

윤수환 Suhwan Yoon

AI Engineer

Seoul, South Korea | [LinkedIn](#) | [Career Mate](#)

Five years of AI/NLP engineering on financial data (filings, news, stocks), taking LLM agents, RAG, and document extraction from research through deployed APIs. Currently building search and multi-agent systems at Wanted Lab.

EXPERIENCE

AI 기술팀 정규직

(주)원티드랩

Nov 2025 — Present

Seoul, South Korea

Replaced the BM25 position search engine with a SPLADE + Dense + Lexical 3-leg hybrid search, tuned via A/B tests. Post-search bounce rate fell from 35% to 28% and search-driven application conversion rose from 6.7% to 8.4%. Also built the career mentor multi-agent service, where resume diagnosis raised user resume levels, and migrated legacy position agent code to LangChain and LangGraph.

AI Engineer → AI Engineer Leader

웨이커코리아 주식회사

Oct 2022 — Aug 2025

Seoul, South Korea

Built a multi-agent AI chat API that uses MCP tools grouped per agent, LLM orchestration, and a GuardRail module blocking sensitive identifiers. Developed a financial stock analysis agent using Plan-and-Act workflows, a Llama-Index QA agent that validates extracted SEC filing data, and a Kafka-based news pipeline with translation and summarization APIs. Earlier work covered SEC filing extraction with LLM and multimodal models, plus model compression for translation (distillation, pruning, quantization, ONNX).

NLP Developer 정규직

(주)디앤아이파비스

Jul 2021 — Oct 2022

Seoul, South Korea

Built a BM25 patent document search engine on self-hosted Elasticsearch. Domain-tuned Sentence-BERT for problem and solution sentence matching, improving accuracy by 1.2% over vanilla BERT. Developed an LDA-based keyword extraction and clustering API for patent documents.

인턴 연구원

네이버 주식회사

Feb 2021 — May 2021

Seoul, South Korea

Researched training methods for ColBERT in information retrieval. Used Learning-to-Augment to generate hard negative samples and improved the Triplet Loss training approach.

EDUCATION

경희대학교 대학원

컴퓨터공학과 석사 (자연언어처리)

Mar 2019 — Feb 2021

경희대학교

컴퓨터공학과 학사

Mar 2013 — Feb 2019

2018 Junior Paper Competition Encouragement Award (KSC)

PROJECTS

Multi-agent career mentoring service at Wanted, Career Mate 커리어 멘토 에이전트

Six workflows and a multi-agent system covering mentoring, mentor recommendation, resume diagnosis, and nudge message generation. Resume diagnosis raised data quality for 11,626 cumulative users.

Compressing MBART for service deployment, 번역 모델 경량화

Built a half-depth MBART with knowledge distillation, then applied PTQ and attention-based pruning. Exported to ONNX as a static graph for serving.

Attention-based question difficulty model for QA, Question Difficulty Estimation (Applied Sciences)

Co-first-authored a paper on estimating question difficulty with an attention model for question answering.

SKILLS

LangChain · LangGraph · Llama-Index · Agentic AI · Multi-Agent · MCP · Prompt Engineering · GuardRail · Hybrid Search · SPLADE · BM25 · OpenSearch · Elasticsearch · FAISS · Reranking · A/B Testing · PyTorch · Scikit-Learn · Named Entity Recognition · Topic Modeling · Knowledge Distillation · Model Pruning · Post Training Quantization · ONNX · Python · FastAPI · Flask · PostgreSQL · Redis · Apache Kafka · AWS · Docker · Jenkins · GitHub Actions · Git