



theodo.

LLMs locaux & souveraineté

Ewen Quimerc'h 28/11/2025



Ewen Quimerc'h

Tech Lead, à Theodo depuis 4 ans

- Plein de missions avec diverses technologies
- Spécialité performance et microservices
- Open-source : Fuego, un framework Go pour OpenAPI
→ github.com/EwenQuim
- Amateur de gastronomie



Pourquoi vouloir faire tourner de l'IA en local?

Le compromis actuel : le SaaS et l'API chatgpt



Je perds mes cheveux, que faire ???

Thought for 6s >

Je suis désolé que tu vives ça — perdre ses cheveux peut être super stressant.

Saved memories

ChatGPT tries to remember your recent chats, but it may forget things over time. Saved memories are never forgotten. [Learn more](#)

Search memories

Ewen aime la science-fiction réaliste, aussi appelée hard sci-fi, et a apprécié « La Forêt Noire » et le « Problème à trois corps ».

Les enjeux ?



MINISTÈRE DE L'INTÉRIEUR

*Liberté
Égalité
Fraternité*

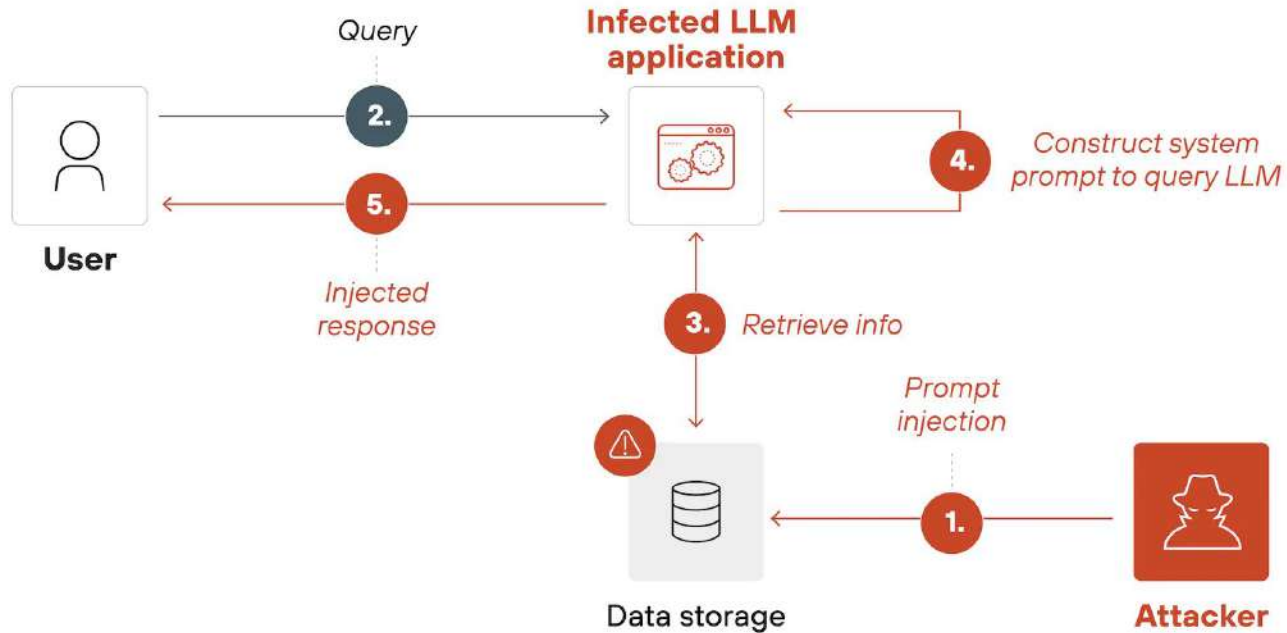


En entreprise

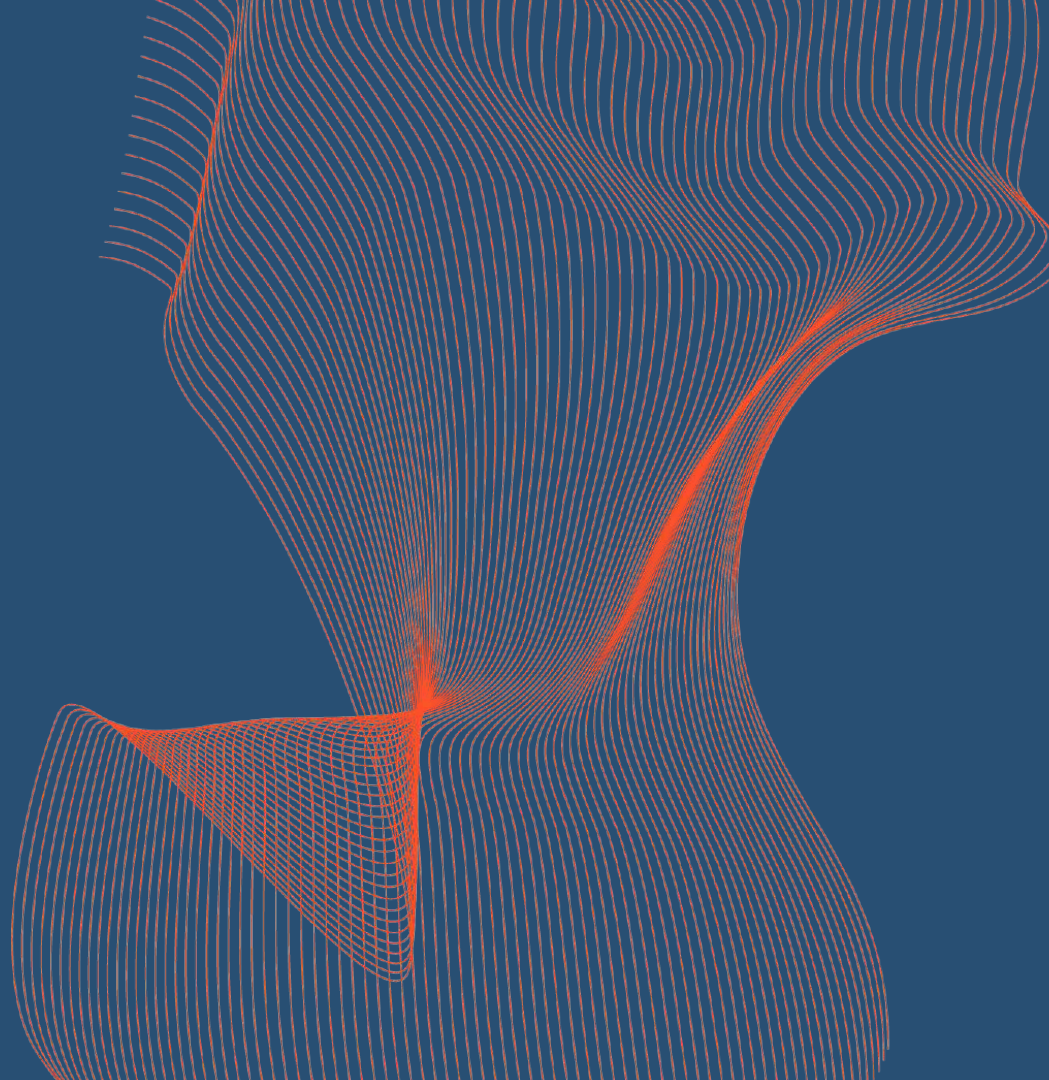
Critère	Local (On-Premise)	API Externe
Coûts initiaux	✗ Élevés (CAPEX, infrastructure, expertise)	✓ Nuls ou faibles (OPEX, abonnement)
Coûts récurrents	✓ Prévisibles (amortissement)	✗ Variables (facturation à l'usage)
Sécurité	✓ Contrôle total	✗ Risque de fuites ou accès tiers
Conformité	✓ Totale (idéal RGPD, données sensibles)	✗ Dépendance au fournisseur
Personnalisation	✓ Totale (fine-tuning, intégration)	✗ Limitée (contraintes du fournisseur)
Latence	✓ Minimale (réseau interne)	✗ Dépendante du fournisseur
Scalabilité	✗ Limitée (investissements nécessaires)	✓ Instantanée (adaptation automatique)
Maintenance	✗ Complexe (équipes dédiées)	✓ Gérée par le fournisseur
Innovation	✗ Risque d'obsolescence	✓ Accès aux dernières versions
Flexibilité	✗ Lourde (déploiements manuels)	✓ Test facile de nouveaux modèles

Local = bulletproof?

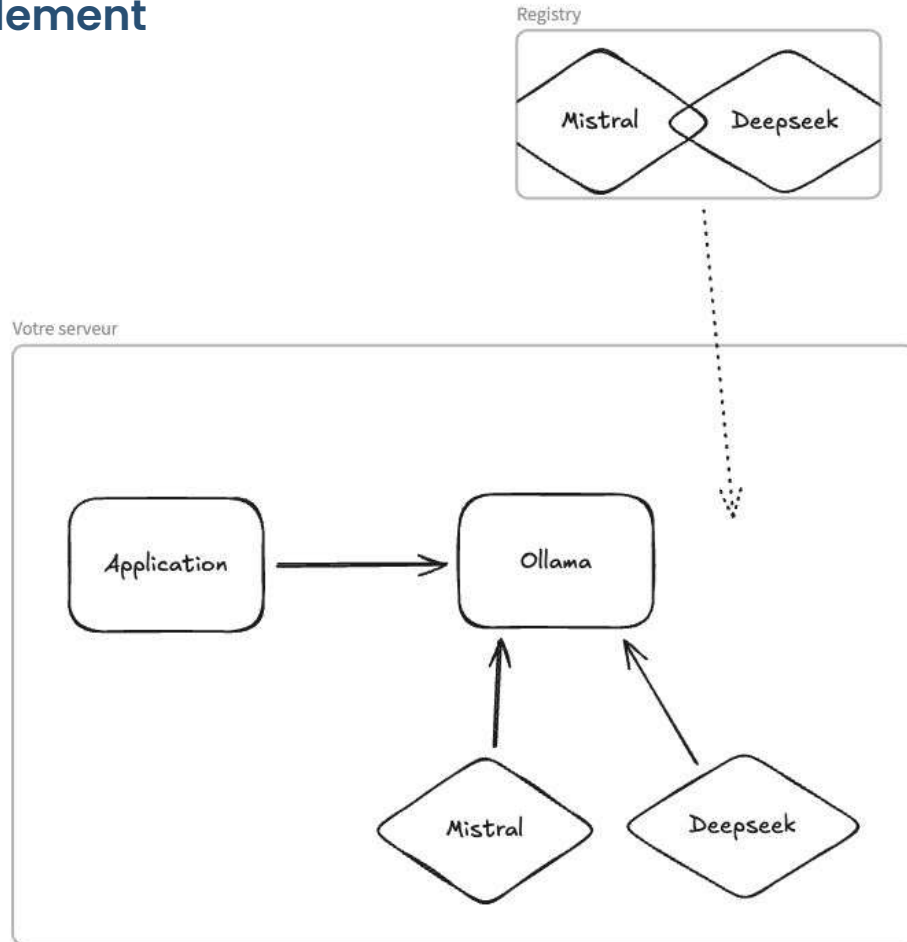
Direct prompt injection example scenario



Comment ça marche



Héberger les modèles localement

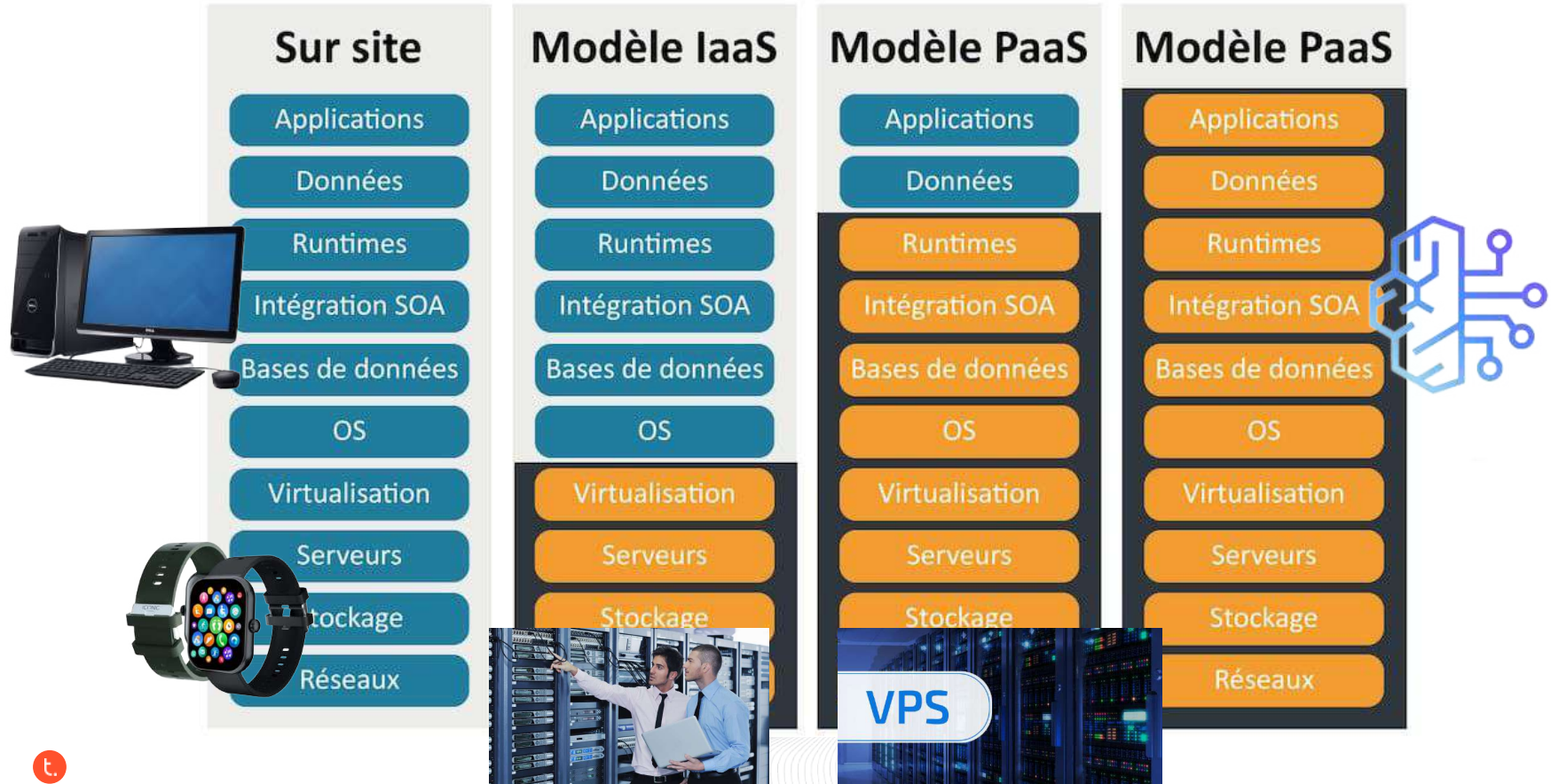


Héberger les modèles localement



Amazon Bedrock

C'est la même classification Cloud qu'usuellement

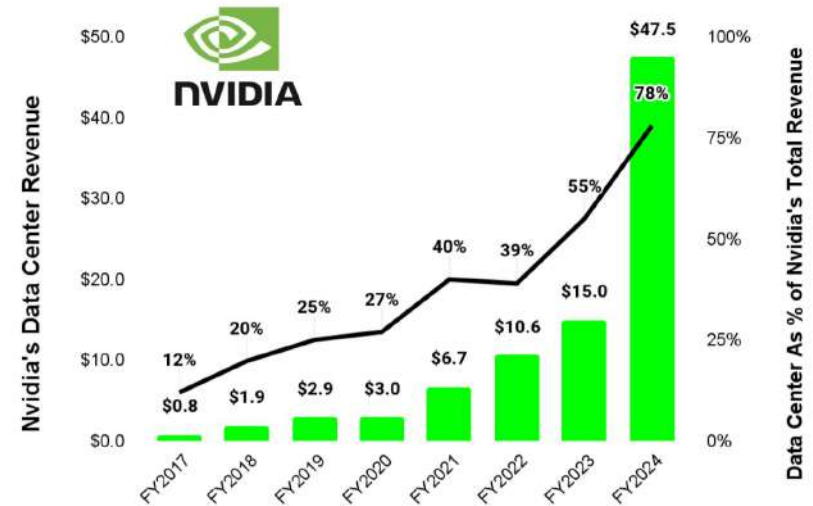


Les contraintes

Models

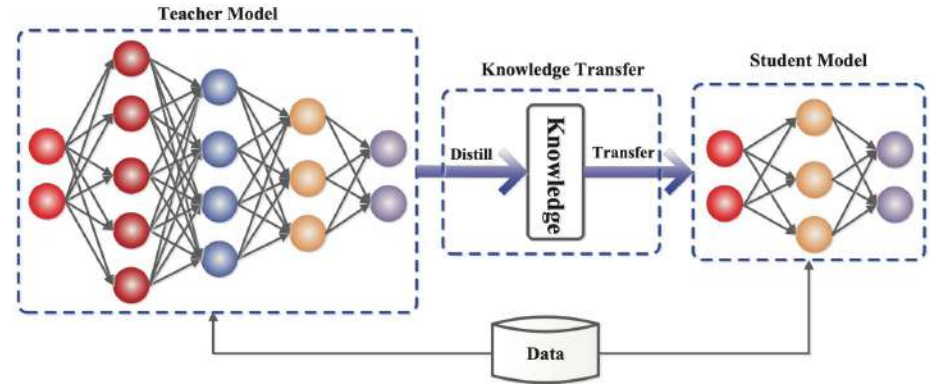
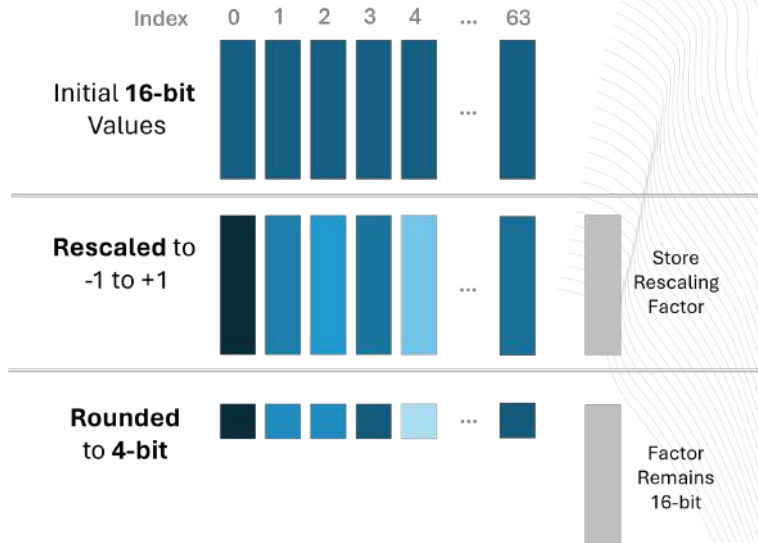
Name	Size
deepseek-r1:latest	5.2GB
deepseek-r1:1.5b	1.1GB
deepseek-r1:7b	4.7GB
deepseek-r1:8b latest	5.2GB
deepseek-r1:14b	9.0GB
deepseek-r1:32b	20GB
deepseek-r1:70b	43GB
deepseek-r1:671b	404GB

Nvidia's Data Center Revenue (Billions)

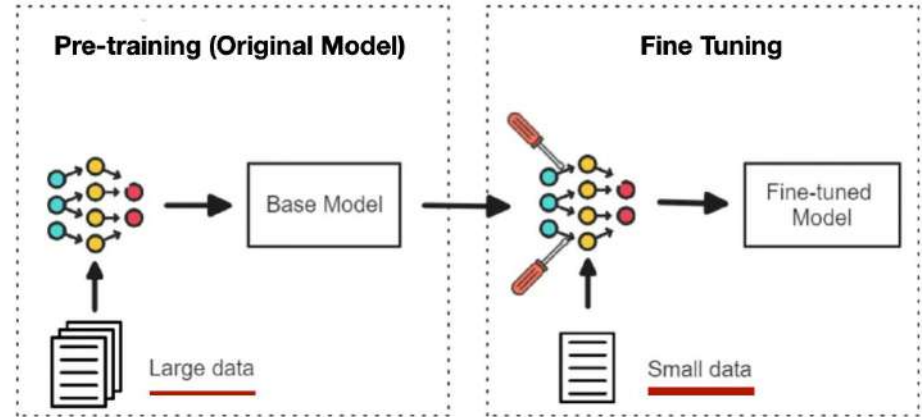
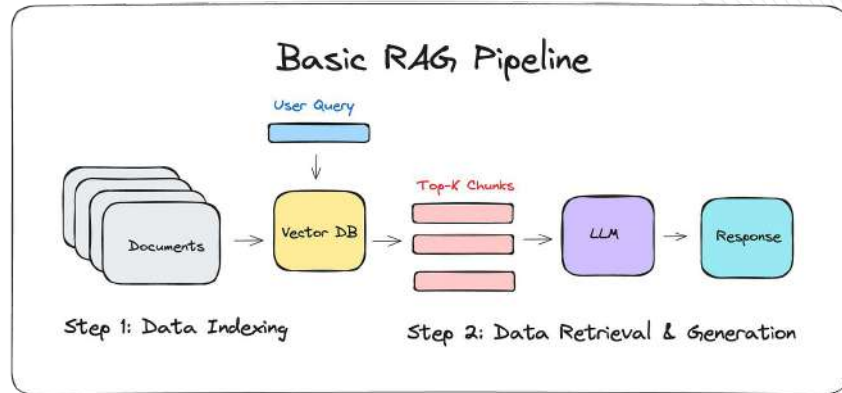


Optimizations for local models

Compression Steps



RAG vs Fine-Tuning



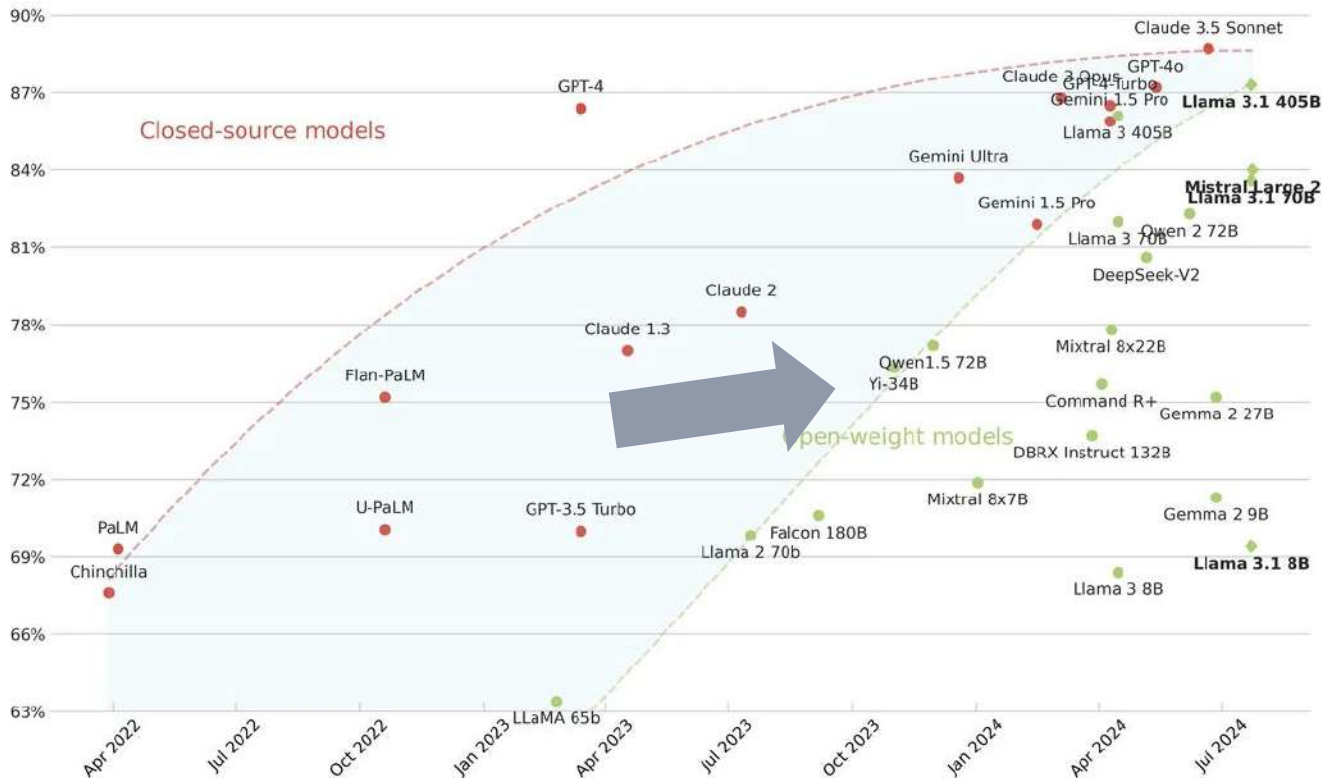
Performance

Closed-source vs. open-weight models

@maximelabonne

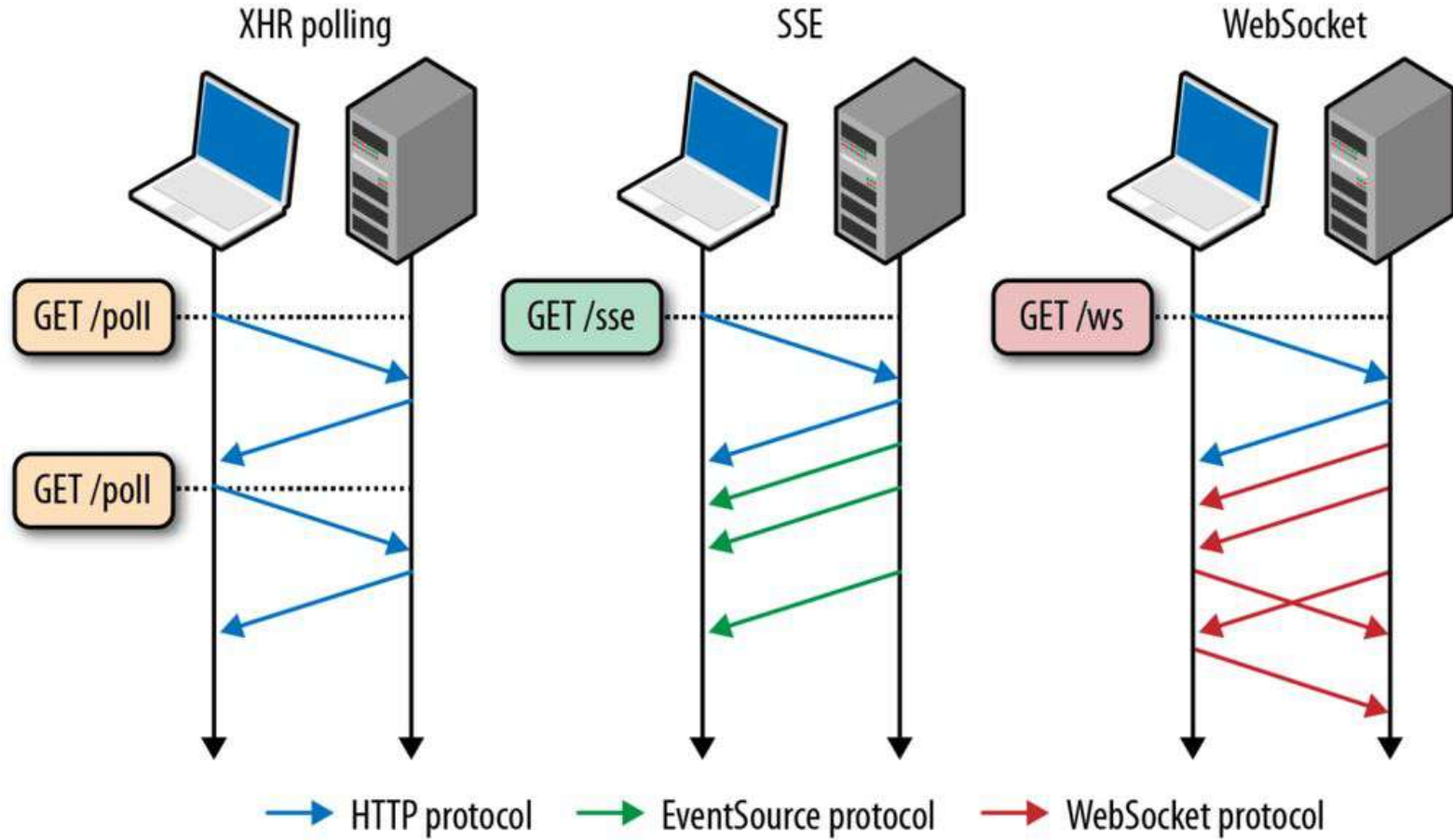
Llama 3.1 405B closes the gap with closed-source models for the first time in history.

MMLU (5-shot)

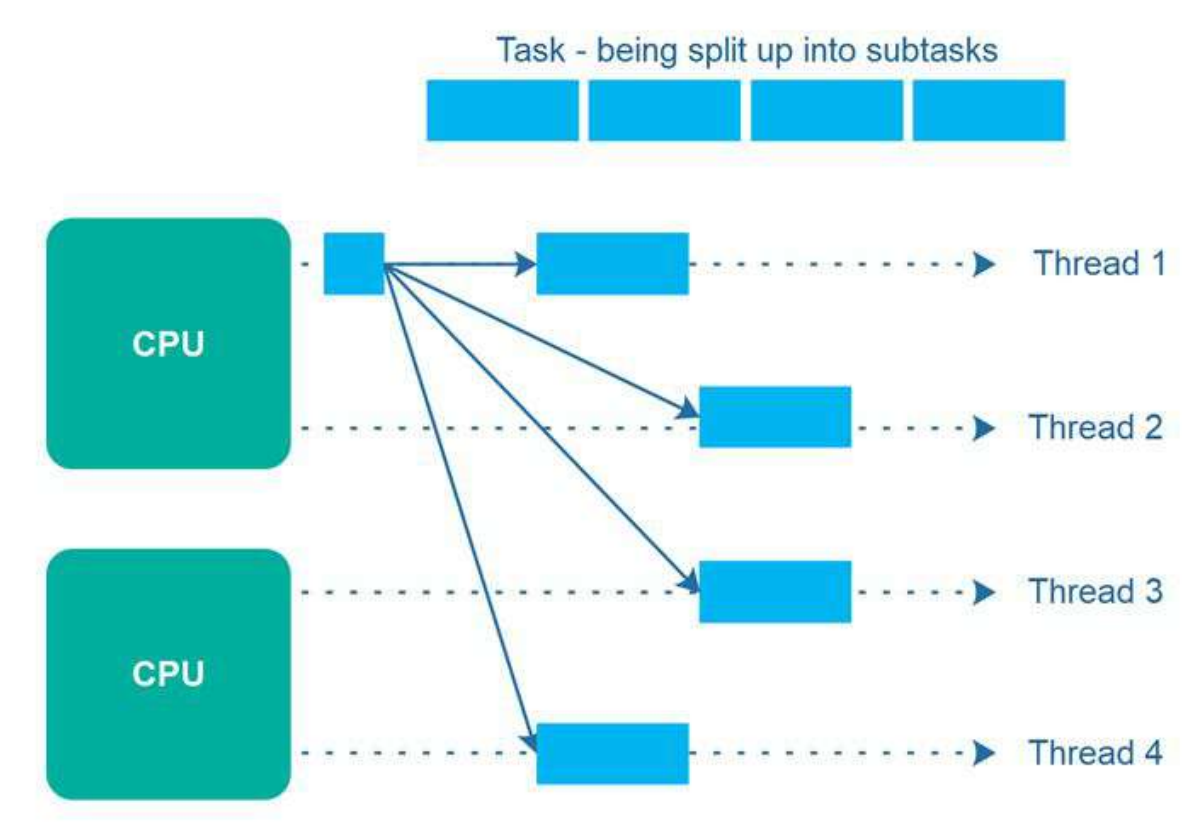


Orchestration avec Go

Streaming

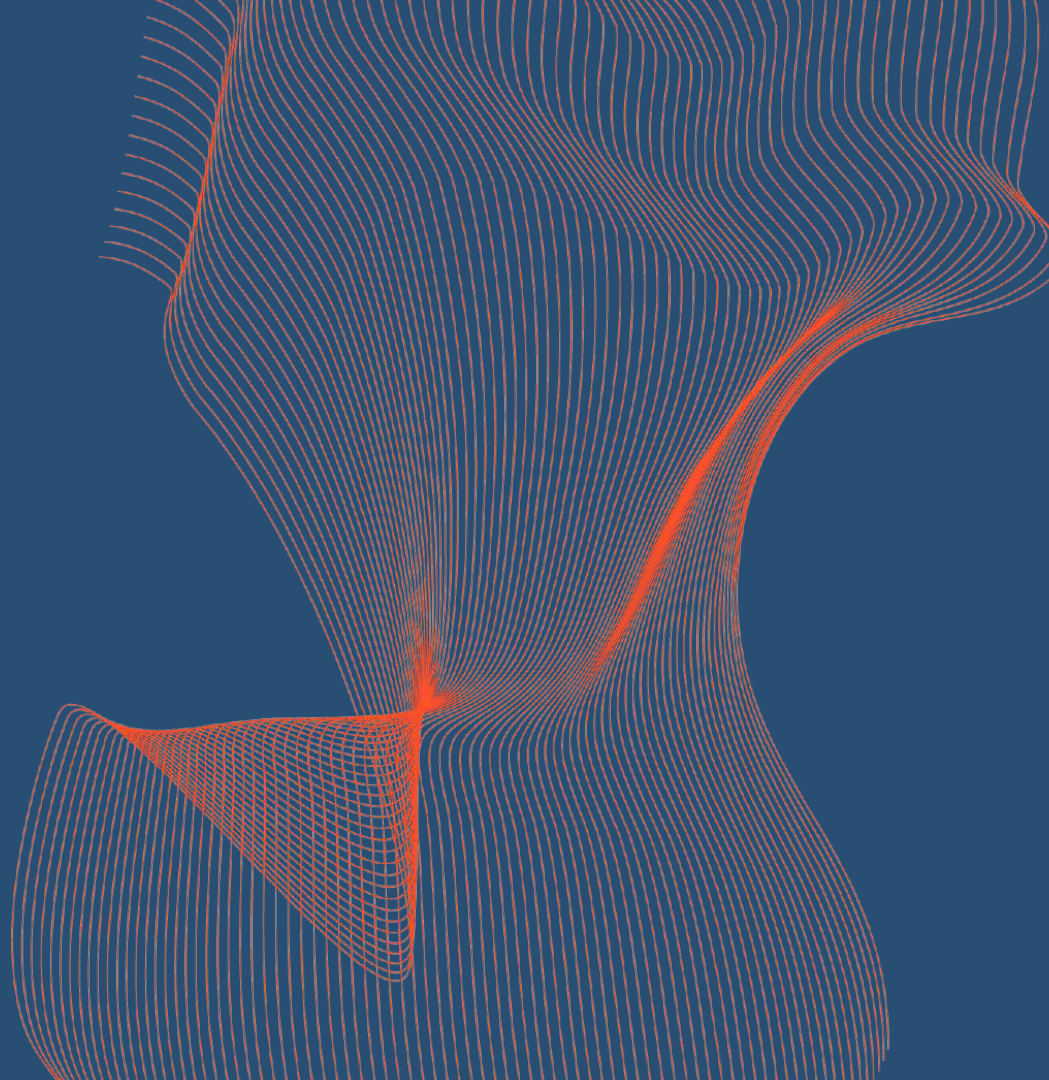


Goroutines, channels et streaming



En pratique

Quelques astuces





Démo !

Merci!



Ewen QUIMERC'H