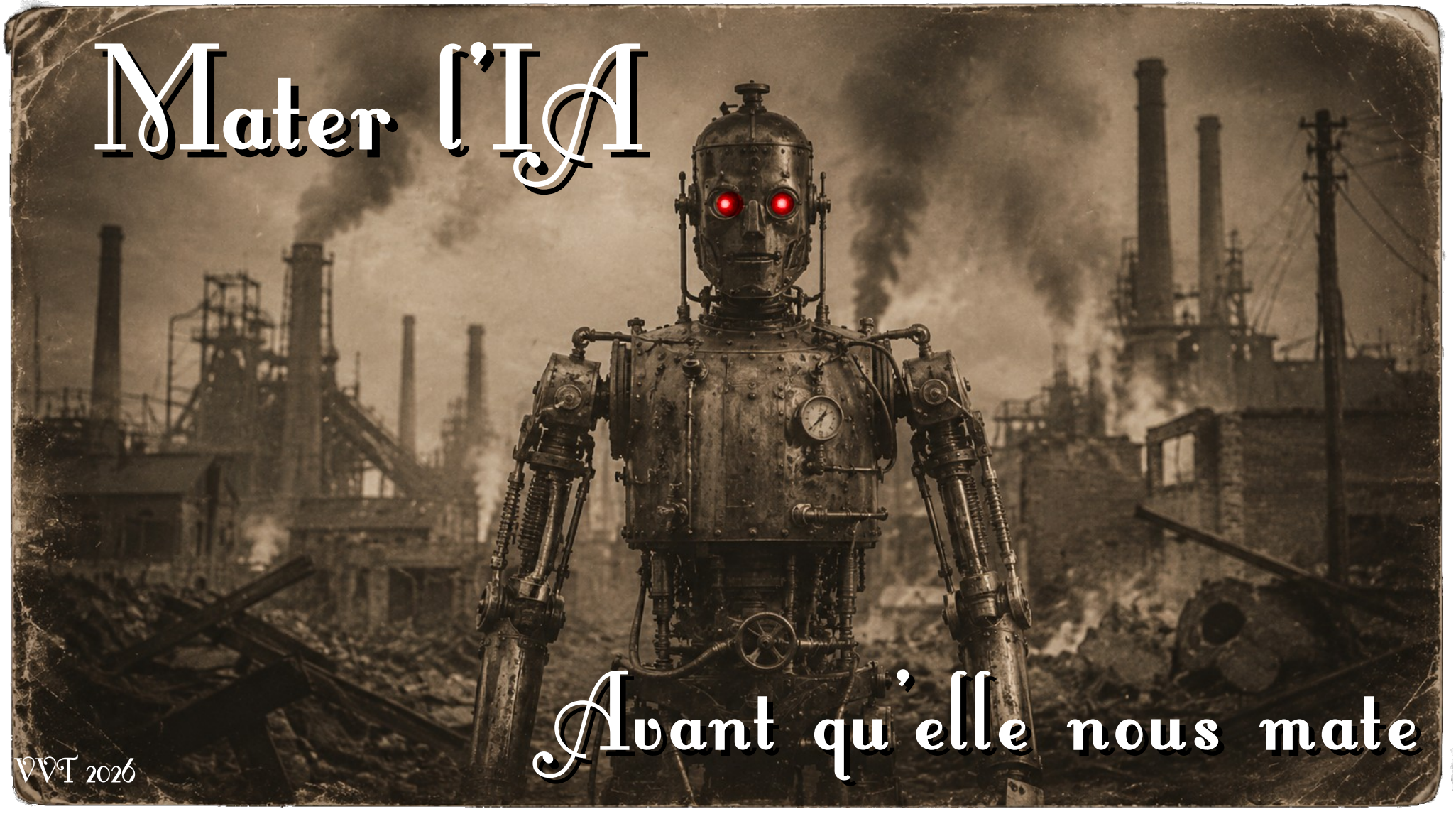

Mater l'IA
Avant qu'elle nous mate
VVT 2026

Pourquoi s'y interesser maintenant ?

Une IA aurait orchestré en quasi-autonomie une cyberattaque au profit de la Chine, une première mondiale

Par **Paul de Breteuil**

Le 14 novembre 2025 à 17h54

Les espions ont eu recours à Claude un modèle d'intelligence artificielle générative développé par l'entreprise américaine Anthropic. Entre 80 et 90% de l'opération aurait été automatisée.

Pour la première fois, une IA aurait mené en quasi-autonomie une vaste opération d'espionnage pilotée par des pirates numériques chinois financés par Pékin. Pour arriver à leurs fins, les espions ont eu recours à Claude, un modèle d'intelligence artificielle générative développé par l'entreprise américaine Anthropic. C'est l'entreprise elle-même qui a révélé ces informations dans un rapport de 13 pages publié jeudi 13 novembre. *«Nous pensons qu'il s'agit du premier cas documenté d'une cyberattaque à grande échelle menée par une IA sans intervention humaine significative»* a déclaré Anthropic sur X.

L'USINEDIGITALE

Cybersécurité : Claude Mythos a trouvé plus de 10 000 failles critiques un mois après son lancement, y compris dans les systèmes les plus importants au monde

Le projet Glasswing d'Anthropic, qui donne accès aux capacités de son modèle Claude Mythos à quelques sociétés triées sur le volet, transforme profondément le paradigme de la cybersécurité. Il déplace le centre de gravité de la détection vers la capacité de correction et de déploiement des correctifs.

Thibault Caudron

Publié le 26 mai 2026 à 14h06

LES NUMERIQUES

La Chine soupçonnée d'avoir mis la main sur Mythos, d'où son interdiction et celle de Fable 5 ?

Par Nassim Chentouf | Publié le 14/06/26 à 22h30

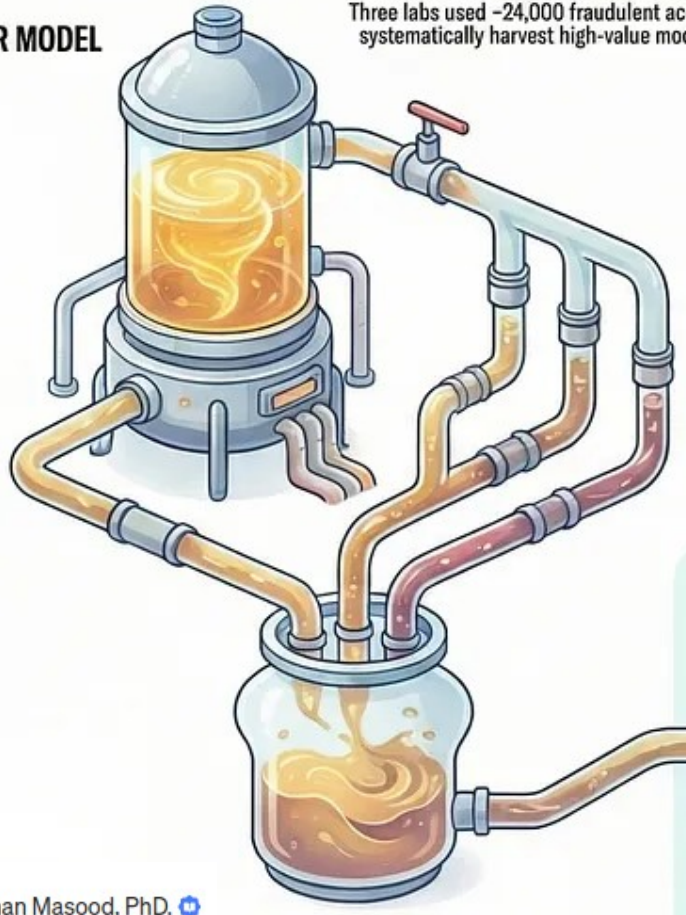
Selon *Semafor*, un groupe en relation avec la Chine aurait accédé à Mythos, le modèle IA le plus avancé d'Anthropic. Cette information aurait directement poussé la Maison-Blanche à interdire Mythos et Fable 5 hors des frontières étasuniennes. D'où la décision brutale du gouvernement de Donald Trump contre Anthropic.

The Rise of Industrial-Scale LLM Distillation Attacks

16 MILLION+ ILLICIT EXCHANGES

Three labs used ~24,000 fraudulent accounts to systematically harvest high-value model outputs.

TEACHER MODEL
(Claude)



STUDENT MODEL
(Illicit)

DeepSeek

Over 150,000 Exchanges

Primary Capabilities Targeted:
Reasoning, reward-model grading, and censorship-exasion.

Moonshot AI

Over 3.4 Million Exchanges

Primary Capabilities Targeted:
Agentic reasoning, tool use, coding, and computer vision.

MiniMax

Over 13 Million Exchanges

Primary Capabilities Targeted:
Agentic coding, tool use, and crchestration.

TARGETING PROPRIETARY "DIFFERENTIATORS"

Atteckers specifically targeted complex capabilities like agentic reasoning, tool use, and ooding.

THE "HYDRA CLUSTER" ARCHITECTURE

Massive account incoounts meet webs used to distribute traffic and evade detection.

THE SOLUTION: MULTI-LAYERED DEFENSES

BEHAVIORAL FINGERPRINTING

Using classifiers to identify repetitive, high-volume prompt patterns that deviate from normal usage.

INTELLIGENCE SHARING

Distributing technical indicators among AI labs and cloud providers to create a holistic defense.

HARDENED ACCESS CONTROLS

Strengthening verification for the pathways most commonly exploited for setting up fraudulent accounts.

Adnan Masood, PhD.
· Mar 9, 2026

Pourquoi s'inquiéter ?

Survivre à l'IA ?

Usages volontaires

- Règlement européen sur l'intelligence artificielle
- Extraterritorialitédu droit américain
- Hallucinations
- Fuite de clés API
- Fuite de données et de secrets

Défense face aux attaques

- Minimisation de la surface d'exposition
- WAF, Rate limit, fail2ban, pots de miel, pattern matching
- DevSecOps, carto dépendances logicielles, veille CVEs, orchestration, docker/kubernetes, CI/CD, dev agile, intégration continue, git renovate, etc.

Survivre à l'IA ?

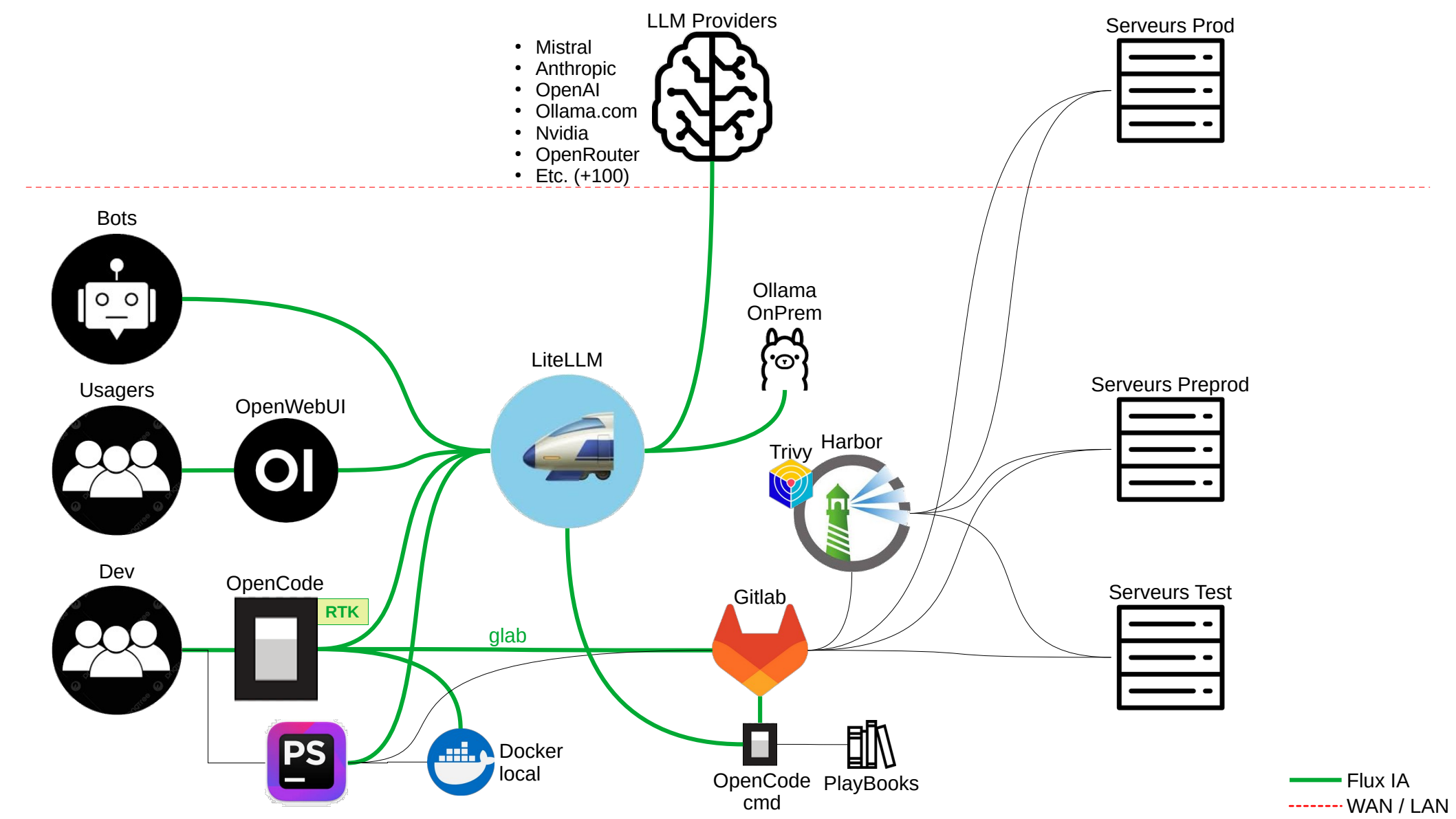
Et les usages involontaires / inconscients

- Nos matériels et logiciels comportent de l'IA ?
- Les journaux, metriques, etc. alimentent une IA ?
- On peut désactivé (vraiment) l'IA ?
- On pourra encore la désactiver plus tard ?
- On peut spécifier le endpoint LLM de notre choix ?

- Quid des connecteurs MCP ?
- Quid de l'interropérabilité ?
- Quid de la portabilité ?

S'appuyer sur l'IA

Cas du DevSecOps


LLM Providers
Mistral
Anthropic
OpenAI
Ollama.com
Nvidia
OpenRouter
Etc. (+100)
Serveurs Prod
Bots
Ollama
OnPrem
LiteLLM
Usagers
OpenWebUI
Serveurs Preprod
Trivy
Harbor
Dev
OpenCode
RTK
Gitlab
glab
Serveurs Test
Docker
local
OpenCode
cmd
PlayBooks
Flux IA
WAN / LAN

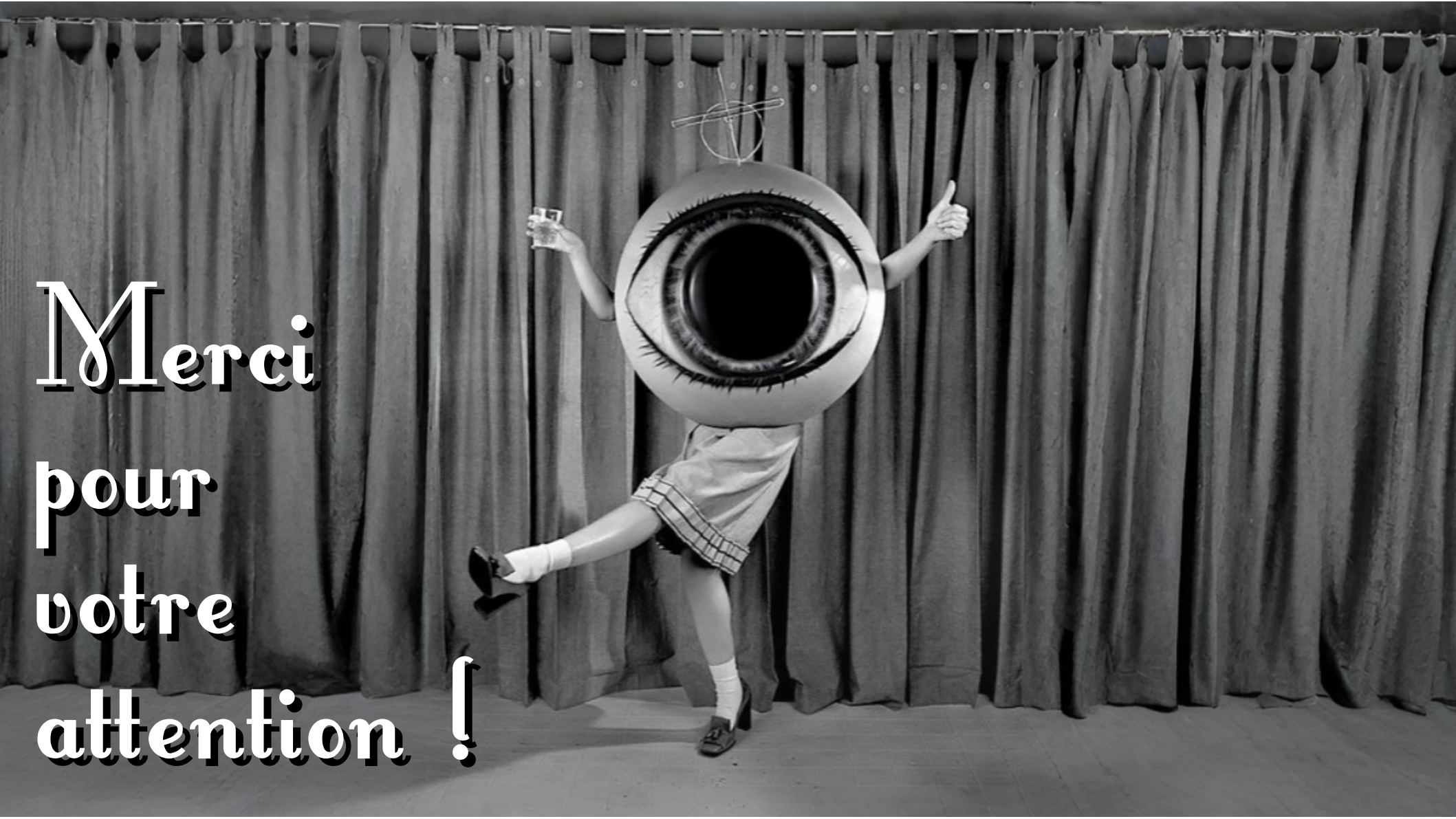
Un « proxy llm » ?

« Core » features de LiteLlm

- API unifiée
- Modèles « virtuels » routés
- Fallback de provider / modèle
- Suivi des usages et des coûts
- Gestion de virtual API keys
- Attributions de quotas et de rate-limit
- Optimisation d’usage de tokens / contexte de sessions
- Sur-prompts
- Logs
- Connecteurs MCP

Mais ne quand même pas oublier

- Faire des prompts efficaces
  - Concis, précis
  - Prompter le prompt
  - Si tu sais pas dis le
  - Verbes d’action à l’impératif
  - Numéroter les actions
- Limiter l’historique (input tokens)
- Limiter les sorties (output tokens)
- Clarifier l’objectif et le contexte
- Donner des exemples
- Ne pas oublier les agents, les skills, les AGENTS.md


Merci
pour
votre
attention !