esup portail

01 AVRIL 2025
Paris et @distance

ESUP-DAYS 39

# PROJET ILAAS

Olivier WONG -- Université de Rennes
Thibault LE MEUR -- Centrale Supélec Paris Saclay

ESUP-DAYS 39

D'OÙ ÇA VIENT
À QUOI ÇA RESSEMBLE
QUI ON EST
COMMENT ON FAIT ?
PREMIERS DC CONNECTÉS
COMMENT REJOINDRE OU UTILISER

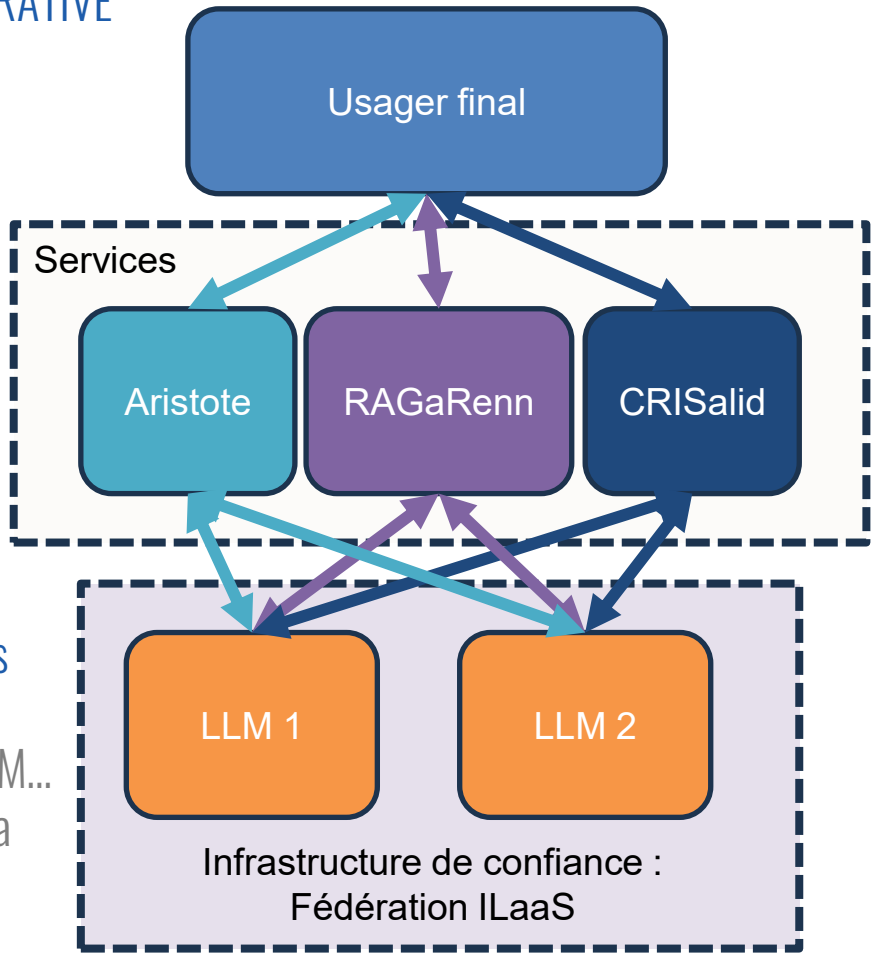
# PROJET ILAAS ... QU'EST-CE QUE C'EST ?



esup portail

# D'OÙ ÇA VIENT ? EXPÉRIMENTATIONS EN 2023 PUIS 2024 : IA GÉNÉRATIVE

- Aristote, CentraleSupélec
  - Brique Aristote-Dispatcher utilisée pour mutualiser les accès aux GPU en gérant des priorités
    - Développée dans le cadre du projet Aristote lors du choix d'utilisation de modèles supérieurs à 8b
      https://github.com/CentraleSupelec/aristote-dispatcher
    - Objectif : mutualiser les ressources de calcul sur différents usages et pour différents "clients"
- RAGaRenn, Université de Rennes
  - Solution de confiance pour expérimenter l'intérêt du RAG pour les personnels volontaires
    - Déploiement de briques open–source: OpenWebUI, ollama, vLLM...
    - Hébergement local : datacenter régional labellisé Eskemm Data
    - Observations sur les usages (qualitatif & quantitatif)



# D'OÙ ÇA VIENT ? CONSTATS PARTAGÉS MI-2024

- **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

Photo de Steven Wright sur Unsplash

Photo de Chris Liverani sur Unsplash

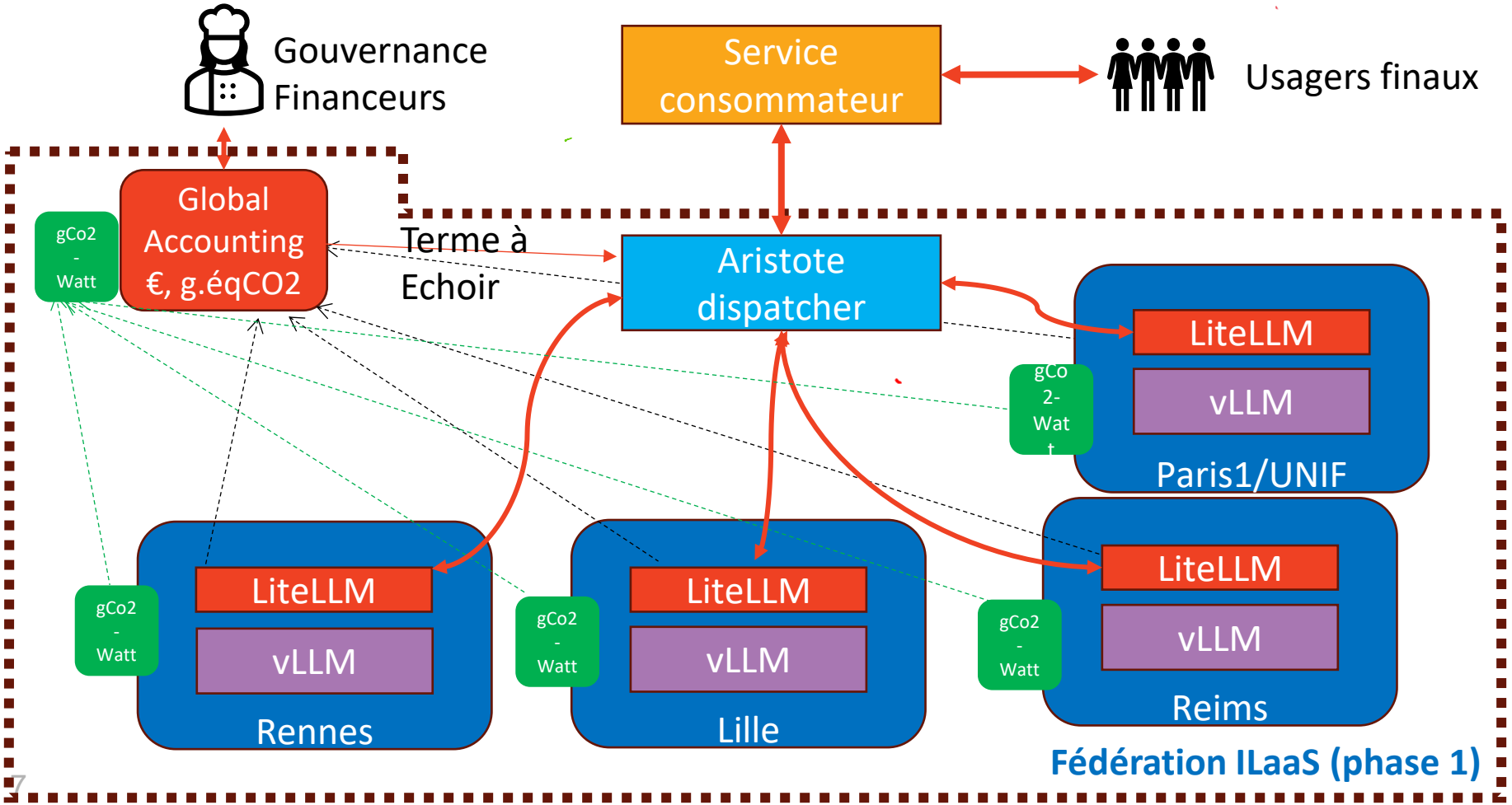
# MUTUALISER LES INFRASTRUCTURES NUMÉRIQUES DES PARTENAIRES POUR LES METTRE À DISPOSITION D'UTILISATEURS BÉNÉFICIAIRES

Objectif du projet ILAAS

esup portail

Photo de Luisa Brimble sur Unsplash

# A QUOI ÇA RESSEMBLE ? (VUE D'ENSEMBLE)



# QUI ON EST ?

- Consortium ouvert, évolutif
  - 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, Université Paris Cité, Université de Lorraine, Université Cote d'Azur, Université de Strasbourg, Université de Haute-Alsace, Université d'Orléans, Université de Tours, Université Marie et Louis Pasteur, BRGM, UPJV, CY Cergy Paris Université, Polytechnique, EHESS,... [INSEREZ LE NOM DE VOTRE ETABLISSEMENT ICI]
- Partenaires potentiels
  - autres DC labellisés
  - mésocentres (lien Mesonet)
  - fournisseurs de cloud souverains
- MESR (DGRI & DGESIP) : financeur et bénéficiaire
- Tiers (entités publiques) : bénéficiaires d'un accès à la fédération

# COMMENT ON FAIT ?

- Objectifs partagés : confiance, souveraineté, soutenabilité, résilience
  - Expérimentation itérative
  - Phases ouvertes selon établissements & collègues souhaitant s'investir
- Projet en plusieurs phases
  - Phase 1 : fin 2024 – début 2025 : expérimenter des usages, dégager des métriques de dimensionnement
    - Déploiement logiciel : aristote-dispatcher
    - Travaux : tableaux de bord, facturation au token, authentification, découplages applicatifs (en lien avec CRISalid)
    - Achat et intégration des GPU : intégrés avec l'existant, exposés en fédération
    - Préparation de l'industrialisation phase 2 : API, mesures (soutenabilité / facturation / modèle économique)
  - Phase 2 : mi-2025 : augmenter l'échelle, industrialiser
    - On en discute… Pour l'instant : support, formation, documentation ; usages & impacts ; consolidation de la puissance de calcul ; modèle économique & circuit financier
  - Phase 3 : à déterminer : sortir de l'expérimentation, selon stratégie & financements

# COMMENT ON FAIT ?

- Mutualisation & échanges
  - GT projet toutes les 2 semaines
  - GT techniques idem
  - Plénière tous les 2 mois
- Esprit constructif
  - Actions constructives
  - Critiques constructives
- Franchise
- Expérimentation
  - Fournir des preuves
  - Droit à l'erreur
  - Bien + Rapide ~~VS Parfait + Lent~~

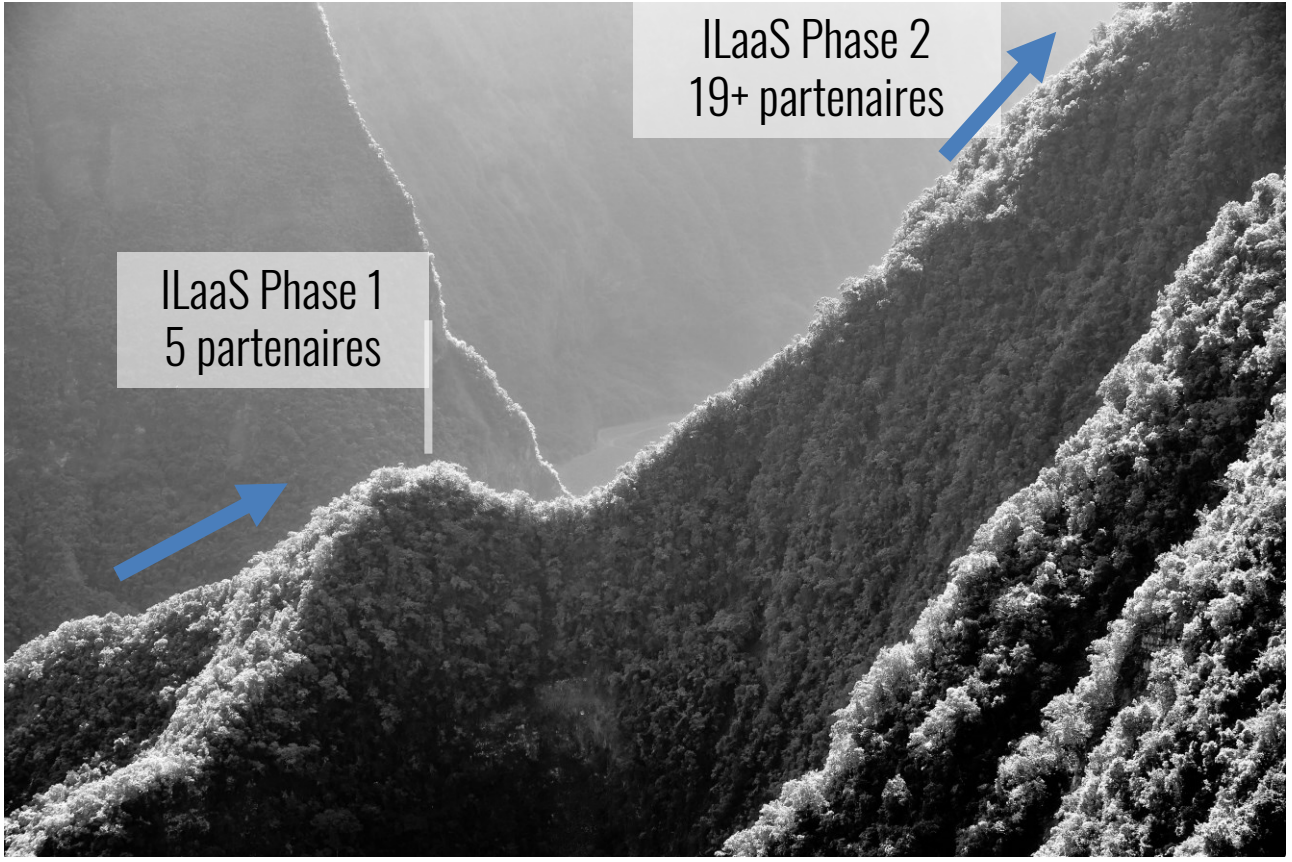


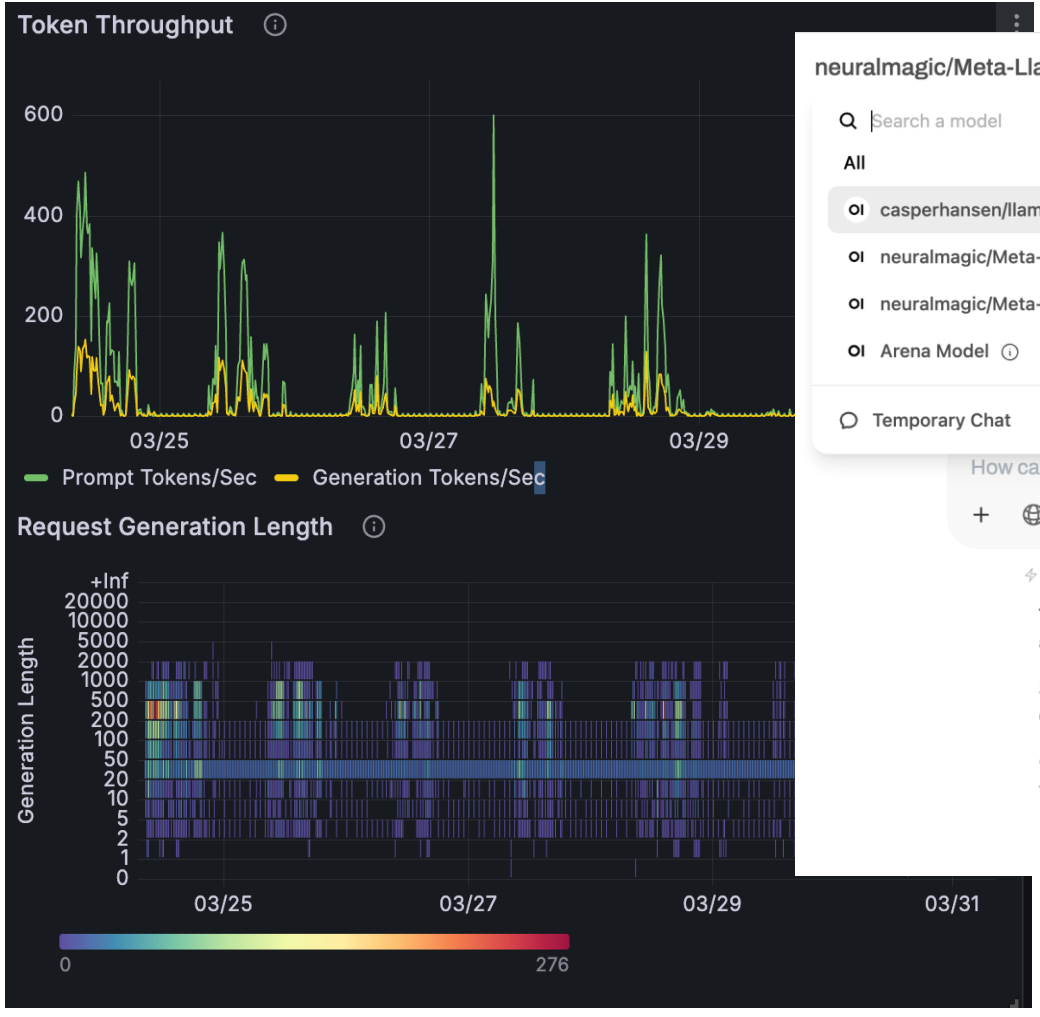

Photo de Philippe Bout sur Unsplash

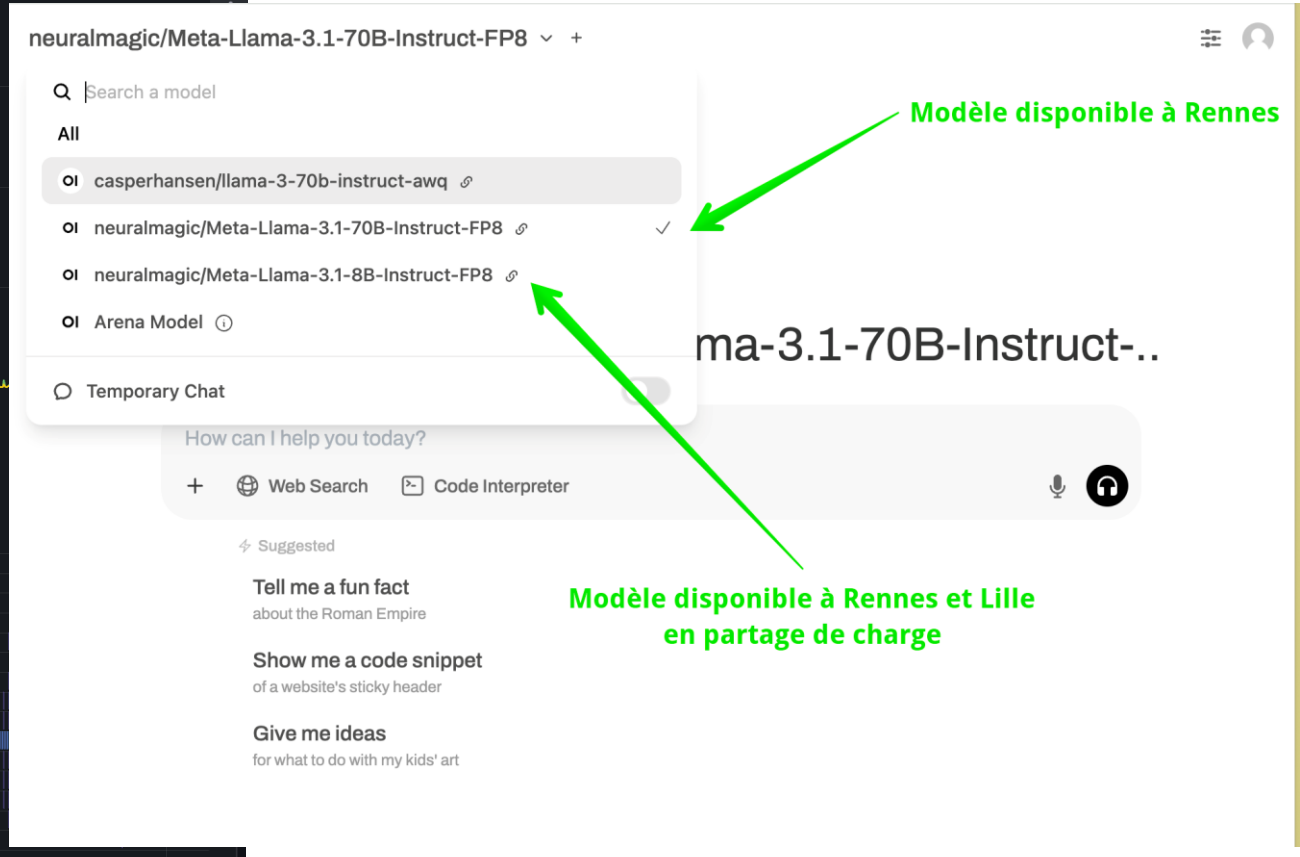
# COMMENT REJOINDRE OU UTILISER ?

- 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 …
  - Recensement par établissement : delta existant VS souhaité (= alimenter la demande de financement phase 2)
  - Engagement au niveau de la gouvernance : VP (numérique), Directeur (DSI)
  - Accord de consortium multi partenaires : à étudier avant que l'encre sèche

- 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 actuels : Inférence LLM as a service (ILaaS)

# PREMIERS DATACENTRES CONNECTÉS



# COMMENT REJOINDRE OU UTILISER ?

- Moyens de communication
  - Liste de diffusion (1-2 messages par mois) : ilaas@groupes.renater.fr
    S'abonner : https://groupes.renater.fr/sympa/subscribe/ilaas?previous_action=info
  - Espace resana dédié pour la documentation
    Sur invitation : https://resana.numerique.gouv.fr/public/perimetre/consulter/1300091
  - Rocket.chat :
    Lien d'invitation : https://rocket.esup-portail.org/invite/nydgcJ
- Pages de présentation
  - Atelier IA esup-portail : https://www.esup-portail.org/wiki/x/GIByW
  - Site web (pour bientôt) : ilaas.fr, en attendant : contact@ilaas.fr
- Echanges synchrones durant les semaines impaires
  - GT technique mardi 16h : https://rendez-vous.renater.fr/auth/ilaas-tech_9f435f-fc08eb-37656d
    prochaine session : mar. 8 avril 2025 (atelier installation)
  - GT projet jeudi 13h : https://rendez-vous.renater.fr/auth/ilaas-proj_c7825d-446d42-b6e721
    prochaine session : jeu. 10 avril 2025 (avancement)

CE SONT RAREMENT LES RÉPONSES QUI APPORTENT LA VÉRITÉ, MAIS L'ENCHAÎNEMENT DES QUESTIONS.

Daniel Pennac, *La fée Carabine*