Jaehyeok Choi

Seoul, South Korea | wowogur12@naver.com | github.com/choijhyeok | linkedin.com/in/choijh0119

Profile

AI Engineer specializing in LLMs, RAG, Agentic AI, and large-scale document processing pipelines. Strong experience in designing and implementing end-to-end generative AI systems, from PoC development to real-world, production-oriented scenarios. —

Work Experience

AI Engineer Jul 2024 — Present
KT *Seoul, South Korea*
- Delivered multiple enterprise-grade GenAI PoCs for legal, finance, media, and global clients
- Designed complex Agent-based systems, including RAG, Web Search Agents, and Realtime Voice Agents
- Built end-to-end solutions on Azure, from data processing to demo web deployment

Data Scientist Apr 2022 — May 2024
AIFactory *Seoul, South Korea*
- Designed and validated AI competitions and enterprise PoCs across LLM, CV, and time-series domains
- Built LLM fine-tuning pipelines, RAG systems, and performance evaluation workflows (RAGAS)
- Implemented large-scale data preprocessing and automated inference pipelines

Data Scientist Jul 2021 — Jan 2022
GSITM *Seoul, South Korea*
- Developed forecasting and optimization models for retail and logistics domains
- Implemented demand forecasting and matching algorithms for real-world business operations

—

Key Projects (KT)

AI Engineer, Large-scale Insurance Policy RAG Pipeline (Meritz) May 2025 — Dec 2025
- Designed and implemented a large-scale RAG pipeline for 280,000 insurance policy documents
- Built an end-to-end indexing workflow using Azure Durable Functions
- Developed custom loaders for PDF, PPT, Excel, and JSON sources
- Implemented table-stable extraction and custom chunking logic using PyMuPDF
- Applied a Parent–Child RAG architecture on Azure AI Search

AI Engineer, AI Branch – Realtime Voice-based AI Banker (Shinhan Bank) Jan 2025 — Apr 2025
- Built a voice-to-voice AI banker PoC using the GPT-4o-Realtime API
- Implemented WebSocket-based real-time streaming pipelines
- Developed loan recommendation agents using Function Calling
- Integrated Azure AI Search for financial product retrieval and reasoning
- Deployed the full system on Azure Container Apps

AI Engineer, Web Search Agent (Giga Genie / JTS Thailand) Sep 2024 — Mar 2025
- Developed multiple Web Search Agent PoCs based on Bing Search
- Designed agent workflows using LangGraph
- Reduced hallucinations using Self-RAG and Corrective-RAG
- Achieved up to 94% answer accuracy with sub-10s response latency
- Built demo applications using Gradio

AI Engineer, Legal RAG System PoC (Korea Forest Service) Aug 2024 — Dec 2024
- Built a legal-domain RAG system PoC for statutes and legal precedents
- Designed a combined LLM fine-tuning + RAG architecture
- Achieved Top-5 retrieval accuracy of 93.51%
- Evaluated performance with RAGAS (Context Precision, Recall, Faithfulness > 90)
- Developed a demo web using Gradio and pdf.js with source highlighting

—

Selected Projects (AIFactory)

Data Scientist, LLM Fine-tuning and RAG Evaluation
- Fine-tuned LLaMA2, Gemma, and EEVE models using QLoRA and DeepSpeed
- Built RAG pipelines and evaluated performance using RAGAS
- Developed crawling pipelines for external data collection

Data Scientist, AI Competition Design (CV / Time-Series)
- Designed challenges for Object Detection, Segmentation, and Pose Estimation
- Built datasets in COCO format and defined evaluation metrics (Macro F1, IoU, MAE)
- Led competition design and validation for multiple enterprise and public clients

—

Selected Projects (GSITM)

Data Scientist, Retail Sales Forecasting System
- Preprocessed convenience store sales data with seasonality analysis
- Built Prophet-based sales forecasting models
- Proposed inventory optimization strategies based on forecast results

Data Scientist, Genetic Algorithm-based Logistics Matching Optimization
- Implemented optimization algorithms using DEAP
- Developed matching logic to maximize profit under weight and cost constraints
- Improved logistics efficiency and operational decision-making

—

Personal Projects & Talks

Personal Project, Gemma Function Calling Assistant
- Fine-tuned Gemma 7B using SFT to enable Function Calling
- Built a personal assistant that executes external tools and delivers results via KakaoTalk

Speaker / Open Source Contributor, LLaMA2 Fine-tuning & RAG Pipeline
- Presented an end-to-end LLaMA2 fine-tuning and RAG pipeline
- Implemented QLoRA-based instruction fine-tuning optimized for Colab T4
- Built an easy-to-use fine-tuning pipeline with a Gradio UI
- Resources: GitHub Repo, YouTube Talk

Speaker, Building RAG-based Services with Streamlit & LangChain
- Speaker at LangChain KR Meetup (2024 Q1)
- Demonstrated deployment of RAG-based AI services using Streamlit
- Explained vector DB separation and QA-chain-based recommendation architectures
- Showcased automated report generation pipelines (HTML → PDF)
- Slides: PPT

Open Source Maintainer, Korean HWP/HWPX Document Parsing Open Source
- Developed Python libraries for parsing HWP and HWPX documents
- Addressed real-world limitations of Korean document processing in Python environments
- Resources: GitHub Repo, LinkedIn Post

—

Education

Gyeongsang National University — South Korea
B.S. in Information Statistics and Computer Science — *Mar 2016 — Feb 2022*

—

Skills

- **Languages**: Python, JavaScript, SQL
- **LLM / GenAI**: RAG, LangChain, LangGraph, Fine-tuning, Function Calling, RAGAS
- **Infrastructure**: Azure Functions, Azure AI Search, Azure Container Apps, Docker
- **Document Processing**: PyMuPDF, OCR, Large-scale Document Pipelines