

# CHO, SEONGLAE

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

## EDUCATION

### UNIVERSITY COLLEGE LONDON

*Artificial Intelligence for Sustainable Development MSc*

London, England, United Kingdom

2024 - Present

### YONSEI UNIVERSITY

*Computer Science BE*

Seoul, South Korea

2017 - 2024

## PUBLICATIONS

- Cho, S., Jang, M., Yeo, J., & Lee, D. (2023). *RTSUM: Relation Triple-based Interpretable Summarization with Multi-level Saliency Visualization*. In Proceedings of the **NAACL 2024** System Demonstrations Track. Association for Computational Linguistics. <https://arxiv.org/abs/2310.13895>

## EXPERIENCE

### KAKAO MOBILITY

*Software Engineer*

Seoul, South Korea

December 2021 - September 2022

- Achieved OS independence of the library by refactoring a C++ 3D algorithm module to Rust, leveraging Node.js NAPI binding
- Led a 3-person team in developing an app for 3D maps as a part of the Autonomous Driving pointcloud data pipeline

### STRYX

*Software Engineer*

Seoul, South Korea

November 2019 - December 2021

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

## AWARDS

### HERMES, 1ST PLACE (£3,000), HOLISTIC AI HACKATHON (2024)

*Team Lead*

London, England, United Kingdom

November 2024 – November 2024

- Fine-tuned Sparse AutoEncoder (SAE) for GPT-2 to identify and analyze correlated features addressing biases in AI Safety
- Achieved 90% stereotyped text generation via Steering Vector, matching fine-tuned model performance without LLM training

### MBTIGPT, 1ST PLACE (₩3,000,000), YONSEI GENAI COMPETITION (2023)

*Team Lead*

Seoul, South Korea

September 2023 - January 2024

- Implemented enterprise-level RAG application of AI personality analyzer using Redis, OpenAI API, Node.js and Faiss
- Reduced OpenAI API costs by 30% by prompt optimization, utilizing code from LLM as optimizers' paper

## PROJECTS

### UNIVERSITY COLLEGE LONDON

*Neural-land*

London, England, United Kingdom

Sep 2024 – Nov 2024

- Extracted 5,000+ features from Mistral 8b by implementing automated interpretability, utilizing LLM as a Neural Explainer

*MCP-Notion*

Jan 2025 – Jan 2025

- Built a Model Context Protocol (MCP) server with markdown support, enabling navigation in Anthropic's Claude Desktop

### YONSEI UNIVERSITY DATA & LANGUAGE INTELLIGENCE LAB

*ReSRer*

Seoul, South Korea

September 2023 - January 2024

- Improved ODQA(Open-Domain Question Answering) performance by 20% with zero-shot LLM prompt engineering
- Completed Wikipedia-scale QA evaluation by indexing 21M Wikipedia passages into a Milvus vector database in 12 hours
- Accelerated LLM training speed using Pytorch's multi-node distributed training FSDP with 4 x RTX3090 with QLoRa

*RTSum*

March 2023 - August 2023

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

### YONSEI UNIVERSITY

*LLaMa2GPTQ*

Seoul, South Korea

June 2023 - July 2023

- Optimized computing and memory cost by 75% using 4-bit GPTQ quantization applied to the LLaMa2 model

*Texonom*

November 2021 - June 2023

- Extended context window size by 400% by building a custom model deploying the e5 ONNX model for the recommender system
- Developed vector search API for RAG by embedding whole 30,000 pages in service into Postgres pgVector database

## SKILLS

- Python | Pytorch | Rust | Typescript | C++ | Transformers | DDP | FSDP | PEFT | Vector Database | Milvus | ODQA | Faiss | Git
- LLM | RAG | PostgreSQL | Redis | CI/CD | Kubernetes | Docker | Docker Compose | Hadoop | Github Action | Distributed Systems
- TGI | TEI | TDB | Model Inference | Model Training | Prompt Engineering | Interpretable AI | Ansible | ETL | Node.js | ONNX