Retour du GT IA

- Morgan Bohn - Université de Strasbourg
- Nicolas Truchaud - Université de Polynésie Française

Pourquoi des ateliers sur l'IA ?

- **Contexte** : Explosion de l'utilisation des IA génératives propriétaires
- **Enjeu** : Besoin de solutions d'IA génératives souveraines dans l'ESR
- **Solution** : Fédérer les initiatives isolées et partager les compétences
- **Méthode** : Favoriser l'open source et l'auto-hébergement

Objectifs des ateliers

- Créer une communauté inter-établissements
- Favoriser l'acquisition et l'approfondissement des compétences techniques
- Partager des outils et librairies communes
- Transmettre ses expériences avec l'IA
- Conjuguer sobriété et souveraineté numérique
- Garantir la sécurité et la protection des données

Canaux de communication

- Liste de diffusion : myia@groupes.renater.fr
- Canal rocket : https://rocket.esup-portail.org/channel/GT-IA
- Chaîne vidéo : https://videos.esup-portail.org/groupes-de-travail/gt-ia/
- Wiki esup : https://www.esup-portail.org/wiki/spaces/ESPADHERENT/pages/1456144390/Atelier+IA

Quelques chiffres

- 1 atelier par mois depuis Novembre 2025
- 1h à 2h par atelier
- 6 ateliers réalisés
- 2 ateliers encore prévus
- ~ 50 participants réguliers
- 170+ participants au pic

Collaboration avec Ilaas

- Site : https://www.ilaas.fr/
- Infrastructure d'IA mutualisée dans l'ESR
- Mettre en commun les ressources des établissements
- Anime certains ateliers

Atelier 1 : Initiation à l'IA locale

Installer et configurer Ollama et OpenWebUI pour créer son premier assistant

- Concepts fondamentaux des LLM
  - Réseaux de neurones, tokenisation et Transformers
- L'écosystème Ollama
  - Moteur d'inférence et usage local avec Python
- L'interface OpenWebUI
  - Installation et prise en main de l'outil

Atelier 2 : Inférence et passerelle LiteLLM

Déployer des modèles en prod et les orchestrer avec LiteLLM

- Panorama des moteurs d'inférence
  - vLLM, SGLang, Infinity, TGE et Speaches
- Focus vLLM : Déploiement, KV Cache et Multi-GPU
- Focus Infinity & Speaches : Embedding, Reranking, STT et TTS
- Orchestration avec LiteLLM
  - API OpenAI, Load Balancing et Fallback
- Intégration OpenWebUI
  - Connexion et tests des fonctionnalités

Atelier 3 : Les bases du NLP

Découvrir les briques essentielles du traitement du langage naturel

- Concepts clés : prétraitements, représentations et architectures
- Liens avec l'IA générative contemporaine
- Pratique : classification et analyse de sentiments

Atelier 4 : Déploiement sur Kubernetes

Industrialiser le déploiement des LLMs en production

- Infrastructure avec OpenShift
- Déploiement GitOps avec ArgoCD
- Live démo : interfaces Kubernetes de l'Unistra

Atelier 5 : OpenWebUI en production

Optimiser et scaler l'interface utilisateur

- Configuration de la production
- Configuration du RAG par défaut
- Déploiement à large échelle

Atelier 6 : Le RAG dans OpenWebUI

Maîtriser la génération à enrichissement contextuel

- Introduction aux concepts du RAG
- Mise en pratique : base documentaire DSI
- RAG avancé : Pipelines et MCP (Model Context Protocol)

A venir

Atelier 7 : Fine-Tuning

- 📅 7 mai 2026 – 🕘 9h à 11h
- Lien : https://bbb.unistra.fr/rooms/boh-fle-qdw-5ns/join

Atelier 8 : Développer un agent IA

- 📅 4 juin 2026 – 🕘 9h à 11h
- Lien : https://bbb.unistra.fr/rooms/boh-fle-qdw-5ns/join

Bilan

- Participants globalement satisfaits
- Beaucoup de questions et d'échanges
- Collaboration naissante (ilaas)
- Bonnes pratiques communes
- Public fidèle

La suite ?

- Pause estivale et reprise des ateliers à la rentrée
- Recherche de futurs intervenants
- Idées d'ateliers :
  - Intégration des services IaaS et d'Albert API
  - Programmation avec OpenCode
  - Développement d'un serveur MCP
  - Autres propositions ?

Merci, questions ?