Umar Ahmed Thameem Ahmed

umar.ahmed.t.30@gmail.com | LinkedIn: umarahmed2000 | GitHub: UmarAhmed30

EDUCATION

University of Colorado Boulder — **Boulder, CO, United States**
*Master of Science in Computer Science – **CGPA: 4.0/4.0*** — *May 2027*
Related Coursework: Systems for Machine Learning, Datacenter Scale Computing

College of Engineering Guindy, Anna University — **Chennai, TN, India**
*Bachelor of Engineering in Computer Science and Engineering – **CGPA: 8.75/10*** — *Apr 2022*
Related Coursework: Data Structures & Algorithms, Database Management Systems, Operating Systems, Cloud Computing, Computer Networks, Machine Learning, Big Data Analytics, Data Mining, Software Engineering

SKILLS

Programming: Python, Ruby, Java, Go, JavaScript, TypeScript, SQL, NoSQL, Bash
Frameworks & Libraries: React, Angular, Flask, FastAPI, PyTorch, LangChain, LangGraph, vLLM
Databases & Caching: MySQL, MongoDB, PostgreSQL, Oracle DB, Qdrant, Redis, Memcached
Cloud, DevOps & OS: Docker, Kubernetes, AWS, GCP, Git, GitHub, CI/CD, ElasticSearch, Nginx
GenAI & Observability: OpenAI API, Gemini API, Anthropic Claude, Langfuse, Prometheus, Grafana

EXPERIENCE

Nueromind Technologies — **Bangalore, KA, India**
Software Development Engineer | Python, FastAPI, ReactJS, MongoDB, Docker, AWS, ELK — *Jan 2024 – Jul 2025*

- Led a team of 4 and owned the core **backend** and **AWS cloud infrastructure** for the **flagship, revenue-driving** AI-powered video conferencing platform. Built a video ingestion pipeline (**WebRTC**, **RTSP**, **Socket.IO**, and **HLS)** and a **Celery**-based **distributed model inference** pipeline delivering **30 FPS** at production load.
- Engineered a **model inference pipeline** for canary-deploying **multi-modal** models (CLIP, ASR, UGround, Wan 2.1) using Modal and Nomad. Integrated **Prometheus** monitoring, reducing inference latency by **40%** for **~2K** monthly users.
- **Fine-tuned** and **quantized** an in-house LLM for social media content creation, improving task accuracy by **8%** and reducing inference latency by **35%** through targeted model-level optimizations.
- Deployed an autonomous browser-based screen-sharing validator using **GPT-4V** and **Anthropic's Computer Use**, achieving **99.99% stream uptime** with automatic detection of black frames/FPS drops and auto recovery.
- Built **NL2SQL** microservices using an instruction-tuned **LLM** to generate **PostgreSQL** queries with embedded business logic, cutting query build time by **30%**. Integrated **Langfuse** for traces/accuracy metrics and **guardrails**.
- Developed an async **S3** prefetching video processor module using **FFmpeg** (**I/O-tuned**, **multi-threaded**) to parallelize chunking/transcodes, reducing end-to-end video processing latency by **40%**.

Infibeam Avenues — **Bangalore, KA, India**
Software Development Engineer | Ruby on Rails, AngularJS, MySQL, Docker, Kubernetes — *Jul 2022 – Jan 2024*

- Engineered a scalable **seller notification system** on **Firebase**, enabling bulk notifications via Excel imports for JioMart (**1M+** active sellers), reducing manual messaging efforts by **40 hours per month** across support teams.
- Developed a suite of **REST APIs** for a **seller ticketing system**, coordinating and integrating with a third–party service, to allow sellers to report and manage issues more efficiently, improving issue resolution time by **25%**.
- Built a stock-approval work-allocation module with a **round-robin** strategy to evenly distribute seller requests across admins, cutting median approval time by **18%** and reducing per-admin queue variance by **60%**.

PROJECTS

HyRA - Hybrid Routing Agent

- Building a cost-aware **LLM routing framework** using **vLLM, LangGraph** and **Flask** on **GCP**, enabling dynamic selection among open-source models and projecting **30 - 40%** reductions in latency and cost over a static baseline.
- Integrated an **LLM-as-a-Judge** feedback loop with a **PostgreSQL**-based Model Registry, improving route-selection accuracy from **71%** to **89%** while keeping overhead **< 1 second** per query.
- Containerized and deployed **vLLM** servers with **Langfuse** profiling, guiding routing weight calibration.

Distributed Log Processing System

- Engineered a **real-time distributed log analytics pipeline** utilizing **Apache Kafka**, **Apache Flink**, **ClickHouse DB**, and **Redis**, enabling **10x** higher log throughput (**~1M** events/min) with **99.9**% message durability.
- Developed a **full stack observability dashboard** (**Next.js** + **FastAPI** + **Redis** + **ClickHouse**) to deliver **-100 ms** query responses by **80%** via top-K caching and optimized analytical queries.

ACHIEVEMENTS & CO-CURRICULAR ACTIVITIES

- **Top 10 (7th Place) Finish** among 50 teams at the **AWS GameDay** Hackathon, leveraging **Amazon Nova Pro/Lite**, **Titan Text Embeddings V2**, **Lambda**, and **Step Functions** within a pre-deployed GenAI pipeline. Optimized retrieval and orchestration logic using **Transcribe** and **OpenSearch**, with guardrails implemented for prompt safety.
- Served as the **Web Developer** for the **Computer Science and Engineering Association** and **CEG Tech Forum**, spearheading the development of high-traffic websites for **technical symposiums**, serving **50K+** users globally and automating registration and event workflows.