



# Abdelhafid Dahhani

## AI Architect & Lead AI Product Engineer

AI Engineer with 8+ years of experience designing and deploying production AI systems, specializing in Large Language Models (LLMs), Agentic AI, Natural Language Processing (NLP), Speech Processing, Multimodal AI, and Software Engineering. Experienced in building end-to-end AI platforms, scalable cloud architectures, and intelligent systems for conversational intelligence and large-scale data processing.



+33-755-927-837  
abdelhafid.dahhani@gmail.com  
Eragny (95) - France  
[/in/abdelhafid-dahhani/](#)  
<https://gitlab.com/abdelhafid.dahhani>

## Profile

**PhD:** Information Processing and Data Science

**Experience:** 8+ years of experience in Artificial Intelligence, Software Engineering, Speech Processing, and Natural Language Processing (NLP).

**Entrepreneurship:** Ex-Founder & CTO of FusionLex.

**Business & Consulting:** Experience engaging directly with C-level executives to understand business challenges, identify high-value AI opportunities, and translate business needs into concrete technical solutions.

**Research:** Author of four peer-reviewed scientific publications and former Innovation Engineer at INRIA.

**Recognition:** Hello Tomorrow Emerging Pioneers Finalist 2026.

## Education

### PhD in Information Processing

University of Savoie Mont Blanc (USMB), **France**  
2020 – 2023

#### Studies Areas

- Artificial Intelligence & Machine Learning
- Knowledge Representation
- Ontology and knowledge Graph
- Natural Language Processing (NLP)
- Graph Algorithms
- Software Engineering
- Multi Agent Architecture

### Visiting Researcher (Doctoral Research Stay)

University of Sherbrooke, **Canada**  
2023

#### Studies Focus

- Knowledge Modeling
- Human-Computer Interaction (HCI)

### Master's Degree in Software Engineering

Hassan First University, **Morocco**  
2015 – 2017

Graduated with Honors – GPA: 15.06/20

## Languages

**French** - Fluent

**English** - Fluent

**Arabic (Classical & Moroccan)** - Fluent

## Technical Skills

### Artificial Intelligence & Machine Learning

- Deep Learning (ANNs, RNNs, Autoencoders)
- Natural Language Processing (NLP)
- Large Language Models (LLMs) Integration & Deployment
- LLM Application Engineering (LangChain, LangGraph, LangSmith)
- Agentic AI & Multi-Agent Systems
- Retrieval-Augmented Generation (RAG)
- Prompt Engineering
- Multimodal AI, Conversational AI
- Behavioral Intelligence & Modeling

### Speech & Signal Processing

- Speech Processing & Speech Analytics
- Digital Signal Processing (DSP)
- Time-Series Modeling
- Similarity Analysis
- Speech-to-Text (Whisper, DeepGram)

### Data Engineering & Analytics

- Exploratory Data Analysis (EDA)
- Knowledge Discovery in Databases (KDD)
- Data Preparation & Feature Engineering
- SQLAlchemy, Vector Database (Milvus)

### Software Engineering

- Python, TypeScript
- Flask, FastAPI, RESTful APIs
- Software Architecture
- Microservices, Object-Oriented Design
- SOLID Principles
- MySQL, PostgreSQL, UML
- Design Patterns
- **Custom Text Generation Inference (TGI)**
- **gRPC & Rust**

### MLOps, Cloud & DevOps

- Docker, Git, Jenkins, Terraform, CI/CD
- AWS Cloud, ML Pipelines,
- MLFlow, OpenTelemetry, Kubernetes

### AI & Data Science Libraries

- PyTorch, TensorFlow, Keras
- Scikit-learn, NumPy, Pandas
- Matplotlib, Seaborn

### Semantic Technologies

- RDF, RDFS, OWL, Protégé
- Knowledge Graphs

### Mathematics

- Linear Algebra
- Graph Theory
- Numerical Modeling
- Probability & Statistics

# PhD Research & Publications      Certifications

## Scientific Publications

- Towards a Semantic Model for Wise Systems (**ADVCOMP 2022**) - (Valencia - Spain)
- A Graph Algorithm (**International Journal on Advances in Intelligent Systems, 2023**) - (Valencia - Spain)
- A Graph Matching Algorithm (**FedCSIS 2023**) - Warsaw (Poland)
- Ontology Alignment: **First Step Towards Voice Control (IWCMC 2024)** - Ayia Napa (Cyprus)
- Distance Matching: Towards Near-Human Decisiveness (**Principal scientific contribution of the doctoral thesis - Chapter 5**)

- Natural Language Processing with Deep Learning** – Udemy (2024)
- Develop LLM-Powered Applications with LlamaIndex** – Udemy (2024)
- Finetuning Large Language Models** – DeepLearning.AI (2024)
- Natural Language Processing with Classification and Vector Spaces** – DeepLearning.AI (2024)
- Natural Language Processing with Sequence Models** – DeepLearning.AI (2024)
- AWS – The 12 Core Cloud Services** – Udemy (2023)
- Deep Learning with TensorFlow and Keras** – Udemy (2023)
- Linear Algebra and Feature Selection** – 365 Data Science (2023)
- General Mathematics** – 365 Data Science (2023)
- DevOps: Introduction to Key Foundations and Principles** – Udemy (2022)

# Professional Experience

## Software Architecture

The projects presented below were designed and developed following Clean Architecture principles (Robert C. Martin), emphasizing clear separation of concerns, modularity, scalability, maintainability, and testability.

The solutions include data processing pipelines, RESTful APIs, and analytical components for data extraction, transformation, machine learning, and large-scale data analysis, with a strong focus on production-ready software engineering.

 **Founder - Senior AI Lead**  
FusionLex - **France**

Mai 2024 - June 2026

**2 years**

Designed and deployed a production-ready AI platform supporting daily conversational analysis for a team of 10 sales representatives, processing approximately 300 conversations per day through an asynchronous cloud-native architecture.

## Key Achievements

- Designed and implemented the end-to-end production architecture using Clean Architecture and microservices principles, with a strong focus on scalability, reliability, maintainability, and production readiness.
- Built a cloud-native asynchronous ML inference platform for large audio workloads, using AWS S3 for storage, Redis-based job queuing, background workers, and scalable inference services.
- Took AI models (Emotion & Language Model) from research and experimentation through production deployment, integrating speech-to-text, speaker diarization, NLP, behavioral intelligence, and LLM models into customer-facing products.
- Developed and deployed production REST APIs using Python and FastAPI, exposing AI/ML capabilities through reliable, versioned service interfaces.
- Built a custom **Rust-based Text Generation Inference (TGI)** server for efficient production serving of Llama 3.3 70B, optimizing model serving for latency, throughput, and resource utilization **through t Batching Algorithm**.
- Designed model-serving architectures supporting asynchronous workloads, large audio files, and concurrent inference requests, with a focus on reliability and scalable processing.
- Integrated and orchestrated multiple foundation models including Llama, Whisper, and Hugging Face models, selecting and routing models according to latency, quality, and workload requirements.
- Containerized AI/ML services using Docker and deployed cloud workloads on AWS, applying production deployment and infrastructure best practices.
- Implemented CI/CD and MLOps practices covering model versioning, reproducible deployments, API versioning, experiment tracking, and automated delivery workflows.
- Built production monitoring, logging, and observability pipelines using MLflow/LangSmith/Prometheus, structured logging, and operational metrics to monitor model and service performance.

- Implemented model evaluation and benchmarking pipelines, tracking response quality, inference performance, model behavior, and regressions throughout the development lifecycle.
- Developed advanced RAG and semantic retrieval systems using Milvus, including document ingestion, semantic chunking, embeddings, retrieval, reranking, and grounded generation.
- Designed retrieval strategies for both text documents and long audio transcripts, optimizing context utilization and retrieval quality for production LLM applications.
- Built customer-facing Angular dashboards and interfaces for conversational analytics, behavioral insights, and AI-generated intelligence.
- Designed conversational intelligence metrics and statistical indicators to translate model outputs into actionable product insights for business users.
- Worked directly with business stakeholders to prototype, validate, deploy, and iterate AI capabilities, owning the transition from technical research to production product features.
- Owned the full AI product lifecycle, from model selection and experimentation to production deployment, monitoring, performance optimization, and continuous improvement.

#### Technologies & Expertise

- Artificial Intelligence & Machine Learning
- Speech Processing & Speech Analytics
- Behavioral Intelligence
- Emotion & Human Energy Prediction
- Multimodal AI
- Natural Language Processing (NLP)
- Large Language Models (LLMs)
- Python, Rust, Angular
- REST APIs, Redis & Microservices
- Cloud Computing (AWS EC2, VPC, CloudWatch)
- Data Engineering & Analytics,
- MLFlow, OpenTelemetry, Kubernetes, Docker, Jenkins, CD/CI

### Innovation Engineer

INRIA - **France**

Mai 2024 - October 2025

**1.5 year**

**Context:** Contributed to the design and development of an AI platform for conversational intelligence and human-computer interaction as part of a research and innovation project.

#### Key Achievements

- Designed and developed software components for a conversational AI and analytics platform.
- Deployed and optimized AI models on NVIDIA Tesla V100 GPUs using CUDA, improving inference and training performance.
- Integrated Natural Language Processing (NLP) and Speech Processing technologies into production-oriented AI workflows.
- Contributed to the software architecture and technology stack definition for the platform.
- Designed and implemented technical prototypes to validate research hypotheses and business use cases.
- Collaborated with industrial and academic partners to align technical solutions with market needs and product requirements.

#### Technologies & Expertise

- Artificial Intelligence & Machine Learning
- Natural Language Processing (NLP)
- Pythonn NVIDIA GPUs (Tesla V100), CUDA
- AI Model Deployment, Rapid Prototyping and Research & Innovation

### Software Engineering Consultant

Adneom

March 2020 - September 2020

**6 months**

Client: Amundi - **France**

#### Context:

Contributed to the development of AM'STORE, an internal platform designed to centralize the management of resources, applications, and access rights within Amundi's information system.

### Key Responsibilities

- Developed backend components using Python and PHP/Symfony 3.4.
- Contributed to the development of REST APIs for resource and access rights management.
- Fixed software defects and implemented new features according to business requirements.
- Produced functional and technical documentation for developed features.
- Participated in code reviews and software quality improvement activities.
- Studied and worked with the internal Camelion framework.
- Integrated regression testing into the development pipeline.
- Documented APIs using Swagger/OpenAPI and managed source code with Git.

### Technical Environment

Python • PHP • Symfony 3.4 • REST APIs • Git • Swagger/OpenAPI • Docker • Jenkins • JIRA • Confluence

## Software Engineering Consultant

ACENSI  
Client: I@D - France

October 2019 - March 2020

6 months

### Context

Contributed to the development of I@D's real estate digitalization platform, based on a RESTful microservices architecture.

### Key Responsibilities

- Developed backend microservices using Python (Flask) and PHP (Symfony 3.4).
- Designed and implemented a centralized logging system based on Monolog.
- Developed a microservice for converting and merging Base64-encoded documents.
- Centralized email delivery using Symfony EventDispatcher and Subscribers.
- Fixed software defects and implemented new features according to business requirements.
- Participated in code reviews and improved software quality following Clean Code principles and PSR standards.
- Optimized the Docker deployment process and automated the cloning of project submodules.
- Documented REST APIs using Swagger/OpenAPI.

### Technical Environment

Python • Flask • PHP • Symfony 3.4 • REST APIs • Microservices • Docker • Git • Monolog • Symfony EventDispatcher • JIRA • Confluence

## Software Engineer

RevoTrendz (Consultant), OCP(Consultant), Vili  
Technology (Consultant), Albanki (Inter) - Morocco

July 2015 - October 2019

4 years

### Key Responsibilities

- Developed application log visualization and historical data analysis tools to identify usage patterns.
- Applied the K-Means algorithm for application log classification.
- Fixed software defects according to business requirements.
- Conducted code reviews to ensure compliance with design patterns and software engineering best practices.
- Designed and transformed UML class diagrams into database schemas using the Doctrine ORM.
- Applied design patterns to improve API maintainability and extensibility.
- Installed and configured a Jabber/XMPP server for high-performance instant messaging and real-time database support for push notifications.
- Developed a Node.js architecture for delivering notifications to offline users.
- Developed RESTful web services using Symfony 3 and FastAPI, including JWT-based authentication.
- Prepared and transformed datasets for mathematical optimization models using Gurobi Optimizer.
- Integrated Gurobi Optimizer into a scientific computing application.
- Redesigned a scientific computation application based on mathematical optimization models.
- Modeled a recruitment testing platform for Vili Technology.

### Project Configuration & Infrastructure

- Used Jenkins for Continuous Integration (CI).
- Containerized APIs using Docker.
- Administered development and testing servers on AWS.