Forum 55: IA générative pour le HPC : une nouvelle rupture ?

ILaaS – Une infrastructure d'IA mutualisée pour l'ESR

ILaaS
Infrastructure d'IA mutualisée

www.ilaas.fr

Thibault Le Meur
Responsable du développement des usages numérique (CentraleSupélec)
Co-animateur du GT-Services ILaaS

Laurent Morin
Ingénieur de Recherche (IRISA)
Co-animateur du GT-Technique ILaaS

ILaaS
Infrastructure d'IA mutualisée
www.ilaas.fr

Pourquoi ILaaS ?

Photo de Chris Liverani sur Unsplash

MUTUALISER LES INFRASTRUCTURES NUMÉRIQUES DES PARTENAIRES

POUR LES METTRE À DISPOSITION D'UTILISATEURS BÉNÉFICIAIRES

Objectif du projet ILAAS

Mutualiser des infrastructures et faciliter les usages

ILaaS
Infrastructure d'IA mutualisée

Mutualiser des infrastructures de calcul pour l'inférence IA

- Des ressources GPU pré-existantes dans les DSI
- Des ressources GPU disponibles même ponctuellement sur des "slots" dans des Clusters
- Consolider l'hébergement dans des DataCentres labellisés
- Pour les nouvelles ressources de calcul: mutualiser les conseils d'achat pour les pérenniser
- Partager des bonnes pratiques et mutualiser les connaissances

Faciliter divers usages

- Présenter une interface de programmation « standard »
- Multiplexer usages interactifs ou asynchrones
- Multiplexer des usages d'inférence différents: LLM généralistes ou spécialisés, STT, …

ILaaS
Infrastructure d'IA mutualisée
www.ilaas.fr

Projet ILaaS: partenaires du consortium

DC labellisés et établissements : 5 en phase 1 ; **plus de 35 en phase 2**
avec le **soutien de la DGRI**

- Phase 1 : Université de Rennes (porteur), Université Reims Champagne Ardennes, Université Paris I Panthéon Sorbonne, Université de Lille, CentraleSupélec Université 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

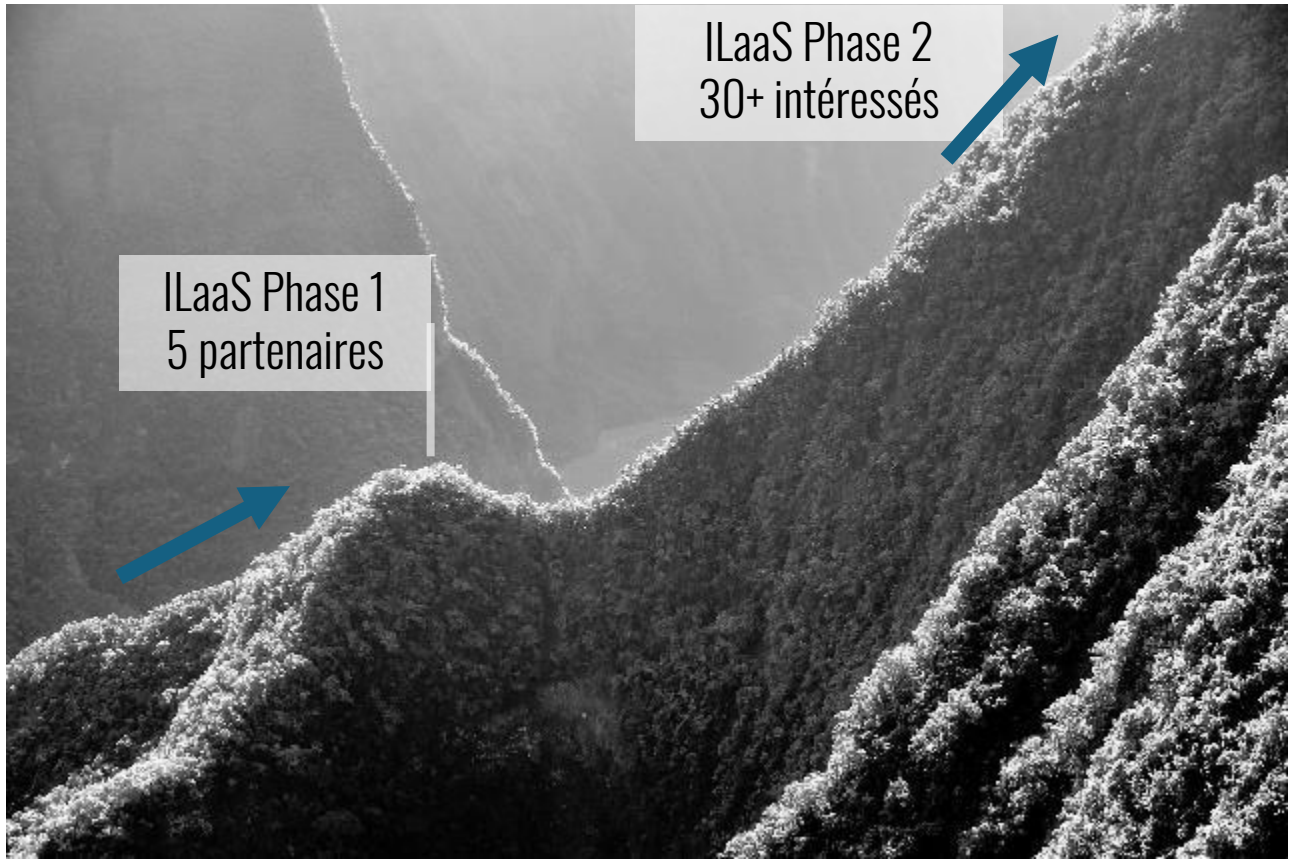


Photo de Philippe Bout sur Unsplash

Développer des usages responsables

- Soutenabilité
  - Équilibre budgétaire: maitrise de coûts
  - 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

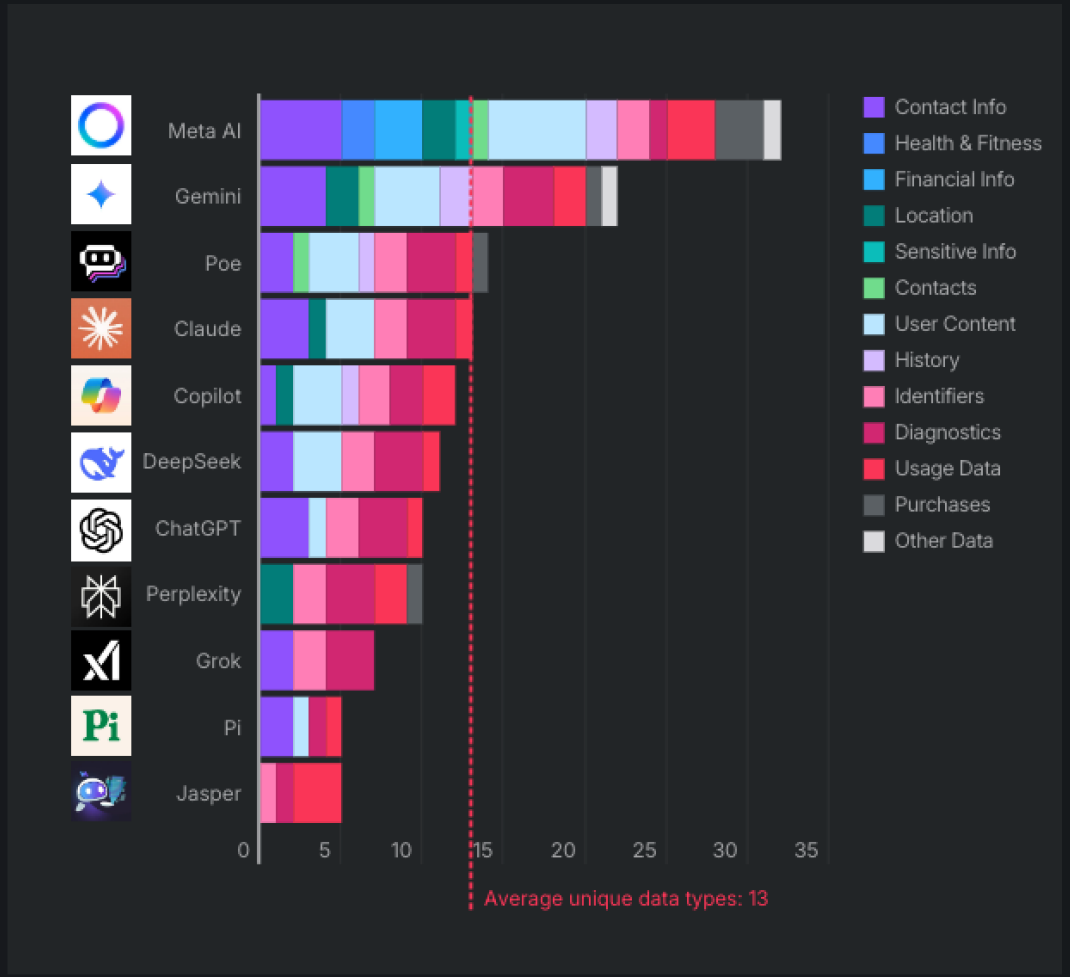
Photo de Steven Wright sur Unsplash

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.



Share

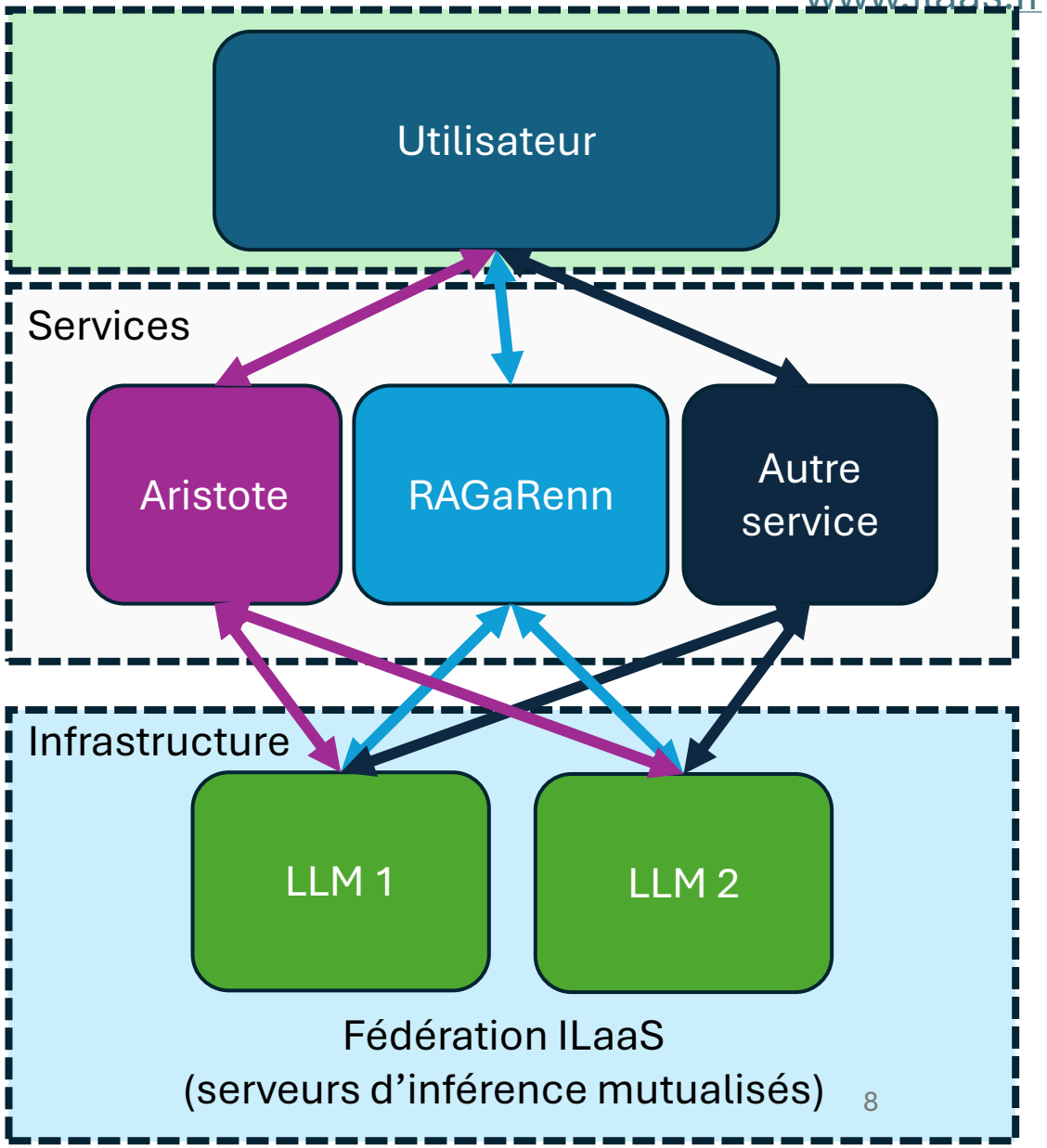
Surfshark®

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

Stratégie ILaaS: ouverture et maîtrise

ILaaS Infrastructure d'IA mutualisée
www.ilaas.fr

- Principe : séparer l'accès à une IA et l'accès à des services utilisant l'IA
- Approche agnostique
  - Choix du/des modèles utilisé(s) (LLM ou STT, embedding, …)
  - 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 maitrisées
  - Mesure des usages réels & impacts associés



ILaaS
Infrastructure d'IA mutualisée
www.ilaas.fr

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
  - Facturation & relations contractuelles : étude en cours

ILaaS
Infrastructure d'IA mutualisée
www.ilaas.fr

Comment utiliser les services d'ILaaS ?

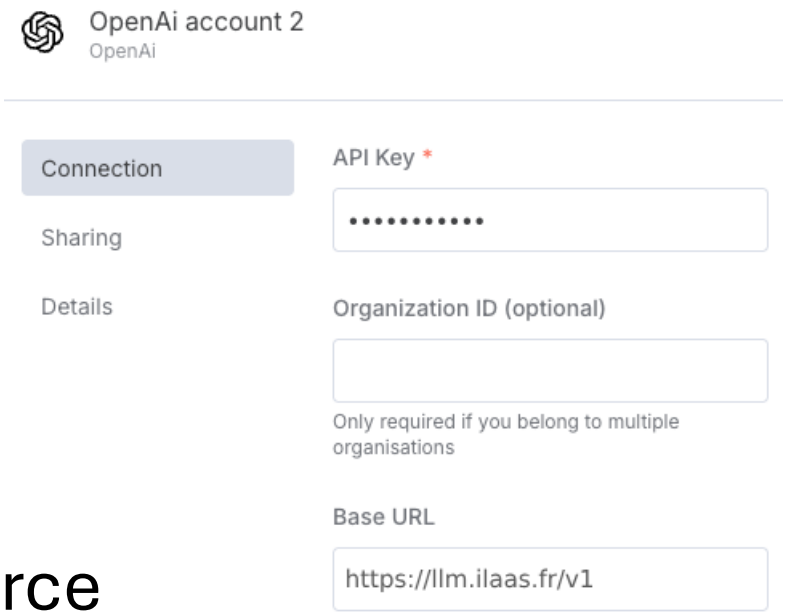
ILaaS dans vos applications, ou bien vos applications dans ILaaS

ILaaS Infrastructure d'IA mutualisée
www.ilaas.fr

- ILaaS dans vos applications
  - 2 interfaces de programmations
    - LLM: complétions
    - STT: génération de transcriptions vidéo, de métadonnées, de comptes-rendus, quiz
  - Prochainement:
    - Embedding
    - Gestion de Collections de documents
    - OCR
- Vos applications dans ILaaS (Marketplace phase 2)
  - Services IA open-source et propriétaire instanciés & gérés pour les clients
  - Exemples
    - Rag@Rennes
    - TIPS aunege
    - En cours de raccordement
      - LibreChat (Université Paris-Cité)
      - Capytale

Usage API d'inférence LLM

- 2 endpoints disponibles:
  /models, /chat/completions
- Une URL: https://llm.ilaas.fr/v1
- Une clé d'API pour y accéder



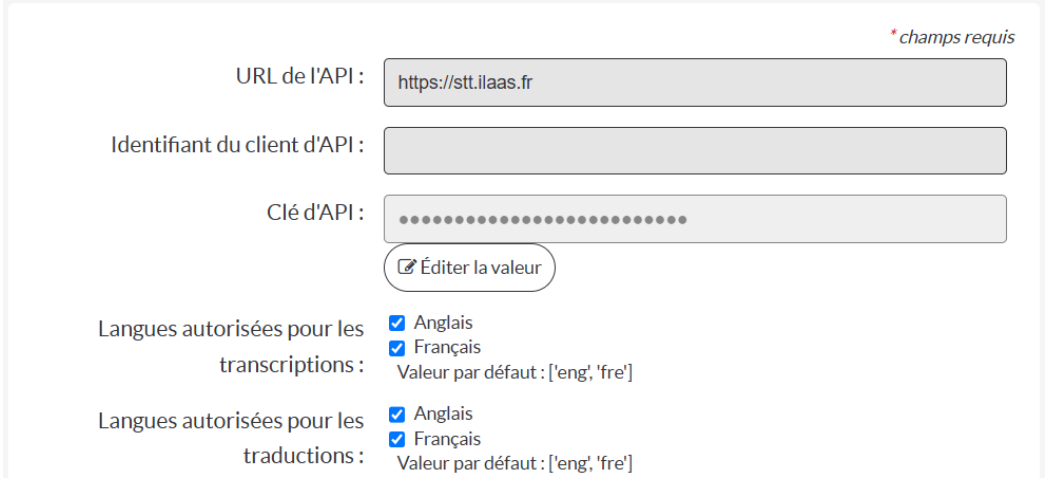
- Opensource
  https://github.com/CentraleSupelec/aristote-dispatcher

```python
import requests

# Configuration
api_key = "APIKEY"
base_url = "https://llm.ilaas.fr/v1"
headers = {
    "Authorization": f"Bearer {api_key}",
    "Content-Type": "application/json"
}

# Récupération des modèles
response = requests.get(f"{base_url}/models",
                        headers=headers)
models = response.json()
model_name = models['data'][0]['id']

# Demande de complétion
prompt = "Bonjour, comment allez-vous ?"
payload = {
    "model": model_name,
    "messages": [{"role": "user", "content": f"{prompt}"}],
    "stream": False,
    "max_tokens": 100
}

response = requests.post(f"{base_url}/chat/completions",
                         headers=headers, json=payload)
result = response.json()

# Affichage de la réponse
print(prompt)
print(model_name)
print(result['choices'][0]['message']['content'])
```

ILaaS
Infrastructure d'IA mutualisée
www.ilaas.fr

API de Transcription

- Une API dédiée pour accéder au service
  - URL de base
  - Client_id et Secret
- Comment l'utiliser
  - Configuration dans un logiciel compatible (Ubicast Nudgis, ESUP-Pod)
  - Implémentation d'un client
- Opensource



https://pod.esup-portail.org/

https://github.com/CentraleSupelec/aristote-api/blob/main/docs/AristoteAPI_Client_guide.md

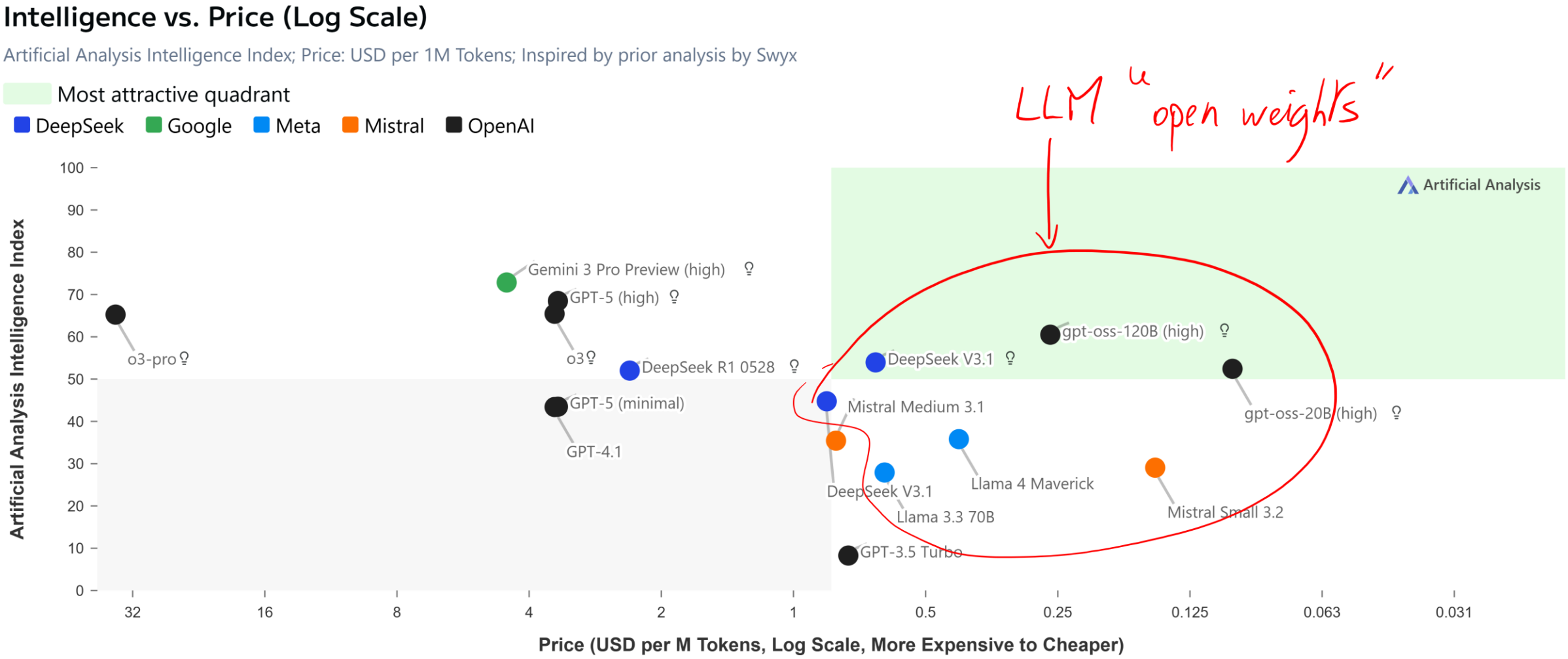
ILaaS
Infrastructure d'IA mutualisée
www.ilaas.fr

Cycle de vie des modèles gérés au niveau de la fédération

Choisir ses modèles d'inférence

Des modèles suffisamment performants pour un impact raisonnable (budgétaire, environnemental)

ILaaS
Infrastructure d'IA mutualisée
www.ilaas.fr



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 26 novembre 2025

Gouvernance nécessaire sur la sélection des modèles

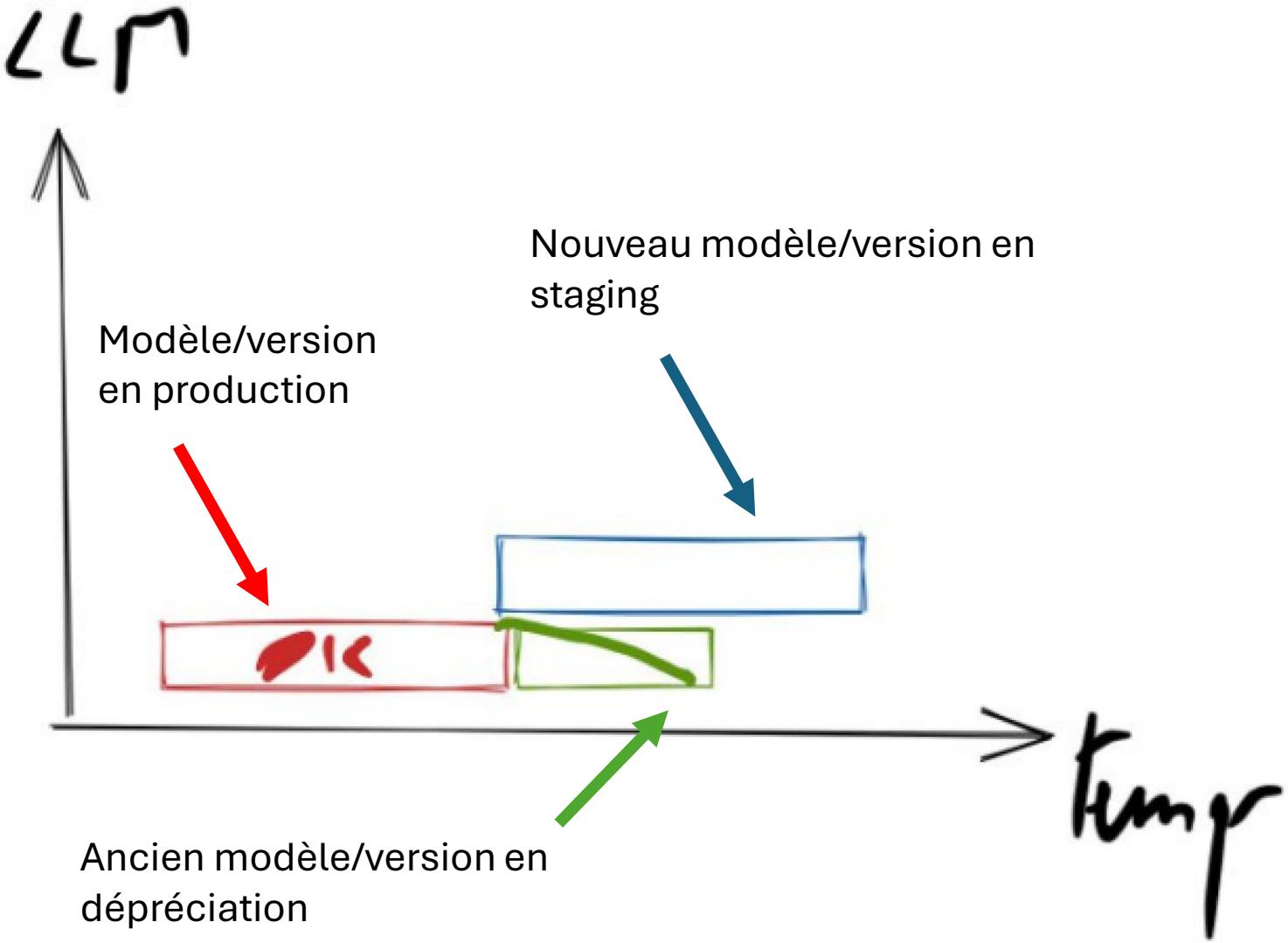
ILaaS
Infrastructure d'IA mutualisée
www.ilaas.fr

- Choix d'un ensemble de modèles pour couvrir les fonctionnalités
  - LLM généraliste
  - LLM multimodal
  - LLM de raisonnement
  - LLM pour réaliser des agents d'IA

- Choisir les paramétrages en fonction des usages
  - Quantification,
  - Taille de la fenêtre de contexte,
  - Template spécifiques aux modèles (notamment pour l'agentique)

- Offrir une continuité de service
  - Disposer du même modèle sur plusieurs DC
  - Assurer qu'un modèle sera suivi dans le temps (un changement pouvant avoir des conséquences sur les applications)

Cycle de vie des modèles

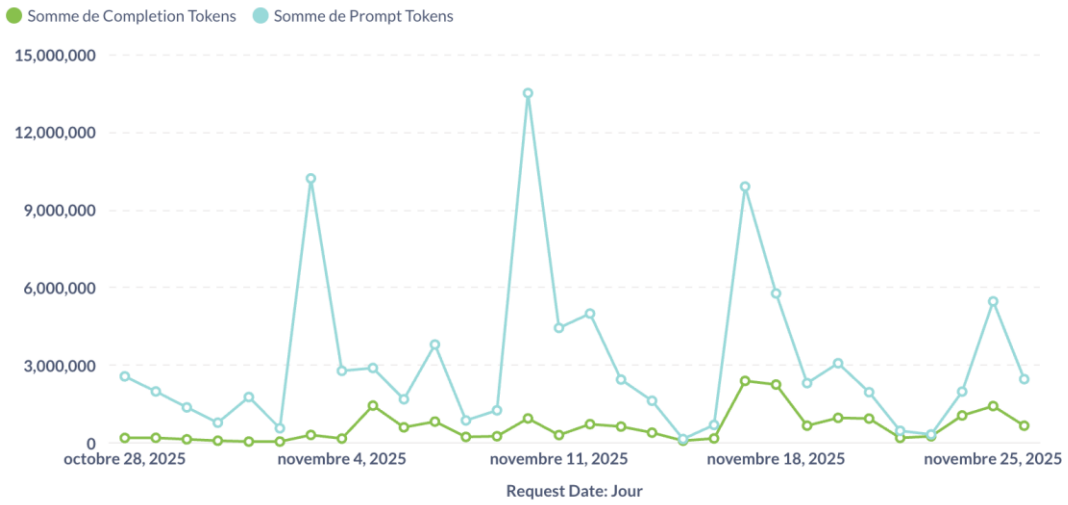


- Catégories de modèles
  - généraliste, multimodale, raisonnement, agentique
  - 6 modèles
    - gpt-oss-120b
    - mistral-small-3.2-24b
    - llama-3.1-8b
    - llama-3.3-70b
    - qwen-2.5-3b
    - qwen-3-30b
- Déploiement
  - 7 serveurs d'inférence
  - 5 universités: Rennes, Reims, Paris1, Côte d'Azur, Strasbourg
  - > 128000 requêtes / 60j

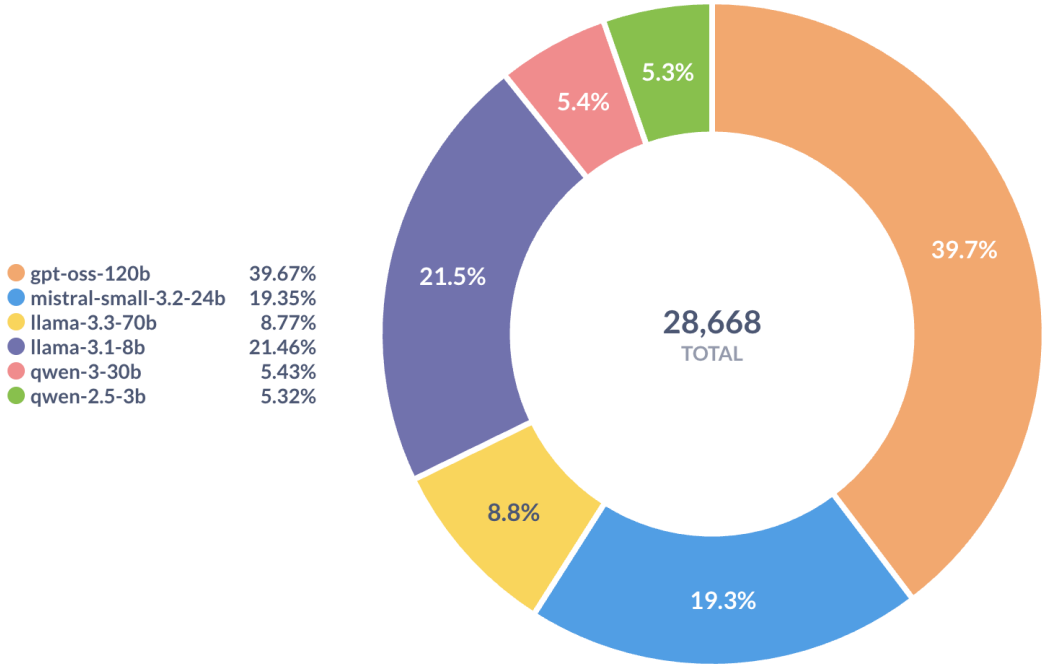
ILaaS
Infrastructure d'IA mutualisée
www.ilaas.fr

Quelques statistiques d'usage

Evolution du traitement en jetons sur 30 jours



Répartition du nombre de requêtes par modèle sur 10 jours



ILaaS
Infrastructure d'IA mutualisée
www.ilaas.fr

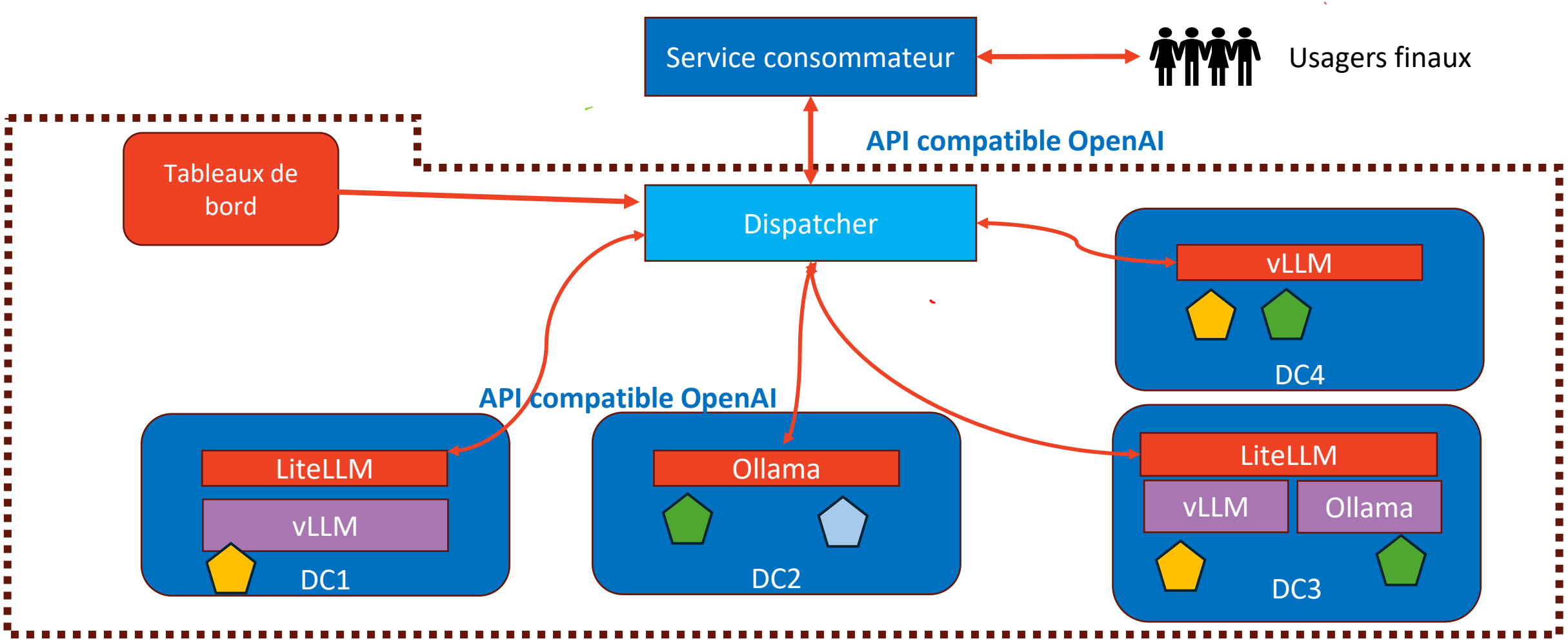
Sous le capot du service d'inférence LLM

LLM.ilaas.fr

ILaaS unifie l'accès à plusieurs modèles hébergés sur plusieurs Datacentres

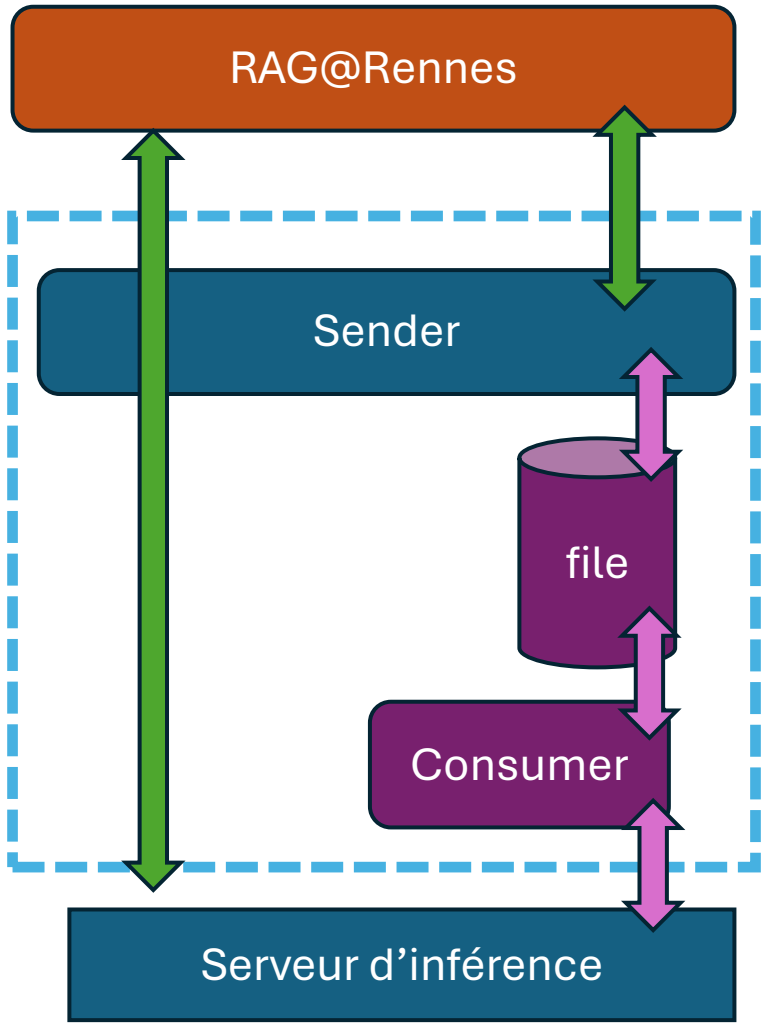
ILaaS Infrastructure d'IA mutualisée
www.ilaas.fr

- Serveurs d'inférence compatibles OpenAI



Le dispatcher est un service de mise en relation

- entre un « utilisateur »
- qui utilise un point d'entrée unique
- … par l'intermédiaire de files d'attentes
- … gérées par un « sélecteur »
- et un serveur d'inférence

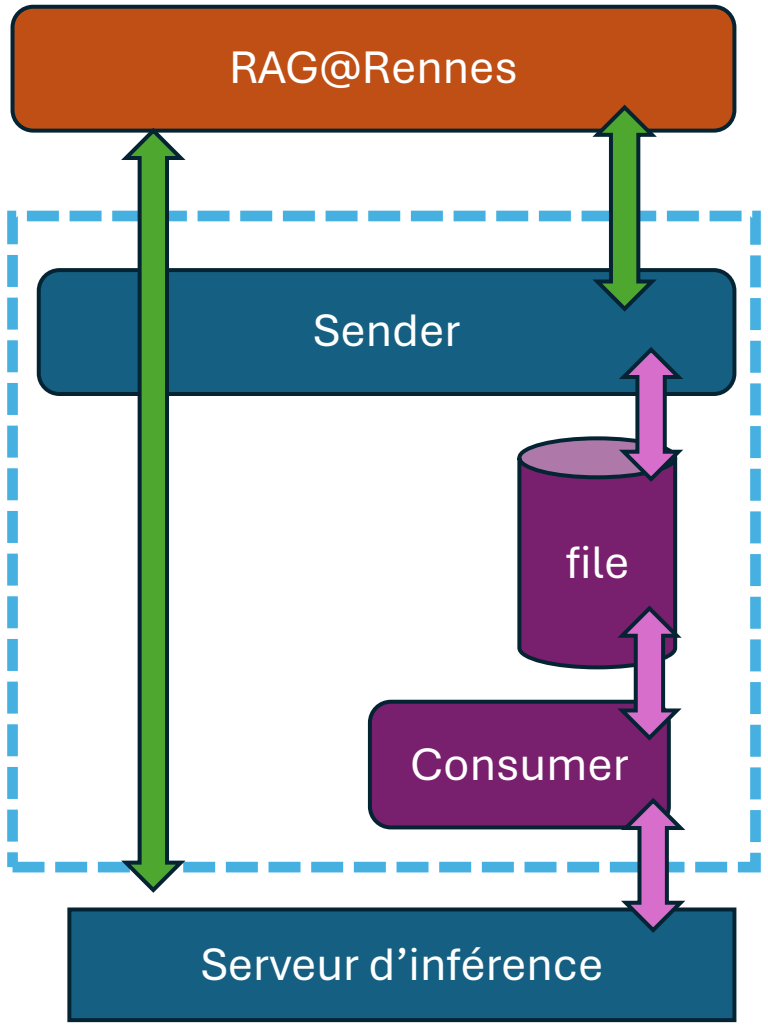


ILaaS
Infrastructure d'IA mutualisée
www.ilaas.fr

Exemple du déploiement Rag@Rennes

- Déploiement Applicatif : RAGaRenn Apps = instances Open Web UI
  - 358 instances mono-usager (legacy, 35 actifs)
  - 98 instances pour des groupes d'usagers 73 actifs dont :
  - 24 en SSO ESR France
  - 1 EduGain
- Utilisateurs
  - +1968 inscrits dans l'instance principale ESR
  - > 3000 inscrits au global

- Déploiement Inférence : RAGaRenn LLM
  - Administration par Tâches (Slurm)
  - 8 GPUs en propre
    - 2xH200, 2xH100 (80Gb), 4xV100
    - 4 de plus en débordement (H100)
  - Utilisé par des tiers prestataires de services
  - Un des nœuds de la fédération ILaaS

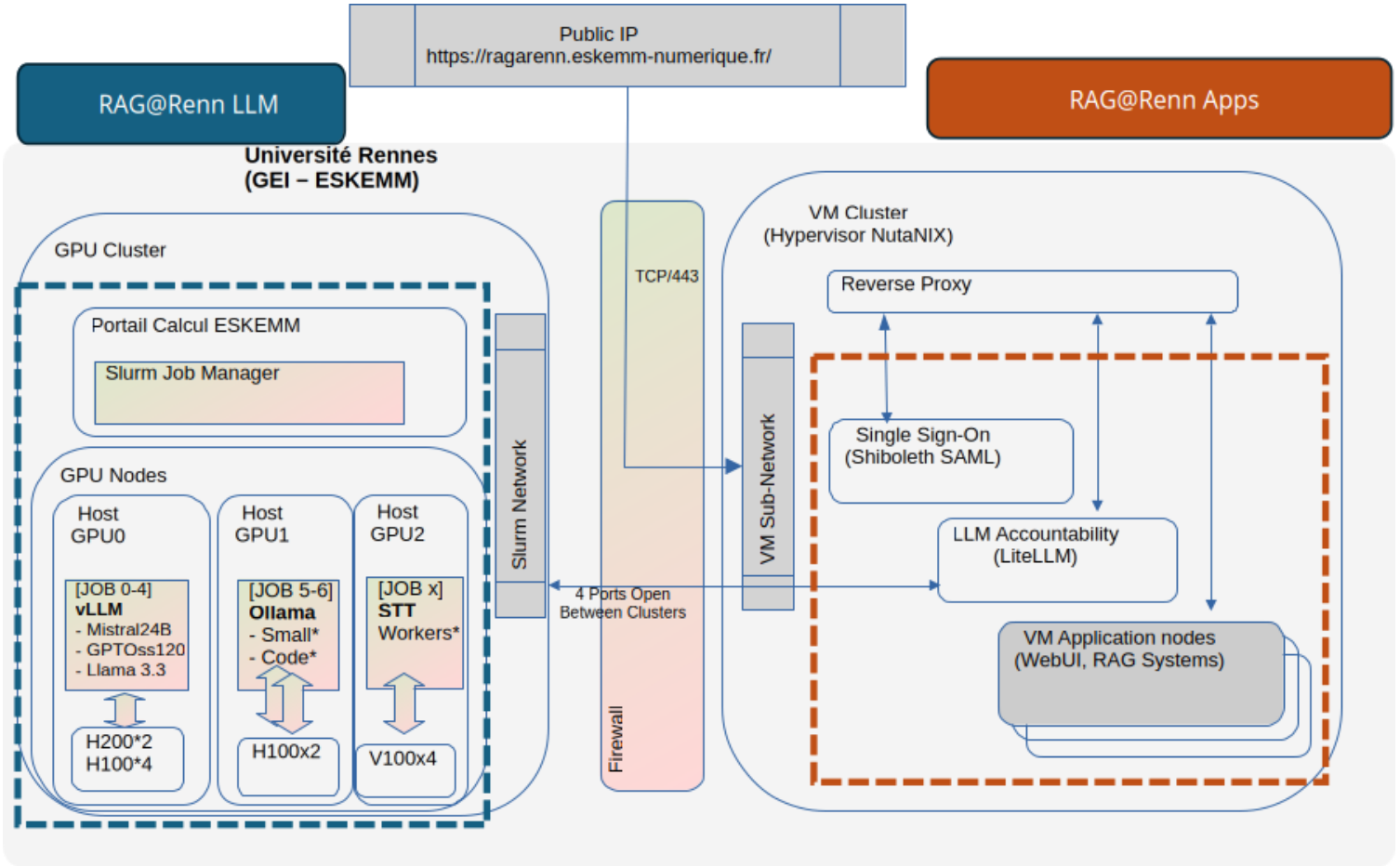


https://ragarenn.eskemm-numerique.fr/

Rag@Rennes: Déploiement sur 2 clusters

- Accès et contrôle d'accès avec un proxy unique (HaProxy)

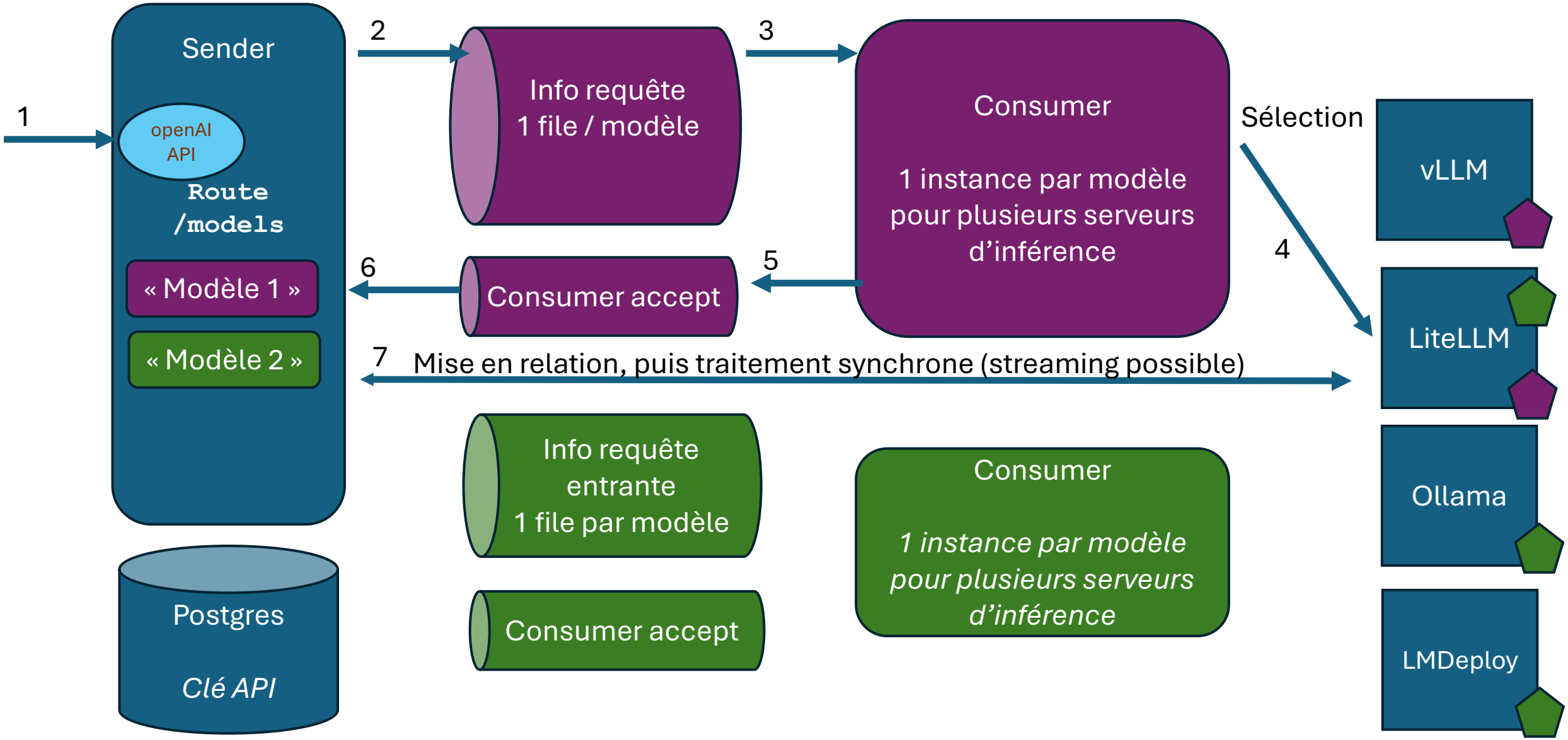
- Cluster VM (NutaNix)
  - Gestion LLM (LiteLLM)
  - Gestion SSO Renater (ESR, EduGain)
  - Déploiement automatique d'instances OpenWebUI
  - Débordement Ilaas intégré
- Cluster GPU
  - Déploiement Modules et containers et (Python et Apptainer)
  - LLMs alloués par des politiques statiques
  - Transcription/Traduction Audio – (STT) alloués dynamiquement selon disponibilités (« spot »)
  - Accès par 4 lignes TCP dédiées pour chaque



Traitement simplifié d'une requête

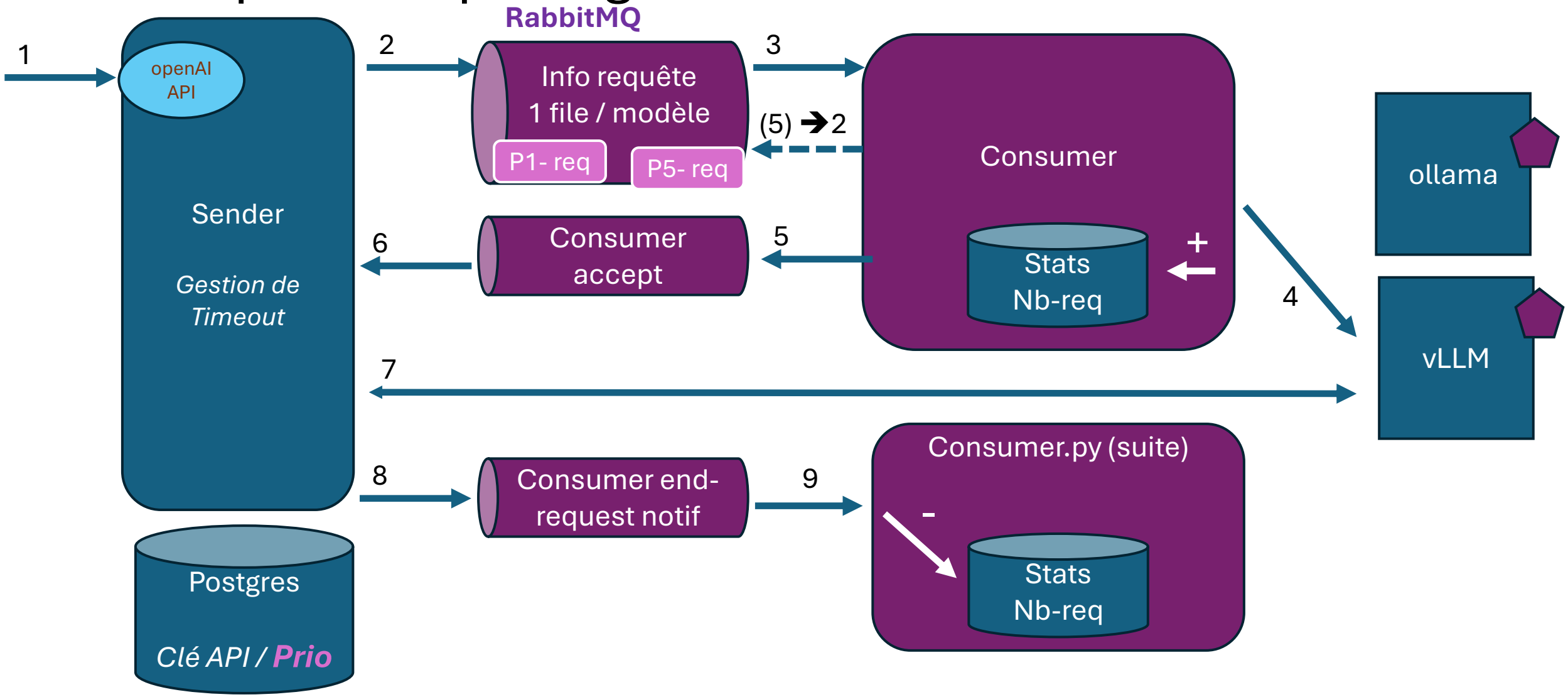
Stratégie round-robin sans requeuing

ILaaS
Infrastructure d'IA mutualisée
www.ilaas.fr



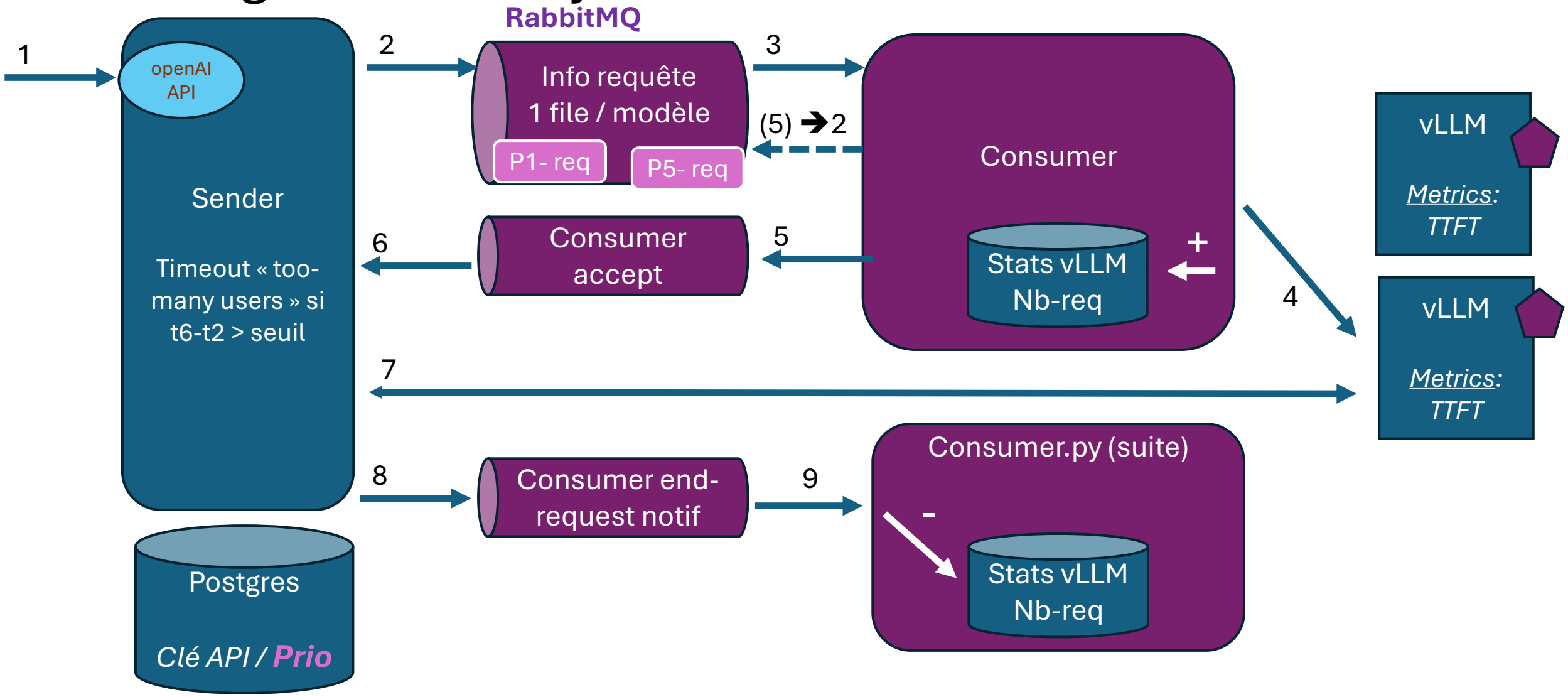
Gestion de priorités

Politique de requeuing



Optimisation du routage

Stratégie Least-Busy



Stratégies et Politiques de requeue

ILaaS
Infrastructure d'IA mutualisée
www.ilaas.fr

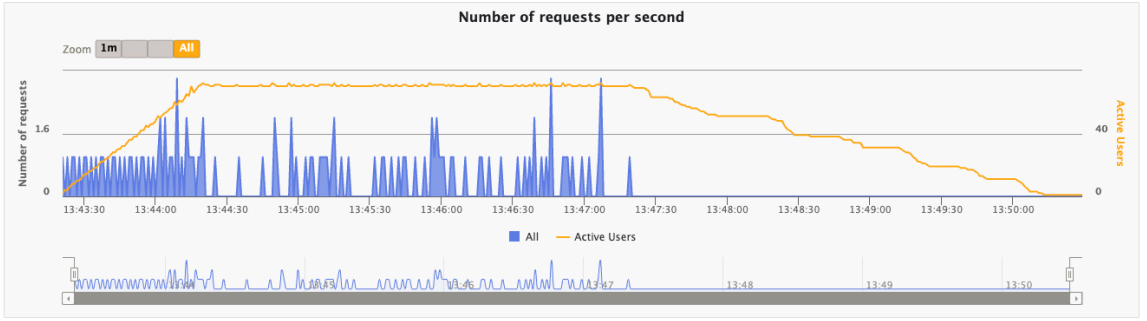
Fonctionnement	Stratégie de sélection	Politique de requeue
Round Robin simple Gestion de la disponibilité Répartition de charge simple	Round Robin	Sans requeuing
Round Robin Gestion de la disponibilité Répartition de charge simple Gestion de priorités	Round Robin	Basé sur le nombre de requêtes simultanées
Least busy (vLLM) Gestion de la disponibilité Répartition de charge optimisée Gestion de priorités Garantie de performances dans certaines limites	Basé sur statistiques vLLM	Basé sur statistiques vLLM et sur le nombre de requêtes simultanées

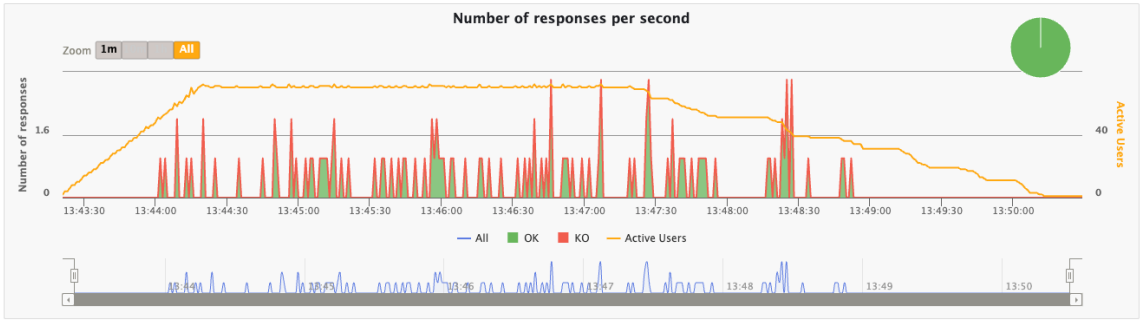
Architecture simulation de la gestion des priorités

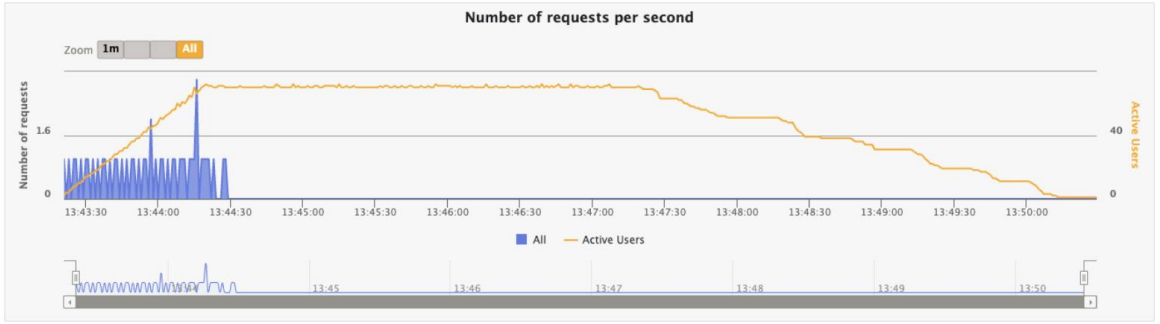
Exemple de test réalisé

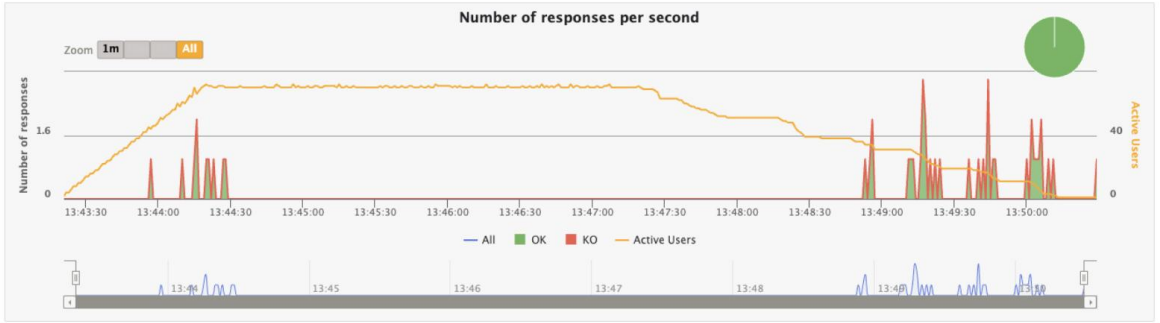
- Rampe de ~1mn pour passer de 0 à 50 utilisateurs simultanés
- Plateau à 50 utilisateurs actifs pendant ~3mn
  - connexion ouverte, nouvelle requête à chaque réponse obtenue
- Rampe finale diminuant le nombre total d'utilisateurs
  - Pas de nouvelle requête après la dernière réponse

➔ permet de combiner usages interactifs et batch





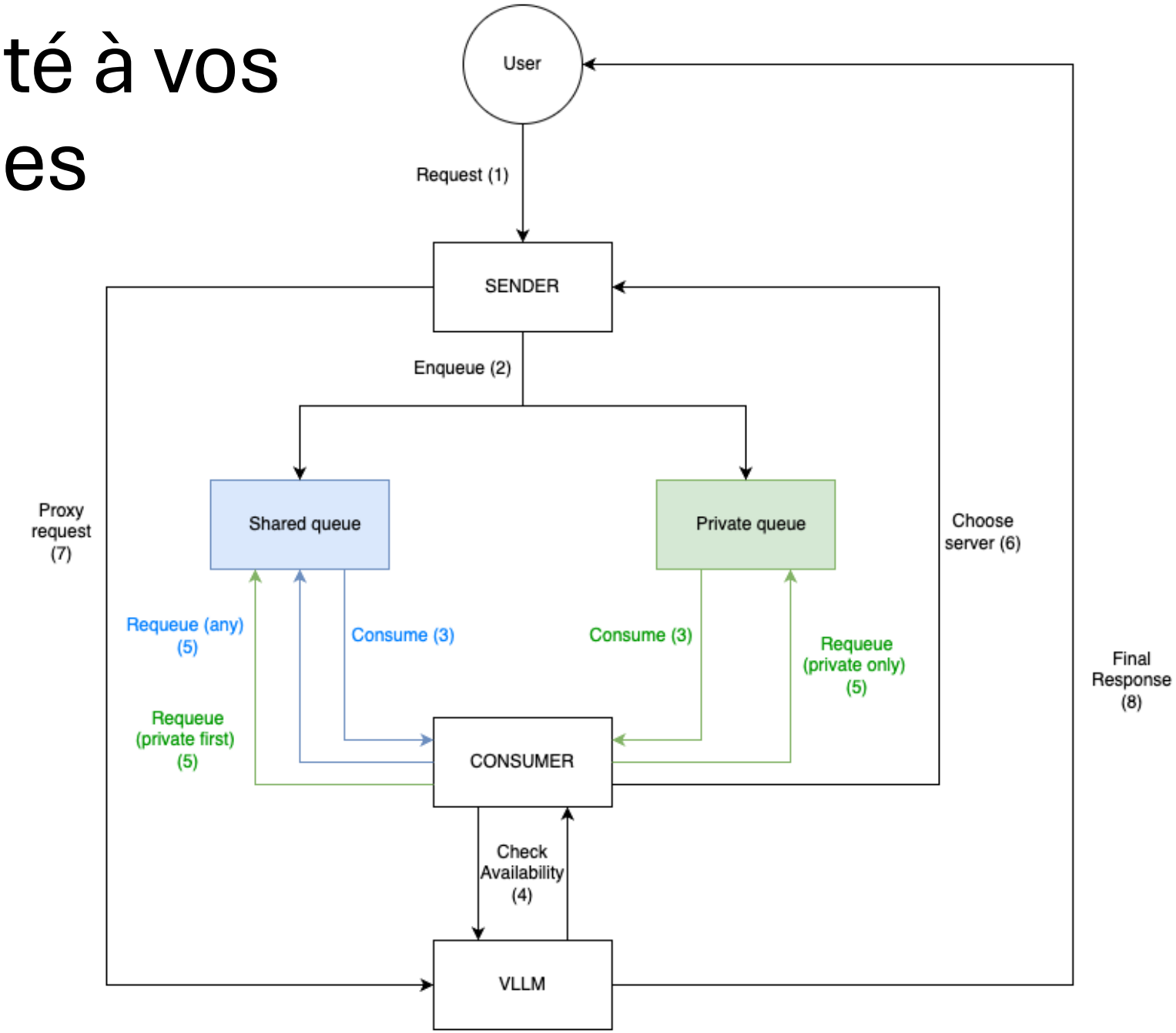




Accéder en priorité à vos propres ressources

- Vous fournissez de la puissance d'inférence…
- Vous utilisez llm.ilaas.fr…

➔ Vous avez un coupe-file pour un accès privilégié à vos serveurs



ILaaS
Infrastructure d'IA mutualisée
www.ilaas.fr

Hébergement et ajout/suppression de modèle

- Les composants Dispatcher sont hébergés dans un cluster Kubernetes (actuellement OVH , hébergement ESR en cours d'étude)
  - Sender, Consumer, BdD, Dashboards

- Ressources de calcul sont hébergées dans les DC Labellisés ESR
  - Pour mutualiser un serveur d'inférence
    - Remplir un formulaire technique: type de serveur, nom du modèle, quantification, taille de la fenêtre, option de tools-calling, URL d'accès et clé d'accès
    - Ouverture de flux @IP des nœuds du cluster Kubernetes ILaaS vers le serveur
    - Déploiement de la nouvelle configuration via Helm

ILaaS
Infrastructure d'IA mutualisée
www.ilaas.fr

Hébergement et ajout/suppression de modèle

- Values de déploiement ILaaS (Kubernetes)

```
models:

- name: mistral31-24b
  model: "mistralai/Mistral-Small-3.1-24B-Instruct-2503"
  vllm_servers:
    - url: https://dc1.univ-xxx.fr/vllm
      token: <secret>
- url: https://dc1.univ-xxx.fr/vllm
      token: <secret>
  routing_strategy: least-busy

- name: llama31-8b2
  model: "meta-llama/Llama-3.1-8B-Instruct"
  vllm_servers:
    - url: https://dc2.univ-yyy.fr
      token: <secret>
  routing_strategy: round-robin
```

Least-Busy limité à des serveurs vLLM actuellement

Indépendant des serveurs

ILaaS
Infrastructure d'IA mutualisée
www.ilaas.fr

Sous le capot du service d'inférence STT

STT.ilaas.fr

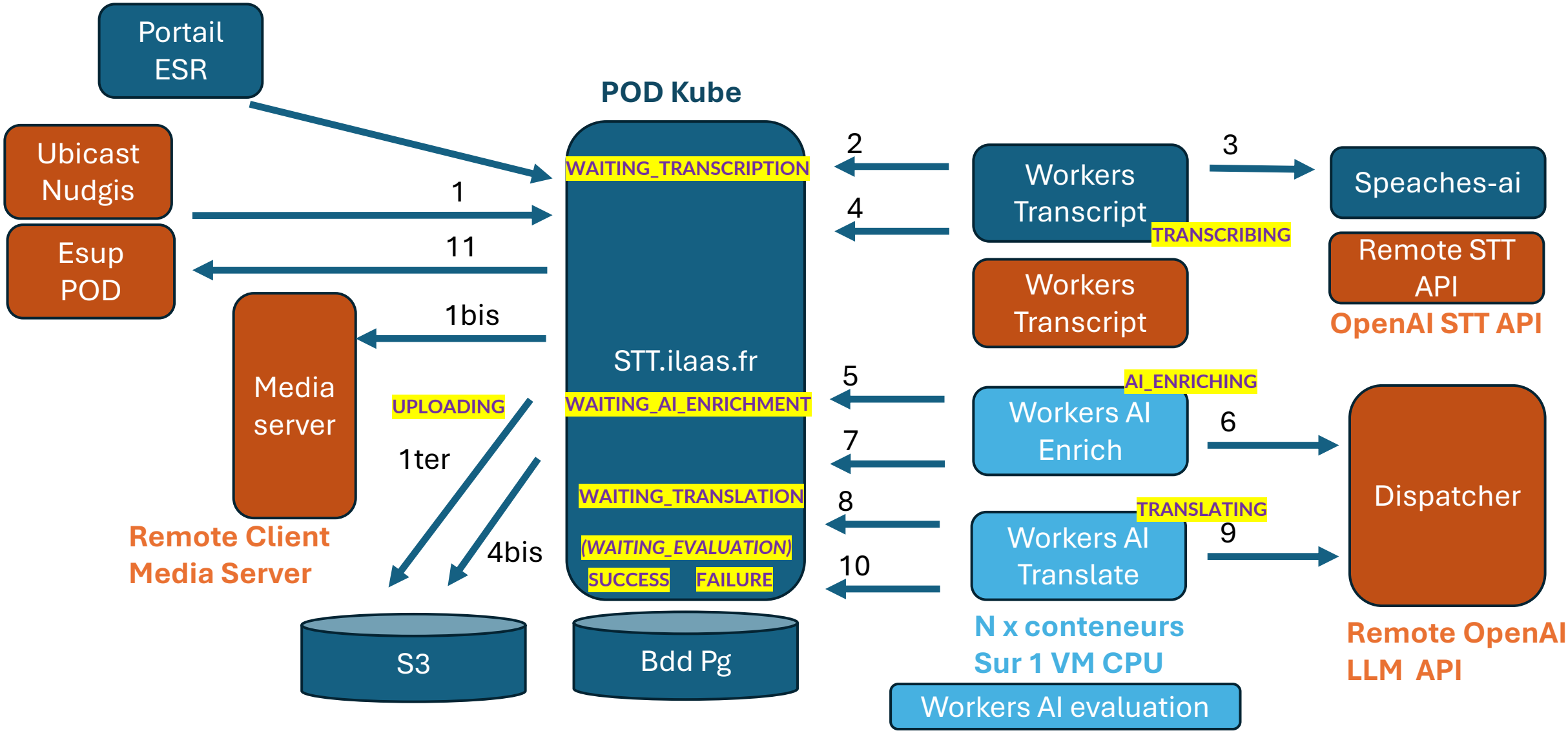
ILaaS
Infrastructure d'IA mutualisée
www.ilaas.fr

Retranscription audio/vidéo et plus

- Service de retranscription asynchrone: produit des transcripts et sous-titrages (VTT, SRT)
- Permet aussi:
  - La traduction
  - L'utilisation d'un LLM pour générer optionnellement
    - des métadonnées (titre, description, mots clés)
    - Des comptes rendus ou des quiz
- Basé sur la technologie Aristote
- Hébergé dans un cluster Kubernetes (actuellement chez OVH)

STT.ilaas.fr

ILaaS Infrastructure d'IA mutualisée
www.ilaas.fr



ILaaS
Infrastructure d'IA mutualisée
www.ilaas.fr

Conclusion

ILaaS
Infrastructure d'IA mutualisée
www.ilaas.fr

Ouverture de nouveaux services API

- Instanciation d'un nouveau portail d'API pour
  - L'embedding
  - La gestion de collections de documents (base de données Qdrant)
  - L'OCR

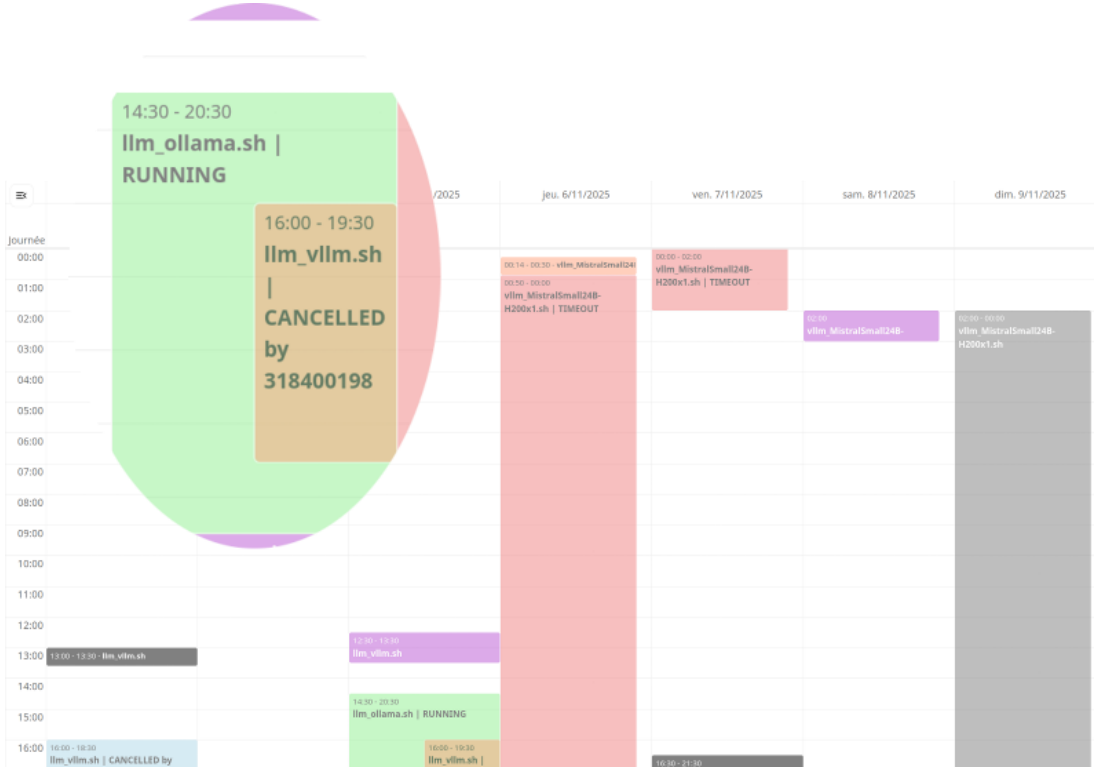
- Basé sur la technologie OpenGateLLM qui propulse AlbertAPI (DINUM) https://github.com/etalab-ia/OpenGateLLM

- Et à terme, ajout/suppression dynamique de modèles
  - Possibilité d'ajouter un serveur d'inférence pour les membres du consortium authentifiés
  - Gestion des « modèles critiques » (service minimum)

ILaaS
Infrastructure d'IA mutualisée
www.ilaas.fr

Travaux en cours

- Transfert de la gestion des capacités à la DSI
- Déploiement de services métiers (agents, Pipeline RAG, …)
- Etoffer l'offre applicative directe aux utilisateurs
- Intégrer des modèles spécialisés et frugaux sur ILaaS



Exemple de déploiement automatiques de LLMs avec Slurm

ILaaS
Infrastructure d'IA mutualisée
www.ilaas.fr

Tous différents, tous ensemble

- Les ressources disponibles et les déploiements sont différents
- ILaaS vise à fédérer des ressources, mais aussi des compétences et des usages
- Diversification des hébergements
  - ILaaS sait aussi fédérer des approches Cloud / HPC

L'union fait la force: rejoignez-nous !

ILaaS
Infrastructure d'IA mutualisée

Questions