CHO, SEONGLAE

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

EDUCATION

UNIVERSITY COLLEGE LONDON London, England, United Kingdom
Artificial Intelligence for Sustainable Development MSc 2024 - Present

YONSEI UNIVERSITY Seoul, South Korea
Computer Science BE 2017 - 2024

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://arxiv.org/abs/2310.13895

EXPERIENCE

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

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

STRYX Seoul, South Korea
Software Engineer November 2019 - December 2021

- Reduced build time by 70% and simplified dependency management by merging multiple repositories into a mono-repository, while coordinating with DevOps teams
- Enhanced internal server efficiency by cutting bandwidth and TTFB by 80%, introducing multi-layer Redis caching
- Downsized the Docker image by 90%, from 2GB to 180MB, by applying multi-stage builds in CI, elevating team productivity

PROJECTS

YONSEI UNIVERSITY DATA & LANGUAGE INTELLIGENCE LAB Seoul, South Korea
Yokhal March 2024 - April 2024

- Accelerated LLM training speed using Pytorch's multi-node distributed training FSDP with 4 x RTX3090 with QLoRa

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

- Improved ODQA(Open-Domain Question Answering) performance by 20% with zero-shot LLM context manipulation
- Completed large-scale QA benchmarks by indexing 21M Wikipedia passages into a Milvus vector database in 12 hours
- Boosted LLM evaluation by 40% by introducing a multi-GPU local inference server, Huggingface TGI with asynchronous batch processing

YONSEI UNIVERSITY Seoul, South Korea
MBTI GPT 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

YONSEI UNIVERSITY DATA & LANGUAGE INTELLIGENCE LAB Seoul, South Korea
RTSum March 2023 - August 2023

- Published as the first author, designed Knowledge Graph (KG)-based experiment for validating Interpretable AI framework
- Increased OpenIE5's NLP triple extraction speed by 300% by deploying a reverse proxy and Docker container replicas

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

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

YONSEI UNIVERSITY Seoul, South Korea
Texonom November 2021 - June 2023

- Extended context window size by 400% by building a custom transformer 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