CHO, SEONGLAE

Gower St, London WC1E 6BT
+44 7356-161248 | seonglae.cho.24@ucl.ac.uk

SKILLS

- Service: Full stack | TypeScript | Node.js | PostgreSQL | Redis | Rust | Vector Database | Faiss | Milvus | Vite | React | Streaming
- AI: Python | PyTorch | AI Agent | AI Cost Optimization | DDP | FSDP | ONNX | Pydantic AI | RAG | LangGraph | AI Evaluation
- Infra: Kubernetes | Linux | Git | CI/CD | Docker | Docker Compose | Ansible | ETL | Github Action | Hadoop | Distributed Systems

EXPERIENCE

HOLISTIC AI London, United Kingdom
AI Research Engineer Intern May 2025 - Present

- Implemented an evaluation pipeline for Deep Research AI Agent using OpenSSF baseline metrics to assess method performance

KAKAO MOBILITY Seoul, South Korea
Software Engineer, Digital Twin Team December 2021 - September 2022

- Led a 3-person team in developing a national-scale 3D mapping service as a part of the Autonomous Driving pointcloud pipeline
- Ported a C++ 3D-projection algorithm to Rust with Node.js bindings, making the library cross-platform

STRYX Seoul, South Korea
Software Engineer, 3D Mapping Team November 2019 - December 2021

- Reduced build time by 70% and simplified dependency management by merging multiple repositories into a monorepo
- Downsized the Docker image by 90%, from 2GB to 180MB, by applying multi-stage builds in CI, elevating team productivity

EDUCATION

UNIVERSITY COLLEGE LONDON London, England, United Kingdom
Artificial Intelligence for Sustainable Development MSc September 2024 – June 2025
YONSEI UNIVERSITY Seoul, South Korea
Computer Science BE March 2017 - August 2024

AWARDS

HERMES, 1ST PLACE (£3,000) , HOLISTIC AI HACKATHON (2024) London, England, United Kingdom
Team Lead November 2024 – November 2024

- Fine-tuned Sparse AutoEncoder (SAE) for GPT-2 to identify and steer correlated features for multiple biases for AI Safety
- Reduced stereotypical text generation by 20% from an initial 90% rate by applying a Steering Vector derived from the SAE

MBTIGPT, 1ST PLACE (₩3,000,000), YONSEI GENAI COMPETITION (2023) Seoul, South Korea
Team Lead September 2023 - January 2024

- Built an end-user AI service that employs RAG on user chat history by an MBTI personality analyzer with Redis and Faiss
- Acquired over 1,000 users within a month, with even paid purchases, by optimizing free-tier model and reducing costs by 30%

PUBLICATIONS

- Cho, S., Jang, M., Yeo, J., & Lee, D. (2023). *RTSUM: Relation Triple-based Interpretable Summarization with Multi-level Salience Visualization*. In Proceedings of the **NAACL 2024** System Demonstrations Track. Association for Computational Linguistics. https://aclanthology.org/2024.naacl-demo.5/
- Cho, S. (2025). SAE Training Dataset Influence in Feature Matching and a Hypothesis on Position Features. *AI Alignment Forum*. https://www.alignmentforum.org/posts/ATsvzF77ZsfWzyTak/dataset-sensitivity-in-feature-matching-and-a-hypothesis-on-1

PROJECTS

UNIVERSITY COLLEGE LONDON London, England, United Kingdom
MCP-Notion January 2025 – February 2025

- Built a Model Context Protocol (MCP) SSE server that searches Notion pages and converts them to markdown for interoperability

YONSEI UNIVERSITY DATA & LANGUAGE INTELLIGENCE LAB Seoul, South Korea
ReSRer September 2023 - January 2024

- Indexed 21M Wikipedia-scale corpus into a Milvus vector database and accelerated LLM training through distributed training

RTSum March 2023 - August 2023

- Published as the first author, designed Knowledge Graph (KG)-based summarization experiment for Interpretable AI framework

YONSEI UNIVERSITY Seoul, South Korea
LLaMa2GPTQ June 2023 - July 2023

- Reduced LLaMa2 memory cost by 75% with 4-bit GPTQ quantization and integrated RAG vector search for Local LLM

Texonom November 2021 - June 2023

- Built an ANN-based vector retrieval API by embedding all 30,000 content pages in service into Postgres pgVector database