DARIUS GROS

Ingénieur ML / IA · NLP · Systèmes ML en production

Lille, France · Télétravail · darius.gros@ourkat-technologies.fr · +33 6 29 34 41 54

ourkat-technologies.fr · linkedin.com/in/darius-gros · dariusgros.dev

PROFIL

Ingénieur ML/IA avec 5+ ans d'expérience en mise en production de modèles. Focus NLP, systèmes LLM et fine-tuning (HuggingFace Transformers, PyTorch). Conception de pipelines ML de bout en bout : génération de données, suivi d'expériences, déploiement sous contraintes réelles (RGPD, infrastructure souveraine, données sensibles).

Compétent en architectures RAG, workflows agentiques (LangGraph) et systèmes de recherche hybrides. Base solide en data engineering à grande échelle (Spark, Delta Lake, Airflow) et bonnes pratiques d'ingénierie logicielle (architecture hexagonale, TDD, SOLID).

Autonome, curieux, efficace en équipes techniques resserrées.

COMPÉTENCES CLÉS

Machine Learning & Deep Learning : Python, PyTorch, scikit-learn, classification, régression, clustering, feature engineering, augmentation de données, cross-validation, optimisation de seuil, fonctions de perte, prévision, tests A/B

IA / LLM / NLP : Fine-tuning (CamemBERT, QLoRA), transfer learning, RAG, LangGraph, LangChain, prompt engineering, embeddings, pgvector, ChromaDB, Claude API, Ollama, HuggingFace Transformers, analyse de sentiment, classification de texte

MLOps & Expérimentation : MLflow, suivi d'expériences, versioning de modèles, feedback loops, entraînement config-driven, sweep de seuils, prévention du data leakage, reproductibilité, packaging Docker, monitoring

Data Engineering : Apache Spark, Delta Lake, Databricks, Airflow, PostgreSQL, BigQuery, Redshift, pipelines ETL, modélisation de données

Backend & Infrastructure : FastAPI, Pydantic, API REST, WebSocket, architecture hexagonale, TDD, SOLID, Docker, Traefik, MFA/OTP, conformité RGPD

Cloud & DevOps : AWS, GCP, GitHub Actions, CI/CD, déploiement VPS

Langages : Python, SQL, Scala, R

EXPÉRIENCES PROFESSIONNELLES

Ingénieur ML / IA (Freelance), Ourkat Technologies Fév 2026 – présent

Systèmes IA de bout en bout pour PME. 3 projets ML + 2 sites web livrés.

- Conception d'un pipeline ML complet : ingestion, segmentation documentaire, embeddings, recherche par similarité avec pgvector, évaluation et génération. Orchestration LangGraph en architecture hexagonale.
- Système RAG hybride (BM25 + recherche de proximité pgvector) pour citation automatique de sources dans les rapports générés. Architecture deux couches : squelette de rapport généré par code + module de rédaction IA optionnel activé après validation métier.
- Industrialisation avec packaging Docker, intégration FastAPI, PostgreSQL, déploiement sur infrastructure souveraine et CI/CD.
- MLOps : observabilité avec Langfuse, fine-tuning QLoRA, optimisation des coûts d'inférence et de la latence API.

→ Pipeline RAG pour citation automatique de documents techniques. Rédaction IA conditionnelle avec validation métier.

Co-fondateur & Lead IA/Backend, Korus 2026 – présent

Produit d'augmentation événementielle temps réel. Pipeline NLP en production sur infrastructure souveraine.

- Fine-tuning DistilCamemBERT pour classification binaire de contenu toxique (recall > 0.95). 6 catégories de détection, seuil optimisé à 0.3, weighted CrossEntropyLoss.

- Pipeline de génération de données multi-source : données synthétiques Ollama, web scraping, augmentation adversariale (leetspeak, unicode confusables, zero-width chars). Split train/test/val avant augmentation. Zéro data leakage.
- Framework d'expérimentation config-driven (YAML + MLflow), versioning de modèles avec rollback. Feedback loop production : faux négatifs signalés par opérateur réintégrés au training.
- Backend temps réel avec FastAPI, WebSocket et PostgreSQL, déployé sur VPS souverain UE, conforme RGPD, sans dépendance cloud US.

→ Modèle fine-tuné en production. Pipeline de modération 3 couches dimensionné pour 100 à 5 000 participants.

Consultant IA, Automatisation documentaire, Client PME (confidentiel) Déc 2025 – Fév 2026

- Déploiement d'un système hybride parsing + LLM pour conversion automatique de rapports PDF techniques (géotechnique, multi-fournisseurs) en JSON/Excel structurés. Prompts modulaires de 300+ lignes avec validation par contraintes physiques du domaine.
- Optimisation tokens : extraction text-based via pdfplumber (pas de vision API), traitement page par page. Stratégie de rejet conservateur privilégiant la précision.
- Infrastructure complète : VPS, Docker, intégration API LLM, authentification MFA/OTP, conformité RGPD. Interface Gradio adoptée au quotidien par les équipes métiers.

→ Traitement manuel intégralement automatisé. ~30 % d'économies sur les coûts API LLM.

Data Engineer Confirmé, Pricing & Géomarketing, Decathlon 2023 – 2025

- Industrialisation de pipelines Spark et Delta Lake avec CI/CD complet (GitHub Actions, Airflow, Databricks), en appliquant les bonnes pratiques transposables au cycle de vie ML : versioning, reproductibilité et fiabilité.
- Traduction des besoins métiers en solutions data, coordination transverse avec deux directions métiers et l'équipe data plateforme.
- Migration de stacks legacy, optimisation des coûts cloud, mentorat de data engineers juniors.

→ Temps de déploiement réduit de 2h à 17 min (–85 %). ~20 % d'économies sur l'infrastructure cloud.

Data Engineer / Data Scientist Junior, Decathlon 2021 – 2023

- Pipelines ETL pour l'analytique produit et marketing, modèles de prévision (scikit-learn) et tests A/B en production.
- Segmentation client. Dashboards décisionnels (Tableau, Power BI) utilisés par 3 équipes métiers.

FORMATION

Master Data Science (MIASHS, parcours MQME), Université de Lille

Data Mining, inférence bayésienne (MCMC, Monte Carlo), apprentissage grande dimension (Ridge/LASSO/SVM), séries temporelles (ARMA/Box-Jenkins), économétrie du risque, analyse spatiale

Bachelor of Science, Mathématiques, Wingate University, Caroline du Nord, USA

Mineure Économétrie · Bourse académique et sportive · All-American Scholar · 5 ans aux États-Unis

CERTIFICATIONS

- Deep Learning A-Z, Hands-On Artificial Neural Networks, Udemy / SuperDataScience (22,5 h)
- Machine Learning A-Z, Hands-On Python & R in Data Science, Udemy / SuperDataScience (42,5 h)

LANGUES

- Français, Natif
- Anglais, Courant (C1), 5 ans d'études et de vie aux États-Unis