Barış Deniz Sağlam

Senior AI Engineer · AI Researcher

Ankara, Turkey · open to relocation · bdsaglam@gmail.com · github.com/bdsaglam · linkedin.com/in/bdsaglam · scholar.google.com/citations?user=ItgIKcYAAAAJ

Senior engineer with 6+ years building and owning production AI systems end-to-end — architecture, system design, evaluation, and delivery — across health tech, insurance, legal tech, and aerospace. Now pursuing a PhD, with research interests in foundation models, continual learning, and neuro-symbolic AI.

Experience

Senior AI Engineer — Independent Contractor · 2025–present · Remote *Bosch Legends Lab, Novee AI, Genie AI*

- Own architecture, system design, and tech-stack decisions for production AI systems — backend, infrastructure, and the evaluation pipelines behind RAG systems and agents.
- **Bosch Legends Lab:** De-facto architecture owner of an AI expert-sourcing platform. Introduced Temporal workflow orchestration to replace an event-driven design unfit for long-running, human-in-the-loop AI workflows, and led a microservices → modular-monolith migration (RFCs, team alignment). Built the pipeline end-to-end: CV parsing, profile generation, a web-exploration research agent, and a hybrid-search + re-rank + RAG expert matcher.
- **Novee AI:** Architected LLM structured-extraction systems for insurance underwriting (energy-asset, war/terrorism perils, general risk), combining agentic and RAG approaches over heterogeneous schemas; built the domain-specific evaluation frameworks, failure-mode loops, and human-in-the-loop labeling that anchor iteration; shipped a contract-drift detection algorithm.
- **Genie AI:** Prototyped LLM agents for a legal-writing assistant.

AI Software Engineer — Peppy Health / Euphoric · 2023–2025 · London, UK (remote)

- Core contributor and senior engineer on the AI platform that became Euphoric, now a standalone company.
- Designed and built production RAG systems — a benefits navigator and health chatbot — with pgvector, Weaviate, FastAPI, and OpenAI.
- Owned evaluation as the product's backbone: pipelines combining classical IR metrics, domain-specific measures, and expert feedback; applied A/B testing and statistical analysis to LLM features.
- Ran internal knowledge-sharing on information retrieval (keyword → hybrid search → re-ranking).

Senior Software Engineer — Tractable · 2021–2023 · London, UK (remote)

- Delivered a production ML system for insurance claim-leakage assessment that directly enabled expansion into new European markets.

Software Development Engineer — Amazon Web Services · 2020–2021 · Berlin, Germany (hybrid)

- Built highly available, low-latency web services for cloud infrastructure.

Machine Learning Engineer — Mediate (MIT spin-off) · 2018–2020 · Boston, USA (remote)

- Founding engineer at an MIT spin-out building Supersense, an assistive-tech app for blind and visually impaired users; established the ML and mobile architecture end-to-end.
- Built a deep-learning document edge-detection system that outperformed classical computer vision in cluttered real-world scenes, a MobileNet intent classifier, and a Core ML 3D document-orientation system; deployed on-device with TensorFlow Lite and Core ML.

R&D Engineer — Turkish Aerospace Industries · 2013–2017 · Ankara, Turkey

- Built in-house flight-test data analysis and validation software (Python, NumPy/SciPy/Pandas, MATLAB), replacing costly tooling and improving team productivity.

Education

PhD, Information Systems — METU · 2025–present Deep learning, language models, reinforcement learning, information retrieval.

MSc, Data Informatics — METU · 2021–2024 · CGPA 3.93 Thesis: *Knowledge Graph Augmented Multi-Hop Question Answering Using Large Language Models.*

BSc, Mechanical Engineering — METU · 2006–2012

Selected projects

- **Implicit Program Synthesis for Abstract Reasoning (epiq)** — an LLM agent solving ARC-AGI puzzles by implicit program synthesis with grounded symbolic execution and holdout verification. *github.com/bdsaglam/epiq*
- **Fine-tuning vs. Prompting for LLM Adaptation** — showed DSPy prompt optimization can match LoRA fine-tuning at far lower cost on entity–relation extraction.
- **Reinforcement Learning for Multi-Hop QA** — GRPO-trained Llama-3.1-8B for agentic retrieval; answer F1 0.30 → 0.48 on MuSiQue.

Publications & awards

- **MSc Thesis** — *Knowledge Graph Augmented Multi-Hop Question Answering Using Large Language Models.* METU, 2024.
- **DocVQA Challenge** — joint winner, ICDAR 2026.

Skills

- **AI engineering** — production AI products end-to-end: RAG systems, agents, and LLM fine-tuning, serving, and post-training.
- **Evaluation** — systematic evaluation of AI systems: evaluation pipelines, offline/online metrics, human/judge alignment, and statistical rigor.
- **Architecture** — system design and architecture for AI products; backend, infrastructure, and observability.
- **Research** — rigorous experimental methodology: strong baselines, ablations, and honest measurement.
- **Leadership** — technical direction, tech-stack decisions, and cross-functional collaboration.