Christian Segnou

Senior AI Engineer / MLOps & LLMOps – RAG & Optimisation

📍 **Île-de-France, France**
📱 **Téléphone :** +33 6 66 48 47 74
📧 **Email:** chris.segnou@gmail.com
🔗 **LinkedIn:** linkedin.com/in/christian-segnou
💻 **CV Mode RAG:** https://christian-segnou-cv.streamlit.app/

Résumé

Expert en Intelligence Artificielle et en Science des Données, avec plus de 5 ans d'expérience dans la conception, l'optimisation et l'industrialisation de solutions IA / GenAI en production.

Spécialisé en LLMs, RAG, NLP et LLMOps (On-Premise & Cloud), j'interviens sur l'ensemble du cycle de vie des modèles en garantissant performance, scalabilité, sécurité et conformité.

Je maîtrise les environnements d'inférence avancés (vLLM, LiteLLM, quantification INT4/INT8, caching, optimisation GPU), la mise en production via Docker, Kubernetes, MLflow, et l'automatisation CI/CD avec Jenkins.

J'assure également une observabilité complète des systèmes à grande échelle grâce à Prometheus & Grafana (monitoring, métriques, latence, throughput, alerting).

Mon approche combine une expertise technique approfondie et une orientation impact métier, acquise dans des secteurs exigeants tels que l'aéronautique, la logistique et l'automobile.

Compétences Clés

Langages : Python, SQL, Linux

Bases de données & Vector Stores : PostgreSQL, MongoDB, SQL Server, PineCone, ChromaDB

Visualisation de données : Pandas, Matplotlib, Seaborn, Plotly

Frameworks IA : PyTorch, TensorFlow, Scikit-learn, HuggingFace, LangChain, LangGraph

MLOps / DevOps / CICD : Docker, Kubernetes, ArgoCD, Jenkins, GitHub Actions, Gitlab CI, Hydra, MLflow, DVC

API: FastAPI, Flask

Cloud Computing : AWS, Azure, Google Cloud (Vertex AI)

LLMops / Perf : vLLM, LiteLLM, COMET, Optimisation (quantification, cache, speculative decoding)

IA / GenAI / NLP : LLMs, RAG, Agents, RLHF, NLP, IA Générative, Vision multimodale

Soft Skills : Leadership technique, communication claire, pédagogie, gestion de projets complexes

---

Expérience Professionnelle

Aquila Data Enabler *(Juin 2025 – Présent)*

(Novembre 2025 - Présent)

Senior LLMOps on-prem : exploitation et optimisation de la plateforme IA générative

client DGFiP/DTNUM : Au sein de la DTNum, la DGFiP a conçu une plateforme d'IA générative souveraine pour répondre à la montée en charge des usages internes. L'objectif de la mission est d'assurer le fonctionnement optimal, la scalabilité et la performance de la plateforme, tout en maîtrisant les coûts GPU et la bande passante.

Réalisations clés :

- Déploiement et supervision d'environnements vLLM / LiteLLM pour la mise en production de modèles de langage et multimodaux.
- Optimisation des performances via quantification, pruning, cache management, speculative decoding, multi-instance et désagrégation.
- Paramétrage des modèles selon les besoins métiers (Text, Multimodal reasoning, reranking, batch processing).
- Mise en œuvre d'un plan de monitoring et d'optimisation continue (logs, métriques, scalabilité).
- Collaboration étroite avec le Tech Lead LLMOps, le Product Manager et l'équipe Data Science pour l'alignement entre performance, disponibilité et coûts.
- Support à la bande passante IA générative pour garantir la continuité de service sans extension matérielle.

Stack technique : vLLM, LiteLLM, Python, FastAPI, Docker, Jenkins, ArgoCD, Kubernetes, GitLab CI/CD, cloud interne souverain, observabilité (Prometheus/Grafana).

(Juin 2025 – Octobre 2025)

Senior AI Engineer – LLM/RAG agentique (multimodal), LLMOps on-prem

Client TECHNIP : Développement d'un RAG agentique multimodal (Agents multi-systèmes) qui comprend des PDF industriels complexes (tables géantes multi-pages, PFD/P&ID riches en symboles) et route les requêtes entre Text-to-SQL et recherche sémantique.

L'objectif était de remplacer la lecture manuelle de PDF industriels XXL (datasheets multi-pages à très grandes tables, PFD/P&ID très denses) qui ralentissait les équipes. Les assistants génériques (tel que copilot) n'arrivaient pas à "converser" avec ces documents. J'ai conçu un RAG agentique multimodal: l'utilisateur pose sa question, des agents spécialisés (routeur, SQL, sémantique, vision, juge) collaborent et retournent une réponse sourcée en quelques secondes. La solution a été conteneurisée et déployée on-prem (GitLab + Docker) pour évaluation client.

Missions :

Architecture agentique: Orchestration LangGraph d'un système multi-agents (communication, états, transitions), collaboration routeur ↔ SQL ↔ sémantique ↔ juge, fallback bidirectionnel.

Multimodalité avancée:

- Gestion de très grandes tables (nombreuses colonnes, en-tête/sous en-tête multiple complexe, états/couleurs), normalisation d'en-têtes, consolidation multi-pages, export CSV/SQLite et requêtage direct.
- Compréhension de très grands schémas PFD/P&ID complexes via vision tuilée (haute résolution, overlap, consolidation), description d'images et index d'images optionnel (CLIP/SigLIP).
- Text-to-SQL: Sélection intelligente de tables, génération SQL guidée LLM, exécution SQLite, réponse naturelle avec citation précise "Source: fichier, table n".
- Recherche sémantique: Fusion sémantique + mots-clés; prompts modulaires pour réduire tokens/coûts et améliorer la pertinence.
- Interface: UI Gradio (sélection de documents, upload, détection de fichier dans la question, préfixes "datasheet:"/"semantic:"), liens sources cliquables.
- Déploiement on-prem & Ops: Image Docker, registry GitLab, pipeline CI (build/push), déploiement sur serveurs internes (Docker/Compose), authentification, variables d'environnement et secrets, persistance (volumes/Chroma), logs/health-checks.
- Qualité: Suite PyTest, scripts de build d'index et démos vision, journaux détaillés et outils de debug.

Stack technique:

LLM/RAG: LangGraph, LangChain, OpenAI (GPT-4o), Ollama (Qwen2.5/vision), Chroma

Documents: PyMuPDF4LLM, pdfplumber, Pillow; sentence-transformers (CLIP/SigLIP) pour index image

Web/Apps: Gradio, FastAPI/Uvicorn (optionnel)

Data/SQL: pandas, SQLite

Build & tooling: Python 3.10+, Poetry, pytest, logging, YAML config

Impact:

Vitesse & fiabilité: réponses sourcées en secondes (vs heures), robustifiées par collaboration multi-agents et fallback.

Efficience: prompts modulaires et routage agentique → réduction tokens/coûts.

Scalabilité documentaire: ingestion/indexation d'un corpus PDF volumineux avec tables géantes et diagrammes industriels complexes.

Expleo France *(Septembre 2022 – mai 2025)*

(Mai 2024 – Mai 2025)

AI Engineer & Computer Vision Engineer

Mission Airbus (Aéronautique) : Développement d'un chatbot intelligent et automatisation de l'évaluation des sorties RAG via un LLM expert

- Conception d'un chatbot basé sur des **LLMs avec LangChain et RAG** pour l'interaction avec des documents techniques complexes.
- Extraction de textes depuis des documents complexes à l'aide de **YOLO et OCR avancés.**
- **Développement et mise en place d'une pipeline MLOps** complète pour l'entraînement et le **fine-tuning** d'un LLM expert avec **RLHF**, orchestrée via **GitLab CI/CD** et déployée sur **Azure.**

🔹 **Pipeline MLOps pour RLHF**

1. Préparation des données

- Extraction des documents depuis **Azure Blob Storage** et segmentation en chunks**.**

- Génération automatique de **questions-réponses à** partir des chunks en utilisant **GPT-4 (Azure OpenAI).**
- Stockage des **prompts, réponses, annotations et métadonnées** dans une base **MongoDB,** assurant la traçabilité des sources**.**

2. Entraînement du LLM Expert pour l'évaluation des sorties RAG

- Déploiement d'une instance **AzureML** pour entraîner un LLM expert chargé d'annoter les réponses du **RAG** comme un **évaluateur humain.**
- Suivi des hyperparamètres et monitoring via **MLflow (stockage PostgreSQL).**
- Versionnement des modèles et stockage des artefacts sur **Azure Blob Storage.**
- Déploiement du modèle expert via **FastAPI,** permettant l'évaluation automatisée des sorties **RAG.**

3. Fine-tuning d'un LLM avec RLHF

- Utilisation des annotations du LLM expert pour créer un dataset d'apprentissage par renforcement.
- **Fine-tuning** du LLM principal avec la méthode **RLHF (Reinforcement Learning from Human Feedback)** en exploitant les feedbacks générés par le modèle expert.
- Implémentation d'un pipeline **PPO (Proximal Policy Optimization)** avec **Transformers Reinforcement Learning (TRL).**
- Intégration de **MLflow** pour le suivi des expériences et des performances du modèle après chaque phase de **fine-tuning.**

4. Automatisation et orchestration

- Déclenchement dynamique de la pipeline via **GitLab CI/CD** lorsque **:**
    - Un nombre critique de nouveaux documents est détecté sur **Azure Blob Storage**.
    - Les notes du LLM expert baissent, indiquant une dégradation de la qualité des réponses du RAG.
- **Versionnement et orchestration :**
    - **DVC** pour la gestion des données**, Hydra** pour la configuration.
    - Déploiement **Kubernetes,** chaque composant étant exécuté dans un container **Docker** stocké sur **Azure ECR.**

Impact & Résultats

- **Automatisation complète** de l'évaluation des réponses générées par le **RAG** grâce à un **LLM expert autonome.**
- **Amélioration continue** de la qualité du RAG via un **fine-tuning progressif en RLHF,** réduisant la nécessité d'évaluations humaines**.**

- Optimisation des performances et de la scalabilité grâce à une pipeline **MLOps** entièrement orchestrée et versionnée.

(Avril 2023 - Avril 2024)

Mission Stellantis (Automobile) : Automatisation et contrôle de conformité avec des agents IA multimodaux & Sécurité et efficacité des données

Contrôle de la qualité rédactionnelle :

- Développement d'agents IA pour contrôler si les rapports d'inspection automobile respectent les règles définies dans le guide rédactionnel.
- Utilisation de **RAG** et de bases de connaissances vectorielles pour une analyse précise et rapide des contenus textuels.

Sécurité et efficacité des données :
Extraction d'informations sensibles avec Llama 3.2 : 3B, déployé localement dans un environnement Docker, avec stockage sur une base MongoDB on-premise.

Le serveur ne disposait que d'un GPU de 20 Go de VRAM et j'ai donc utilisé un modèle quantifié en INT4. Grâce à son architecture GQA, la mémoire du cache KV était d'environ 0,11 MiB par token. En pratique, cela permettait de gérer environ 150 000 tokens en simultané, soit près de 75 requêtes concurrentes de 2 000 tokens, tout en maintenant une latence stable et une consommation maîtrisée.

- Traitement de documents PDF non structurés (tableaux débordants sur plusieurs pages, cellules non alignées, etc)
- Utilisation de Llama 3.2:3B pour des extractions précises avec gestion locale des données
- Conteneurisation avec Docker et mise en œuvre d'une base de données MongoDB.
- Déploiement de la solution en local

(Septembre 2022 – Mars 2023)

Mission Logistique : Vision par ordinateur et OCR pour le suivi des équipements

- Automatisation de 11,000 fichiers PDF avec YOLOv8 et bases relationnelles.
- Conception et mise en place d'un pipeline MLOps :
    - Tracking des expérimentations et des modèles avec MLflow.
    - Versionnement des données avec DVC et stockage des modèles sur Azure.
    - Déploiement des modèles via FastAPI pour une intégration continue.

- Monitoring des performances et dérives avec Evidently AI.

Environnement technique : Python, LangChain, RAG, YOLOv8, Azure, Docker, MongoDB, FastAPI, MLflow, DVC, Evidently.

ORINOKO *(Septembre 2024 – Décembre 2024, Freelance – Activité parallèle)*

Formateur LLMs
Accompagnement des entreprises dans l'acquisition et l'implémentation des technologies LLM.

- **Conception et animation de formations personnalisées** :
  - Modules couvrant les fondamentaux de l'IA, le fine-tuning et le déploiement des LLMs.
  - Techniques avancées en NLP (Word2Vec, embeddings de mots) et architecture Transformer.
- **Pratiques innovantes en LLMs :**
  - Utilisation de bases de données vectorielles et RAG pour les solutions augmentées.
  - Enseignement du fine-tuning avec RLHF via Vertex AI (Google Cloud).
  - Déploiement et gestion de modèles en production avec des outils comme Comet (LLMOps).
- **Projets réels :**
  - Chatbots personnalisés, outils de résumé de texte, systèmes Q&A interactifs.

Environnement technique : Python, PyTorch, LangChain, HuggingFace, Vertex AI, Comet, Bases de données vectorielles.

AbilyCare (Santé) *(Mars 2024 – Septembre 2024, Freelance – Activité parallèle)*

Formateur Machine Learning & Deep Learning

- **Introduction à la science des données et au machine learning :**
  - Enseignement des fondamentaux statistiques et algorithmes supervisés (régression, forêts aléatoires).
  - Applications concrètes avec des projets sectoriels (analyse santé, logistique).
- **NLP avancé :**

    - Exploration des réseaux RNN, LSTM, Transformers, et embeddings.
- **Encadrement de projets :**
    - Aide à la mise en œuvre de pipelines complets en machine learning.

Environnement technique : Python, pandas, seaborn, TensorFlow, CNNs, Transformers.

Université de Paris Saclay *(Octobre 2018 – Juillet 2022)*

Chargé de Recherche Scientifique & Enseignant de Python

- Recherche sur la dynamique des parois magnétiques pour optimiser le stockage.
- Conception de cours en programmation Python et IA pour Licence 1 à Licence 3.
- Application d'algorithmes de machine learning pour des projets scientifiques.

Environnement technique : Python, Scikit-learn, Pandas, OpenCV, ImageJ.

Formations

- **Doctorat en Sciences des Matériaux** – Université de Paris Saclay (2022)
- **Master en Nano Sciences** – Université de Paris Saclay (2018)

Certifications

- Concepteur développeur en Sciences des Données (2022)
- Python for Data Science, AI & Development (2022)
- ETL and Data Pipelines with Shell, Airflow and Kafka (2022)
- Hands-on to Linux Commands and Shell Scripting (2021)
- Databases and SQL for Data Science with Python (2021)

Langues

- **Français :** Langue maternelle
- **Anglais :** Courant

Projets Notables

- **Chatbot Airbus** : Conception d'un chatbot intelligent basé sur des LLMs pour automatiser l'interaction avec des documents techniques complexes.

- **Extraction de données Stellantis** : Création d'un pipeline sécurisé pour extraire et structurer des données sensibles sans compromission de la confidentialité.
- **Gestion d'équipements logistiques** : Automatisation avec YOLO et Airflow, réduisant de 40 % le temps de gestion.