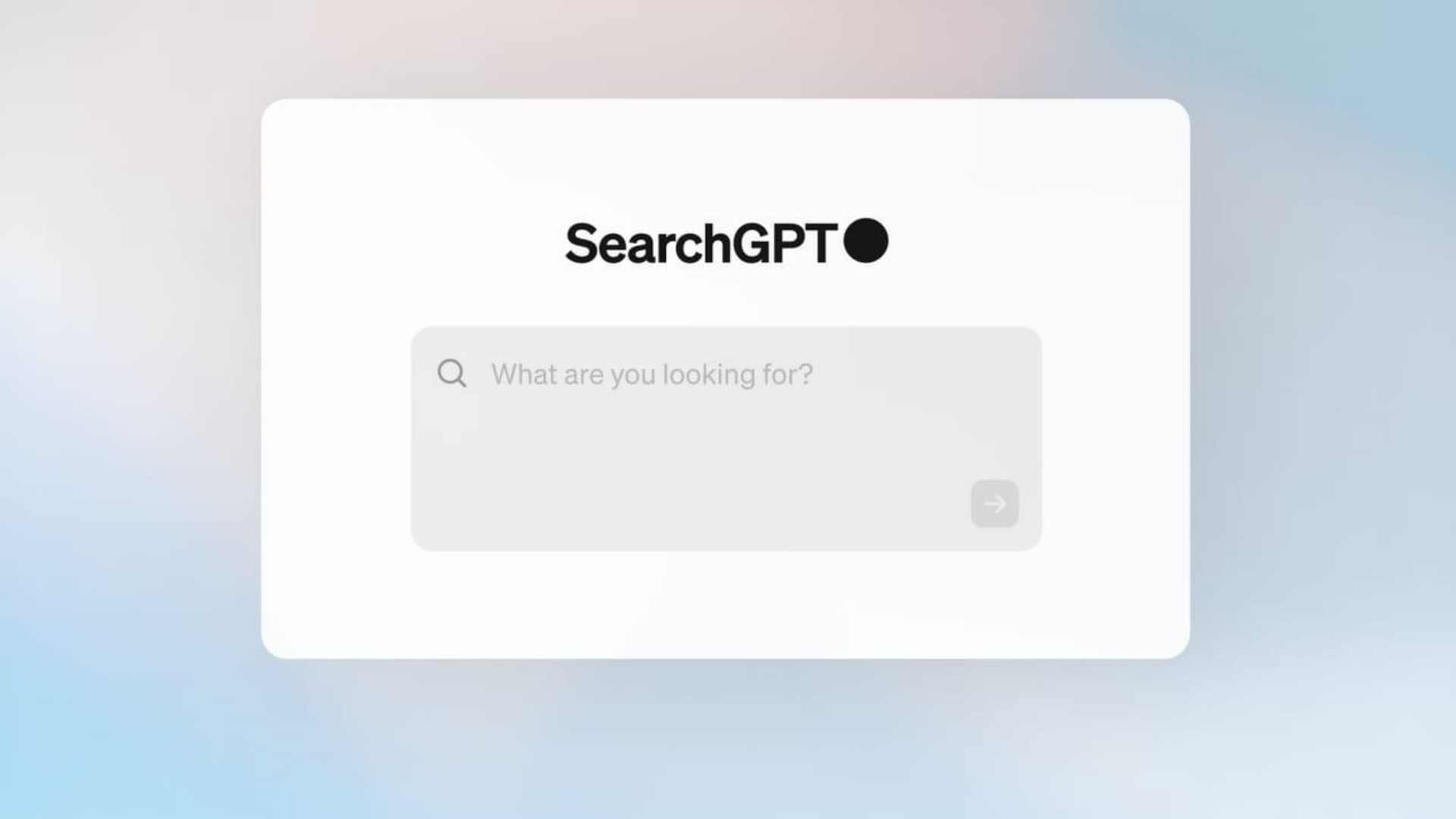

OpenAI o1
Reasoning


Competition Math
(AIME 2024)
accuracy
100
80
60
40
20
0
13.4
56.7
83.3
gpt4o
o1
preview
o1
Competition Code
(Codeforces)
percentile
11.0
62.0
89.0
PhD-Level Science Questions
(GPQA Diamond)
56.1
78.3
78.0
69.7
expert
human


Accuracy of o1-mini
Digits in Number 1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Digits in Number 2
1 100 100 100 100 96.2 100 96.2 100 96.2 100 92.3 100 88.5 92.3 96.2 88.5 92.3 80.8 92.3 88.5
2 100 100 100 100 100 100 100 96.2 100 92.3 100 84.6 69.2 76.9 80.8 69.2 65.4 80.8 80.8 46.2
3 100 100 100 92.3 96.2 92.3 100 100 88.5 84.6 76.9 84.6 73.1 57.7 57.7 65.4 53.8 34.6 42.3 26.9
4 100 100 100 100 100 100 92.3 92.3 88.5 92.3 84.6 73.1 53.8 42.3 50.0 46.2 46.2 30.8 11.5 26.9
5 100 100 100 92.3 96.2 92.3 88.5 76.9 76.9 69.2 57.7 38.5 65.4 61.5 34.6 23.1 26.9 30.8 7.7 3.8
6 96.2 96.2 92.3 100 92.3 84.6 69.2 73.1 61.5 57.7 61.5 46.2 19.2 15.4 15.4 23.1 11.5 0.0 15.4 7.7
7 96.2 100 92.3 100 80.8 76.9 61.5 73.1 50.0 57.7 46.2 46.2 26.9 11.5 11.5 7.7 3.8 11.5 3.8 7.7
8 96.2 100 88.5 92.3 84.6 69.2 65.4 61.5 57.7 61.5 34.6 26.9 7.7 23.1 3.8 0.0 3.8 0.0 3.8 0.0
9 100 100 100 80.8 57.7 57.7 50.0 50.0 53.8 19.2 34.6 19.2 3.8 3.8 3.8 0.0 15.4 0.0 0.0 0.0
10 96.2 96.2 96.2 80.8 73.1 50.0 30.8 34.6 19.2 3.8 0.0 7.7 0.0 3.8 0.0 0.0 0.0 0.0 0.0 0.0
11 96.2 96.2 84.6 73.1 57.7 42.3 23.1 26.9 11.5 3.8 7.7 7.7 3.8 3.8 0.0 0.0 0.0 0.0 0.0 0.0
12 92.3 92.3 69.2 69.2 50.0 23.1 3.8 7.7 15.4 11.5 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
13 88.5 80.8 84.6 57.7 42.3 26.9 7.7 0.0 3.8 7.7 3.8 0.0 3.8 0.0 0.0 0.0 0.0 0.0 0.0 0.0
14 84.6 76.9 73.1 65.4 38.5 11.5 3.8 3.8 3.8 0.0 0.0 0.0 3.8 0.0 0.0 0.0 0.0 0.0 0.0 0.0
15 96.2 84.6 73.1 34.6 7.7 26.9 3.8 3.8 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
16 88.5 76.9 57.7 23.1 19.2 7.7 7.7 3.8 3.8 0.0 0.0 0.0 0.0 0.0 3.8 0.0 0.0 0.0 0.0 0.0
17 84.6 88.5 65.4 38.5 23.1 11.5 11.5 3.8 3.8 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
18 92.3 69.2 53.8 30.8 11.5 19.2 3.8 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
19 80.8 61.5 34.6 15.4 3.8 3.8 3.8 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
20 84.6 61.5 26.9 23.1 3.8 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
Accuracy (%)
100
80
60
40
20
0


Accuracy of gpt-4o-2024-08-06
Digits in Number 1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Digits in Number 2
1 100 100 100 100 100 100 92.3 100 100 96.2 96.2 100 84.6 92.3 100 88.5 84.6 92.3 96.2 92.3
2 100 100 96.2 80.8 88.5 53.8 57.7 50.0 53.8 61.5 42.3 46.2 34.6 34.6 26.9 42.3 34.6 42.3 50.0 23.1
3 100 96.2 65.4 19.2 30.8 15.4 7.7 19.2 15.4 0.0 3.8 3.8 7.7 3.8 11.5 7.7 11.5 3.8 0.0 0.0
4 100 96.2 30.8 23.1 3.8 7.7 3.8 3.8 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
5 100 80.8 30.8 0.0 3.8 0.0 0.0 0.0 0.0 0.0 0.0 3.8 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
6 96.2 76.9 11.5 3.8 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
7 96.2 61.5 7.7 0.0 3.8 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
8 100 69.2 11.5 7.7 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
9 100 65.4 19.2 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
10 88.5 50.0 3.8 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
11 96.2 38.5 15.4 3.8 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
12 84.6 57.7 11.5 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
13 92.3 53.8 3.8 3.8 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
14 88.5 57.7 11.5 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
15 88.5 46.2 11.5 7.7 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
16 80.8 26.9 7.7 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
17 96.2 38.5 7.7 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
18 88.5 46.2 11.5 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
19 88.5 53.8 19.2 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
20 100 23.1 7.7 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
Accuracy (%)
100
80
60
40
20
0

https://x.com/dallaslones/status/1834313445677605256


SearchGPT
What are you looking for?

https://x.com/OpenAI/status/1816536290822881780


iPhone 16 Pro
Hello, Apple Intelligence.
PRO


A more personal Siri
Your mom's flight lands at 11:18 AM.
Private Cloud Compute
Clean Up in Photos
Summaries in Messages
Writing Tools
Reduce Interruptions in Focus
Apple Intelligence
Image Playground
Genmoji
Image Wand
Audio recording
summaries
Priority messages in Mail
Inbox
Priority notifications
Natural language search

https://x.com/joffrey/status/1829081299576914091

https://x.com/GuangyuRobert/status/1831006762184646829

sakana.ai

首页
探索
资产
个人主页

AI 工具

图片生成
智能画布
视频生成
故事创作 Beta

消息中心

反馈
邀请

AI 作图

轻松实现创意图片

图片生成
智能画布

AI 视频

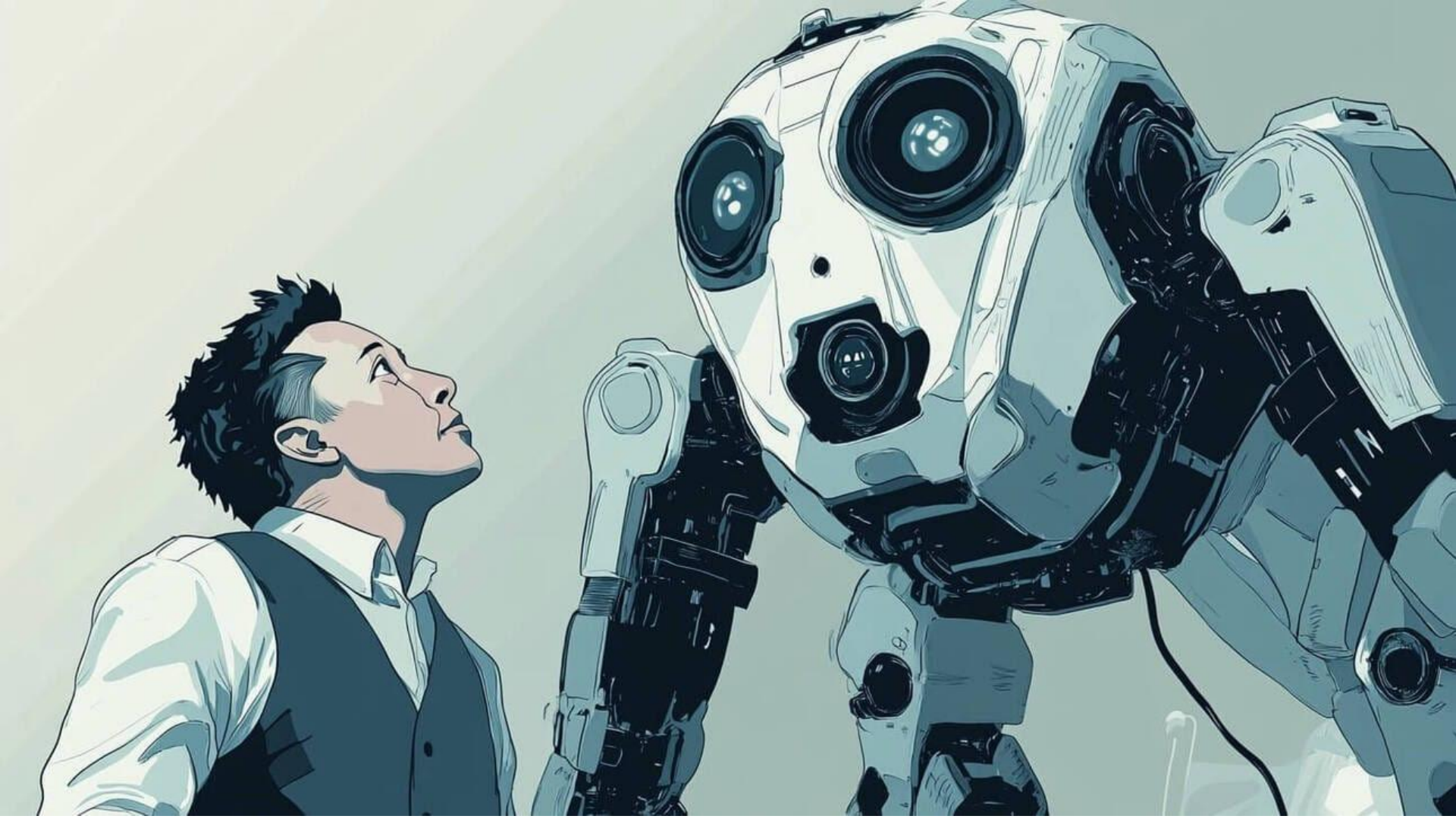
让你的创意动起来

视频生成

图片 视频

热门 摄影 插画 设计

https://www.youtube.com/shorts/pqTX4Gl4uGc?app=desktop



Overall ELO Scores on Chatbot Arena
ELO Rating
1330
1320
1310
1300
1290
1280
1270
1260
1250
1240
1230
1220
1210
1200
ChatGPT-4o-latest-2024-08-08
Gemini-1.5-Pro-Exp-0801
GPT-4o-2024-05-13
sus-column-r (Early Grok-2)
GPT-4o-mini-2024-07-18
Claude 3.5 Sonnet
Gemini-Advanced-0514
Meta-Llama-3.1-405b-Instruct
Gemini-1.5-Pro-API-0514
Gemini-1.5-Pro-API-0409-Preview
GPT-4-Turbo-2024-04-09
GPT-4-1106-preview
Mistral Large 2
Claude 3 Opus
Athene-70b
Meta-Llama-3.1-70b-Instruct
GPT-4-0125-preview
Yi-Large-preview
Gemini-1.5-Flash-API-0514
reka-core-20240722
Deepseek-v2-API-0628
Gemma-2-27B-it
deepseek-coder-v2-0724
Yi-Large
Bard (Gemini Pro)

Meta

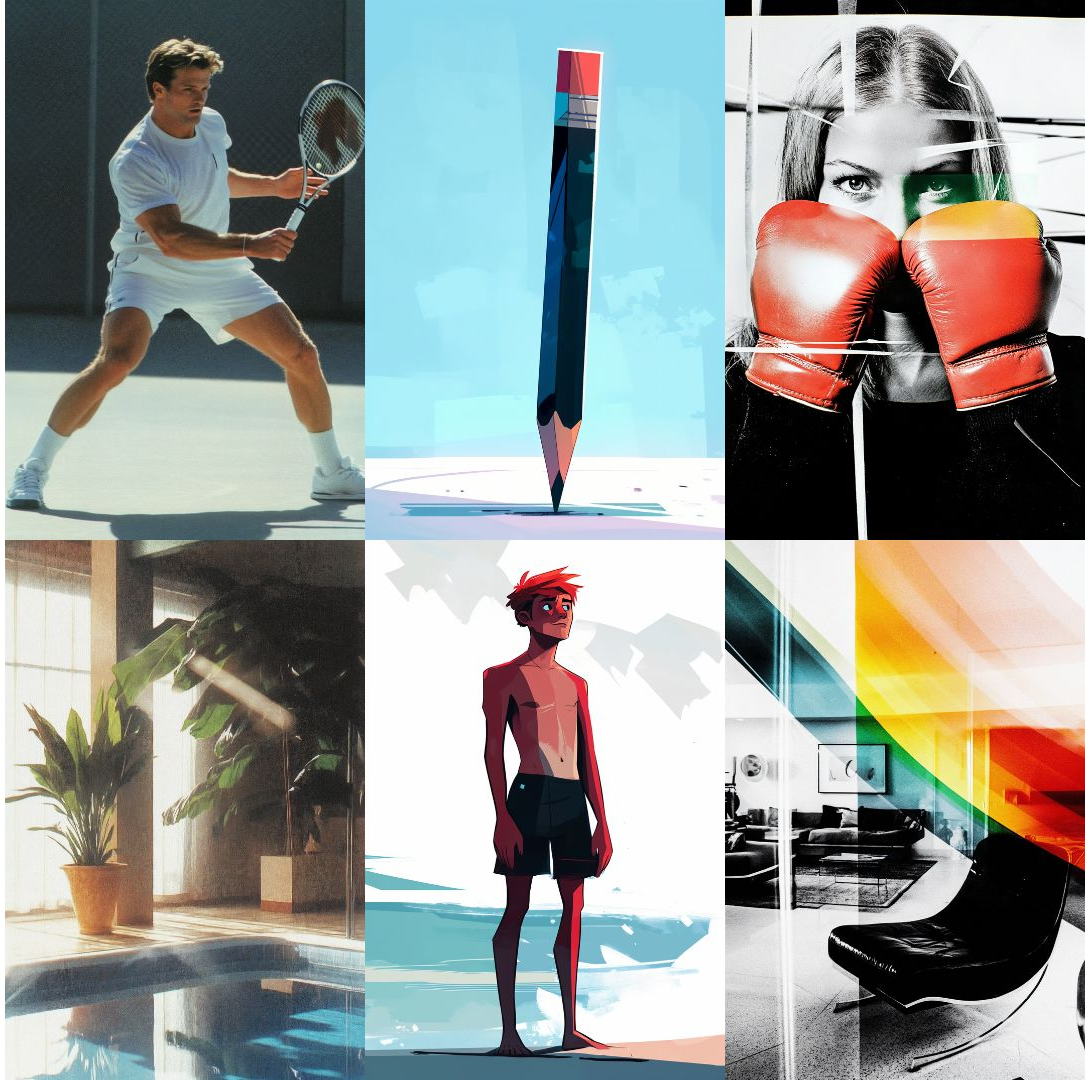
Llama 3.1

FLUX

PAR BLACK FOREST

RESSOURCES MIDJOURNEY

CODES SREF



CHIC AVANT-GARDE EDITORIAL PHOTO

--SREF 3752650426

PROMPT : **PHOTO OF [SUJET], BLACK AND WHITE, WHITE --SREF 3752650426 --AR 2:3**

--sref 44243

--sref 44242

--sref 33464

--sref 88019

--sref 87023

--sref 85169

--sref 56321

--sref 65982

--sref 69145

CONF’

“Les dessous de Llama 3”

Xavier Martinet

AI Research Engineer @ Meta

Meta

Le **18/09/2024** à **19h**

Au Palace, 4 rue Voltaire, Nantes

_icilundi

Screeb RCA [sfeir] lonestone