



ALLONIA

FAST TRACK TO AI

FONCTIONNALITÉS & ARCHITECTURES

ALLONIA AUTOMATISE LA CREATION D'AGENTS IA POUR LES ENTREPRISES

DEUX PRODUITS



La plateforme de création d'agents, connectée, avec des modules d'action pour créer automatiquement des workflows qui transforment les processus



La plateforme qui permet d'entraîner des modèles personnalisés avec les données de l'entreprise pour garantir la meilleure qualité et l'adaptation aux différents métiers.

3 TECHNOLOGIES PROPRES

MODELES INDUSTRIELS

Garantir la qualité des réponses en entraînant des modèles sur des données sectorielles.

Allonia simplifie la capacité des entreprises en fournissant les outils nécessaires: data prep, entraînement, alignement

MODELES D'INTENTION ET D'ACTION

Pour comprendre les tâches complexes, Allonia a développé un modèle d'intention qui découpe les questions en actions à réaliser (*synthétiser, lancer une recherche, remplir une base de données, envoyer des emails ...*)

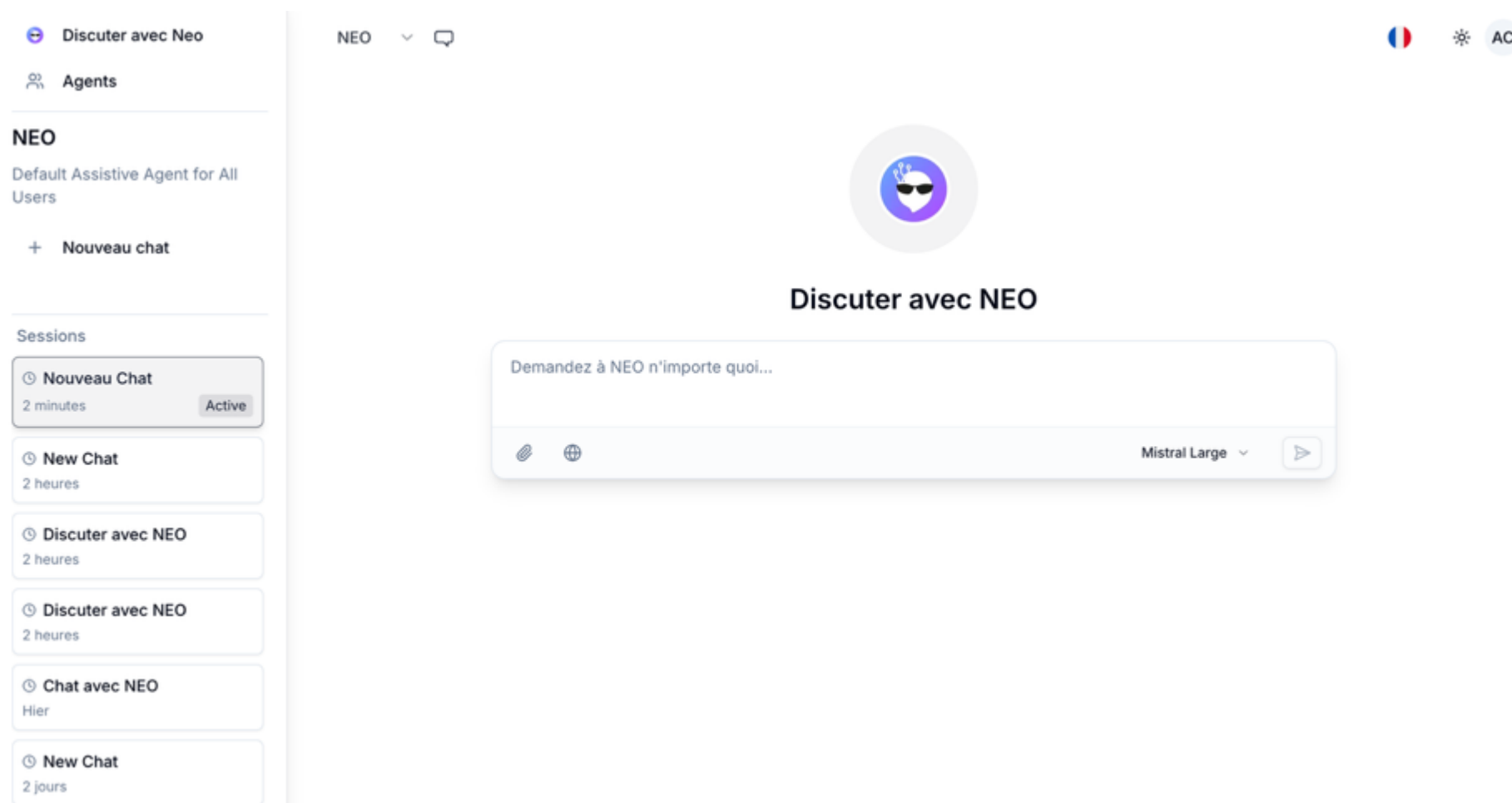
AUTOMATISATION

Pour les processus complexes, Allonia permet d'automatiser la création des workflow unitaires, et de les exécuter.



NÉO

**UNE PLATEFORME DE CRÉATION
D'AGENTS POUR LES MÉTIERS**

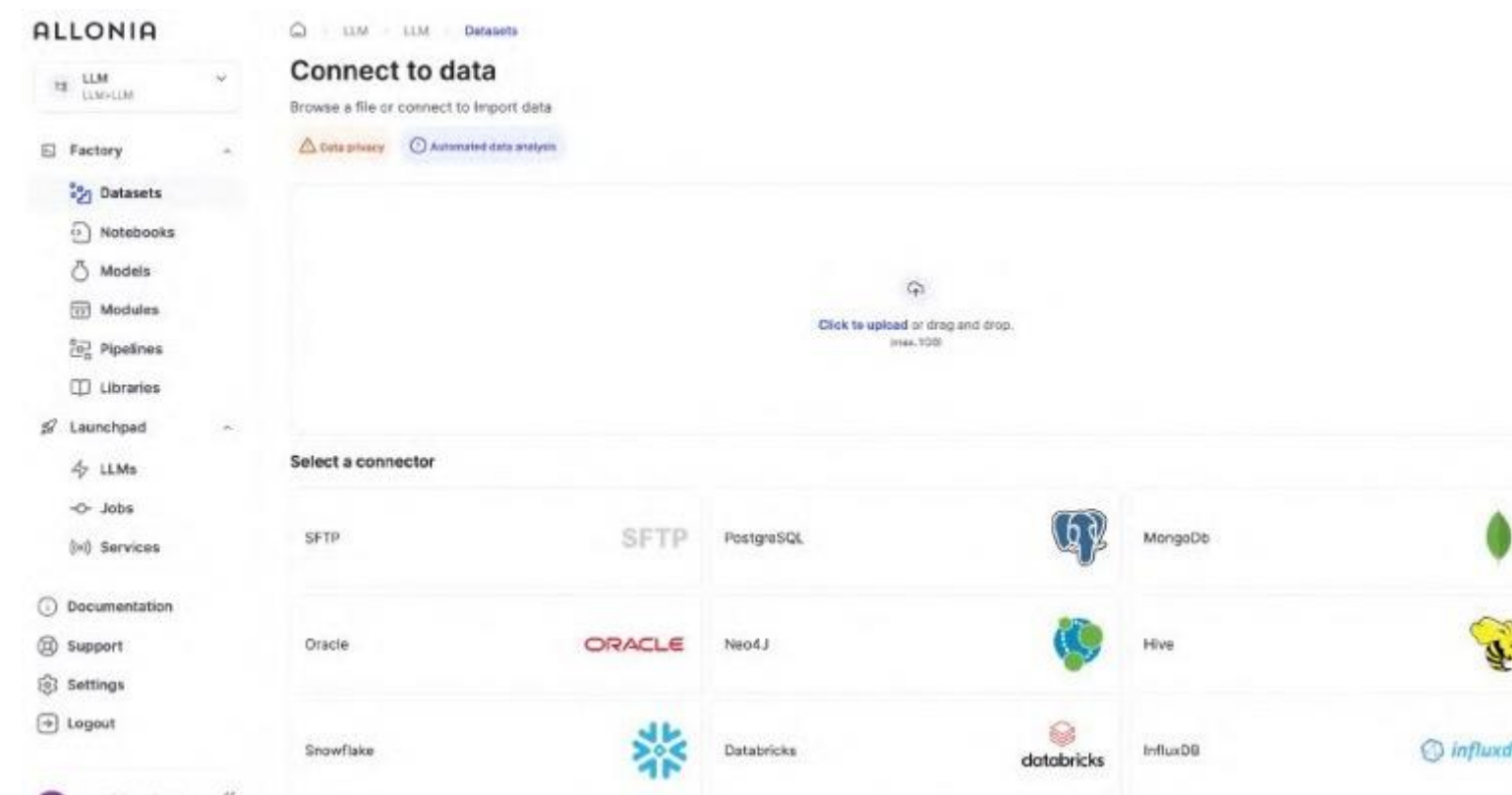


**LE CHAT QUI PERMET DE GÉRER L'ACCES AUX
DIFFÉRENTS MODÈLES, D'INTÉGRER SES DOCUMENTS,
DE FAIRE DES RECHERCHES, ...**



NEXUS

**UNE PLATEFORME DE DÉVELOPPEMENT
ET DE DÉPLOIEMENT**



**TOUS LES OUTILS D'ENTRAÎNEMENT, LA
CAPACITÉ DE MIXER AI & GENAI,
LA PARTIE OPS POUR INDUSTRIALISER**



NÉO CHAT & AGENTS

Fonctionnalités de chat

- Texte, interprétation images, transcript sons
- Interrogations bases de données
- Génération de code
- Multimodèles

Fonctionnalités d'agents

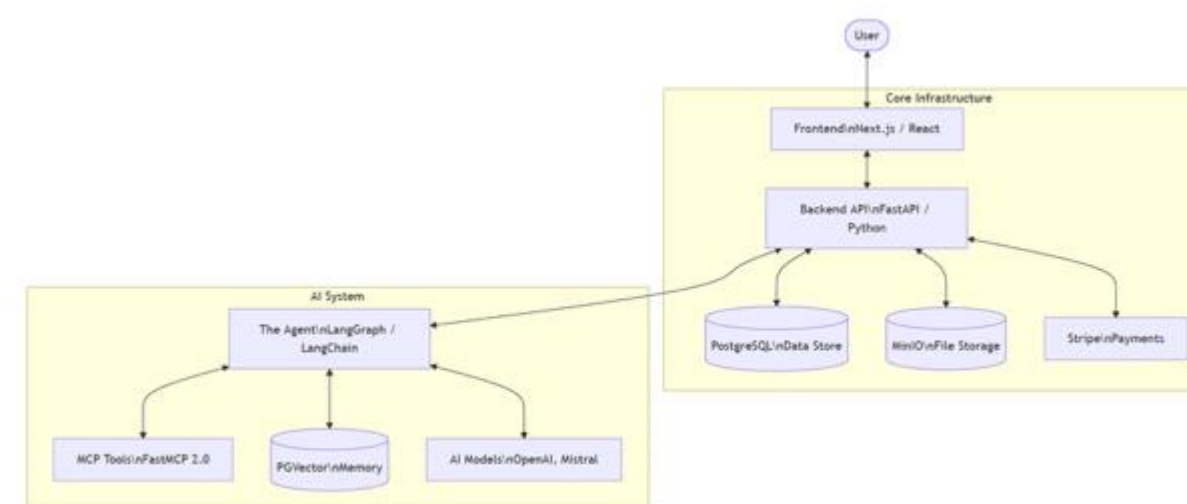
- Intégration de préprompt
- Intégration de corpus
- Connexions bases de données
- Connexions server MCP
- Modèles d'intentions
- Bibliothèques de tools (*search, emails, ppt...*)
- Avancés :
 - Embedder, tokenizer dédiés
 - API de traitement dédiés

Administration

- RBAC
- Collaboration
- Tableau de bord d'usages (agents, modèles...)
- Paiement Stripe intégré

Technologies

- Frontend : Nextjs, React
- Backend : Python
- Agent : Langchain, Langgraph





NEXUS

PLATEFORME ENTRAÎNEMENT & DÉPLOIEMENT

Gestion des projets

- Création de projets / sous projets (Tasks)
- Cycle de développement par projet
(*entraînement / déploiement*)

Fonctionnalités d'entraînement

- Connexions bases de données
- Notebook en ligne (VSCode peut être ajouté)
 - Librairies préinstallées
 - Template AI, Optim, GenAI

- Gestion des modèles créés
- Création de modules

Fonctionnalités de déploiement

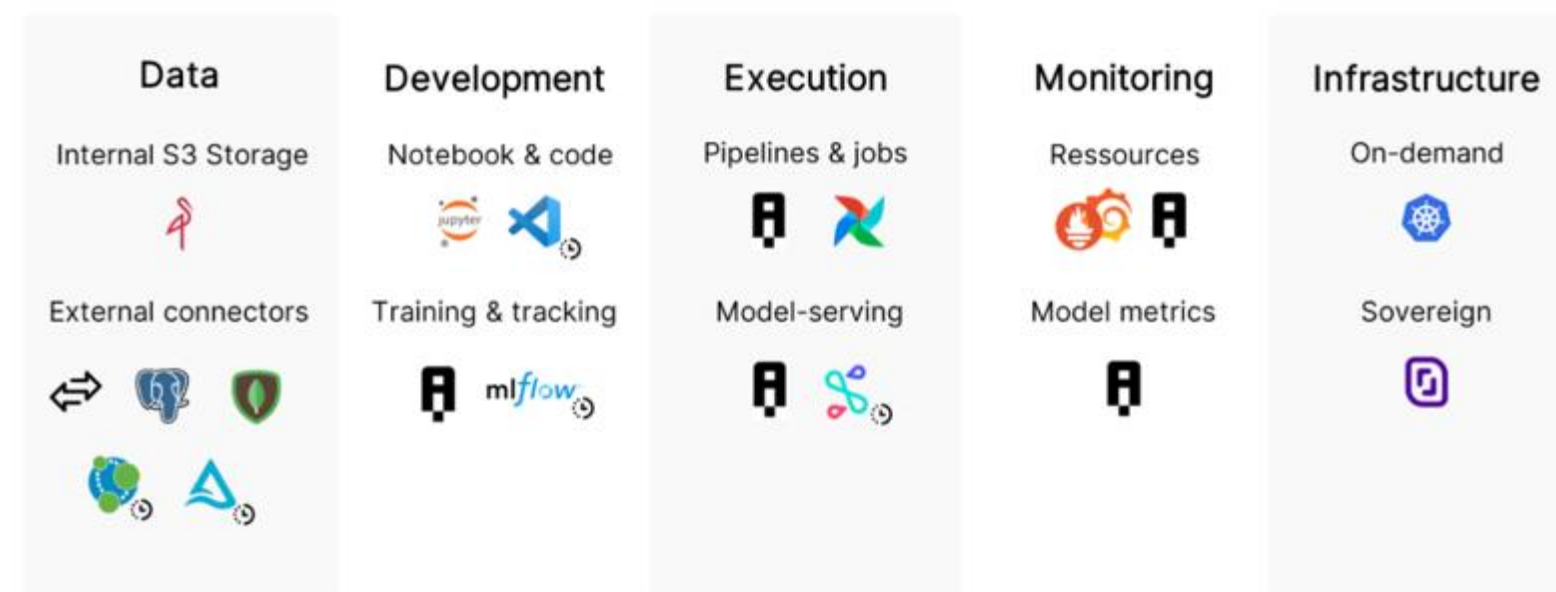
- Création de pipelines
- Génération de user services
- Lancement de jobs

Administration

- RBAC par Projets / sous projets
- Collaboration

Technologies

- Frontend : Nextjs, React
- Backend : Python
- Infra : Kubernetes



UNE STACK IA & IA GEN COMPLETE

CHATS & AGENTS



Agents prédéfinis

Agents personnalisés

Création de documents

Recherche web & enrichissement

OUTILS d'entraînement



Data standardisation

Embeddings

function calling

finetuning

evaluation & alignment

bibliothèques de modèles



AI & GenAI
opensource libs

opensource hosted
models

image models
voice models

scoring models
optim models

custom models &
RAGs

Gestion de l'infrastructure



Conteneurs

Scaling

Inference & API

Sécurité

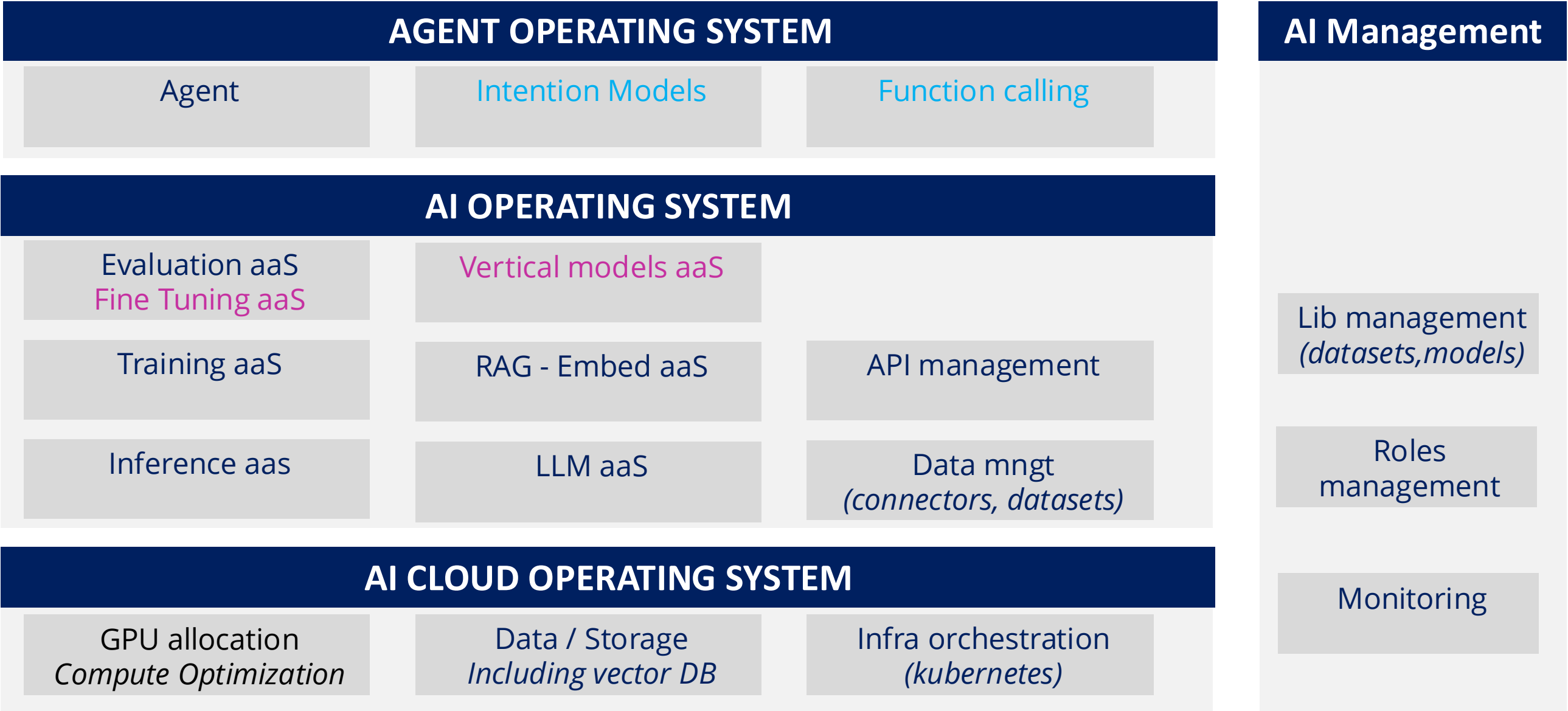
AVEC UNE STACK AGENTIQUE ASSOCIÉE

Une stack techno déjà entièrement disponible au travers des produits Neo et Nexus.

Une roadmap importante sur les sujets de Function calling et de modèle d'intention pour augmenter de plus en plus les possibilités d'automatisation des processus d'entreprises.

NÉO

NEXUS



DES TECHNOLOGIES D'EXCELLENCE DEVELOPER EN COLLABORATION AVEC LES MEILLEURES LABOS EN IA

DES LABOS EN IA POUR LE SOCLE



Agentification et modèles personnalisés



Optimisation des coûts d'inférence



Scoring et clustering



DES LABOS SPECIALISES POUR LES SUJETS INDUSTRIELS



Maintenance prédictive notamment dans le nucléaire



Simulation et génération avec de l' Adversarial resilience learning



Optimisation logistique et hybridation avec de l'IA générative

DES COLLABORATIONS RÉALISÉES GRÂCE AUX PROGRAMMES D'INNOVATION DE FRANCE 2030 AVEC LES PROJETS FERMAT, SCRIBE, RENOVAITE, AUTEND, CYLVIA

UNE ENTREPRISE AU COEUR DE L'ECOSYSTEME POUR LA SOUVERAINETÉ EUROPÉENNE

ECOSYSTEMES PUBLICS



Le calcul intensif au service de la connaissance

EuroHPC
Federation Platform

GRANDS PROGRAMMES EUROPEENS

AI Factory AI2F
Avec GENCI



Lauréat de Simpl avec Sopra
(contrat total de 150M€)



FEDERATIONS





ALLONIA
FAST TRACK TO AI

PARTENAIRES





UN MODELE D’AFFAIRE EPROUVE ET PRET A SCALER

UN BUSINESS MODEL DE LICENCES & D’ACCOMPAGNEMENT

- Les produits NEO et NEXUS sont vendus sous forme de licence produit
- Avec des clients publics et privés
- Un sweet spot pour des organisations entre 1000 et 10000+ collaborateurs

AVEC UN POTENTIEL INTERNATIONAL

- Des premiers démarrage avec Sopra Belgique et NTTData Espagne
- Un support de nos investisseurs en Allemagne et en Suisse permettant des débuts de discussions (BMW...)



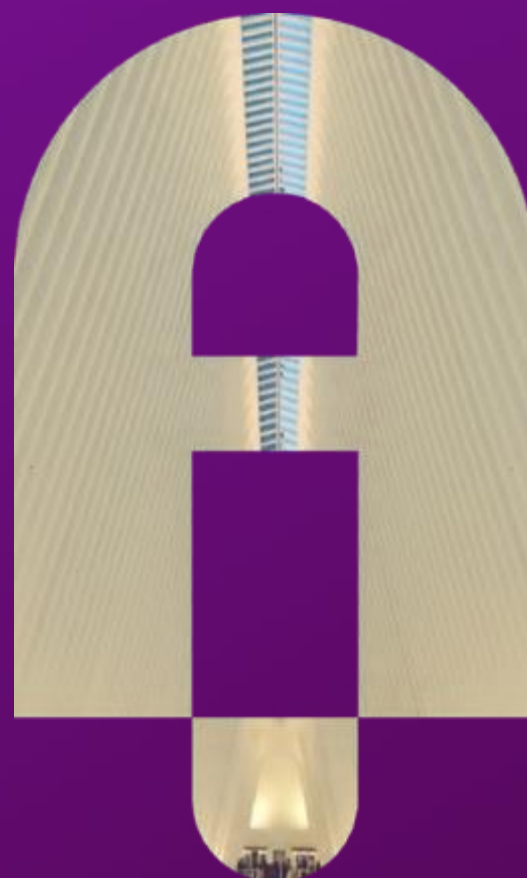
UNE AMBITION FORTE SUR UN MARCHÉ EN PLEINE CONSTRUCTION

UN MARCHÉ IMPORTANT EN CONSTRUCTION

- Un marché des agents B2B estimé à 300Md€ dont 70Md€ en Europe
- Une croissance potentielle de 25-30% par an
- Et surtout, un marché non établi :
 - Fortes évolutions technologiques
 - Forts impacts de transformation
 - Pas d'acteurs établis
 - Trop favorables aux grands acteurs « de confiance »

UNE FORTE AMBITION POUR STRUCTURER LE MARCHÉ IA

- L'ambition de proposer la stack IA de référence en Europe grâce à des projets comme Simpl, les collaborations avec les foundations models et nos travaux avec les cloud providers
- Une promesse claire de transformation des processus par l'IA alliant Produits et Services pour rassurer et offrir de la rentabilité



CHAT GÉNÉRIQUE



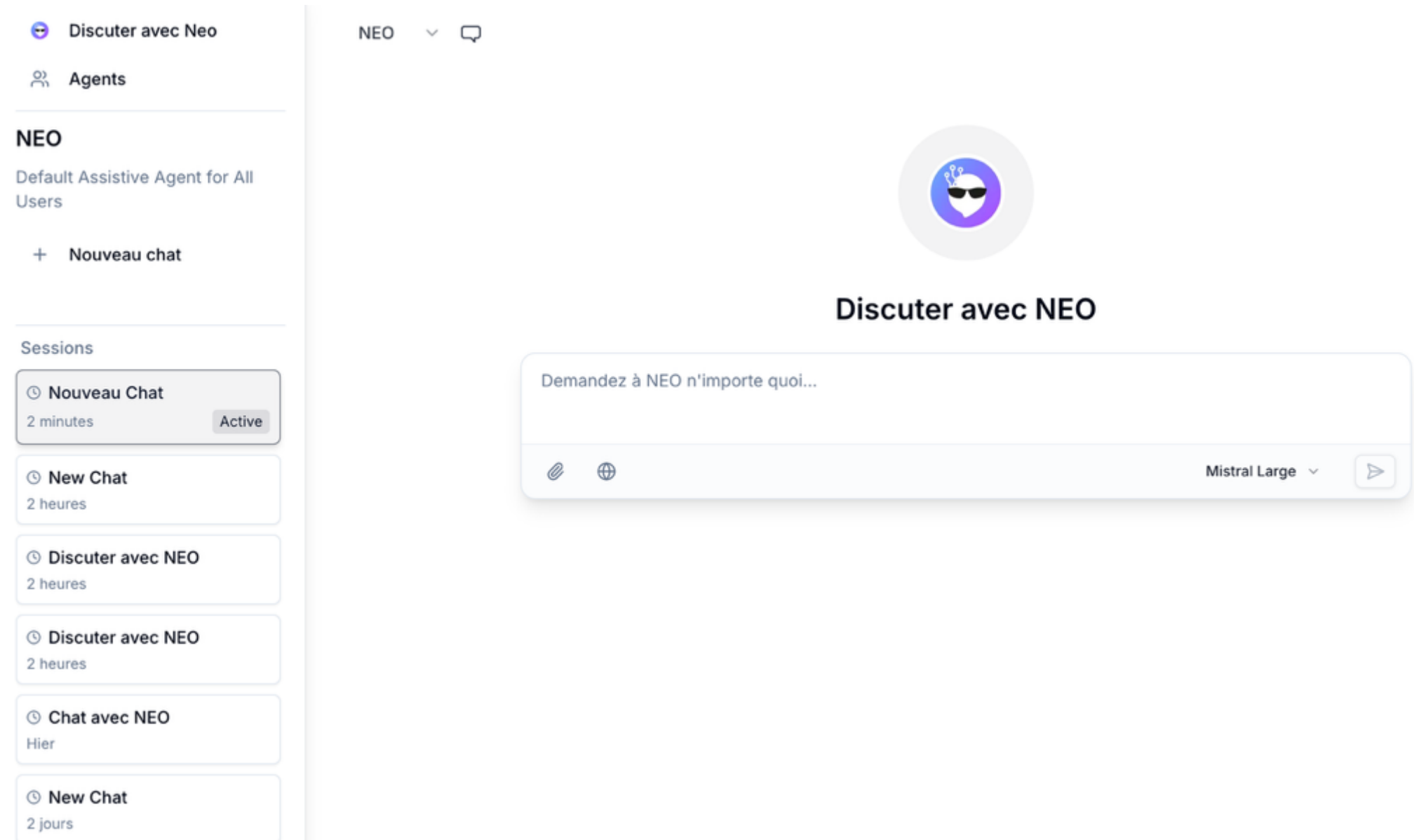
NÉO

L'accès à un chat générique avec les fonctionnalités classiques :

- Synthèse, réécriture, génération ... de texte
- Recherche d'information sur le web
- Upload de documents et interaction avec ces documents

En complément : Une offre multimodale

- Interprétation d'images et de graphiques
- Transcription de fichier audio (y compris Teams)
- Des premières capacités d'action : envoi des réponses par email



EXEMPLE SIMPLE DE FONCTIONNEMENT

DEMANDE DE CONSTRUCTION D'UN WORKFLOW DANS NEO

GENERATION AUTOMATIQUE DES CODES AVEC NEXUS



Chat Corpus - Agent Chat Tools Dev

Accueil / Chat Corpus / Chat Tools Dev

Chat_Tools

magistral-small



Nouvelle Session

peux tu m'envoyer les 10 dernieres actualités sur l'intelligence artificielle à antoine.couret@allonia.com ?

L'email a bien été envoyé avec succès, souhaitez-vous exécuter d'autres actions ?

Generated by Mixtral-8x7B-Instruct-v0.1



1 Source



Workflow

Page : 1

Texte :

```
[
  {
    "name": "chat",
    "parameters": {
      "query": "Peux-tu me donner"
    }
  },
  {
    "name": "send_email",
    "parameters": {
      "recipient": "antoine.couret@allonia.com",
      "recipient_name": "Antoine",
      "subject": "10 dernières actualités sur l'intelligence artificielle",
      "body": "Voici les 10 dernières actualités sur l'intelligence artificielle à antoine.couret@allonia.com ?"
    }
  }
]
```

L'agent va automatiquement créer les bonnes actions, construire le workflow, se connecter au serveur d'email et envoyer l'email.

Ce workflow peut également intégrer d'autres actions : création de documents, de graphiques, trigger pour des actions répétitives ...

UNE ARCHITECTURE AGENTIQUE NATIVE

Une architecture avec les fonctionnalités disponibles pour l'automatisation de processus

Qualité des réponses

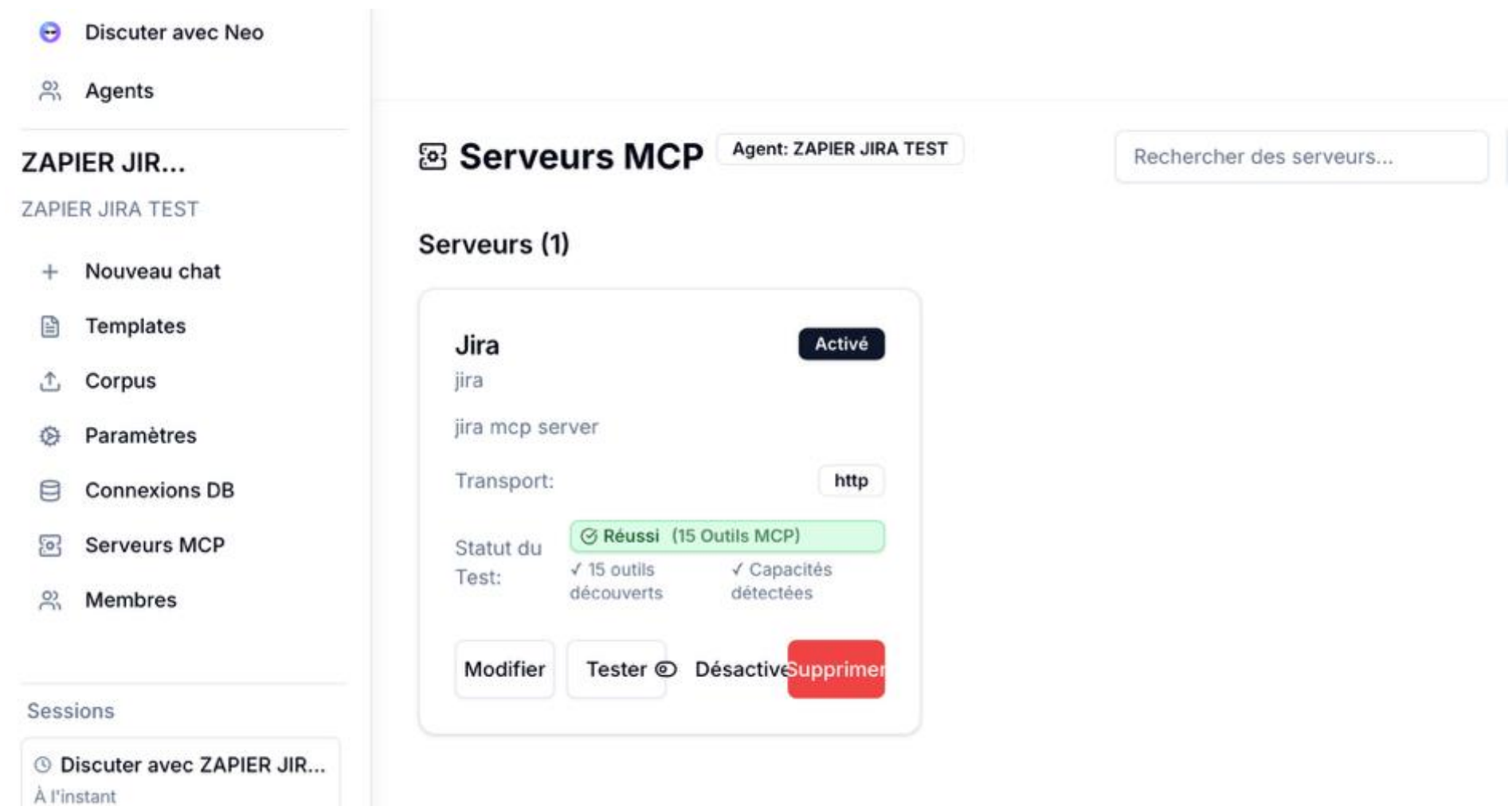
- Intégration de prompts génériques
- Outils d'évaluation et de benchmarking
- Gestion textes, SQL, images, sons, code

Connecteurs

- Connexion aux bases de données
- Connecteurs MCP
- Création de tools

Administration

- Déploiement services par agent, direction, entreprise
- RBAC
- Monitoring des usages



UNE OFFRE D’AGENTS MODULAIRE SUR ETAGERE ET CONNECTEE A L’AI STUDIO

L'ensemble de l'architecture est modulaire et offre des possibilités de personnalisation en fonction des différents agents créés :

Agents

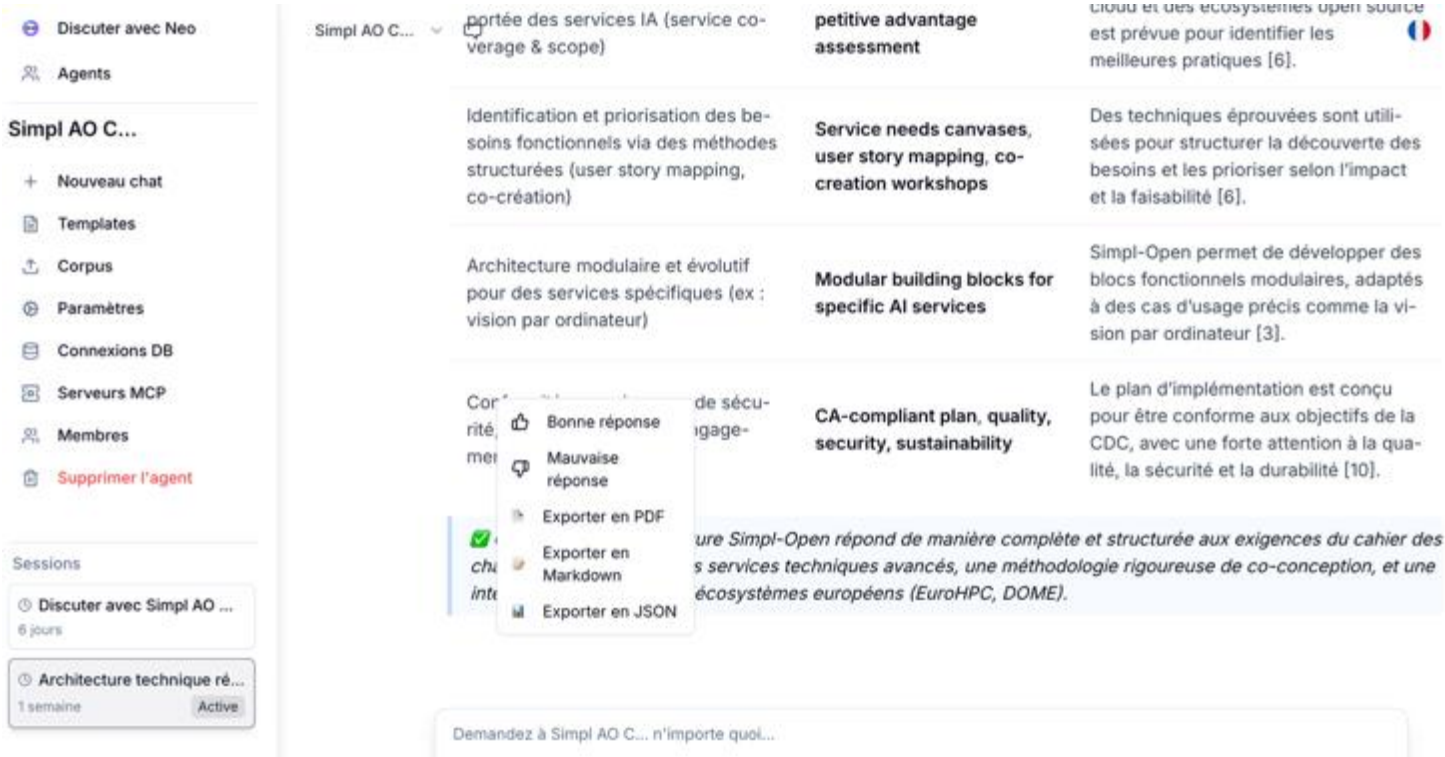
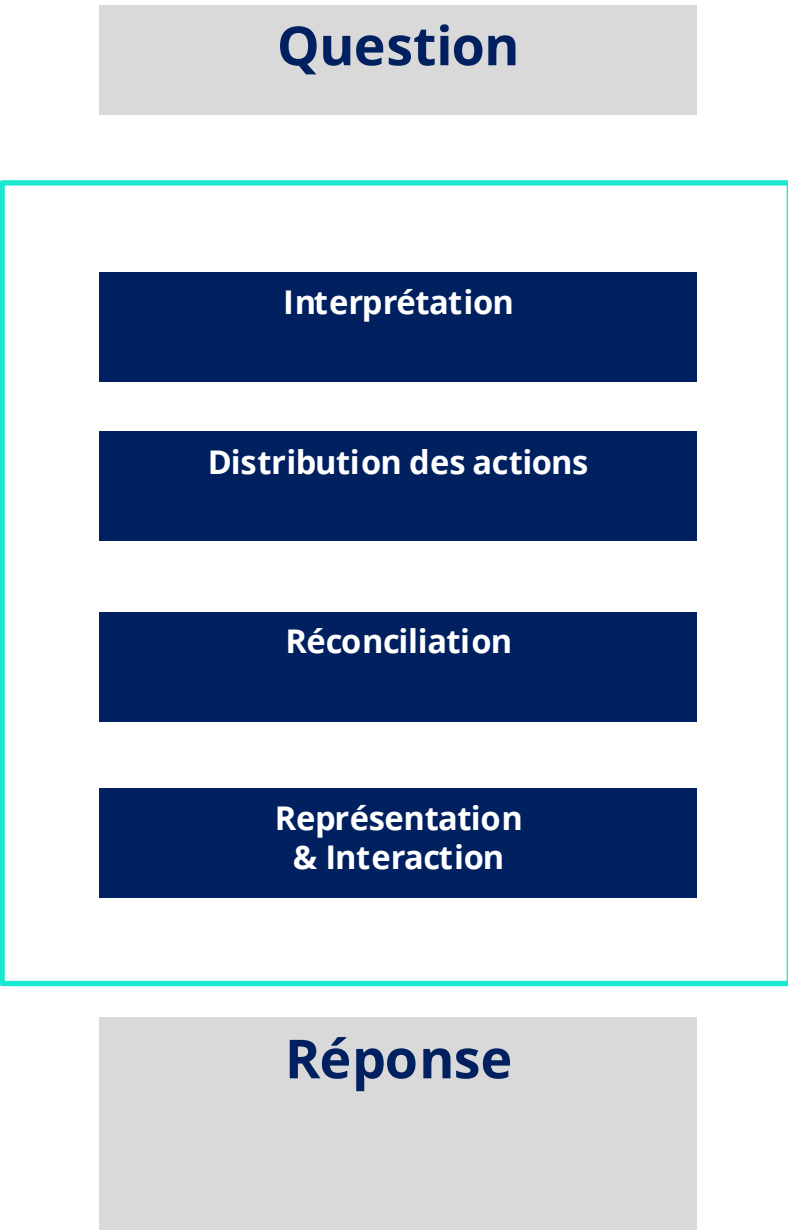
- Embedder personnalisable
- Tokenizer
- Tools avec serveur MCP

Pipeline

- Notebook AI & GenAI
- Pipeline complexe
- Catalogues modèles, API,...

Composants

- Modèles de base hébergés
- Base de données vectorielle



UN AI STUDIO QUI PERMET LE DÉVELOPPEMENT ET LE DÉPLOIEMENT D'AGENTS

Une architecture basée sur kubernetes qui garantit une industrialisation facile

Entraînement

- Notebook avec toutes les librairies
- Création d'embedder
- Accès aux modèles de fondation

Pipeline to agents

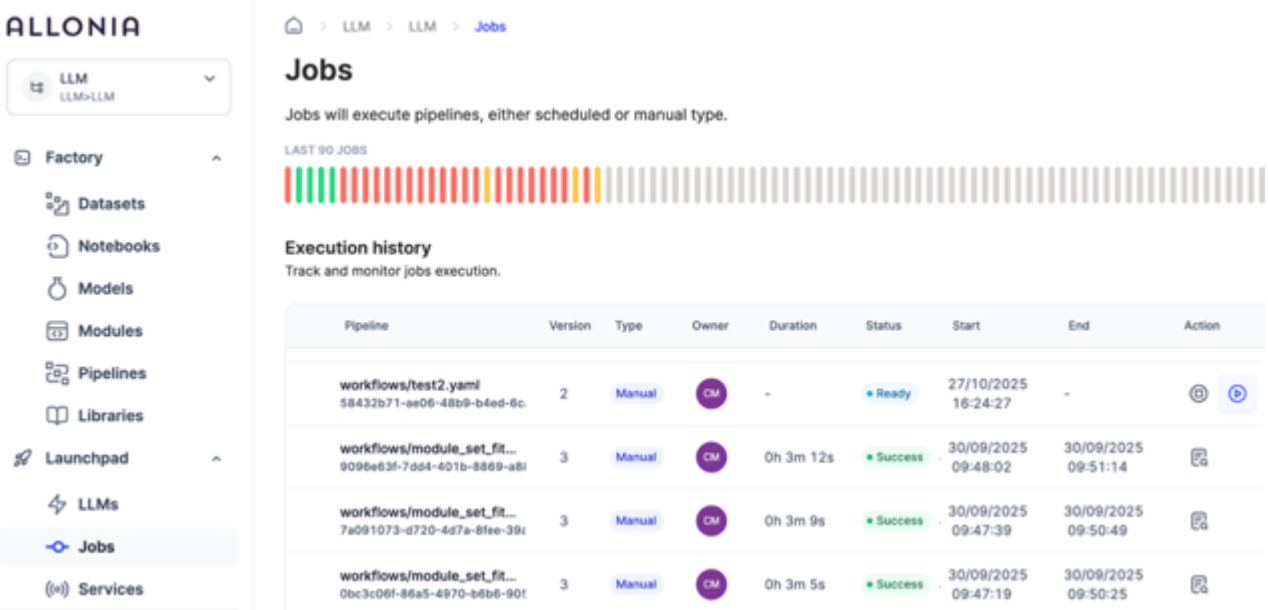
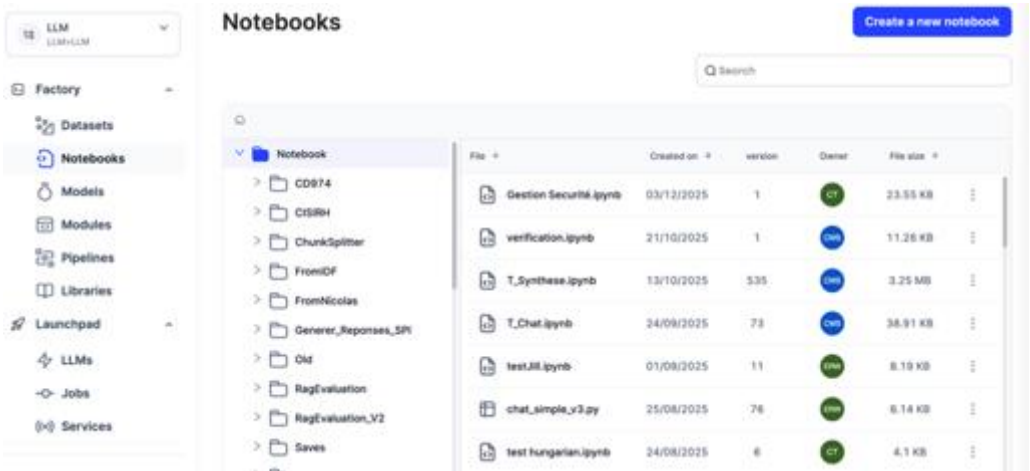
- Création de pipeline d'agents
- Génération de user services

Monitoring

- Démarrage / arrêt de services
- Suivi de la santé des jobs
- Usage des agents / services

Gestion de l'infrastructure

- Pré & auto provisioning
- Accès EuroHPC - Genci
- Balancing
- Alerting



UNE PLATEFORME PENSÉE POUR L'INDUSTRIALISATION

Une architecture basée sur kubernetes qui garantit une industrialisation facile

Pipeline to Job

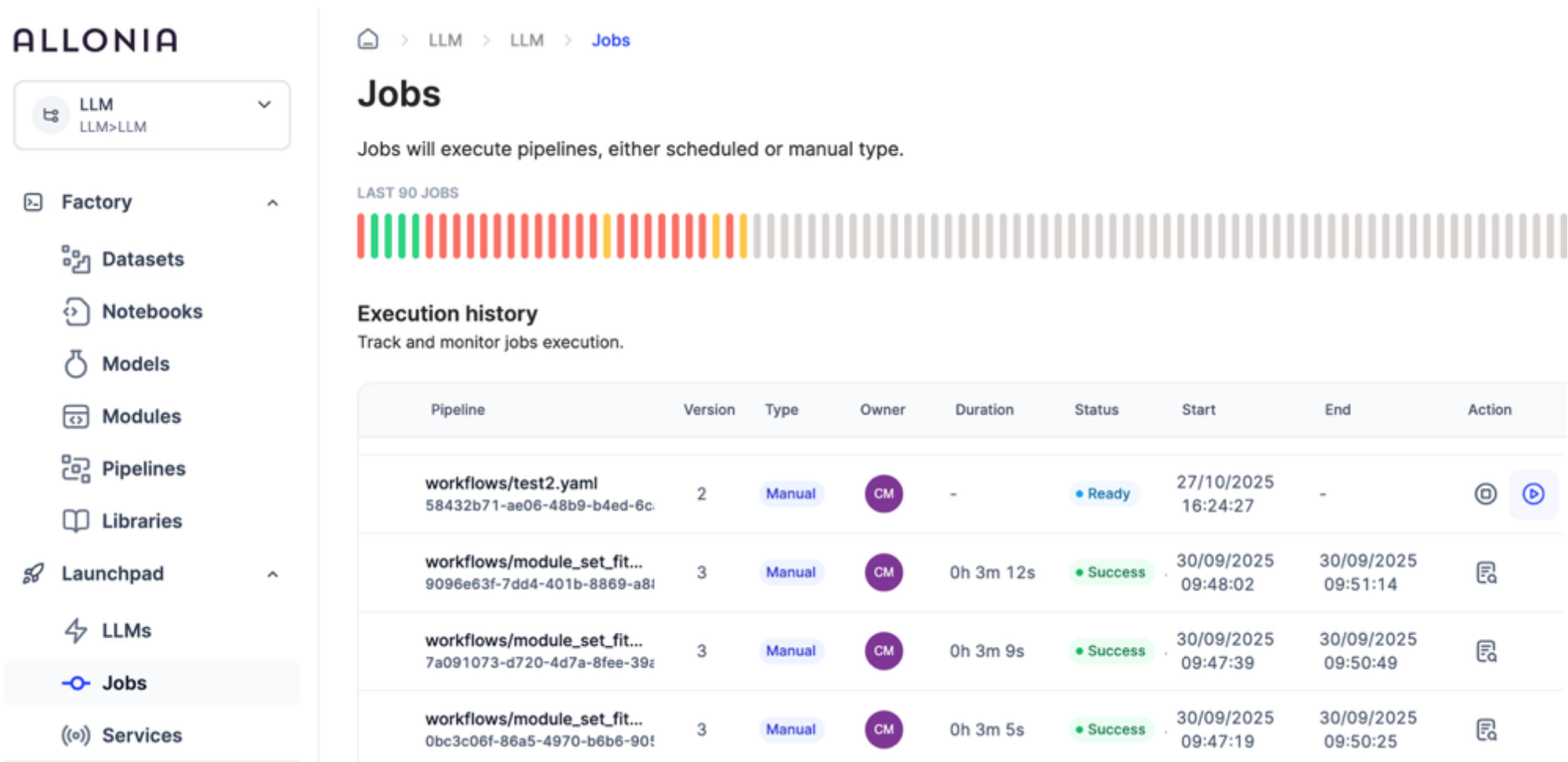
- Génération de user services
- Création de jobs

Monitoring

- Démarrage / arrêt de services
- Suivi de la santé des jobs
- Usage des agents / services

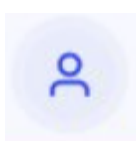
Gestion de l'infrastructure

- Pré & auto provisioning
- Balancing
- Alerting



Offre modèles génériques open source en SecNumCloud

EXEMPLE
DE PRIX

 Embedder (Camembert...)	 LLMaas Jusqu'à 30B	 LLMaaS Jusqu'à 70B	 LLMaaS Jusqu'à 130B
2 850€ / mois	3 580€ / mois	7 000€ / mois	14 000€ / mois
Perfect for Expert agent LLM project	Perfect for experimentation and expert agent LLM	Perfect for large use cases LLM project	Perfect for large use cases LLM project
✓ 1 VM Linux – v5c8r16p1 + 1 GPU P100 ✓ Optimized infrastructure deployment ✓ Cost effective TCO ✓ Standard support	✓ 1 VM Linux - v7c32r64p1 + 1 GPU L40 ✓ Optimized infrastructure deployment ✓ Cost effective TCO ✓ Standard support	✓ 1 VM Linux - v7.c32r64p1 + 2 GPU H100-80 ✓ Optimized infrastructure deployment ✓ Cost effective TCO ✓ Dedicated support	✓ 2 VM Linux - v7.c32r64p1 + 4 GPU H100-80 ✓ Optimized infrastructure deployment ✓ Cost effective TCO ✓ Extended support

**Les modèles peuvent être des modèles textes, images, voix, code*

Offre AI Studio open source en SecNumCloud

EXEMPLE
DE PRIX

 Starter	 Pro	 Enterprise	 Enterprise
5 000€ / mois	10 000€ / mois	25 000€ / mois	60 000€ / mois
Perfect for medium teams : collaboration, inference, mo nitoring	Perfect for large teams : collaboration, inference, mo nitoring	Perfect for large teams : collaboration, inference, mo nitoring	Perfect for large teams : collaboration, inference, mo nitoring
✓ Usage illimité ✓ 1 modèle opensource 8B 1 image 12B 1 voix ✓ 1 VM Linux - v7c32r64p1 + 1 GPU L40 ✓ 500 GB Performance (gp2) ✓ Api EndPoint to LLM ✓ Online support	✓ Usage illimité ✓ 1 modèle opensource 30B 1 image 12B 1 voix ✓ 1 VM Linux - v7c32r64p1 + 2 GPU L40 ✓ 1 TB Performance (gp2) ✓ Api EndPoint to LLM ✓ Extended support	✓ Usage illimité ✓ 1 modèle opensource 70B 1 image 12B 1 voix ✓ 1 VM Linux - v7c32r64p1 + 4 GPU L40 ✓ 1 TB Performance (gp2) ✓ Api EndPoint to LLM ✓ Extended support	✓ Usage illimité ✓ 1 modèle opensource 130B 1 image 12B 1 voix ✓ 1 VM Linux - v7c32r64p1 + 8 GPU L40 ✓ 1 TB Performance (gp2) ✓ Api EndPoint to LLM ✓ Extended support

Offres de services

EXEMPLE
DE PRIX



Support, Maintenance, Exploitation de bundles HGX + Modèles

5 -20 000 € / déploiement

- ✓ Mise en place infra
- ✓ Déploiement plateforme
- ✓ Suivi complet (étude, conception déploiement dans archi & infra client)

5 000 € / mois +

- ✓ Monitoring plateforme déployée
- ✓ Maintenance sur infra dédiée
- ✓ Support administrateurs 2j par mois
- ✓ Support selection de modèles



Création d'agents



BUILD

- 1/2j de formation
- Création des comptes
- Prise en main
- Premiers tests sur des cas reels
- Recommandation de prompts

2 500€
(jusqu'à 10 personnes)



ONBOARDING

- Suivi sur les deux premières semaines
- Prise en compte des retours
- Amélioration des prompts

1 500€
(jusqu'à 10 personnes)

- Support mensuel (1/2h par pers/sem)

3 000€/mois
(jusqu'à 10 personnes)



CREATION D'AGENTS

- Agents simples (multimodal, nb documents)
- Agents complexes (multisources, industriels)
- Automatisation de processus (hors refonte processus)

À partir de ...

... 20 000€

... 40 000€

... 60 000€

ALLONIA INVESTIT DE FAÇON IMPORTANTE EN R&D



*Programme de recherche soutenu par France 2030
dans le programme
« Biens communs pour l'IA générative »*

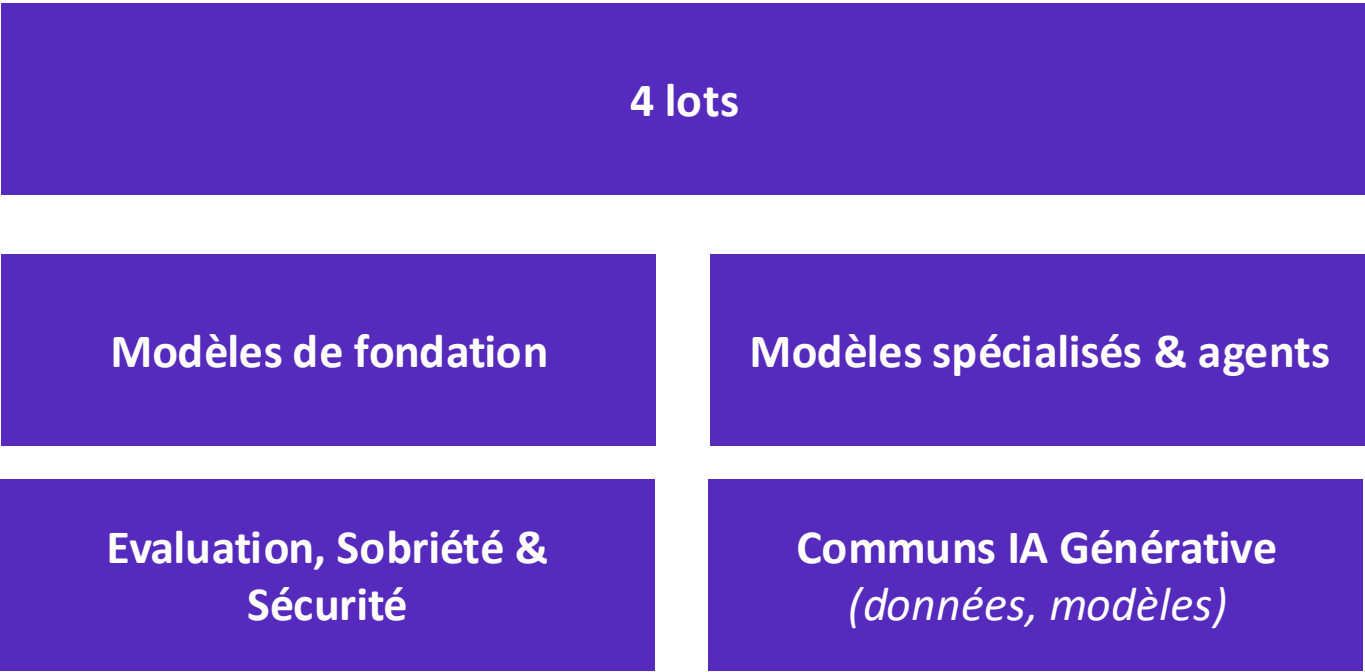
OBJECTIFS

- Développer des modèles souverains avec plus de français et des modèles spécialisés (*entreprises, process*)
- Garantir sécurité, évaluation et sobriété
- Créer une communauté IA sectorielle française

Inria



Lighton
ALLONIA



POUR AUGMENTER LA QUALITÉ DES AGENTS

AMÉLIORER LA CHAÎNE MONOLITHIQUE ACTUELLE

Question

Interprétation

Distribution des actions

Réconciliation

Représentation
& Interaction

Réponse

PROGRAMME DE R&D

- **Modèles spécialisés** : entraînement de modèles d'intention pour mieux refléter les processus métiers – Inria Allonia
- **Construction de modèles d'intention** : améliorer la qualité de compréhension des questions des utilisateurs et augmentation des actions (tools) possibles - Allonia
- **Développements d'outils** d'évaluation, de safety et d'efficacité – Allonia IDRIS
- **Développement d'un framework agentique open source orienté par l'évaluation** : volonté de codéveloppement avec des partenaires « industriels »

UNE STACK AI CIBLE AVEC EUROSTACK



AI	Training •Connectors •ML training (NB, ...) •Finetuning, Embedding... •Feature Store	Inference •Model exposition •Inference server •Model Serving	Agentic •LLM aaS (text, voice, image, code) + deploy yours •Embedders, OCR •Agent, MCP	AI Mngt •Model Monitoring •API generation •API mngt •CI CD	Observability •AI Act compliance	Economics •per token •per CPU / GPU •Per tenant •...
	Services •Connectors •Ingestion •Streaming •Batch •Integrity check	Data Ops •Data Prep •Data Pipeline •Data Lab •Orchestration •Synthetic data	Analytics •Vizualisation : connection to BI tools •Annotation	Data Mngt •Data Catalogue •Query •Mesh	Security •RBAC •Integrity	Observability •GDPR Compliance •Connectors
LAKE HOUSE	Orchestration & Cluster services . Kub . Containers . Airgap			Management Services . Multi Cloud . Hybrid . Edge . Mutli tenant		
	Compute •CPU •GPU	Storage / DB •S3 •Postgre, Postgre vector •Mongo •Graph •Vector ...	Network •Routing •Encoding •CDN •vrack	Access Mngt •IAM •API Gateway / Ingress / Egress. •Federation. •Access Mgt	Security •IAM. •API Gateway / Ingress / Egress. •Federation. •Access Mgt	Observability •Monitoring •Audit •Login •Alerting
CLOUD						

AI COVERAGE

AI infrastructure

AI Clusters



Custom-build Clusters
(CPU or GPU)



On-Demand Clusters



On-Demand Large Clusters
for training (up to 5,000 GPUs)

AI GPU aaS



NVIDIA H100
(SXM & PCIe)



AMD Instinct
MI300 Series



NVIDIA L40S



NVIDIA L4



NVIDIA Tesla P100

AI Services

AI Training



Data connectors & Catalog



Training auto-provisioning
(CPU, GPU)



Notebook including libraries
(scikit, pytorch, tensorflow ...)



Gen AI tools (*data standardisation, Finetuning, Embedder creation, tokenizer creation, Evaluation, Alignment*)

AI Models



Scoring, default detection, optimisation ,...



Geolocation, optimisation



LLMs : text, image, voice, code
OCR models



Deploy your own foundation
models

AI Ops



Inference auto-provisioning
(CPU, GPU)



AI workload



Pipeline & orchestration



User services, APIs



Monitoring

AI Agents



Tools : RAG, search, text to sql



MCP servers & connectors



Document generation with
templates



Document generation with
templates