Gérome Ferrand

AI Systems Architect

LLM Agent Systems • RAG Architectures • AI Evaluation

Marseille, France

gerome.ferrand@gmail.com | +33 6 78 53 19 37 | gerome ferrand.vercel.app | linkedin.com/in/gerome-ferrand

AI Systems Architect specialized in large-scale language model systems and AI-driven architectures. Experienced in designing **end-to-end LLM systems** integrating retrieval, structured knowledge sources, and tool-based reasoning workflows.

My work focuses on rapid prototyping, **technical feasibility assessment**, and translating product requirements into robust AI system architectures. With a research background in machine learning and representation learning, I emphasize rigorous evaluation, system reliability, and scalable design for real-world AI deployments.

CORE EXPERTISE

- AI Systems Architecture
- Large Language Model Systems
- Retrieval-Augmented Generation (RAG)
- LLM Agents and Tool Integration
- Knowledge Graph + LLM Integration
- AI System Evaluation and Observability
- Rapid AI Prototyping
- Synthetic Dataset Generation

TECHNICAL EXPERIENCES

ML Engineer/AI Systems Engineer, Free Pro, Iliad Group, Marseille — Oct 2023 – Present

- Led applied AI initiatives within the Free Pro Innovation Lab, designing **AI system architectures** and prototyping **LLM-powered capabilities** aligned with product and operational needs.
- Designed and implemented a company-wide **Retrieval-Augmented Generation (RAG) architecture** supporting internal knowledge access and support automation, integrating hybrid retrieval, vector databases, graph-based memory, and tool-enabled LLM agents.
- Built modular **LLM pipelines** combining retrieval, structured knowledge sources, and agent orchestration patterns to enable scalable AI-assisted workflows.
- Developed **evaluation pipelines for LLM systems**, including automated quality assessment, structured output validation, and robustness testing to support system reliability targets.
- Created synthetic and adversarial datasets to stress-test retrieval pipelines and improve embedding robustness and recall stability.
- Implemented **LLM-based evaluation agents** to monitor generation quality, reasoning consistency, and structured outputs in safety-sensitive workflows.
- Contributed to **multimodal AI initiatives** including speech and conversational pipelines (ASR, diarization, TTS) supporting future conversational assistant capabilities.
- Authored documentation and R&D reports supporting national innovation programs (CIR, AMI – BPI France).

Independent Projects — Oct 2022 – Oct 2023

- Developed a YOLO-based computer vision system for precision agriculture drone POC.
- Built a job search automation system for offer collection, normalization, and similarity-based analysis.
- Deployed a personal website for professional visibility and educational outreach.

Data Science Research Intern, LIRIS, CNRS, Lyon — Feb 2022 – Sep 2022

- Built dynamic citation networks with citation context using TF-IDF, word2vec, doc2vec.
- Performed network topology analysis, clustering, feature selection, classification.
- Investigated GNNs and graph embeddings: GCN, graph2vec, node2vec, edge2vec.

Academic Research Project, UCBL, Lyon Nov 2020 – Jun 2021

- Constructed citation context networks from scientific *XML* corpora.
- Enriched citation graphs with semantic context and analyzed structure.

Early Experiences: 2018 – 2020

- Academic Data Visualization Project, UCBL, Lyon: web scraping, data cleaning, statistical analysis.
- Network Infrastructure Intern, Perpignan Hospital Center, Perpignan: deployment of devices, remote supervision, security hardening.
- NLP Research Project, UCBL, Lyon: text classification, summarization, sentiment analysis, ensemble methods, state-of-the-art review.

Real Estate Contract Management (work-study), SNCF, Marseille Oct 2015 – Aug 2018

- Tender management, budget oversight, risk analysis, compliance monitoring.

TECHNICAL SKILLS

- **Programming:** Python, C++, Java, Bash, SQL
- **AI / ML:** PyTorch, Hugging Face, scikit-learn, Pandas, NumPy
- **LLM Systems:** Retrieval-Augmented Generation, agent workflows, structured outputs, LLM evaluation, Pydantic, LangChain
- **Knowledge & Data Systems:** FAISS, Qdrant, Chroma, MongoDB, Neo4j, NetworkX
- **Infrastructure:** Docker, Scaleway, NVIDIA GPUs (H100, L40s, L4)
- **Multimodal / Speech:** Whisper, WhisperX, NVIDIA MSDD, Demucs, TTS, ASR, diarization
- **Tools:** Git, Jupyter, LaTeX

EDUCATION

MSc in Artificial Intelligence, UCBL, Lyon 2020 – 2022
*Co-accredited with **ENS Lyon** and **Centrale Lyon**. Covered advanced topics in machine learning, deep learning, data mining, knowledge discovery, robotics, graph theory, game theory, IoT.*

BSc in Mathematics & Computer Science, UCBL, Lyon 2019 – 2020
Covered OO-Programming, databases, web, networks, concurrent programming, algorithmic complexity, and formal logic. Courses also included applied linear algebra, probabilities and statistics.

DipHE in Mathematics & Computer Science, UCBL, Lyon 2018 – 2019
Covered, among others, foundations in computer science & mathematics, imperative and recursive programming, OS, scientific computing, computer architecture, networks, web, algebra, analysis, probability.

LANGUAGES AND PERSONAL INTERESTS

Languages: French (native), English (professional).
Soft Skills: Clear and effective communication, system-level thinking, cross-functional collaboration & leadership.
Sports: Pàdel, Skiing, Rugby.
Culture: Science, Mathematics, Philosophy, Psychology, SF, Technology, Geopolitics.