Burak Gözütok

Senior AI Engineer · LLM agents, RAG, evaluation and production AI
Antalya, Turkey (UTC+3) · Independent contractor through my own company, invoicing in USD or EUR
burakgztk@gmail.com · linkedin.com/in/burakgozutok · github.com/burakgzt · burakgozutok.com · Google Scholar

SUMMARY

Senior AI engineer with 7+ years of shipping machine learning and LLM systems to production. I build LLM agents and RAG pipelines, check them against real outcomes, and deploy them with the backend and AWS infrastructure they run on. I built and ran a multi-model LLM SaaS on my own, was a Senior Principal Software Engineer and team tech lead at Trilogy, and led the data science team of a 10M-user platform. Author of 7 peer-reviewed ML papers.

SKILLS

LLM and agents: AI coding agents in daily work (Claude Code, Cursor), OpenAI, Anthropic and Gemini APIs, open-weight models, LangGraph, LangChain, tool use, multi-agent orchestration, structured outputs (Pydantic), prompt caching, model routing, streaming, server-side guardrails, LLM call logging for cost and quality

RAG and search: vector databases (pgvector, Qdrant, FAISS, Pinecone, ChromaDB), Elasticsearch, graph databases, search-grounded generation, reranking

ML: PyTorch, TensorFlow, Hugging Face Transformers, BERT, LoRA fine-tuning, CNNs, recommendation systems, text-to-speech, SageMaker

Backend and data: Python (FastAPI, Django), Strawberry GraphQL, Node.js, TypeScript, PostgreSQL, Temporal, REST, WebSockets, Playwright

Cloud and DevOps: AWS (Lambda, Fargate, SageMaker, API Gateway, DynamoDB, S3, CloudFront, Cognito, IAM), Pulumi, AWS CDK, Docker, GitHub Actions, Railway

Frontend and mobile: React, Next.js, Svelte/SvelteKit, React Native (Expo), Ionic/Capacitor, PowerSync (offline-first sync)

EXPERIENCE

Senior Software Engineer · Independent contractor

Client in France (remote) · Sept 2024 – present

- AI and cloud-native applications with LLMs and RAG; NLP text classification with multilingual BERT models and LLM classifiers, served serverless on AWS SageMaker and Lambda with batch and parallel processing.
- Large-scale crawler and browser-automation pipelines (Playwright), with background workers orchestrated by Temporal and autoscaled on AWS Fargate.
- Backend services in Python (Django with Strawberry GraphQL and subscriptions, FastAPI), PostgreSQL, immudb, AWS Cognito and Stripe billing, containerized with Docker.
- Web and mobile apps in Next.js, SvelteKit, React Native and Ionic/Capacitor, with offline-first sync via PowerSync.
- Infrastructure as code with Pulumi on AWS and Railway; my deployment pipelines doubled deployment speed.

AI Engineering Lead · Donanımhaber (contract)

Istanbul, Turkey (remote) · 2024 – present

- Kept on as a contractor after leading the data science team (below); I lead the AI projects: RAG and LLM applications with intent classification, backed by graph databases.
- Text-to-speech and LoRA fine-tuning of open-weight models.
- Satellite image processing.

Co-founder and sole developer · Contentrare

Remote · Dec 2024 – Oct 2026 (closed)

- Built the whole product alone (540 commits): a multi-tenant SaaS that writes long-form SEO articles in 20+ languages.
- Each article runs as a state machine of about 20 LLM steps (keyword research, outline, section-by-section writing, internal linking, title and meta), each returning a typed Pydantic object, so a failed step retries on its own instead of restarting the article.

- Multi-model across OpenAI, Anthropic and Gemini, with a search-grounded research step that checks drafts against sources; every LLM call logged per article to track cost and output quality.
- FastAPI, Strawberry GraphQL, PostgreSQL and SvelteKit, deployed with Pulumi on AWS Lambda; teams, SSO, MFA and credit-based billing (Paddle).

Senior Principal Software Engineer (AI Super Engineer) · Trilogy
United States (remote) · Jan 2024 – Aug 2024

- Technical lead of the Content Generation Platform team: delivered an AI-powered SAT prep platform that generated 100K+ personalized questions, videos and quizzes with LLMs and generative models, across text, image, voice and video generation, the backend and the AWS deployment. Recommendation letter from Trilogy's VP of Software Engineering available.
- Built an AI customer support agent that automates about 80% of user queries.
- Designed RAG pipelines with vector databases for knowledge-grounded answers, and internal AI knowledge bases that help engineers find project context and documentation.
- Built AI agents for developer workflows that cut manual engineering effort by about 40%.
- Delivered AWS infrastructure with CDK (Lambda, DynamoDB, API Gateway, S3): fully automated deployments with 99.9% uptime; backend in Node.js and TypeScript.

Data Science Team Lead · Donanımhaber R&D Center
Istanbul, Turkey · Mar 2019 – Jan 2024

- Led the data science team of a forum and social platform with 10M+ users, and publicly funded R&D projects (TÜBİTAK 1501, Eureka, ITEA) from research to production.
- NLP models (BERT, Transformers, FastText) for toxic content detection at 95%+ accuracy.
- Recommendation and feed-ranking systems that raised engagement by about 20%; search reranking and ad matching.
- Sensitive-image classification with CNNs, with Elasticsearch and FAISS for retrieval; LLM and LoRA-based deployments, all running as Docker services. The R&D produced 6 of my peer-reviewed papers.

Co-founder and Full-Stack Developer · Algelsin E-Commerce Platform
Antalya, Turkey · 2018 – Feb 2021

- React Native apps (iOS and Android) and backend services (.NET, PHP, MongoDB) handling 50K+ real-time requests per minute; ML recommendation and notification systems that raised customer engagement by about 25%.

PERSONAL PROJECTS

Built on my own time, alongside the client work above, to try new agent and LLM techniques end to end.

- **Multi-agent stock research system** (personal, 2026, LangGraph): five analyst agents (market, fundamentals, news, social, macro) run in parallel; bull and bear researchers debate and a manager agent judges; a trader agent's plan is challenged by three risk agents before a final call. ChromaDB vector memory of past cases; works with OpenAI, Anthropic or Gemini; a tracker scores every call against later prices. Research tool, not financial advice.
- **askrolog, consumer AI chat app for iOS** (personal, 2026): built for low LLM cost, with a cheap model by default and a stronger one only for long premium sections, prompt caching on shared context, and a per-section answer cache keyed by a hash of the input; streaming, server-side guardrails. Supabase Edge Functions, Claude via OpenRouter, Expo. Built and tested; not published.

PUBLICATIONS

7 peer-reviewed papers (2020–2023): recommendation systems (SIU 2020, SIU 2022), search reranking (ICMI 2022), Turkish text classification and adult-content image classification (INISTA 2020, two papers), message-product matching for advertising (SIU 2023) and epileptic seizure prediction (SIU 2021). Full list on Google Scholar.

EDUCATION AND LANGUAGES

- **MSc, Biomedical Engineering**, Boğaziçi University, Istanbul · 2019–2021 · GPA 3.82/4.00. Thesis: epileptic seizure prediction with machine learning and deep learning.
- **BSc, Computer Engineering** (GPA 3.78) **and Electrical and Electronics Engineering** (GPA 3.65), double major, Istanbul Şehir University · 2014–2019. All programmes taught in English.
- **Languages:** English (fluent, professional), Turkish (native).