

# Arbaz Siddiqui

Principal Software Engineer · AI Infrastructure and Distributed Systems

arbaz00@gmail.com · +91 8963874828 · [arbazsiddiqui.me](https://arbazsiddiqui.me) · [github.com/arbazsiddiqui](https://github.com/arbazsiddiqui) · [hf.co/arbazsiddiqui](https://hf.co/arbazsiddiqui) · [linkedin.com/in/arbazsiddiqui](https://linkedin.com/in/arbazsiddiqui)

## SUMMARY

---

Backend engineer with ten years of experience building consumer products at scale. At Turnip I built and ran the live-streaming and social platform serving **25M users** through four pivots (game streaming, live audio, a creator social network, and now an AI design suite), and spent the last two years on the AI infrastructure around it. Outside work I fine-tune open-weights models, build mobile apps, and use generative AI to run autonomous pipelines, including an open 12B model that trended #1 on Hugging Face and content pipelines at **600K monthly views**.

## EXPERIENCE

---

**Turnip** · *Principal Software Engineer*

Jan 2021 – Present

*Four pivots in one startup, from game streaming to live audio to a creator social network to Zaps, an AI design suite.*

- Built and ran the backend serving **25M users**, 20 microservices on Kubernetes exposed through Kong and talking internally over Kafka, with MongoDB as the primary store, Neo4j, Redis and Elasticsearch by use case, and real-time systems over MQTT. Sole backend engineer for the first two years.
- Wrote the live video streaming stack, **50K peak concurrent streams** and about **100TB** of video a day on RTMP and FFmpeg at **sub-2s latency**, holding sticky sessions over TCP through HAProxy ingress and packing streams onto shared Kubernetes pods for efficiency and cost.
- Implemented live audio rooms on Agora with MQTT holding presence and state for **2M monthly users** over **70M call minutes**.
- Designed the social network's feed and graph, weighting every interaction by a user and their friends into a per-user ranking cached and served at **sub-100ms**, with the graph on **Neo4j** powering second-degree friend discovery and suggestions.
- Shipped the AI editing suite of face swap, image enhancement and styles on **ComfyUI** with open-weights models, served on in-house GPU inference and then moved to a distributed GPU network, cutting GPU cost **80%** at about **2 seconds** an image and around **100K HD generations a day**.
- Replaced template tag search with semantic retrieval over **40K** embedded templates and carousels, served by vector search at **p95 under 50ms**.
- Set up an in-house AI agent harness for internal tooling where each project carries its own memory, TDD requirements, deployment skills and guardrails, a black box to the product team. **Ten non-engineers shipped 17 internal tools** with zero engineer commits.
- Migrated core services from Node to Go behind a strangler fig, with byte-parity responses under an enforced **100% test-coverage** gate.

**MouseBelt** · *Senior Backend Engineer (Contract)*

Apr 2020 – Jan 2021

- Built a centralized cryptocurrency exchange end to end, accounts, KYC, markets and order flow on a **matching engine written in C** with gRPC and WebSocket layers.
- Ran custody across **six chains** (Bitcoin, Ethereum, EOS, Cosmos, ICON, BNB Chain) and built a multi-chain **DEX aggregator** with price scrapers across BSC, ICON and Moonbeam behind one interface and an on-chain indexer.

India's biggest event ticketing platform.

- Led **seven engineers** across backend and mobile on the movie ticketing platform holding **500+ requests a second** on sale nights, where integrating Vista, the slow and error-prone legacy theatre software, meant redesigning the whole checkout from synchronous to asynchronous.
- Designed the logistics architecture, from internal tooling for the operations team to user-facing order tracking and delivery communication, integrated with logistics partners.
- Built the commerce flow end to end across cart, payments, inventory and taxation, **30 microservices** on EKS with SQS for async communication, Lambda and Kinesis.
- Moved on-demand workloads to serverless on Lambda and Kinesis, cutting cost by over **70%**.

## SELECTED PROJECTS

---

- **Ozan-v1-12B**. Fine-tuned and released a 12B open-weights long-form fiction writing model with QLoRA on one A100 for about \$40. Lowest slop of any 12B on EQ-Bench, beating bigger SOTA models like Gemma 3 27B. **#1 trending on Hugging Face**, 3K+ monthly downloads. [hf.co/arbazziddiqui](https://huggingface.co/arbazziddiqui)
- **The Footnote**. An autonomous pipeline that researches, scripts, voices, illustrates, renders and publishes at **600K monthly views**, built entirely on open-weights models and without any human supervision. [@thefootnotemedia](https://twitter.com/thefootnotemedia)
- **Iris**. A privacy-first cycle tracker, **4.9★** on the App Store in **twelve languages**, with Apple Health sync and partner sharing. Organic installs come from autonomous women's-health content pipelines on Instagram and Pinterest. [App Store](#) · [Pinterest](#)
- **Mercuro**. A logic puzzle game with its own constraint solver and generator, **660 curated puzzles** on iOS and Android, featured in gaming newsletters. [App Store](#)
- **Iru-cache-node**. An O(1) LRU cache with expiry for Node.js, **119★**, featured in Node Weekly. [github](#)
- **anabran**. An HTTP load balancer and reverse proxy written in Go. [github](#)

## SKILLS

---

|                       |                                                                                                                                                                                                                                                                     |
|-----------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Languages</b>      | Go, TypeScript, JavaScript, Python                                                                                                                                                                                                                                  |
| <b>AI</b>             | RAG, agent orchestration (LangGraph, MCP, Claude Code), fine-tuning (QLoRA), model serving, GPU inference, evals, PyTorch, ONNX, Triton, MLflow, generative image pipelines (ComfyUI), diffusion models (Qwen, MiniMax, Flux), workflow automation (n8n), llama.cpp |
| <b>Backend</b>        | Node.js, Kafka, REST, gRPC, WebSockets, microservices, event-driven architecture, RabbitMQ, SQS, Kinesis, MQTT, FastAPI, RTMP, FFmpeg                                                                                                                               |
| <b>Data</b>           | MongoDB, PostgreSQL, Redis, Elasticsearch, Neo4j, Cassandra, Firebase RTDB                                                                                                                                                                                          |
| <b>Infrastructure</b> | Kubernetes (GKE, EKS), Docker, AWS, GCP, Terraform, Helm, Kong, Lambda, CI/CD                                                                                                                                                                                       |
| <b>Observability</b>  | Prometheus, Grafana, ELK, OpenTelemetry, New Relic, Sentry                                                                                                                                                                                                          |
| <b>Crypto</b>         | Order matching engines, Multi-chain custody, DEX aggregation, NFT systems                                                                                                                                                                                           |

## EDUCATION

---

**B.Tech, Computer Science and Engineering** · The LNM Institute of Information Technology    2013 – 2017