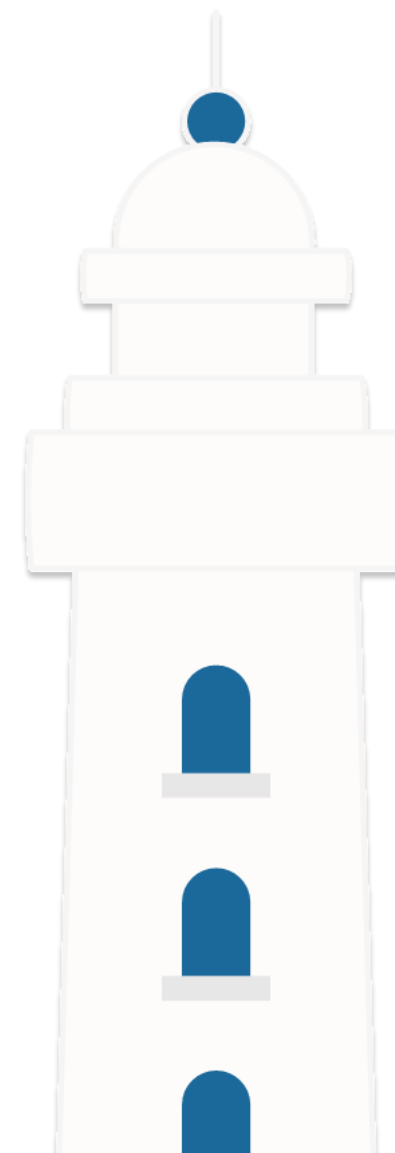




CSIESR

Assises 2026



Coopérer pour une infrastructure d'IA générationnelle dédiée à l'ESR

Olivier WONG HEE KAM

olivier.wong-hee-kam@univ-rennes.fr

Coordinateur projet



Vice-président numérique



Université
de Rennes

Président



C'est quoi ?

4

Pour quoi ?

12

Comment ?

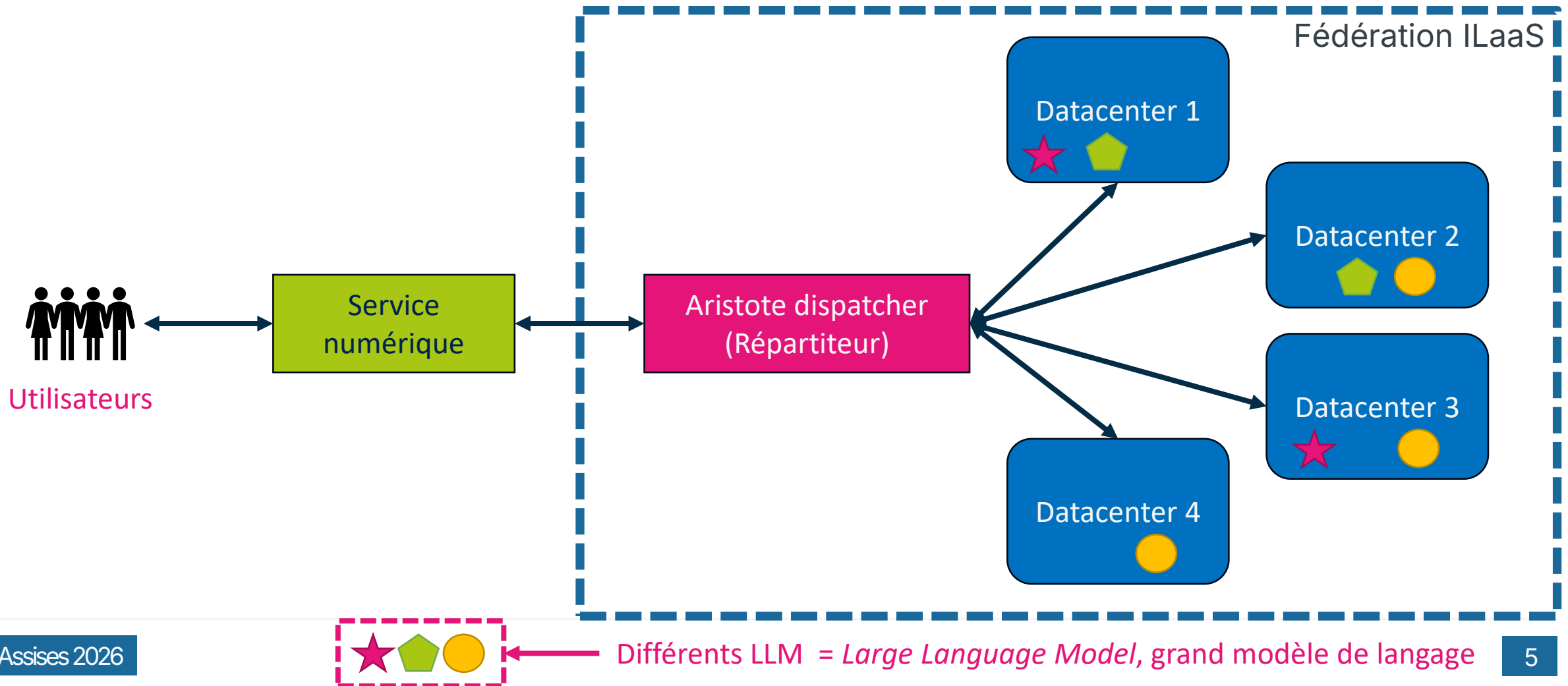
19

Merci

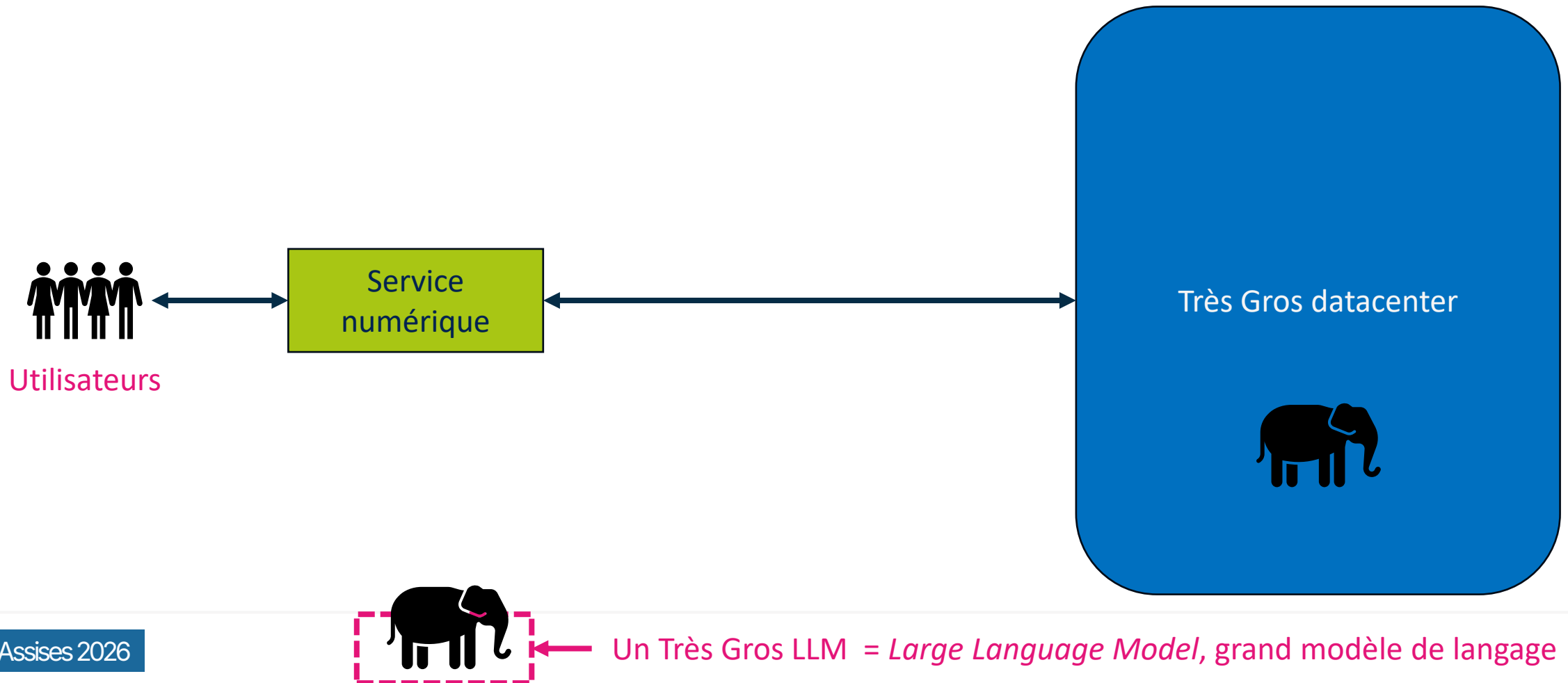
28

C'est quoi ?

Infrastructure d'IA générative pour l'inférence de LLM « *as a Service* »



Ce serait beaucoup plus PERFORMANT comme ça, non ?



Oui.
Ce serait PERFORMANT et
« qui peut le plus peut le moins »
mais est-ce pertinent ?



Transporter un seul croissant dans la remorque du camion sur 100m, est-ce bien la performance rêvée ?

Que serait une
stratégie de robustesse ?

Robustesse

Capacité d'un système à se maintenir
sa stabilité (à court terme)
et sa viabilité (à long terme)
malgré les fluctuations



Objectif du projet

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

Pour quoi ?



Analogie du casino

L'objectif du casino est de gagner le plus d'argent possible, en incitant les joueurs à jouer plusieurs fois et longtemps. Les machines à sous du casino ont un comportement aléatoire.

Pour un joueur :

- Il est difficile de prévoir combien ça coûtera pour obtenir un résultat donné
- À chaque essai, il est difficile de prédire si le résultat sera favorable, mais l'espoir d'une bonne surprise incite à essayer

C'est plus facile de tirer le levier d'une machine à sous en espérant un gain hypothétique, que de sortir du casino pour réellement travailler contre rémunération

Dans de rares cas aléatoires qui existent, une machine est capable de produire un résultat démesurément élevé : c'est le « jackpot » célébré comme un moment magique

Des situations d'addiction sont possibles et peuvent conduire à une mise en danger pour les joueurs concernés

Souveraineté numérique ? Autonomie stratégique ?

Souveraineté numérique

« Maîtrise complète allant du matériel au logiciel »

Avis de la Cour des Comptes : une souveraineté absolue serait
« coûteuse et contre-productif »

ou

Autonomie stratégique

« Maîtriser ses dépendances afin d'accroître les choix possibles »

RAPPORT

Souveraineté numérique

De l'urgence d'organiser
la coopération entre le public,
le privé et les communs
numériques



8 piliers de la souveraineté numérique (autonomie stratégique)

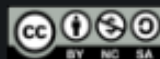
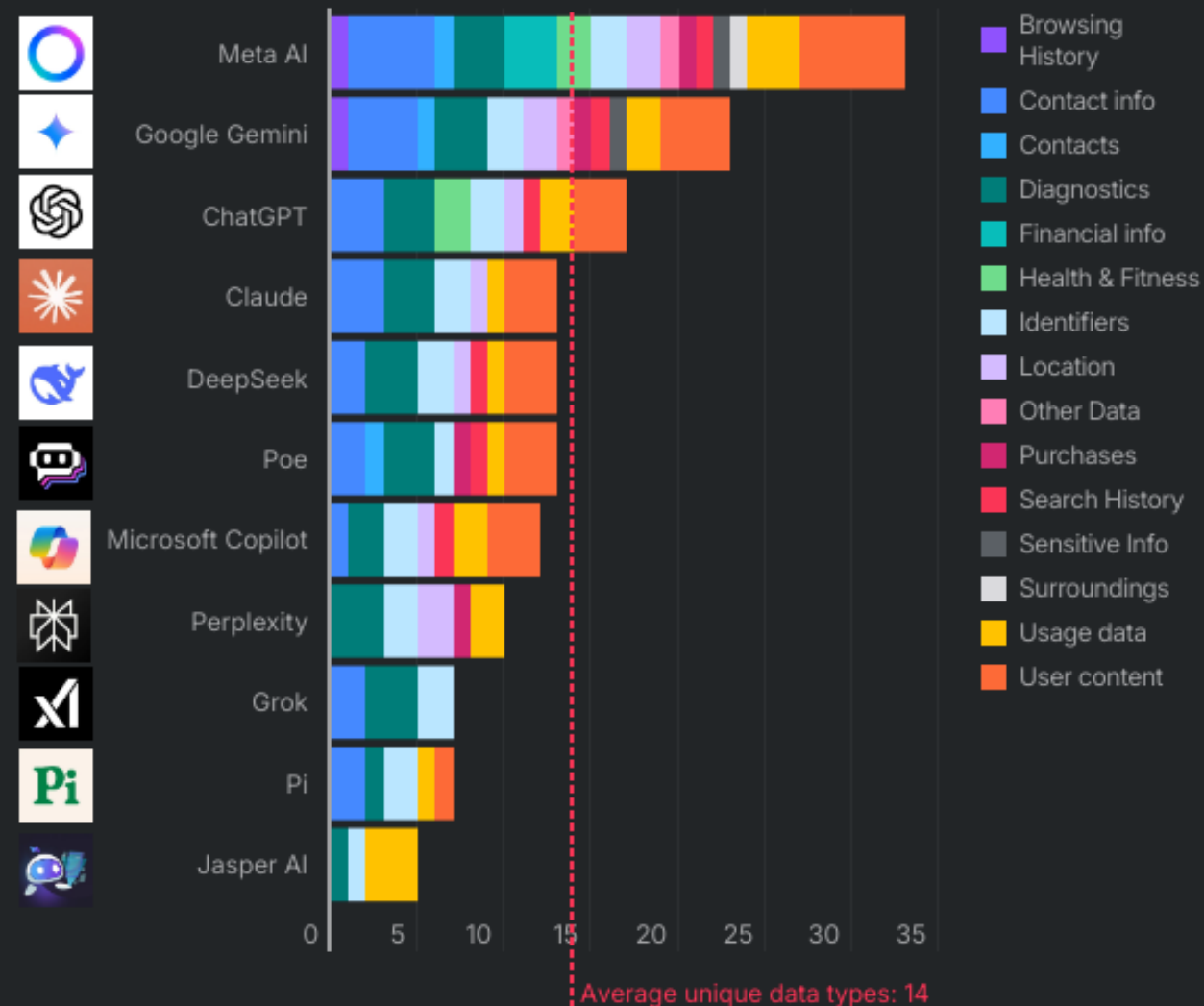
- Contrôle et liberté de choix : capacité d'un utilisateur (Etat, entreprise, citoyen) à prendre ses propres décisions concernant les technologies sans y être contraint par un tiers
- Réversibilité : capacité à changer de fournisseur / infra sans coût prohibitif, ni perte de données
- Diversification : éviter une dépendance monopolistique, via une stratégie multifournisseurs
- Sécurité : protection au juste niveau des données selon leur sensibilité & la réduction des vulnérabilités face aux cyberattaques et à l'espionnage
- Régulation : capacité à imposer des règles strictes aux acteurs opérant sur le territoire national ou traitant des données de ressortissants nationaux
- Sobriété : éliminer les fonctionnalités inutiles, promouvoir des solutions répondant précisément aux besoins des utilisateurs, afin de concilier usages numériques et engagements climatiques & budgétaires
- Formation : soutenir la recherche, l'attraction des talents, et le renforcement des transferts entre la recherche et l'industrie

Services d'IA générative & collecte de données

DATA UPDATE: MARCH, 2026

Meta AI collects over 90% of user data types

ChatGPT may now collect 17 out of 35 data types, a 70% increase from the 10 data types identified in last year's AI chatbots review



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



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

- Maîtrise 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)

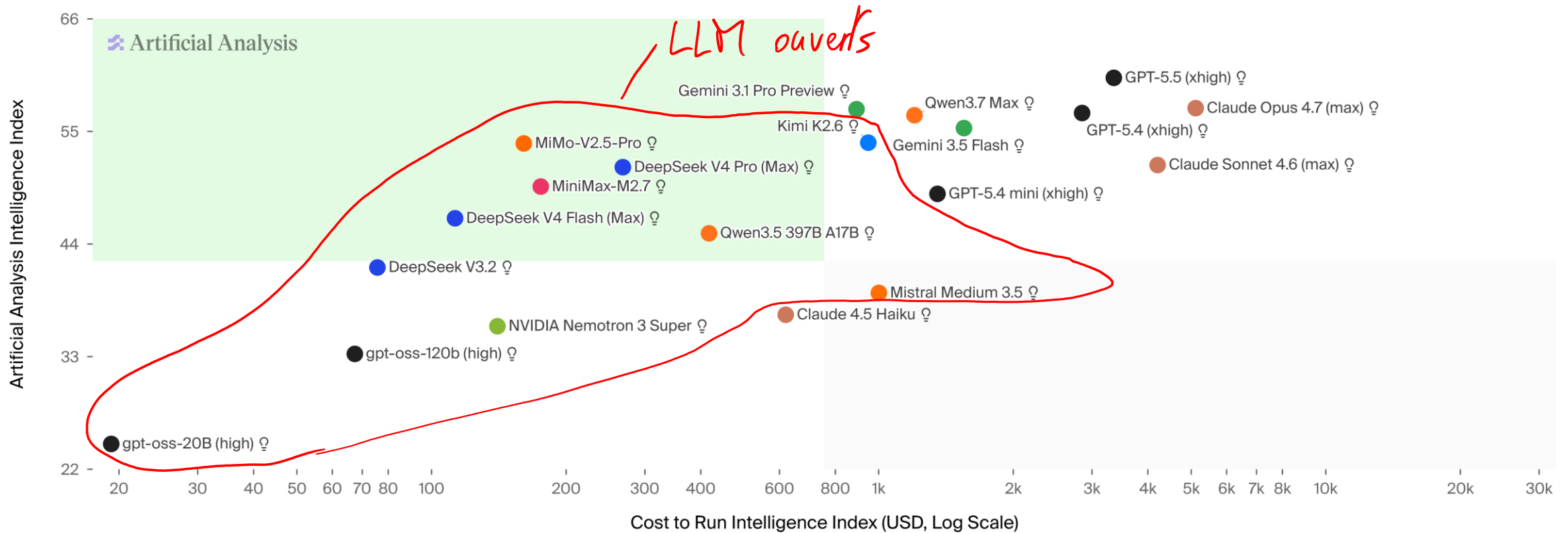
Comment ?

Intelligence vs. Cost to Run Artificial Analysis Intelligence Index

Artificial Analysis Intelligence Index · Cost to run Intelligence Index

Most attractive quadrant

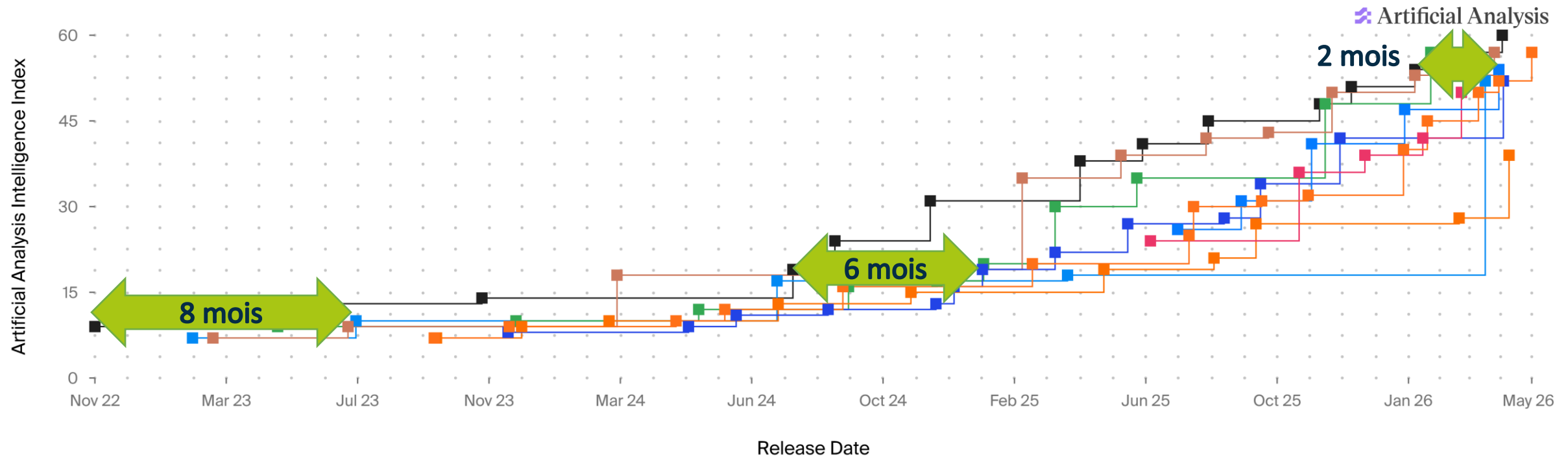
● OpenAI ● Google ● Anthropic ● Mistral ● DeepSeek ● MiniMax ● NVIDIA ● Kimi ● Xiaomi ● Alibaba



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

● OpenAI ● Meta ● Google ● Anthropic ● Mistral ● DeepSeek ● MiniMax ● Kimi ● Alibaba



Projet : genèse & avancement

Lancement du projet fin 2024 à 5 établissements

Point d'étape mars 2026 :

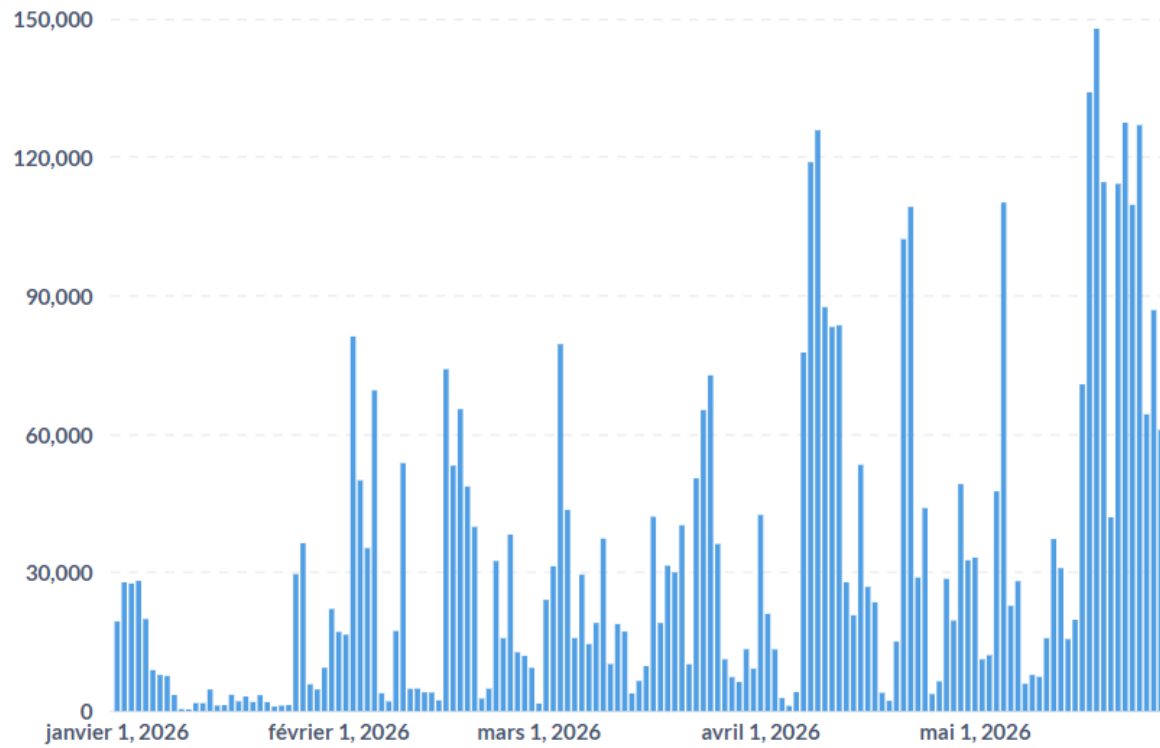
- 211 clés distribuées pour des services opérés par 88 établissements
- Fédération opérée par 6 établissements dans leurs Data Centres Labellisés en Région de l'ESR



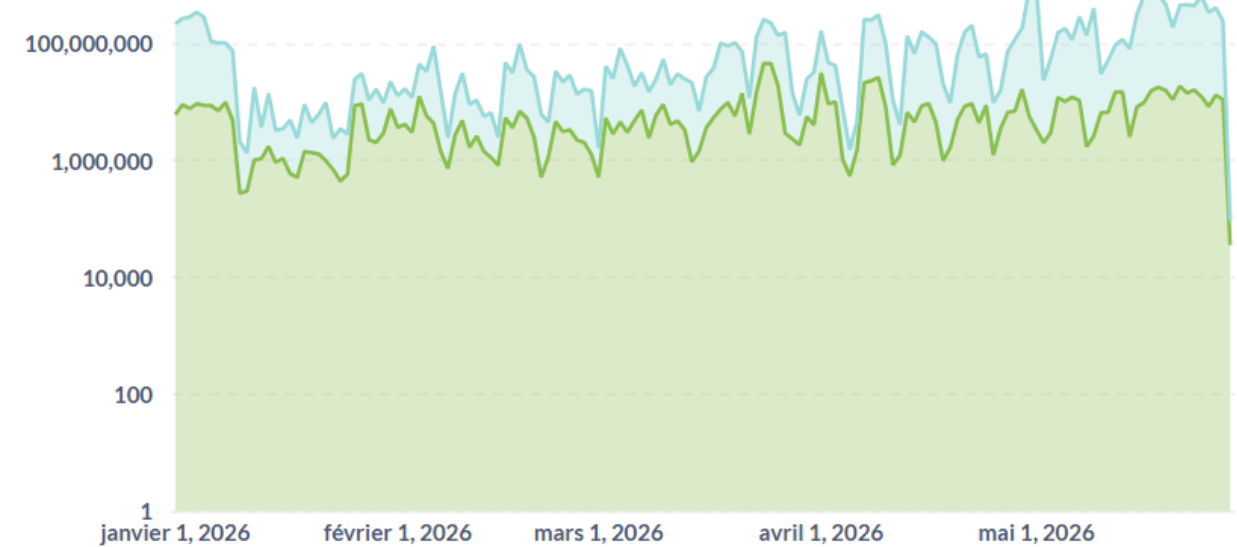
Capacité	Matériel	Répartition
Socle permanent	8 H200 répartis sur 4 nœuds	4 nœuds
Variable selon besoins et disponibilité	8 RTX 6000 Pro + 4 L40(s) 16 H100 + 11 H200	16 nœuds
Prochainement (sous réserves vu le contexte d'inflation)	8 B300 14 H200 ou équivalent	A déterminer

Usages constatés (mai 2026)

Nombre de requêtes par jour (150k max par jour)
Du 1^{er} janvier au 28 mai 2026



Nombre de tokens en entrée et en sortie
(max par jour : ~ 834k entrée / 45k sortie)
Du 1^{er} janvier au 28 mai 2026
Echelle logarithmique



Feuille de route 2026

Structuration

- Accord de consortium entre 24 établissements : Université de Rennes, Université Reims Champagne-Ardennes, Université Paris 1 Panthéon-Sorbonne, Université Paris Saclay, Université de Lille, Université Côte D'Azur, Université Paris Cité, Université de Tours, Université d'Orléans, Université de Technologie de Compiègne, Université de Strasbourg, Université de Lorraine, La Rochelle Université, Université de Picardie Jules Verne, Université Clermont Auvergne, Institut Mines-Telecom, Nantes Université, Université de Haute Alsace, CRIANN, Université d'Artois, Inria, Hcéres, Renater, CY Cergy Paris Université, Université Gustave Eiffel
- Consortium 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, ...)

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

- Concertation stratégique
- Soutien financier (fonds d'amorçage CoSIN, DGRI)
- Webinaires en commun IA Sup (DGESIP)



Offre API ilaas.fr (documentation sur le site)

Inférence LLM par API llm.ilaas.fr gateway compatible API OpenAI

- LLM
 - 6 open weights : Gemma 4 31B, Mistral Small 4 119B, Qwen 3.6 35B A3B
 - Arrêtés fin septembre 2026 : GPT oss 120B, Llama 3.x (8B et 70B), Mistral Small 3.2 24B,
 - 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](https://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 : www.ilaas.fr/rag-ilaas swagger rag-api.ilaas.fr/docs

- Modèles bge-m3 (emdeddings) et bge-reranker-v2-m3 (reranker)

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

Applications disponibles sur ILaaS

[Connexion universitaire](#)

Dialogue

Publiez des bots (chat, mail, voix)
combinant arbres guidés et agents d'IA en
utilisant un éditeur visuel.

[Info](#)

LibreChat

Tchatez avec un bot personnalisable et
interopérable

[Info](#)

Aristote

Générez titres, mots-clés et quiz en
plusieurs langues à partir de vos vidéos
Pédagogiques

[Info](#)

n8n

Automatisez vos workflows en connectant
services et API via une interface visuelle
low code

[Info](#)

OI

Rag@Renn

Travaillez sur vos documents avec un bot
intégrant assistants et RAG (Retrieval
Augmented Generation)

[Info](#)

TIPS (Technologie Interactive pour Pratiques de Scénarios)

Dynamisez la méthode des cas et engagez
vos étudiants dans une expérience
d'apprentissage interactive et moderne.

[Info](#)

Questions - réponses

Quelles sont les modalités pour entrer dans le consortium ILaaS ?

Il faut contribuer au projet : proposer de la puissance de calcul, ou des développements (code) liés au projet.

Mon établissement ne peut pas contribuer au projet, mais souhaite pouvoir consommer des services proposés par la fédération ILaaS. Comment faire ?

Dans ce cas, il faudra signer une convention bilatérale, moins engageante que de rentrer dans le consortium des partenaires.

J'aimerais avoir un accès individuel à l'offre ILaaS. Comment faire ?

Il faut d'abord vous diriger vers votre établissement pour bénéficier d'un accès (API ou app). L'accès API est destiné aux applications gérées par un établissement, pas pour un usage individuel (pour des tests par exemple) mais nous réfléchissons à une solution gérable pour tout type d'établissement.



Merci

ILaaS

Infrastructure
d'IA mutualisée

