Naman Mudgal

Greater Noida ,India
+91-9521463164 naman.mudgal2001@gmail.com https://www.linkedin.com/in/naman-mudgal-6735931ba/
https://github.com/MudNam

Technical Skills

Languages: C++,Python, Golang, Java, Data Structures and Algorithms, TypeScript (React, Node.js),Nextjs
Machine Learning: MiniLM, BERT, VQ-RAG, Retrieval-Augmented Generation (RAG), Model Quantization (INT8)
Technologies/Frameworks: Kubernetes (GKE, EKS), FastAPI, Gin, gRPC, Prometheus, Grafana, Terraform, ArgoCD

Experience

Walm **Nov 2024 – Present**
Software Engineer *Gurgaon*

- Designed and deployed a high-performance NLP inference pipeline using MiniLM embeddings with dimensionality reduction, achieving 93.4% faster startup via Docker multi-stage builds and model pre-caching. Enabled zero-downtime blue-green deployments on GCP MIG with GitHub Actions CI/CD, automated health checks, SLO-based traffic shifting, and N-1 rollback support, delivering 99.98% availability for entity recognition and semantic search services.
- Spearheaded a multi-modal Agent-Based AI conversational system leveraging fine-tuned **BERT**-derived emotion embeddings (8-dimensional affective space) with vector-quantized Retrieval-Augmented Generation **(VQ-RAG)**, reducing hallucinations by 91.3% across 2,300+ daily customer interactions. Collaborated on MapReduce-powered Market Basket Analysis with XGBoost for real-time product recommendations, achieving 14.6 AOV increase while engineering deterministic conversation flows via parameterized prompt templates that elevated completion rates from 76 to 94.2% with 38ms p95 latency.
- Built and optimized a React Native commerce application as part of a 7-engineer team, integrating custom-distilled LLMs deployed on RunPod using ONNX quantization (INT8), achieving significant lower inference latency versus cloud alternatives while maintaining high accuracy parity. Implemented context-aware feature flags and comprehensive testing strategies to ensure quality while enabling frequent production releases.
- **Tools: PyTorch, Typescript, Kubernetes, Transformers, Pinecone, Langchain, Python, Postgress, GCP, Python, GitHub Actions, RAG FastAPI, Prometheus**

Qonto **May 2024 – Oct 2024**
Software Engineer *Paris, Île-de-France*

- Architected and implemented a fault-tolerant banking transaction processing system featuring atomic multi-account operations with robust deadlock prevention and comprehensive audit trails.
- Engineered a horizontally scalable microservice architecture with circuit breakers, rate limiting, and automated failover mechanisms to handle peak financial processing demands.
- Designed secure API endpoints with token-based authentication, comprehensive six-level logging infrastructure, and extensive automated testing using database simulation for reliable verification.
- **Tools: C++, Golang, PostgreSQL, Docker, gRPC, Jenkins, Gin, AWS, Kubernetes, EKS, Argocd**

Tradable **Dec 2023 – Mar 2024**
Software Engineer Intern *Noida,UP*

- Engineered high-performance auction platform handling 100,000+ concurrent bids with (99.6%) transaction success rate, scaling to support 10,000 registered users with (87.3%) retention
- Implemented secure user authentication with role-based access control reducing unauthorized attempts by (94.2%), while designing interactive dashboards with real-time activity tracking increasing user engagement by (76.8%)
- **Tools: Node.js, React, Typescript, Docker, AWS, Kubernetes, Prometheus, GitHub Actions, Kafka**

Projects

Multimodal AI Video Analysis App | *ML, Pytorch* **November 2023**

- Built an end-to-end multimodal AI system using PyTorch that analyzes video content for sentiment and emotion classification by integrating video frame processing, audio feature extraction, and text embedding with sophisticated fusion techniques.
- Developed a complete SaaS platform with user authentication, API key management, and quota tracking that seamlessly integrates with AWS SageMaker endpoints for scalable inference, providing real-time analysis through a modern web interface.
- **Tools: PyTorch, Next.js, React, Tailwind CSS, AWS (S3, SageMaker, IAM), Auth.js, TensorBoard, BERT, T3 Stack, TypeScript, Prisma**

Education

Indian Institute of Tachnology (IIT) Patna **May 2020 – Aug 2024**
Bachelor of Technology -Minor in AI *Patna.India*