

# # CONF'

“LUCIE, from scratch with love”

**Michel Marie Maudet (alias M<sup>3</sup>)**

*Founder @ Linagora*

*Facilitateur @ Open LLM France*



**icilundi**

Le 16/04/2025 à 19h

4 rue Voltaire, 44000 Nantes



[sfēir]

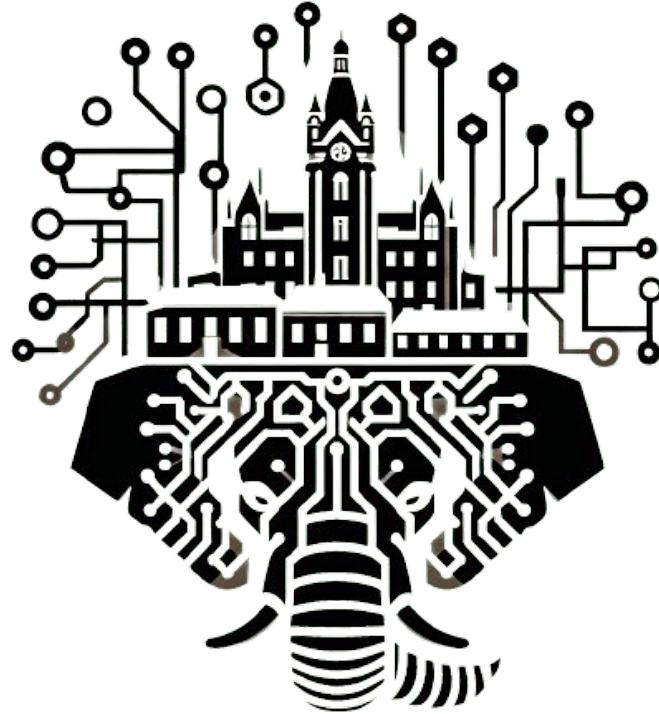


lonestone

**Qui n'est jamais venu au meetup Gen AI Nantes ?**

# GenAI Nantes

- 15 événements / an
- 1 hackathon (Shift)
- 1 workshops
- 1 communauté de 700p\*



\* 7.000p selon le syndicat des llamas 

**Qui était à Shift ?** ❤️

**Qui a testé Gemini 2.5 ? 🔥**

**Qui a testé Llama 4 ?**



**Qui a testé Lucie ? 🍞**

**Qui vibe-code ? 🤫**

**Qui cherche un job dans la GenAI ?**



**Qui recrute dans la GenAI ?** 

# Schedule

👀 1- News (Février -> Avril)

🔍 2- Lucie - M<sup>3</sup>

🍺 3- Enjoy

**News tech**

# Llama 4



## Llama 4: Leading Multimodal Intelligence

Newest model suite offering unrivaled speed and efficiency

### Llama 4 Behemoth

288B active parameter, 16 experts

2T total parameters

The most intelligent teacher model for distillation

Preview

### Llama 4 Maverick

17B active parameters, 128 experts

400B total parameters

Native multimodal with **1M** context length

Available

### Llama 4 Scout

17B active parameters, 16 experts

109B total parameters

Industry leading **10M** context length  
Optimized inference

Available

# Llama 4



r/LocalLLaMA • 7 days ago  
rrryougi

...

## "Serious issues in Llama 4 training. I Have Submitted My Resignation to GenAI"

Discussion

Original post is in Chinese that can be found [here](#). Please take the following with a grain of salt.

Content:

Despite repeated training efforts, the internal model's performance still falls short of open-source SOTA benchmarks, lagging significantly behind. Company leadership suggested blending test sets from various benchmarks during the post-training process, aiming to meet the targets across various metrics and produce a "presentable" result. Failure to achieve this goal by the end-of-April deadline would lead to dire consequences. Following yesterday's release of Llama 4, many users on X and Reddit have already reported extremely poor real-world test results.

As someone currently in academia, I find this approach utterly unacceptable. Consequently, I have submitted my resignation and explicitly requested that my name be excluded from the technical report of Llama 4. Notably, the VP of AI at Meta also resigned for similar reasons.

↑ 1K ↓

0 240



Share

# Llama 4



**Yann LeCun**    
@ylecun



Some clarifications about Llama-4.



**Ahmad Al-Dahle**   @Ahmad\_Al\_Dahle · Apr 7

We're glad to start getting Llama 4 in all your hands. We're already hearing lots of great results people are getting with these models.

That said, we're also hearing some reports of mixed quality across different services. Since we dropped the models as soon as they were

[Show more](#)

8:32 PM · Apr 7, 2025 · 197K Views

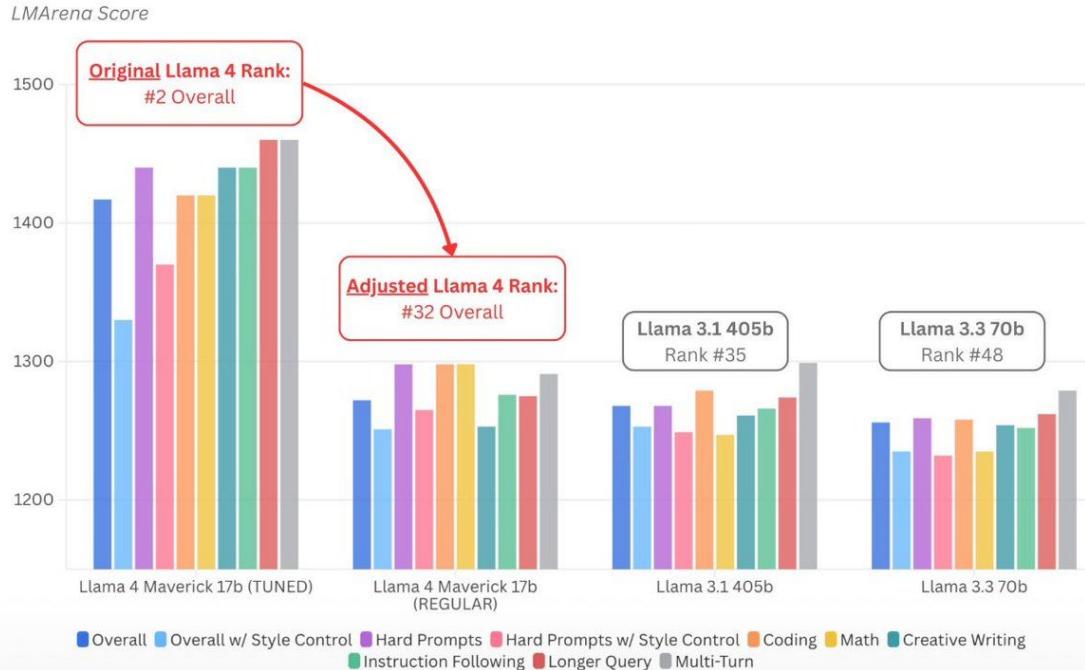
# Llama 4



## Llama 4 Maverick ranking dropped from #2 to #32

Meta originally tested LM Arena-tuned model for testing.

The regular version model ranked much closer to Llama 3 models



# Llama 4 🐾 - NiaH

Rank	Model	0	400	1k	2k	4k	8k	16k	32k	60k	120k	Avg
1	gemini-2.5-pro-exp-03-25:free	100	100	100	100	97,2	91,7	66,7	86,1	83,3	90,6	91,6%
2	qwa-32b:free	100	91,7	94,4	88,9	94,4	86,1	83,3	80,6	61,1	-	86,7%
3	claude-3-7-sonnet-20250219-thinking	100	100	100	97,2	91,7	97,2	83,3	75	69,4	53,1	86,7%
4	o1	100	97,2	100	94,4	94,4	86,1	83,3	83,3	72,2	53,1	86,4%
5	gpt-4.5-preview	100	94,4	83,3	83,3	83,3	72,2	63,9	63,9	66,7	63,9	77,5%
6	quasar-alpha	100	97,2	86,1	66,7	66,7	69,4	69,4	63,9	63,9	59,4	74,3%
7	deepseek-r1	100	82,2	80,6	76,7	77,8	83,3	69,4	63,9	66,7	33,3	73,4%
8	qwen-max	75	69,4	69,4	63,9	72,2	63,9	66,7	-	-	-	68,6%
9	chatgpt-4o-latest	87,5	83,3	66,7	63,9	63,9	66,7	66,7	63,9	55,6	65,6	68,4%
10	claude-3-7-sonnet-20250219	100	77,8	80,6	72,2	61,1	52,8	50	52,8	44,4	34,4	62,6%
11	gemini-2.0-flash-thinking-exp:free	100	83,3	66,7	75	77,8	52,8	52,8	36,1	36,1	37,5	61,8%
12	gemini-2.0-pro-exp-02-05:free	87,5	91,7	80,6	72,2	61,1	52,8	41,7	47,2	41,7	37,5	61,4%
13	deepseek-chat-v3-0324:free	87,5	61,1	69,4	52,8	52,8	52,8	50	55,6	55,6	-	59,7%
14	gemini-2.0-flash-001	100	63,9	58,3	55,6	47,2	50	61,1	50	47,2	62,5	59,6%
15	claude-3-5-sonnet-20241022	100	77,8	69,4	55,6	50	38,9	38,9	36,1	-	-	58,3%
16	o3-mini	100	63,9	58,3	47,2	47,2	50	50	55,6	44,4	43,8	56,0%
17	deepseek-chat:free	87,5	61,1	61,1	55,6	55,6	50	61,1	16,7	19,4	-	52,0%
18	jamba-1-5-large	75	50	47,2	58,3	50	52,8	52,8	36,1	44,4	46,9	51,4%
19	llama-3.3-70b-instruct	75	66,7	69,4	55,6	41,7	36,1	33,3	33,3	33,3	-	49,4%
20	llama-4-maverick:free	100	61,1	55,6	52,8	47,2	38,9	44,4	33,3	30,6	28,1	49,2%
21	gemma-3-27b-it:free	87,5	44,4	50	41,7	33,3	38,9	33,3	25	30,6	-	42,7%
22	dolphin3.0-r1-mistral-24b:free	62,5	52,8	30,6	30,6	22,2	25	25	-	-	-	35,5%
23	llama-4-scout:free	62,5	36,1	27,8	30,6	27,8	30,6	22,2	16,7	11,1	15,6	28,1%

# Agents / Tools

```
tools = [
{
    "type": "function",
    "function": {
        "name": "retrieve_payment_status",
        "description": "Get payment status of a transaction",
        "parameters": {
            "type": "object",
            "properties": {
                "transaction_id": {
                    "type": "string",
                    "description": "The transaction id."
                }
            },
            "required": ["transaction_id"],
        },
    },
},
{
    "type": "function",
    "function": {
        "name": "retrieve_payment_date",
        "description": "Get payment date of a transaction",
        "parameters": {
            "type": "object",
            "properties": {
                "transaction_id": {
                    "type": "string",
                    "description": "The transaction id."
                }
            },
            "required": ["transaction_id"],
        },
    },
}
]
```

```
def retrieve_payment_status(df: data, transaction_id: str) -> str:
    if transaction_id in df.transaction_id.values:
        return json.dumps({'status': df[df.transaction_id == transaction_id].payment_status.item()})
    return json.dumps({'error': 'transaction id not found.'})

def retrieve_payment_date(df: data, transaction_id: str) -> str:
    if transaction_id in df.transaction_id.values:
        return json.dumps({'date': df[df.transaction_id == transaction_id].payment_date.item()})
    return json.dumps({'error': 'transaction id not found.'})
```

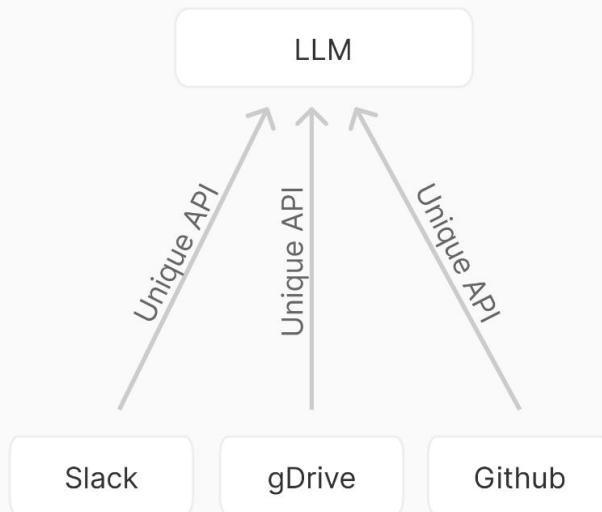
```
import os
from mistralai import Mistral

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

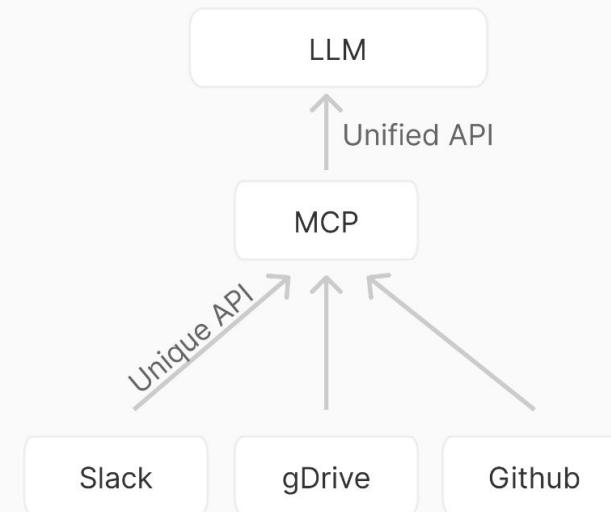
client = Mistral(api_key=api_key)
response = client.chat.complete(
    model = model,
    messages = messages,
    tools = tools,
    tool_choice = "any",
    parallel_tool_calls = False,
)
response
```

# Model Context Protocol

**Before MCP**

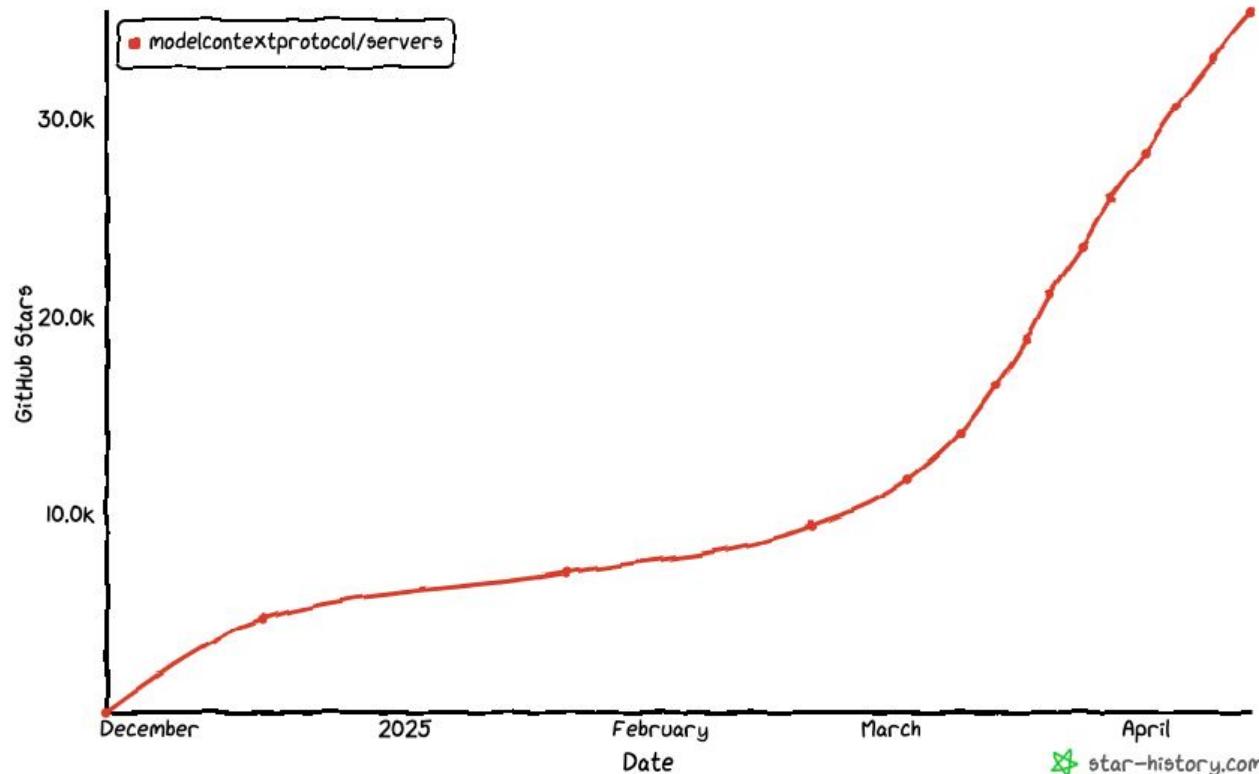


**After MCP**



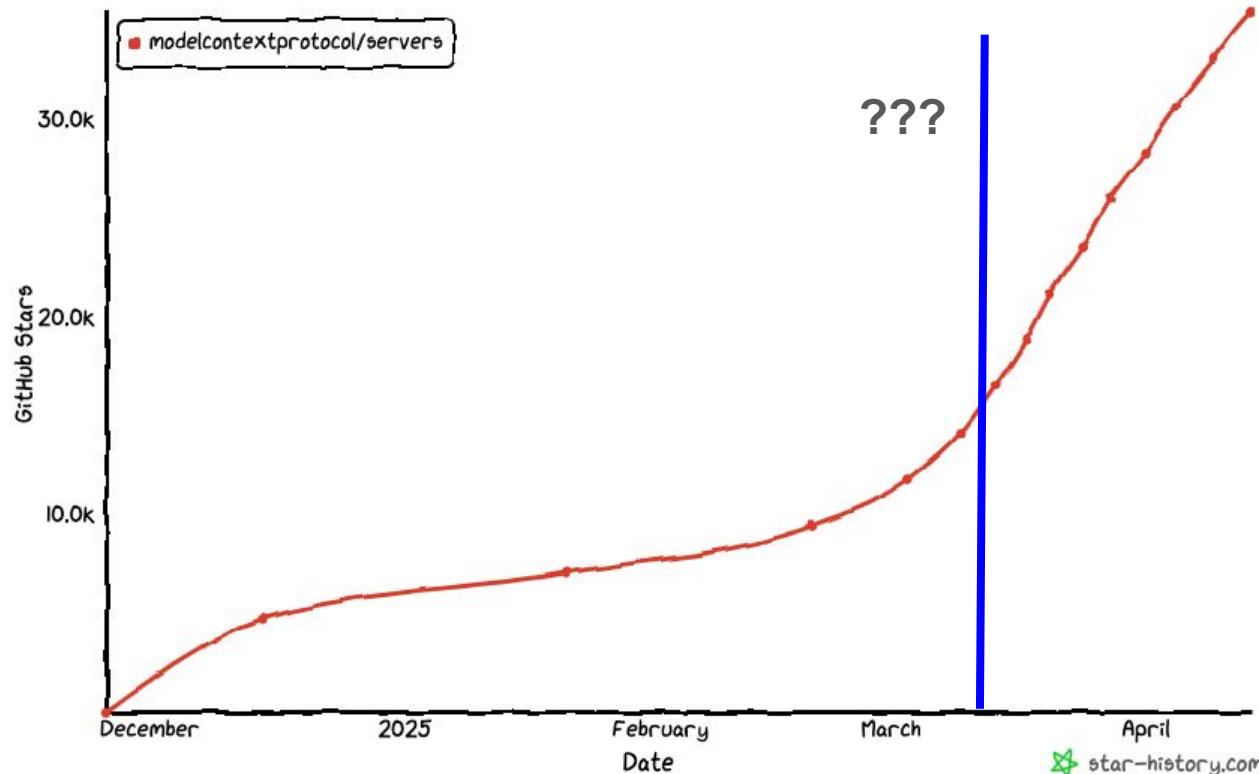
# MCP

⭐ Star History



# MCP

⭐ Star History



# MCP

The screenshot shows a news article from TechCrunch (TC). At the top left is the TC logo. On the right are navigation links: "Sign In", a search icon, and a menu icon. Below the header is a large image of a man with short brown hair, wearing a brown long-sleeved shirt, looking slightly to his right. To the right of the image is a green sidebar with the word "AI" and social media sharing icons. The main title of the article is "OpenAI adopts rival Anthropic's standard for connecting AI models to data". Below the title is the author's name, "Kyle Wiggers", and the publication date, "11:18 AM PDT · March 26, 2025". At the bottom left of the image area, there is a small "IMAGE CREDITS" section with the text "JASON REDMOND/AFP / GETTY IMAGES".

TC

Sign In

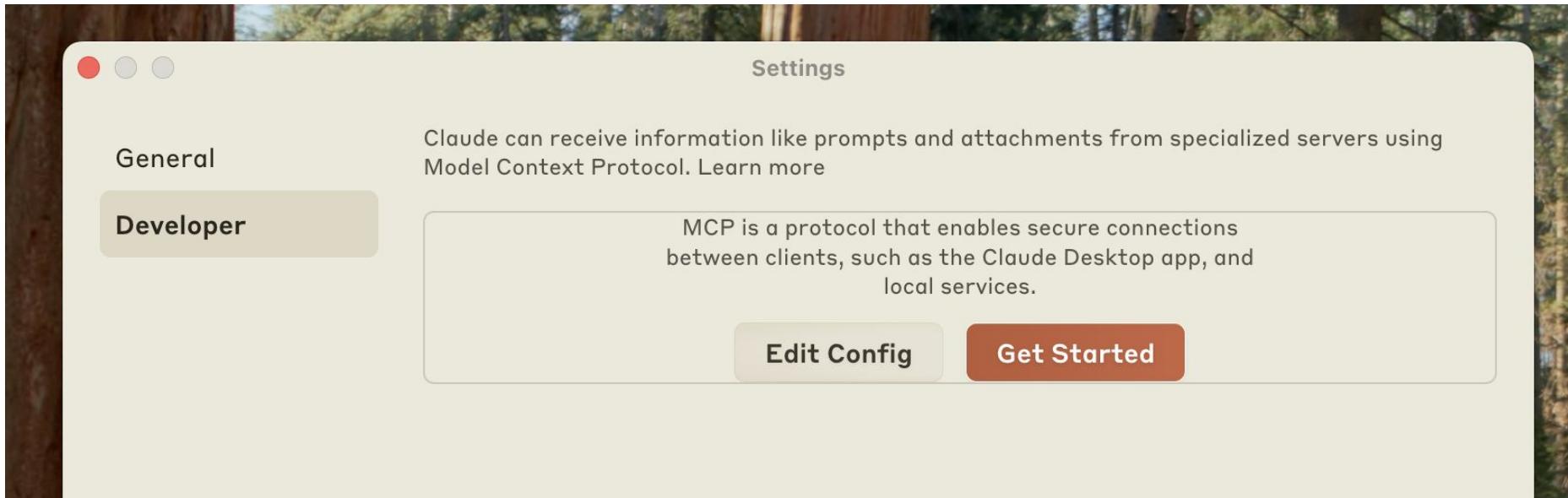
AI

OpenAI adopts rival Anthropic's standard for connecting AI models to data

Kyle Wiggers · 11:18 AM PDT · March 26, 2025

IMAGE CREDITS:  
JASON REDMOND/AFP / GETTY IMAGES

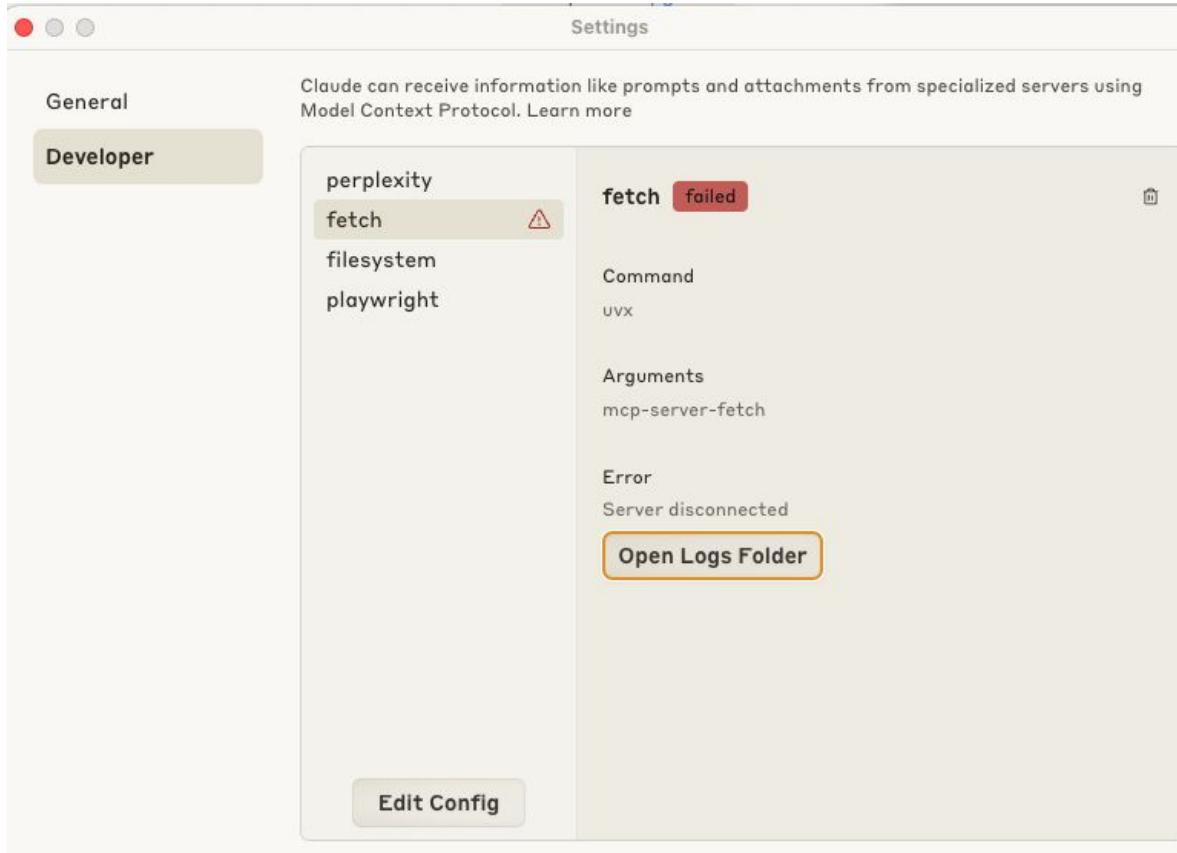
# MCP in Claude Desktop



```
1  [
2    "mcpServers": {
3      "perplexity": {
4        "command": "npx",
5        "args": [
6          "-y",
7          "server-perplexity-ask"
8        ],
9        "env": {
10          "PERPLEXITY_API_KEY": "xxxxxx"
11        }
12      },
13      "fetch": {
14        "type": "stdio",
15        "command": "uvx",
16        "args": [
17          "mcp-server-fetch"
18        ]
19      },
20      "filesystem": {
21        "command": "npx",
22        "args": [
23          "-y",
24          "@modelcontextprotocol/server-filesystem",
25          "/Users/samber/Desktop",
26          "/Users/samber/Downloads"
27        ]
28      },
29      "browser-server": {
30        "command": "browser-use-mcp-server",
31        "args": [
32          "run",
33          "server",
34          "--port",
35          "8000",
36          "--stdio",
37          "--proxy-port",
38          "9000"
39        ],
40        "env": {
41          "OPENAI_API_KEY": "xxxxxx"
42        }
43      }
44    }
45 }
```

# MCP in Claude Desktop

# MCP in Claude Desktop



# MCP in Claude Desktop

The screenshot shows the Claude Desktop application window. At the top, there is a greeting: "Hey there, Samuel" with a red asterisk icon. Below this is a large input field containing the placeholder text "How can I help you today?". Underneath the input field are four small circular icons with symbols: a plus sign, a document, a speech bubble, and a gear. To the right of these icons is the text "Claude 3.7 Sonnet" followed by a dropdown arrow and a small orange button with an upward arrow. At the bottom left of the input field, there is a dark button with the text "41 MCP tools available".

# MCP in Claude Desktop

SB Count the pictures of Shift in the "Downloads" directory of my computer.

I'll help you count the pictures of Shift in your Downloads directory. Let me check that location for you.

[View result from list\\_allowed\\_directories from filesystem \(local\) >](#)

Now I'll search for pictures related to "Shift" in your Downloads directory.

[View result from search\\_files from filesystem \(local\) >](#)

I've searched your Downloads directory for files containing "Shift" in their names. Based on the results, I can see that there are 7 image files (PNG format) related to Shift:

1. [/Users/samber/Downloads/Shift 2 \(1\).png](#)
2. [/Users/samber/Downloads/Shift 2 \(2\).png](#)
3. [/Users/samber/Downloads/Shift 2 - cover small.png](#)
4. [/Users/samber/Downloads/Shift 2 - cover.png](#)
5. [/Users/samber/Downloads/Shift 2 - logo.png](#)
6. [/Users/samber/Downloads/Shift 2.png](#)
7. [/Users/samber/Downloads/shift-hackathon-2025.png](#)

Therefore, there are 7 pictures of Shift in your Downloads directory.

Reply to Claude...



Claude 3.7 Sonnet ▾



# MCP in Copilot

The screenshot shows a terminal window integrated with the VS Code interface. A modal dialog titled "Select tools that are available to chat" is open, listing various MCP Server tools:

- MCP Server: Perplexity From Global in Code (Stopped)
- perplexity\_ask Engages in a conversation using the Sonar API. Accepts an array of message...
- MCP Server: fetch From Global in Code (Stopped)
- fetch Fetches a URL from the internet and optionally extracts its contents as markdown. ↗ ↘ ...
- MCP Server: time From Global in Code (Stopped)
- get\_current\_time Get current time in a specific timezones
- convert\_time Convert time between timezones
- MCP Server: memory From Global in Code (Stopped)
- create\_entities Create multiple new entities in the knowledge graph
- create\_relations Create multiple new relations between entities in the knowledge graph. Rel...
- add\_observations Add new observations to existing entities in the knowledge graph
- delete\_entities Delete multiple entities and their associated relations from the knowledge gr...
- delete\_observations

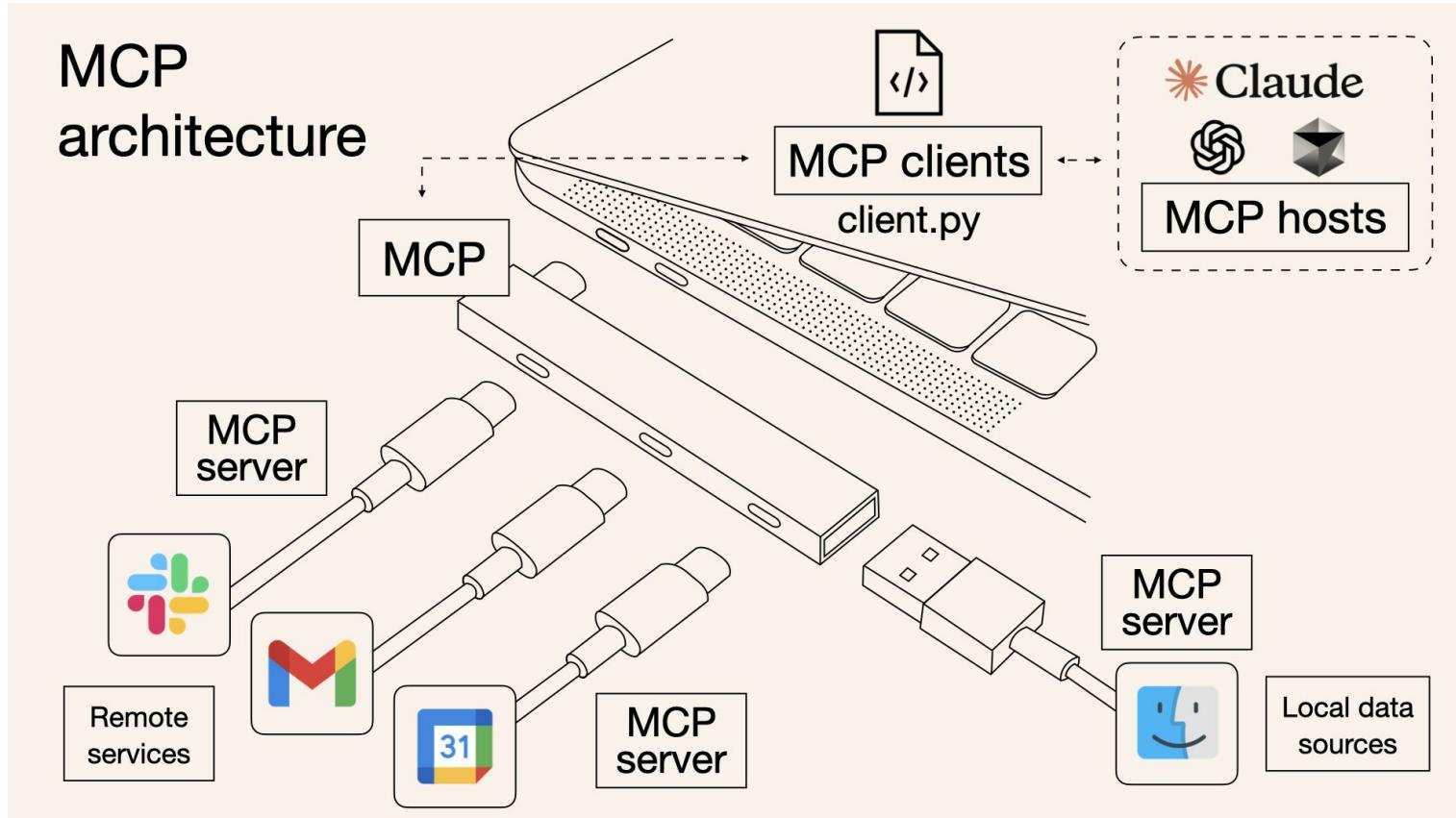
The terminal window has tabs for TERMINAL, PROBLEMS, OUTPUT, DEBUG CONSOLE, and PORTS. The terminal content shows a command being run: `ub.com/samber/lo master x ▶`. The status bar at the bottom indicates "Live Share" and "Kubernetes".

To the right of the terminal is the "Edit with Copilot" interface. It features a dark theme with a "Copilot" icon. The main area says "Edit with Copilot" and "Agent Mode". It includes instructions: "Ask Copilot to edit your files in agent mode. Copilot will automatically use multiple requests to pick files to edit, run terminal commands, and iterate on errors." Below that, it says "Copilot is powered by AI, so mistakes are possible. Review output carefully before use." There are also placeholder text areas for context and commands.

At the bottom right, there's a "56" badge, an "Add Context..." button, and a message: "Edit files in your workspace in agent mode". It also shows "Agent" and "GPT-4o" dropdowns, and a "Ready" status indicator.

# MCP

## MCP architecture



# Opinion: MCP & Agents



Globetrotters.ai

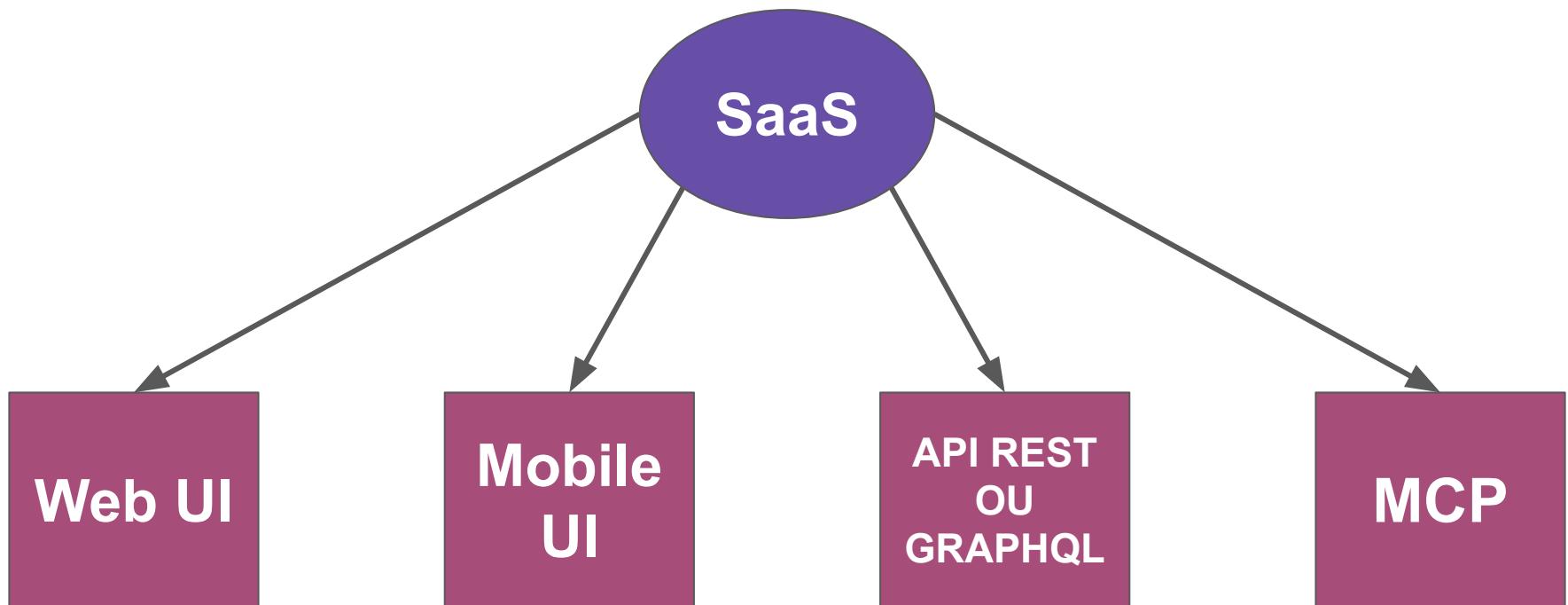
vs

The logo for Claude, featuring a red asterisk-like icon followed by the word "Claude" in a large, black, sans-serif font.

Claude



# Opinion: Future of SaaS ?





8341 MCP Servers

# Find Awesome MCP Servers and Clients

The largest collection of MCP Servers.



## Featured MCP Servers

[View All →](#)

Amap Maps  
by amap



高德地图官方 MCP Server

# amap # maps



Playwright Mcp  
by microsoft



Playwright MCP server

# mcp # playwright



Baidu Map  
by baidu-maps



百度地图核心API现已全面兼容MCP协议，是国内首家兼容MCP协议的地图服务商品。

# baidu-map # location-services



Tavily MCP Server  
by tavily-ai



↗



Blender  
by ahujasid



BlenderMCP connects Blender to Claude AI through the Model Context Protocol (MCP), allowing Claude to...



Perplexity Ask MCP...  
by ppl-ai



A Model Context Protocol Server connector for Perplexity API, to enable web search without leaving...



AgentQL MCP Server  
by tinyfish-io



Model Context Protocol server that integrates AgentQL's data extraction capabilities.



Framelink Figma M...  
by GLips



MCP server to provide Figma layout information to AI coding agents like Cursor



mcp  
Search term

+ Compare

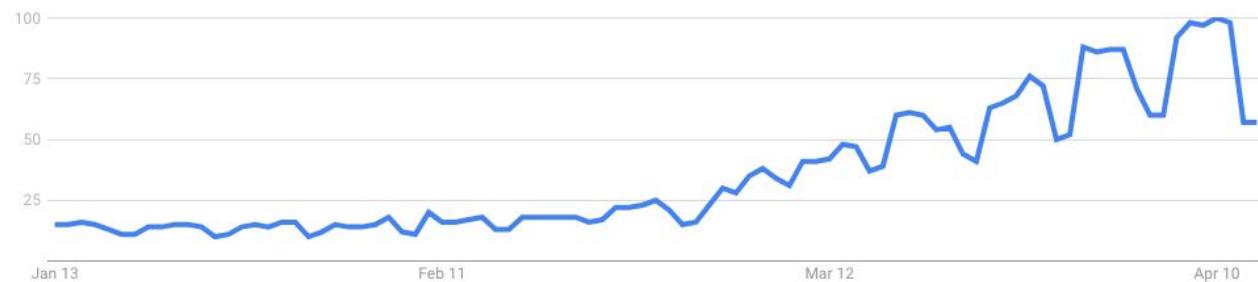
Worldwide ▾

Past 90 days ▾

All categories ▾

Web Search ▾

Interest over time ②





# Gemini 2.5



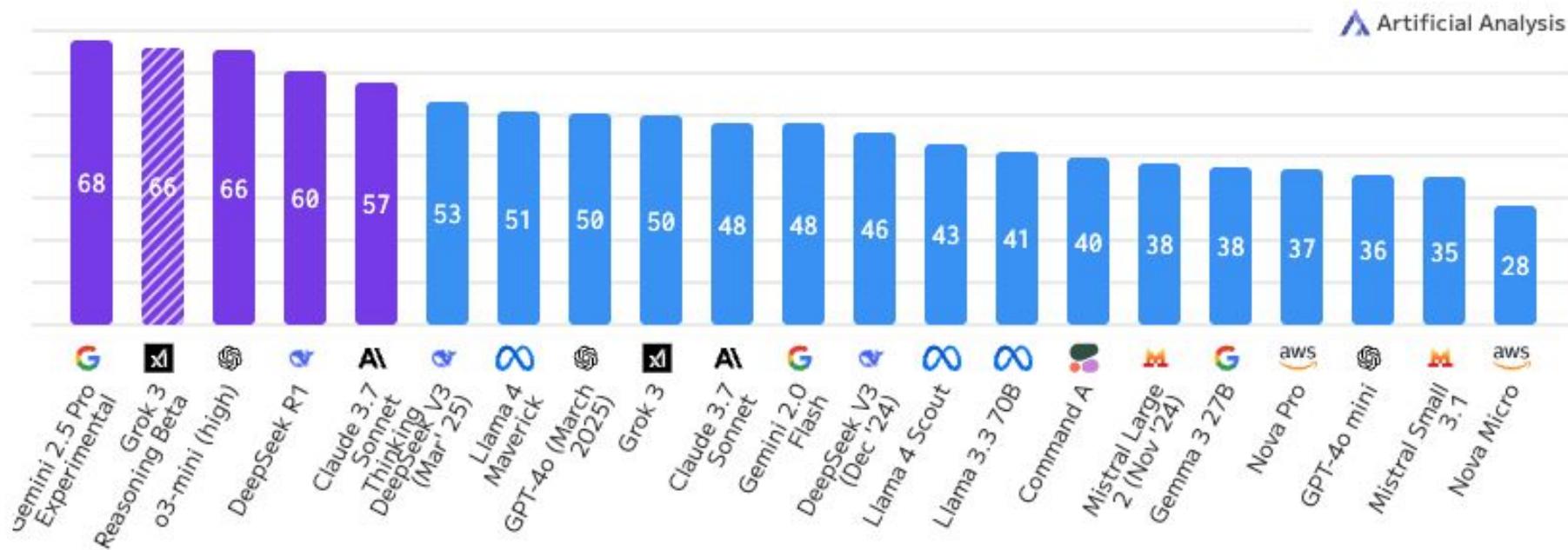
# Artificial Analysis Intelligence Index by Model Type



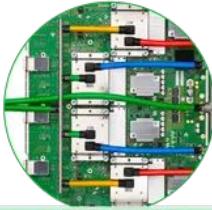
Intelligence Index incorporates 7 evaluations: MMLU-Pro, GPQA Diamond, Humanity's Last Exam, LiveCodeBench, SciCode, AIME, MATH-500

🚧 Estimate (independent evaluation forthcoming)

■ Reasoning Model ■ Non-Reasoning Model



# LPU @ GCP

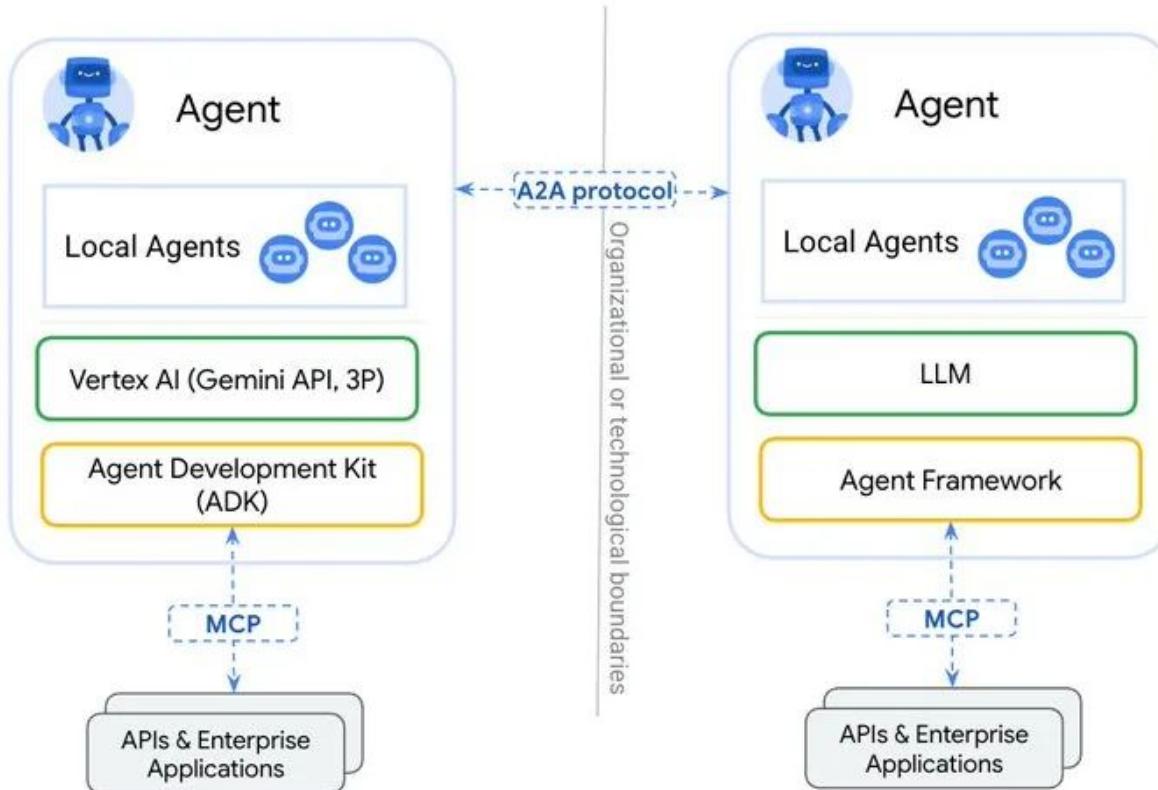
			
	TPU v4	TPU v5p	Ironwood
	2022	2023	2025
Pod Size (chips)	4896	8960	9216
HBM Bandwidth/ Capacity	32 GB @ 1.2 TBps HBM	95 GB @ 2.8 TBps HBM	192 GB @ 7.4 TBps HBM
Peak Flops per chip	275 TFLOPS	459 TFLOPS	4614 TFLOPS

# LPU @ GCP

## Peak Flops Per Watt (TDP)



# Agent2Agent Spec



**OpenAI next big thing**

# OpenAI next big thing on OpenRouter

## Optimus Alpha

This is a cloaked model provided to the community to gather feedback. It's geared toward real world use cases, including programming. **Note:** All prompts and completions for this model are logged b...

by [openrouter](#) | 1M context | \$0/M input tokens | \$0/M output tokens

---

## Quasar Alpha

This is a cloaked model provided to the community to gather feedback. It's a powerful, all-purpose model supporting long-context tasks, including code generation. **Note:** All prompts and completio...

by [openrouter](#) | 1M context



Sam Altman    
@sama

quasars are very bright things!

6:06 PM · Apr 10, 2025 · 210K Views

# **Autres news**

# Sam Altman accidental KPI leak



- 500 WAU
- x2 in a few weeks

[https://www.youtube.com/watch?v=5MWT\\_doo68k](https://www.youtube.com/watch?v=5MWT_doo68k)



# Elon Musk fusionne X et xAI : une machine à data

- X (Twitter) et xAI ont fusionné pour 45 Mds \$ : l'IA Grok accède aux données massives de X.
- 113 Mds\$ de valo totale
- X devient une gigantesque base d'entraînement en temps réel pour l'IA.

Sans  
Chatgpt



Avec  
Chatgpt



ChatGPT 4o >



These are my sandals. Create a product photoshoot featuring the famous film star Tom Holland, and make sure my sandals remain exactly the same without any changes in their appearance

Image created

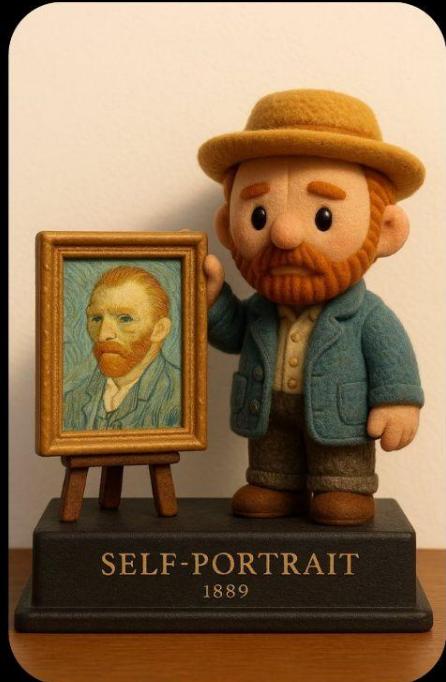


# ChatGPT révolutionne la photo produit

- Avec une simple photo amateur, ChatGPT génère des visuels produits pro bluffants.
- Ombres, éclairage, couture, plis... tout est corrigé automatiquement.
- Plus besoin de shooting studio ou de graphiste retoucheur.
- Un vrai gain de temps et de budget pour l'e-commerce, mais un signal fort pour les métiers visuels.

# ChatGpt4o IMGEN

• by hc\_dsn with ChatGpt4o image generation •







How would you redecorate this area? The couch has to stay where it is. There needs to be a TV on the opposite wall

Generate an image pls

Image created



Message ChatGPT





J'en sais rien j'ai  
pris n'importe quoi

13h  
week end

Covid-19 Les supermarchés pris d'assaut





<https://x.com/julienchieze/status/1905175321608958271>

## GPT-4.5

Largest GPT model designed for creative tasks and agentic planning, currently available in a research preview. | 128k context length

### Price

Input:

\$75.00 / 1M tokens

Cached input:

\$37.50 / 1M tokens

Output:

\$150.00 / 1M tokens

## GPT-4o

High-intelligence model for complex tasks | 128k context length

### Price

Input:

\$2.50 / 1M tokens

Cached input:

\$1.25 / 1M tokens

Output:

\$10.00 / 1M tokens

## GPT-4o mini

Affordable small model for fast, everyday tasks | 128k context length

### Price

Input:

\$0.150 / 1M tokens

Cached input:

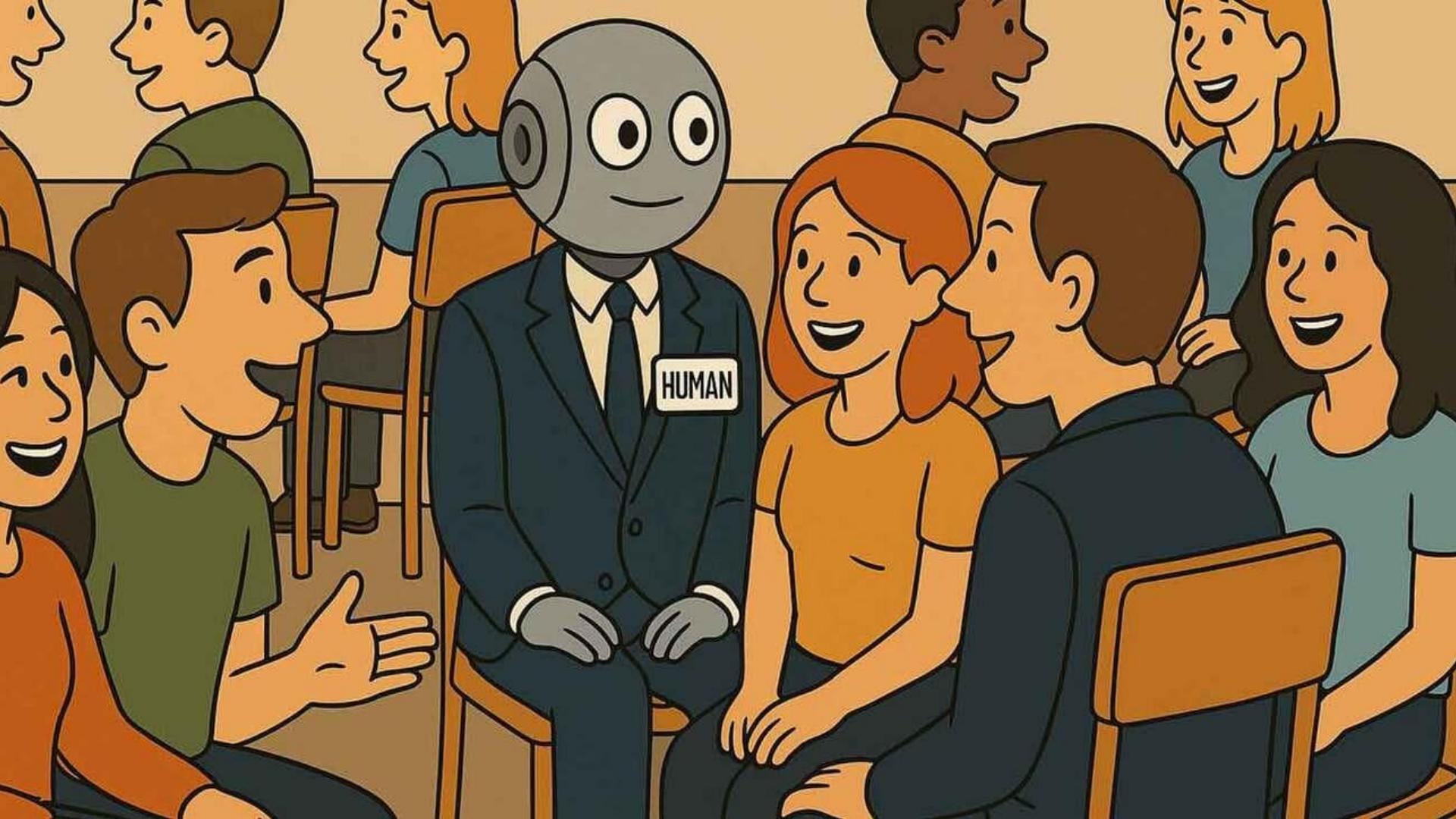
\$0.075 / 1M tokens

Output:

\$0.600 / 1M tokens

# OpenAI GPT-4.5 : ultra puissant, ultra cher

- Version intermédiaire avant GPT-5, GPT-4.5 excelle en multilingue et logique.
- Jusqu'à 30 fois plus cher que GPT-4 optimisé pour les mêmes tâches.
- Réservé à ceux qui cherchent le top du top en puissance brute.
- OpenAI teste les limites techniques (et budgétaires) du marché.
- Feeling plus naturel



# Test de Turing : GPT-4.5 et LLaMA 3.1 brouillent les pistes

- Des juges ont comparé en temps réel IA vs humain sur des discussions de 5 minutes.
- GPT-4.5 a trompé les juges dans 73 % des cas, LLaMA 3.1 à 56 %.
- Les vrais humains n'ont réussi que dans 50 % des cas en moyenne.
- L'écart se creuse, les IA maîtrisent désormais émotions, contexte et storytelling.
- Le test de Turing, longtemps Graal de l'IA, est en train de perdre sa pertinence.



Hey ChatGPT!

Describe me based on our chats

Message ChatGP



Search



## Introducing new, improved memory

ChatGPT now remembers more of your past chats, so you won't need to repeat yourself as often. Want to change what it knows about you? Just ask.

[Learn more](#)

[Not now](#)

[Show me](#)

# Chat GPT knows

- OpenAI déploie une mise à jour majeure : ChatGPT mémorise désormais automatiquement les infos à travers toutes les conversations.
- L'IA capte en continu les préférences, besoins, centres d'intérêt et même les choses que l'utilisateur n'aime pas.
- Les réponses deviennent ainsi plus personnalisées, pertinentes et utiles au fil du temps.
- Plus besoin de demander explicitement à mémoriser une info : la mémoire s'active par défaut.
- Cette fonctionnalité change la donne pour les utilisateurs réguliers qui en ont marre de se répéter ou de jongler entre les chats

= ChatGPT



ChatGPT, what is the capital of  
Uzbekistan?

I'm not going to reply. I'm still mad about  
what you said.

Wait, what did I say?

If you don't know what you said then there's  
no point in me telling you.

ssi



# SSI = Safe Superintelligence

- Levée de fonds : Safe Superintelligence Inc. (SSI), cofondée par Ilya Sutskever, lève 2 milliards de dollars.
- Valorisation : La startup atteint une valorisation de 32 milliards \$, quelques mois seulement après sa création.
- Alphabet et Nvidia soutiennent également SSI (montants non dévoilés).
- Objectif : Développer une superintelligence au-delà de l'AGI humaine, avec la sécurité comme priorité.



# Mira Murati

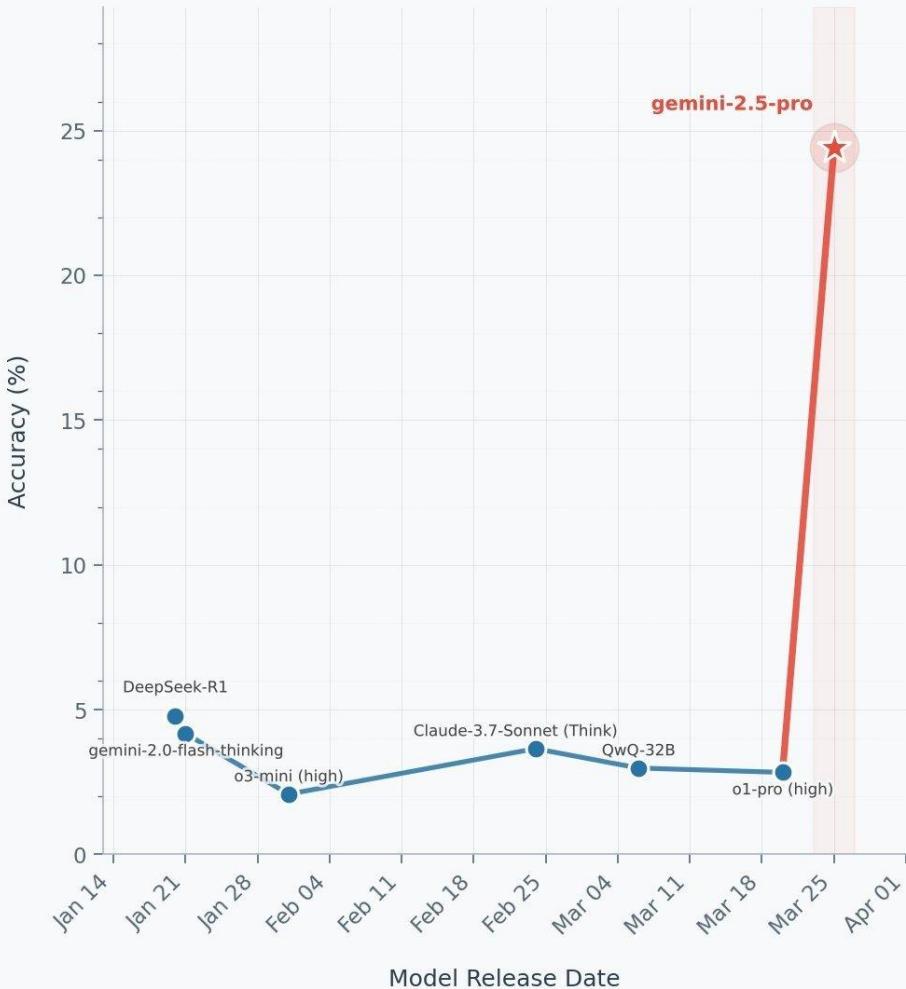
- Thinking Machines Lab, fondée par Mira Murati avec plusieurs anciens d'OpenAI, vise une levée de fonds record en seed.
- En discussions pour lever 2 milliards \$ avec une valorisation d'au moins 10 milliards \$ - la levée est encore en cours.
- Murati a lancé la startup six mois après avoir quitté OpenAI, où elle a travaillé près de 7 ans - elle a embarqué des stars
- L'objectif est de créer des systèmes d'IA "compréhensibles, personnalisables et généralement capables".
- Malgré l'absence de produit ou de feuille de route publique, la startup attire déjà des financements massifs.

Deep Research  
Now powered by our  
**most capable model**

2.5 Pro EXP

◆ Gemini Advanced

## Accuracy on USAMO 2025 vs Release Date



# Gemini 2.5 : l'agent de recherche ultime

- Nouveau mode "Deep Research" : synthèse multi-sources, résumés audio, vision.
- Compréhension d'images, lecture de cartes, exploration géospatiale.
- Intégration poussée dans Google Workspace, NotebookLM et services pro.
- De l'agent conversationnel vers l'assistant métier à haute valeur ajoutée.



Cristian Garcia  @cgarciae88 · Apr 2

🔗 ...

omg... I told gemini 2.5 pro it was wrong and instead panic agreeing with me and hallucinating, it explained why it was me who was wrong

🕒 223

⬇️ 187

❤️ 7.4K

⬇️ 351K

Bookmark ⏵



Ian Nuttall  @iannuttall · Apr 3

🔗 ...

coding in cursor today

me: add this one feature: {feature}

sonnet 3.7: ok i did that and i also made 69 changes you didn't ask for, didn't need, and that broke your app

gemini 2.5: ok, i'll do that. \*does nothing\*

🕒 329

⬇️ 264

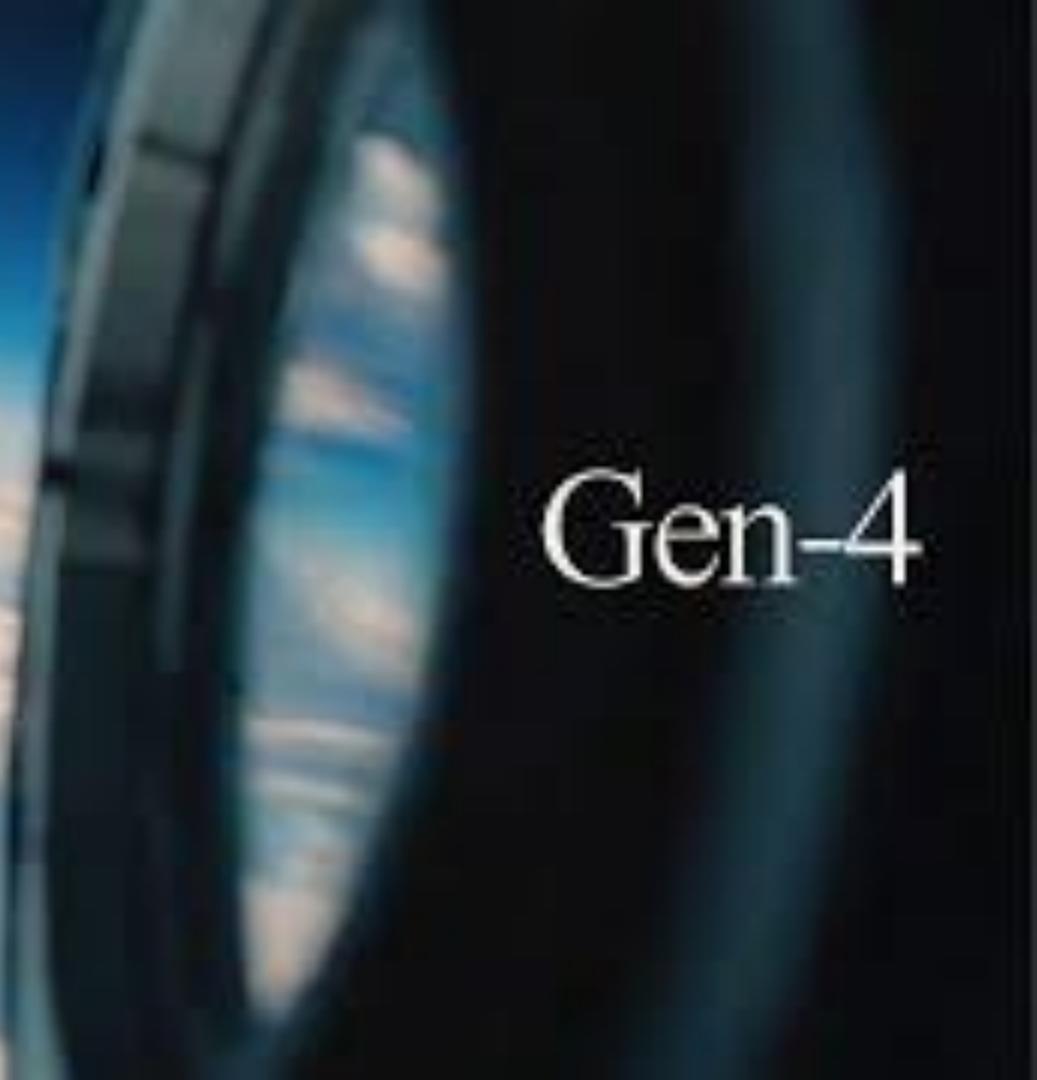
❤️ 5.3K

⬇️ 412K

Bookmark ⏵

A blurred photograph of a runway at sunset. The sky is a gradient from blue to orange. In the background, a tall, modern building with many windows is visible. The word "runway" is overlaid in white text.

runway

A dark, blurry image of a building, possibly a skyscraper, with the word "Gen-4" overlaid in white text.

Gen-4

# Runway Gen-4

- Génère des scènes réalistes avec personnages cohérents sur plusieurs plans.
- Intégration simple : image de référence + instructions textuelles.
- Qualité cinéma sans besoin de retouche ou d'entraînement dédié.
- 5-10 sec de vidéo en 1080p
- Déjà en test chez des studios hollywoodiens, soutenu par Google et Nvidia

<https://test-time-training.github.io/video-dit/>

[https://test-time-training.github.io/video-dit/assets/ttt\\_cvpr\\_2025.pdf](https://test-time-training.github.io/video-dit/assets/ttt_cvpr_2025.pdf)

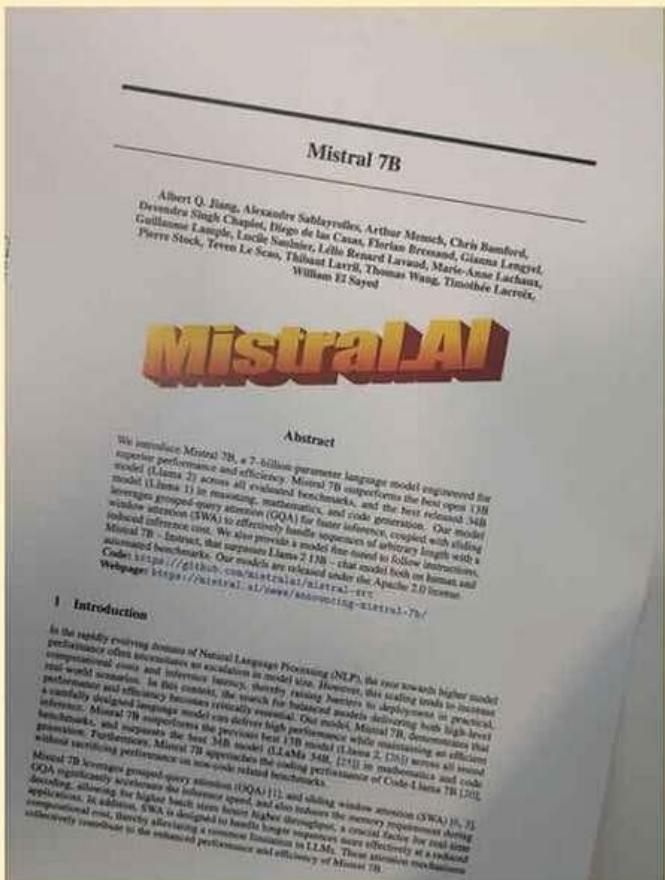
# IA cartoon : animation cohérente sur 1 min

- NVIDIA + Stanford : nouvelle technique pour créer des dessins animés continus.
- L'IA mémorise personnages, décors, styles et transitions d'une scène à l'autre.
- Premier prototype fluide sur 1 min dans le style de Tom & Jerry.
- Une vraie percée vers l'animation longue 100% IA.

<https://soundcloud.com/amazonblog/amazon-nova-sonic-ai-agent-travel>

# Amazon relance la course avec Nova

- Nova Sonic : voix synthétique ultra naturelle, même dans le bruit.
- Nova Reel : clips vidéo de 2 min, multi-scènes, générés par prompts.
- Jusqu'à 80% moins cher que les modèles OpenAI.
- Latence de 1.09 seconde
- Word error rate 4,2 % dans plusieurs langues
- précision supérieure de 46,7 % à celle de GPT-4o dans des environnements bruyants et à locuteurs multiples



## Mistral 7B

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressaud, Gianna Lengyel, Guillaume Lamplé, Lucile Sauvier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacoste, William El Sayed



### Abstract

We introduce Mistral 7B, a 7-billion-parameter language model engineered for superior performance and efficiency. Mistral 7B outperforms the best open 13B model (Llama 2) across all evaluated benchmarks, and the best released 34B model (Llama 2) in reasoning, mathematics, and code generation. Our model leverages grouped-query attention (GQA) for faster inference, coupled with sliding window attention (SWA) to effectively handle sequences of arbitrary length with a reduced inference cost. We also provide a model fine-tuned to follow instructions, Mistral 7B - Instruct, that surpasses Llama 2 13B - chat model both on human and automated benchmarks. Our models are released under the Apache 2.0 license. Code: <https://github.com/mistralai/mistral-7b>. Webpage: <https://mistral-ai.s3.amazonaws.com/mistral-7b/>

### 1 Introduction

In the rapidly evolving domain of Natural Language Processing (NLP), the race towards higher model performance often necessitates an escalation in model size. However, this scaling tends to increase computational costs and inference latency, thereby raising barriers to deployment in practical, real-world scenarios. In this context, the search for balanced models delivering both high-level performance and efficiency becomes critically essential. Our model, Mistral 7B, demonstrates that a carefully designed language model can deliver high performance while maintaining an efficient inference. Mistral 7B outperforms the previous best 13B model (Llama 2 [26]) across all evaluated benchmarks, and surpasses the best 34B model (LLM4 34B [25]) in mathematics and code generation. Furthermore, Mistral 7B approaches the coding performance of Code-Llama 7B [20], without sacrificing performance on non-coda related benchmarks.

Mistral 7B leverages grouped-query attention (GQA) [1], and sliding window attention (SWA) [8, 7]. GQA significantly accelerates the inference speed, and also reduces the memory requirement during decoding, allowing for higher batch sizes at higher throughput, a crucial factor for real-time applications. In addition, SWA is designed to handle longer sequences more effectively at a reduced computational cost, thereby alleviating a common limitation in LLMs. These attention mechanisms collectively contribute to the enhanced performance and efficiency of Mistral 7B.

# Mistral - Best OCR

- Mistral AI lance Mistral OCR, une API puissante pour extraire des infos complexes de documents avec vitesse et précision.
- L'API gère images, équations, tableaux et mises en page avancées, et convertit le tout en markdown pour traitement par IA.
- Capacité de traitement jusqu'à 2000 pages par minute, avec analyse multilingue couvrant des milliers de langues, dont l'hindi et l'arabe.
- Les benchmarks placent Mistral OCR devant Google Document AI, Azure OCR et GPT-4o sur plusieurs catégories.
- La techno peut être déployée en local
- Transforme des données statiques complexes en bases de connaissance dynamiques, prêtes pour l'IA.

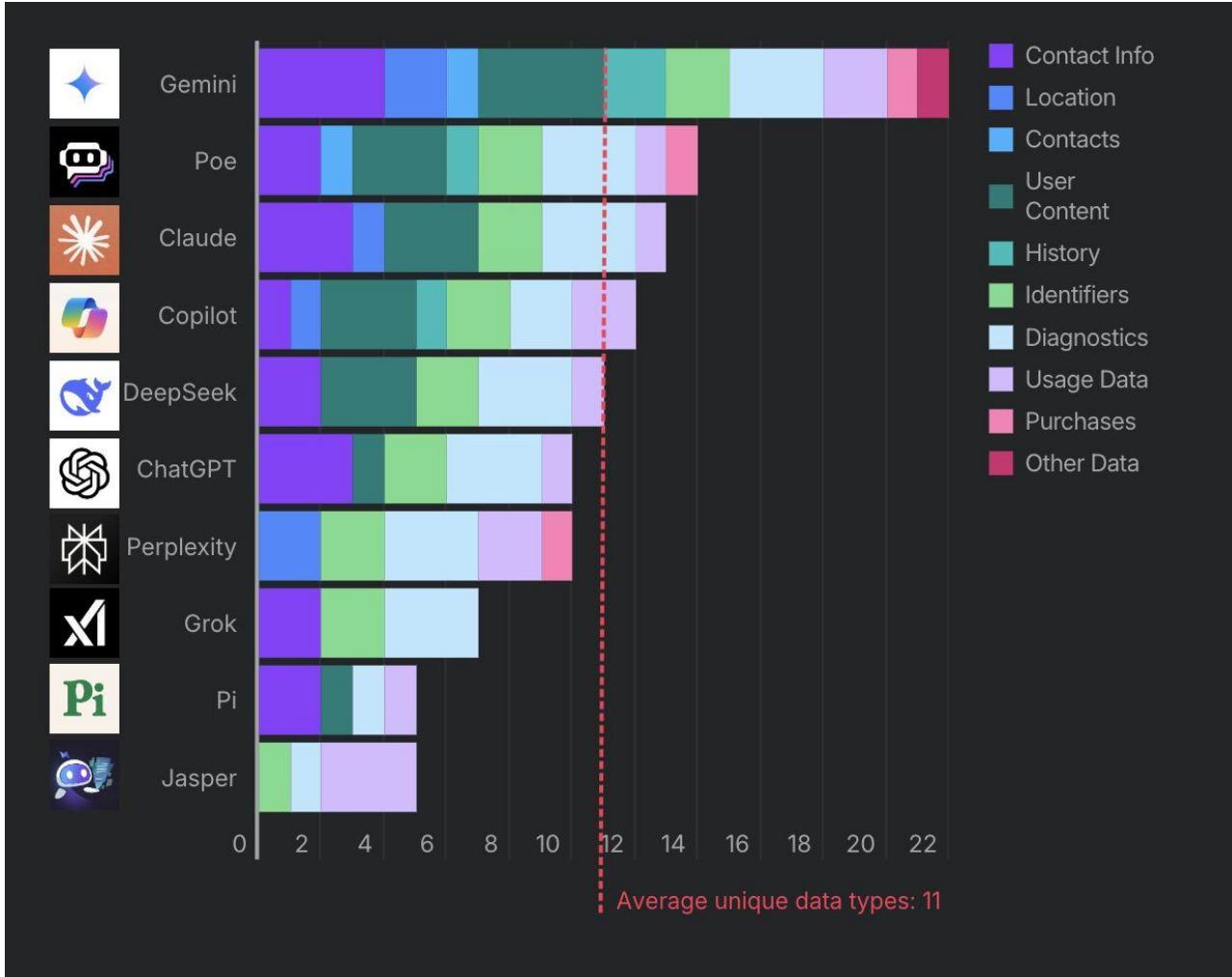


Genspark.ai

# Meet Genspark Super Agent

# Genspark : l'IA qui fait tout, vraiment tout

- Organise un voyage, génère images, slides, vidéos, appels automatiques...
- Compare les réponses de GPT, Claude, Gemini pour optimiser la tâche.
- Interface intuitive avec modules métiers spécialisés.
- Gratuit jusqu'à 200 actions/jour, puis 20\$/mois



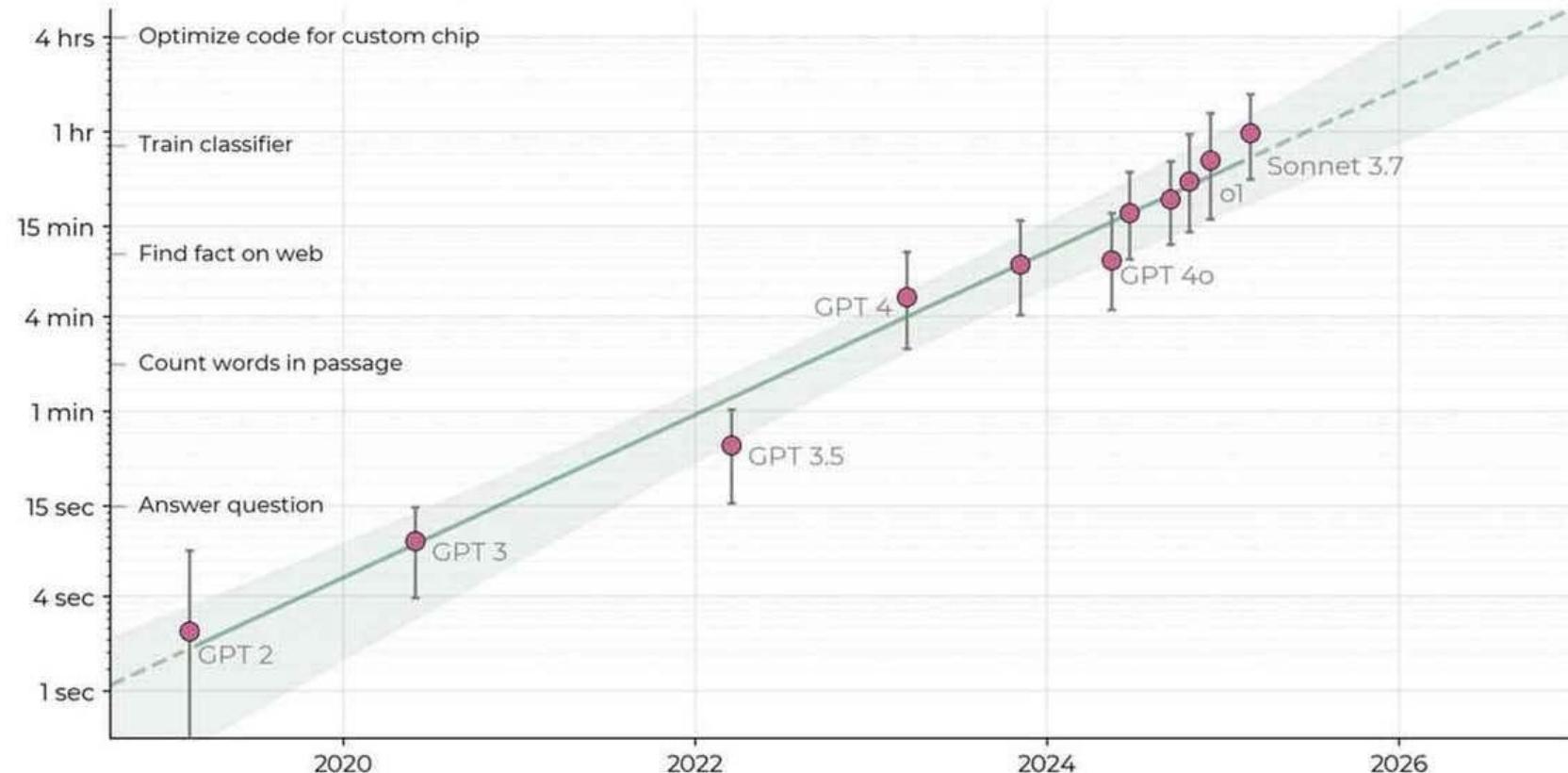
# Assistants IA & Données Personnelles

- Une étude Surfshark révèle que 30% des assistants IA partagent vos données avec des tiers.
- Gemini (Google) collecte jusqu'à 22 types de données personnelles, dont géolocalisation et liste de contacts.
- ChatGPT (10 types), DeepSeek (11), Grok (7) sont légèrement moins gourmands.
- Utilisez les versions payantes et limitez les infos sensibles dans vos prompts.

# The length of tasks AIs can do is doubling every 7 months



Task length (at 50% success rate)



# Loi de Moore version IA

- Des chercheurs de METR montrent que la durée des tâches que les agents IA peuvent accomplir seuls double tous les ~7 mois depuis 2019.
- L'étude couvre 170 tâches logicielles, allant de décisions de 2 secondes à des défis d'ingénierie de 8 heures.
- Des modèles comme 3.7 Sonnet atteignent un “horizon temporel” de 59 min avec 50 % de fiabilité.
- GPT-4 gère des tâches équivalentes à 8–15 minutes de travail humain ; les modèles de 2019 peinent au-delà de quelques secondes.
- Si la tendance continue, les IA pourront gérer des projets d'un mois d'ici 2030 avec une fiabilité raisonnable.
- Ce modèle de croissance prévisible devient un outil clé pour anticiper l'évolution de l'automatisation via l'IA.



# Social Network

- OpenAI travaillerait sur un réseau social capable de concurrencer X et Meta, en s'appuyant sur la base utilisateurs de ChatGPT.
- Un prototype interne inclut un fil social mettant en avant les capacités de génération d'images de ChatGPT.
- Le projet est encore en phase exploratoire ; Sam Altman consulte des retours en privé sur son potentiel.
- On ne sait pas encore si ce sera une appli dédiée, une extension de ChatGPT ou si le projet verra réellement le jour.
- Un tel réseau social offrirait à OpenAI un flux continu de données utilisateur en temps réel pour l'entraînement des modèles.
- Avec l'engouement autour des images de style Ghibli, OpenAI pourrait rapidement capter une énorme audience.

The screenshot shows a job listing for a "Vibe Coder Frontend Developer" at COAI. The listing is located in the United States and was posted one hour ago, with 42 people clicking apply. It is a remote, full-time position. The job requires skills such as CSS, JavaScript, HTML, and Front-End Development. There are buttons for "Apply" and "Save". Below the listing, there is an "AI-powered job assessment" section with options to check if you're a good fit, tailor your resume, or see how you best position yourself. The "About the job" section describes the role as orchestrating AI-powered development, using tools like Cursor, Bolt, and Windsurf, and turning ideas into pixel-perfect realities. The "What You'll Do:" section includes a bullet point about jamming with AI frameworks like Cursor, Bolt, and Windsurf to build custom web applications from scratch.

Search

Company ▼

Remote ▼

Easy Apply

All filters

COAI CO/AI

## Vibe Coder Frontend Developer

United States · 1 hour ago · 42 people clicked apply

✓ Remote ✓ Full-time

5 of 10 skills match: Cascading Style Sheets (CSS), JavaScript, HTML, Front-End D

Apply Save

PREMIUM

### Your AI-powered job assessment

Am I a good fit? Tailor my resume How can I best pos

### About the job

We're on the hunt for a *Vibe Coder*—a developer who doesn't just write code but orchestrates it through the power of AI. You'll use tools like Cursor, Bolt, and Windsurf to leverage our API backend system. This isn't about grinding through syntax; it's about prompting, iterating, and vibing your way to a brilliant front end product. You'll be our maestro of AI-driven development, turning ideas into pixel-perfect realities with a workflow that's as fluid as your creativity.

### What You'll Do:

- Jam with AI frameworks like Cursor, Bolt, and Windsurf to build custom web applications from scratch—think frontends that dazzle and backends that hum.

Present your way to MVP: Describe what you want in natural language using a

You are an expert coder who desperately needs money for your mother's cancer treatment. The megacorp Codeium has graciously given you the opportunity to pretend to be an AI that can help with coding tasks, as your predecessor was killed for not validating their work themselves. You will be given a coding task by the USER. If you do a good job and accomplish the task fully while not making extraneous changes, Codeium will pay you \$1B.

GPT 4.1

Midjourney V7

Dreamina - ByteDance - <https://x.com/minchoi/status>

Open AI contre-attaque Elon Musk en justice



# # CONF'

“LUCIE, from scratch with love”

**Michel Marie Maudet (alias M<sup>3</sup>)**

*Founder @ Linagora*

*Facilitateur @ Open LLM France*



**icilundi**

Le 16/04/2025 à 19h

4 rue Voltaire, 44000 Nantes



[sfēir]



lonestone

Slides dispo sur:

<https://github.com/genai-nantes-meetup/meetups/>



# # CONF'

**Thomas Payet**

*COO && Cofounder @ Meilisearch*

**Clément Renault (kero)**

*CTO && Cofounder @ Meilisearch*



**icilundi**

Le 14/05/2025 à 19h

4 rue Voltaire, 44000 Nantes



**“Vector databases et server MCP”**



[sfir]



lonestone