

Arbaz Siddiqui

Principal Software Engineer · AI Infrastructure and Distributed Systems

arbaz00@gmail.com · arbazsiddiqui.me · github.com/arbazsiddiqui · hf.co/arbazsiddiqui · linkedin.com/in/arbazsiddiqui

SUMMARY

Backend engineer, nine years, mostly consumer products at scale. At Turnip I built and ran the platform behind 25M users, then spent two years on the AI infrastructure around it, including an 80% cut in GPU spend with generation still near two seconds. Outside that I ship my own work, an open 12B language model, various apps, and content pipelines at 500K+ monthly views.

EXPERIENCE

Principal Software Engineer Turnip

Jan 2021 – Present

Four pivots in one startup. Game streaming, live audio, a creator social network, now Zaps, an AI design suite.

- Built and ran the backend behind 25M users, and owned the cloud spend. Twenty-one services on GKE, fronted by Kong and autoscaled under load, with Kafka and Redis carrying traffic and the core since rewritten in Go.
- Built the streaming stack on RTMP and FFmpeg, 50K streams a day under five seconds of latency, then live audio rooms on Agora with MQTT and Firebase holding state for 2M monthly users across 70M call minutes, then a social feed with second-degree discovery over Neo4j, indexed live by Elasticsearch on Kafka.
- Shipped AI photo editing on ComfyUI, orchestrated with LangGraph then n8n. Moving generation to a distributed GPU network cut GPU spend 80% at roughly two seconds an image; autoscaling and serverless took 70% off infrastructure.
- Replaced template tag search with semantic search over embeddings, retrieved by vector similarity with LangChain.

Senior Backend Engineer MouseBelt

Apr 2020 – Jan 2021

- Built the exchange backend. Accounts, KYC, markets, order flow, fee schedules, referral commissions, API keys with IP allowlists, failover. Also the low-latency core, a matching engine in C with HTTP and WebSocket layers, price feeds and trade history.
- Multi-chain custody across Bitcoin, Ethereum, EOS, Cosmos, ICON and BNB Chain, plus a DEX aggregator and on-chain indexer.

Lead Backend Engineer Paytm Insider

2017 – Apr 2020

- Led a team of seven on the movie ticketing platform, which held 500+ requests a second on sale nights, and designed the logistics architecture behind delivery opt-in, order tracking and shipment tooling.
- Built the commerce flow of cart, payments, inventory and taxation, run as AWS microservices on Kubernetes with Lambda and Kinesis carrying the on-demand work.

SELECTED PROJECTS · ALL SHIPPED

Ozan-v1-12B · hf.co/arbazsiddiqui. Trained and released with QLoRA SFT and anti-slop DPO on one A100 for about \$40. Slop 5.30 on EQ-Bench Creative Writing v3, lowest of any 12B and below Gemma 3 27B at 7.03, 2.3K downloads in month one.

The Footnote · @thefootnotemedia. Researches, scripts, voices, illustrates, renders and publishes twice a day with nobody in the loop, at 500K monthly views. The same pipeline runs a horror channel and Iris's marketing.

Open source · github.com/arbazsiddiqui. A decade of public code, 200+ stars across it. lru-cache-node, an O(1) LRU cache with expiry, featured in Node Weekly and still on npm, and anabranh, an HTTP load balancer and reverse proxy in Go.

Iris and Mercuro · [App Store](https://AppStore). A privacy-first cycle tracker at 4.9 stars in twelve languages, and a logic puzzle game with its own constraint solver and generator, on iOS and Android.

SKILLS

AI: fine-tuning, RAG (LangChain, embeddings, pgvector), diffusion (ComfyUI, Flux, Qwen, Wan 2.2), agent orchestration (LangGraph, MCP servers, Claude Agent SDK), workflow automation (n8n), llama.cpp, distributed GPU inference

Backend: Go, Node.js and TypeScript, Python · Kafka, Redis, MQTT, WebSockets, RTMP and FFmpeg, Postgres, MongoDB, Elasticsearch, Neo4j, Cassandra · event-driven and hexagonal architecture

Infrastructure: Kubernetes (GKE), Kong, Docker, GCP, AWS, serverless (Lambda, Kinesis, SQS), CI/CD, Prometheus, ELK

Crypto: order matching engines, multi-chain custody, DEX aggregation, NFT systems