Rishi Ravula

571-660-8500 | rishi.ravula@gmail.com | New York, NY, USA | https://rishiravula.fyi | github.com/RishiRavula

EDUCATION

Duke University **August 2020 - May 2024**
Bachelor of Science Engineering, Electrical Computer Engineering & Computer Science *GPA: 3.78*

CERTIFICATIONS & AWARDS

AWS Solutions Architect Associate, Capital One Enterprise-Wide Hackathon Winner (1/700). HackDuke 2021 Health Sector Winner (1/300)

PROFESSIONAL EXPERIENCE

Capital One **New York, NY, USA**
Senior Associate Software Engineer *August 2024 - Present*

- Building **MCP-based LLM Agents** to perform continuous, near-real-time monitoring: **decreasing Mean Time to Acknowledge to minutes**.
- Built low-latency filtering APIs over DynamoDB, enabling real-time queries on **400+ datasets**; **$1M+ in data monetization at 60K+ QPS.**
- Spearheading MVP for vector and graph databases, enabling low-latency access in read-distributed systems; developing from 0 to 1.
- Integrated **Glue ETL pipelines and LLMs** in recruitment systems, achieving **90% accuracy** and **reducing human bias** by **37%** in backtesting.
- Optimized low-latency data store, **slashing latency by 70%** using Elasticache/Redis for real-time ML fraud detection.
- Built an automated data loader using S3, Glue, and Lambda, to **restore and recover TBs of data during disasters; cut recovery time by 90%.**

Natter **London, UK**
Software Engineer Contract *August 2023 - January 2024*

- Built a highly available, scalable video conferencing platform supporting **10,000+ concurrent users** under heavy traffic.
- Developed **fine-tuned, open-source LLMs**, trained on proprietary data, utilizing zero-shot inferencing and novel embedding techniques.
- **Increased attentive engagement by 40%** through AI-driven user matching and predictive analytics, enhancing interaction quality.

Capital One **New York, USA**
Software Engineering Intern *May 2023 - August 2023*

- Conducted the migration of a core account, Java Spring Boot service from an EC2 instance to a serverless Fargate infrastructure.
- Migration **outperforms the legacy service 45% faster in response times** and tolerates **2.6 million monthly transactions** in production.
- Orchestrated Fargate clusters, Load Balancers, IAM roles and more, including cross-region failovers and Docker containerization.

Tuned.com **Scottsdale, AZ, USA**
Software Engineering Intern *May 2022 - August 2022*

- Developed a high-frequency trading alert system for the crypto automated trading platform using JavaScript, Kotlin, and GraphQL.
- Alerts were sent as client-facing notifications with low-latency routing mechanisms and with up to **an 85% success rate per execution**.
- Contributed to the platform's success prior to its acquisition, alerts supplementing over **1 million trades** and **$6 billion in transaction volume.**

PROJECTS

Promptful.ai

- Building Promptful, a full-stack app to concurrently compare LLM outputs from OpenAI, Claude, etc. using Go and Typescript.
- Leading 3 developers to eliminate the friction of prompt testing across LLMs by centralizing access, state management, model comparison.
- Integrated Hugging Face Transformers and TRL to support user low-code fine-tuning and local inference with Dockerized self-hosted models.
- Built Go-based backend microservice and Typescript UI for model comparison, prompt tuning, and personalized usage feedback.

End-to-End LLMs For Production Course

- Launched an 8-week LLM production curriculum, attracting 150+ pre-enrollments covering topics from data ingestion to model deployment.
- Engineered real-time pipelines (RabbitMQ CDC, Flink, Qdrant) delivering sub-second embeddings and adopting MLOps for 99% model uptime.
- Secured early investors, marketing to top universities and enterprises for widespread adoption.

Capital One Enterprise-Wide Hackathon

- Architected and led a team to develop a RAG-powered Slack/VSCode agent using Docker, pgvector, and LLaVa models.
- Unanimously won 1st place; solution adopted for production by Director of Slack with me as team lead overseeing implementation.
- Estimated to save $1M during production, increasing question answer search accuracy by 80% and impacting 30,000 employees.

AI Celebrity Replica

- Fine-tuned Phi-3 and Mistral 8B for celebrity emulation: LangFuse, PostgreSQL and Tavily for 40% response accuracy improvement.
- Orchestrated AWS services (EC2, ELB, Route 53, Docker) for 99.9% uptime/scalability to enable retrieval augmented generation.
- Sold for 6 figures and delivered to Sensay, securing 2 full-time roles and a 30% engagement boost.

SKILLS

Languages: Python, TypeScript, Scala, Kotlin, Java, SQL, Swift, C, HTML, CSS

Technologies & Tools: AWS, Docker, Kubernetes, LLMs, GraphQL, REST APIs, Redis, PostgreSQL, Jenkins, Git,

Libraries & Frameworks: PyTorch, FastAPI, Numpy, Pandas, LangGraph, LangChain, HuggingFace Transformers, TRL