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

26 novembre 2025
Open Source Days @UGA

Olivier WONG

Vice-président numérique

Université de Rennes

Président du bureau

VP-Num
L'association des Vice-présidents numérique dans l'Enseignement Supérieur

IA : Intelligence Artificielle

Enjeux du numérique dont l’IA

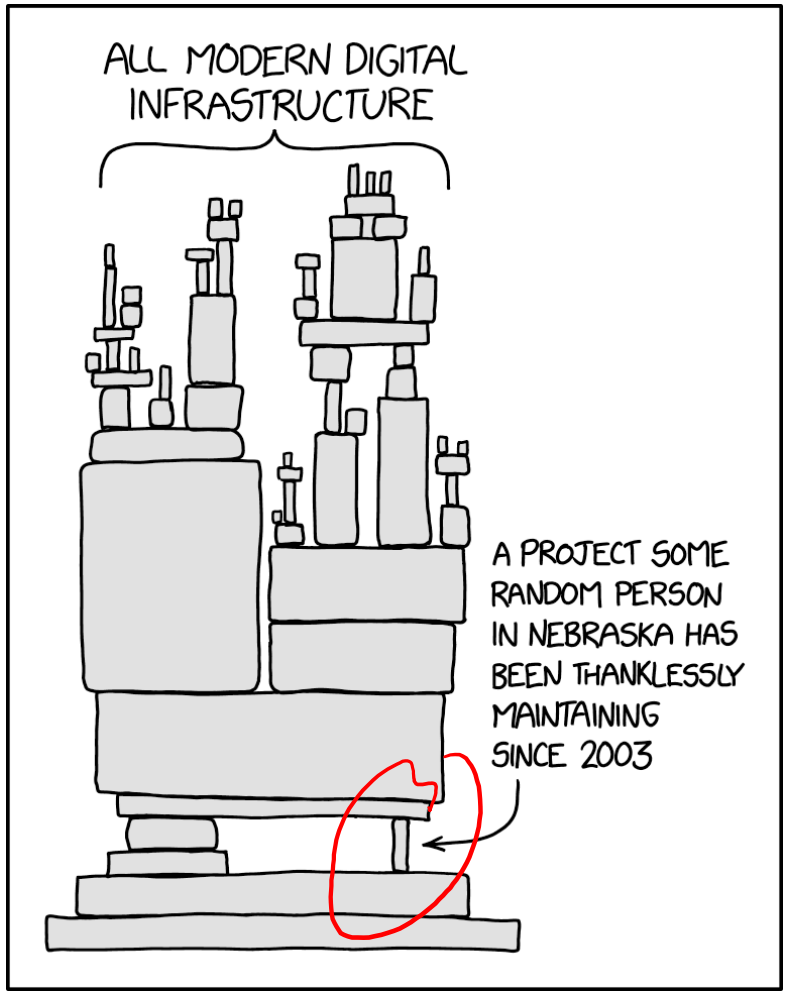
- Souveraineté
- Sécurisation des données
- Impact environnemental

Photo de James Toose sur Unsplash

Souveraineté numérique : capacité à faire des choix technologiques

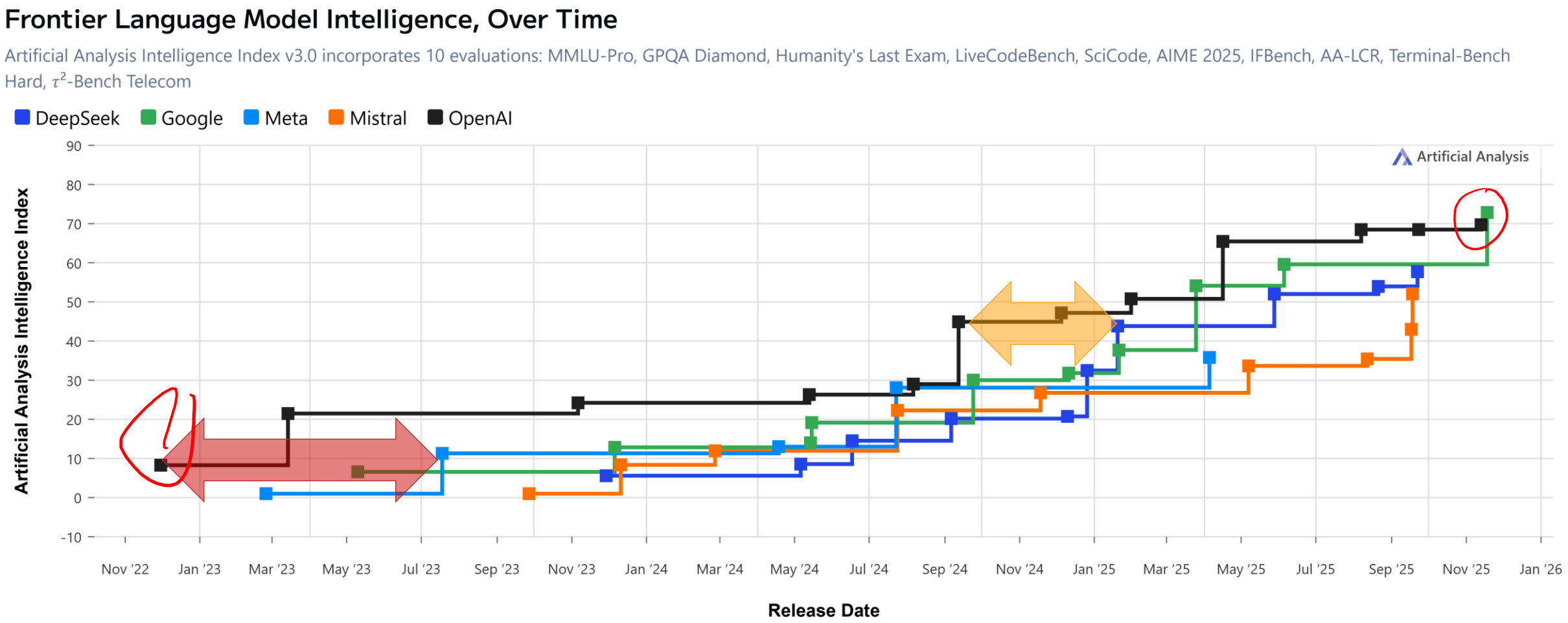
- **Choisir et diversifier nos dépendances**
- Infrastructures et matériel : Etats-Unis, Asie (Taiwan, Chine)
- Logiciels et services : marché mondial
- Usages : captation et ancrage par les plateformes

- **Quelles dépendances ?**
- Chaine d'approvisionnement
- Répartition de la valeur entre acteurs
- Déplacement des enjeux sur les données



Dependency https://xkcd.com/2347/

Performance des LLM de pointe dans le temps



https://artificialanalysis.ai/#frontier-language-model-intelligence-over-time 25 novembre 2025

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.

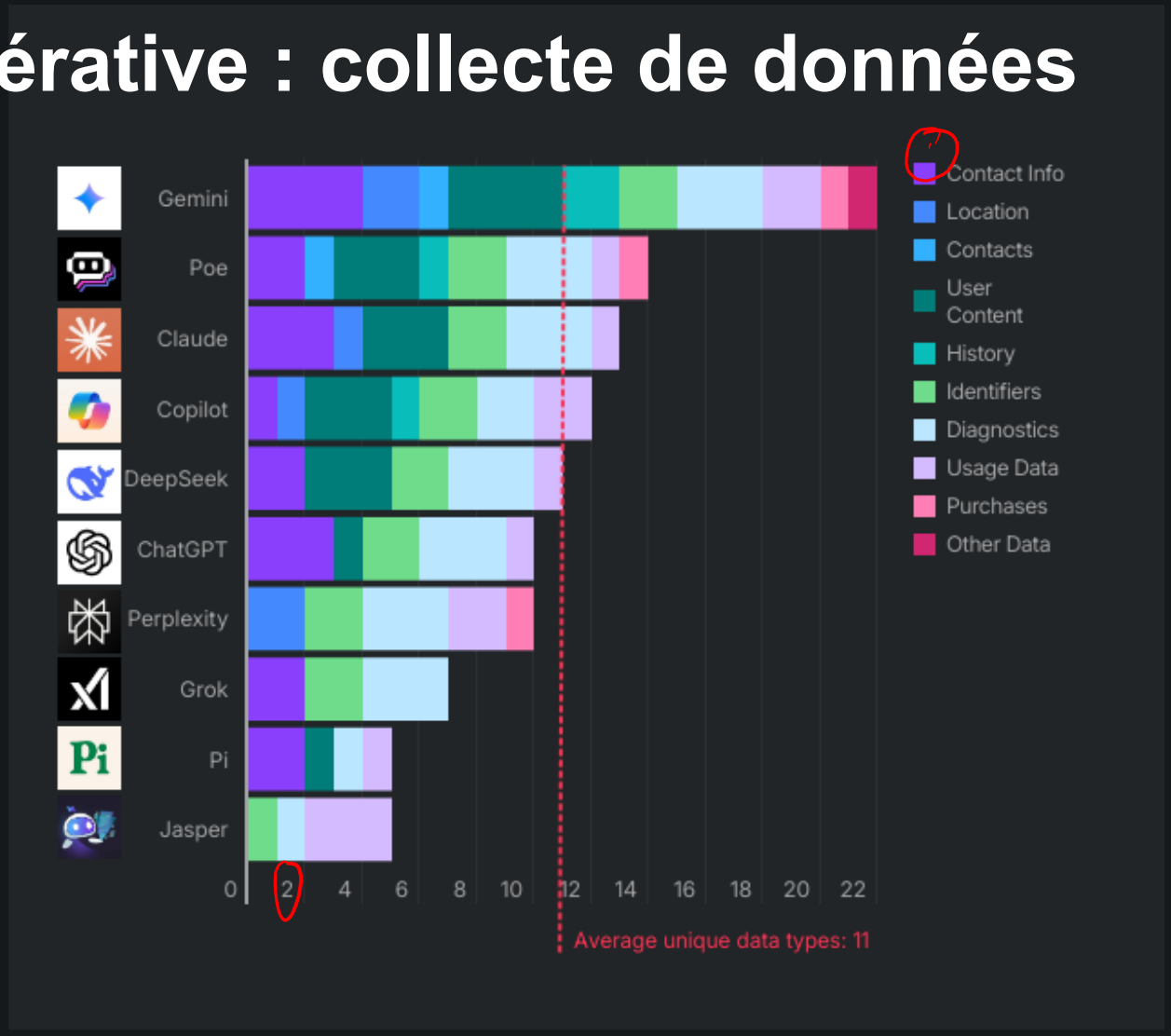
Ars Technica, 06 juin 2025
https://arstechnica.com/tech-policy/2025/06/openai-says-court-forcing-it-to-save-all-chatgpt-logs-is-a-privacy-nightmare/

Services gratuits d'IA générative : collecte de données

DATA COLLECTED: February 12, 2025

40% of popular AI chatbots collect user location

Google Gemini collects the most user data among AI chatbots, gathering 22 out of 35 types of data

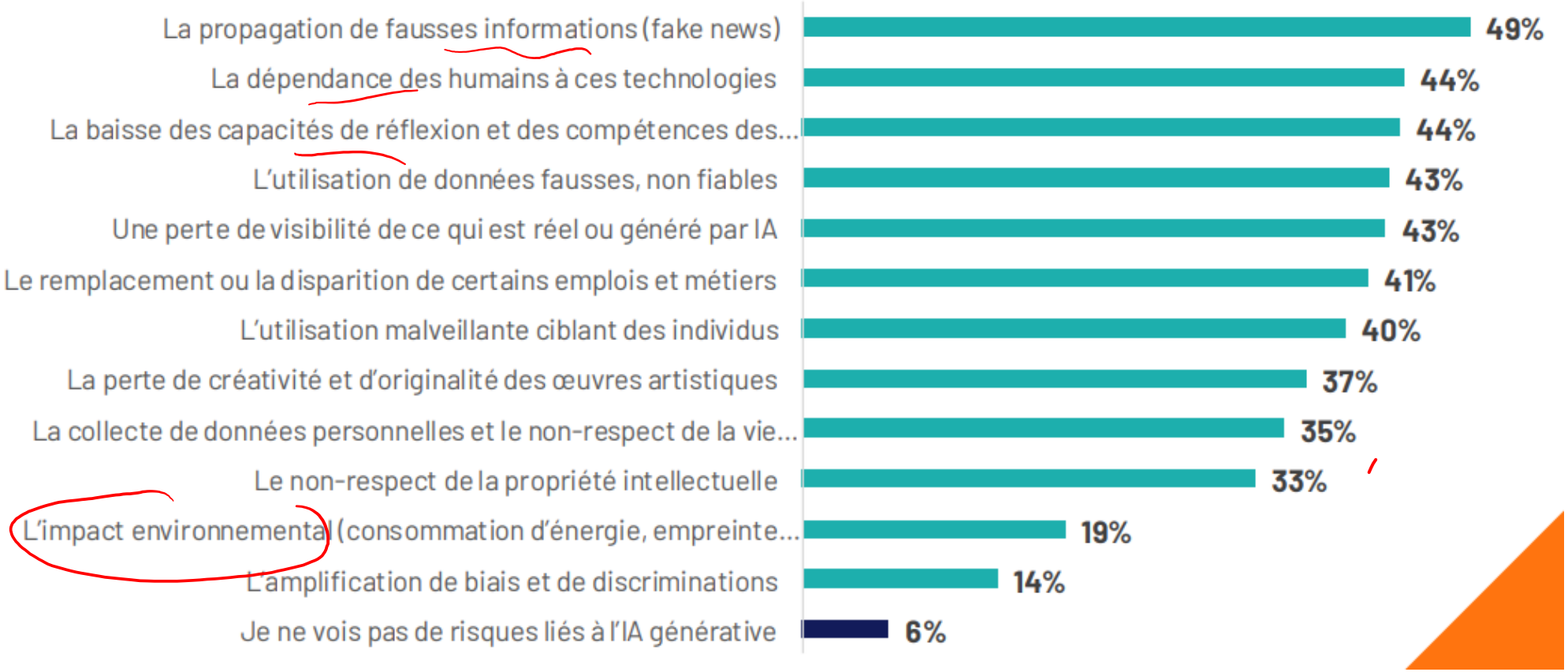


Surfshark®

https://surfshark.com/research/chart/ai-chatbots-privacy

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

Ressources matérielles

ChatGPT, 30 nov. 2022

Impact environnemental du numérique en France (actualisé)

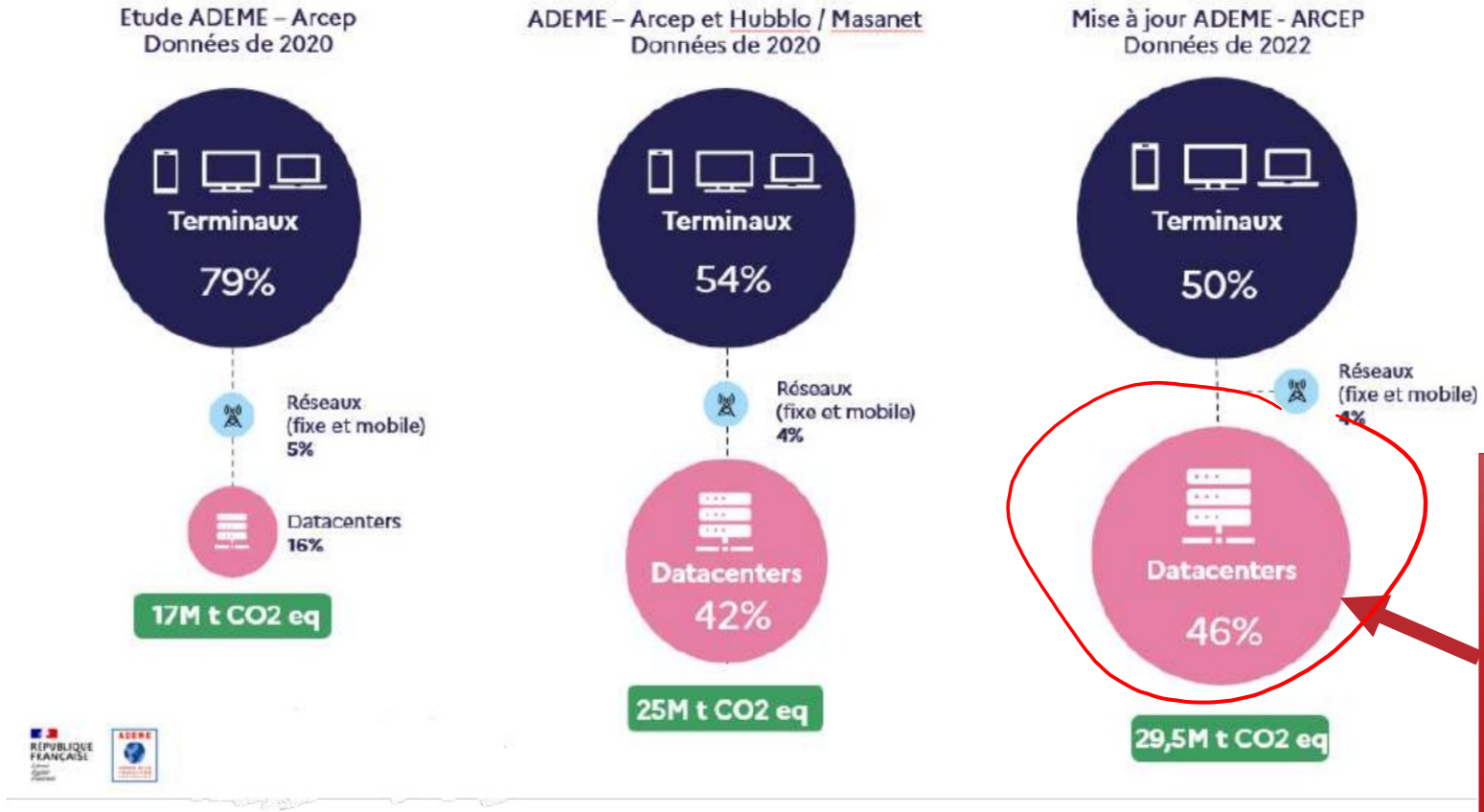


Figure 4 – Comparaison des résultats avec les études précédentes – par tier

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

L'empreinte environnementale du numérique
https://ecoresponsable.numerique.gouv.fr/docs/2024/etude-ademe-impacts-environnementaux-numerique.pdf

Trajectoire énergétique : prospective

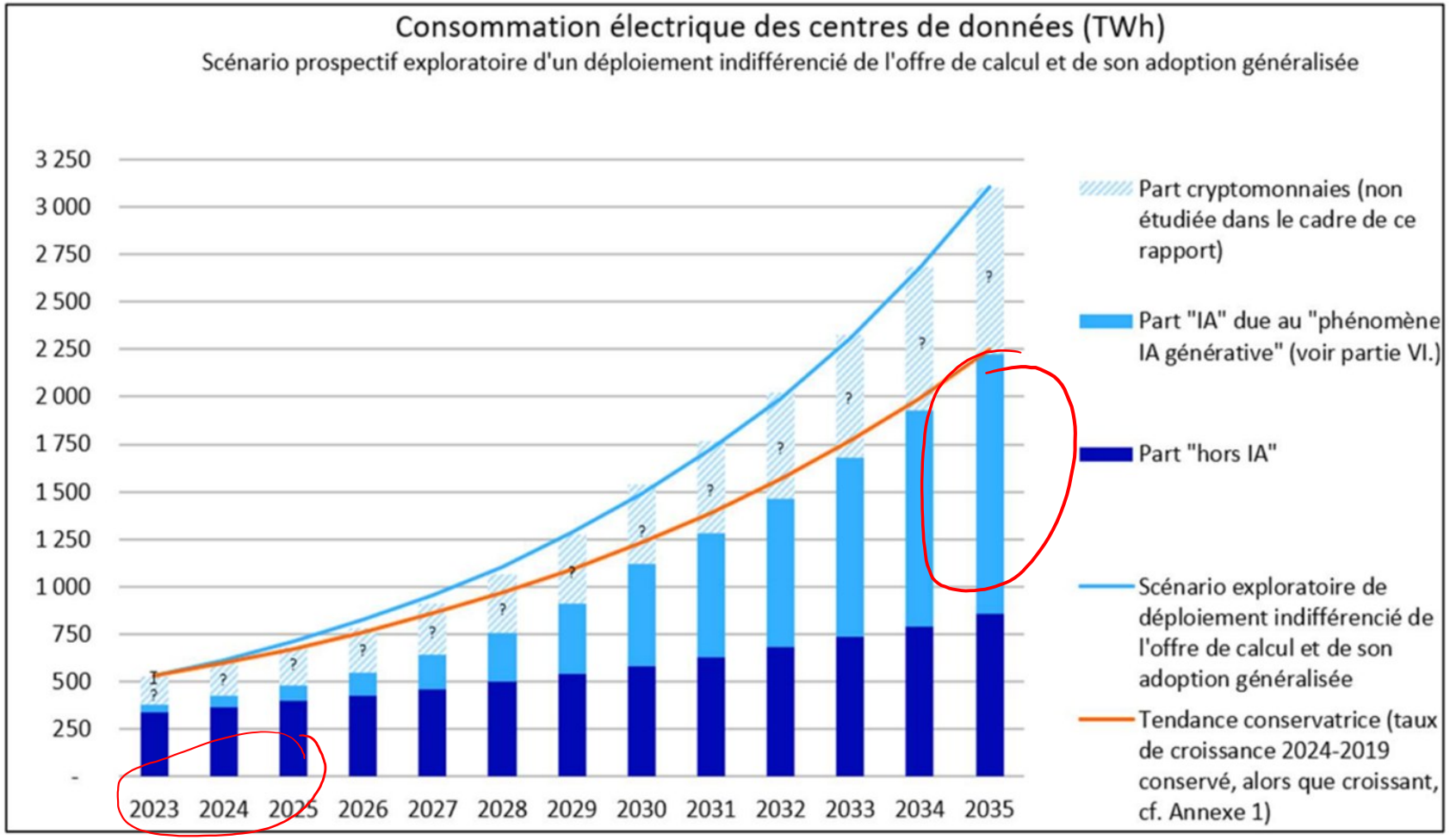
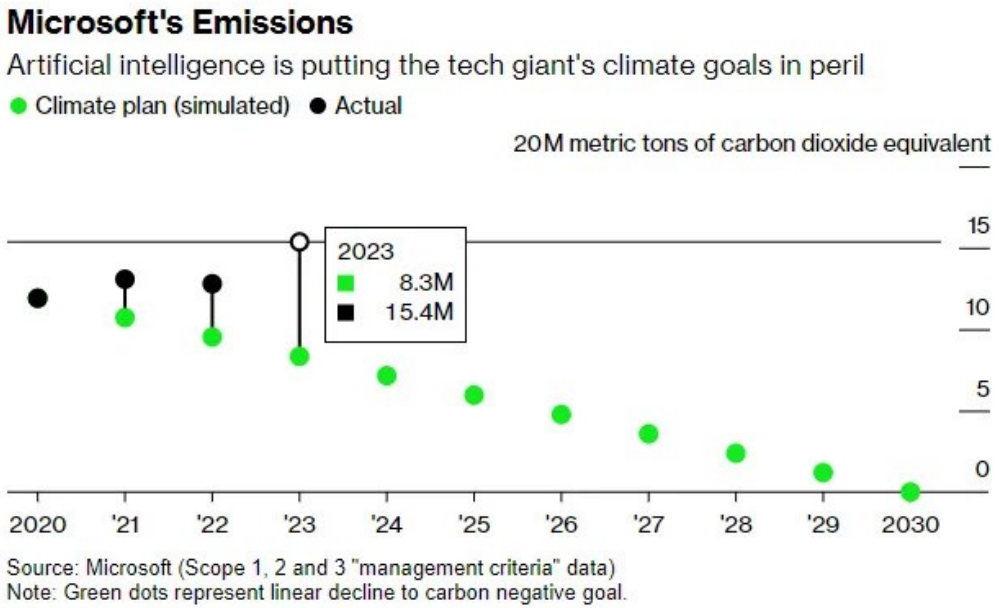


Figure 10 - Consommation d'électricité des centres de données en phase d'usage (TWh) de notre scénario prospectif et exploratoire d'un déploiement indifférencié de l'offre de calcul et de son adoption généralisée en comparaison à une tendance conservatrice. Source : (The Shift Project, 2025a)

Intelligence artificielle, données, calculs : le rapport final du Shift - The Shift Project, 1er octobre 2025

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.”



https://www.itforbusiness.fr/europe-chine-carbone-des-nuages-gris-au-dessus-de-microsoft-76735
https://financialpost.com/pmn/business-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

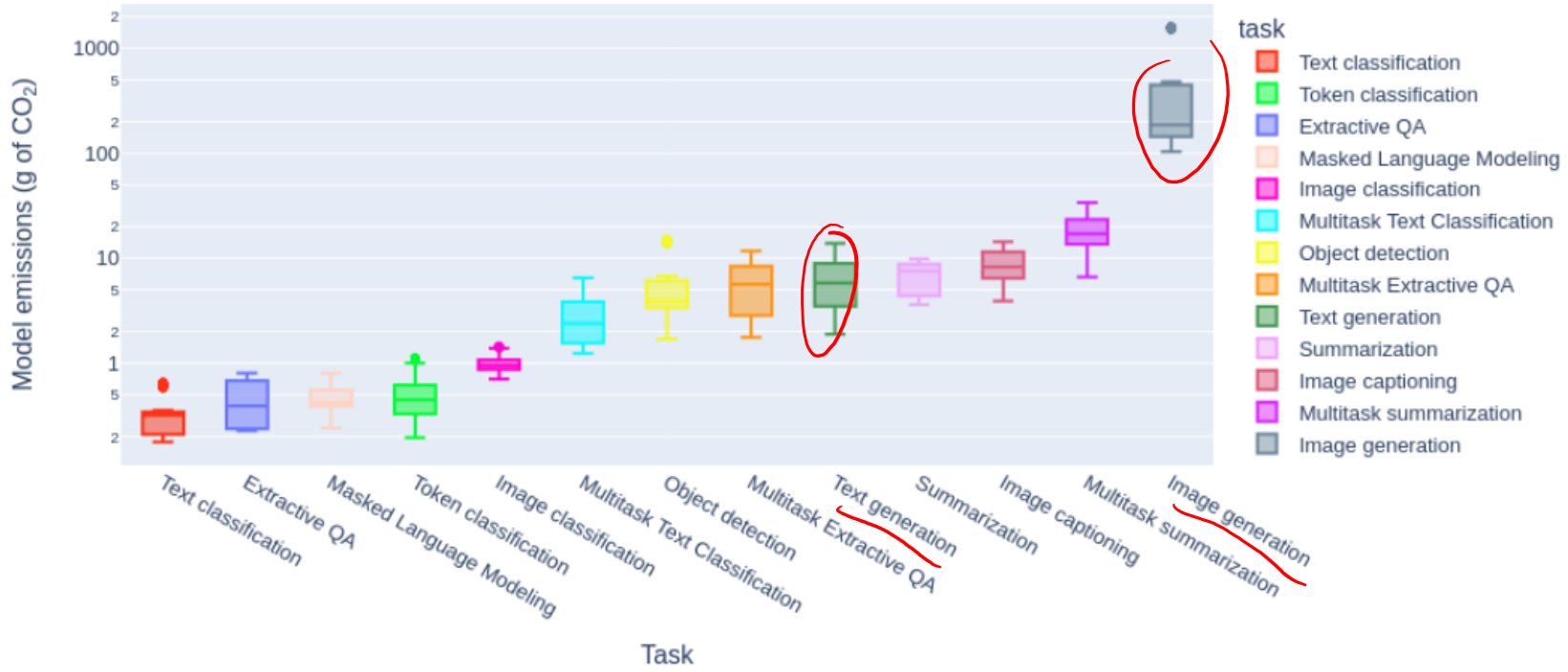


Fig. 1. The tasks examined in our study and the average quantity of carbon emissions they produced (in g of CO_2) for 1,000 queries.
N.B. The y axis is in logarithmic scale.

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

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

Analyse du cycle de vie : service frugal d'IA

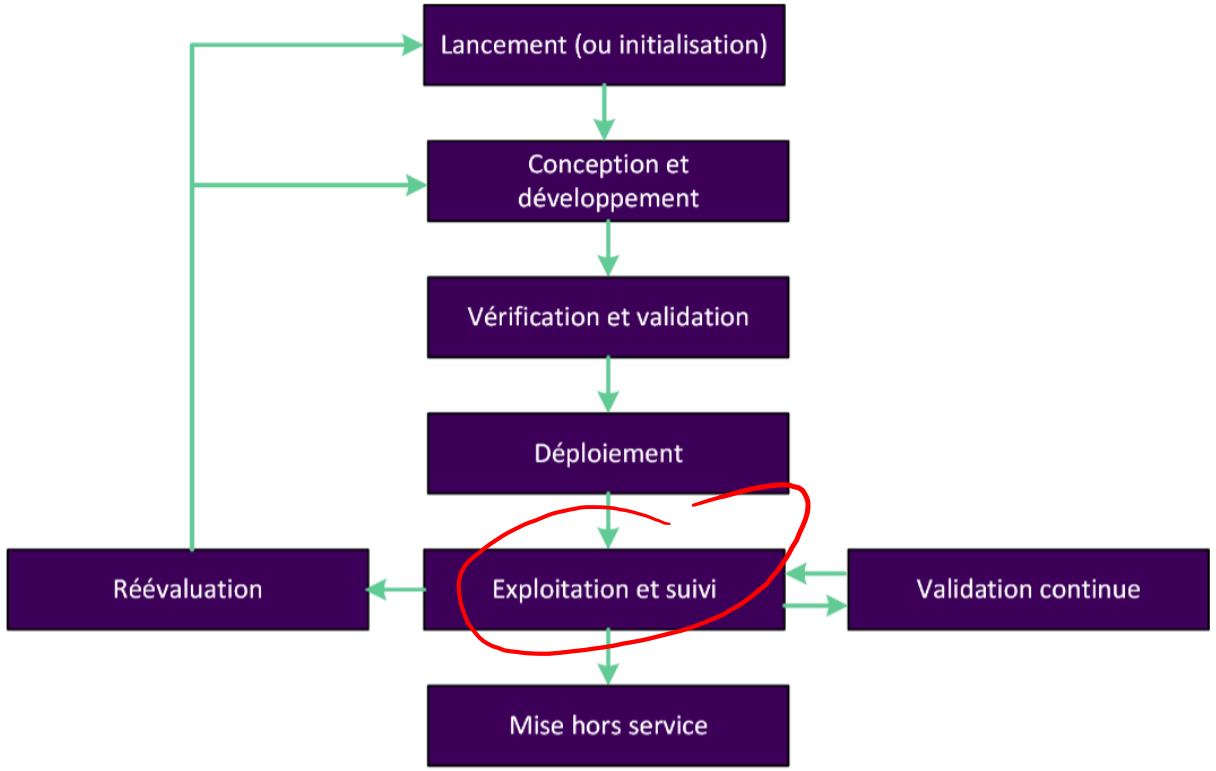


Figure 2 — Cycle de vie d'un système d'IA

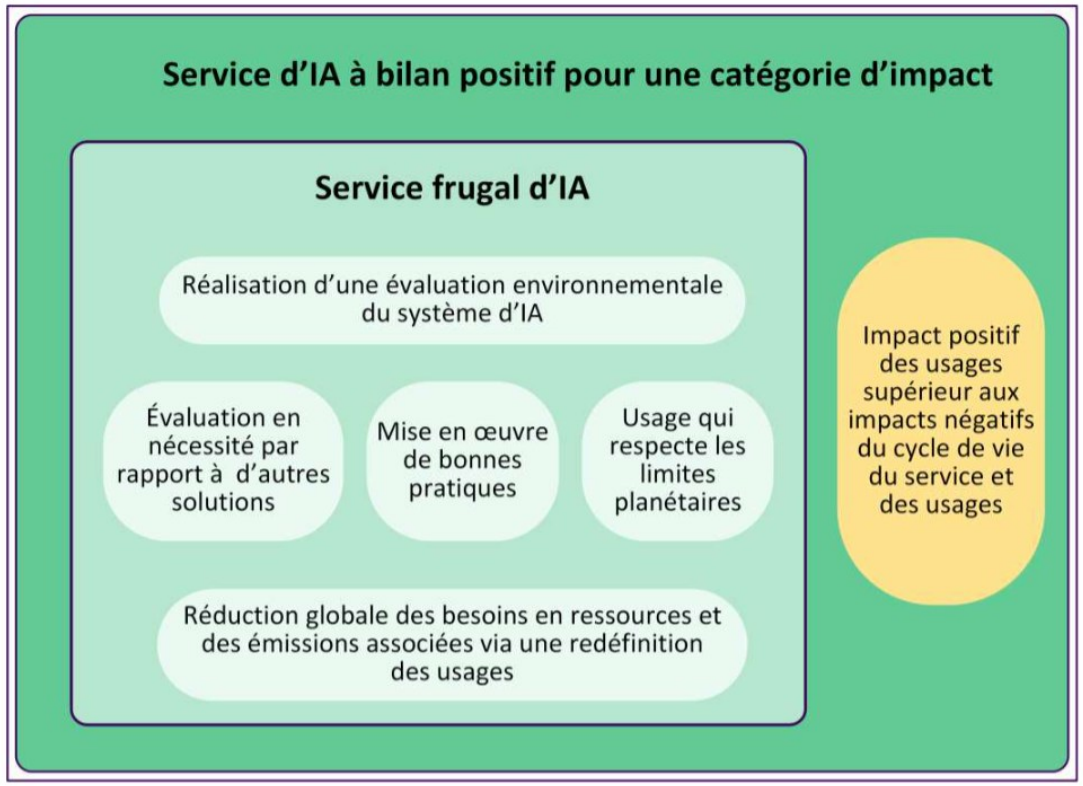


Figure 3 — Concepts de service frugal d'IA et de service à bilan positif sur une catégorie d'impact

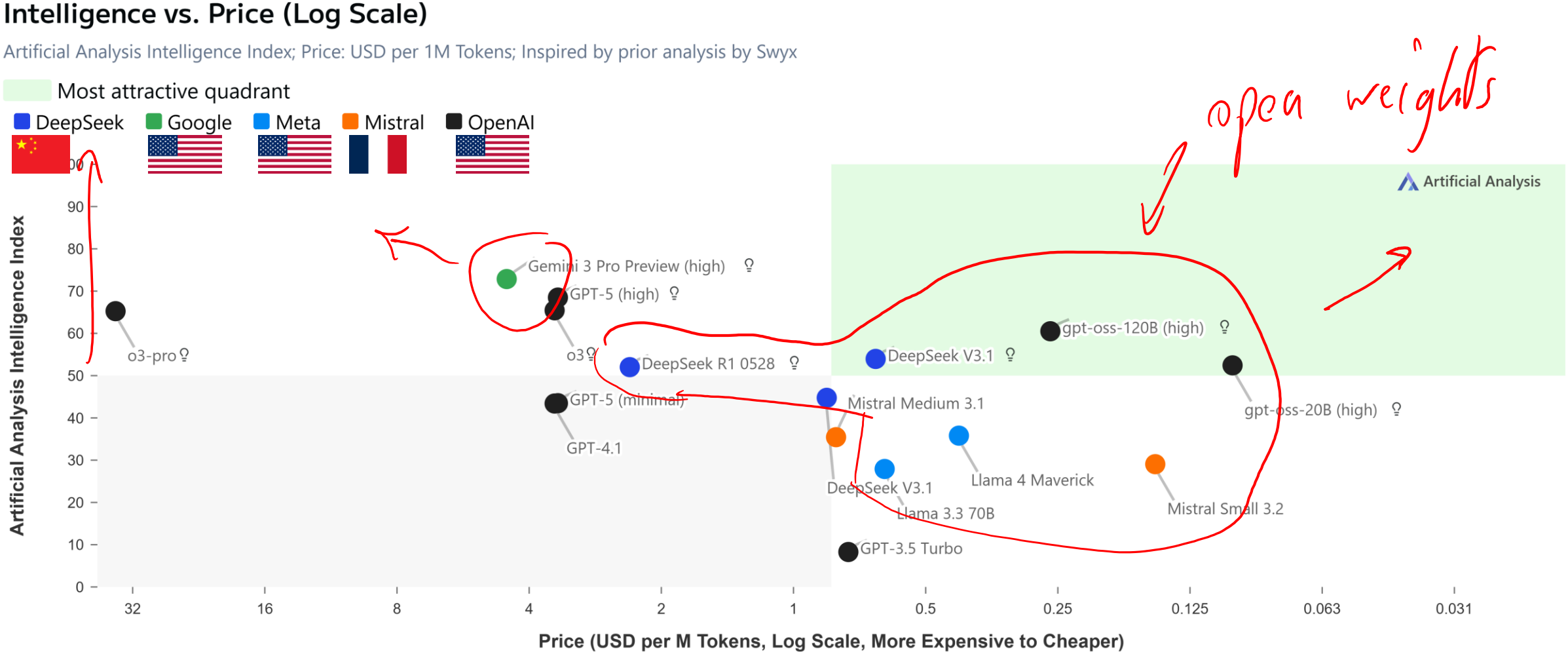
AFNOR SPEC 2314 Référentiel général pour l'IA frugale, juin 2024

Efficience, frugalité

	Définition	Notions connexes	Raisonnement	Approche	Précisions
Efficience	Aptitude à optimiser les moyens alloués pour atteindre un résultat défini	Efficacité, optimisation	En relatif/par unité d'usage Le besoin prime : optimisation d'une solution jugée celle répondant le mieux au besoin	Recherche d'un optimum local ou d'un compromis sur un niveau de résultat fortement contraint	Prise en compte des effets de premier ordre pour les minimiser Prise en compte des parties prenantes de l'IA
Frugalité	Aptitude à se contenter d'un niveau de résultat jugé suffisant en redéfinissant les usages et les besoins	Sobriété (ou *Sufficiency* [10] en anglais)	En global La contrainte sur les ressources prime : recherche de la solution utilisant le moins de ressources possible et apportant une réponse satisfaisante au besoin	Recherche d'un optimum global ou d'un compromis large sur un niveau de résultat, ce qui nécessite d'élargir ou d'assouplir le besoin	Prise en compte des effets de premier ordre et de second ordre pour minimiser les impacts environnementaux négatifs Prise en compte de tous les acteurs au-delà des seules parties prenantes de l'IA

AFNOR SPEC 2314 Référentiel général pour l'IA frugale, juin 2024

Performance et impact (budgétaire, environnemental)



https://artificialanalysis.ai/?models=gpt-oss-20b%2Cgpt-oss-120b%2Cgpt-5-minimal%2Cgpt-5%2Cgpt-4-1%2Co3%2Co3-pro%2Cllama-4-maverick%2Cgemini-2-5-flash-reasoning%2Cgemini-2-5-pro%2Cmistral-small-3-2%2Cmistral-medium-3-1%2Cdeepseek-r1%2Cdeepseek-v3-1%2Cdeepseek-v3-1-reasoning%2Cgpt-35-turbo#intelligence-vs-price-log-scale 18 septembre 2025

Stratégie numérique

- Expérimentation RAGaRenn
- Mutualisation projet ILaaS

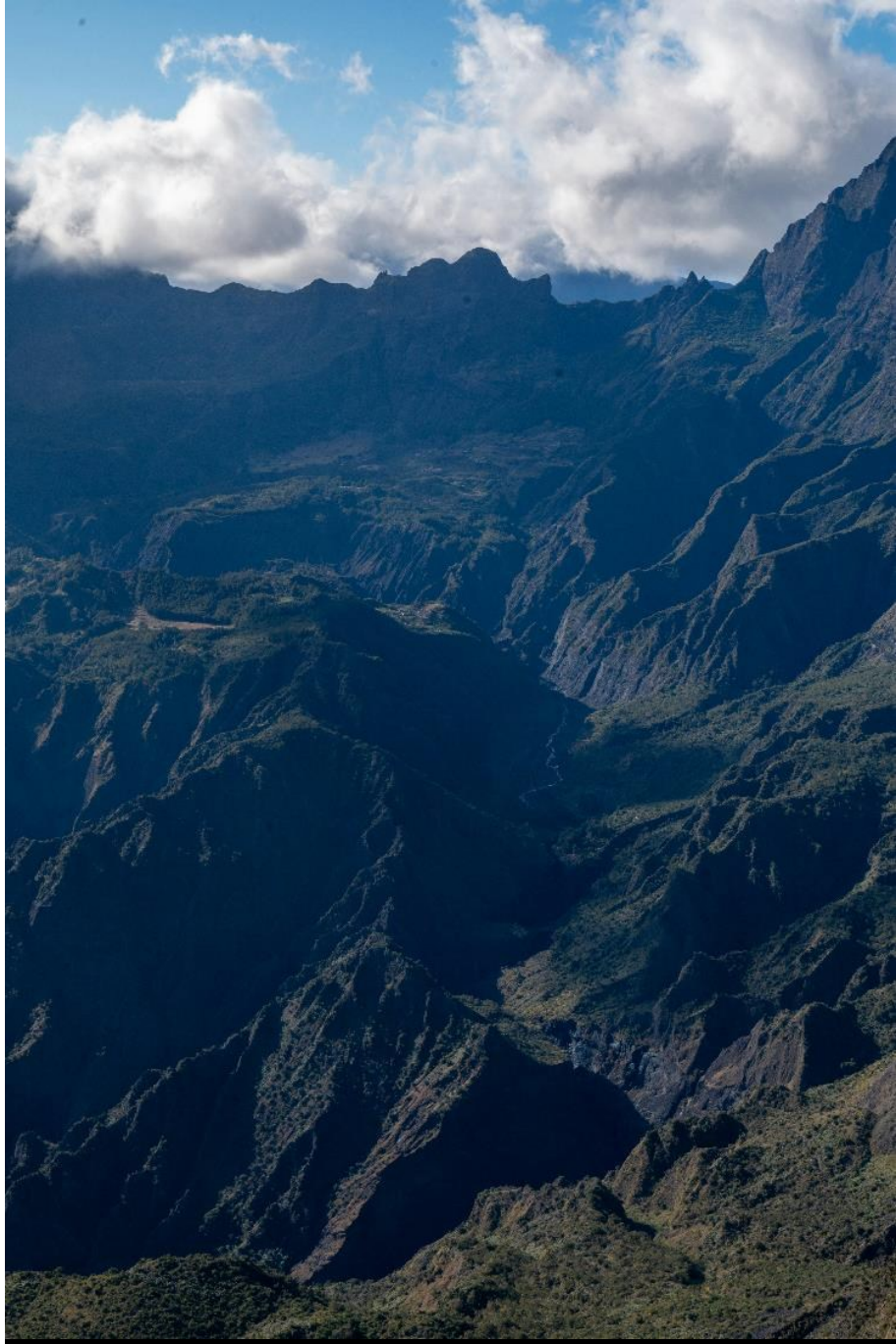


Photo de Oscar Brouchot sur Unsplash

RAG = Retrieval Augmented Generation

Sans RAG

Avec RAG

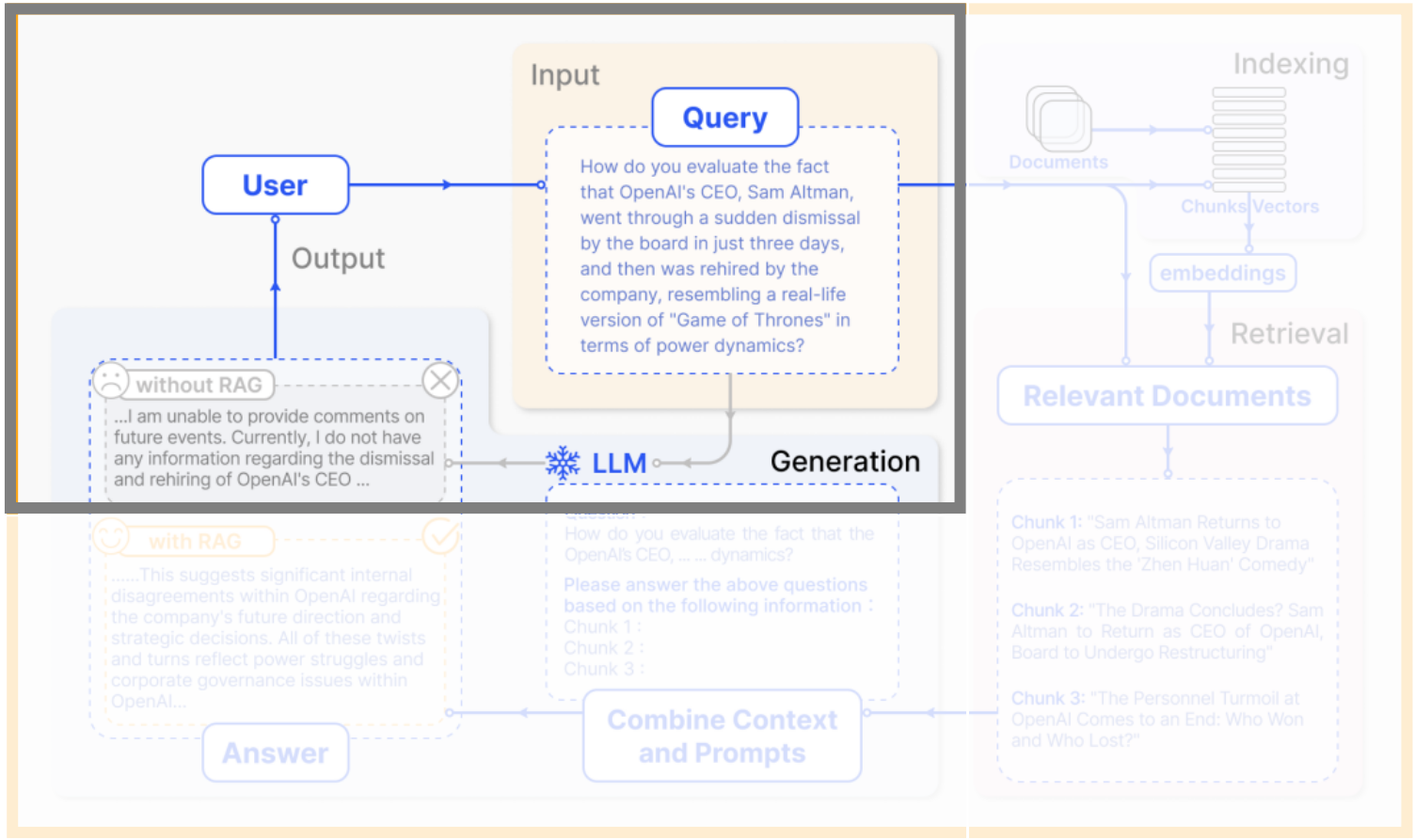
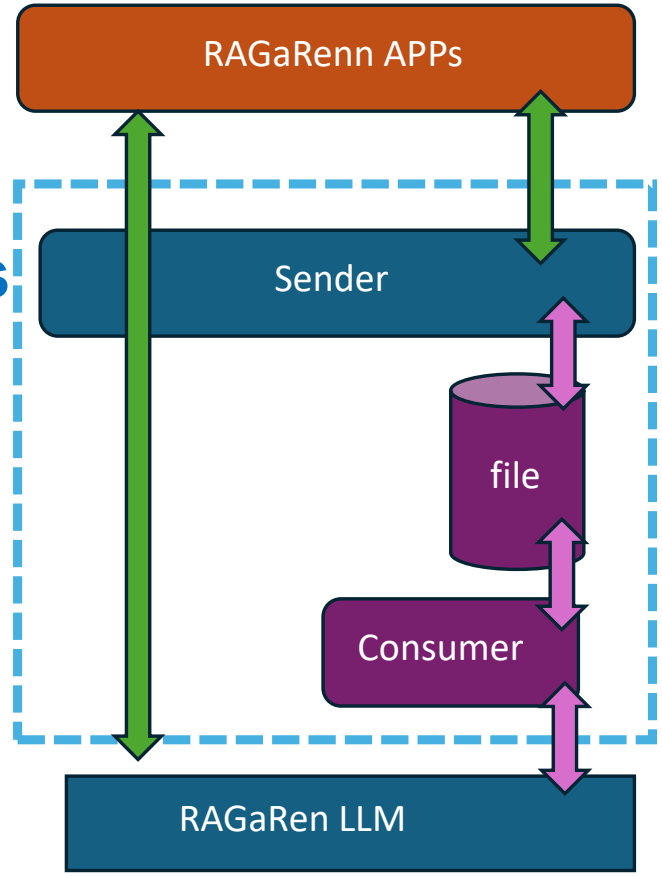


Fig. 2. A representative instance of the RAG process applied to question answering. It mainly consists of 3 steps. 1) Indexing. Documents are split into chunks, encoded into vectors, and stored in a vector database. 2) Retrieval. Retrieve the Top k chunks most relevant to the question based on semantic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

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., … Wang, H. (2024). Retrieval-Augmented Generation for Large Language Models: A Survey https://doi.org/10.48550/arXiv.2312.10997

RAGaRenn : utiliser des sources connues



Déploiement applicatif : RAGaRenn Apps = instances

- 358 instances mono-usager (legacy, 35 actifs)
- 98 instances pour des groupes d'usagers
  73 actifs dont :
  - 24 en SSO : dédiées à 1 université + 1 ESR France
  - 1 EduGain

Utilisateurs depuis mars 2024

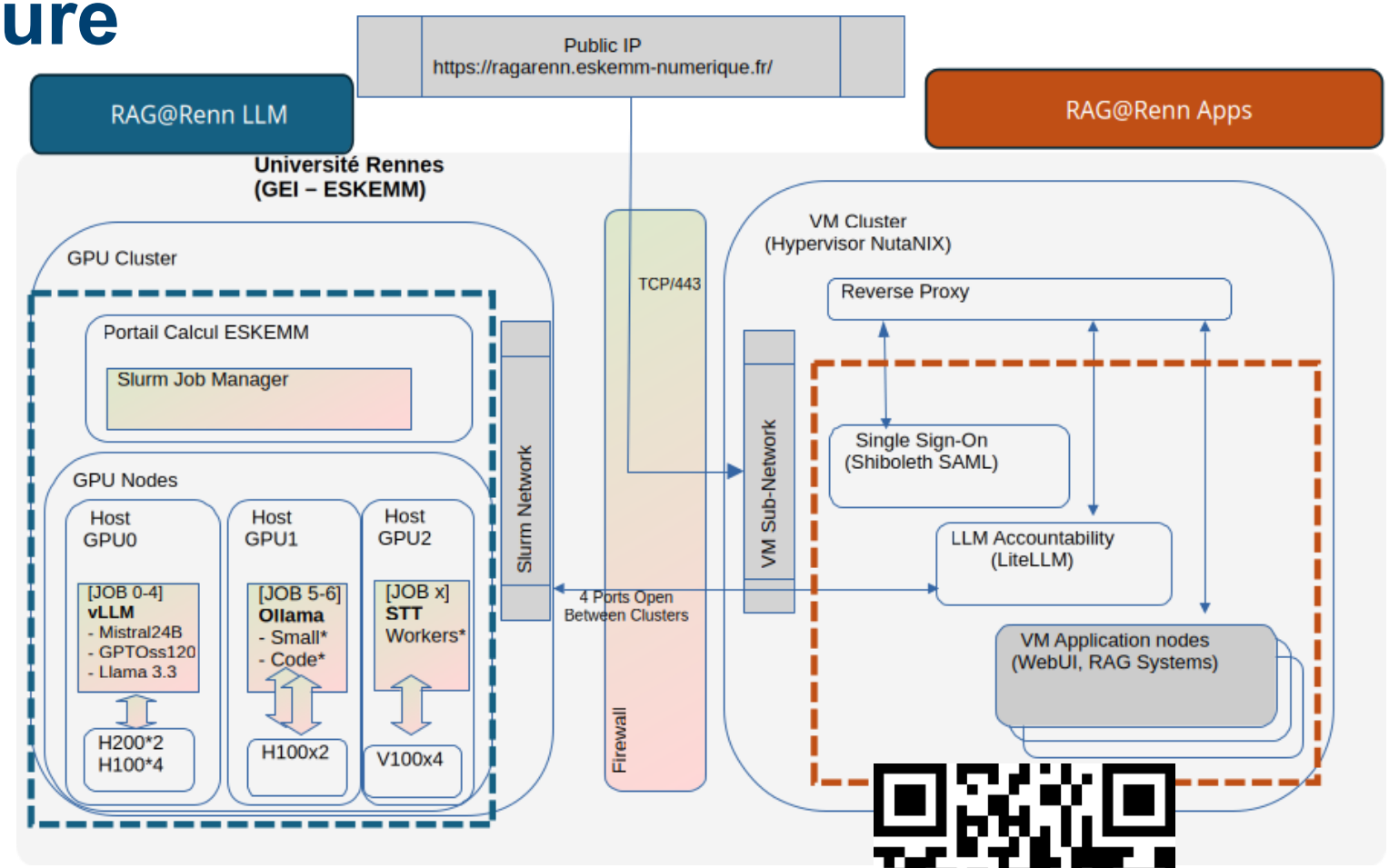
- +1968 inscrits dans l'instance principale ESR
- > 3000 inscrits au global
- 155369 requêtes sur un an
- Classification requêtes (Dewey) – énergie + temps

https://ragarenn.eskemm-numerique.fr/

RAGaRenn infrastructure

Déploiement inférence

- Administration par Tâches (Slurm)
- 8 GPUs en propre : 2xH200, 2xH100 (80Gb), 4xV100
- 4 H100 de plus si besoin de débordement



https://ragarenn.eskemm-numerique.fr/

Stratégie numérique : IA générative

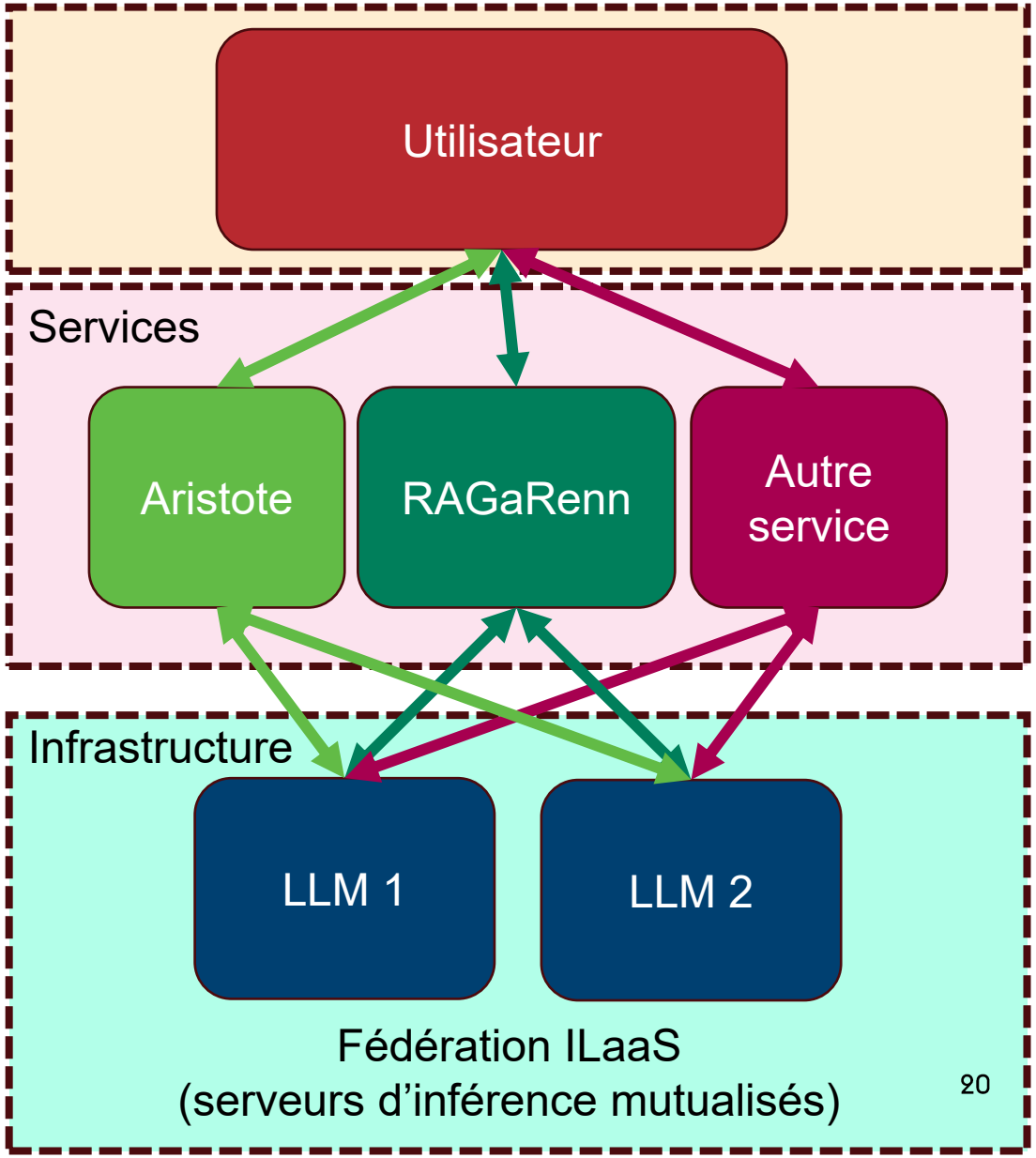
Principe : séparer l'accès à un LLM et l'accès à des services utilisant un LLM

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 maitrisées
- 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)

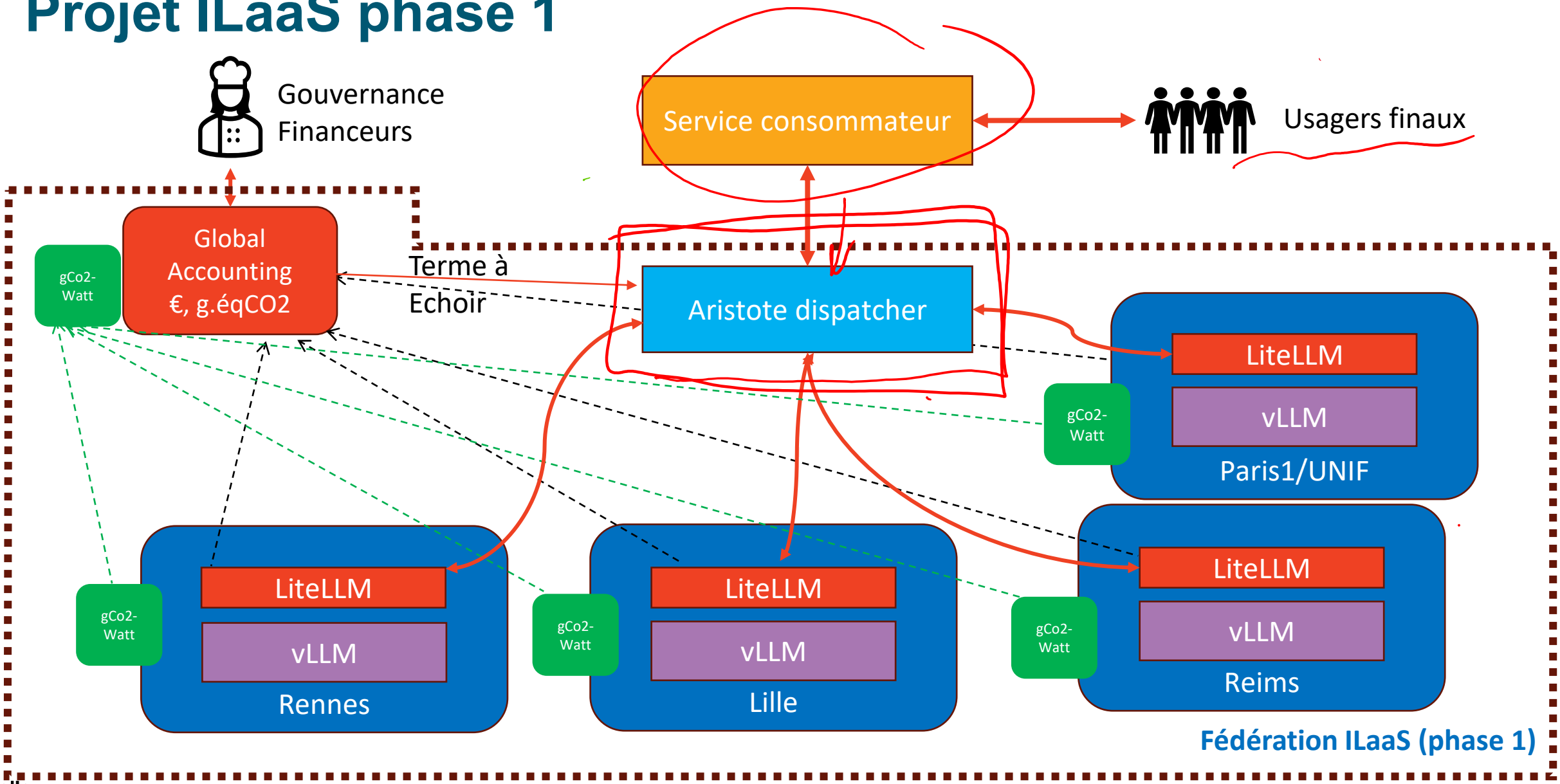
- Déploiement : briques open source (OpenWebUI, vLLM, ollama)
- Hébergement local : datacenter régional labellisé Eskemm Data
- Observations sur les usages (qualitatif & quantitatif)

Infrastructures : projet ILaaS*

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

* ILaaS = Inférence LLM as a service

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**

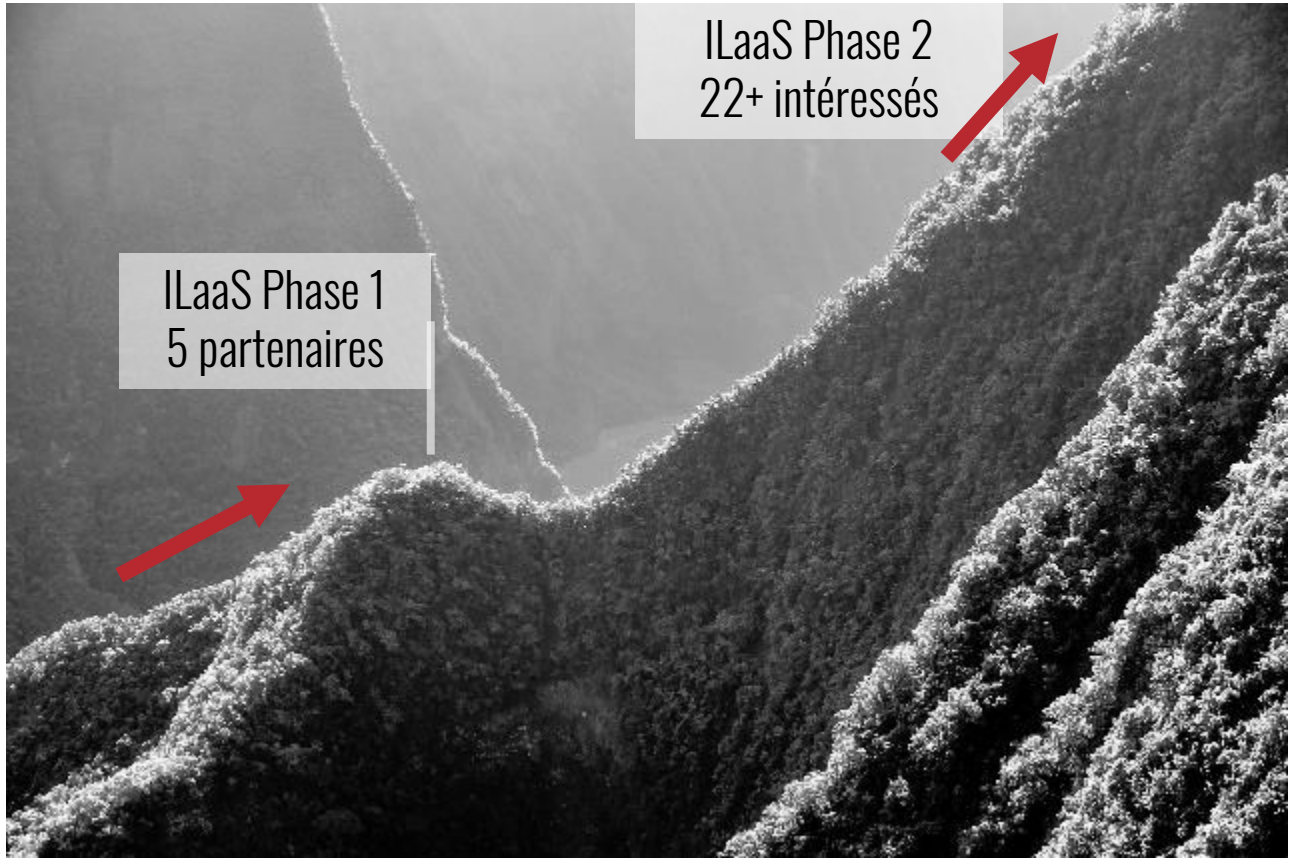


Photo de Philippe Bout sur Unsplash

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 (maintenant), Speech-to-text (bientôt)
- Facturation & relations contractuelles : étude en cours

marketplace

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