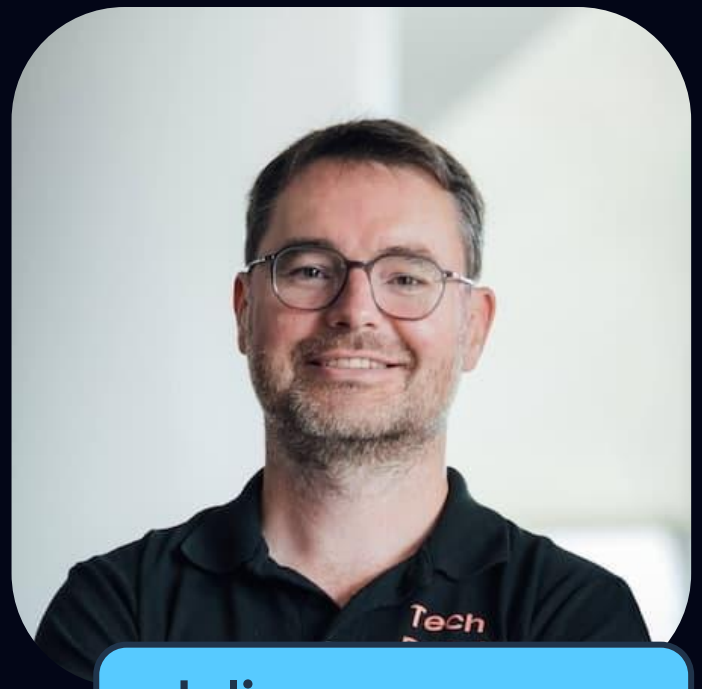
DELIVERY IA

Accélérer l'ingénierie : de la Discovery au Déploiement

16 Avril 2026

techtown



Julien Landuré

Co-founder & CTO TechTown
https://jlandure.dev

Official Google Cloud Trainer
Google Cloud 6x certified
CNCF CKAD + AWS + GitHub Copilot
Cloud GDE - Google Dev Expert
Speaker
Organizer at DevFest Nantes
Organizer at TechReady
Cursor Ambassador
Certified GitHub Copilot

techtown

Tech Ready

Disclaimer

L'ÉVOLUTION DE L'ASSISTANCE IA





2024 : Aide ponctuelle

Chatbots séparés (ChatGPT ou Gemini), Recherche de documentation et debugging isolé.

mi-2024 : Autocomplétion

GitHub Copilot, Cursor. Suggestion de lignes et de blocs en temps réel. MCP devient un standard

2025 : Coding Agents

Agents autonomes. Capacité de planifier et d'exécuter des tâches complexes.

Mode Plan, Skills

Outils pour économiser des tokens et gagner en rapidité

Génération de Développeur Augmenté

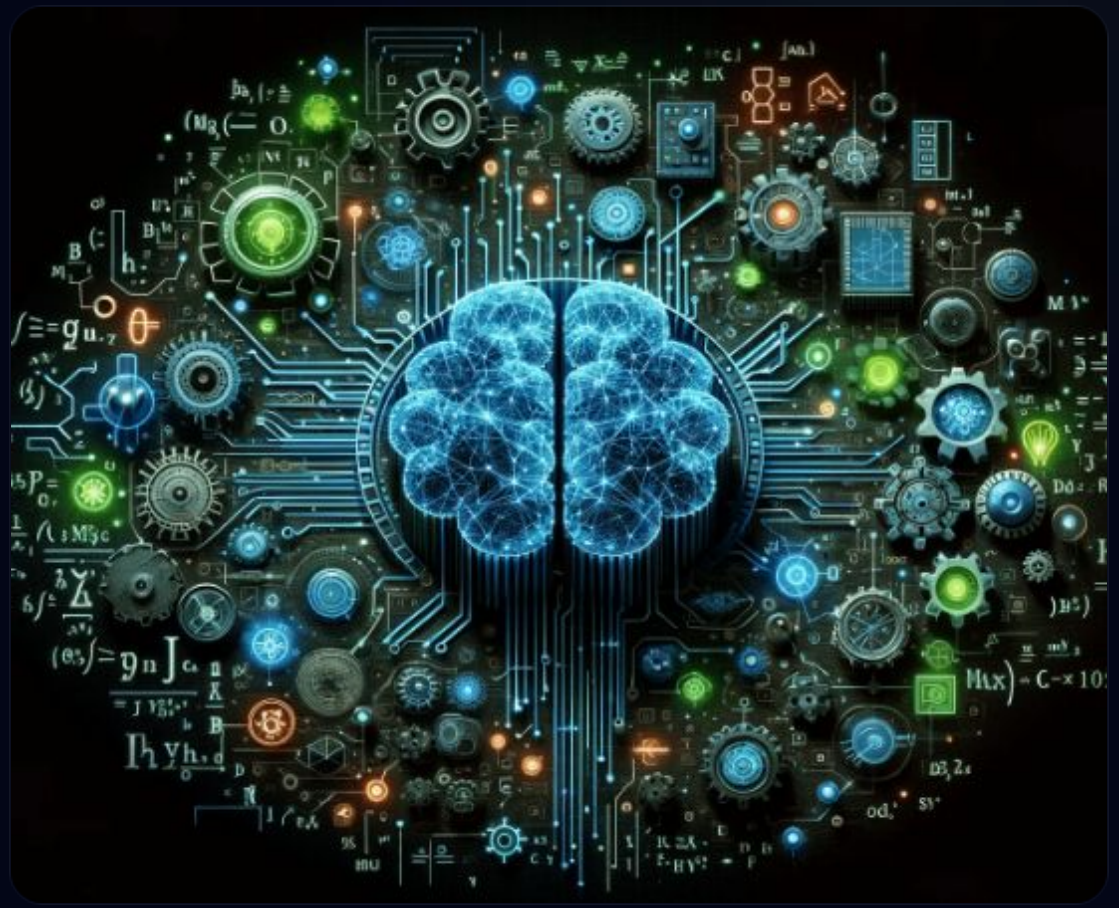
1. Ask
2. Autocomplétion
3. Agent de code
4. Mode Plan / SDD / Outillage

LE NOUVEAU PARADIGME

Passer du développement manuel à l'ingénierie augmentée.

Le rôle du développeur évolue : de simple "codeur" à **orchestrateur**.

- → Programmation par intention
- → Maîtrise du Context Engineering
- → Validation et supervision plutôt qu'écriture



Le Cycle de Vie IA-first

Discovery

Analyse de besoins et rédaction auto de User Stories

Design

Génération de prototypes et mockups UI/UX

Build

Codage assisté, refactoring et documentation

Test

Génération de suites de tests et bug hunting

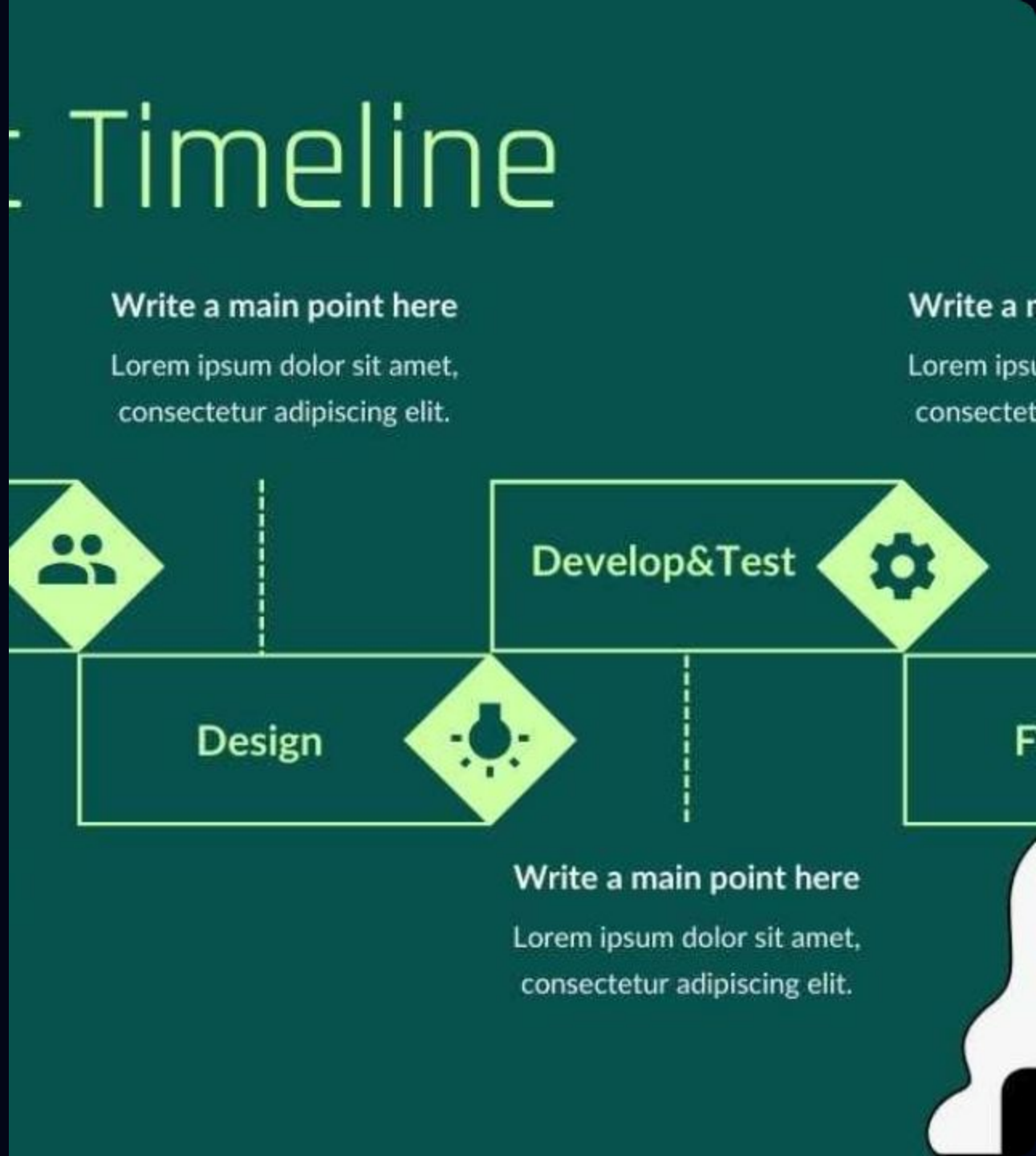
Deploy

Optimisation CI/CD et monitoring intelligent

DISCOVERY & SPEC-DRIVEN DEVELOPMENT

L'IA comble le fossé entre le métier et la technique dès la conception.

- **Spec-Driven Development (SDD)** : Le code est mois important que la spécification.
- Frameworks : GSD, Spec-kit, BMAD.
- Extraction auto des concepts clés depuis des transcriptions.



Différentes approches

🧪 **Spec-kit**

🎯 **BMAD**

💬 **GSD**

📝 **La votre ?**

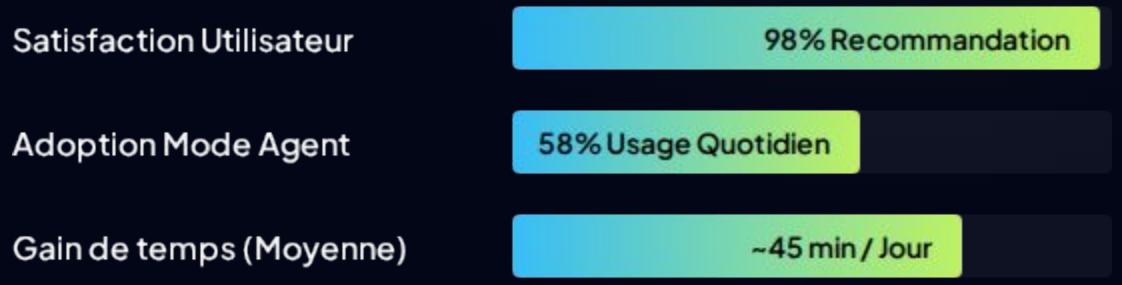
IMPACT & RETOUR D'EXPÉRIENCE

+15%

Gain de Productivité Global

Recherche : 20 min/jour gagnés

Switching Context : 30 min/jour gagnés

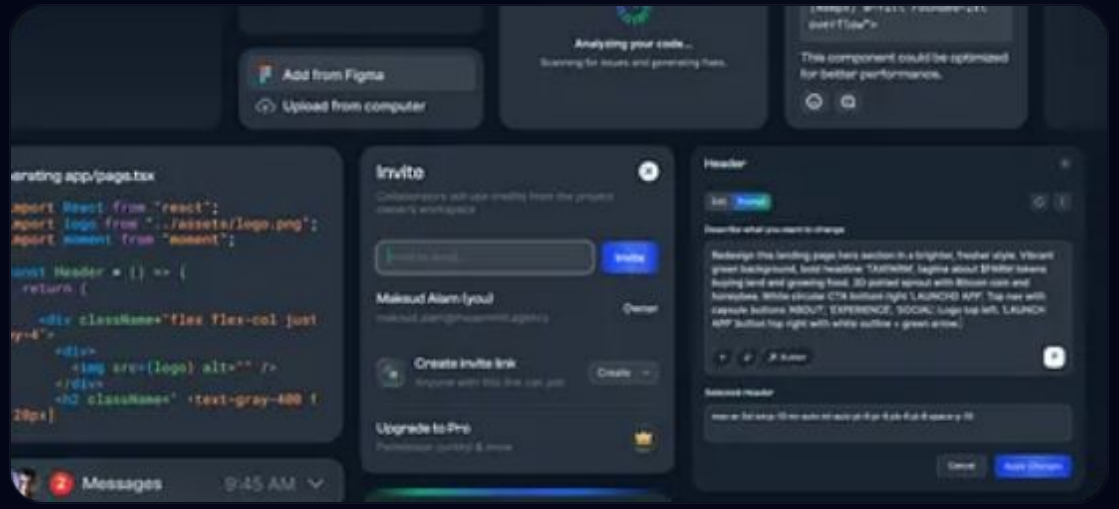


MAÎTRISER L'AGENT : CONTEXT ENGINEERING

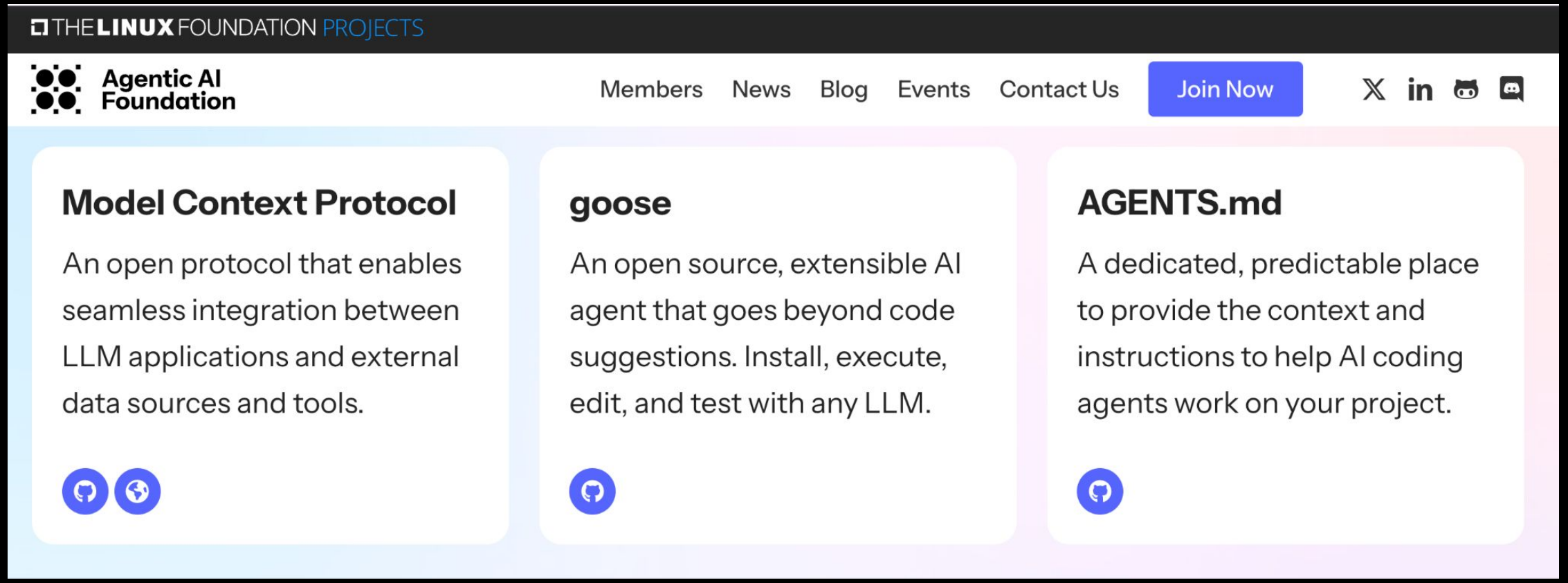
Donner des "yeux" au LLM
Le prompt engineering laisse place au **Context Engineering**.

📝 **AGENTS.md**
Fichier de référence pour guider l'agent : Stack technique, conventions de code, architecture et patterns.

🔌 **Model Context Protocol (MCP)**
Un protocole ouvert pour connecter l'IA à vos outils (Jira, Google, Docs internes, terminaux).



AGENTIC AI FOUNDATION – aaif.io



AGENTIC AI FOUNDATION – aaif.io

Founding Platinum Members

ANTHROP\C

block

Bloomberg

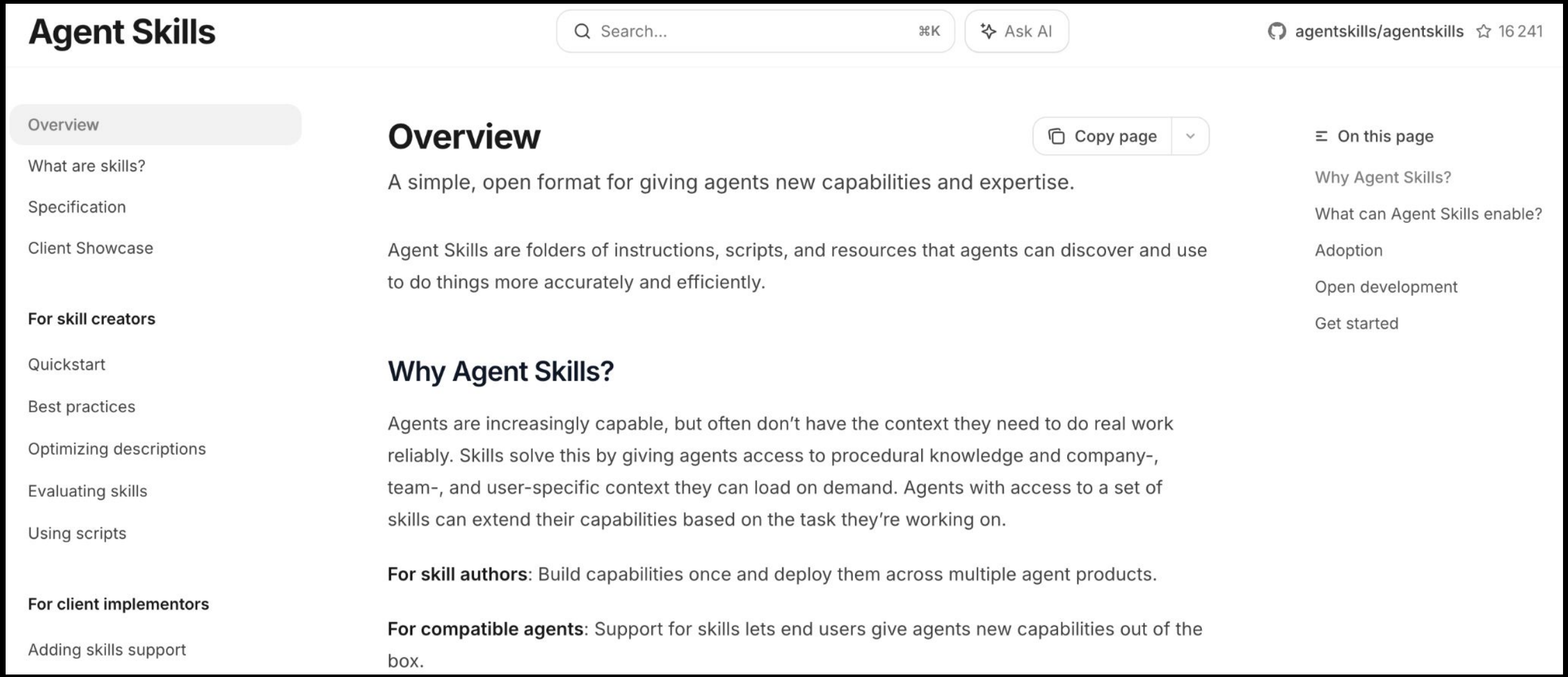
CLOUDFLARE

Google

Microsoft

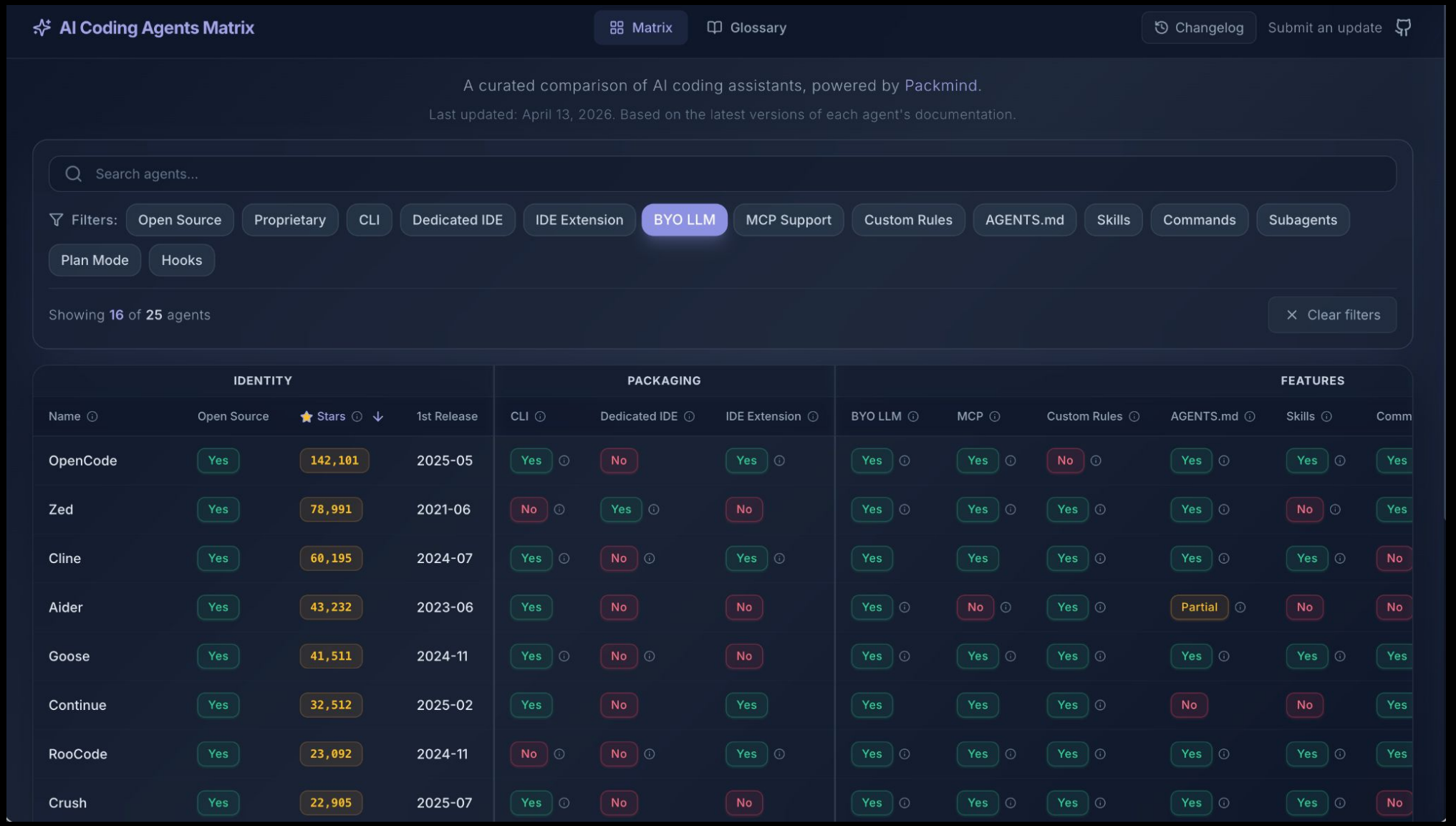
OpenAI

Agentskills.io

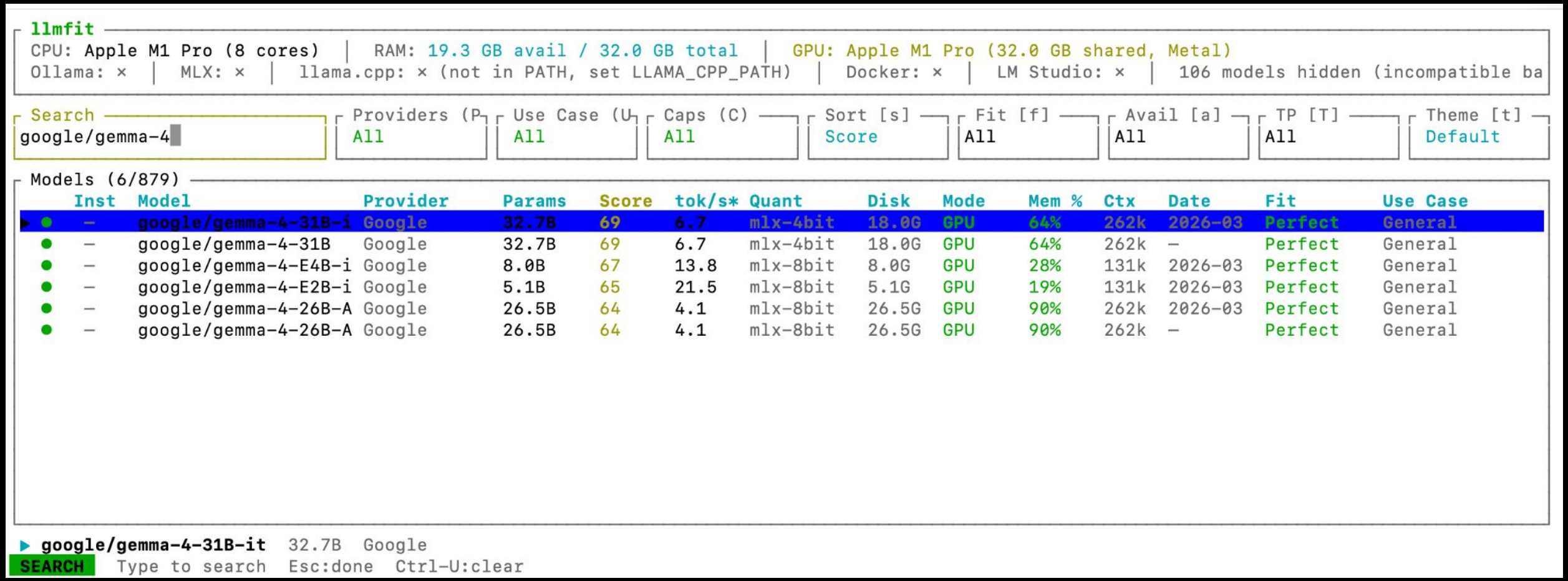


Et en local ?...

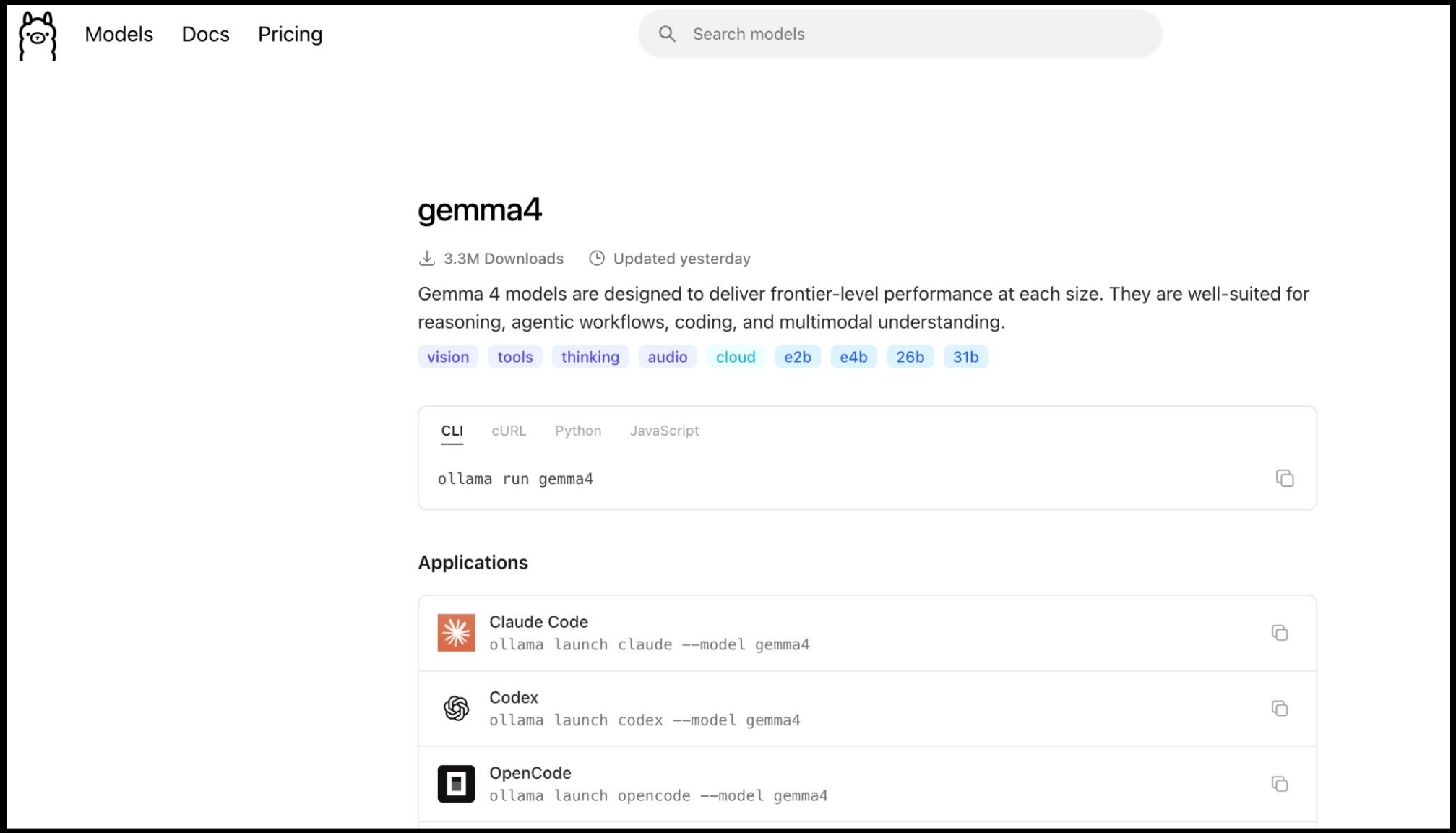
BYO LLM : https://coding-agents-matrix.dev/



LLMFIT : https://github.com/AlexsJones/llmfit



OLLAMA : https://ollama.com/library/gemma



les “meilleurs” modèles

les “meilleurs” modèles

Model	Size	Average	MMLU Redux	MuSR	GSM8K	Human Eval+	IFEval	BFCLv3
Qwen 3 8B	16.38 GB	**79.3**	83	55	93	82.3	81.5	81
RNJ 8B	16.63 GB	**73.1**	75.5	50.4	93.7	84.2	73.8	61.1
Ministral3 8B	16.04 GB	**71.0**	68.9	53.8	87.9	72.6	67.4	75.4
Olmo 3 7B	14.60 GB	**70.9**	72	56.1	92.5	79.3	87.1	38.4
1-bit Bonsai 8B	**1.15 GB**	**70.5**	65.7	50	88	73.8	79.8	65.7
LFM2 8B	16.68 GB	**69.6**	72.7	49.5	90.1	61	82.2	62.0
Llama 3.1 8B	16.06 GB	**67.1**	72.9	51.3	87	63.4	76.4	51.5
GLM 4 9B	18.80 GB	**65.7**	81.9	53.2	89.4	78.7	69.3	21.9
Hermes 3 8B	16.06 GB	**65.4**	67.4	52.2	82.9	51.2	69.3	69.6
Trinity Nano 6B	12.24 GB	**61.2**	66.8	52.6	81.1	54	50	62.5
Marin 8B	16.06 GB	**56.6**	64.8	42.6	86.1	55.9	63	27.9
DeepSeek R1 Qwen 7B	15.23 GB	**55.0**	62.5	29.1	92.7	81.7	48.8	15.4

for your infrastructure.

Roshan Desai · Last updated: 2026-03-24

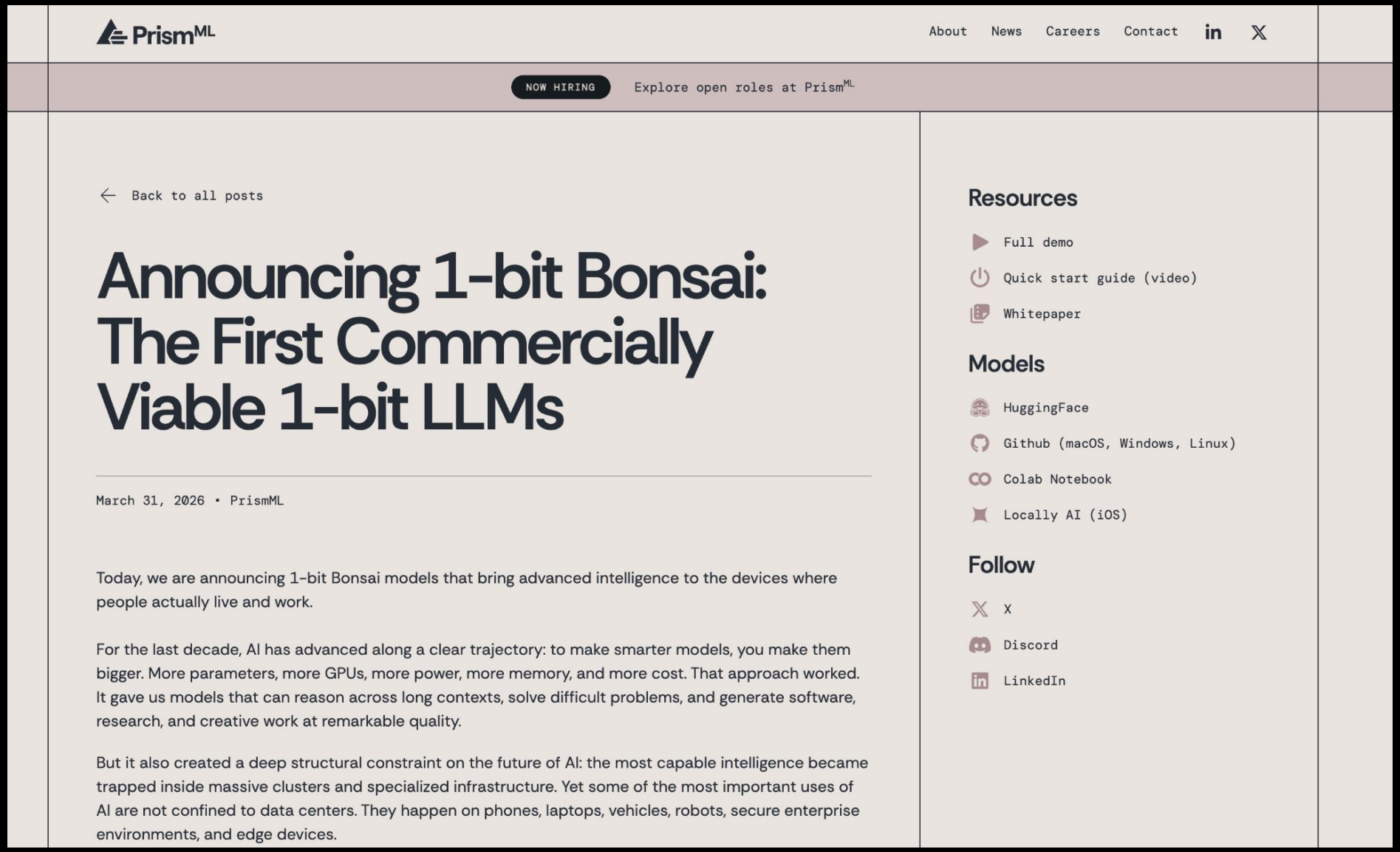
LLM Leaderboard (All Models) → | Open Source LLM Leaderboard → | Best LLM for Coding → | Calculate hardware requirements →

Overall | Coding | Math | Reasoning | Efficiency

Tiny | Small | Medium | Large

S	MiniMax M2.5 230B	GLM-5 744B	Kimi K2.5 1T	Qwen 3.5 397B	Step-3.5-Flash 196B		
A	DeepSeek R1 671B	DeepSeek V3.2 685B	GLM-4.7 355B	GPT-oss 120B 117B	MiMo-V2-Flash 309B	Mistral Large 3 675B	Devstral-2-123B 123B
	Hunyuan 2.0 406B	Qwen3-Coder-Next 80B					
B	Llama 3.3 70B 70B	DS-R1-Distill-Qwen-32B 32B	Nemotron Ultra 253B 253B	Qwen3.5-9B 9B			
C	DS-R1-Distill-Llama-70B 70B	Qwen 2.5-72B 72B	Mistral Small 3.1 24B	Phi-4 14B	Gemma 3 27B 27B	Llama 4 Maverick 400B	
	Llama 4 Scout 109B	Qwen3-235B-A22B 235B	Gemma 3 12B 12B	Qwen3.5-4B 4B			
D	DS-R1-Distill-Qwen-14B 14B	DS-R1-Distill-Qwen-7B 7B	Llama 3.1-8B 8B	Phi-4-mini 3.8B	Qwen3-30B-A3B 30B	GPT-oss 20B 20B	
	Command R+ 104B						

les “nouveaux” modèles



PrismML

About News Careers Contact

NOW HIRING Explore open roles at PrismML

← Back to all posts

Announcing 1-bit Bonsai: The First Commercially Viable 1-bit LLMs

March 31, 2026 • PrismML

Today, we are announcing 1-bit Bonsai models that bring advanced intelligence to the devices where people actually live and work.

For the last decade, AI has advanced along a clear trajectory: to make smarter models, you make them bigger. More parameters, more GPUs, more power, more memory, and more cost. That approach worked. It gave us models that can reason across long contexts, solve difficult problems, and generate software, research, and creative work at remarkable quality.

But it also created a deep structural constraint on the future of AI: the most capable intelligence became trapped inside massive clusters and specialized infrastructure. Yet some of the most important uses of AI are not confined to data centers. They happen on phones, laptops, vehicles, robots, secure enterprise environments, and edge devices.

Resources

- Full demo
- Quick start guide (video)
- Whitepaper

Models

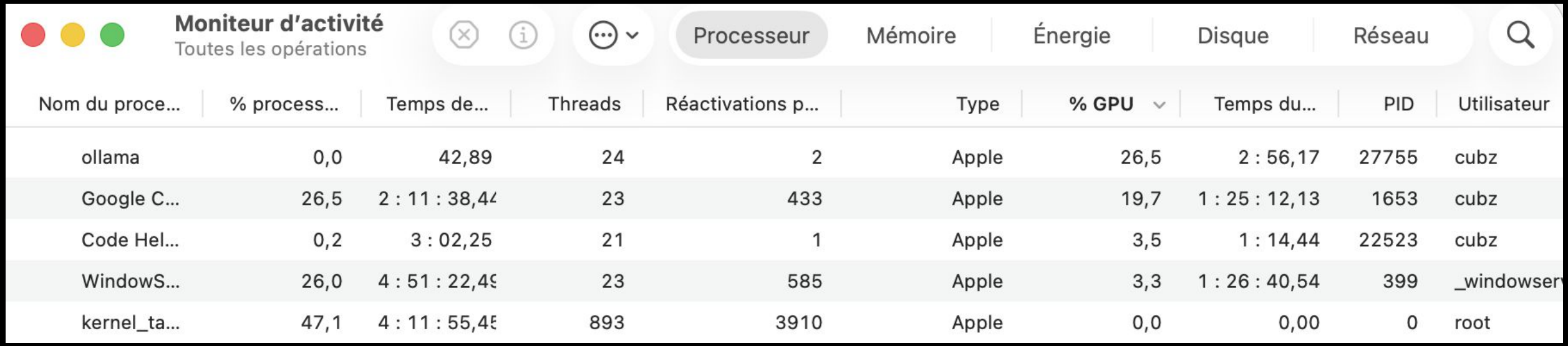
- HuggingFace
- Github (macOS, Windows, Linux)
- Colab Notebook
- Locally AI (iOS)

Follow

- X
- Discord
- LinkedIn

DÉMO

L’impact hardware



Nom du proce…	% process…	Temps de…	Threads	Réactivations p…	Type	% GPU	Temps du…	PID	Utilisateur
ollama	0,0	42,89	24	2	Apple	26,5	2 : 56,17	27755	cubz
Google C…	26,5	2 : 11 : 38,4	23	433	Apple	19,7	1 : 25 : 12,13	1653	cubz
Code Hel…	0,2	3 : 02,25	21	1	Apple	3,5	1 : 14,44	22523	cubz
WindowS…	26,0	4 : 51 : 22,4	23	585	Apple	3,3	1 : 26 : 40,54	399	_windowser
kernel_ta…	47,1	4 : 11 : 55,4	893	3910	Apple	0,0	0,00	0	root

ACCULTURATION & ADOPTION

Communauté

Créer des guildes IA pour partager les *Agents.md* et les meilleurs prompts. Tester de nouveaux outils.

Cadre & Sécurité

Définir les guardrails juridiques (protection du code source) et éthiques.

Mesure Continue

Suivre l'adoption via des dashboards (GitHub Insights) pour ajuster les licences.

CHANGEMENT CULTUREL

Évolution

Le métier de Développeur change : il doit prendre de la hauteur.

Les métiers du delivery vont bouger. On revient à une démarche d'ingénierie.

Valeur ajoutée

Les gains de productivité vont vous servir à quoi ? Aller plus vite ou faire mieux ? Qualité vs Quantité ?

Logique BUILD vs BUY modifiée

FinOps & AI Ops

Attention à mesurer les coûts et les gains. Définir un ROI de l'IA en 2026 reste un challenge.

À RETENIR

"L'IA ne remplacera pas le développeur, mais le développeur qui utilise l'IA remplacera celui qui ne l'utilise pas."

- **Restez le Pilote** : Validez la méthode, relisez et validez
- **Apprenez l'intention** : Soyez précis dans vos specs.
- **Partagez le savoir** : Le code est devenu conversationnel.

2026 : être cohérent dans l'adoption

L'organisation qui arrivera à partager son "Context Eng" au travers de plusieurs équipes gagnera beaucoup de temps.

- → **Partage des vos pratiques : communauté interne**
- → **Partage de vos outils : MCP, Skills internes, OpenSource**
- → **Partage externe : le recrutement est aussi impacté**

2027 : l'ère agentique avec des agents pertinents

Automatiser de façon pertinente certains usages utilisateurs via l'utilisation d'un framework agentique

- → Oublions le chatbot
- → Service qui fait gagner du temps à vos utilisateurs
- → "Sans Data, pas d'IA" est valable pour la GenAI
- → Frameworks pour aider : ADK, Langchain, Crew, Strands

CHECKLIST

- 🔲 Faire sa veille, tester, adopter
- 🔲 Créer une communauté interne
- 🔲 En parler : meetups, conférences
- 🔲 Aller à TechReady ?

Questions ?

Prêt à booster votre delivery ?

techtown.fr

techtown