Soo Kyung Ahn

Software Engineer
Email: sahnsookyung@gmail.com
Site: https://sookyungahn.com

SUMMARY

Backend software engineer transitioning into AI systems, with 3.5 years building reliable, observable services for high-volume banking and digital goods platforms. Hands-on experience integrating LLMs with caching/retries, and experimenting with both local inference and GPU-hosted inference workflows. Strengths include Kubernetes-based deployments, incident-ready observability, and pragmatic engineering practices applied to real-time, user-facing AI experiences.

TECHNICAL SKILLS

AI & ML: LLM Orchestration (OpenAI, Gemini, Anthropic), Local Inference (Ollama, vLLM, GGUF), Voice Pipelines (Kokoro TTS, Faster-Whisper, Silero VAD), Prompt templating.
Interactive Applications: Unity 3D (C#), VR Development.
Frontend Engineering: TypeScript, React.js, Web Components, TailwindCSS, Three.js, Vite
Backend Engineering: Python, Java, Event-Driven Architecture (Kafka, SQS), REST APIs, WebSockets, Databases(MySQL/PostgreSQL), Playwright (web scraping)
Infrastructure: AWS(EKS, EC2, S3, Lambda, RDS, Route53, SQS/SNS), Kubernetes, Docker, Istio, Observability Stack (ELK, Prometheus, Grafana), CI/CD (Github Actions, CircleCI).

AI ENGINEERING PROJECTS

Locally Hosted Conversational Voice Agent | Python, Kokoro TTS, Whisper, Ollama

github.com/sahnsookyung/LLM-voice-chatbot

- Built an end-to-end local voice agent using Silero VAD, Faster-Whisper (STT), Ollama (LLM), and Kokoro TTS.
- Implemented streaming TTS playback and non-blocking audio capture; profiled latency bottlenecks across STT/LLM inference and audio chunk boundaries.
- Implemented sliding-window conversation history and prompt templating to control context growth across multi-turn chats.
- Ran quantized LLMs via Ollama (e.g., Gemma-3-27b GGUF) and tuned for local memory/quality trade-offs.

LLM-Based Data Expansion Pipeline | Python, Gemini API, RunPod

github.com/sahnsookyung/parts-data-cleaning

- Built a batch pipeline for dataset expansion with caching, retry logic, and exception logging to handle rate limits and transient API failures.
- Benchmarked RunPod-hosted Gemma-27b vs managed LLM APIs; migrated to Gemini 2.5 Flash to improve speed/quality per cost.
- Streamed responses to maintain throughput when provider-side batch APIs were unreliable.

FRONTEND PROJECTS

Interactive Portfolio Website | TypeScript, Web Components, Tailwind CSS, Three.js

https://github.com/sahnsookyung/jiko-shoukaisho-website · Live: ***sookyungahn.com***

- Single-page interactive portfolio using clickable SVG regions, overlay viewers, and tooltips to present professional and personal background.
- TypeScript Web Components for UI modules (tooltips, overlays, navigation), keeping the codebase modular and easy to extend.
- 3D visual effects and animations with Three.js, optimized via code-splitting, effect toggles, and SVG asset minification.
- Built with Vite and Tailwind CSS to enable fast iteration on layout and styling with a lightweight, framework-free stack.

PROFESSIONAL EXPERIENCE

Software Engineer | Electrum Payments

Digital Goods and Services Team | Cape Town, South Africa | 2022 – 2025

Built and maintained services supporting millions of transactions per day across digital goods products for banks and national retailers.

- Led the development and maintenance of backend services for digital goods (airtime, vouchers, money transfers, lotto), powering high-volume transactions across South Africa's top banks and national retailers.
- Engineered modular, plugin-based service components that enabled rapid feature delivery and robust testability across a shared payments platform.
- Service deployment and lifecycle management on AWS-backed Kubernetes (EKS), owning Istio routing, rollout orchestration, and runtime diagnostics.
- Partnered with DevOps to integrate core infrastructure including cloud messaging (Kafka, SQS/SNS), DNS (Route53), and observability stacks into developer workflows.
- Designed Kibana dashboards and Elasticsearch queries for real-time operational visibility and business-critical monitoring.
- Played a key role in delivering complex, multi-product client launches in collaboration with cross-functional engineering teams.
- Prototyped internal AI-powered tools including an LLM-based code review bot and automated ticket triage system to streamline engineering and support workflows.
- Development practicing pragmatic engineering principles, driving test automation (JUnit, Mockito), code quality (SonarQube), and CI/CD reliability across the team.

EDUCATION

B.Sc. Honours in Computer Science (First Class GPA) | University of Cape Town | 2021

- Thesis Focus: Virtual Reality (Unity) & Human-Computer Interaction. Built a VR environment to study user engagement in nature simulations.

B.Sc. Computer Science & Applied Statistics (Distinction) | University of Cape Town | 2018 – 2020

- Capstone Project: Created a visualization environment for complex molecules in the context of drug design.
- Dean's Merit List