CLEMENT PALEZIS

Talence / Bordeaux, France | 07 60 40 08 11 | palezis.c@gmail.com
LinkedIn: https://www.linkedin.com/in/cpalezis | GitLab: https://gitlab.com/Nhkp | Portfolio: https://clement-palezis.dev

AI Software Engineer - GenAI, LLM Agents & MLOps

LLM Applications | LLM Agents | RAG | APIs | ML Evaluation | Performance Computing

SUMMARY

AI Software Engineer with experience building ML pipelines, backend services, APIs, dashboards, and data applications. Strong background in Python, machine learning, HPC, and energy-efficient computing, with a growing focus on GenAI applications including LLM agents, RAG systems, semantic search, and model deployment. Interested in building practical AI tools that improve internal operations, automate document processing, reduce support workload, and support customer-facing workflows.

GENAI & BACKEND SKILLS

GenAI & LLM Applications: LLMs, RAG, Agentic AI, LLM agents, prompt engineering, tool calling, LangChain, LangGraph, Hugging Face, OpenAI API, Mistral API, Gemini API

Retrieval, Evaluation & Observability: Embeddings, semantic search, FAISS, Milvus, vector databases, document processing, LLM/RAG evaluation, retrieval quality, prompt quality, latency/cost monitoring, continuous optimization

Backend & Interfaces: Python, FastAPI, REST APIs, Streamlit, Docker, backend services, API development, data dashboards, POC development

Machine Learning & Data: PyTorch, Scikit-learn, Pandas, Polars, NumPy, PySpark, ML pipelines, benchmarking, credit scoring models

MLOps & Monitoring: MLflow, BentoML, Airflow, Evidently, Grafana, model deployment, model monitoring, drift analysis, CI/CD for ML applications

GENAI & MLOPS PROJECTS

RAG-Based Cultural Recommendation Assistant

Product-oriented prototype | LangChain, FAISS, FastAPI, Streamlit, OpenAgenda, LLM APIs (OpenAI / Mistral / Gemini)

- Built a RAG-based cultural event recommendation assistant powered by OpenAgenda data and natural-language queries.
- Implemented event ingestion, embeddings, FAISS vector retrieval, and response generation with LangChain to recommend relevant events based on user preferences and context.
- Exposed the assistant through a Streamlit test interface and FastAPI-ready architecture to evaluate prompts, retrieval quality, latency, and user interactions.

Credit Scoring Model Monitoring & Drift Analysis

MLOps project | MLflow, Grafana, Evidently

- Designed a monitoring workflow for a credit scoring model using MLflow for experiment tracking, model metadata, and versioned evaluation runs.
- Built Grafana dashboards to monitor prediction quality, operational metrics, and inference behavior over time.
- Used Evidently to analyze data drift, prediction drift, feature distribution changes, and potential performance degradation to support continuous model improvement.

WORK EXPERIENCE

R&D Engineer - Denergium

November 2024 - November 2025 | Energy-efficient computing, ML pipelines, backend services

- Built Python-based tools, backend services, APIs, and dashboards to analyze HPC workload performance and energy metrics.
- Developed ML analysis pipelines and benchmarking frameworks for LLMs and large-scale simulation workloads.
- Designed evaluation workflows to measure workload performance, identify inefficiencies, and support continuous optimization for more efficient compute operations.
- Collaborated on applied research topics involving AI, data processing, and sustainable computing.

Software Engineer - Scalian DS / Thermo Fisher Scientific

September 2022 - November 2024 | R&D engineering, cryo-electron tomography workflows

- Worked as an R&D engineer on cryo-electron tomography workflows for Thermo Fisher Scientific.
- Developed GPU-based 3D template matching approaches that reduced tomography processing time from days to minutes.
- Built state-of-the-art template matching approaches, including GPU-accelerated, tensor-based, and AI-based methods.
- Applied the Segment Anything Model to nanometer-scale tomography segmentation and collaborated with Universidad de Murcia and the Max Planck Institute.

Software Engineer Intern - CEA

March 2022 - August 2022 | Heterogeneous architectures, simulation code

- Developed software around programming models for heterogeneous architectures in a simulation code.
- Worked on numerical simulation, data visualization, parallel/task programming, and heterogeneous cluster computing.

EDUCATION

Master's Degree - AI Engineering | OpenClassrooms | 2026 - Present

Engineering and Data Science Expert - machine learning optimization, model lifecycle management, model performance, CI/CD deployment for ML applications.

Master's Degree - Computer Science | Universite de Bordeaux / ENSEIRB | 2020 - 2022

High Performance Computing and Data Science - architectures, programming models, HPC and machine learning algorithms, execution and visualization environments.

Bachelor's Degree - Computer Science | Universite de Bordeaux | 2019 - 2020

ADDITIONAL SKILLS

Performance Computing: C, C++, CUDA, OpenMP, MPI, SIMD, CMake, Bash, GPU programming, HPC workloads.

Core Expertise: GenAI engineering, LLM applications, chatbots, support automation, document automation, API serving, internal tools, POCs, evaluation workflows, observability.

Languages & Behavioral Skills: French (native), English (professional fluency), adaptability, empathy, curiosity, collaboration, continuous learning.