

Ben Younes Sadok Aziz

AI/LLM Engineer

 sadokazizbenyounes.me  sadokaziz.benyounes@ensi-uma.tn  benyounes-sadokaziz  benyounes-sadokaziz

Education

National School of Computer Science (ENSI), Manouba, Tunisia

Graduated Sept 2026

Engineering Degree in Computer Science

Preparatory Scientific Cycle: Mathematics – Physics

Sept 2021 – Jun 2023

Higher School of Science and Technology of Hammam Sousse (ESSTHS)

Experience

Applied AI & MLOps Engineering Internship – Ora Studio

Apr 2026 – Aug 2026 (PFE)

- Built Voice Vibe, a real-time multi-agent voice system with six specialized LangGraph agents and streaming WebSocket STT/TTS, achieving ~1s p50 first-audio latency — mid-pack among cascade voice platforms, while native voice-to-voice models are faster but can't orchestrate agents.
- Designed an end-to-end MLOps pipeline (CRISP-DM) for a suggestion-acceptance recommender: leak-free feature engineering, calibrated LightGBM/Logistic Regression models tracked with MLflow and DVC, and a deterministic promotion gate with shadow → gate → promote deployment.
- Hardened the system with prompt-injection defenses, WebSocket rate-limiting, and a GitHub Actions CI/CD pipeline with automated testing and dependency auditing.
- Technologies:** FastAPI, LangGraph, Mistral, DeepSeek, PostgreSQL, LightGBM, MLflow, DVC, Docker, GitHub Actions, pytest.

AI and LLM Engineering Internship – BA-Consulting

Jun 2025 – Aug 2025

- Built custom MCP servers (Calendar, Gmail) and a QLoRA-based TinyLlama fine-tuning pipeline (4-bit quantization) orchestrated with ZenML/MLflow and Docker, deploying Gen AI Toolbox database tools on AWS Lambda for serverless operation.
- Technologies:** FastMCP, LangChain, Docker, AWS Lambda, DynamoDB, SQLite, Unsloth, ZenML, MLflow.

Full-Stack Software Development Engineering Internship – Core Techs Solutions

Jul 2024 – Aug 2024

- Optimized backend data structures with hash tables and built real-time dashboard components turning complex vehicle data into actionable visualizations.
- Technologies:** Spring Boot, Angular, PostgreSQL.

Projects

Intelligent Traffic Monitoring System – Computer Vision

Jul 2025 – Oct 2025

- Implemented 9-point homography calibration with outlier rejection to convert YOLOv8 + ByteTrack detections into real-world vehicle speed estimates.
- Built a plate-quality assessor and per-class violation thresholds, keeping one clean, readable license plate crop per vehicle instead of hundreds of duplicate frames.
- Technologies:** YOLOv8, ByteTrack, FastAPI, Redis, Celery, Docker, Prometheus, Grafana, MLflow, Streamlit, WebSocket

Sign Bridge – Computer Vision and NLP

Jan 2025 – May 2025

- Built two real-time sign-language translation systems — a FastAPI/NLP backend and a Unity avatar client — driven by OpenPose skeletal keypoints across 3 sign language datasets.
- Applied BERT-based NLP to translate speech and text at the level of meaning rather than word-for-word transcription, driving accurate 3D avatar sign generation.
- Technologies:** FastAPI, JavaScript, Unity, OpenPose, PostgreSQL, BERT, Transformers, NLTK, Pandas

Skills

Programming: Java, Python, JavaScript, TypeScript

Databases: PostgreSQL, MySQL, DynamoDB, SQLite

Frameworks & GenAI: Angular, Spring Boot, FastAPI, LangChain, LangGraph, FastMCP

Machine Learning & CV: PyTorch, YOLO, OpenCV, ByteTrack, TensorFlow, Scikit-learn, LightGBM, NLTK, OCR (Tesseract/EasyOCR), Pandas, BERT, Transformers, Unsloth, NumPy

MLOps: Docker, GitHub Actions, Prometheus, Grafana, CI/CD, ZenML, MLflow, DVC, pytest

Tools & Protocols: AWS, Redis, WebSocket, Google Gen AI Toolbox, CUDA, bitsandbytes

Languages: English (Advanced), French (Advanced), Arabic (Native)