Joshua NANCEY

Solutions Architect & Forward Deployed AI Engineer

Ingénieur UTC • Systèmes LLMs en Production, Architecture Full-Stack & Déploiement Critique

joshua.nancey@etu.utc.fr • +33 7 49 15 69 60
github.com/joshdeutc • joshdeutc.github.io
linkedin.com/in/joshua-nancey
Paris / Europe (Mobile) • Permis B



Expériences Professionnelles & Déploiement Terrain

▸ Epiconcept — Équipe R&D Data Science & Santé Publique Européenne *Mars 2026 – Sept. 2026 (6 mois)*

Ingénieur R&D IA & Solutions LLM — Du Cadrage Métier à la Mise en Production *Paris, France*

- **Architecture d'extraction clinique par LLMs locaux (Projet Européen ECDC) :** Cadrage et déploiement d'un pipeline automatisé sur dossiers patients MIMIC-IV sous contrainte d'infrastructure hospitalière stricte (**1× NVIDIA A100 40 Go VRAM**, 100% on-premise pour conformité RGPD). Benchmark de 20 configurations (**MedGemma-4B/27B**, **Gemma-2-27B**, **Qwen-2.5-32B**, **Mistral-Small-24B**) atteignant **F1 = 0,91**.
- **Garde-fous déterministes anti-hallucination & multilinguisme :** Conception d'un vérificateur formel imposant un schéma JSON strict et des citations exactes (`evidence_is_grounded`) vérifiées mot à mot dans le texte source ; validation du transfert zéro-shot vers des hôpitaux croates (**F1 = 0,88**) et rédaction d'un manuel technique de 6 chapitres (**Quarto**).
- **Vision médicale haute résolution & contrôle du risque :** Fine-tuning du réseau **GMIC** sur 300+ Go de mammographies 16 bits (**AUC = 0,889**) et implémentation d'un triage sélectif garantissant FNR ≤ 5% et FPR ≤ 5% (bornes de Hoeffding).
- **Enablement & formation technique :** Co-création d'une capsule interactive de 8 chapitres (**Docker** / **Jupyter**) pour former les data scientists et épidémiologistes au déploiement de modèles profonds fiables.

▸ Aubay — Pôle Innovation *Sept. 2024 – Févr. 2025 (6 mois)*

Ingénieur Solutions Data Science & Full-Stack IA *Boulogne-Billancourt, France*

- **Plateforme SaaS d'audit et détection d'anomalies :** Conception de bout en bout d'une application web industrielle (**API REST Flask**, interface **React**, connecteur cloud **AWS S3**, conteneurisation **Docker**) permettant aux équipes clientes d'auditer la qualité de leurs bases de données.
- **Moteur d'IA tabulaire hybride :** Intégration de **PyTorch Tabular** (**GANDALF**, **FT-Transformer**, **TabTransformer**) et **Isolation Forest** / **DIF** pour détecter les anomalies contextuelles sur données mixtes (**F1 = 0,903**, **ROC-AUC = 0,985**).

▸ Laboratoire BMBI (CNRS UMR 7338) & Complétude *2022 – 2026*

Ingénieur Recherche Modélisation 3D/4D (5 mois) & Enseignant Scientifique (4 ans) *Compiègne / Paris, France*

- **Modélisation 4D par réseaux implicites (BMBI) :** Reconstruction spatio-temporelle de capsules déformables par réseaux **SIREN** (**IoU = 99,65%**), régularisation Eikonale et calcul analytique de courbures par `autograd` d'ordre 2 (Gauss-Bonnet vérifié à 0,025%).
- **Enseignement & vulgarisation (Complétude, depuis 2022) :** Accompagnement personnalisé en mathématiques et physique-chimie ; capacité éprouvée à traduire des concepts abstraits pour des interlocuteurs non-spécialistes.

Architectures Logicielles & Projets Déployés

▸ Simulation-Go (IA04) — Système Multi-Agents Autonomes Concurrents & API Temps Réel *Go (Golang), Concurrency, REST API, Web UI*

- Moteur de simulation multi-agents concurrent en **Go (goroutines, channels)** modélisant la négociation et la prise de décision décentralisée d'agents autonomes avec serveur HTTP embarqué et visualisation temps réel.

▸ MonLycée Cloud Auth Relay — Broker de Session Zero-Trust & Extension Manifest V3 *FastAPI, Playwright, AWS/Oracle*

- Microservice Cloud déportant l'authentification SSO Keycloak via Chromium headless (**Playwright**), avec verrouillage de concurrence (`asyncio.Lock`), cache mémoire TTL (< 50 ms) et injection transparente de 16 cookies via Service Worker **Edge Manifest V3**.

▸ Jarvis & Custos — Moteur RAG Multi-Juges (SQLite FTS5) & Suite Sécurité Android/Linux *Python, SQLite FTS5, Kotlin, systemd*

- Moteur d'ingestion multi-API (51 connecteurs ATS européens) et pipeline d'évaluation à 2 passes et 3 LLMs distincts avec validation déterministe de citations verbatim ; suite de contrôle d'accès **Android Device Owner** (Kotlin) et daemons Linux (`systemd`).

Formation Académique

▸ Diplôme d'Ingénieur en Génie Informatique — Filière IAD & Label Mod'Math *2023 – 2026*

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

- **Ingénierie Logicielle & Data :** Architecture orientée objet (C++17 / Qt 6 — *LO21*), Entrepôts de données & pipelines ETL (*NF26*), Apprentissage statistique (*SY09*), Vision par ordinateur (*SY32*).
- **Label Mod'Math :** Optimisation convexe & combinatoire (*RO04*), Recherche opérationnelle & processus stochastiques (*RO05*).

▸ Classes Préparatoires aux Grandes Écoles (CPGE PTSI / PT) *2020 – 2023*

Lycée Raspail *Paris, France*

Compétences Techniques

LLMs & IA Appliquée : Inférence locale (MedGemma, Qwen, Mistral), JSON contraint, Grounding anti-hallucination, RAG (FTS5/BM25), PyTorch

Ingénierie & Cloud : Python, FastAPI, Flask, React, JS (Manifest V3), C++17, Kotlin, SQL, Docker, Linux (systemd), AWS S3, Git/CI

Architecture & Client : Cadrage technique, Démos & PoC rapides, Documentation Quarto, Pédagogie client

Langues : Français (natif), Anglais (professionnel courant B2+, rédaction & présentations techniques)