Nicholas Goh

AI Full Stack Engineer

gohn0004@e.ntu.edu.sg
https://www.linkedin.com/in/nicholas-goh-19ba1b194/
https://www.nicholas-goh.com
+6596958068

Work Experience

Software Engineer (AI Engineering)

Aug 2023 - Present

Klass Engineering and Solutions

- Designed system and implemented an app to showcase LLM Orchestration Capabilities. This opened opportunities for future work with AI Agents delegating and executing tasks
- Modularizing UI into a reusable base Chatbot framework to reduce technical debt and streamline development, cutting UI 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 of efficiency
- Implement database caching of chunking and vectorization stages of RAG for production environments, improving efficiency by at least 100%

Projects

Customer Service Automation

Feb 2025 - Mar 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, minimizing disruptions, 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

Sep 2024 - Nov 2024

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

Core Skills

Langchain, Langgraph, Typescript, Docker, PostgreSQL, PGVector, Milvus, Python, CPP, React, Cassandra, Terraform

Education

Nanyang Technological University

Aug 2019 - May 2023

Bachelor of Science Data Science and Artificial Intelligence