ILaaS
Infrastructure
d'IA mutualisée

ilaas.fr

Souveraineté

Olivier WONG

Coordinateur projet
vice-président numérique

Université
de Rennes

Contenu sous licence
Creative Commons BY-NC-SA 4.0
Photo de Philippe Bout sur Unsplash

Plan

- Enjeux stratégiques
- Projet : historique et feuille de route 2026
- Infrastructure & usages
- Offre

Photo de Edi Libedinsky sur Unsplash

ILaaS
Infrastructure d'IA mutualisée

Enjeux stratégiques

- Sécurisation des données
- Impact environnemental
- Souveraineté

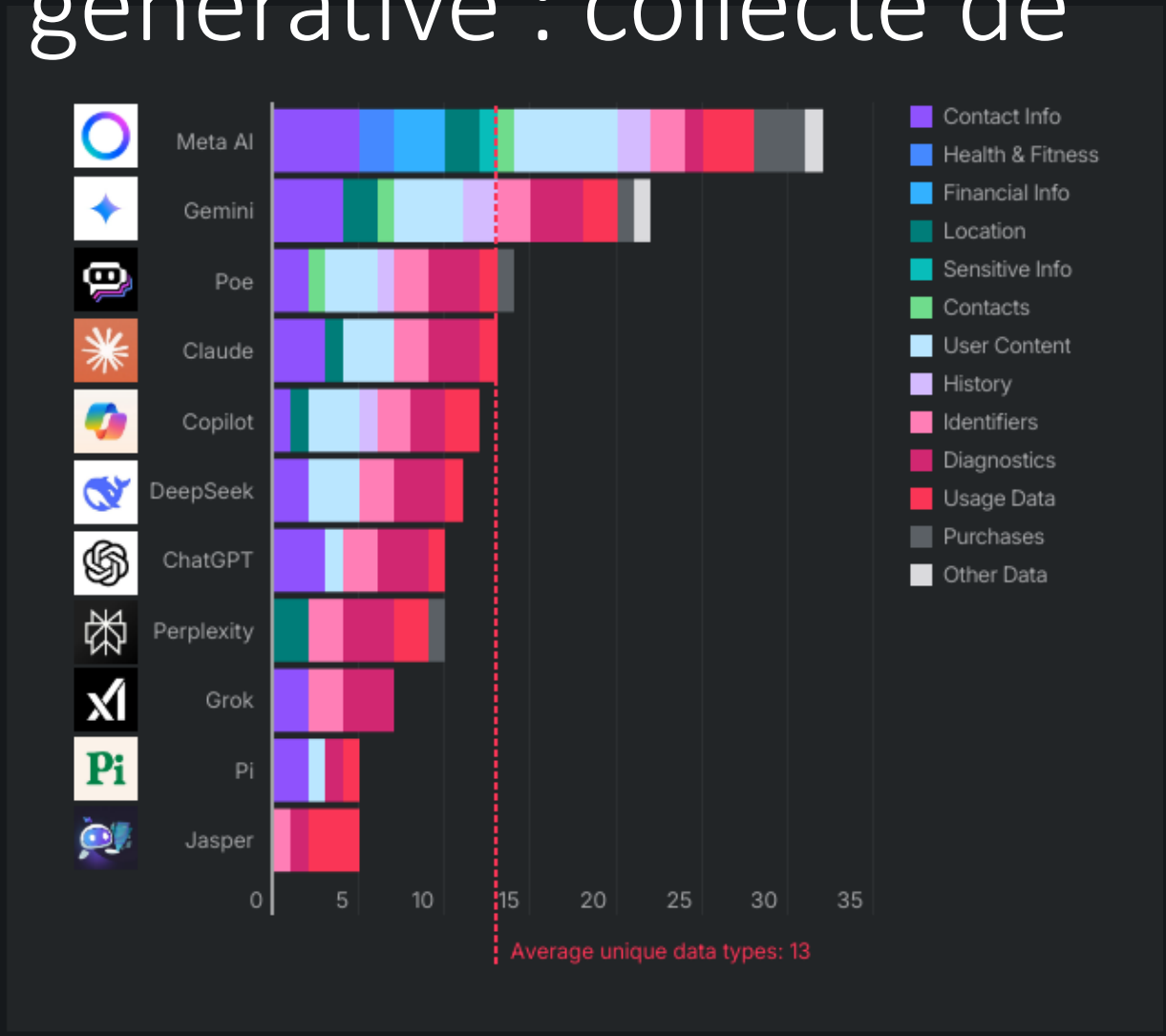
Photo de James Toose sur Unsplash

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

Surfshark®

https://surfshark.com/research/chart/ai-chatbots-privacy, 20 mai 2025

Impact environnemental du numérique en France

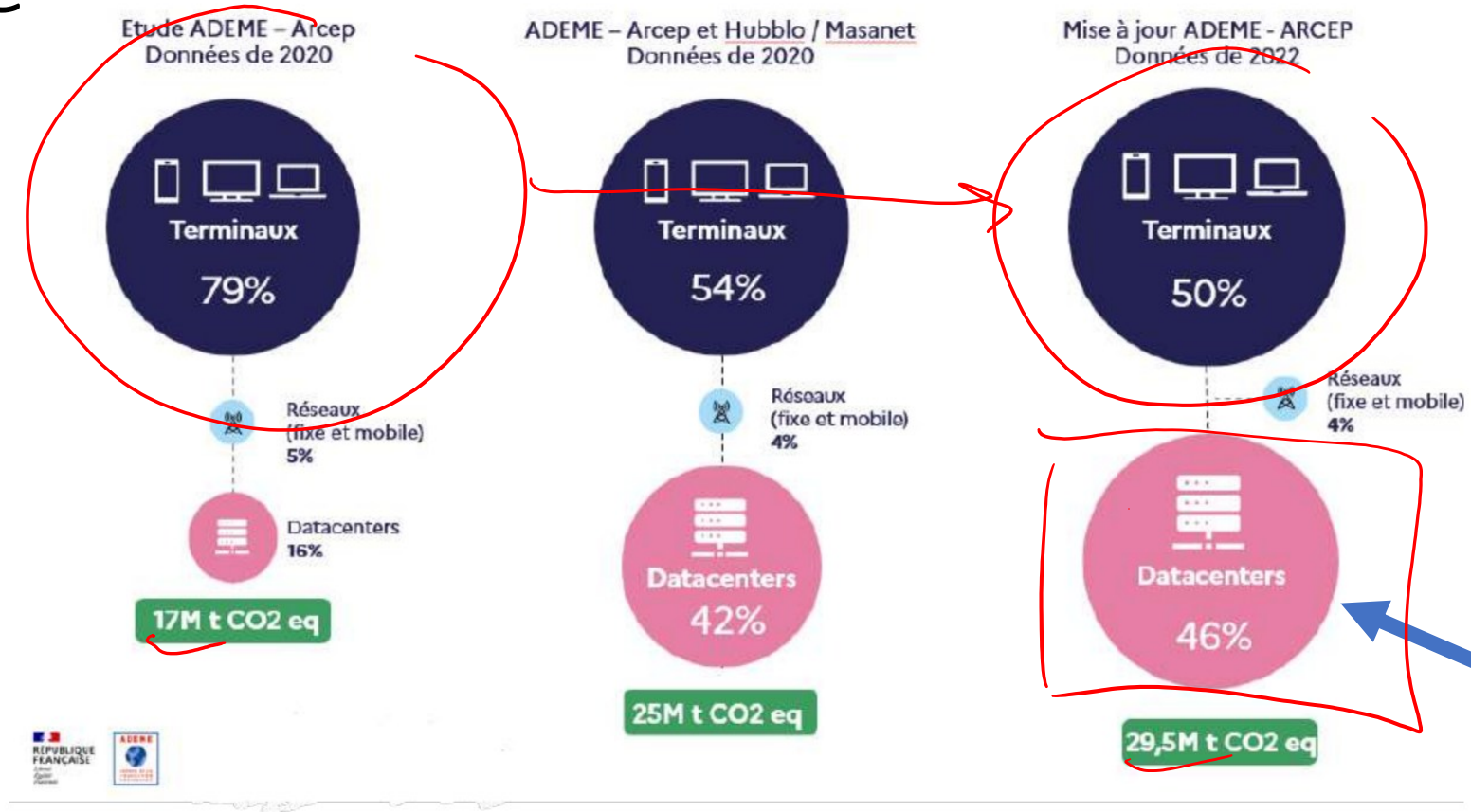


Figure 4 – Comparaison des résultats avec les études précédentes – par tier

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

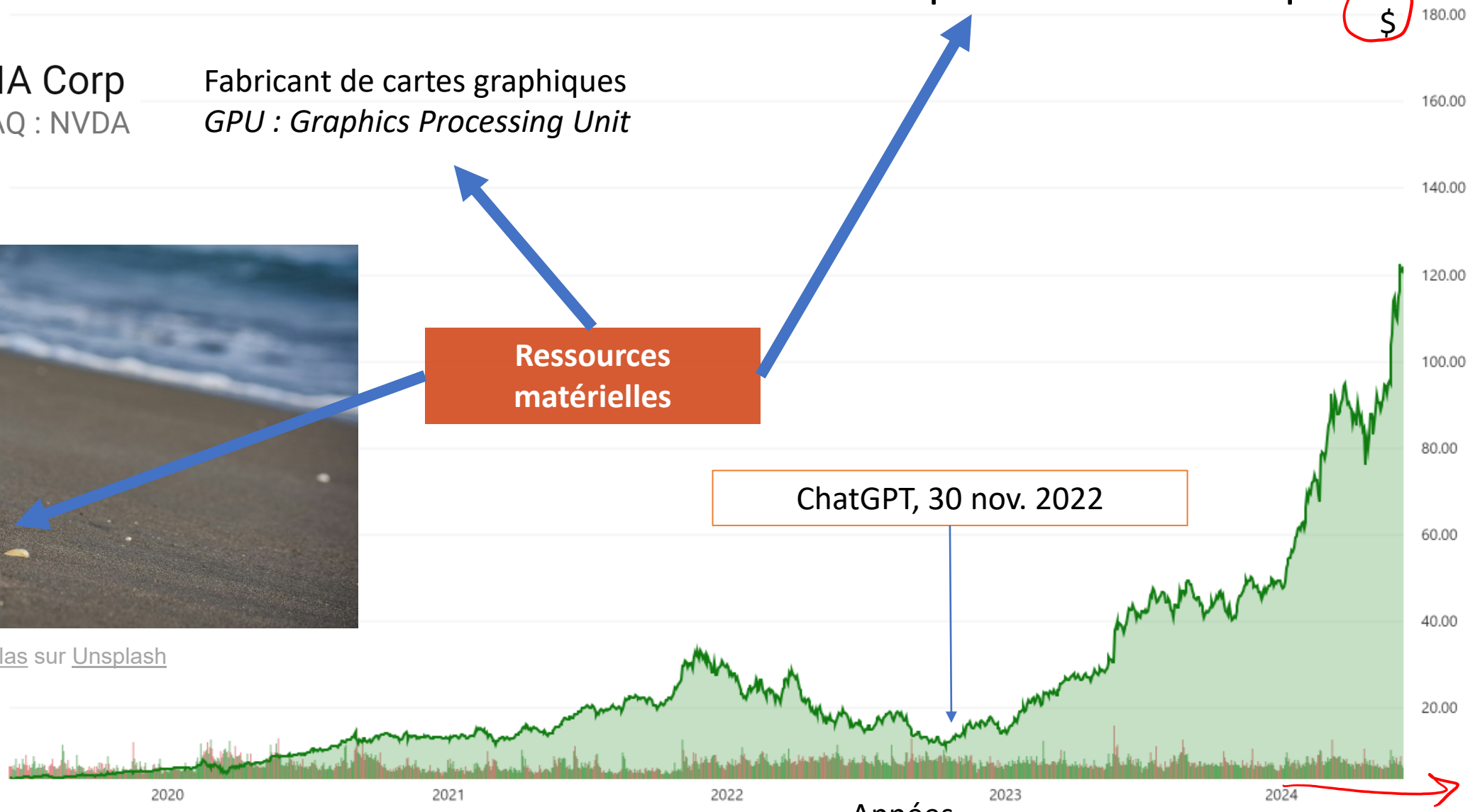
Pendant la ruée vers l'or … vendez des pelles et des pioches

NVIDIA Corp
NASDAQ : NVDA

Fabricant de cartes graphiques
GPU : Graphics Processing Unit

Photo de Alejandro Alas sur Unsplash

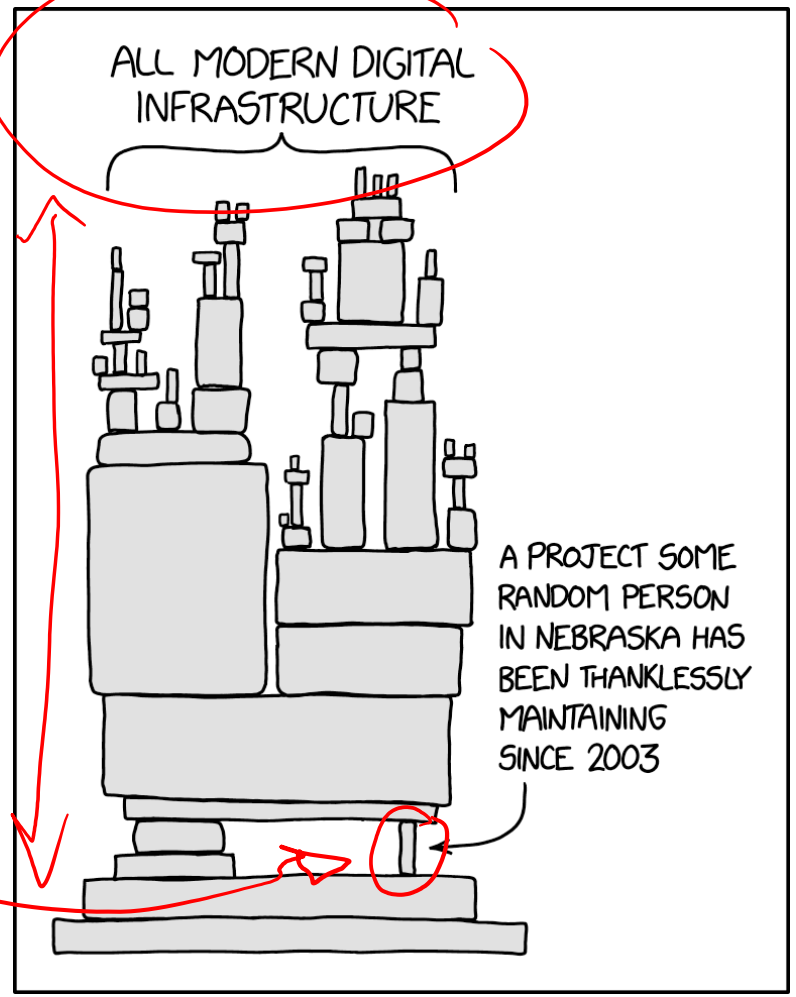
Ressources matérielles



Souveraineté numérique : capacité à faire des choix technologiques

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

- Quelles dépendances ?
  - Chaine d'approvisionnement (matérielle et logicielle)
  - Répartition de la valeur entre acteurs
  - Déplacement des enjeux sur les données



Dependency https://xkcd.com/2347/ CC-BY-NC 2.5

Photo de Chris Liverani sur Unsplash

Objectif

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

Enjeux stratégiques : infrastructure mutualisée

- Soutenabilité : numérique responsable
  - Optimisation de l'infrastructure et des services : utiles, utilisés, utilisables
  - Accompagnement des usages vertueux
  - Modèle économique soutenable
- Sécurité
  - Niveau de confiance partagé dans l'ESR
  - Offre mutualisée autour des datacenters régionaux labellisés
- Souveraineté
  - Maitrise du déploiement et du traitement LLM
  - Maintien et développement des compétences techniques en IA
- Résilience et robustesse
  - Qualité de service, lissage des pics, gestion des indisponibilités
  - Distribution régionale & lien avec les centres nationaux (CINES)

Projet : genèse & avancement

- Fin 2024 : Lancement du projet à 5 établissements

- Fin mars 2026
  - 211 clés distribuées pour des services opérés par 88 établissements
  - Fédération opérée par 6 Data Centres Labellisés en Région de l'ESR

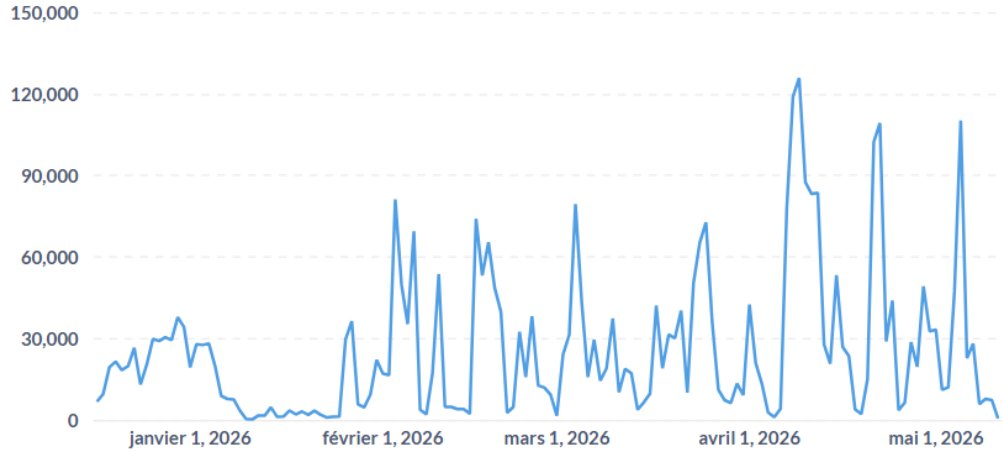
Feuille de route 2026

- Structuration
  - Accord de consortium entre 17 établissements, ouvert au fil de l'eau
  - Modèle économique, facturation, dimensionnement de l'infrastructure cible
  - Consolidation infrastructure socle

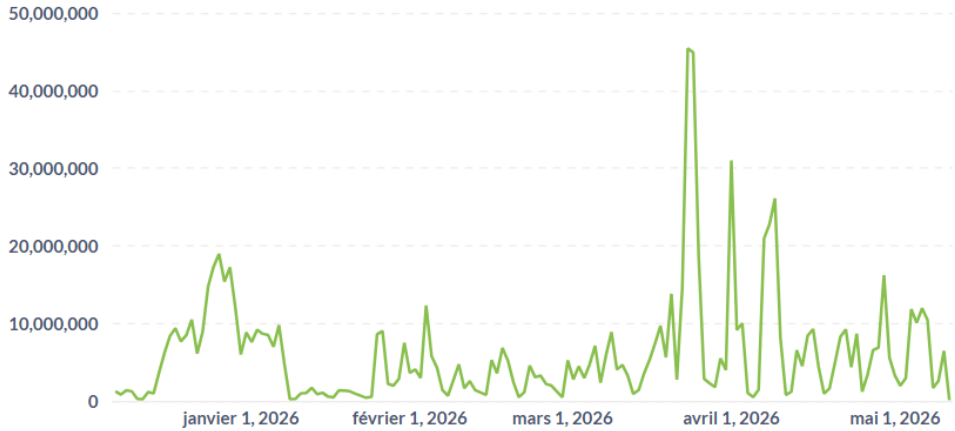
- Portail applications ILaaS : accès à un large panel d'applications IA
  - Logiciels EdTech sur des serveurs maîtrisés (ex : Nolej ; Pyxiscience, …)
  - Solutions Open Source développées au sein de l'écosystème ESR
  - Environnement de développement agentiques (N8N, Dialogue, …)

Infrastructure & usages

Capacité	Matériel	Répartition
Socle permanent	8 H200 répartis sur 4 nœuds	4 noeuds
Variable selon besoins et disponibilité	8 RTX 6000 Pro + 4 L40(s) 16 H100 + 11 H200	16 noeuds
Prochainement (sous réserves vu le contexte d'inflation)	8 B300 14 H200 ou équivalent	A déterminer



Nombre de requêtes par jour (pic constaté : 120 000)



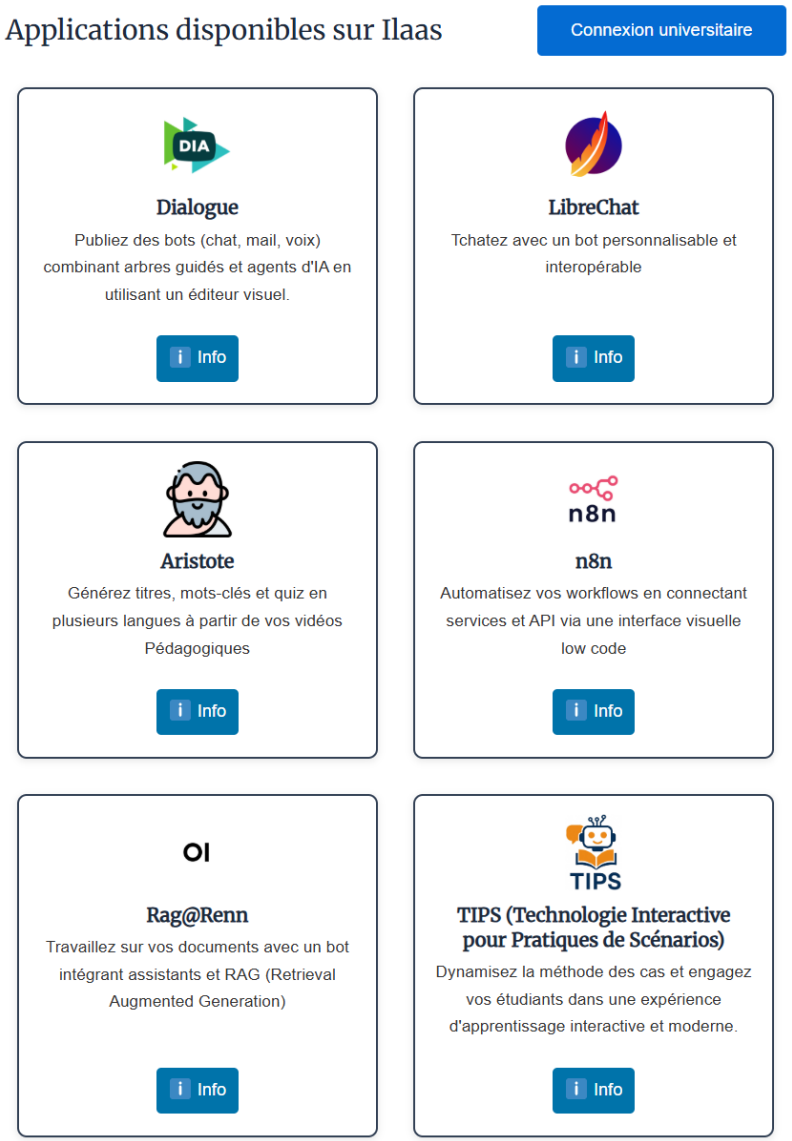
Nombre de *tokens* par jour (pic constaté : 44 millions)

Partenariats & soutiens

- AMUE : projet IA souveraine
  - Contribution technique
  - Offre API Mistral Medium (ILaaS / Rennes)
  - Plateforme « Le Chat » (CINES & ILaaS)
- Esup
  - Participation aux journées techniques
  - Collaborations atelier IA
- MESRE
  - Soutien financier et concertation stratégique
  - Webinaires en commun IA Sup (DGESIP)

Offre API ilaas.fr

- **Inférence LLM par API** llm.ilaas.fr gateway compatible API OpenAI
  - LLM
    - 6 open weights : Gemma 4, GPT oss 120B, Llama 3.x (8B et 70B), Mistral Small 3.2 24B, Qwen 3 30B
    - 1 commercial : Mistral Medium 3
  - Haute-disponibilité : répartition multi datacenters
  - Répartition (charge, file d'attente)
  - Accès prioritaire pour les contributeurs de ressources (coupe-file)
  - Confidentialité : pas d'accès au contenu des requêtes (simple relais)
- **Transcription Speech-to-text** (STT) asynchrone stt.ilaas.fr
  - Basée sur Whisper
  - Workers distribués dans les DC ILaaS
  - Fonctionnalités : génération de métadonnées (titres, résumés, mots-clés), traduction, sous-titrage paramétrable (nb de lignes, nb de caractères), génération de Quiz et comptes-rendus
  - Confidentialité : suppression après traitement (vidéos, transcripts), anonymisation de l'utilisateur final
- **Embeddings** : bientôt

Offre portail apps.ilaas.fr

- Vue publique & authentifiée (fédération Renater)
- Applications
  - open source ou commerciales
  - Utilisant les inférences ILaaS
- Plateformes
  - mutualisées (ex: Dialogue) : raccordement par permissions
  - instances dédiées (ex: n8n) : déploiement par organisation
- Traitement des demandes
  - Validation politique par établissement (VP-Num)
  - Validation du GT-Projet