# Samir MAHAMAT SALEH

## AI & Data Engineer

GenAI · RAG & agents · Data engineering · Évaluation de modèles

Île-de-France | Permis B | mahamatsaleh.samiradam@gmail.com
github.com/MoSaleh47 | linkedin.com/in/samir-adam-mahamat-saleh

## PROFIL

Diplômé Bac+5 en Intelligence Artificielle et Science des données (aivancity, Programme Grande École), **Microsoft Certified: Azure AI Apps and Agents Developer Associate**, et auteur d'une prépublication scientifique sur l'évaluation de LLM. Deux ans d'alternance à la DGAC/DSAC à concevoir et livrer des outils d'IA générative utilisés par de vraies équipes métier, dans un environnement public sensible et réglementaire. Profil à double entrée : d'un côté la construction de traitements de données de bout en bout (ingestion multi-sources, parsing, nettoyage, qualité, APIs, restitution), de l'autre la conception et surtout **l'évaluation rigoureuse** de systèmes RAG, d'agents et de modèles de classification. Poursuite en MSc Data Engineering & Cloud Computing à partir de novembre 2026.

## COMPÉTENCES TECHNIQUES

**Langages**: Python, SQL, TypeScript
**GenAI, LLM et agents**: RAG dense et hybride, embeddings, stratégies de chunking (dont parent-enfant), bases et index vectoriels (FAISS, Azure AI Search), recherche lexicale BM25 / TF-IDF, reciprocal rank fusion, reranking, prompt engineering (zero-shot, few-shot, chain-of-thought), sorties JSON structurées, agents et orchestration multi-passes, self-consistency, Model Context Protocol (MCP), fine-tuning léger LoRA / PEFT
**Évaluation et fiabilité des modèles**: RAGAS, construction de jeux de questions de référence, mesure du taux de réponses non fondées, fidélité des citations, calibration, métriques F0.5 et précision-rappel, intervalles de confiance bootstrap et Wilson, diagnostic de surprédiction, données très déséquilibrées, tests d'injection de prompt, tracing Langfuse
**Machine learning et NLP**: Scikit-learn, TensorFlow, classification, extraction d'information, OCR, réseaux antagonistes génératifs (Pix2Pix), métriques de qualité d'image SSIM / PSNR, Hugging Face, modèles ouverts (Qwen) via Ollama
**Frameworks et modèles**: LangChain, Azure OpenAI, Azure AI Foundry, OpenAI API, Mistral, Anthropic, Groq
**Data engineering**: Pandas, NumPy, DuckDB, ingestion de sources hétérogènes, parsing XML / PDF / JSON, nettoyage, data quality, contrôles de cohérence, pipelines batch, consolidation multi-sources, automatisation de traitements récurrents
**Backend et interfaces**: FastAPI, Flask, APIs REST, Streamlit, Gradio, PyQt5, React 18 avec Vite, Tailwind CSS, React Flow, Recharts
**Cloud et DevOps**: Azure, Docker, Git / GitHub, CI/CD (GitHub Actions, Azure DevOps), tests et logging, développement assisté par IA (GitHub Copilot, Cursor, Claude Code)
**Analyse et restitution**: SQL analytique, Power BI, Tableau, Matplotlib, Excel, documentation technique et fonctionnelle, vulgarisation auprès d'équipes non techniques
**Gouvernance et conformité**: RGPD, règlement européen sur l'IA (AI Act), IA responsable, confidentialité des données, contexte réglementaire public
**En cours d'acquisition (MSc 2026-2027)**: Spark, Kafka, MongoDB, Databricks, AWS, algorithmique distribuée

## CERTIFICATION ET PUBLICATION

**Microsoft Certified: Azure AI Apps and Agents Developer Associate** - Microsoft, juillet 2026 (valide jusqu'en juillet 2027). Credential ID 3F9FED23D318BD9C, certification D6K6AD-CB45D0, vérifiable en ligne. Périmètre : conception d'applications et d'agents IA sur Azure AI Foundry et Azure OpenAI, RAG, orchestration, évaluation et déploiement responsable.
**Prépublication scientifique** - *VCBench* (SSRN, 2026) : étude de méthodologie d'évaluation sur les modes d'échec et la calibration de LLM ouverts exécutés localement pour la prédiction d'événements rares. Résultat négatif assumé, documenté avec diagnostics de calibration. ssrn.com/abstract=6832218

## EXPÉRIENCE PROFESSIONNELLE

**Chargé d'Études en Data Science - DGAC / DSAC** Oct. 2024 - Sept. 2026
*Direction Générale de l'Aviation Civile, Direction de la Sécurité de l'Aviation Civile* *Paris, France*

- Conception d'une solution IA de traitement des déclarations d'événements de sécurité aérienne, transformant des comptes rendus en texte libre en **48 champs structurés au format JSON**, avec nettoyage, validation d'intégrité et contrôles de cohérence automatisés.
- Automatisation du pré-remplissage de formulaires réglementaires PDF à partir des données extraites, ramenant le temps de traitement d'une déclaration de **plus de 30 min à environ 5 min**, soit plus de 80 % de temps gagné sur la saisie et la recherche d'informations.
- Développement d'une chaîne applicative complète en Python comprenant extraction LLM, nettoyage et validation JSON, génération de formulaires, backend FastAPI avec plusieurs endpoints REST et interface web métier.
- Évaluation de la solution sur **une centaine d'occurrences** de sécurité aérienne, avec vérification des sorties structurées, des champs réglementaires et de la cohérence des données avant utilisation.
- Mise à disposition et tests auprès de 5 agents métier, avec recueil des retours utilisateurs, corrections, documentation technique et fonctionnelle et amélioration itérative de l'outil.

**Stagiaire Développeur Python - UBA Tchad** Juin 2023 - Août 2023
*United Bank for Africa - traitement de données bancaires multi-agences* *Tchad*

- Développement d'une application PyQt5 permettant aux équipes métier d'importer, nettoyer et consolider les fichiers Excel provenant de plusieurs agences bancaires.
- Automatisation du nettoyage des données avec Python et Pandas, puis génération de visualisations Matplotlib à l'appui de l'analyse et du reporting.

**Data Analyst Junior - AI Fashion** Janv.-Mars 2022 ; Juil.-Août 2022
*Projet start-up et stage de validation d'année, Clinique de l'IA aivancity* *Remote*

- Collecte et extraction de données web depuis des sites e-commerce et des marques de parfums et de vêtements, segments luxe et streetwear inclus, pour alimenter des analyses produit.
- Nettoyage, structuration et stockage des jeux de données avec Excel et Python / Pandas, puis production de visualisations et d'enseignements exploitables.
- Construction d'un benchmark olfactif fondé sur les ingrédients et les notes, avec analyse comparative des composantes de parfums.

**Data Analyst Junior - Bunqr** Sept. 2021 - Déc. 2021
*Projet start-up, Clinique de l'IA aivancity* *Remote*
- Projet d'automatisation comptable : lecture de factures par OCR et classification algorithmique des documents dans un plan comptable.

## PROJETS TECHNIQUES ET RECHERCHE

**PFE - Benchmark de question-réponse en réglementation aéronautique** | *RAG, FAISS, LLM, LoRA/PEFT* 2025 - 2026
- Projet de fin d'études : chaîne complète d'ingestion du règlement UE 376/2014 et des règles EASA, avec parsing XML et PDF, nettoyage et découpage en 133 chunks, puis construction d'un index vectoriel FAISS.
- Constitution d'un jeu de 300 questions de référence et protocole d'évaluation comparant approches few-shot, RAG et fine-tuning léger.
- Analyse de la fidélité des citations juridiques produites par les modèles et mesure du taux de réponses non fondées.

**AI Act Evidence Desk - agent RAG de conformité** | *RAG hybride, agents, MCP, RAGAS, Langfuse, Gradio* 2026
- Projet open source transformant un cas d'usage IA en cartographie de preuves sourcée au regard du règlement européen sur l'IA, à destination de responsables produit et d'analystes conformité.
- Architecture de recherche : découpage parent-enfant (enfants de 200 mots, parents de 800 mots), recherche hybride BM25 et TF-IDF combinée par reciprocal rank fusion, puis reranking avant assemblage du contexte.
- Raisonnement : synthèse en trois passes avec validation par self-consistency (k=3) et étape de critique explicite du résultat.
- Industrialisation : serveur MCP à trois outils, tracing Langfuse, évaluation RAGAS avec résultats mesurés versionnés, tests de sécurité couvrant cinq scénarios d'injection de prompt, mode hors ligne pour les tests. github.com/MoSaleh47/AI-Act-Evidence-Desk

**VCBench - évaluation de LLM sur données très déséquilibrées** | *Qwen3, Ollama, calibration, TF-IDF* 2026
- Code compagnon de la prépublication SSRN : évaluation de LLM ouverts exécutés localement (Qwen3-32B via Ollama, variante quantifiée GGUF) sur une tâche de classification binaire à très faible taux de positifs.
- Comparaison de stratégies de prompt (vanilla, chain-of-thought, few-shot, hybride) et mesure par F0.5, précision, rappel, taux de prédictions positives, courbes précision-rappel, intervalles bootstrap et Wilson, avec baseline régression logistique sur TF-IDF.
- Conclusion documentée : les gains de F0.5 en petit échantillon sont instables sans diagnostics de calibration, et la baseline classique reste supérieure. github.com/MoSaleh47/vcbench-improvement

**YAKOOTA - assistant d'IA générative** | *Azure OpenAI, LangChain, embeddings, RAG* 2024
- Chatbot documentaire de bout en bout fondé sur le parsing de documents, les embeddings et un workflow de recherche contextualisée, avec tests de robustesse et itérations sur la qualité des réponses.

**EuroMotion Model Factory - simulation de chaîne d'approvisionnement** | *React 18, TypeScript, Vite, React Flow, Recharts* 2026
- Application web de dynamique des systèmes pour la planification stratégique d'une chaîne d'approvisionnement de véhicules électriques et logiciels : moteur de simulation piloté par configuration JSON, diagrammes interactifs et restitutions graphiques. github.com/MoSaleh47/EuroMotion-Model-Factory

**Cartoonizer GAN - traduction d'images** | *TensorFlow, Pix2Pix, OpenCV, PyQt5* 2024
- Réseau antagoniste conditionnel Pix2Pix convertissant des photographies en illustrations de style cartoon : pipeline d'entraînement complet, modèle pré-entraîné livré, application de bureau PyQt5 et évaluation quantitative par SSIM et PSNR. github.com/MoSaleh47/cartoonizer-gan

## FORMATION

**MSc Data Engineering & Cloud Computing - aivancity** Nov. 2026 - 2027
*Titre RNCP Niveau 7 : Expert infrastructure et traitement des données massives. Programme enseigné en anglais* *Villejuif, France*
- Contenu : Spark, Kafka, MongoDB, Databricks, AWS et Azure, algorithmique distribuée, gestion de projets orientés données, certifications cloud.

**Programme Grande École - aivancity School of AI & Data for Business & Society** 2021 - 2026
*Diplôme d'études supérieures en Intelligence Artificielle et Science des données (RNCP38584 Niveau 7, Bac+5)* *Paris, France*
- Cursus hybride associant intelligence artificielle, science des données, culture business, éthique de l'IA, RGPD et AI Act, gestion de projet, innovation et transformation digitale.
- Projets menés au sein de la Clinique de l'IA, dispositif de mise en situation réelle sur des besoins d'entreprises et de start-up.

## LANGUES ET QUALITÉS

**Langues**: Français courant et professionnel. Anglais niveau C1, cursus de master suivi en anglais, lecture quotidienne de littérature technique et scientifique anglophone.
**Qualités**: rigueur méthodologique et goût de la mesure plutôt que de l'intuition, autonomie, curiosité technique, capacité à vulgariser auprès d'interlocuteurs non techniques, documentation systématique, travail en équipe pluridisciplinaire.