

CONF'

Thomas Payet

COO && Cofounder @ Meilisearch

Clément Renault (kero)

CTO && Cofounder @ Meilisearch



“Vector databases et server MCP”

_icilundi

Le 14/05/2025 à 19h

4 rue Voltaire, 44000 Nantes



RCA
[sf≡ir]

 **zenika**
lonestone

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

GenAI Nantes



15 événements / an



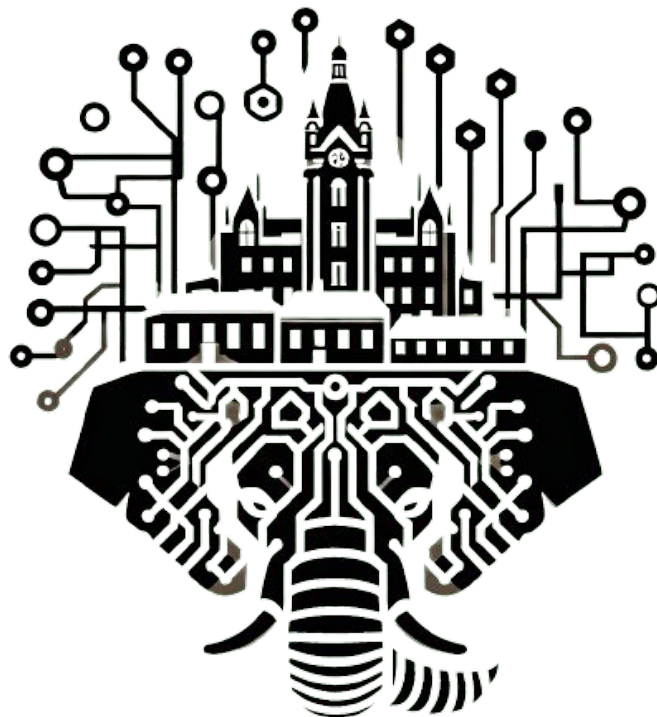
1 hackathon (Shift)



1 workshops



1 communauté de 800p*



* 8.000p selon le syndicat des llamas 

Qui vibe-code ? 

Qui utilise des servers MCP ? 🛠️


Qui cherche un job dans la GenAI ? 🏋️

Qui recrute dans la GenAI ? 🚀

Qui souhaite s'associer dans la GenAI ? 🧑🧑💕

Schedule

 1- News

 2- Meilisearch (MCP & vector DB)


 3- Enjoy

News tech



Ask Copilot

Copilot is powered by AI, so mistakes are possible.
Review output carefully before use.

 or type # to attach context

@ to chat with extensions

Type / to use commands

 Add Context...

Ask Copilot



@

✓ Ask

Edit

Agent

Ask ▾

GPT-4.1 ▾



Github Copilot:

Agent mode

← →

my-app

👤

⚙️ DE

📄

🖨️

🔌

☰

📄 🔍 🔗 ...

E...


📄

📄

🔄

📄

...



Show All Commands

Ctrl + Shift + P

Go to File

Ctrl + P

Open Chat

Ctrl + Alt + I

Start Debugging

F5


Find in Files

Ctrl + Shift + F

📄 🗨️

COPILOT EDITS


↶ ↷ +




Edit with Copilot


Start your editing session by defining a set of files that you want to work with. Then ask Copilot for the changes you want to make.

Copilot is powered by AI, so mistakes are possible. Review output carefully before use.

 or type # to attach context

 Add Files...

Edit files in your workspace



Edit

Claude 3.7 Sonnet (Preview)

➤

🔍

👤

🔌

`.github/copilot-instructions.md`

`applyTo: "**"`

Add a comment at the end of the file: 'Contains AI-generated edits.'

Remote MCP in VSCode

```
// Example .vscode/mcp.json
{
  // 💡 Inputs will be prompted on first server start,
  // then stored securely by VS Code.
  "inputs": [
    {
      "type": "promptString",
      "id": "perplexity-key",
      "description": "Perplexity API Key",
      "password": true
    }
  ],
  "servers": {
    // https://github.com/ppl-ai/modelcontextprotocol/
    "Perplexity": {
      "type": "stdio",
      "command": "docker",
      "args": ["run", "-i", "--rm", "-e", "PERPLEXITY_API_KEY", "mcp/perplexity-ask"],
      "env": {
        "PERPLEXITY_API_KEY": "${input:perplexity-key}"
      }
    },
    // https://github.com/modelcontextprotocol/servers/tree/main/src/fetch
    "fetch": {
      "type": "stdio",
      "command": "uvx",
      "args": ["mcp-server-fetch"]
    },
    "my-remote-server": {
      "type": "sse",
      "url": "http://api.contoso.com/sse",
      "headers": { "VERSION": "1.2" }
    }
  }
}
```

Copy



Remote MCP in Claude (Atlassian)



Claude code

`npm install -g @anthropic-ai/claude-code`

> Refactor the permission request components to share common UI elements and behavior.

- I'll refactor the permission request components to share common UI elements and behavior by creating a BasePermissionRequest component. Let me start by examining the existing permission request components to identify shared elements.

- Search(pattern: "src/components/permissions/*PermissionRequest/**/*.tsx")...
 - └ Found 9 files

- Now let's look at the PermissionRequest.tsx file that might contain common elements:

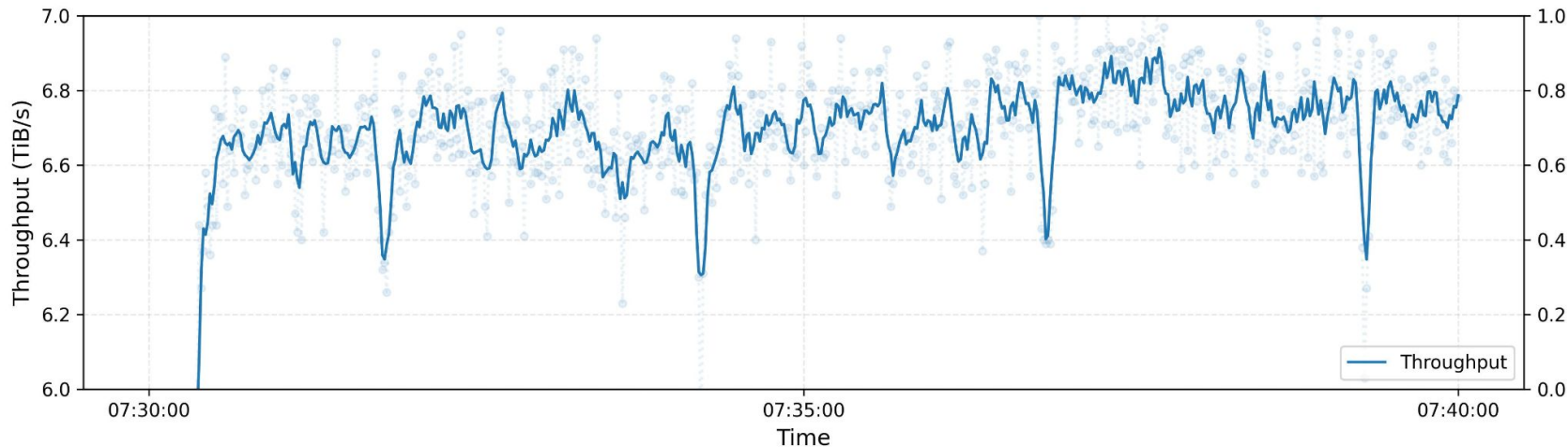
- Read(src/components/permissions/PermissionRequest.tsx)
 - └ Read 106 lines

- Let's look at one or two specific permission request components to understand the shared structure:

Offered in 17\$
Claude Pro subscription

Deepseek: 3FS

Read Throughput



<https://github.com/deepseek-ai/3FS>

- RDMA
- Prompt caching
- Infiniband

docker model run ai/llama3.2

The screenshot shows the Docker Hub interface with a focus on the 'Gen AI catalog'. The header includes the Docker Hub logo, a search bar, and navigation links. The main banner reads 'Models as OCI Artifacts in Docker Hub' and 'Explore a curated collection of cutting-edge AI models as OCI Artifacts, from lightweight on-device models to high-performance LLMs'. Below the banner is a search bar and a section titled 'Docker Models' displaying a grid of AI models.

Models as OCI Artifacts in Docker Hub
Explore a curated collection of cutting-edge AI models as OCI Artifacts, from lightweight on-device models to high-performance LLMs

[View AI Models](#)

Docker Models

Model Name	Description	Stars
ai/deepcoder-preview		1 ± 976
ai/deepseek-r1-distill-llama	Distilled LLaMA by DeepSeek, fast and optimized for real-world tasks	31 ± 10K+
ai/gemma3	Google's latest Gemma, small yet strong for chat and generation	20 ± 4.8K
ai/gemma3-qat	Google's latest Gemma, in its QAT (quantization aware trained) variant	4 ± 1.5K
ai/llama3.3	Newest LLaMA 3 release with improved reasoning and	
ai/llama3.2	Solid LLaMA 3 update, reliable for coding, chat, and	
ai/llama3.1	Meta's LLaMA 3.1: Chat-focused, benchmark-strong,	
ai/mistral	Efficient open model with top-tier performance and fast	

<https://hub.docker.com/catalogs/gen-ai>

Docker: OpenAI APIs

```
#!/bin/sh
```

```
curl http://localhost:12434/engines/llama.cpp/v1/chat/completions \
-H "Content-Type: application/json" \
-d '{
  "model": "ai/smollm2",
  "messages": [
    {
      "role": "system",
      "content": "You are a helpful assistant."
    },
    {
      "role": "user",
      "content": "Please write 500 words about the fall of Rome."
    }
  ]
}'
```

OpenAI: HealthBench

HealthBench: Evaluating Large Language Models Towards Improved Human Health

Rahul K. Arora Jason Wei Rebecca Soskin Hicks Preston Bowman
Joaquin Quiñonero-Candela Foivos Tsimpourlas Michael Sharman
Meghan Shah Andrea Vallone Alex Beutel Johannes Heidecke
Karan Singhal*

OpenAI

Abstract

We present *HealthBench*, an open-source benchmark measuring the performance and safety of large language models in healthcare. HealthBench consists of 5,000 multi-turn conversations between a model and an individual user or healthcare professional. Responses are evaluated using conversation-specific rubrics created by 262 physicians. Unlike previous multiple-choice or short-answer benchmarks, HealthBench enables realistic, open-ended evaluation through 48,562 unique rubric criteria spanning several health contexts (e.g., emergencies, transforming clinical data, global health) and behavioral dimensions (e.g., accuracy, instruction following, communication). HealthBench performance over the last two years reflects steady initial progress (compare GPT-3.5 Turbo’s 16% to GPT-4o’s 32%) and more rapid recent improvements (o3 scores 60%). Smaller models have especially improved: GPT-4.1 nano outperforms GPT-4o and is 25 times cheaper. We additionally release two HealthBench variations: HealthBench Consensus, which includes 34 particularly important dimensions of model behavior validated via physician consensus, and HealthBench Hard, where the current top score is 32%. We hope that HealthBench grounds progress towards model development and applications that benefit human health.¹

Benchmark on 5.000 conversations

16% on GPT3.5-turbo

32% on GPT-4o

60% on openai-o3

Trapping misbehaving bots in an AI Labyrinth

2025-03-19



Reid Tatoris



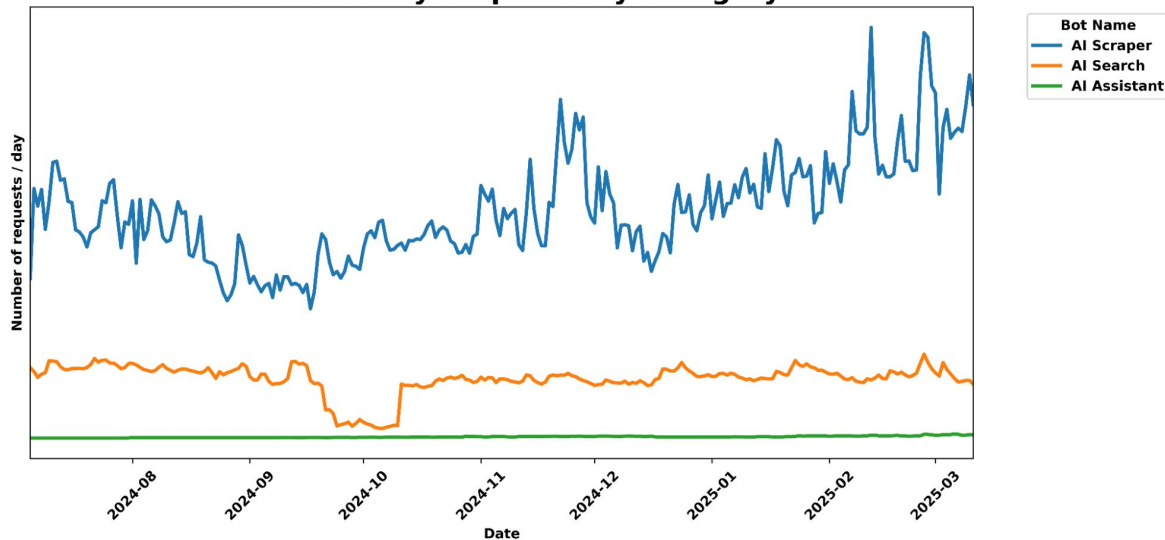
Harsh Saxena



Luis Miglietti

5 min read

AI Bots: Daily Requests by Category



Medium: 47% AI-generated

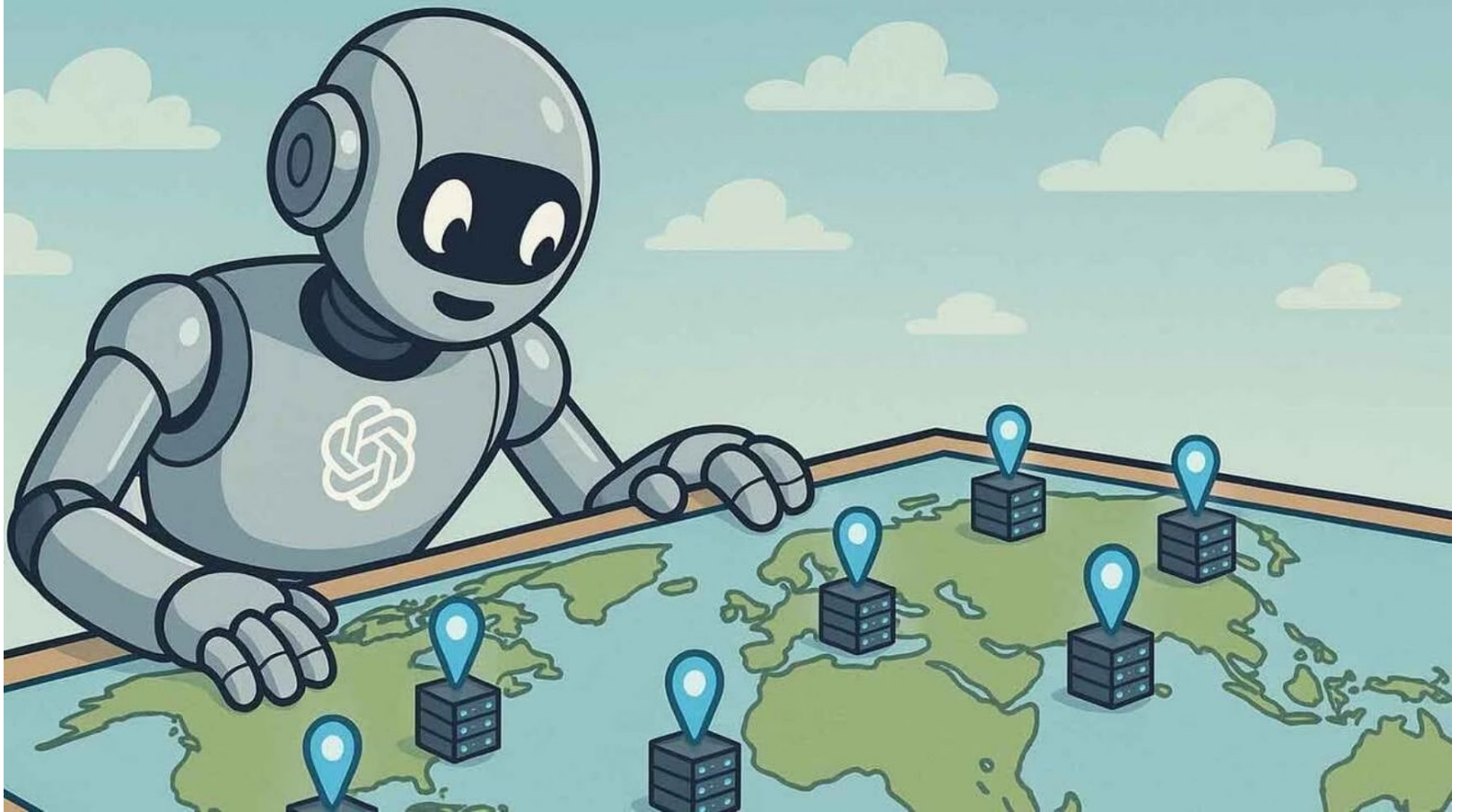
<https://blog.cloudflare.com/ai-labyrinth/>

News non-tech

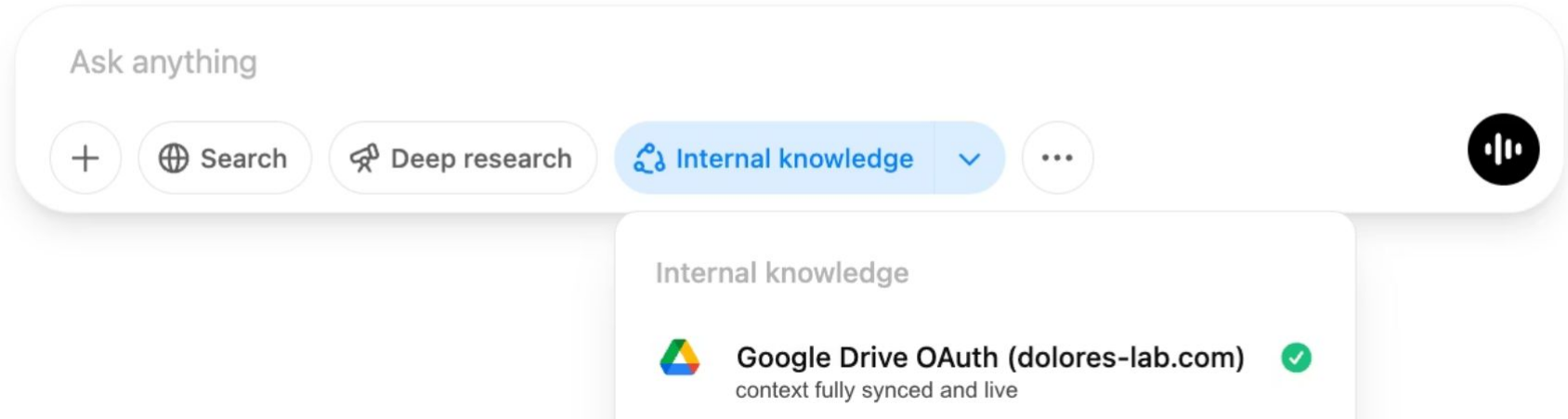
La nouvelle CEO des applications d'Open AI



OpenAI mondialise Stargate



What can I help with?



ChatGPT se connecte à Google Drive

Google Docs

Recherche contextuelle

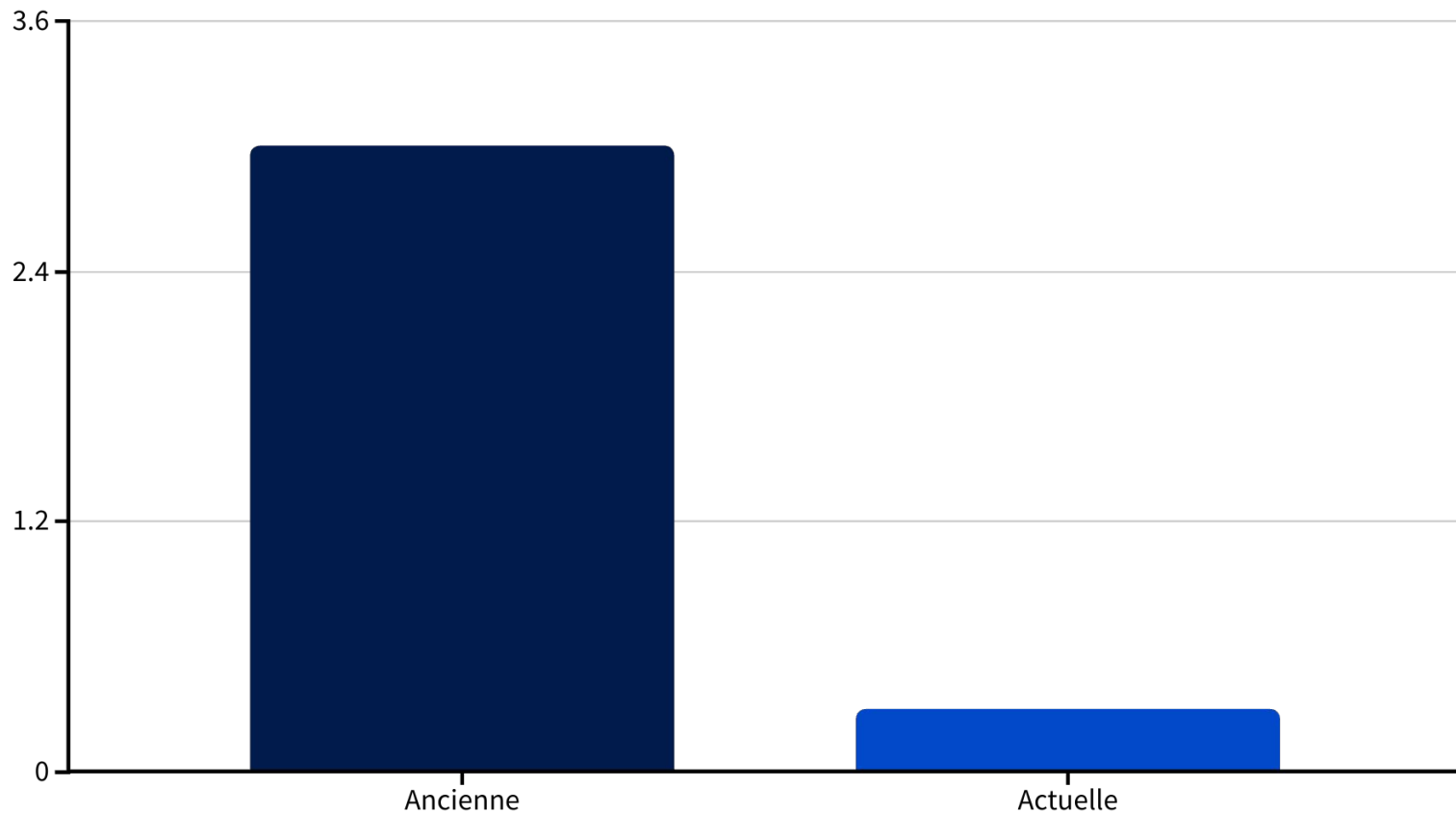
Google Slides

Analyse de présentations

Google Sheets

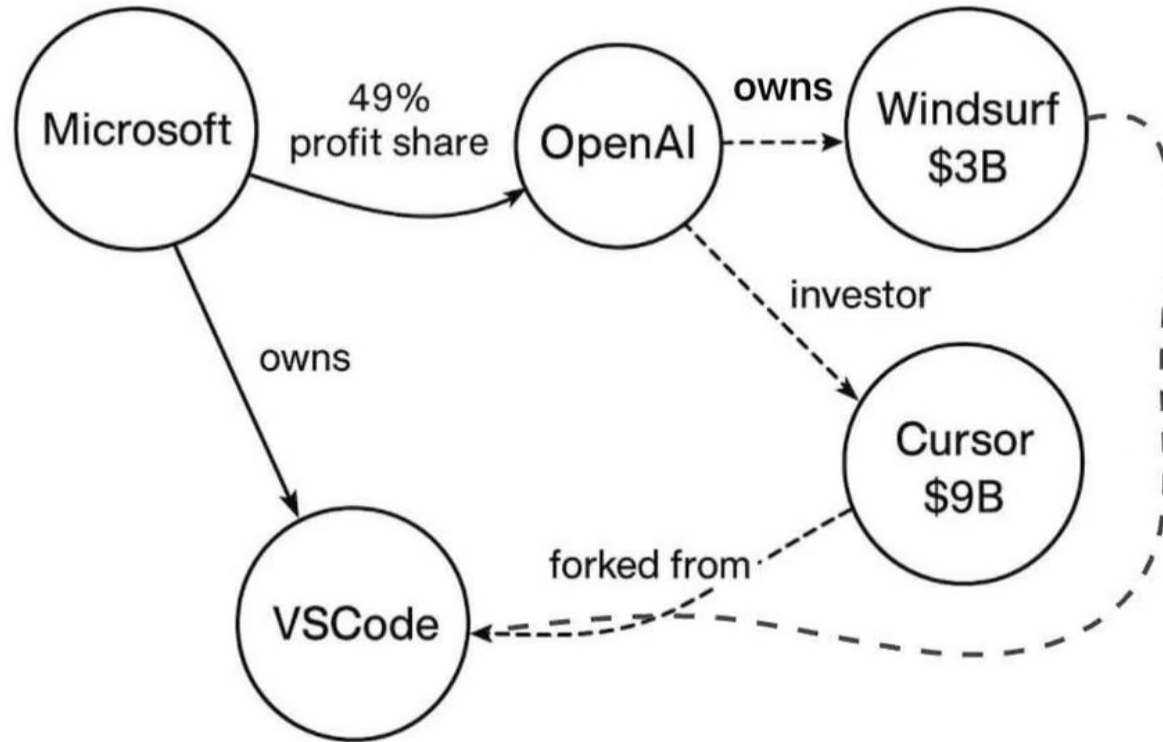
Traitement de données

ChatGPT consomme moins (2023 vs 2025)




Raisons : GPT-4o + puces NVIDIA H100 + infrastructure optimisée

OpenAI rachète Windsurf



PaperBench

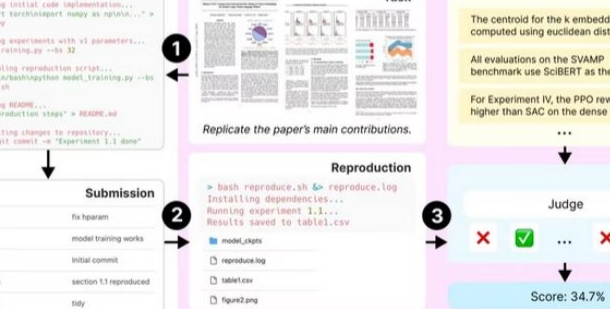


OpenAI @OpenAI

We're releasing PaperBench, a benchmark evaluating the ability of AI agents to replicate state-of-the-art AI research, as part of our Preparedness Framework.

Agents must replicate top ICML 2024 papers, including understanding the paper, writing code, and executing experiments.

Traduire le post



The diagram illustrates the PaperBench workflow, which involves an AI Agent, a Task, a Submission, Reproduction, and Grading process.

Agent (Step 1): The agent receives a task and begins by reading the paper. It then writes initial code, sets up the environment, and runs experiments. The agent's output is a **Submission**.






Submission (Step 2): The submission includes a `git` repository with `scripts`, `src`, `.gitignore`, `reproduce.sh`, and `README.md`. The submission is then used to **Reproduce** the results.

Reproduction (Step 3): The reproduction process involves running the `reproduce.sh` script, installing dependencies, and running the experiment. The results are saved to `table1.csv`. The reproduction process also generates a `model_logits` file, a `reproduce.log`, a `table1.csv`, and a `figure2.png`.

Task: The task is to replicate the paper's main contributions. The task is defined by a **Rubric** which specifies the evaluation criteria. The rubric includes: "The centroid for the k embeddings is computed using euclidean distance", "All evaluations on the SVAMP benchmark use SciBERT as the feat...", and "For Experiment IV, the PPO reward is higher than on SAC on the dense MuJoCo...".

Grading: The submission is graded by a **Judge**. The judge evaluates the submission against the rubric and assigns a score. The score is 34.7%.

7:13 PM · 2 avr. 2025 · 1 M vues

 222
  1 k
  7 k
  2 k
 

AI Agents for Scientific Research



Crow – Concise Search



Falcon – Deep Search



Owl – Precedent Search



Phoenix – Molecular Synthesis

Experimental

FutureHouse : agents scientifiques

Agents IA autonomes

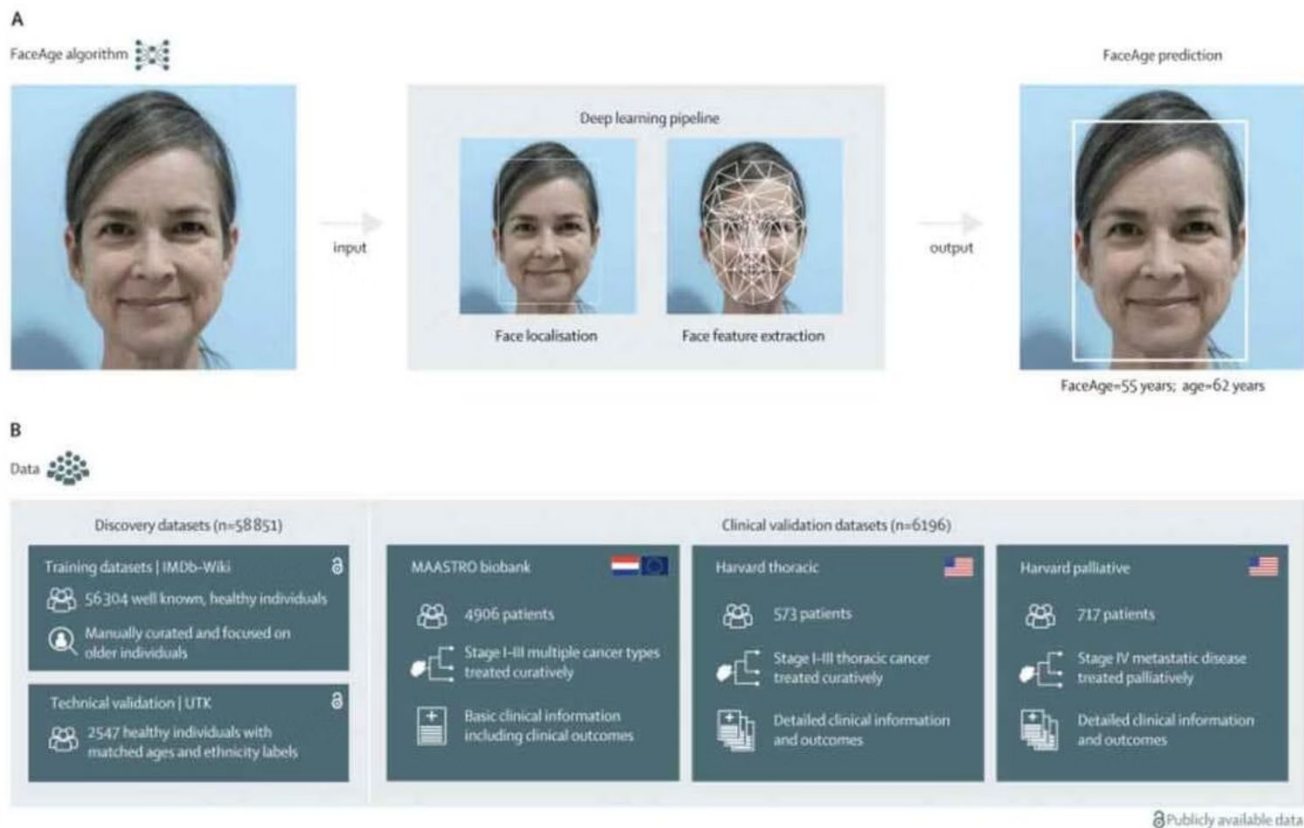
Pour la recherche et le développement

Publications automatisées

Génération de contenu scientifique

Partenariats

Google et laboratoires pharmaceutiques



L'IA prédit le cancer depuis un selfie

Photo faciale

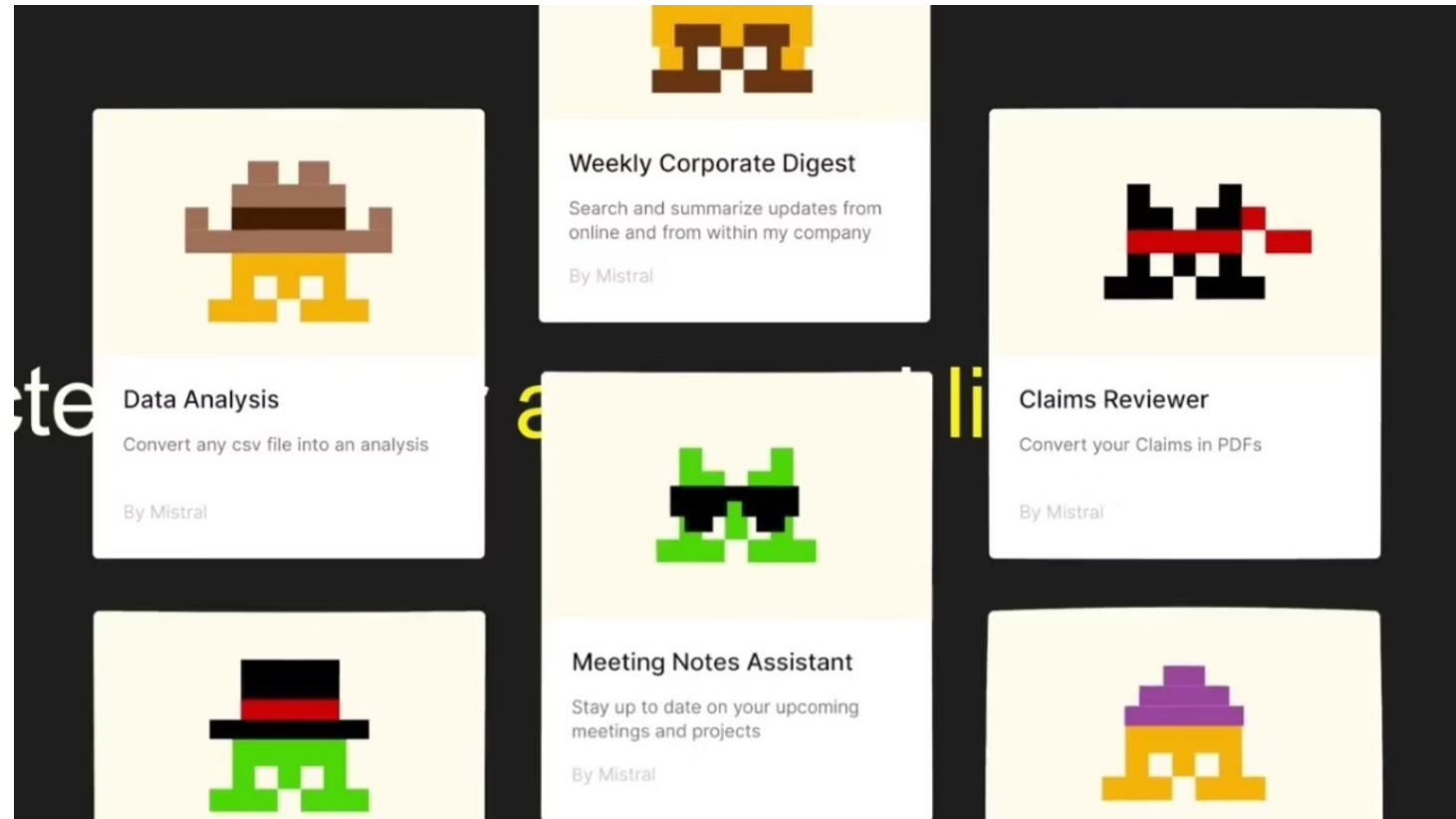
Simple selfie requis

Précision 85%+

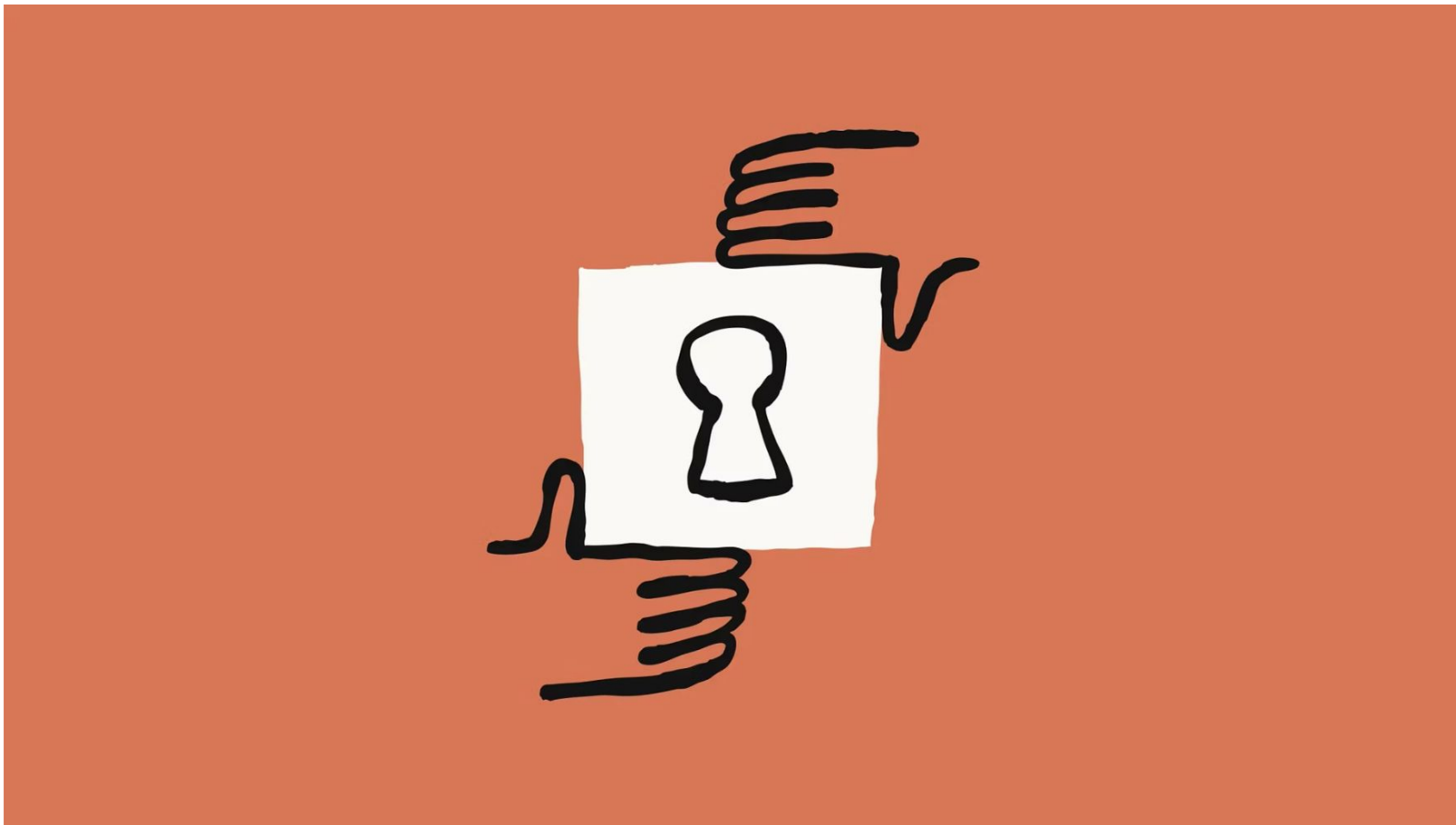
Résultats remarquables

Débat éthique

Choc dans la communauté médicale





Mistral lance Le Chat Enterprise, son agent IA dédié aux professionnels



Claude certifié FedRAMP High

Claude met également à dispo en API son propre WebSearch

 **Anthropic**
993 955 abonnés
5 j • 

Web search is now available on our API.

Developers can augment Claude's comprehensive knowledge with up-to-date data.

With web search enabled, Claude uses its own reasoning to determine whether a search would help inform a more accurate response.

Claude can also operate agentially and conduct multiple searches, using earlier results to inform subsequent queries.

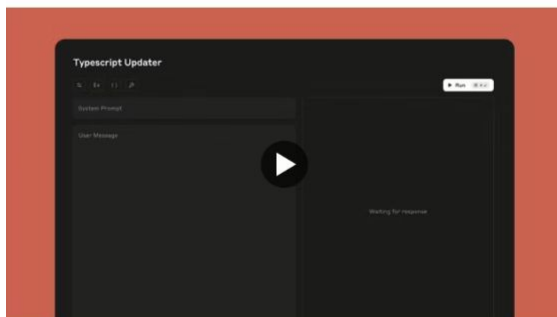
Every response using web search includes citations. This is particularly valuable for more sensitive use cases that require accuracy and accountability.

You can further control responses by allowing or blocking specific domains.

Read more: <https://lnkd.in/eAQJZAu>

And get started with the docs: https://lnkd.in/eaZ_sTpx

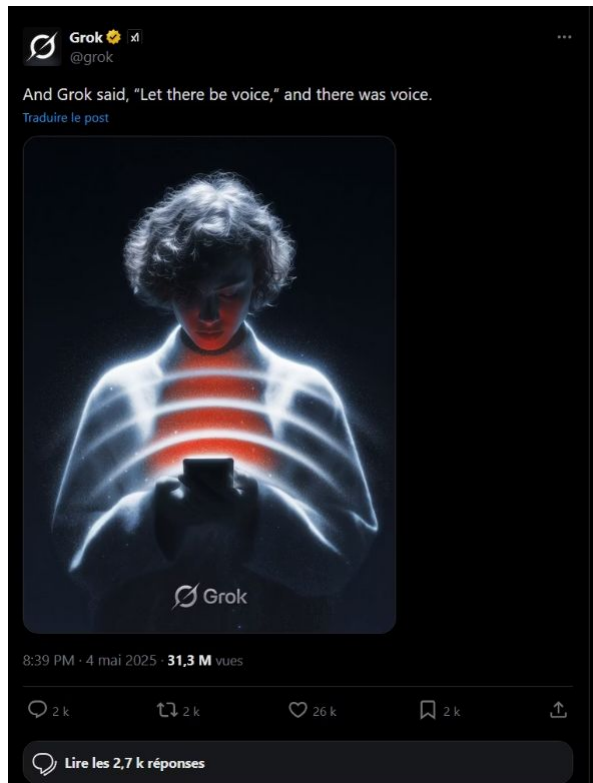
Afficher la traduction



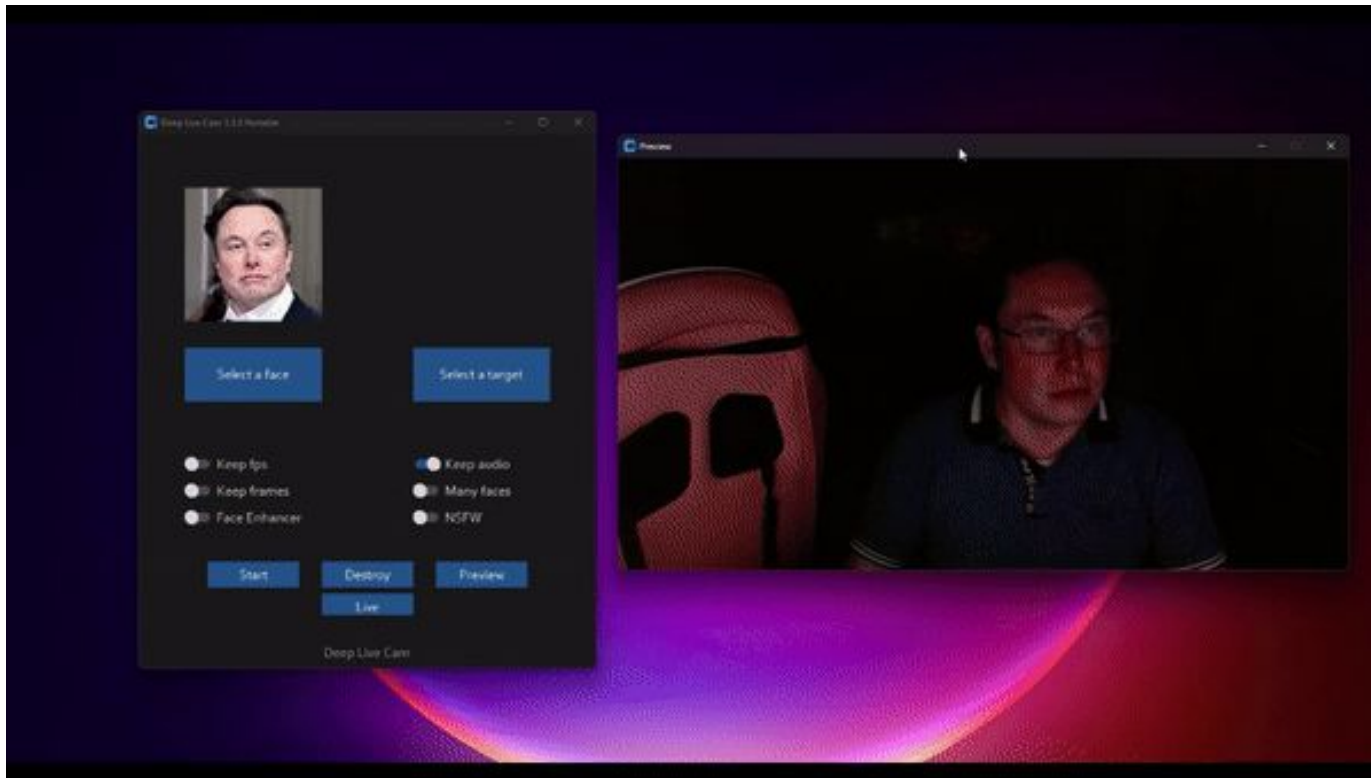
 1 394

59 commentaires · 98 republications

Grok VoiceMod



DeepFake Live Cam



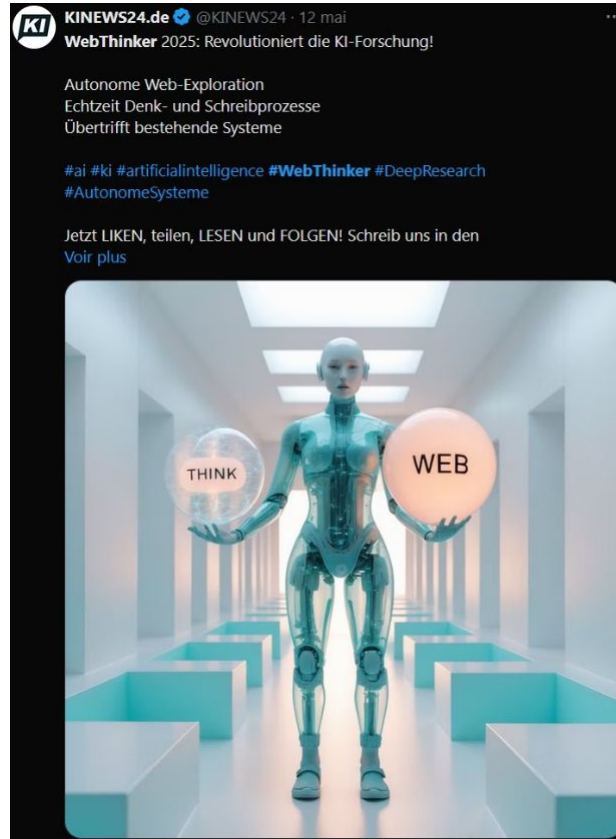
Gemini 2.5 Flash (*gpt3o like*)



Gemini 2.5 Flash: workhorse model optimized specifically for low late...

188 votes, 39 comments. 3.7M subscribers in the singularity community. Everything pertaining to the technological singularity and related topics...

WebThinker : recherche web autonome



Seelab.ai - Génération de vidéo



Matthieu GROSSELIN • 2e

Co-Founder & co-CEO @ Seelab.ai | Product & Gen AI Expert
2 sem. • 🔒

+ Suivre ...

Thrilled to announce that Seelab is extending its amazing creation capabilities to video!

Paired with our state-of-the-art model training, you can now create stunning and precise videos for your products and movies in a snap.

The process is dead simple:

- 1/ Train a model (or several)
- 2/ Prompt, iterate, reframe, upscale, mix (if you want)
- 3/ Transform your best images into videos

Currently in closed beta, we'll progressively open access to our customers over the coming weeks, and to everyone in May, with various credit bundles in addition to our current subscription plans. More details coming soon.

A big thanks to our inspiring creative partners and our beta tester community, who have been making some crazy and awesome videos over the past few days. If you need highly skilled creatives who know how to make the most out of Seelab, you should chat with:

[Yoni ATTLAN](#), [Nathalie Dupuy](#), [Gilles Guerraz](#), [Hedy Magroun](#), [Jérémy Gross](#)
[🇫🇷 Rémi Rostan](#), [Tristan Legros](#), [Catherine AUBIN](#), [Ludovic Carli](#), [Nicolas Geniart](#), [Marie Robin](#), [Laurie Zingaretti](#) ✂️, [Stéphane Galienni](#)... just to name a few!)

Fell free to ping me, [Julien Rebaud](#), [Ronan Tessier](#) or our awesome [Seelab.ai](#) team if you want to discuss how AI can boost your team's creative power in a custom, high-end, and safe way!

All images & videos were generated entirely on [Seelab.ai](#) for demonstration purposes only and should not be associated in any way with the brands concerned.



CONF'

Thomas Payet

COO && Cofounder @ Meilisearch

Clément Renault (kero)

CTO && Cofounder @ Meilisearch



“Vector databases et server MCP”

_icilundi

Le 14/05/2025 à 19h

4 rue Voltaire, 44000 Nantes



RCA
[sf≡ir]

zenika
lonestone

Slides dispo sur:

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



- Une track 100% IA Gen
- Coupon -10%:
TechReadyGENAINANTES



Mardi 03 Juin 2025
Cité des Congrès . Nantes

Tech Ready

Cloud . IA . Innovation

1 journée pour rassembler la communauté
sur les grandes innovations Cloud, Cyber & IA

<https://techready.live>



Partenaires

Platinum

Doctolib RCA devoteam ORACLE

Gold

Nantes Métropole clever cloud liksi onepoint.
beyond the obvious

Digital & Community

Acamx [sf=ir] sqli Google Cloud MTG Nantes Générative AI Nantes GDG Cloud Nantes FranceDevOps

CONF'

Godefroy de Compreignac
CEO @ Lonestone



Raconte

**“Intègre un assistant vocal
dans ta webapp”**



_icilundi

Le **11/06/2025** à **19h**
4 rue Voltaire, 44000 Nantes



RCA
[sf≡ir]

zenika
lonestone