

De RAGaRenn à ILaaS Plateformes d'IA du local au national

21 janvier 2026
3^{ème} journée d'étude ARDOISE

Olivier WONG

Vice-président numérique

Président du bureau



IA : Intelligence Artificielle

Ce contenu *sous licence ouverte* [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/) a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du Programme d'Investissements d'Avenir portant la référence ANR-21-DMES-0001

Enjeux du numérique dont l'IA

- Souveraineté
- Sécurisation des données
- Impact environnemental



Souveraineté numérique : capacité à faire des choix technologiques

Guillaume Poupard, dir. ANSSI , 01 oct. 2016

« Il est tout d'abord essentiel de permettre au politique de garder sa liberté de manœuvre et son autonomie de décision (...) le maintien de cette liberté de manœuvre s'appuie sur une autonomie scientifique, technologique et capacitaire, à la fois nationale et européenne »

<https://www.caia.net/revue-auteurs-rubriques-numeros/article/souverainete-numerique-l-heure-du-choix/433>

Vincent Strubel, dir. ANSSI, 06 jan. 2026

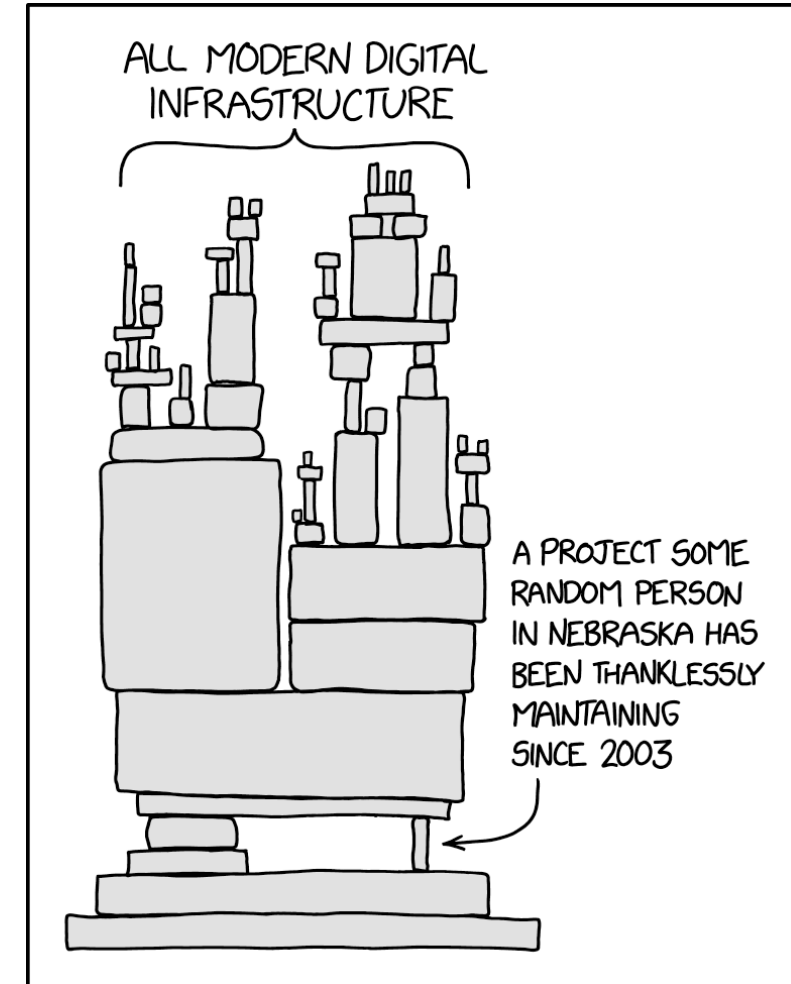
« la notion de souveraineté numérique renvoie à au moins trois enjeux distincts et complémentaires :

1. Ne pas être une victime facile dans nos usages numériques, à la merci de la première cyberattaque venue ;
2. Faire appliquer nos règles, plutôt que subir celles des autres sans avoir voix au chapitre ni possibilité de recours ;
3. Disposer d'une liberté de choix éclairé et d'usage dans le recours aux technologies indispensables à nos missions (en disposant de solutions européennes, en maîtrisant des briques open-source, ou en s'appuyant sur une diversité d'offres suffisantes pour ne pas être dans une dépendance critique). »

<https://www.linkedin.com/pulse/secnumcloud-en-pas-si-bref-vincent-strubel-505ge/>

Souveraineté numérique : capacité à faire des choix technologiques

- **Choisir et diversifier nos dépendances**
- Infrastructures et matériel : Etats-Unis, Asie (Taiwan, Chine)
- Logiciels et services : marché mondial
- Usages : captation et ancrage par les plateformes
- **Quelles dépendances ?**
- Chaîne d'approvisionnement
- Répartition de la valeur entre acteurs
- Déplacement des enjeux sur les données



Ecosystèmes numériques ouverts européens

Stratégie numérique européenne

- approche stratégique du secteur des logiciels libres dans l'UE
- cadre stratégique et opérationnel visant à renforcer l'utilisation, le développement et la réutilisation des ressources numériques ouvertes

**Contribution ouverte jusqu'au
03 février 2026**

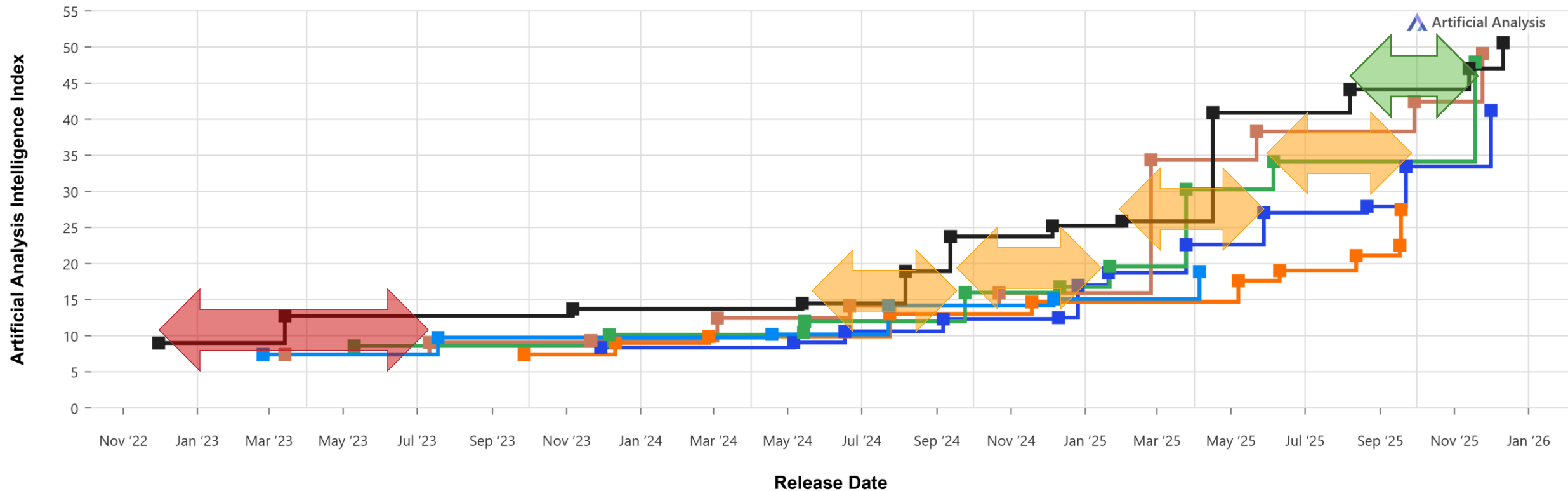


Performance des LLM de pointe dans le temps

Frontier Language Model Intelligence, Over Time

Artificial Analysis Intelligence Index v4.0 incorporates 10 evaluations: GDPval-AA, τ^2 -Bench Telecom, Terminal-Bench Hard, SciCode, AA-LCR, AA-Omniscience, IFBench, Humanity's Last Exam, GPQA Diamond, CritPt

Anthropic DeepSeek Google Meta Mistral OpenAI



<https://artificialanalysis.ai/#frontier-language-model-intelligence-over-time#frontier-language-model-intelligence-over-time> 20 janvier 2026

Fuite de données

Apple a bloqué l'utilisation d'outils d'IA pour certains de ses employés, devenant ainsi la dernière grande entreprise à restreindre l'utilisation de plateformes d'IA générative sur le lieu de travail, en raison de la crainte que les employés ne divulguent des données internes sensibles.

Forbes, 21 mai 2023

<https://www.forbes.fr/technologie/apple-rejoint-une-liste-croissante-dentreprises-qui-repriment-lutilisation-de-chatgpt-par-leurs-employes/>

... Production

Intelligence artificielle : un accord de partenariat entre « Le Monde » et OpenAI
Cet accord pluriannuel, le premier entre un média français et un acteur majeur de l'IA, permettra à la société de s'appuyer sur le corpus du journal pour établir et fiabiliser les réponses de son outil ChatGPT, moyennant une source significative de revenus supplémentaires.

Le Monde, 13 mars 2024

https://www.lemonde.fr/le-monde-et-vous/article/2024/03/13/intelligence-artificielle-un-accord-de-partenariat-entre-le-monde-et-openai_6221836_6065879.html

... Protection ?

OpenAI is now fighting a court order to preserve all ChatGPT user logs—including deleted chats and sensitive chats logged through its API business offering—after news organizations suing over copyright claims accused the AI company of destroying evidence.

Ars Technica, 06 juin 2025

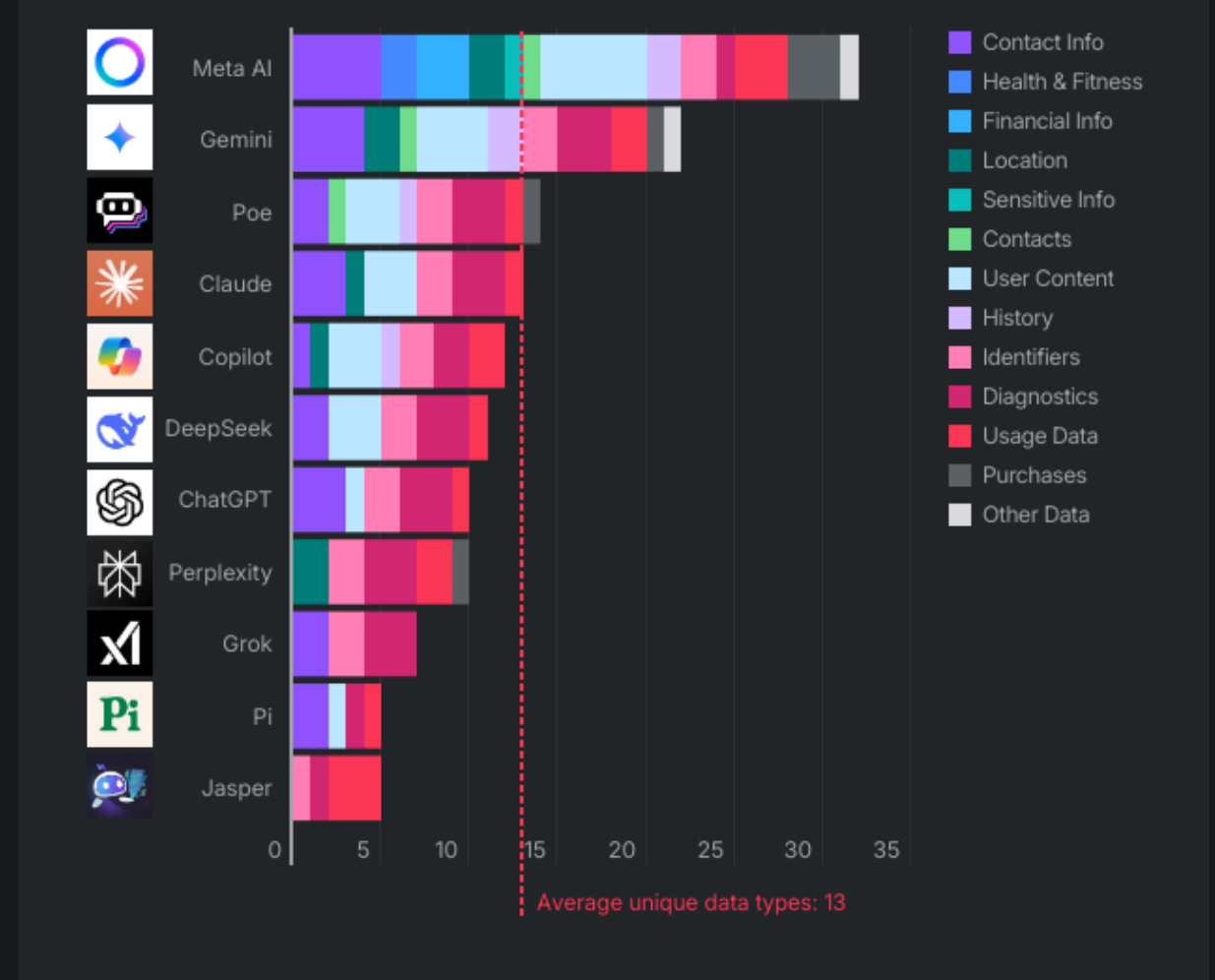
<https://arstechnica.com/tech-policy/2025/06/openai-says-court-forcing-it-to-save-all-chatgpt-logs-is-a-privacy-nightmare/>

Services gratuits d'IA générative : collecte de données

DATA UPDATE: MAY 20, 2025

Meta AI collects over 90% of user data types

Among AI chatbots, Meta AI collects the most user data, with 32 out of 35 data types, which is more than double the average of 13.

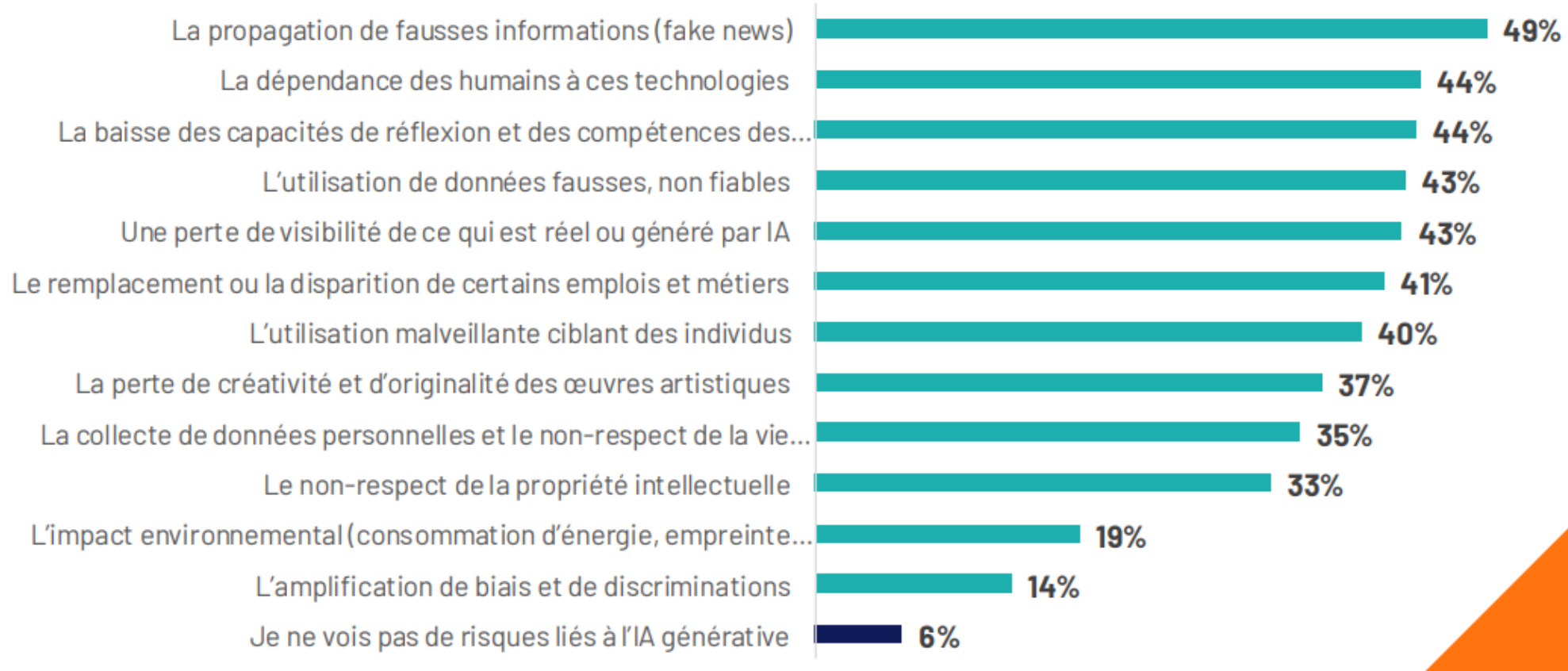


This image is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported license - <https://creativecommons.org/licenses/by-nc-sa/3.0/>



Etude Ipsos « L'usage de l'IA par les français » fév. 2025

Selon vous, quels sont les principaux risques liés à l'utilisation des IA génératives ?



Base = 155 utilisateurs professionnels / 328 utilisateurs privés

[L'usage de l'intelligence artificielle par les Français](#), février 2025

Pendant la ruée vers l'or vendez des pelles et des pioches

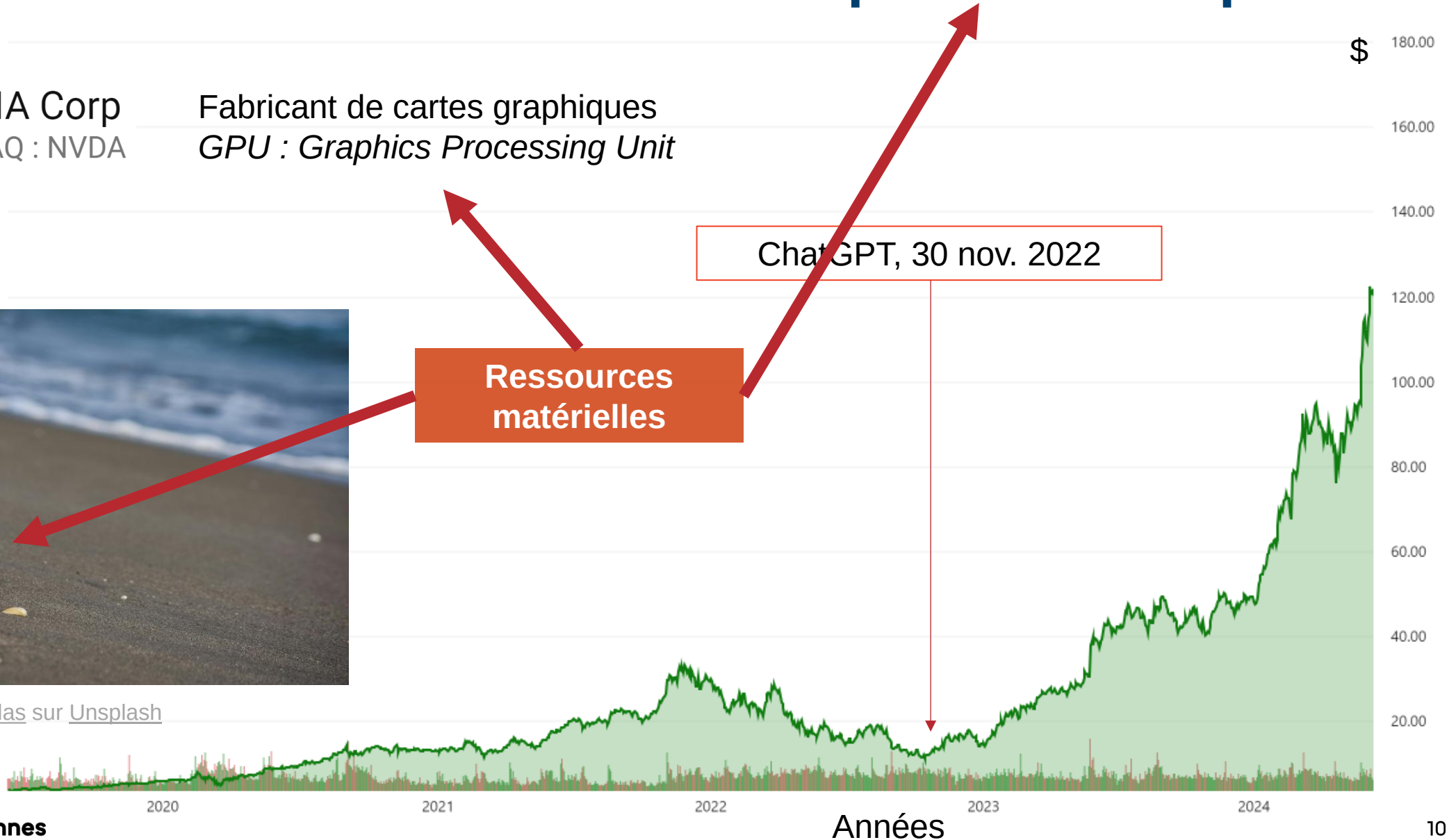


NVIDIA Corp
NASDAQ : NVDA

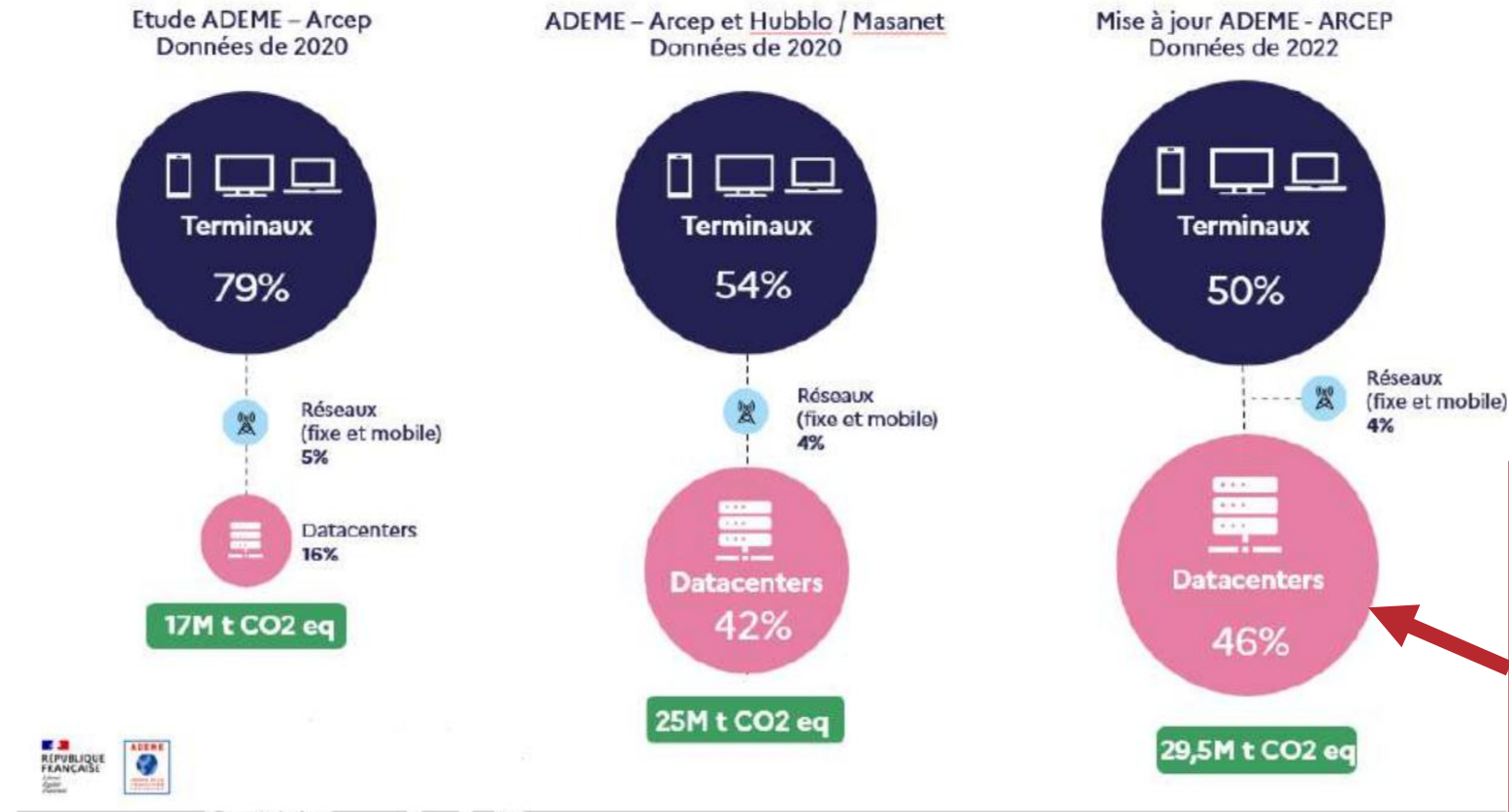
Fabricant de cartes graphiques
GPU : Graphics Processing Unit



Photo de [Alejandro Alas](#) sur [Unsplash](#)



Impact environnemental du numérique en France (actualisé)



En prenant en compte les data centers hébergés à l'étranger mais /\ datacenters IA récents non pris en compte

Figure 4 – Comparaison des résultats avec les études précédentes – par tier

Trajectoire énergétique : prospective

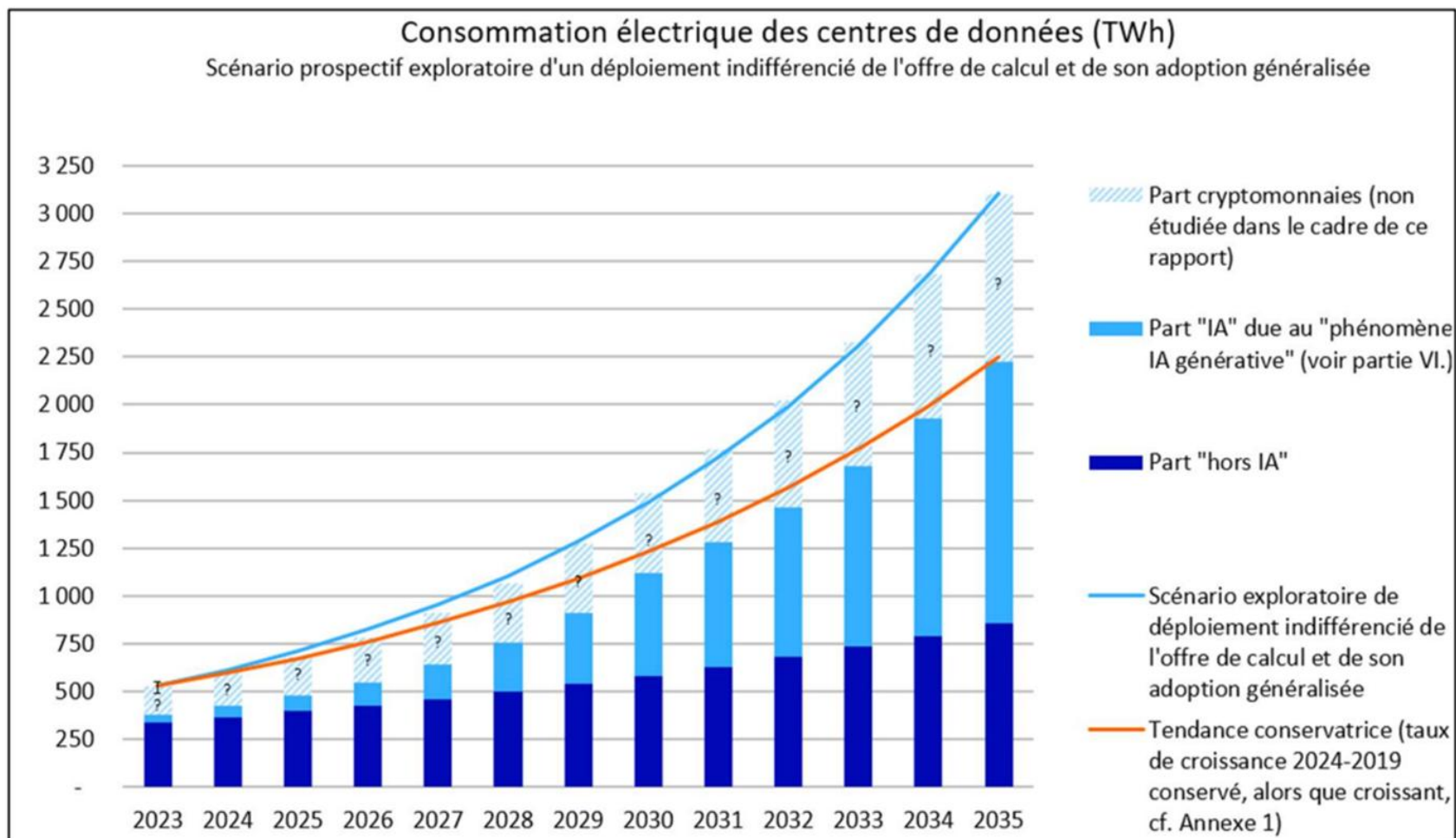


Figure 10 - Consommation d'électricité des centres de données en phase d'usage (TWh) de notre scénario prospectif et exploratoire d'un déploiement indifférencié de l'offre de calcul et de son adoption généralisée en comparaison à une tendance conservatrice. Source : (The Shift Project, 2025a)

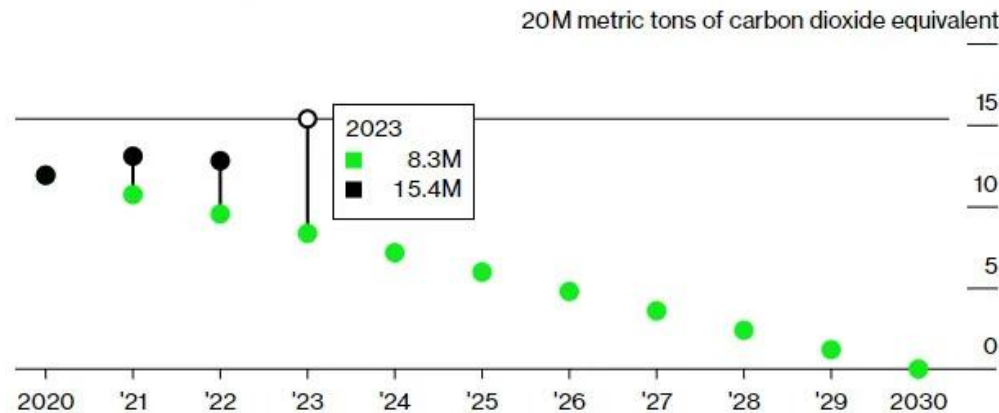
Microsoft's AI Push Imperils Climate Goal as Carbon Emissions Jump 30%

“The company’s goal to be carbon negative by 2030 is harder to reach, but President Brad Smith says the good AI can do for the world will outweigh its environmental impact.”

Microsoft's Emissions

Artificial intelligence is putting the tech giant's climate goals in peril

● Climate plan (simulated) ● Actual



Source: Microsoft (Scope 1, 2 and 3 "management criteria" data)

Note: Green dots represent linear decline to carbon negative goal.

<https://www.itforbusiness.fr/europe-chine-carbone-des-nuages-gris-au-dessus-de-microsoft-76735>

<https://financialpost.com/pmnbusiness-pmn/microsofts-ai-push-imperils-climate-goal-as-carbon-emissions-jump-30>

Impact environnemental de l'IA générative

IA générative vidéo = 30 fois génération image = 2000 fois génération texte
Charge smartphone à 100% = 2 images générées

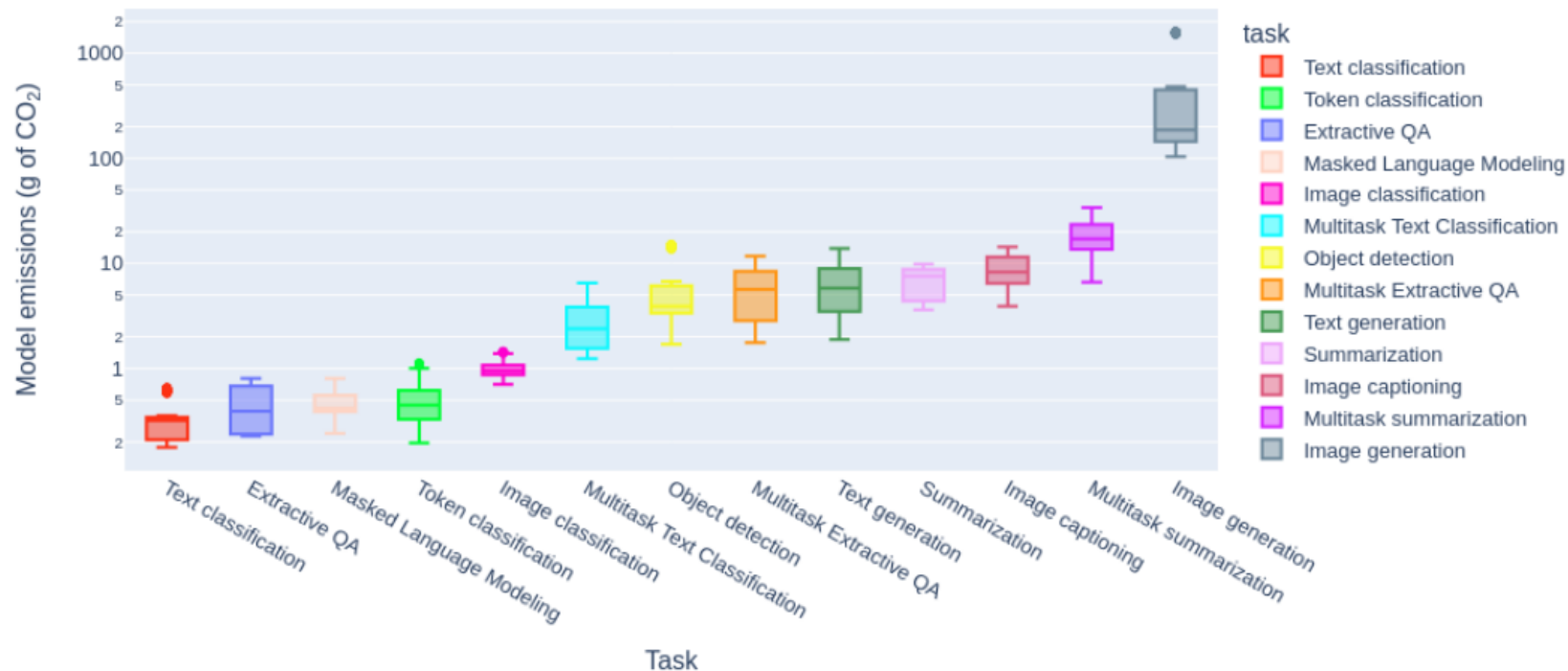


Fig. 1. The tasks examined in our study and the average quantity of carbon emissions they produced (in g of CO₂) for 1,000 queries.
N.B. The y axis is in logarithmic scale.

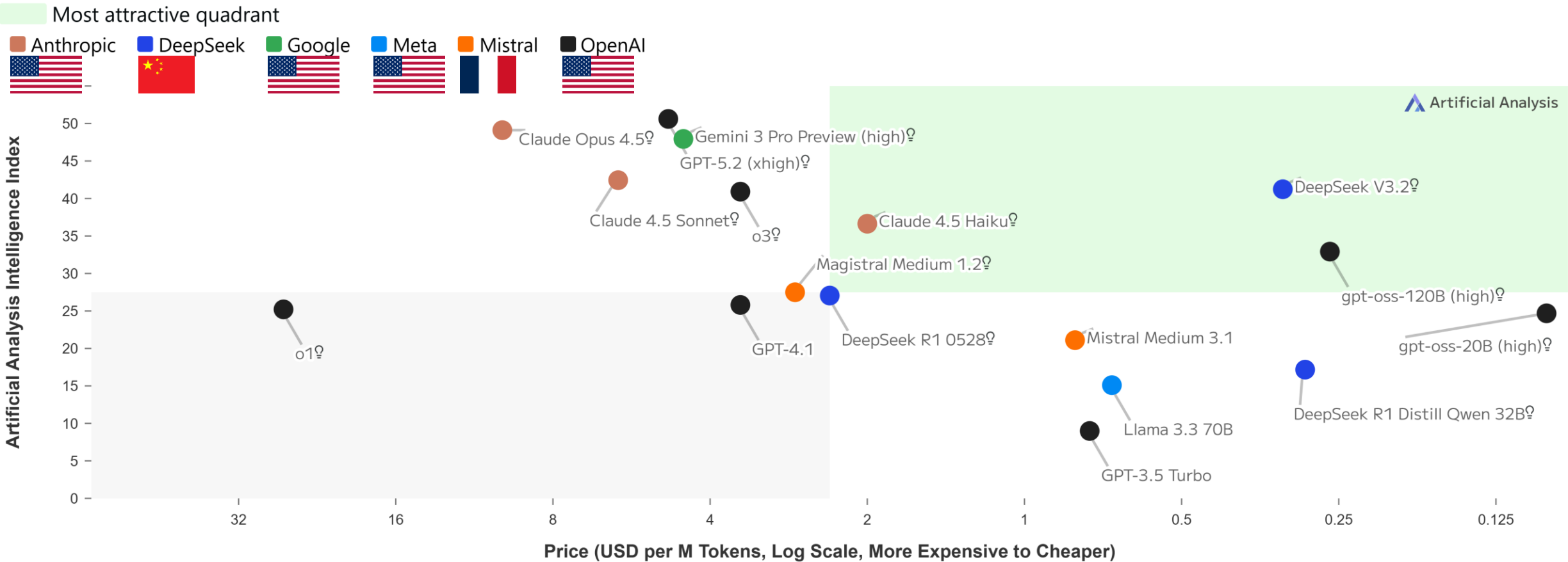
Luccioni, S., Jernite, Y., & Strubell, E. (2024, June). Power Hungry Processing: Watts Driving the Cost of AI Deployment? *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. [doi:10.1145/3630106.3658542](https://doi.org/10.1145/3630106.3658542)

Delavande, J., Pierrard, R., & Luccioni, S. (2025, September). Video Killed the Energy Budget: Characterizing the Latency and Power Regimes of Open Text-to-Video Models. *arXiv [Cs.LG]*. Retrieved from <http://arxiv.org/abs/2509.19222>

Performance et impact (budgétaire, environnemental)

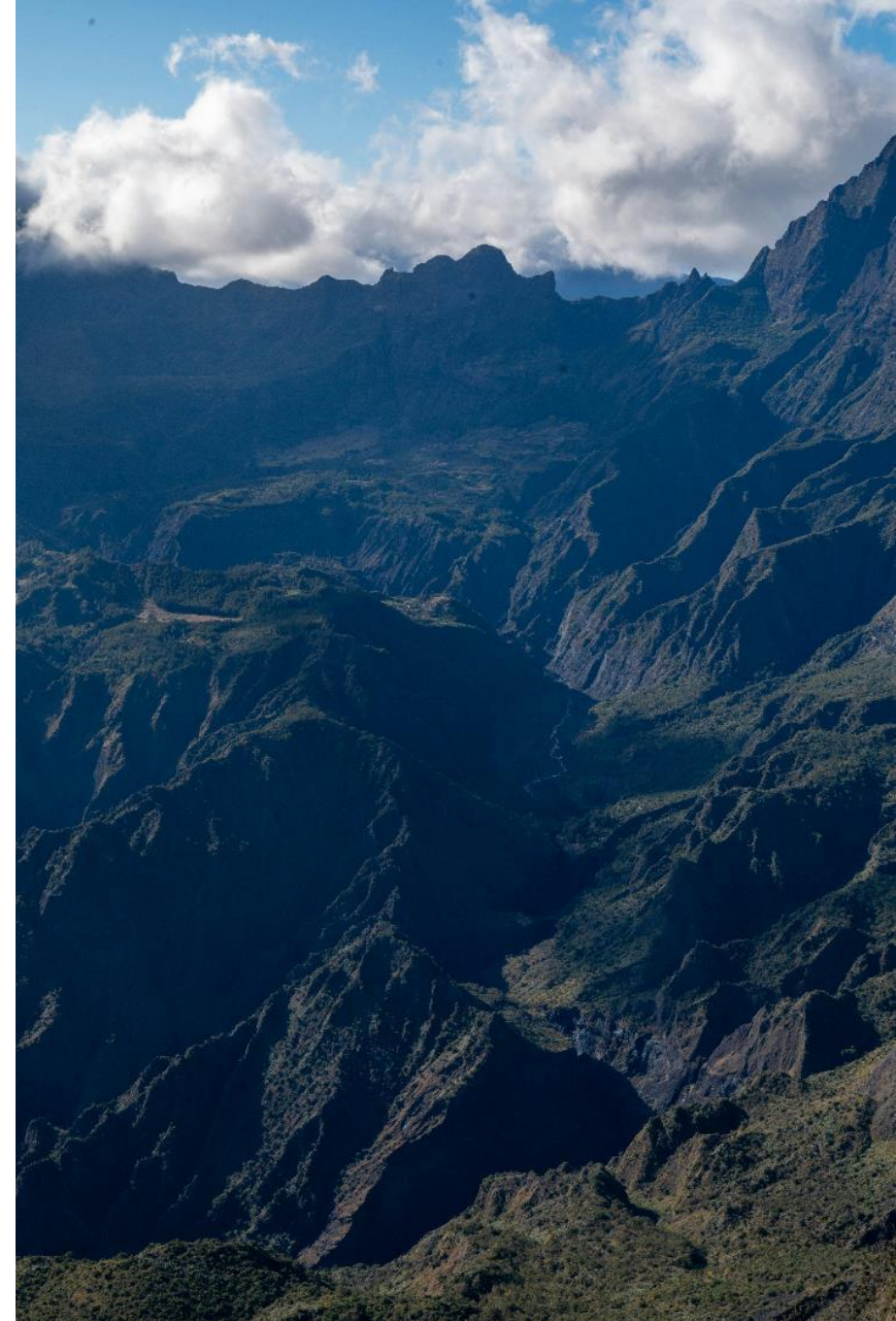
Intelligence vs. Price (Log Scale)

Artificial Analysis Intelligence Index; Price: USD per 1M Tokens; Inspired by prior analysis by Swyx



Stratégie numérique

- Expérimentation RAGaRenn
- Mutualisation projet ILaaS



RAG = Retrieval Augmented Generation

Sans RAG

Avec RAG

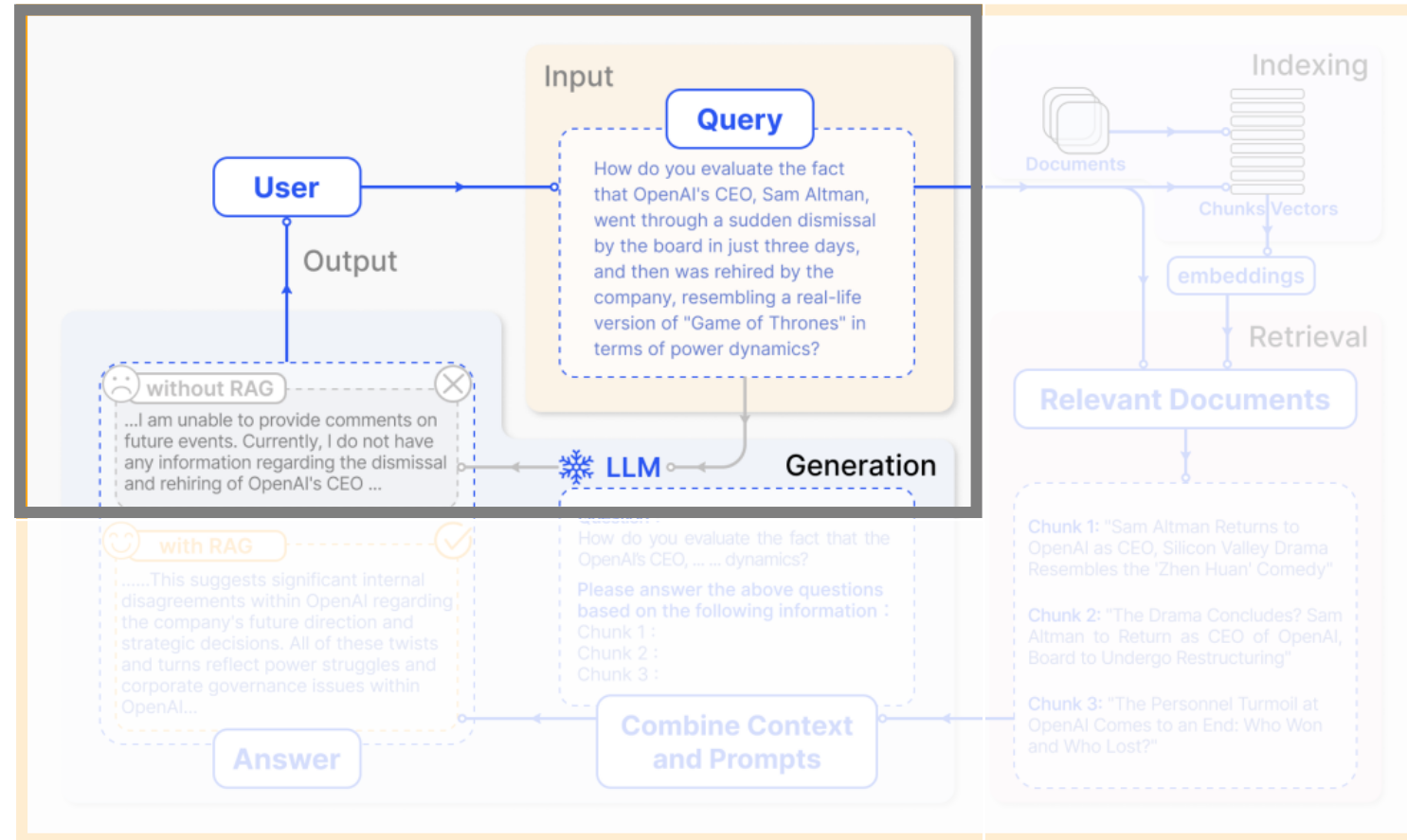


Fig. 2. A representative instance of the RAG process applied to question answering. It mainly consists of 3 steps. 1) Indexing. Documents are split into chunks, encoded into vectors, and stored in a vector database. 2) Retrieval. Retrieve the Top k chunks most relevant to the question based on semantic similarity. 3) Generation. Input the original question and the retrieved chunks together into LLM to generate the final answer.

RAGaRenn : utiliser des sources connues



Sans source (LLM seul)



You

Comment utiliser ma carte Korrigo ?

OI **Mixtral:Latest** 27/03/2024 23:07

La carte Korrigo est une carte de transport utilisée dans certaines régions de France, notamment en Bretagne. Pour utiliser votre carte Korrigo, vous devrez suivre ces étapes :

1. Charger votre carte : Avant de pouvoir utiliser votre carte Korrigo, vous devez la charger avec des titres de transport valides. Vous pouvez le faire en ligne sur le site web de Korrigo, dans une borne de rechargement automatique ou dans un point de vente Korrigo.
2. Valider votre carte : Lorsque vous montez à bord d'un bus ou d'un train, vous devez valider votre carte en la passant sur le lecteur de cartes situé à proximité des portes d'entrée. La validation permet



Avec source (RAG)



You



<https://docinfo.univ-...>
Document

Comment utiliser ma carte Korrigo ?

OI **Mixtral:Latest**

Pour utiliser votre carte Korrigo, vous devez l'associer à votre compte informatique Sésame sur un copieur Toshiba en passant votre carte et en saisissant vos identifiant et mot de passe Sésame. Si vous changez de carte, il faudra renouveler cette opération. Sans carte, vous pouvez toujours imprimer en saisissant vos identifiants Sésame à chaque utilisation d'un copieur.

La carte Korrigo vous permet également d'accéder à différents services universitaires tels que le service de photocopie et d'impression, l'inscription au SIUAPS et l'inscription aux BU. Lors de sa première création, la carte Korrigo est offerte, mais en cas de perte, vol ou détérioration, son renouvellement sera payant.



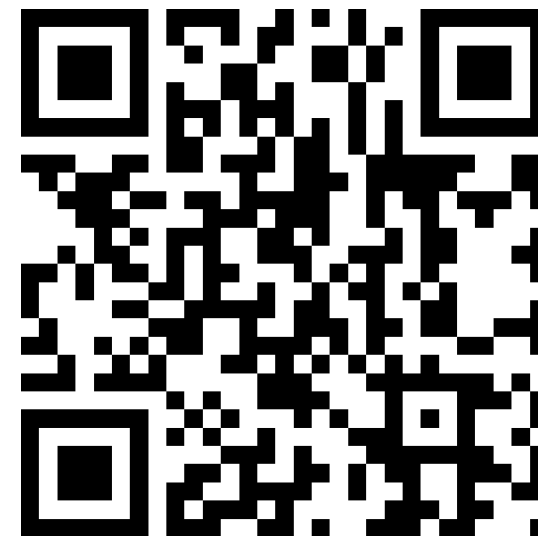
RAGaRenn : utiliser des sources connues

Déploiement applicatif Open Web UI

- **105 instances pour des groupes d'utilisateurs**
 - Majorité : gestion locale (réglages, comptes, données)
 - 31 authentification unique ESR (Renater) dont 1 ouverte à tout l'ESR
 - 1 authentification EduGain (international)

Utilisateurs (mars 2024 - 20 janvier 2026) :

- **2570 inscrits dans l'instance principale ESR**
- **4449 inscrits issus de 352 institutions**
- **155369 requêtes sur 2025**



Stratégie numérique : IA générative

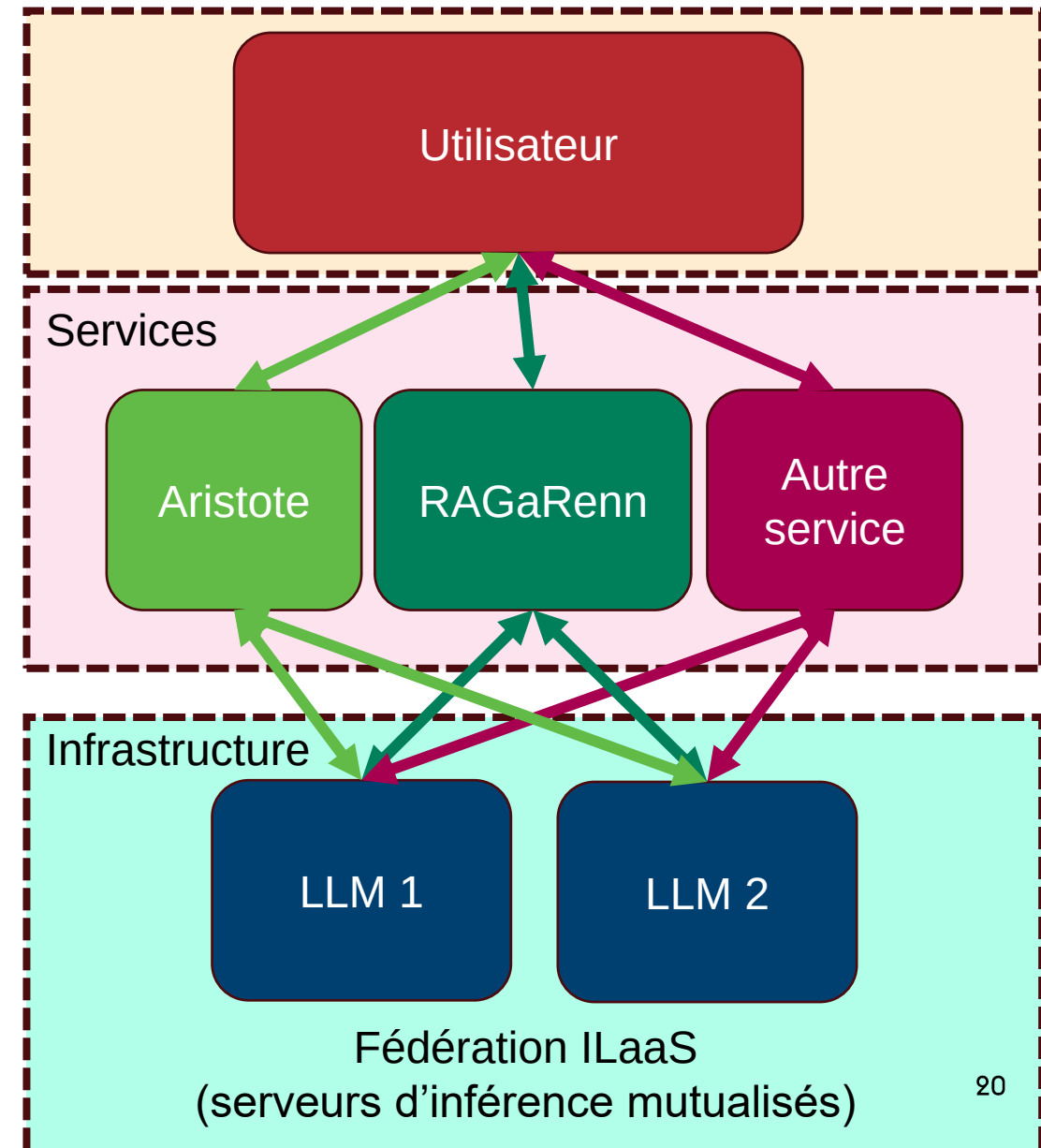
Principe : séparer l'accès à un LLM et l'accès à des services utilisant un LLM

Approche agnostique

- Choix du/des LLM(s) utilisé(s)
- Compatible avec différentes infrastructures (DC labellisés, Méso, SecNumCloud, etc.)

Maitrise des flux de données, des impacts environnementaux & budgétaires

- Stockage et flux de données sécurisés
- Facturations indépendantes (services // LLM) et maîtrisées
- Mesure des usages réels & impacts associés



Collaboration ESR

Aristote, Centrale Supélec Paris Saclay (2023)

- Aristote-dispatcher : mutualiser les GPU en gérant les priorités
- <https://github.com/CentraleSupelec/aristote-dispatcher>

RAGaRenn, Université de Rennes (2024)

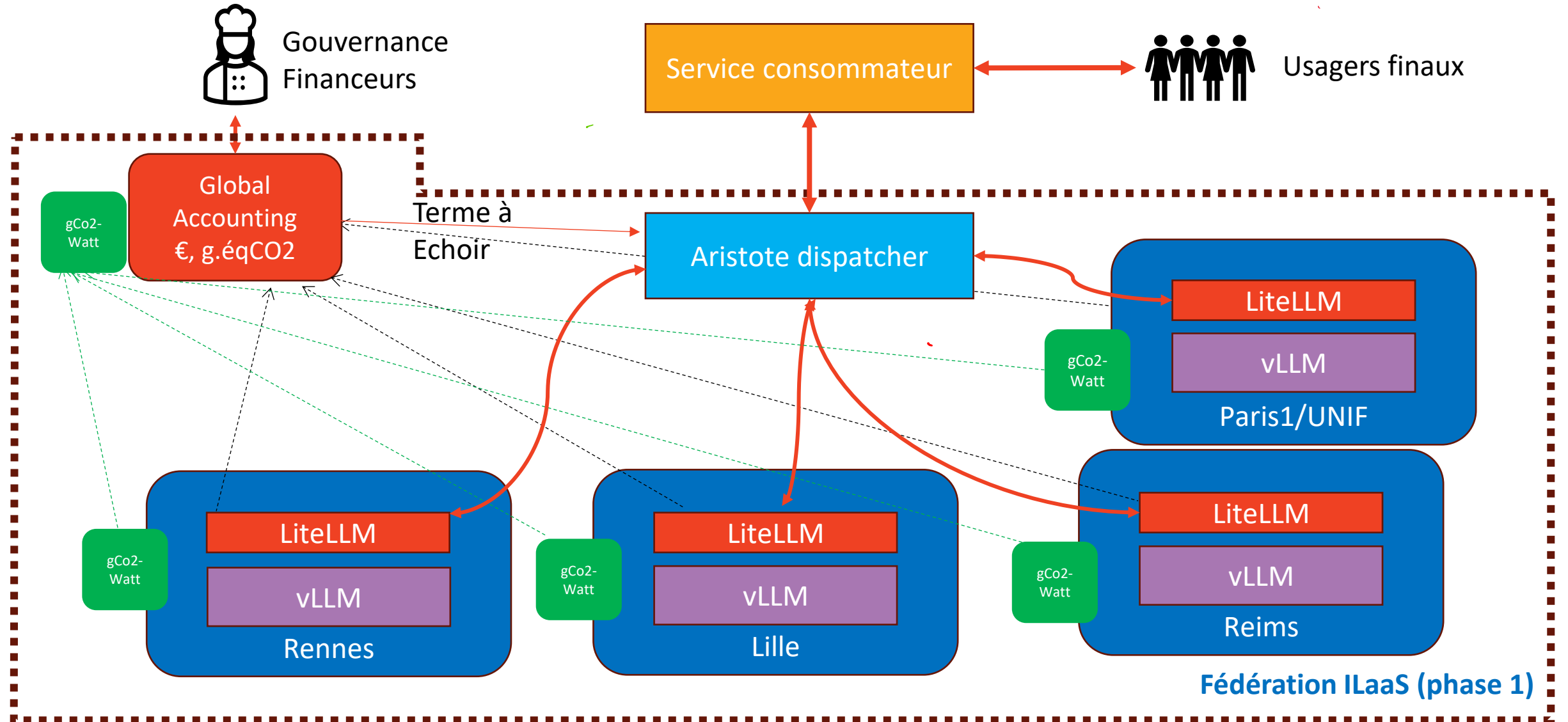
- Déploiement : briques open source ([OpenWebUI](#), [vLLM](#), [ollama](#))
- Hébergement local : datacenter régional labellisé Eskemm Data
- Observations sur les usages (qualitatif & quantitatif)

Infrastructures : projet ILaaS*

**Mutualiser les
infrastructures numériques
des partenaires pour les
mettre à disposition
d'utilisateurs
bénéficiaires**



Projet ILaaS phase 1



Problématiques partagées

Soutenabilité

- Équilibre budgétaire et sobriété numérique : impact environnemental de l'inférence selon usages

Résilience

- Qualité de service, lissage des pics, gestion des indisponibilités

Confiance

- Niveau de confiance partagé, sécurisation raisonnable, facilite l'émergence de nouveaux services

Souveraineté

- Ouvrir les choix possibles, améliorer la robustesse

Projet ILaaS phase 2

Phase 1 : Université de Rennes (porteur),
Université Reims Champagne Ardennes,
Université Paris I Panthéon Sorbonne,
Université de Lille, Centrale Supélec Paris
Saclay

Phase 2 intéressés : Université de Nantes,
Université d'Angers, Le Mans Université,
Université Paris Cité, Université de Lorraine,
Université Cote d'Azur, Université de
Strasbourg, Université de Haute-Alsace,
Université d'Orléans, Université de Tours, La
Rochelle Université, Université Marie et
Louis Pasteur, UPF, UTC, UPJV, CY Cergy
Paris Université, Polytechnique, EHESS, Aix-
Marseille Université, Université de
Bourgogne Europe... **Insérez le nom de
votre établissement ici**



Photo de [Philippe Bout](#) sur [Unsplash](#)

Comment rejoindre ou utiliser ?

Entrer dans la fédération = contribuer

- Différentes possibilités de contribution : héberger un nœud de calcul, produire du code, aider à conduire le projet, documenter, mener des études ...
- Engagement au niveau de la gouvernance : VP (numérique), Directeur (DSI)
- Signer l'accord de consortium entre partenaires

Bénéficier de la fédération = consommer, payer

- Approbation du consortium pour accéder au(x) service(s) de la fédération
- Tout contributeur peut être consommateur en « circuit court » = les requêtes locales sont prioritaires vs celles qui sont externes
- Services : LLM, Speech-to-text, ... Bientôt : applications
- Facturation & relations contractuelles : étude en cours

Conclusion

IA : tendance de fond

IA générative : viser utile, utilisable, utilisée... Sobriété numérique

Stratégie « juste nécessaire » à établir

- Collaborer : communautés internes, acteurs publics
- Mutualisation infrastructures : fédération ILaaS
- Déterminer les cas d'usages pertinents
- Cadrage [charte & guide de recommandations](#) (cf. travaux ESR)

Agir & construire ensemble

- Expérimenter solutions et pratiques
- Partage retours, cas d'usages, réflexions

