Joshua NANCEY

Forward Deployed AI Engineer & Solutions Architect

M.Eng. UTC • Production LLM Systems, Full-Stack Architecture & Technical Delivery

joshua.nancey@etu.utc.fr • +33 7 49 15 69 60
github.com/joshdeutc • joshdeutc.github.io
linkedin.com/in/joshua-nancey
Paris, France (Open to Relocation: EU / CH / UK)



Professional Experience & Technical Delivery

▸ Epiconcept — R&D Data Science & European Public Health

March 2026 – Sept. 2026 (6 months)

Forward-Deployed AI Research Intern — Clinical LLM Pipelines & High-Stakes Vision — *Paris, France*

- **End-to-End Clinical LLM Architecture (ECDC European Project):** Scoped, architected, and deployed an on-premise clinical extraction pipeline on MIMIC-IV records under strict hospital GDPR and hardware constraints (**single NVIDIA A100 40 GB GPU**). Benchmarked 20 LLM setups (**MedGemma-4B**/**27B**, **Gemma-2-27B**, **Qwen-2.5-32B**, **Mistral-Small-24B**), achieving **F1 = 0.91**.
- **Deterministic Anti-Hallucination Guardrails & Cross-Lingual Rollout:** Engineered a strict JSON schema & substring verifier (`evidence_is_grounded`) enforcing verbatim clinical evidence quotes; validated zero-shot cross-lingual transfer to Croatian hospital reports (**F1 = 0.88**) and authored a 6-chapter engineering playbook (**Quarto**).
- **Trustworthy Medical Vision & Risk Triage:** Fine-tuned the interpretable **GMIC** network on 300+ GB of 16-bit mammograms (**AUC = 0.889**) and deployed a selective abstention layer guaranteeing FNR ≤ 5% and FPR ≤ 5% via Hoeffding bounds.
- **Technical Enablement:** Co-authored an 8-chapter containerized training capsule (**Docker** / **Jupyter**) to onboard partner epidemiologists and data scientists onto deep learning and risk-controlled deployment.

▸ Aubay — Innovation Lab

Sept. 2024 – Feb. 2025 (6 months)

AI Solutions & Full-Stack Data Science Intern — *Boulogne-Billancourt, France*

- **Full-Stack Data Quality & Anomaly Detection Platform:** Architected and delivered an end-to-end web application (**Flask REST API**, **React** frontend, **AWS S3** cloud storage, **Docker**) enabling enterprise stakeholders to audit tabular datasets interactively.
- **Deep Tabular Modeling:** Integrated **PyTorch Tabular** (**GANDALF**, **FT-Transformer**, **TabTransformer**) with **Isolation Forest** / **DIF** to detect contextual anomalies in heterogeneous data (**F1 = 0.903**, **ROC-AUC = 0.985**).

▸ BMBI Lab (CNRS UMR 7338) & Complétude

2022 – 2026

3D/4D Neural Modeling Researcher (5 months) & STEM Instructor (4 years) — *Compiègne / Paris, France*

- **4D Implicit Neural Surfaces (BMBI):** Reconstructed deforming micro-capsules via **SIREN** networks (**IoU = 99.65%**), Eikonal regularization, and 2nd-order `autograd` curvatures (Gauss-Bonnet theorem verified to 0.025%).
- **Technical Communication (Complétude, since 2022):** 4+ years of 1-on-1 mathematics and physics tutoring; proven ability to translate complex quantitative trade-offs for diverse technical and non-technical stakeholders.

Production Systems & Full-Stack Engineering Projects

▸ Simulation-Go (IA04) — Concurrent Autonomous Multi-Agent System & Web Dashboard

Go (Golang), Concurrency, REST API, Web UI

- Built a concurrent multi-agent simulation engine in **Go (goroutines, channels)** modeling decentralized agent negotiation, perception filters, and environment dynamics with an embedded HTTP API and real-time web UI.

▸ MonLycée Cloud Auth Relay — Zero-Trust Session Broker & Manifest V3 Extension

FastAPI, Playwright, AWS/Oracle

- Engineered a cloud microservice offloading **Keycloak SSO** login to headless Chromium (**Playwright**) with mutex concurrency (`asyncio.Lock`), < 50 ms RAM session caching, and 1-click cookie injection into Microsoft Edge via a **Manifest V3 Service Worker**.

▸ Jarvis & Custos — Grounded Multi-Judge RAG Engine (SQLite FTS5) & OS Security Suite

Python, SQLite FTS5, Kotlin, systemd

- Built a parallel multi-ATS ingestion engine (51 European boards) and a 2-pass, 3-LLM evaluation pipeline enforcing deterministic substring grounding; engineered a native **Android Device Owner** security suite (Kotlin) and Linux `systemd` watchdogs.

Education

▸ Master of Engineering (M.Eng.) in Computer Science, AI & Mathematical Modeling

2023 – 2026

Université de Technologie de Compiègne (UTC) — *Compiègne, France*

- **Software & Data Engineering:** Object-Oriented Architecture (C++17 / Qt 6 — *LO21*), Data Warehousing & ETL Pipelines (*NF26*), Statistical Learning (*SY09*), Computer Vision (*SY32*).
- **Mod'Math Honors:** Convex & Combinatorial Optimization (*RO04*), Operations Research & Stochastic Modeling (*RO05*).

▸ Preparatory Classes for Grandes Écoles (CPGE PTSI / PT — Intensive Math & Physics)

2020 – 2023

Lycée Raspail — *Paris, France*

Technical Skills

Applied AI & LLMs : Local LLM Inference (MedGemma, Qwen, Mistral), Constrained JSON, Verbatim Grounding, RAG (FTS5/BM25), PyTorch

Full-Stack & Systems : Python, FastAPI, Flask, React, JS (Manifest V3), C++17, Kotlin, SQL, Docker, Linux (systemd), AWS S3, Git

Solutions & Delivery : Technical Scoping, Rapid Prototyping, Customer Enablement, Quarto Docs

Languages : French (Native), English (Professional Working Proficiency B2+, technical writing & presentations)