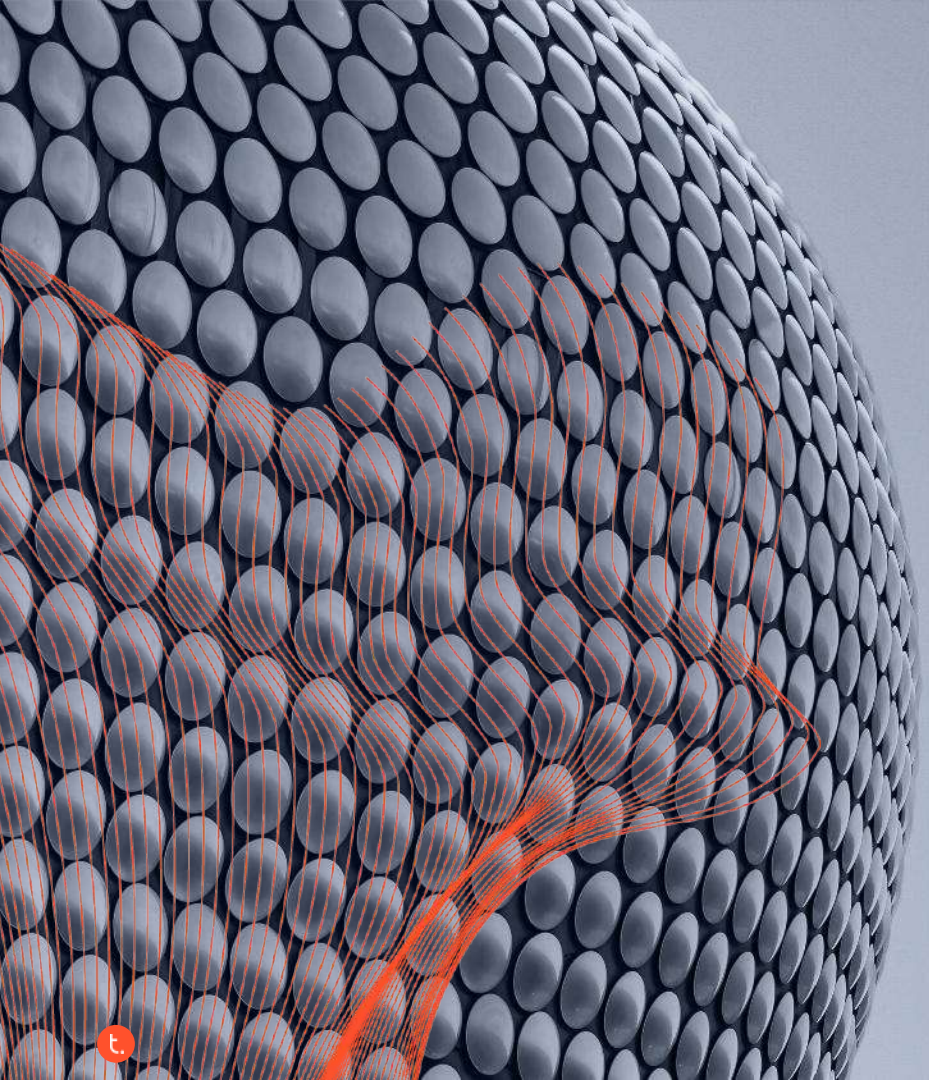


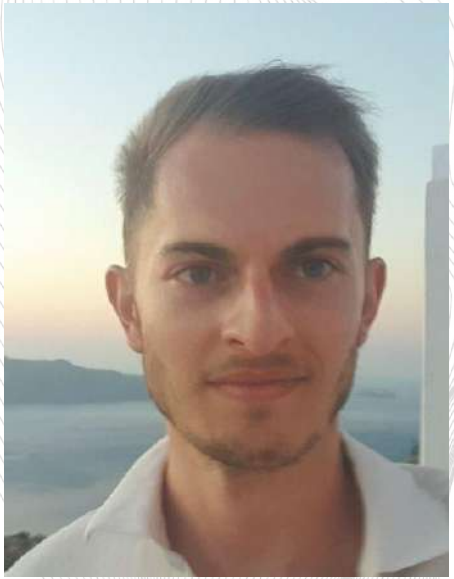
theodo.

LLMs locaux & souveraineté

Ewen Quimerc'h 28/11/2025

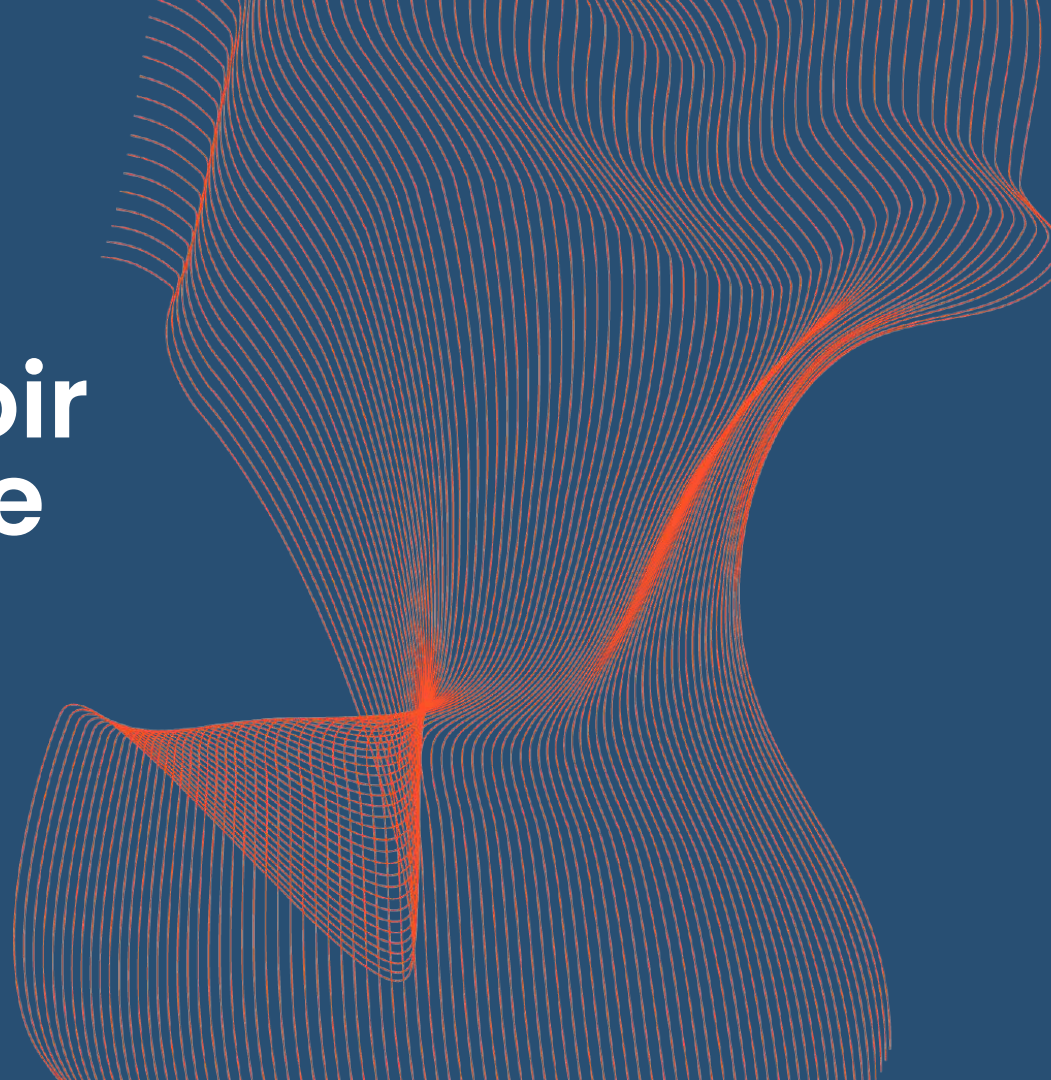
Ewen Quimerc'h

Tech Lead, à Theodo depuis 4 ans

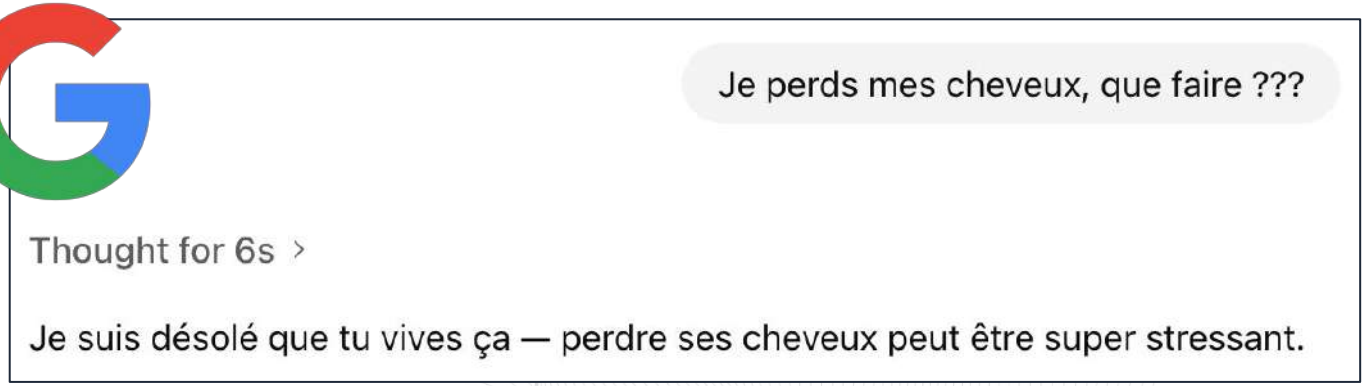


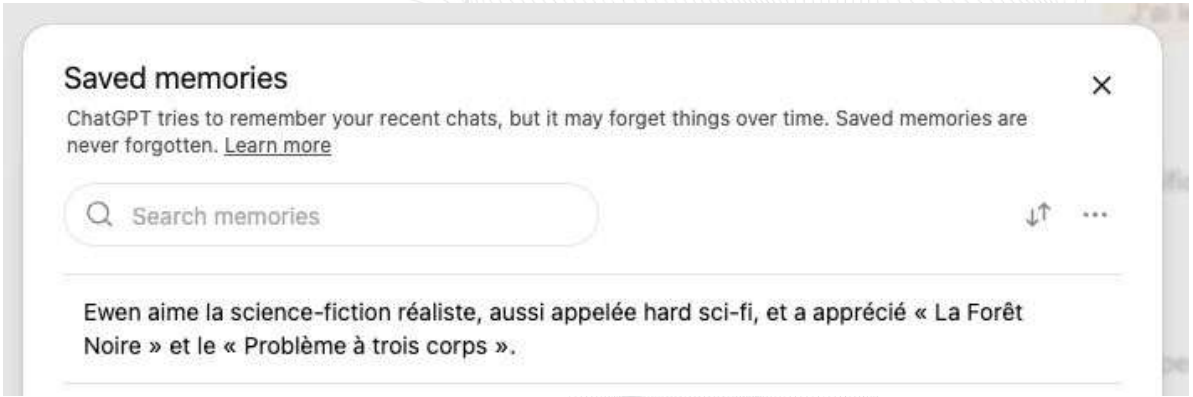
- Plein de missions avec diverses technologies
- Spécialité performance et microservices
- Open-source : Fuego, un framework Go pour OpenAPI
  → github.com/EwenQuim
- Amateur de gastronomie

Pourquoi vouloir faire tourner de l’IA en local?



Le compromis actuel : le SaaS et l'API chatgpt





Les enjeux ?

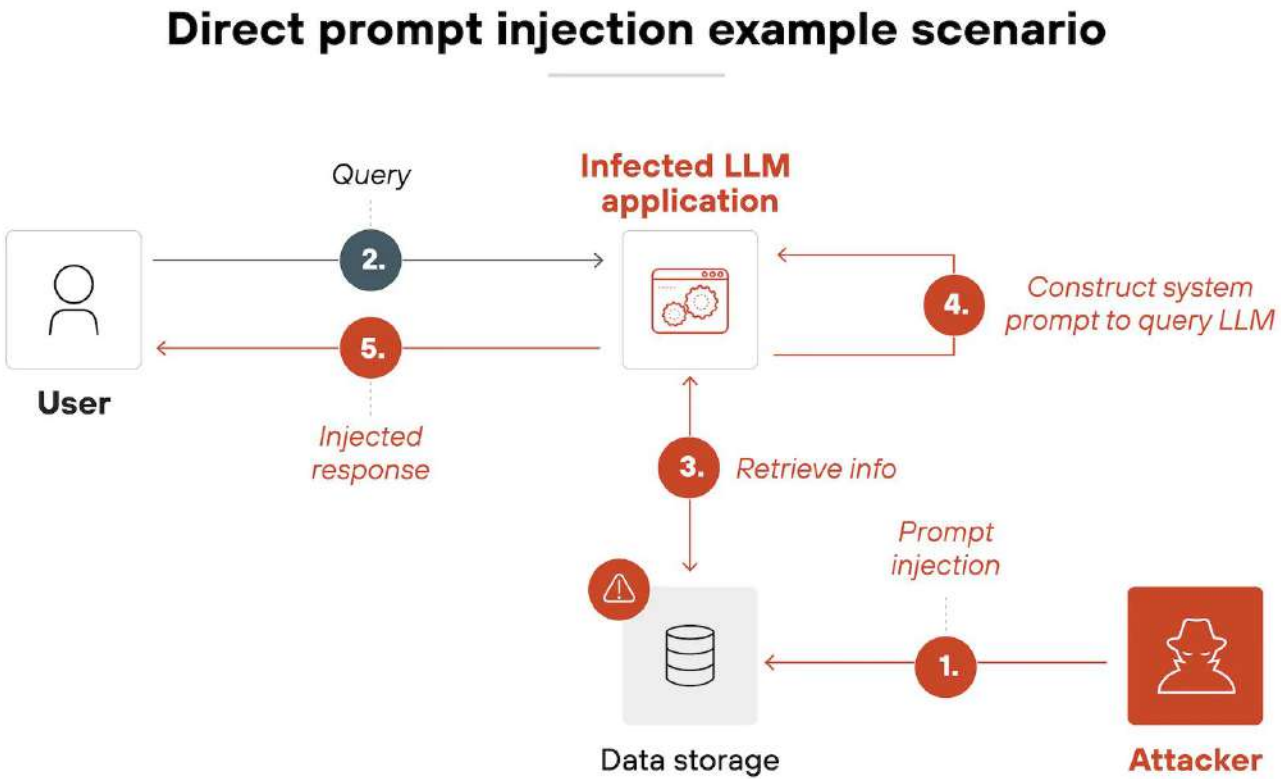
MINISTÈRE
DE L'INTÉRIEUR

Liberté
Égalité
Fraternité

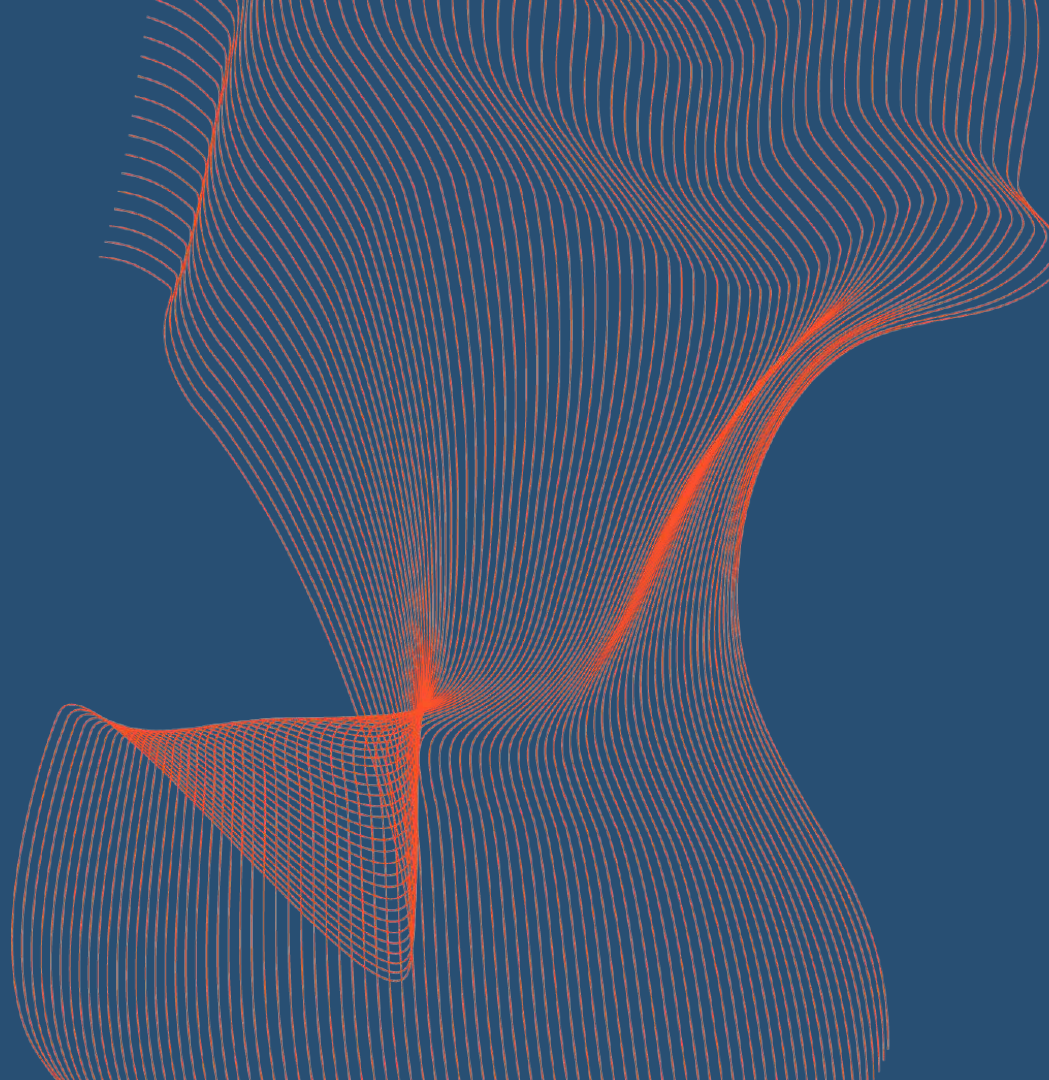
En entreprise

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?

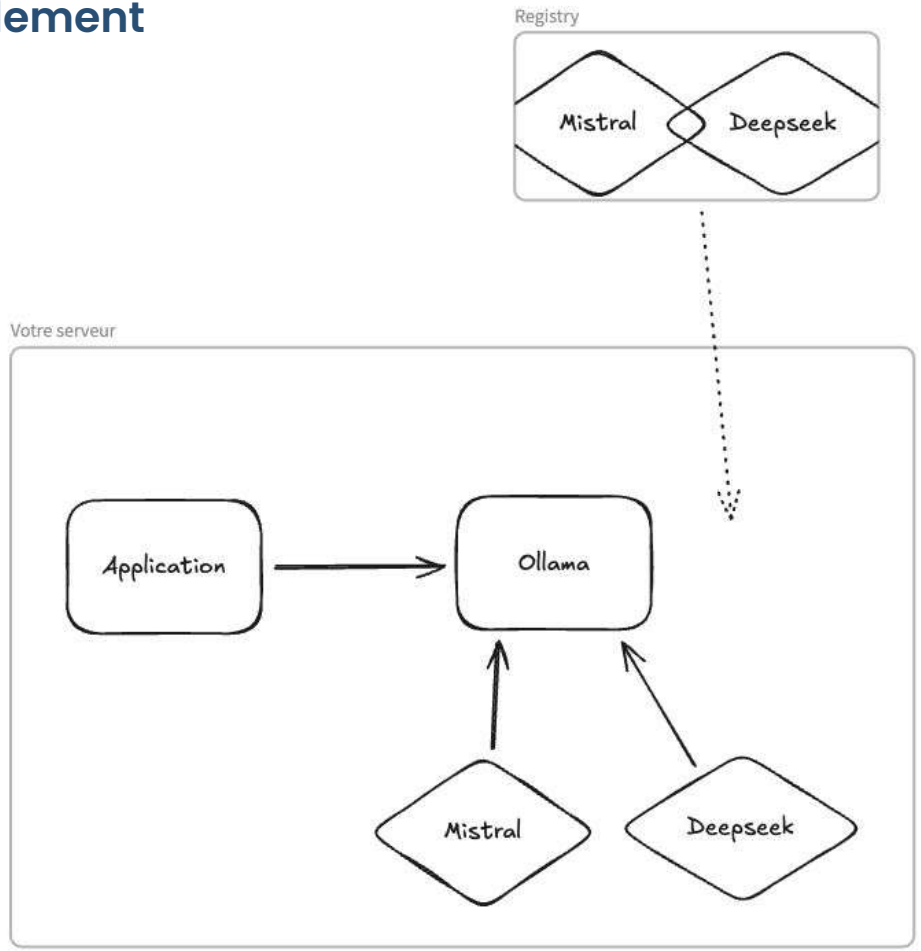


Comment ça marche



Héberger les modèles localement

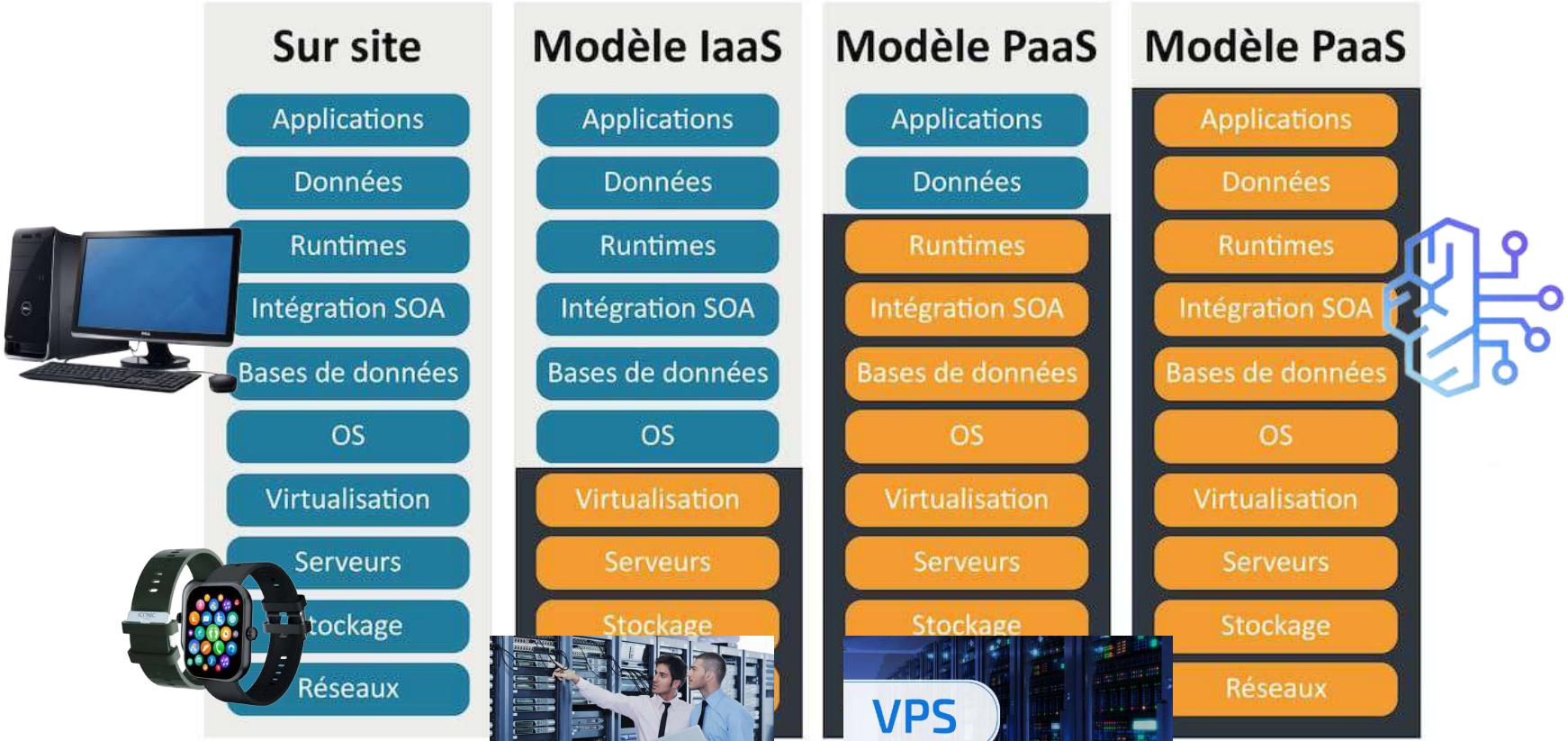
Ollama



Héberger les modèles localement

Amazon Bedrock

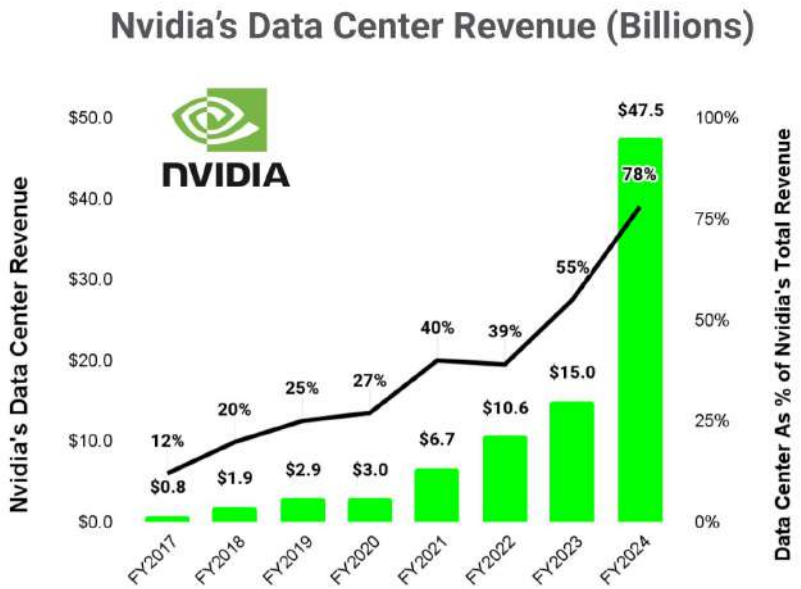
C’est la même classification Cloud qu’usuellement



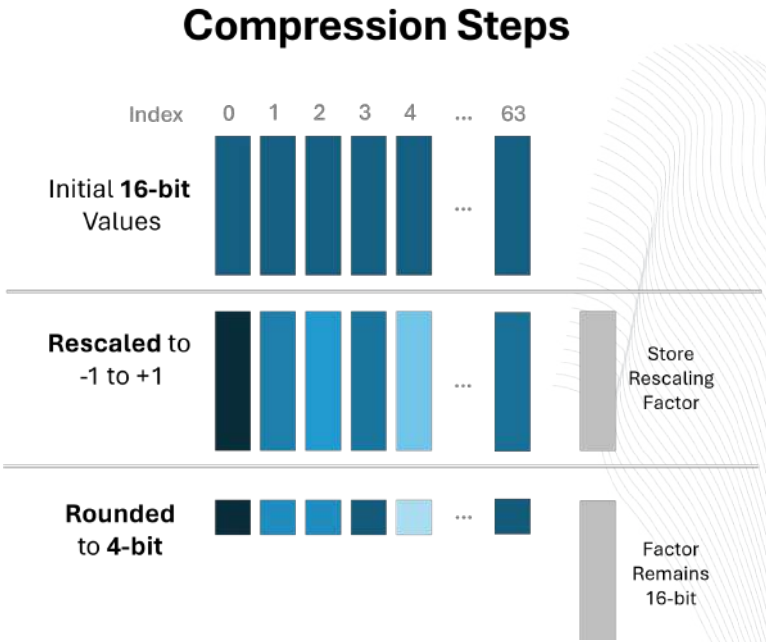
Les contraintes

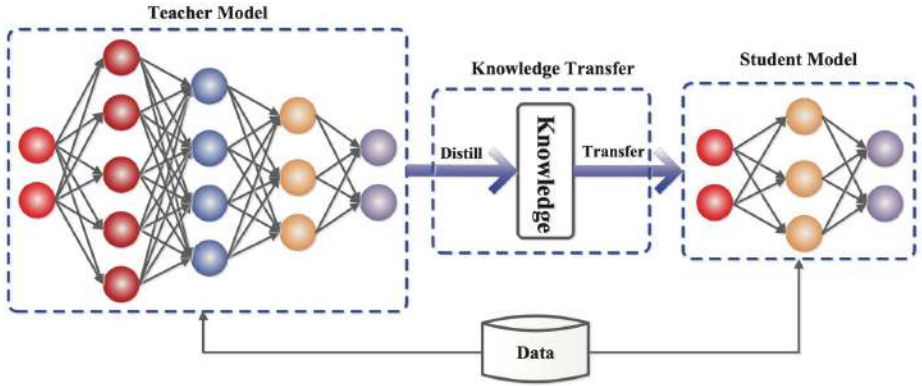
Models

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

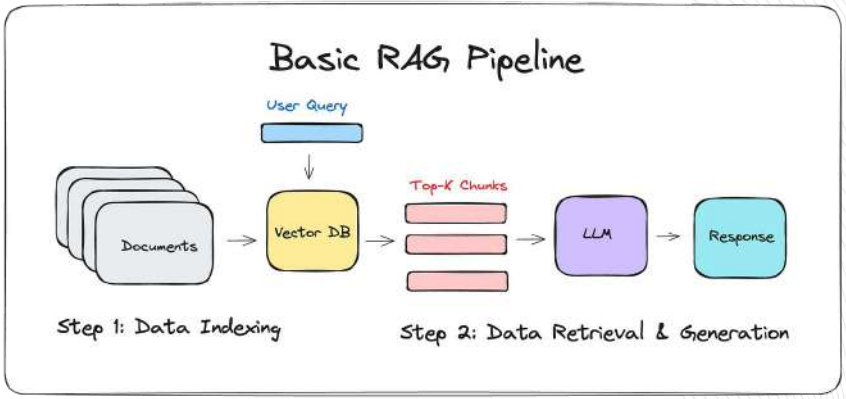


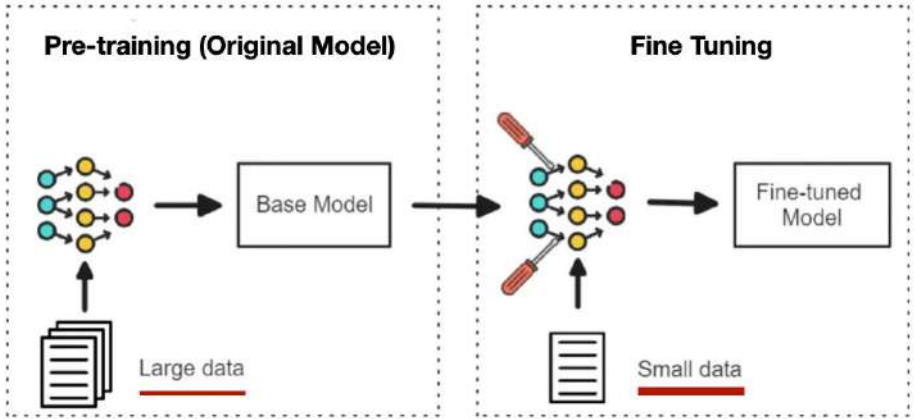
Optimizations for local models



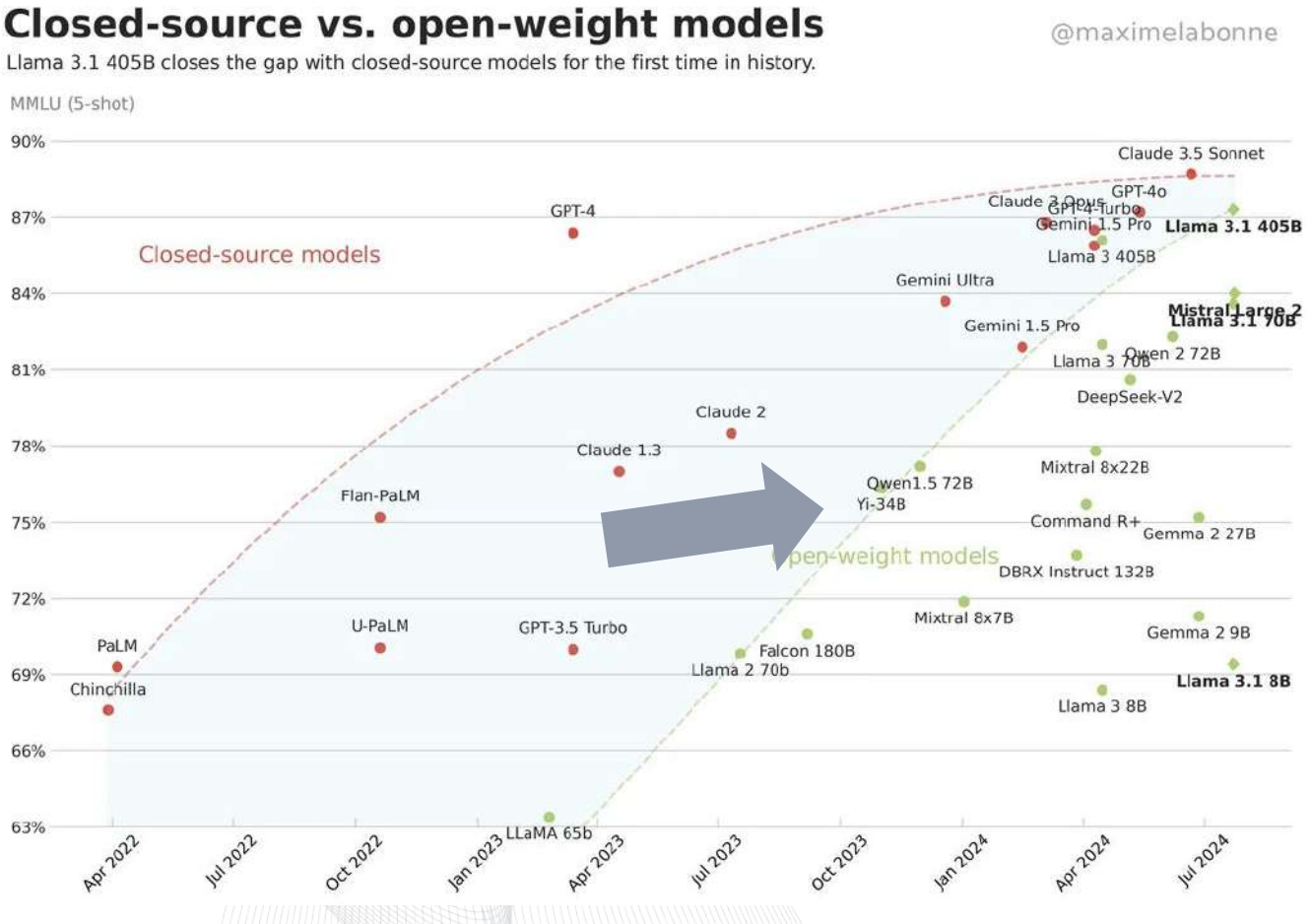


RAG vs Fine-Tuning

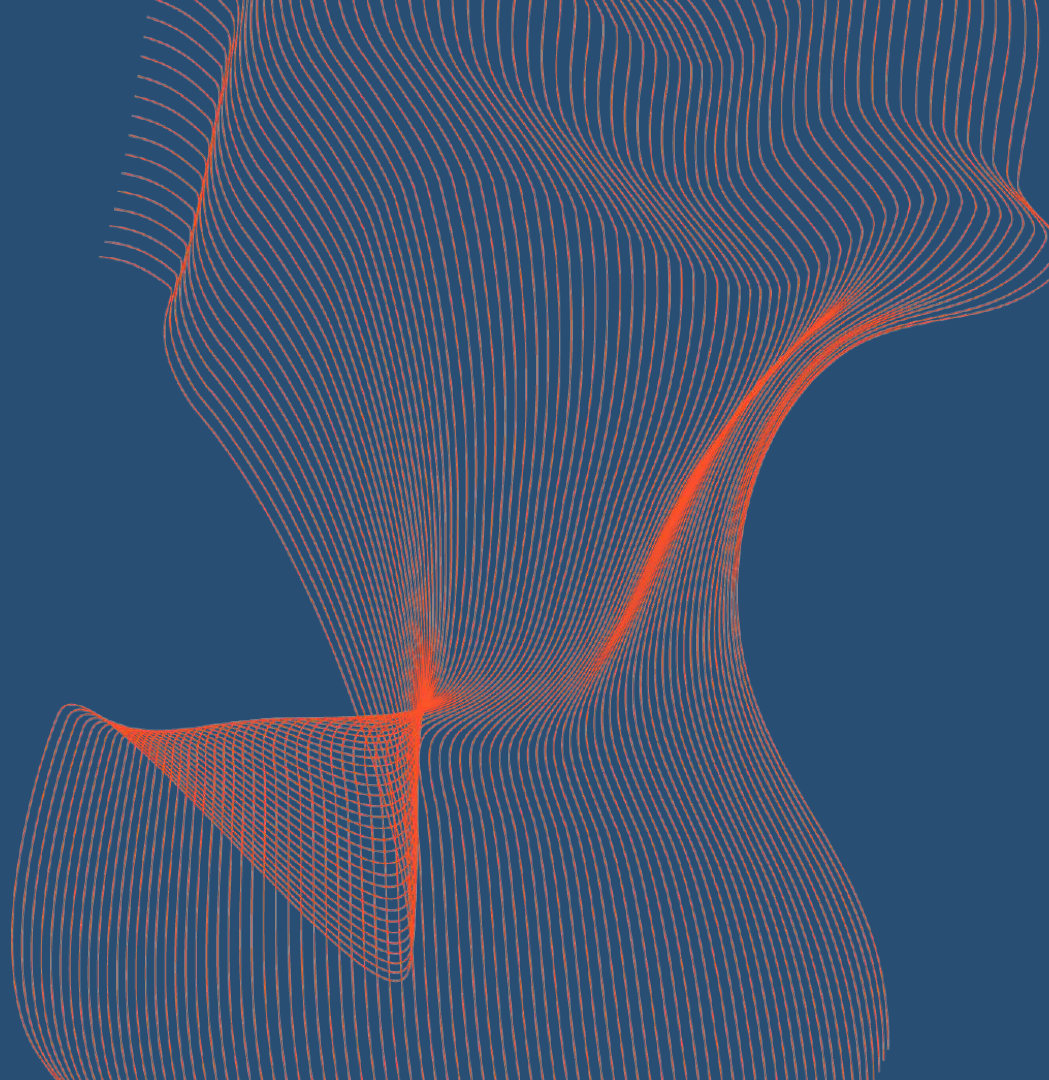




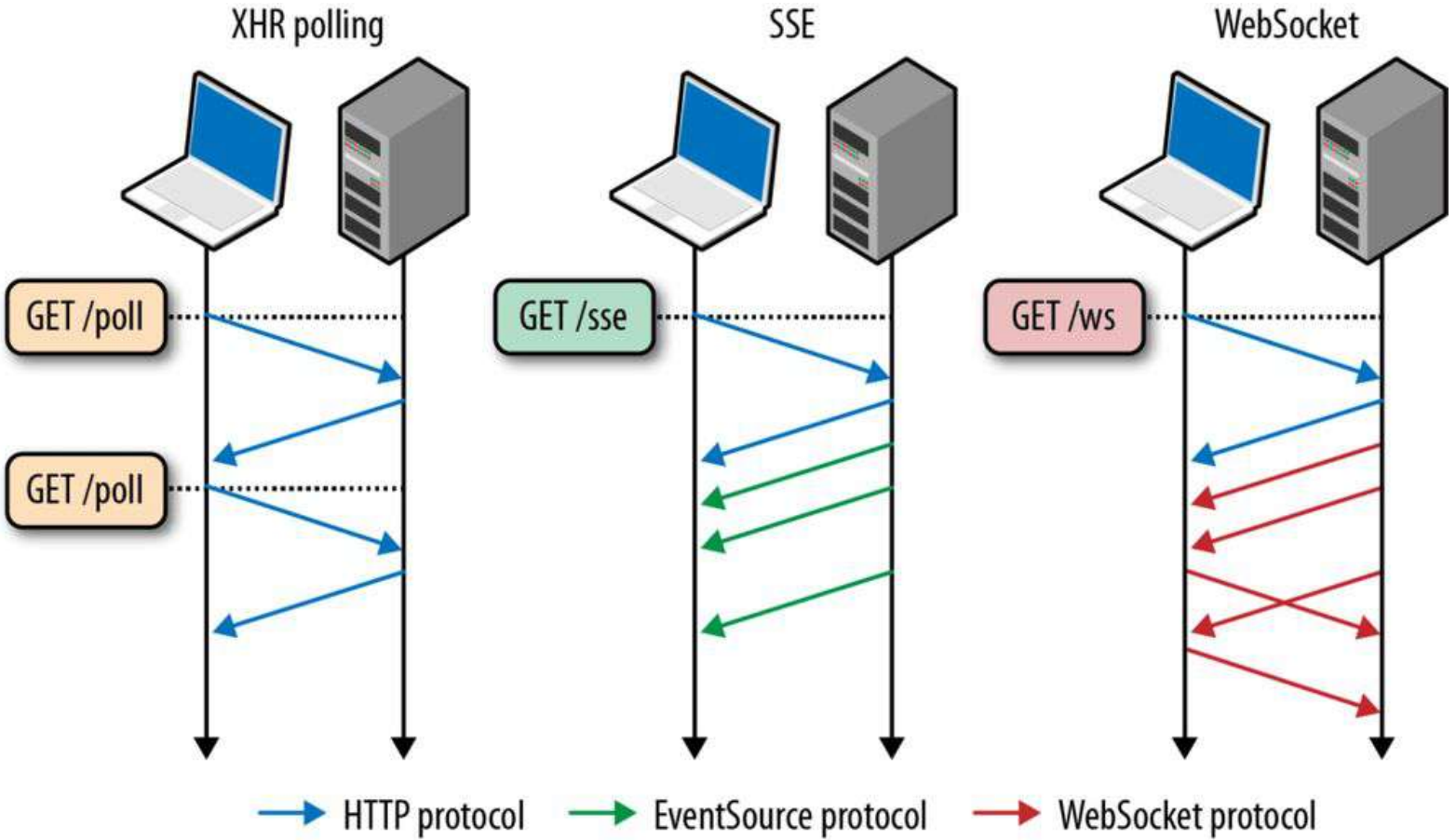
Performance



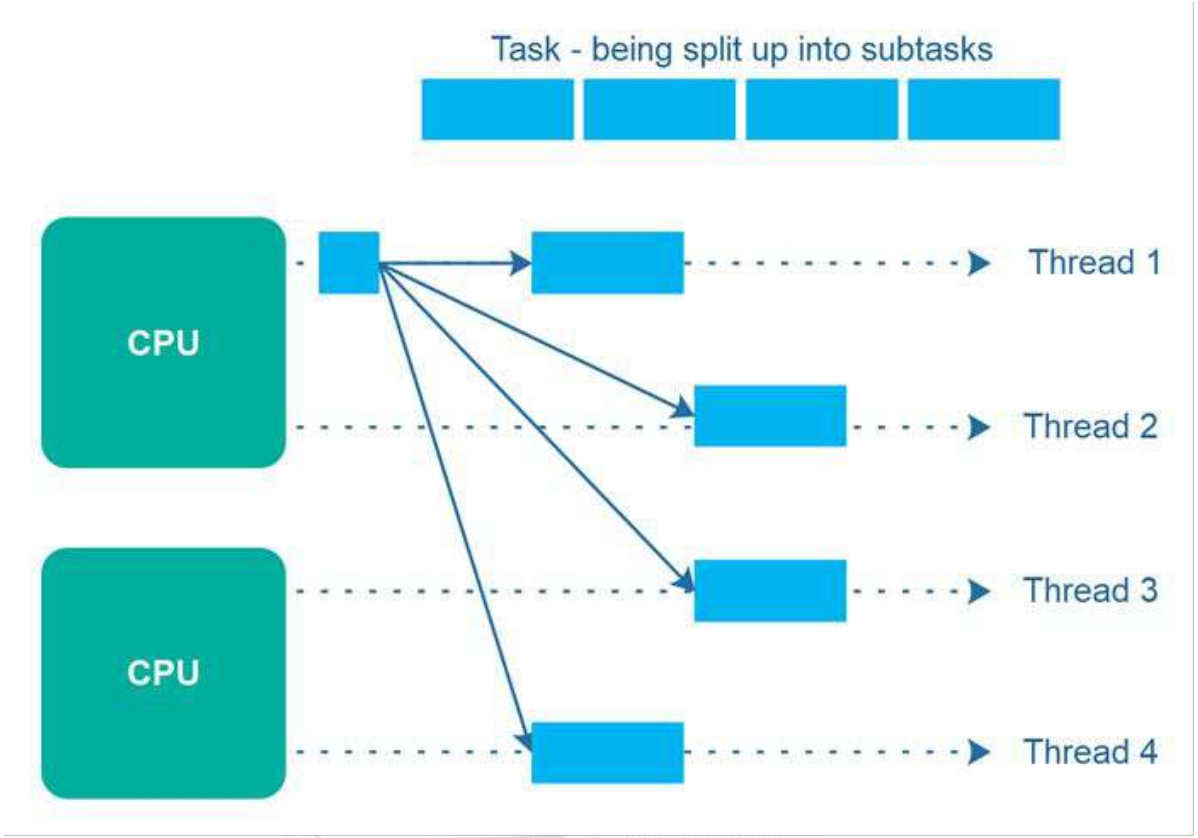
Orchestration avec Go



Streaming

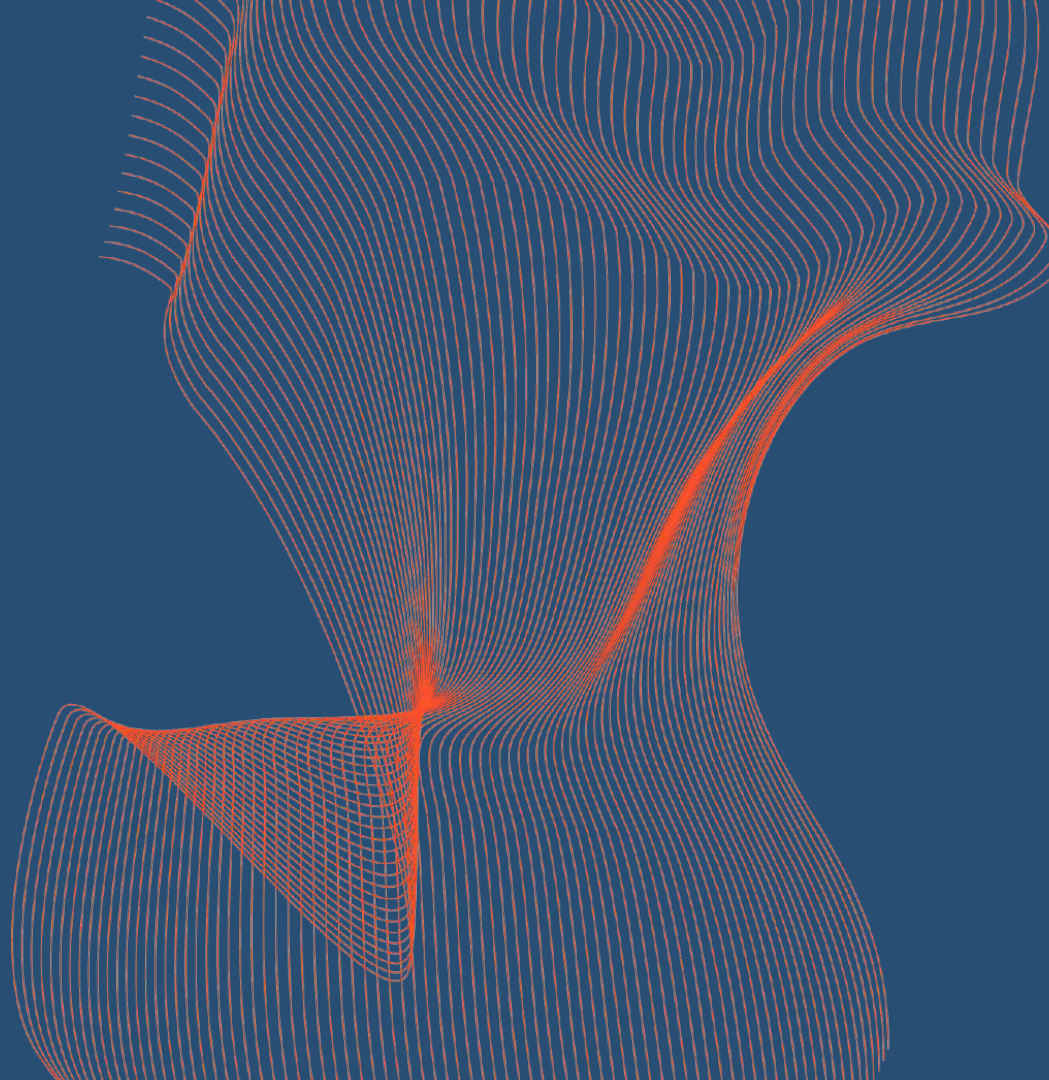

XHR polling
SSE
WebSocket
GET /poll
GET /poll
GET /sse
GET /ws
HTTP protocol
EventSource protocol
WebSocket protocol

Goroutines, channels et streaming



En pratique

Quelques astuces



Démo !

Merci!

Ewen QUIMERC'H