Onur TAŞKIN

AI Engineer — Agentic Systems & Backend

Ankara, Turkey
onurtaskin180@gmail.com
www.onurtaskin.com
onurtaskinn
onurtaskinn

Profile Summary

Software engineer building production AI systems: agents, RAG pipelines and scalable Python backends. I ship LLM services end to end, fine-tune and evaluate models, and optimize pipelines for cost and latency. Computer Engineering graduate from METU.

Work Experience

09/2024 – Present **Software Engineer — Applied AI**, *Enocta*

- Built AI microservices with FastAPI, JWT and SQLAlchemy, deployed on-premise with Docker Compose and nginx.
- Built a framework-free tool-calling agent with Gemini Function Calling for a production learning-path generator.
- Designed a RAG pipeline with HyDE, Weaviate and Cohere reranking that grounds every recommendation in real catalog data.
- Developed Slide & Podcast Studio, multimodal content pipelines producing ~50 media assets daily.
- Built a subtitle service with parallel processing and structured outputs that translated a 10,000-video catalog.
- Fine-tuned Gemma 4 12B with QLoRA and evaluated it on a gold set using a blind LLM judge.
- Reduced cost and latency through model cascading, provider fallback and per-request cost tracking.

07/2024 – 08/2024 **Software Engineering Intern**, *ASELSAN*

Built a template management feature for an internal engineering tool with React and Java Spring, including offline save and restore.

07/2022 – 09/2022 **Software Engineering Intern**, *TÜBİTAK BİLGEM İLTAREN*

Built a full-stack data bank application with a .NET Core Web API and a cross-platform .NET MAUI client.

Projects

10/2026 – 11/2026 **Hermo – Multi-Tenant Agent Platform**, *Python, Docker, .NET 8, Claude, OpenAI GPT-Live*

- Built a Microsoft Teams assistant that gives each user an isolated, sandboxed agent container.
- Added real-time voice calls where the voice agent hands tasks off to the user's agent, with ~1 s response latency.

2025 – Present **Capter – AI Subtitling SaaS**, *FastAPI, PostgreSQL, Gemini, FFmpeg, React, Docker*

- Built a live subtitling product supporting 102 languages, with Stripe payments.
- Processed media concurrently with asyncio and added model fallback with exponential backoff.

2025 **Intelligent Document Processing Service**, *FastAPI, Gemini, OpenAI, Claude, Weaviate, PostgreSQL*

- Built a service that extracts structured data from scanned documents using OCR and LLMs.
- Added a multi-provider LLM router, RAG-selected examples and per-request cost logging.

2025 **Agent Framework Benchmark**, *LangGraph, CrewAI, pydantic-ai, smolagents*
Compared four agent frameworks on the same task, tools and retrieval layer.

2025 **OVec – Vector Database from Scratch**, *Python, NumPy*
Built a vector database with k-means++ clustering and an IVF index.

Education

2021 – 2025 **Bachelor's Degree, Computer Engineering**, *Middle East Technical University*, Ankara, Turkey

- GPA: 3.41/4.00
- Key Courses: Introduction to Deep Learning, Natural Language Processing, Data Structures and Algorithms, Database Management Systems, Operating Systems, Computer Networks, Software Engineering, Object-Oriented Programming

2023 **Exchange Program**, *Autonomous University of Madrid*, Madrid, Spain

Skills

Programming Languages	Python (asyncio, concurrency, OOP), C#, JavaScript/TypeScript, Java, C++
AI & Agents	Tool Calling, Agent Orchestration, Structured Outputs, RAG, HyDE, Reranking, LangGraph, CrewAI, pydantic-ai, smolagents, OpenAI / Anthropic / Gemini APIs
Machine Learning	PyTorch, Transformers, PEFT/TRL (QLoRA), scikit-learn, Ollama, Embeddings, Semantic Search
Evaluation & Observability	Gold Sets, LLM-as-Judge, Benchmarking, Cost/Latency Tracking, Langfuse, Logfire
Backend	FastAPI, REST, SSE, JWT, SQLAlchemy, Microservices
Databases	PostgreSQL, MS SQL Server, SQLite, Weaviate, Qdrant
DevOps	Docker, Docker Compose, nginx, Caddy, Linux, Azure

Language Skills

Turkish	Native	
English	Advanced	*Proficient in writing, speaking, and listening*
Spanish	Beginner	*A1*