# Déploiement d'un chatbot d'IA générative souverain à INRAE : retour d'expérience

**Jocelyn DE GOËR** – UMR 0346 EPIA (INRAE-VetAgroSup)

13ème rencontre annuelle du réseau AuDACES

RÉPUBLIQUE FRANÇAISE
*Liberté*
*Égalité*
*Fraternité*

INRAE

# Contexte institutionnel

# Depuis 2022 développement de l'IA générative

- **Apporte de nouvelles façons de travailler**
  - Aide rédactionnelle
  - Aide à la compréhension de textes techniques
  - Recherche et synthèse d'information facilitées
  - Brainstorming
  - Aide au développement (assistance au codage ou vibe coding)
- **Permet d'adresser des problèmes complexes**
  - Automatisation de tâches répétitives et difficiles
  - Modélisation et simulation
  - Aide à la décision
  - Exploration de données textuelles, sons et images

# Des risques multiples

- **Risques réglementaires** liés au non-respect du règlement européen sur l'IA, du droit d'auteur, du droit des affaires, du RGPD, etc
- **Résultats de qualité variable**
  - Des biais à maîtriser
  - Une performance à évaluer
  - La transparence à améliorer
- **Un impact mal évalué sur les communautés**
  - Des rythmes d'adoptions variables avec des inégalités des communautés face aux usages.
  - Des métiers en évolution avec des risques de fractures sociales
- **Risque de « shadow AI »**
- **Un impact environnemental à maîtriser**

# 1 Priorité 1 : une cellule IA pour proposer et porter la politique générale autour de l'IA

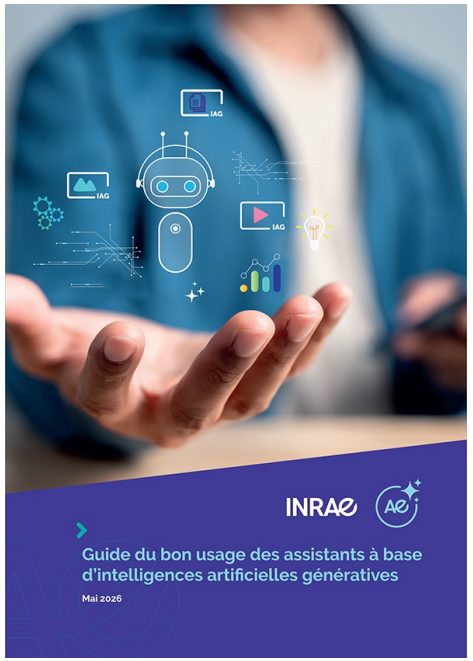
- Une **gouvernance commune à l'échelle de l'institut** au sein d'une cellule IA qui rend compte au Collège de direction
  - Fixer des priorités claires en matière de stratégie IA globale
  - Mettre en place une comitologie adaptée s'inscrivant dans la gouvernance et les pratiques actuelles du numérique

- **Priorisation des cas d'usages** en fonction des bénéfices attendus pour l'institut

# 2 Accompagnement de l'adoption des bonnes pratiques d'utilisation de l'IA

- **Diffuser les bonnes pratiques d'usage de l'IA** à vocation d'assistant et accompagner l'acculturation aux risques juridiques, déontologiques et éthiques

- **Publication le 17/11/2025**: un guide du bon usage des assistants à base d'Intelligences Artificielles Génératives
  - Des recommandations générales
  - Des recommandations par cas d'usage




**hal-05382841**

# 3 Priorité 3 : un accompagnement RH renforcé

- **Une prise en compte et un accompagnement** des impacts de l'IAG sur les conditions et l'organisation du travail, les compétences voire les métiers et les relations interpersonnelles et collectives
- **Un plan de formation** vise à permettre à tous nos agents de monter en compétences en IA, depuis une première acculturation jusqu'à, si besoin, une expertise approfondie :

| Des besoins de sensibilisation | Des besoins d'approfondissement | Un soutien à la transformation |
|---|---|---|
| 1. Apporter des clés pour comprendre **les fondamentaux** de l'IA et les enjeux (*métiers et sociétaux*) contextualisés au cadre INRAE | 2. Accompagner les équipes par la conception et l'organisation d'actions de formation avec une approche différenciée en fonction des métiers, des missions et des publics ciblés<br>Equipes de recherche<br>Métiers de l'appui à la recherche | **3.** Recueillir, analyser, suivre et partager les besoins de montées en compétences et de formations dans le cadre de la cellule IA<br>Soutenir l'action pédagogique de formateurs internes IA |

# Priorité 4 : une offre de service robuste

Proposer une offre de service interne et identifier les pistes de mutualisation à l'échelle de l'ESR

**Mars-septembre 2025 :**

- Construction de scénarios d'une IAG souveraine

**Novembre 2025 – avril 2026 :**

- Mise en place d'une preuve de concept

**Mai 2025 :**

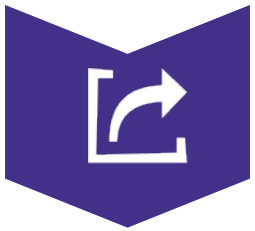
- Ouverture de l'offre de service ARGO

**Étude des pistes de mutualisation avec :**

- Les autres organismes de recherche : BRGM, CEA, Cirad, Ifremer, Inria et IRD notamment
- Les universités notamment via la fédération ILaaS
- Le MESRE

Un **service opéré en interne** afin de maîtriser la sécurité de nos données et les coûts

Un **service robuste et résilient** pour tenir compte d'un contexte technologique en constante évolution



Engager dès que possible une **mutualisation avec des partenaires** au niveau des infrastructures, des compétences, etc.

# 4 Priorité 4 : une offre de service robuste

- **Proposer une offre de service interne et identifier les pistes de mutualisation à l'échelle de l'ESR**
  - Preuve de concept entre novembre 2025 et avril 2026
    - Un premier constat au démarrage des travaux :
    - Pas d'initiatives suffisamment abouties et répondant à nos principaux critères de sécurité et de maîtrise des coûts pour envisager une mutualisation immédiate ;
    - Faute d'alternatives sécurisées, le risque de shadow IA continu

- **Construction de scénarios d'une IAG souveraine**
  - Offre de service ouverte le 11 mai 2026

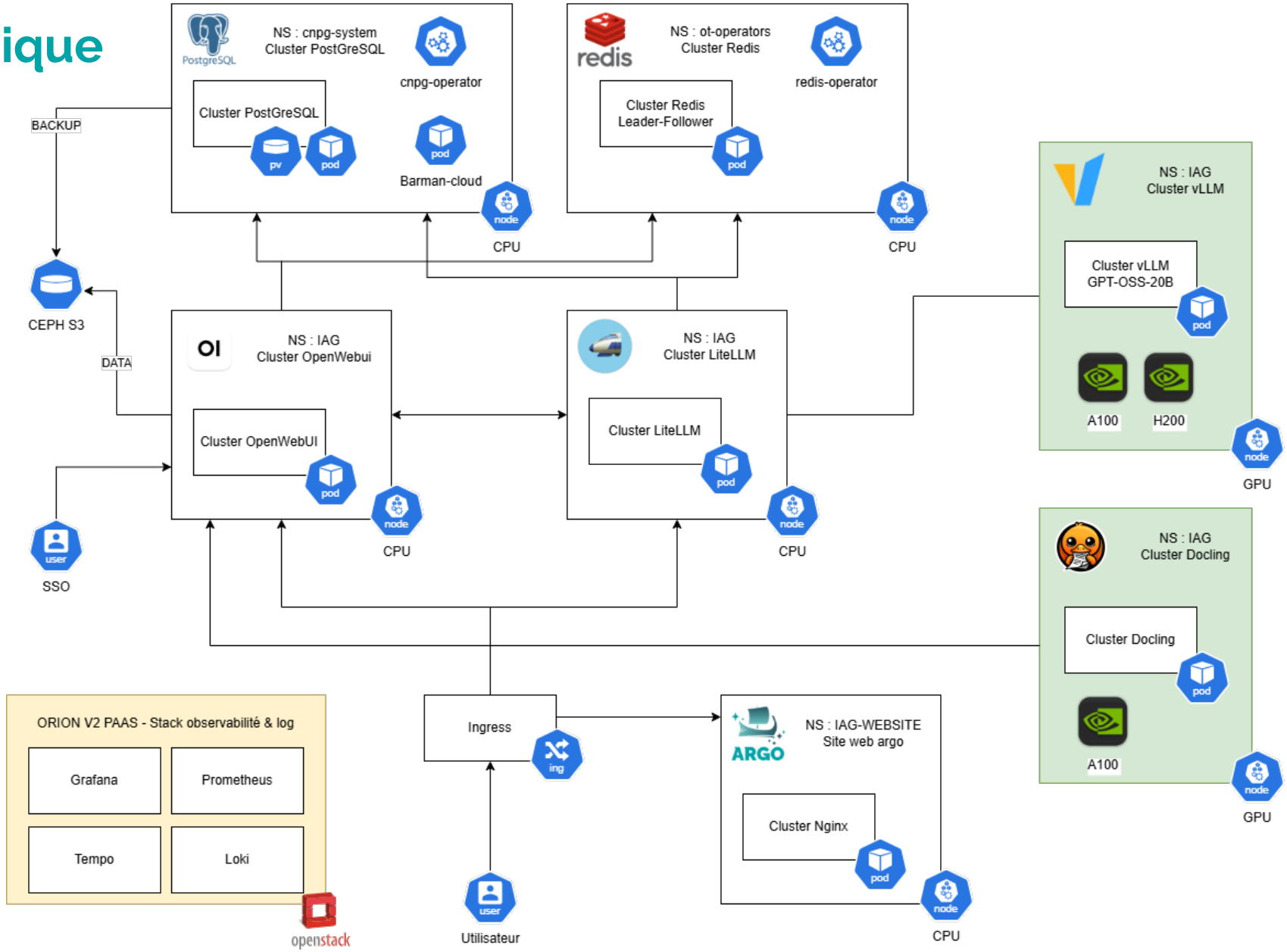
ARGO
CHATBOT IAG INRAE

# Architecture technique

# Architecture technique

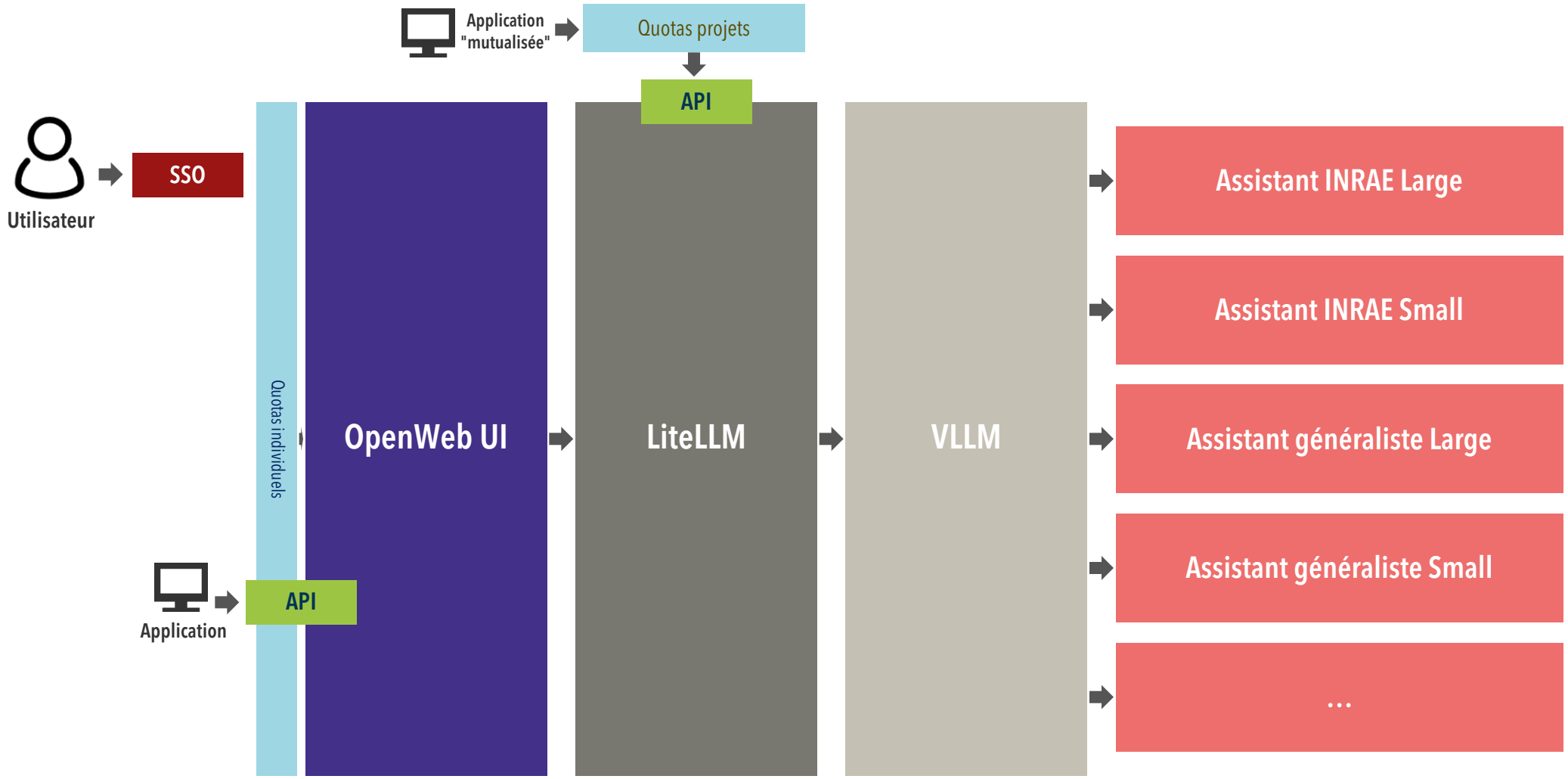
- **PAAS – Kubernetes (Orion V2)**
- **GPU disponibles :**
  - 8x Nvidia A100 40Gb
  - 2x Nvidia H100 96Gb
  - 8x Nvidia H200 128Gb
- **Deux LLM :**
  - GPT-OSS 20B & 120B avec une fenêtre contextuelle de 128k
- **API accessible et ouverte (openai compatible)**

- ⚠ Accessible uniquement sur site INRAE et sur le VPN uniquement




Cluster GPU
PAAS ORION

# Architecture fonctionnelle



# Deux types d'agents conversationnels sont proposés

Agent conversationnel généraliste et non connecté au web

**Aide à la réflexion**
Construire un plan, générer des tableaux récapitulatifs, trouver de nouvelles idées, …

**Traitement des données**
Analyser et exploiter des données et des codes

**Interrogation de documents**
Questionner des documents, synthèse documentaire

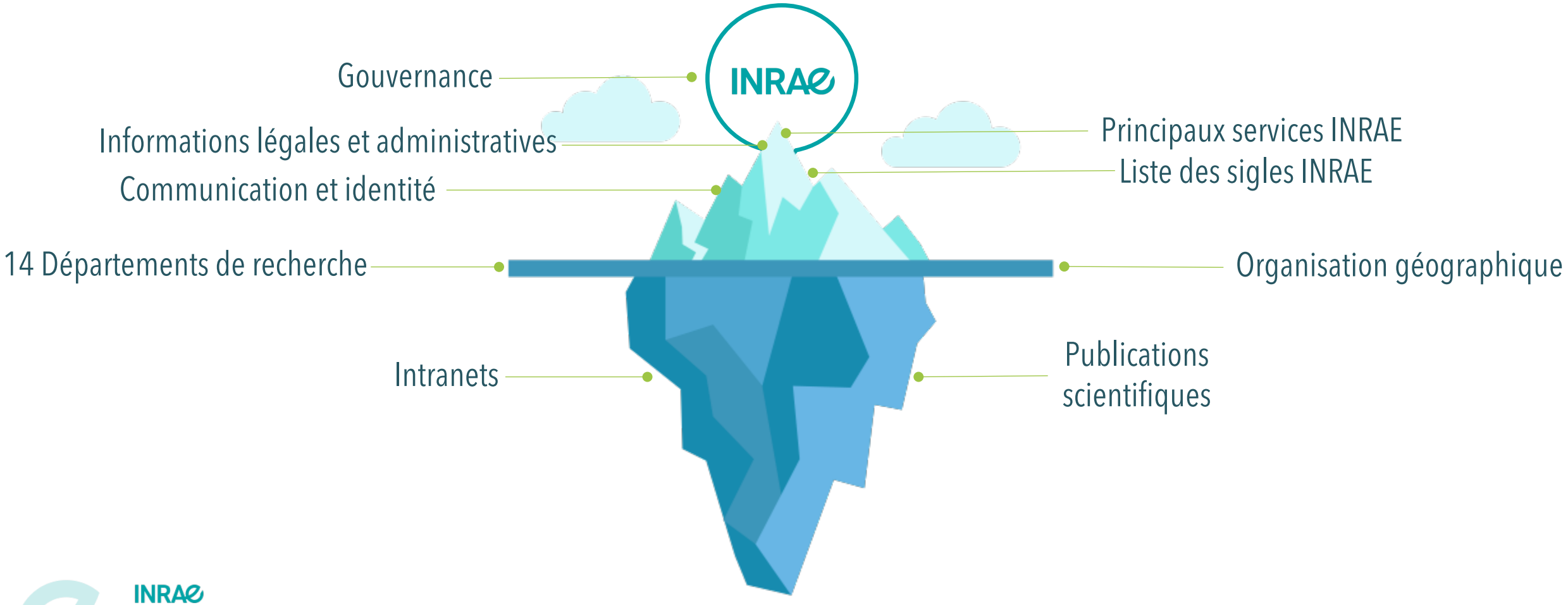
**Traduction**
Traduire un texte ou un document vers de nombreuses langues

**Aide**
Aide au diagnostic, à la résolution de problèmes, identification de dysfonctionnement

**Une interface pour accéder à des modèles généralistes d'intelligence artificielle générative**

# Deux types d'agents conversationnels sont proposés

Agent conversationnel spécialisé sur le contexte INRAE et non connecté au web



# Tests de montée en charge

- **Dimensionnement du service**
  - 12 000 agents INRAE
  - 2 000 utilisateurs simultanés (pic d'utilisation)

| Nb questions simultanées | Taux de réponse | Total tokens/sec | Tokens/sec (individuel) |
|---|---|---|---|
| **500** | 100,00 % | 3 043 | 32,60 |
| **1 000** | 100,00 % | 5 624 | 26,42 |
| **2 000** | 100,00 % | 5 489 | 12,33 |
| **3 000** | 99,60 % | 5 480 | 9,20 |
| **4 000** | 99,70 % | 5 372 | 7,93 |

# La sécurisation des données d'INRAE au cœur des enjeux

**Homologation SSI renforcée**
Centrée sur les risques spécifiques aux IAG

**Pas de réentraînement**
Aucune collecte de données pour réentrainer les modèles

**Maîtrise des données**
Chaque utilisateur a la main sur les conversations et données téléversées sur le chatbot

**Des usages proscrits**
Tels que ceux liés aux domaines RH et pédagogique (entretiens, évaluation, notations)


**Infrastructures**
Hébergement sur le *data center* INRAE de Toulouse, aucune donnée n'est transmise à l'extérieur

**Juridique**
Prise en compte de l'ensemble des exigences juridiques : RIA, RGPD, SSI

**Réseau INRAE**
Accessible sur site INRAE ou via le VPN

**Non connecté au web**
Pour éviter les fuites de données

# Offre de service ARGO

# Fonctionnalités

- **Choix du modèle – sélection libre (GPT-OSS20b et GPT-OSS120b)**
  - Assistant INRAE (Small et Large)
  - Assistant généraliste (Small et Large)
- **Interrogation via API**
  - API compatible OpenAI
- **Deux modes de conversation**
  - Conversations enregistrées : Historique réutilisable, titres automatiques, propriété de l'utilisateur
  - Conversations temporaires : échanges sans trace dans l'historique
- **Discuter avec un document**
  - PDF, txt, Microsoft Office
- **Usage encadré**
  - Mise en place de quotas (10 requêtes/min)

# Mise en production : mai 2026 (premiers retours)

- **Plusieurs profils d'utilisateurs identifiés :**
  - Des utilisateurs en forte demande d'une solution souveraine
  - Des néophytes complets, sans expérience préalable
  - Des utilisateurs inquiets, avec des craintes fondées comme infondées
  - Des utilisateurs avancés
  - Des utilisateurs aux attentes parfois irréalistes

- **Vu par les chiffres :**
  - 3 300 utilisateurs sur 12 000 ont testé la solution un mois après la sortie, portés par un réel effort de communication interne
  - 1 300 utilisateurs actifs sur les 7 derniers jours
  - 200 connexions simultanées récurrentes (pic à 400)

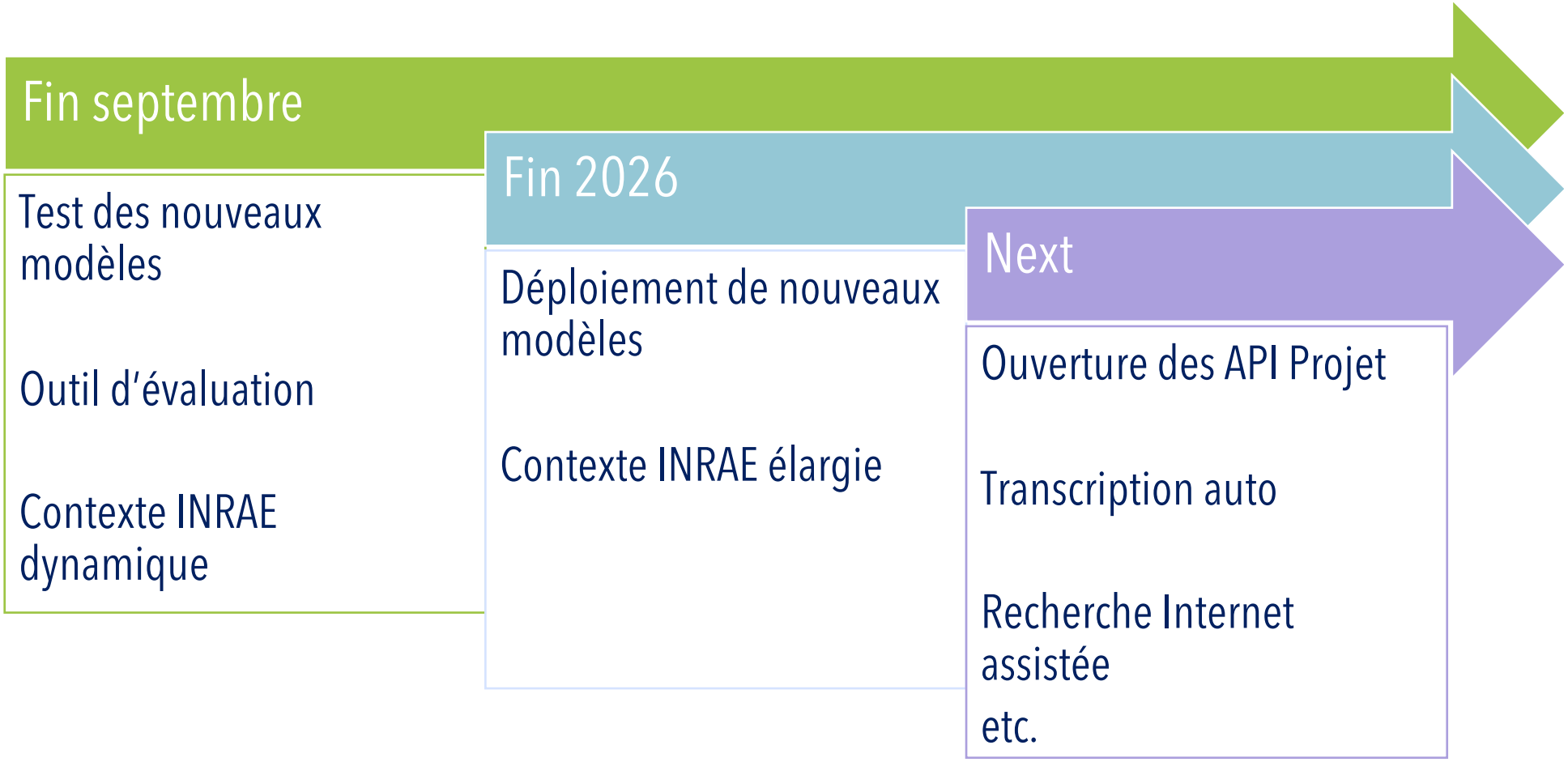
# Mise en production : mai 2026 (premiers retours)

- **Des attentes parfois mal calibrées**
  - De fortes attentes liées à une connaissance approfondie de l'institut, en particulier chez les utilisateurs peu familiers de l'IA

- **Des usages au contraire très matures, portés par une communauté plus restreinte :**
  - Extraction de données issues de sources variées pour alimenter une base de données
  - Assistance au codage
  - Utilisation de fonctions pour le traitement de données issues d'API
  - Un intérêt pour les accès par API plus fort que prévu

- **Des besoins légitimes mais difficiles à satisfaire**
  - Notamment une offre de RAG pour étendre les possibilités des modèles

# Calendrier prévisionnel 2026



# Conclusion

- **ARGO : une IA générative souveraine, désormais opérationnelle**
  - Un service souverain et sécurisé, opéré en interne et ouvert à tous les agents depuis mai 2026
  - Une adoption large, des néophytes aux usages les plus avancés
  - Des besoins qui se précisent : API et RAG en tête
  - **Une trajectoire claire** : nouveaux modèles, contexte INRAE enrichi et mutualisation à l'échelle de l'ESR


# Merci de votre attention

## Une équipe INRAE composée de personnes issues des Directions d'appui et d'Unité de recherche

**Direction des Systèmes d'Information INRAE**

Gaëtane Desgeorge, Aurélien Djian, Sylvain Duchene, Eric Maldonado, Nicolas Raidelet et Herve Toureille

**Direction pour la Science Ouverte INRAE**

Martin Souchal et Alban Thomas

**UMR EPIA**

Jocelyn De Goër et Laurent Cournède

# Déploiement d'un chatbot d'IA générative souverain à INRAE : retour d'expérience

**Jocelyn DE GOËR** – UMR 0346 EPIA (INRAE-VetAgroSup)

13ème rencontre annuelle du réseau AuDACES

RÉPUBLIQUE FRANÇAISE
*Liberté*
*Égalité*
*Fraternité*

INRAE