

Mohamd Ahmed Taleb Bouhemadi

APPLIED AI & LLM ENGINEER · PYTHON BACKEND · MODEL EVALUATION

bouhemaditaleb@gmail.com · +90 501 342 24 54 · Istanbul, Türkiye
linkedin.com/in/mohamd-bouhemadi · github.com/adel6-ahmed · huggingface.co/adeltlb

PROFILE

Computer Engineering M.Sc. student and software developer focused on applied AI and LLM applications — function calling, RAG, and Python backend systems. I turn AI capabilities into working product features and backend workflows, from model fine-tuning and evaluation to deployed APIs. Seeking an Applied AI / LLM internship or junior role.

CORE SKILLS

Applied AI / LLM: LLMs, Function Calling / Tool Use, RAG, AI Agents, LoRA / QLoRA / PEFT, Model Fine-Tuning, Model Evaluation

Backend: Python, FastAPI, REST APIs, PostgreSQL, Supabase, SQL, Docker, Git, Linux

Frameworks: PyTorch, Hugging Face, Pydantic, llama.cpp (*learning*)

Computer Vision (supporting): Computer Vision, YOLO, OpenCV, ByteTrack

EXPERIENCE

Software Developer — DAKAEi October 2024 - June 2026

- Contributed to the Python/FastAPI backend, database, REST APIs, and AI-driven features of Umsiyeti, a digital services product released on the Apple App Store.
- Contributed to backend workflows, database integration, and external API integrations for AI-enabled product features.
- Supported model and API integrations, debugging, and production delivery.

Mentor — Software Engineering, Cybersecurity and AI — HET Academic Training & Development June 2024 - August 2024

- Mentored international students in Software Engineering, Cybersecurity, and Artificial Intelligence.
- Helped participants understand technical concepts and structure their learning activities.

RESEARCH

What Actually Matters for Small-Model Function Calling?

- Fine-tuned Qwen2.5-1.5B-Instruct with QLoRA on 4,258 glaive-function-calling-v2 examples (18.46M trainable params, 1.18%) on a single NVIDIA T4 GPU (24.8 min, 3.64 GB peak).
- Exact-match accuracy rose from **8.58% to 53.73%** on 271 held-out examples — a 45.15-percentage-point (526% relative) improvement (McNemar $p < 10^{-20}$), stable across three seeds ($54.35\% \pm 1.41\%$); adapters and code released publicly.

PROJECTS

Umsiyeti Multilingual WhatsApp Customer Assistant

- Built a multilingual WhatsApp customer assistant within the broader Umsiyeti ecosystem, connected to a live service database via Supabase / PostgreSQL.
- Integrated the WhatsApp Cloud API with LLM function calling, implementing backend workflows that connect AI-generated actions to real service data.

Hassaniya Dialect LoRA — LoRA fine-tuning to adapt a language model to Hassaniya Arabic, a low-resource Arabic variety.

M.SC. THESIS (IN PROGRESS)

Reliability-Aware Compression of Small Language Models for On-Device Function Calling under Intermittent Connectivity — studying how quantization (INT8/INT4) affects function-calling reliability, with grammar-constrained decoding and reliability-aware QLoRA for recovery on edge hardware.

EDUCATION

M.Sc. Computer Engineering — Kocaeli University, Türkiye Sep 2025 - Expected Jun 2027
Grade: 3.6 / 4.0

B.Sc. Computer Engineering — Bolu Abant İzzet Baysal University, Türkiye Sep 2020 - Jun 2024

LANGUAGES

Arabic — Native · Turkish — C1 · French — C1 · English — Intermediate (B1), actively improving