

# Umar Ahmed Thameem Ahmed

[umar.ahmed.t.30@gmail.com](mailto:umar.ahmed.t.30@gmail.com) | (720) 232-0077 | [linkedin.com/in/umarahmed2000](https://www.linkedin.com/in/umarahmed2000) | [github.com/UmarAhmed30](https://github.com/UmarAhmed30) | [umarahmed.codes](https://umarahmed.codes)

## EDUCATION

### University of Colorado Boulder

Boulder, CO, United States

Master of Science in Computer Science (CGPA: 4.0/4.0)

Aug 2025 – May 2027

Related Coursework: Distributed Systems, Systems for ML, Datacenter Scale Computing, Big Data Architecture, Computer Vision

### College of Engineering Guindy, Anna University

Chennai, India

Bachelor of Engineering in Computer Science and Engineering (CGPA: 8.75/10)

Aug 2018 – Apr 2022

Related Coursework: Data Structures, Algorithms, Database Systems, Operating Systems, Object-Oriented Programming, Computer Networks, Cloud Computing, Machine Learning, Big Data Analytics, Data Mining, Software Engineering

## EXPERIENCE

### Nueromind Technologies

Bangalore, India

Founding Software Engineer | Python, React, MongoDB, Docker, AWS

Jan 2024 – Jul 2025

- Led a 7-member cross-functional team through the full SDLC and code reviews, architecting backend systems from the ground up for a **\$240K+ ARR** AI video conferencing platform with **99.9%** uptime on scalable AWS microservices
- Reduced video preprocessing latency from **180s** to **50s** across **~150GB/day** of video data by revamping the FFmpeg pipeline into a maintainable, multithreaded, I/O-tuned architecture with asynchronous S3 prefetching for real-time WebRTC streams
- Achieved **24 FPS** AI inference under high-concurrency production load by delivering a Celery-based distributed inference system
- Improved deployment reliability for a **2K MAU** multimodal inference platform by implementing canary rollouts on a Modal and Nomad setup with Prometheus-based monitoring and alerting

### Infibeam Avenues

Bangalore, India

Software Development Engineer | Ruby on Rails, AngularJS, MySQL, Docker, ELK, Kubernetes

Jul 2022 – Jan 2024

- Shipped a Firebase-based seller notification system serving **1M+** sellers, driving a **15%** increase in seller engagement through real-time push notifications within an Agile team
- Saved support teams **120+ hours/month** by spearheading a suite of REST APIs in Ruby on Rails to automate seller ticketing workflows, collaborating with third-party partners and integrating external services
- Delivered a round-robin work allocation system and clearly documented its logic to distribute seller stock approvals across admins, reducing approval time by **25%** and preventing bottlenecks

### Totallr Technologies

Chennai, India

Software Development Intern | Angular, Spring Boot, MySQL, AWS

Apr 2021 – Jun 2021

- Shipped a full-stack Point-of-Sale system using Spring Boot and Angular, reducing checkout times and automating invoicing across **20+** enterprise clients at launch
- Identified and optimized slow database queries, cutting latency from **1100 ms** to **600 ms** under peak transaction loads through refactored access patterns, triggers, stored procedures, and rigorous testing
- Built an analytics dashboard using Chart.js, enabling fast access to sales insights for forecasting and inventory planning

## PROJECTS

### Distributed Log Processing System

- Accelerated end-to-end log analysis throughput by **10x** to **~1M events/min** compared to a synchronous baseline by engineering an ETL pipeline with Apache Kafka, Apache Flink, ClickHouse, and Redis
- Achieved **<100 ms** average query latency for an observability dashboard with a Next.js frontend and FastAPI backend, powered by Redis and ClickHouse via caching and analytical query optimization

### HyRA - Hybrid Routing Agent

- Developed a LLM routing service with vLLM, LangGraph, and Flask optimized for latency, accuracy, and cost, boosting efficiency **~30%** over a static baseline
- Improved routing accuracy from **71%** to **89%** by designing a PostgreSQL-backed model registry and integrating an LLM-as-a-Judge feedback loop with iterative prompt engineering
- Productionized vLLM model serving on GCP with Langfuse monitoring and performance profiling for distributed LLM workloads

## SKILLS

**Languages & Databases:** Python, Ruby, Java, TypeScript/JavaScript, HTML/CSS, MySQL, PostgreSQL, Oracle DB, MongoDB, Redis

**Frameworks & Libraries:** React, Angular, Ruby on Rails, Flask, FastAPI

**Cloud & Infrastructure:** AWS, GCP, Docker, Kubernetes, Nginx, ELK Stack, Prometheus, Grafana, Linux

**LLM & Developer Tools:** OpenAI API, Gemini API, Claude, Langfuse, Visual Studio, Cursor, Git/GitHub, Postman