Ayush Kumar

Software Engineer | AI Platform and Backend Systems
Hyderabad, India | +91 81025 06748 | ayush.krsh9@gmail.com
linkedin.com/in/ayusshkumr | github.com/TheArchitect19

SUMMARY

Founding engineer who builds and ships AI and data platforms end to end. Currently leading a 9-person team at Quper, a multi-cloud FinOps platform in production at The Chatterjee Group and ABP. Previously built voice AI agents over SIP telephony, AI workflow orchestration, and human-in-the-loop systems at Steps AI, alongside large-scale data pipelines on PostgreSQL and Redshift. Works directly with clients from requirements through SaaS and BYOC deployment.

EXPERIENCE

Lead Founding Engineer / Forward Deployed Engineer, Quper *(India)* **Jun 2026 – Present**

- Lead a 9-person engineering team that took Quper, a FinOps platform, from architecture to a commercially sold product.
- Built three product lines: FinOps (cloud cost), DataFinOps (data-platform cost), and Vidura, an AI module.
- Built data pipelines that collect cost and usage data from AWS, GCP, Azure, Databricks, and Amazon Redshift, orchestrated with Temporal and deployed on AWS ECS Fargate.
- Delivered both SaaS and bring-your-own-cloud (BYOC) deployment models across Kubernetes, Docker Swarm, and AWS ECS, with Amazon ECR and OIDC-based authentication.
- Own client communication and requirements as forward deployed engineer; led production deployments at The Chatterjee Group (multi-billion-dollar conglomerate operating in 45 countries) and ABP (Indian digital media).

AI Engineer / Platform Engineer, Steps AI *(Hyderabad, India)* **Oct 2025 – Jun 2026**

- Built a DAG-based AI workflow engine using Kahn's topological sort, with per-step error policies (fail, skip, retry), a dead-letter queue for failed runs, and 139+ integrations across 24 providers.
- Built a voice AI agent platform on LiveKit that handles real PSTN calls end to end, using SIP trunking from Exotel into LiveKit inbound and outbound trunks, routing each call to the right tenant's agent by dialed number (DID).
- Architected a two-plane voice design separating call transport, session state, and billing from the LLM reasoning service, with Redis-cached runtime config and idempotent call finalization plus LiveKit webhook reconciliation for crash safety.
- Built Temporal-driven outbound calling (cart recovery, COD confirmation) with DNC enforcement, Sarvam speech-to-text and text-to-speech with automatic language detection, and per-call credit billing.
- Designed a human-in-the-loop interrupt system: agents pause on low confidence or user escalation, requests queue on AWS SQS, support agents join via WebSocket, and end users get live status over Server-Sent Events.

Software Engineer, Cricut **Nov 2025 – Apr 2026**

6-month contract via Xequalto Analytics | South Jordan, UT, USA (Remote)

- Built a Redshift monitoring tool with Apache Airflow and AWS Glue tracking cluster health and query performance.
- Created an AI-powered Redshift performance advisor that analyzes long-running queries and recommends optimizations.
- Deployed and managed services on Kubernetes and AWS ECS Fargate.
- Collaborated remotely with Cricut's US-based data engineering team across time zones.

Software Development Engineer, Xequalto Analytics *(Remote, India)* **Aug 2024 – Oct 2025**

- Designed a fault-tolerant ETL pipeline ingesting 1B+ rows into PostgreSQL with batched transformation and loading.
- Architected hash-based horizontal sharding in PostgreSQL for writes, replicated to ClickHouse for OLAP reads, with fan-out queries aggregated at the application layer.
- Improved serialization performance by 60% by adopting Protocol Buffers for service-to-service contracts, and built GraphQL APIs for client-driven querying across internal services.
- Delivered zero-downtime deployments using Kubernetes, ArgoCD, OIDC, and a CI/CD pipeline with Amazon ECR.

PROJECTS

Small Language Model: LinkedIn Post Generator *PyTorch, Metal (MPS)*

- Designed and trained a 300M-parameter decoder-only transformer from scratch on Apple Silicon (M5, Metal/MPS), with no pre-trained base or Hugging Face model weights.
- Implemented a custom BPE tokenizer from scratch, including corpus-frequency analysis, iterative merge rules, and special-token handling.
- Architected a custom hybrid transformer: decoder-only autoregressive backbone with RoPE positional embeddings for length generalization, multi-head causal self-attention with KV-cache support, SwiGLU activations in the feed-forward network, and pre-RMSNorm for training stability.
- Engineered the full training loop on MPS with bfloat16 mixed precision, gradient clipping, and a cosine learning-rate scheduler with warmup, working around MPS-specific kernel limitations absent in CUDA environments.

TECHNICAL SKILLS

- **Languages:** Python, TypeScript, JavaScript, Go, C++, Java, C, SQL
- **Backend:** Node.js, NestJS, FastAPI, Django, Go Fiber, GraphQL, REST, Protocol Buffers, WebSocket, SSE, Temporal
- **Voice and Telephony:** LiveKit, SIP trunking, PSTN, Exotel, speech-to-text, text-to-speech, voice AI agents
- **Cloud and DevOps:** AWS (Redshift, Glue, ECS, Fargate, SQS, ECR, RDS, DynamoDB), GCP, Azure, Databricks, Kubernetes, Docker, Docker Swarm, ArgoCD, CI/CD, OIDC, Linux
- **Data:** PostgreSQL, Amazon Redshift, ClickHouse, Apache Airflow, AWS Glue, MongoDB, DynamoDB, Redis, ETL pipelines, database sharding, FinOps
- **AI/ML:** PyTorch, Transformers, LLM training, tokenization (BPE), AI agents, workflow orchestration
- **Frontend and Tools:** React, Next.js, Git, GitHub, Postman, Swagger

ACHIEVEMENTS AND OPEN SOURCE

- **ICPC Regionalist 2023** (National Rank 73); **LeetCode** Top 7% globally.
- **AskSenior** (Software Development Engineer): scaled the backend to support 10,000+ users.
- **Keploy** (open source): refactored services to NestJS and built repository analytics.
- **CloudCV (EvalAI)** (open source): wrote 100+ tests, upgraded PostgreSQL, and improved API documentation.

EDUCATION

Bachelor of Engineering, Computer Science, Chandigarh University *(Punjab, India)* **Jun 2021 – May 2025**