# Umar Ahmed Thameem Ahmed

umar.ahmed.t.30@gmail.com | (720) 232-0077 | LinkedIn: umarahmed2000 | GitHub: UmarAhmed30 | 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, Cloud Computing, Computer Networks, Machine Learning, Big Data Analytics, Data Mining, Software Engineering*

## SKILLS

**Programming Languages:** Python, Ruby, Java, JavaScript, TypeScript, SQL
**Web Development:** 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, India**
*Founding Software Engineer* | Python, FastAPI, React, MongoDB, Docker, AWS, ELK *Jan 2024 – Jul 2025*

- Led a 7-member **cross-functional** team and architected **backend systems** for a **$240K+ ARR AI video conferencing platform**, achieving **99.9% uptime** through scalable **AWS microservices**
- Reduced video pre-processing latency by **35%** by engineering a **multithreaded**, **I/O-optimized FFmpeg** pipeline that **parallelized** chunking and transcoding with asynchronous **S3** prefetching for **real-time WebRTC streams**
- Achieved **24 FPS AI inference** under high-concurrency production load by architecting a **Celery**-based distributed inference system for asynchronous task orchestration
- Improved **deployment reliability** for a **~2K MAU** multimodal inference platform by implementing **canary rollouts** on a **Modal** and **Nomad** setup with **Prometheus**-based monitoring for latency, token usage, and error rates

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

- Shipped a scalable **Firebase-based seller notification system** serving **1M+** sellers, driving a **15%** increase in seller engagement through real-time push notifications within an **Agile** team
- Developed a suite of **REST APIs** in **Ruby on Rails** to **automate** seller ticketing workflows and **integrate** third-party service APIs, partnering with support and business teams to reduce manual efforts by **120+ hours/month**
- Built a **round-robin work allocation system** to evenly distribute seller stock approval requests across admins, reducing stock approval times by **15%** and preventing workload bottlenecks

**Totallr Technologies** **Chennai, India**
*Software Development Intern* | Angular, Spring Boot, MySQL, AWS *Apr 2021 – Jun 2021*

- Shipped a high-throughput **full-stack Point-of-Sale (PoS)** system using **Spring Boot** and **Angular**, reducing checkout times and automating invoicing across **15+** enterprise clients at launch
- Identified and optimized slow database queries by refactoring access patterns and utilizing **database triggers** and **stored procedures**, cutting the overall query latency from **1100 ms** to **700 ms** during peak transaction loads
- Built an **analytics dashboard** using **Chart.js**, enabling fast access to sales insights for forecasting and inventory planning

## PROJECTS

**Distributed Log Processing System**

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

**HyRA - Hybrid Routing Agent**

- Built a **distributed LLM routing service** with **vLLM**, **LangGraph**, and **Flask**, 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 with iterative **prompt engineering** to evaluate routing decisions, improving accuracy from **71%** to **89%**
- Deployed **vLLM**-based model serving on GCP with **Langfuse** for **production-grade monitoring** and **profiling**