

# Kayvan Mazaheri

✉ KayvanMazaheri@gmail.com | ☎ +98 913 8894689 | 📍 Tehran, Iran | 🔗 www.linkedin.com/in/kmazaheri

## Professional Summary

---

Senior backend engineer with 7+ years building high-throughput systems at Cafebazaar (51M users) and Divar (30M). Led the multi-year rebuild of Cafebazaar's ad platform — live monolith-to-microservices migration, 50K events/sec Kafka + Redis event pipeline, multi-tier caching that cut p99 latency from 700ms to 25ms — sustaining 10× ad revenue growth over two years. Currently shipping production AI (RAG, agent platforms, multi-tenant LLM systems) for B2B SaaS clients, with strong opinions about observability and what breaks first.

## Experience

---

### Independent

Remote

#### AI & Backend Engineer (Contract)

Apr 2025 – Present

- Built a multi-tenant RAG system for an AI legal-tech startup: ingested PDF, DOCX, and plain-text legal documents, designed the embeddings pipeline, and migrated the vector store from Pinecone to a self-hosted Elasticsearch cluster on GCP for cost and indexing control. Shipped a citation-anchored chatbot and a document-derived case-timeline feature.
- Production AI and agent platforms for B2B SaaS clients (NDA). Backend, architecture, and AI work focused on multi-tenancy, evals, and predictability under load.
- Practicing AI-assisted SDLC in real engineering work: Claude Code, agentic dev loops, evals; pressure-testing what holds up under production constraints.

#### Cafebazaar (The leading Android app store in the region with 51M+ users)

Tehran, Iran

##### Technical Team Lead

Mar 2022 – Apr 2025

- Led the ad-platform rebuild at 51M-user scale: live monolith-to-microservices migration with no downtime, service-mesh consolidation, doubled minimum ad pricing, and A/B testing infrastructure for both ad-placement variants and ML-model variants. Sustained 10× ad revenue growth over two years.
- Reduced p99 latency on the real-time ad-estimation hot path from 700ms to 25ms (same endpoint, more traffic) via multi-tier caching with async refresh, request deduplication, batch inference, and graceful fallback to heuristic estimates.
- Built the Kafka + Redis event pipeline at the core of the platform: 50K events/sec daily peak, idempotency keys with Redis-tracked deduplication, batched DB writes, at-least-once delivery with effective exactly-once semantics at the application layer.
- Designed the ML fallback path so a slow or degraded model couldn't kill ad serving — timeout-driven and degradation-driven fallbacks that salvaged revenue when the ML service was unreliable.
- Shipped a Video-on-Demand (VOD) platform end-to-end in one quarter. Owned the backend; led cross-functional delivery with iOS, Android, and web teams on streaming protocols and client-side caching.
- Raised team sprint output ~30% and cut average code-review cycle time ~20% by introducing automated CI/CD checks, clearer review standards, and tooling that removed friction from the merge path. Mentored 7+ junior engineers; 5 promoted internally during the role.

##### Senior Software Engineer

Oct 2020 – Mar 2022

- Delivered and operated 15+ backend APIs sustaining 99.9% uptime (SLO-driven) across 51M users on a distributed-systems architecture. Uptime measured per-request via Grafana dashboards over Prometheus data, across 6h/1d/1w/1mo windows.
- Split the legacy payment monolith into 3 independently deployable microservices. Deployment-related downtime dropped ~40% once releases stopped requiring the whole monolith to redeploy.
- Reworked the stats aggregation service's recovery path. Data loss caused by infrastructure instability went from ~15% to under 1% — more than a 10× reduction, with automatic recovery from outages instead of manual intervention.
- Launched an internal telemetry and distributed-tracing tool adopted by other engineering teams to improve observability and debugging — horizontal impact across multiple teams.

#### Divar (The largest classifieds marketplace in the Middle East, serving 30M+ users)

Tehran, Iran

##### Technical Team Lead

Feb 2020 – Oct 2020

- Built Divar's Real Estate vertical from a blank slate: backend services, domain modelling, and the first revenue-driving features. Grew into one of the marketplace's major revenue lines.
- Stepped into Tech Lead as the team grew. Established the engineering standards — code review, on-call, deployment, testing — that outlasted the role.

##### Software Engineer

Oct 2018 – Feb 2020

- Designed Divar's first company-wide feature-flag and A/B testing service. Presented the proposal to the CTO; gained design approval; built and shipped the service. Enabled controlled rollouts and A/B tests across millions of users; outlasted the role as standard tooling.
- Shipped 50+ APIs powering the Real Estate vertical at Divar. >90% test coverage was the bar for the vertical, not just a personal one.
- Partnered with product and cross-functional engineering teams to deliver revenue-driving features on top of those APIs.

#### Cafebazaar (The leading Android app store in the region with 51M+ users)

Tehran, Iran

##### Software Engineer

Jul 2018 – Oct 2018

- Built a rule-based fraud detection system on the P2P payment platform: velocity, amount-pattern, device-fingerprint, and account-age signals scoring transactions for auto-decision or moderator queue. Manual review burden dropped from 100% of transactions to ~30%; moderator workload down ~70%.
- Shipped 20+ back-office REST endpoints for the payment platform.
- Built a percentage-based bill-splitting and payment flow. Finance edge cases — rounding, partial settlements, retries — covered by a comprehensive test suite to ensure transaction correctness at scale.
- Partnered with Android and iOS teams on API contract design. Built an internal moderation dashboard for the fraud team to investigate and act on system output.

Skills

---

Programming Languages:

Backend & APIs:

AI in Production:

Cloud & Infrastructure:

Data & Messaging:

Observability:

Engineering Practices:

Leadership & Collaboration:

Python, Go, JavaScript, C++

Django, FastAPI, Node.js, gRPC, REST APIs, Async APIs

RAG, Embeddings, Vector Stores, Agent Design, LLM APIs, AI-assisted SDLC

Docker, Kubernetes, GCP, AWS, CI/CD, Git

PostgreSQL, Redis, Kafka, RabbitMQ, MongoDB

Grafana, Prometheus, Sentry, Distributed Tracing

Microservices, System Design at Scale, Concurrency, Unit & Integration Testing

Technical Leadership, Mentoring, Stakeholder Management, Agile/Scrum

Certifications

---

IELTS - Band Score 8

British Council

CEFR C1

Mar 2026

Education

---

Sharif University of Technology

MSc in Bioinformatics

Tehran, Iran

2018 - Dropout

Isfahan University of Technology

BS in Computer Software Engineering

Isfahan, Iran

2014 – 2018

Awards

---

Dec 2016

4th Place in 18th ACM-ICPC, Asia Region - Tehran Site

ACM

Dec 2015

5th Place in 17th ACM-ICPC, Asia Region - Tehran Site

ACM

Dec 2014

10th Place in 16th ACM-ICPC, Asia Region - Tehran Site

ACM

Feb 2013

23rd Place in Bayan International Programming Contest, Tehran

Bayan