Kaivalya Swami

Software Engineer – Agent Platforms & Search Systems

+91 9521280121
kaivu1999@gmail.com
kaivu1999.github.io

Education

2017–2021 **B.Tech, Computer Science & Engineering**, *Indian Institute of Technology Delhi*, *GPA – 8.6/10*.

Thesis Low-Resource Neural Machine Translation for 12 Indic Languages (Prof. Parag Singla). Fine-tuned mBART and created a standardised multilingual translation benchmark.

Ranks JEE Advanced 2017 – AIR **561** / 1.5M applicants | KVPY Scholar – AIR 571 (IISc Bangalore)

Olympiads National Astronomy Olympiad – Top 200, selected for INAO | Physics Olympiad – Top 1% in state

Experience

2021–Present **Software Engineer, AI Platform**, *DevRev*, Bengaluru.

Employee #40 – joined in stealth; now a unicorn with 800+ employees.
B2B SaaS platform with AI-native CRM, support agents, and enterprise search.
Core member of the small AI/Search team owning retrieval, ranking, and agent infrastructure.

Agent Skill Design & Tool Contracts

- **Own *search* as the primary retrieval skill** invoked by all customer-facing AI agents. (most-called tool)
- **Redesigned the skill contract** (query, snippets, result packaging) – tool calls per query **6.5** → **3.8** (**43% fewer skill calls**) with no task-success regression.
- **Compass** – semantic clustering of real agent traffic (inspired by Anthropic's *CLIO*); surfaces query patterns + failure modes that inform skill-contract tuning.

Agent Config Versioning & Knowledge Scoping

Knowledge-scoping framework controlling which knowledge bases, tools, and retrieval strategies each customer-facing agent accesses; versioned, auditable, per-customer configs.

Evals & A/B Testing

- **A/B testing framework** – built for reranker rollouts, generalised company-wide; tests ranking strategies, models, retrieval params, and skill contracts on live traffic via the ML config system.
- **Regression detection by design** – rollouts monitored against CTR, MRR, agent-feedback signals; quality drops caught before full ramp.

Reranking & Retrieval Pipeline

- **Two-stage ranking** powering all AI agents: RRF over BM25 + semantic retrieval, then cross-encoder reranker.
- **Reranker model selection** – evaluated Qwen-3 variants; picked ZeRank-1-Small for best latency/quality (**+15% Recall@5** across object types).
- **Versioned ML configs** enable staged rollouts, per-customer overrides, and safe model swaps.

Search Observability & Analytics

End-to-end instrumentation (CDC → Kafka → BigQuery → LookerStudio): Search CTR, MRR, agent-feedback dashboards across all customer orgs; primary data source for PMs, CX, Sales.

Also Owned ML inference service (Python/Go): async rewrite, 50%+ memory-leak reduction, all CVEs resolved. Shipped cross-lingual search to GA. Fixed distributed trace propagation across 3 microservices via gRPC header forwarding.

May–Jul 2020 **Data Engineering Intern**, *APT Portfolio*.

End-to-end alerting pipeline: Kafka message routing, C++/Python producer wrappers, MongoDB, Flask dashboard.

May–Jul 2019 **Research Intern**, ***Nanyang Technological University, Singapore***, Prof. Weichen Liu.

Neural network quantization – benchmarked Binarized NN (BMXNet) and XNOR-Net speedup on AlexNet & VggNet.

Dec 2019–2020 **Core Member**, *Tensor Dynamics (IIT Delhi Startup)*.

Solar irradiance forecasting 3 hrs ahead using optical-flow cloud tracking on satellite imagery.

Publication

- K. Sharma, **K. Swami**, A. Shete, S. Husain, "Can Greenbergian universals be induced from language networks?", *SyntaxFest 2019*.

Teaching & Leadership

2020–21 **Teaching Assistant, DSA**, *IIT Delhi*, Prof. Rahul Garg.
Lab sessions for 20 students; assignment prep and exam grading.

2018–19 **Executive, Dev Club**, *IIT Delhi*.
Built intranet file-sharing service; lectured on Bash & Git to 100+ students.

Technical Skills

Languages	Python, Go, C++, SQL, Bash
ML & AI	PyTorch, LangChain, LlamaIndex, LangGraph, Hugging Face
Infrastructure	Kafka, BigQuery, MongoDB, gRPC, Docker, Kubernetes, Datadog, Langfuse, AWS, Lookerstudio
AI-Augmented	Cursor + Claude daily driver. Will learn your entire codebase before the onboarding call ends :)

Interests

Juggling	**Represented India at the European Juggling Convention (EJC) 2024** – the sole attendee from India.
House Dance	House dancer, have trained for 3 years. Love to learn all street style dance forms.
Racquet Sports	**Trained in table tennis, tennis, and badminton**; avid player of all racquet sports.
Nature	Avid admirer of nature with a curious mind. Love to talk about the cosmos and learn more and more.