

BABAK ZARRINBAL

Cloud & Platform Engineer — Kubernetes, GitOps, LLM Infrastructure

Düsseldorf, Germany · babak.zarrinbal@gmail.com · +49 157 3489 7539 · zarrinbal.org · github.com/babakzarrinbal · linkedin.com/in/babak-zarrinbal

SUMMARY

Platform/infrastructure engineer with 14+ years in software and 10+ years leading engineering teams, now focused on Kubernetes-native cloud platforms and LLM infrastructure. Currently building multi-cloud (Azure/GCP) Kubernetes platforms and agentic AI systems at Douglas. Track record spans self-hosted LLM serving and fine-tuning, RAG pipelines, and production-scale platform migrations. Active upstream contributor to kgateway (CNCF) — 2 merged PRs and a production-scale bug (3,500+ Envoy clusters) root-caused, load-tested, and fixed upstream.

OPEN SOURCE — KGATEWAY (CNCF ENVOY-BASED API/AI GATEWAY)

- PR #14231 (merged) — fixed XListenerSet status validation
- PR #13592 (merged) — exposed headersToClient in ext-authz responses
- Issue #14363 — root-caused a readiness race causing partial-xDS-config traffic on a 3,500+ cluster production gateway; validated fix with a 1M-request load test; fix merged upstream as PR #14380

EXPERIENCE

Douglas — Cloud Ops Engineer

Mar 2025 – Present

- Own multi-cloud IaC (Azure DevOps + Terraform, GCP + Terraform) for production and staging Kubernetes clusters
- Implemented kgateway from scratch and migrated all production routing onto it
- Built standalone agentic applications on Claude and Gemini (via Copilot), with RAG and memory management, under strict internal data-handling constraints — no sensitive data or processes exposed externally
- Built containerized, fully dynamic Azure self-hosted CI/CD agents, autoscaled with Karpenter and KEDA ScaledJob
- Automated secret rotation and containerized DevOps tooling across environments
- Integrated CrowdStrike Falcon across dev/staging clusters
- Manage a team of DevOps engineers

managbl.ai (acquired by casavi, Sep 2025) — Head of DevOps / Development Manager

Apr 2024 – Mar 2025

- Berlin PropTech automating phone accessibility for property managers via AI; technology now integrated into casavi's platform, serving 150+ customers
- Deployed and served Hugging Face models with vLLM; fine-tuned models for domain-specific intent classification on masked/anonymized user data, aligned with data protection requirements
- Built a RAG pipeline to retrieve contract terms, restrictions, and user/provider information for real-time cross-checking
- Migrated infrastructure to Kubernetes using spot instances to cut compute cost while scaling
- Built the speech pipeline (based on faster-whisper) feeding the above models
- Introduced socket-based inter-service communication for lower-latency streaming

Retraced — Team Lead / Technical Manager

Feb 2022 – Mar 2024

- German blockchain-based supply-chain traceability platform for the fashion industry (2,000+ brands and suppliers)
- Led architecture overhaul and rewrote core services into dockerized components
- Shipped an AI-assisted document management feature using OpenAI and Anthropic APIs
- Designed a Corrective Actions / Preventive Actions (CA/PA) feature on a dynamic, open-schema data layer over Oracle DB
- Owned service proposals, architecture, delivery, and technical stakeholder communication

Tecnotree — Technical Manager

Sep 2016 – Feb 2022

- Finnish telecom software vendor; delivered systems including DCBS (Digital Convergent Billing System), used by operators including Ooredoo
- Managed ~25 engineers across 5 rotating teams, a QC team, and a 4-person DevOps team running Rancher-based Kubernetes infrastructure
- Personally owned DCBS and the Catalogue Management System, among other services
- Mentored engineers from junior to senior; ran recruiting and performance processes

Earlier Experience

- Diliviri/Eposense — Team Lead & Senior Full-Stack Developer (2015–2016)
- Rahpooyan — Full-Stack Developer (2013–2015)
- Atlas-Offshore / Knauf Iran — Project Planner/Controller (2008–2013)

TECHNICAL SKILLS

- AI/LLM: vLLM, Hugging Face Transformers, fine-tuning, RAG, agentic apps, memory management, OpenAI, Anthropic (Claude), Gemini
- Platform: Kubernetes, ArgoCD (GitOps), Karpenter, kgateway/Envoy, cert-manager, Kyverno, KEDA, Rancher, Docker, Terraform
- Cloud: GCP, Azure, AWS (ECS, EC2, Lambda, EKS, S3, DynamoDB), Oracle Cloud
- Languages: JavaScript/TypeScript (Node.js, NestJS, React, Vue), Python (Django/Flask), Go, Java (Spring Boot), Bash
- Data: Oracle DB, MongoDB, Cassandra, PostgreSQL, MySQL, Redis
- Messaging: Kafka, RabbitMQ, SNS/SQS, WebSockets
- Security/Observability: CrowdStrike Falcon, secret rotation/management, Datadog, Grafana

EDUCATION

B.A. Industrial Engineering — University of Science and Culture, Tehran

LANGUAGES

English (fluent) · Persian (native) · German (conversational) · Arabic (basic)