Nicholas Goh

AI Full Stack Engineer

✉ gohn0004@e.ntu.edu.sg ☎ +65 96958068 Singapore nicholas-goh.com

linkedin.com/in/nicholas-goh-19ba1b194 github.com/NicholasGoh

Professional Experience

Klass Engineering and Solutions 08/2023 – present

Software Engineer (AI Engineering)

- Designed system and implemented an app to showcase LLM Orchestration Capabilities. This opened opportunities forfuture work with AI Agents delegating and executing tasks
- Modularizing UI into a reusable base Chatbot framework to reduce technical debt and streamline development, cuttingUI development and integration time for future teams.
- Architected and developed an in-house centralized model weights caching, to address scaling and MLOps challenges.This resulted in 3% (2TB/70TB) of disk space savings and 5 days of time savings per developer
- Analyzed third party codebase to identify and debug critical RAM and VRAM leak, resulting in 100% improvement ofefficiency
- Implement database caching of chunking and vectorization stages of RAG for production environments, improvingefficiency by at least 100%

Projects

Customer Service Automation 02/2025 – 03/2025

- Built an AI-powered system that automatically handles multi-part queries, significantly reducing manual effort,improving response times, and increasing operational efficiency in customer interactions
- Implemented automated error handling and conflict resolution, ensuring more reliable booking processes, minimizingdisruptions, and enhancing overall user experience
- Introducted LLM-based testing and Langsmith tracing to ensure high-quality, consistent outputs from AI agents,significantly reducing troubleshooting time and improving overall system stability and performance

Agentic RAG 09/2024 – 11/2024

- Evaluated trade-offs between NoSQL, Milvus, PostgreSQL and PGVector extension, accessing the additionalcomplexity required to implement cross database consistency
- Evaluated trade-offs between using AWS Lambda + Amplify and EC2 for deployment, opting for EC2 to simplify local andserver testing for both backend and UI
- Leveraged IaC with Terraform to automate provisioning of AWS EC2, ECR, and security policies, enabling cost-effectiveand reproducible setup and teardown of cloud resources
- Optimized document ingestion and vectorization on a small EC2 instance, using a rolling window approach to avoidmemory constraints and preserve context between chunks

Skills

Cloud & Devops: GCP, Grafana, Prometheus, GH Actions | **Databases:** Cassandra, Redis, Sqlite |

Languages & Frameworks: Python, Fastapi, Langgraph, Langchain, Typescript, React, CPP