

Krešimir Kralj

Zürich, Switzerland

Email: mail@kralj.kr ▪ LinkedIn: [kralj.kr/In](https://www.linkedin.com/in/kralj.kr/)
GitHub: [kralj.kr/gh](https://github.com/kralj.kr/gh) ▪ Website: kralj.kr

EXPERIENCE

Google

Software Engineer – YouTube Shopping (Viewer Understanding) Mar 2026 – Present
C++, Python, GoogleSQL, Distributed Systems Zürich, Switzerland

- **Cross-Surface Signal Streaming:** Architected real-time streaming ingestion pipelines to bridge high-intent Google Search signals into YouTube's distributed key-value storage, enabling sub-300ms preference lookups in the primary recommendation serving stack.
- **Recommendation Ranking & Feature Engineering:** Engineered real-time query and entity-level feature pipelines for YouTube Shorts ranking models, improving the predictive accuracy of offer click models (pOfferClick) to drive Valued Watch Time and conversion metrics.

Software Engineer – Cloud Anti-Money Laundering AI Nov 2024 – Feb 2026
Python, Go, GCP, Distributed Systems Zürich, Switzerland

- **ML Platform Governance:** Led reliability and quality evaluations for enterprise ML risk-scoring pipelines ([Google Cloud AML AI](#)); designed model validation protocols across multi-version upgrades, guaranteeing zero quality regression on mission-critical risk models.
- **0-to-1 Distributed Data Platform:** Architected an end-to-end event-driven usage intelligence engine from scratch (Protobuf schemas, stream ingestion, data warehousing), replacing ad-hoc distributed log forensics with real-time analytics to steer deprecation strategies.

Software Engineer – Android AICore (On-Device ML Platform) Sep 2022 – Nov 2024
Java, Kotlin, C++, Android IPC/AIDL, ML Infra London, United Kingdom

- **On-Device Model Delivery Platform:** Architected dynamic over-the-air (OTA) model delivery infrastructure for [Android AICore](#), enabling secure, sandboxed distribution of GB-scale foundation models (*Gemini Nano*) and LoRA adapters across millions of active devices (awarded **2024 Google Tech Impact Award**).
- **OS-Level Multi-Tenant IPC Architecture:** Designed cross-profile multi-user service architecture (system-level User0 daemon and AIDL proxying), eliminating multi-gigabyte RAM/storage duplication across enterprise work profiles with zero latency overhead.

Twitter

Software Engineer – Cortex (Media Understanding) Sep 2020 – Aug 2022
Scala, Python, Distributed Systems, Finagle London, United Kingdom

- **High-Throughput Distributed Processing:** Architected distributed stream-processing services in Scala to ingest, decode, and extract multimodal ML features from all media uploads at **8k+ peak QPS** under strict sub-second latency SLAs.
- **Cross-Platform API Infrastructure:** Standardized core media understanding APIs and asynchronous event contracts, serving as the central media intelligence backbone across Trust & Safety, Home Timeline recommendations, and Search relevance.

Axilis

Software Engineer – Distributed Systems Feb 2019 – Aug 2020
Go, Apache Kafka, Distributed Systems, Docker Zagreb, Croatia

- **Distributed Event Streaming:** Architected a high-throughput audit and transaction pipeline in Go and Apache Kafka, processing tens of thousands of real-time events per second with high reliability and zero message loss.
- **System Resiliency & Traffic Shaping:** Engineered dynamic backpressure and adaptive throttling in the ingestion layer, reducing processing latency by **60%** and eliminating downstream service degradation during extreme peak traffic surges.

Amazon

Software Engineering Intern – Alexa NLU Jul 2018 – Sep 2018
Python, Java, Machine Learning Cambridge, United Kingdom

- **NLU Prototyping & Benchmarking:** Partnered with Research Scientists to prototype and benchmark an experimental entity-resolution service for ambiguous voice queries, evaluating accuracy trade-offs and presenting architectural specifications to engineering leadership.

EDUCATION

University of Zagreb, Faculty of Electrical Engineering and Computing Zagreb, Croatia
M.Sc. in Computer Science (Thesis: [Video Action Recognition using Spatio-Temporal Residual Networks](#)) Oct 2017 – Jul 2019
B.Sc. in Computer Science (Thesis: [Image Classification using Deep Convolutional Neural Networks](#)) Oct 2014 – Jul 2017

TECHNICAL SKILLS

Languages:	Go, Python, Scala, Java, C++, SQL, Protocol Buffers
Distributed Systems:	Apache Kafka, gRPC, Distributed Event Streaming, Microservices, Concurrency, High-Throughput Ingestion
ML Systems & Edge AI:	On-Device Model Serving (Gemini Nano, LoRA), ML Pipeline Governance, Evaluation Protocols, Android IPC/AIDL
Cloud & Infrastructure:	Google Cloud Platform (GCP), Multi-Tenant Cloud Architecture, Docker, Kubernetes, IAM & Zero-Trust Security, Linux