Moncef Benaicha

Senior AI Engineer

+49 156 79072442 | contact@moncefbenaicha.com | moncefbenaicha.com | in/moncefbenaicha | github.com/moncefbenaicha

Profile

Senior AI Engineer specializing in NLP, speech recognition, and agentic AI systems, with a strong software engineering foundation. Experienced in building production-grade LLM-based assistants and multi-agent systems using domain-adapted open-weight models and fine-tuned smaller language models. My experience spans applied AI research and production engineering, including leading projects end-to-end across both small and large organizations.

Core Competencies

Programming Languages	Python - C++ - Rust
ML/Data Libraries	PyTorch - HF Transformers - HF PEFT - HF Accelerate - ONNX - DeepSpeed - bitsandbytes - vLLM - TensorRT - MLFlow
LLM Orchestration Frameworks	LangChain - LangGraph - PydanticAI - A2A SDK
Data Orchestration Frameworks	Dagster - Apache Airflow
Parallel/Distributed Programming	OpenMP - MPI - CUDA
APIs and Systems Integration	FastAPI - gRPC - AMQP
Databases and Vector/Search Stores	PostgreSQL - Redis - Qdrant - Azure AI Search
Cloud, CI/CD, and Containerization	Azure - AWS - Docker - Kubernetes - Terraform - GitHub Actions - Azure DevOps

Professional Experience

KI Group - KI Performance GmbH

Cologne, Germany · Hybrid

Senior AI Engineer · Full-time

Jun. 2024 - Present

- Led 15+ AI/ML engagements for major European enterprises across Automotive, Energy, Electronics, Semiconductors, and HR. Driving high client satisfaction and repeat work.
- Designed and delivered multi-agent systems using **LLMs/SLMs** plus speech-to-text **(STT)** and text-to-speech **(TTS)** to automate workflows and power conversational services for customers and internal teams.
- Built a custom recommendation pipeline combining **fine-tuned** embedding-based retrieval with an LLM for **re-ranking** and response generation.
- Partnered with stakeholders in pre-sales and discovery to translate business goals into solution approaches, assess existing tech stacks, and define cloud integration plans.
- Produced and presented multiple architecture options per project, making trade-offs explicit (latency, cost, quality, security) and recommending a path forward.
- Delivered production-grade AI systems end-to-end, mentoring junior AI engineers and coordinating execution across frontend and infrastructure teams.
- Hands-on with modern AI/ML tooling: LangChain, LangGraph, HF Transformers, MLflow, Azure AI Search, Azure AI Foundry, HF Accelerate, vLLM, NVIDIA Triton, TensorRT, Docker, and CI/CD (GitHub Actions, Terraform).

TEQ Capital (10xDNA Capital)

Bonn, Germany · On-site

Machine Learning Engineer · Full-time

Dec. 2023 - Apr. 2024

- Reduced financial analysts workload by ~**80%** by designing a question graph and integrating LLMs **(GPT-4 Turbo)** for question answering on financial data.
- Improved accuracy and completeness of answers over financial documents by building a **GPT-4 Turbo**-based document understanding pipeline that interprets **text, charts, and tables**.
- Grounded **GPT-4 Turbo** responses in up-to-date financial data by building a **Retrieval-Augmented Generation (RAG)** pipeline with re-ranking.
- Enabled scalable ingestion of heterogeneous data sources by building **Python** pipelines with **Dagster** to process documents (PDF, TXT), audio (WAV, MP3), web pages, and external providers via **REST**, **GraphQL**, and **WebSockets**.
- Enabled batch transcription at scale by deploying a transcription pipeline based on the **Whisper** model.

Fraunhofer IAIS

Sankt Augustin, Germany · Hybrid

ML Research Assistant - ASR & NLP · Part-time

Nov. 2021 - Jun. 2023

- Evaluated language representation models (BERT, XLM-R, XLM-V) for **Named Entity Recognition (NER)**, with a focus on cross-lingual transfer learning.
- Built and maintained a data pipeline to support **automatic speech recognition (ASR)** model training and experimentation.
- Developed Spoken Named Entity Recognition models using both cascading (ASR → NER) and end-to-end approaches, leveraging **Wav2Vec2 XLS-R** and **Whisper**.
- Investigated transfer-learning strategies for Spoken NER across low- and high-resource languages to improve multilingual generalization.

Taliox

Hamburg, Germany · Remote

Software Engineer · Part-time

Nov. 2020 - Mar. 2023

- Developed backend web applications and **REST APIs** in **Python** using **Django**, **FastAPI**, and **Flask**.
- Automated recurring workflows with **Python** and **Shell** scripts for task automation, data collection, and data pre/post-processing.
- Built browser automation bots using **Selenium** and **BeautifulSoup (BS4)** for scraping and operational automation.
- Containerized applications with **Docker** to enable consistent cross-platform development environments and reliable deployments.

RWTH Human Language Technology Department - ASR Group

Aachen, Germany · Hybrid

ML Research Assistant - ASR & NLP · Part-time

Dec. 2019 - Nov. 2020

- Investigated how different feature extraction approaches affect **ASR** acoustic model performance.
- Replicated key experiments from the **wav2vec** research line to validate speech representation learning results.

Education

RWTH Aachen
Aachen, Germany

M.Sc. Data Science

- **Thesis:** Spoken Named Entity Processing in Low Resource Scenarios (Supervised by: Dr. Ralf Schlüter, HLTPR Department Sen. Prof. Dr.-Ing. Hermann Ney)
- **Lectures:** Machine Learning (ML) - Natural Language Processing (NLP) - Automatic Speech Recognition (ASR) - High-Performance Computing (HPC) - Parallel and Data-Centric Programming - Combinatorial Optimization

USMBA Fez
Fez, Morocco

M.Sc. Intelligent Systems and Networks

- **Thesis:** Proposal of a ML based approach to estimate the quantity of chemical and organic matter in soil
- **Lectures:** Artificial Intelligence (AI) - Software Engineering - Distributed Computing - System and Network Administration - Information Security and Cryptography

USMBA Fez
Fez, Morocco

B.Sc. Computer Science

Publications

- Leveraging Cross-Lingual Transfer Learning in Spoken Named Entity Recognition Systems - KONVENS 2024 • [Paper] [Github]

Certifications

- DELL EMC - Data Science - Associate
  EAA-007-0715 # Data Science and Big Data Analytics Exam
- Oracle Certified Associate, Java SE 8 Programmer
  1Z0-808 # Java SE 8 Programmer I

Linguistic skills

Arabic	Native
English	C1
French	C1
German	B1