

Ali Ncibi

AI Engineer | Data, Machine Learning, LLM & MLOps

contact@alincibi.page · +33 7 84 89 08 23 · Massy, Île-de-France · Github · LinkedIn
Immediately available in Île-de-France region.

SUMMARY

AI Engineer and PhD, specialising in Machine Learning, NLP, and production AI systems. Experienced in developing full-stack ML solutions, from data engineering and model optimisation to deployment and evaluation. Hands-on experience building scalable LLM infrastructure and agentic architectures. R&D background with practical data-driven software engineering experience.

LANGUAGES

English — Professional proficiency (C1)

French — Bilingual

Arabic — Native

EXPERIENCE

PhD Research Contract (CDD) Dec 2022 – Jun 2026
INRIA — Saclay, France

- Led R&D of AI-driven solutions for human activity recognition (HAR) in smart environments.
- Structured and analysed 13 public sensor-based datasets with initial heterogeneous formats.
- Developed a containerised standard ML pipeline for model development and evaluation.
- Implemented Bayesian hyper-parameter optimisation and evaluated classical and deep learning models for activity feature extraction, segmentation, and classification using rigorous cross-validation and class-imbalance-aware metrics.
- Evaluated a semantic feature extraction strategy transcribing sensor event sequences into text blocks via templates, then embedding them using language models for recognition. Benchmarked 7 transcription strategies and 3 embedding models, identifying 30% improvement in cross-domain recognition under certain conditions.
- Supervised an intern (3 months): scoped the internship on user-objective evaluation metrics for HAR, tracked progress weekly, and reviewed his integration of the metrics into the standardised ML evaluation framework.

Data/NLP Engineer (CDD) Nov 2021 – May 2022
SouthPigalle — Paris, France

- Collaborated with Data Science and Cloud teams to build NLP solutions for a jewellery shopping chatbot.
- Built a slot-filling corpus using templates and a relational database.
- Developed an interactive interface for data visualisation.
- Fine-tuned a BiLSTM-CRF text-tagging model for slot extraction (+7% averaged balanced accuracy per tag), converted the model to ONNX format, containerised the inference API, and deployed it for production integration.

NLP/Machine Learning Intern Mar 2021 – Sep 2021
INSERM — Clamart, France

- Collaborated with medical teams to develop AI prototypes for clinical timeline reconstruction from medical reports.
- Defined a standard temporal (TLINK) annotation scheme and annotated medical reports to reconstruct patient timelines.
- Fine-tuned text-tagging and classification models for medical concept tagging and temporal relation extraction.
- Developed an interactive patient timeline visualisation tool.

PROJECTS

Wassla — AI-Powered Multilingual News Platform Aug 2026 – Present

Personal project

- Built a trilingual (EN/FR/AR) agentic news pipeline ingesting 18 RSS sources, processing articles through a LangGraph multi-agent workflow (translation, categorization, summarization, independent LLM QA, embedding), and storing results in pgvector; delivered via an SSR reader with semantic RAG search and citation-grounded AI answers.
- Designed a two-level LangGraph architecture with per-article fan-out, a 6-node subgraph, centralized tagged LLM calls, task-based routing, and configurable concurrency.
- Engineered LLM infrastructure using a custom OpenAI-compatible BaseChatModel, structured JSON outputs, token tracing, pluggable providers, deterministic test stubs, prompt-injection defenses, and hardened RSS fetching.
- Implemented observability and evaluation with OpenInference/OTLP tracing to local Phoenix and a golden-dataset regression suite combining deterministic checks and LLM-as-judge scoring, tracked through versioned experiments.
- Secured ingestion with auth., db scheduling, a read-only same-origin API proxy, and QA-based article review gating.
- Enforced quality with 100% pipeline and web test coverage, strict typing, linting, and reproducible make workflows.

https://github.com/ali-ncibi/wassla-app

Self-hosted, auto-scalable LLM platform on Kubernetes clusters Sep 2026 – Present

Personal project

- Deployed a prototype of a self-hosted LLM platform on Kubernetes (OVH MKS), integrating Open WebUI, vLLM (serving a small Qwen2.5-0.5B model as deliberate cost-conscious choice), PostgreSQL/pgvector and Redis, with Prometheus/Grafana monitoring and KEDA/HPA autoscaling across a pool of 3 nodes, each having 2 vCPU cores and 8 GiB memory.
- Optimised inference costs through bfloat16, KV-cache sizing and resource tuning.
- Load-tested scaling from 1 to 3 vLLM replicas using live queue metrics, achieving 394/394 successful requests at 24 concurrent users (23 s end-to-end p95); validated automatic scale-down, and identified load balancer's 50 s idle timeout as bottleneck.
- Production-hardened the deployment by diagnosing infrastructure incidents, securing secrets, provisioning dashboards as code, and delivering a reproducible one-command deployment with an operational run-book.

https://github.com/ali-ncibi/llm-stack

TECHNICAL SKILLS

Programming : Python, SQL

Machine & Deep Learning : Scikit-learn, XGBoost, TensorFlow, PyTorch, Optuna

NLP : Flair, Hugging Face Transformers, NLTK, spaCy

Agentic AI : Prompt Engineering, Agentic Patterns (Reflection, Tool Calling, MCP), LangGraph, LangChain

LLM Deployment & Inference : OpenAI/Gemini compatible APIs, vLLM, Ollama

Monitoring & Evals : Grafana, Prometheus, Arize Phoenix + OpenTelemetry

RAG & Model Adaptation : RAG, Parameter-Efficient Fine-Tuning

Data Engineering & Processing : PostgreSQL, pgvector, Pandas

Data Applications & BI : FastAPI, Streamlit, Seaborn, Plotly, Tableau, Power BI

CI/CD, Containerisation & Orchestration : Git, CI/CD, GitHub Actions, Docker (Compose), Kubernetes

Cloud Platforms : AWS (EKS), GCP (Cloud Run, GKE), OVH (KMS), Runpod

Web & Cross-Platform Application : Nuxt (Vue, Nitro), TypeScript/JavaScript, Capacitor

PUBLICATIONS

Towards Domain-Robust Activity Recognition using Textual Representations of Binary Sensor Events

18th International Conference on Agents and Artificial Intelligence, Marbella, Spain, March 2026

https://hal.science/hal-05401764v1

A Framework for Data-Driven Development of Human Activity Recognition Systems using Binary Sensors
21st International Conference on Distributed Computing in Smart Systems and the Internet of Things, Lucca, Italy, June 2025.

https://hal.science/hal-05213603v1

Une plateforme uniformisée pour la reconnaissance des activités humaines à partir de capteurs binaires
25e Conférence francophone Extraction et Gestion des Connaissances, Strasbourg, France, January 2025.

https://hal.science/hal-04944889v1

EDUCATION

PhD in Computer Science / AI Dec 2022 – Jun 2026
University of Versailles/St-Quentin, Paris-Saclay — Versailles, France

Research focused on machine learning, human activity recognition, and AI-driven data processing pipelines.
Defended successfully on the 1st of June 2026.

Engineer in Data Science / AI Sep 2019 – Jun 2021
Télécom Paris, Institut Polytechnique de Paris — Saclay, France

Specialised in data science, AI, and machine learning applications.

Engineer in IT & Telecommunications Sep 2017 – Jun 2019
École Nationale d'Ingénieurs de Tunis — Tunis, Tunisia

Foundational training in IT, telecommunications, and software engineering.

Preparatory Classes for Entrance to the French Grandes Écoles (CPGE) Sep 2015 – Jun 2017
Institut Préparatoire aux Etudes Scientifiques et Techniques — Marsa, Tunisia

Specialised in mathematics, physics, and sciences of the engineer.