

Nolan Cacheux

AI/ML Engineer | Production AI Agents & MLOps

Portfolio: nolanocacheux.com • **GitHub:** github.com/nolanocacheux • **LinkedIn:** linkedin.com/in/nolanocacheux

E-mail: cachnolan@gmail.com • **Téléphone:** +33 6 25 15 62 61 • Paris, France

AI/ML Engineer spécialisé dans les agents d'entreprise en production, les systèmes RAG gouvernés et les pipelines ML dans des environnements cloud, du cadrage client et de l'architecture jusqu'à l'intégration, l'évaluation, le déploiement, la sécurité, l'observabilité et l'adoption.

Disponible immédiatement pour des opportunités à Paris, en France et dans la région EMEA.

FORMATION

JUNIA ISEN - Lille, France

2021 - 2026

Classé 2e sur 190 étudiants en dernière année (Top 1 %)

Diplôme d'Ingénieur, Intelligence Artificielle et Big Data

Spécialisation en AI, Big Data, Data Science & Machine Learning.

- **Cycle préparatoire intégré** (2021 - 2023) · Classé 2e/96 (1re année) · 8e/76 (2e année) · Félicitations des jurys
- **Français:** Langue maternelle · **Anglais:** Courant (TOEIC B2 Certified)

COMPÉTENCES TECHNIQUES

- **AI Engineering:** Python, LangChain, LangGraph, RAG, bases de données vectorielles (Vertex AI Vector Search, Qdrant, Pinecone, Weaviate), évaluation, observabilité, tests de charge, cache, sécurité de l'IA, guardrails, protection des données, confidentialité, mémoire, Agent2Agent (A2A), Gemini Enterprise Agent Platform
- **Machine Learning:** Prévission, Fine-tuning, PyTorch, TensorFlow, scikit-learn, XGBoost, LightGBM, MLOps, MLflow, Weights & Biases
- **Cloud & DevOps:** GCP, AWS, Azure, Docker, Terraform, CI/CD, FastAPI, SonarCloud, Wiz
- **Big Data:** Apache Spark / PySpark, Databricks, SQL, PostgreSQL, NoSQL, Firestore, MongoDB, Redis
- **Travail en équipe:** Git, GitHub, Jira, Confluence, Atlassian, Agile/Scrum

EXPÉRIENCE PROFESSIONNELLE

Decathlon France

Lille, France

AI Engineer

Sept. 2025 - Présent

- Pilotage des réunions de cadrage client et de la livraison de bout en bout de **DAISI** (Decathlon AI Supplier Informations), **Enterprise Agent Google Chat en production** pour les processus fournisseurs et financiers : APIs internes, factures, procédures d'achat, organisations internes et **intégrations Workday Finance**. Conçu pour une population cible de **10 000 utilisateurs** ; **impact : 13 000 h/an économisés**.
- Implémentation de l'architecture ReAct avec LangChain/LangGraph, FastAPI, Vertex AI, Vector Database and Search, Cloud Run, Cloud SQL, MLflow/Databricks et Terraform. Optimisation des coûts, de la latence, du passage à l'échelle et de la sécurité avec Model Armor, DLP, garanties RGPD, tests d'intégration, rapports de validation détaillés, évaluation et observabilité ; test de charge réussi avec **2 000 utilisateurs concurrents**.
- Conception d'une **Agent Factory** interne avec des bibliothèques partagées et des templates de production pour agents, outils, mémoire, interfaces compatibles Agent2Agent (A2A), évaluation, observabilité, sécurité et déploiement, accélérant jusqu'à **5 fois** la mise à disposition de nouveaux agents.
- Pilotage d'**OpsBot**, **Enterprise Agent** pour les processus de maintenance, sécurité et conformité. Mise en place d'un RAG gouverné avec réponses sourcées, réponse de repli sécurisée, traces, évaluation, Vertex AI RAG Engine, RGPD et une librairie interne pour standardiser agents, outils, observabilité et déploiement. **Impact attendu et hypothèse de ROI : 14 500+ h/an réallouées, ~540 k€ de productivité réallouée**.
- Création de **Summarizer**, plateforme de résumé avec Gemini orchestrée par Airflow/MWAA, AWS EKS/S3/ECR, actualisation Delta Databricks, validation à l'exécution et runbooks opérateurs.
- Animation de l'**innovation AI/Tech interne** : newsletter hebdomadaire, guide d'onboarding, documentation/runbooks, **30+ h** de formations GenAI/IA/Tech, hackathons internes et présentations sur le développement d'agents auprès de **100+ personnes**.

- Pilotage d'un **projet stratégique de Machine Learning** ayant remplacé un processus manuel de prévision et industrialisé les prédictions pour **huit KPI de vente clés**, dont le GMV et les articles vendus, segmentés par canal (InStore/OutStore et 1P/3P) pour **64 catégories sportives**.
- Conception et développement d'un **pipeline automatisé de prévision de bout en bout** avec XGBoost, Apache Spark, PySpark, Databricks, AWS S3, Airflow/MWAA et MLflow, couvrant l'ensemble du workflow, de l'ingestion des données brutes à la livraison des prédictions.
- **Amélioration de 15 % de la précision des prévisions** par rapport au processus manuel précédent, mesurée avec le MAPE, grâce à une comparaison rigoureuse de Prophet, XGBoost, LightGBM et Chronos-Bolt, à du feature engineering avancé sur la météo, les jours fériés, les événements, les lags et les fenêtres glissantes, ainsi qu'au traitement parallèle avec joblib. Enregistrement des modèles dans MLflow Model Registry et automatisation des exports Google Sheets et Tableau pour les équipes métier.
- Contribution au lancement du nouveau **ballon de la Jupiler Pro League** avec une performance de freestyle football et la vidéo officielle de présentation avec **Antoine Griezmann**.

Beobank NV/SA
Data Scientist

Brussels, Belgium
Mai 2023 - Août 2023

- Traitement et analyse de données transactionnelles bancaires avec Python, VerticaPy, NumPy, Pandas, SQL et JupyterLab afin d'identifier les lacunes de classification et d'améliorer l'analyse bancaire digitale.
- **Amélioration du taux de catégorisation des transactions de 63 % à 84 %** grâce à l'identification de nouvelles catégories, à l'ajout de règles métier et aux itérations sur la qualité de classification. Support de l'identification des revenus/charges et création de reportings réutilisables pour les analyses en aval.

PROJETS PERSONNELS

Mistral AI Hackathon London 2026 · Ecotopia

GitHub

- Sélectionné parmi **7 000+** candidatures. Hackathon Mistral AI sponsorisé par W&B, NVIDIA, AWS, ElevenLabs et Hugging Face.
Simulation politique par SLMs fine-tunés : 4 modèles QLoRA (NF4), <10 min/modèle. Notre Mistral 8B fine-tuné **surpasse** Mistral Large en sortie structurée avec une **latence 10x inférieure**.
Stack : Spring Boot, Spring AI, TypeScript, Weights & Biases MCP et Hugging Face.

RAG Equity Research Agent

GitHub

- **Prototype personnel de recherche multi-agent déployé** pour l'analyse automatisée d'actions avec revue humaine.
Pipeline RAG évolutif sur des documents SEC avec **reranking** sémantique et agent conversationnel intégré à un canal de messagerie. Déployé sur Azure avec LangGraph, LangChain, Qdrant, Azure OpenAI, FastAPI, Terraform, Docker et Pytest.

AI Product Photo Detector

GitHub

- Pipeline **MLOps prêt pour la production** avec PyTorch et apprentissage par transfert EfficientNet-B0.
Stack : FastAPI, MLflow, Vertex AI Pipelines (KFP), DVC, Prometheus/Grafana, Docker Compose et CI/CD GitHub Actions.

ML Video Comment Analyzer

GitHub

- Application ML full-stack pour la classification automatisée des sentiments avec **BERT**, la **modélisation de thèmes** avec BERTopic et la génération de résumés par LLM.
Développée avec Transformers, Ollama, Next.js, React, FastAPI, SQLAlchemy, Tailwind CSS et Pytest.

CHAMPION DE FREESTYLE FOOTBALL

nolanfree.fr

Athlète **professionnel** et **président** de l'association Nolan-Free depuis 2019 : prestations, ateliers & apparitions médiatiques (M6, France 3).

- **30K+ abonnés** et **35M+ vues** sur les réseaux sociaux.
- **Champion du Monde Junior** (Prague, 2019), **Top 15 Mondial** (2026), **Top 5 Europe** (2024), **2x Vice-Champion de France** (2024, 2025)
- **300+ prestations** : PSG, LOSC Lille, Valenciennes FC, Puma, Umbro, Ferrero
- **Sponsorisé** par Umbro pendant plusieurs années.
- Création de plateformes digitales officielles **WFFA** pour la communauté mondiale du freestyle football : site public, World Ranking, workflows d'événements, fondations judging/admin et flux de données athlètes/membres, utilisées par **1 000+ freestylers dans le monde**.