UMUT ABALI

abaliumut@outlook.com | +90 544-564-8799 | LinkedIn | GitHub | HuggingFace

EDUCATION

Istanbul Medeniyet University *Istanbul, TUR*
B.Sc. Computer Engineering (GPA: 3.20) *2022 – 2026*

EXPERIENCE

Mapin Data *Istanbul, TUR*
Data Engineering Intern *Jan 2026 – Mar 2026*

- Built a data pipeline to extract and structure unstructured information from web pages, PDFs, and QR codes using FireCrawl, OpenCV, and Visual Language Models (VLMs). Reduced token costs by 70% through OpenRouter optimization and custom text-chunking algorithms, keeping accuracy high. Worked alongside the core team to clean and map this parsed data into a production schema, making it ready for downstream AI tasks.

Princess Cruises *Alaska,USA*
Public Areas Attendant *Jun 2025 – Oct 2025*

- Participated in the Work & Travel USA program and worked in a fast-paced environment on a cruise ship; improved English communication skills by talking with international guests daily and worked effectively within a multicultural team to maintain high service standards.

Menatek Technologies *Istanbul, TUR*
Software Engineer Intern *Mar 2025 – May 2025*

- Designed and built a full-stack Inventory management system from scratch using PHP and MySQL to digitalize and centralize asset tracking; architected relational database schemas with optimized indexing to ensure strict data integrity, while implementing secure authentication and Role-Based Access Control to safeguard sensitive operational actions based on user roles.

PROJECTS

- **Mini-CDN**: Built a prototype CDN in Go featuring distributed edge nodes, a round-robin load balancer, and origin fallback mechanisms. Developed the caching layer with TTL expiration, LRU eviction, and manual cache purging APIs. Integrated Grafana metrics to track cache hit/miss ratios and monitor real-time content delivery performance.
- **Agentic RAG System**: Built a multi-agent retrieval system using LangGraph, setting up supervisor and execution agents to handle complex user queries. Developed a hybrid search engine that combines BM25 and dense vectors with RRF fusion and cross-encoder reranking, using Qdrant with multi-tenant isolation. Created the full-stack setup with a Next.js frontend, FastAPI for SSE streaming, JWT authentication, and Redis/Postgres for session memory, containerizing everything via Docker Compose and using vLLM for Qwen3-8B-AWQ inference.
- **Turkish Gemma-4 Fine-tune**: Fine-tuned the Gemma model on a custom dataset of 3,000 Turkish instructions using Unsloth and LoRA to optimize training efficiency. Managed the entire training process on a single A100 GPU, completing 2 epochs in under 45 minutes by configuring roughly 18M trainable parameters. Published the final weights to Hugging Face (uabali/gemma4-e4b-TR), making the model available in LoRA adapter, merged 16-bit, and GGUF formats for the community.
- **Multimodal Agentic Chatbot**: Developed a fully local, GPU-accelerated chatbot powered by Gemma-4 E4B (llama.cpp/GGUF) with a voice-enabled Chainlit streaming UI. Implemented a LangGraph ReAct loop featuring a CRAG web-search fallback and MCP server integration for dynamic GitHub, filesystem, and calendar management. Integrated multimodal PDF ingestion capabilities to seamlessly process and synthesize both text and visual data.

TECHNICAL SKILLS

Prog. Languages: Python, C/C++, Bash/Shell

AI & ML: PyTorch, OpenCV, LangChain, LangGraph, vLLM, llama.cpp

LLMOps & Agents: RAG, Agentic Systems, LLM Orchestration, Vector DB(Qdrant), MCP

Cloud & Backend: AWS, GCP, Docker, FastAPI, Linux

Languages: Turkish , English(Professional Working)