Umar Ahmed Thameem Ahmed

Email: umar.ahmed.t.30@gmail.com | LinkedIn: umarahmed2000 | GitHub: UmarAhmed30 | Portfolio: umarahmed.codes

EDUCATION

University of Colorado Boulder — **Boulder, CO, United States**
Master of Science in Computer Science - ***CGPA: 4.0/4.0*** — *May 2027*
Related Coursework: Distributed Systems, Systems for ML, Datacenter Scale Computing, Big Data Architecture, Computer Vision

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, Object-Oriented Programming, Cloud Computing, Computer Networks, Machine Learning, Data Mining, Software Engineering

SKILLS

Programming: Python, Ruby, Java, JavaScript, TypeScript, SQL
Frameworks & Libraries: React, Angular, AngularJS, Ruby on Rails, Flask, FastAPI
Databases & Caching: MySQL, MongoDB, PostgreSQL, Oracle DB, Qdrant, Redis, Memcached
Cloud & DevOps: Docker, Kubernetes, AWS, GCP, Git, GitHub, CI/CD, Elasticsearch, Nginx
GenAI & Observability: OpenAI API, Gemini API, Anthropic API, vLLM, Langfuse, Prometheus, Kibana, Grafana

EXPERIENCE

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

- Led a **cross-functional** team of 7 and owned the end-to-end **SDLC** for the **production backend** and **cloud infrastructure** of a revenue-critical AI video-conferencing platform using **FastAPI**, **Docker** and **AWS**, while ensuring high availability.
- Shipped an async video ingestion pipeline and a **Celery**-based **distributed model inference** system delivering **30 FPS**.
- Built a canary-deployed multimodal inference pipeline supporting **CLIP**, **ASR**, **UGround**, **Wan 2.1**, using Modal and Nomad, with Prometheus-based observability supporting **~2K MAU**.
- 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*

- Built services in an Agile environment, powering a scalable **seller notification system** on **Firebase** used by **1M+** active sellers, reducing manual efforts by **40 hours/month** across support teams.
- Developed a suite of **REST APIs** in **Ruby on Rails** for a **seller ticketing system**, integrating with a third-party service, which reduced 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 and per-admin queue variance.

PROJECTS

HyRA - Hybrid Routing Agent

- Built a cost-aware, **distributed LLM routing service** using **vLLM**, **LangGraph**, and **Flask** on **GCP**, dynamically routing requests across models to optimize latency, accuracy, and cost, achieving **~30% efficiency gains** over a static baseline.
- Designed a **PostgreSQL-backed model registry** and integrated an **LLM-as-a-Judge** feedback loop to evaluate routing decisions, improving accuracy from **71%** to **89%**.
- Deployed **vLLM**-based model serving infrastructure with production-grade monitoring and profiling using **Langfuse**.

Distributed Log Processing System

- Engineered an end-to-end **real-time distributed log analytics pipeline** utilizing **Apache Kafka**, **Apache Flink**, **ClickHouse**, 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**), reducing average query latency to **<100 ms** via top-K caching and optimized analytical queries.

ACHIEVEMENTS & CO-CURRICULAR ACTIVITIES

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