# Junyeong Song

LLM | Generative AI | RAG Engineer

junyeong.nero@gmail.com github.com/junyeong-nero LinkedIn

## Summary

Engineer focused on end-to-end pipelines for generative AI: LLM fine-tuning, diffusion models, and RAG systems. Independently delivers projects from data collection and synthesis through training, evaluation, and deployment, with a consistent emphasis on Korean-language performance.

## Education

**UNIST (Ulsan National Institute of Science and Technology)** *2020.03 – 2026.07*
B.S. in Computer Science and Engineering | GPA 3.83 / 4.3 | Unistar Scholarship (top entrance)

## Work Experience

**CJ AI Division** — Intern *2025.02 – 2025.10*

- Built parts of a multi-agent pipeline for drama-script analysis (Parser and Scene-Analysis agents)
- Designed a VLM-based multi-format document parser (Qwen2.5-7B-VL); improved Korean OCR **WER from 20% to 7%**
- Adopted a thinking-style model and simplified the workflow; raised scene-analysis **F1 from 0.2 to 0.5 (2.5x)**

**EmbeddedAI Lab, UNIST** — Lab Intern *2024.07 – 2024.09*

- Studied diffusion models and reproduced DDPM / DDIM / CFG papers from scratch

## Key Projects

**Drama Scripts Analysis Pipeline with Multi-Agents** *2025.02 – 2025.10 | Team of 4*
*Stack:* Python, FastAPI, LangChain, Transformers, PyTorch

- Built a multi-format document parser converting PDF / HWP / DOCX scripts into clean TXT
- Replaced traditional OCR with Qwen2.5-7B-VL VLM, improving Korean OCR **WER 20% → 7%**
- Designed a Chain-of-Thought scene-level strength/weakness classifier; F1 **0.2 → 0.5**
- Integrated Parser / Analyzer / Evaluator agents into a unified LangChain workflow

**Synthetic OCR Image Generator** *2025.12 – 2026.02 | Solo*
*Stack:* Python, Transformers, OpenCV

- Built a font-based synthetic dataset pipeline across 5 subsets (Document, Table, Sentence, Markdown, KIE)
- Optimized $O(N^2)$ SSIM bottleneck during similar-character DB construction with a **2-stage pipeline** (embedding-based fast filter → SSIM precise comparison)
- Made visual perturbations (background, blur, contrast, shadow) configurable for edge-case OCR evaluation

**tiny-chatbot-agents — Dual-Stage RAG Chatbot** *2026.01 – 2026.02 | Solo*
*Stack:* Python, Streamlit, vLLM, Ollama, Playwright, ChromaDB

- Built a Playwright-based dynamic crawler for financial T&Cs and QnA data
- Designed a hybrid **Dual-Stage RAG**: QnA-first retrieval (speed) → T&C precise retrieval (precision)
- Combined Vector / Rule / Triplet search with Cross-Encoder reranking; **Top-5 Recall +20%**
- Added a hallucination-verifier agent and an LLM-Judge eval (Accuracy / Faithfulness / Completeness)
- Implemented agentic workflow on top of MCP (Model Context Protocol)

**tiny-stable-diffusion — SD3 From Scratch** *2025.12 – 2026.02 | Solo*
*Stack:* Python, PyTorch, HuggingFace Datasets

- Implemented Stable Diffusion 3's **MMDiT + VAE** architecture from scratch
- Applied joint attention, adaLN-zero, Rectified Flow, logit-normal sampling, and Min-SNR training
- Set up CC3M-wds (3M) streaming-mode training and a continual-training system
- Resolved prompt-overfitting on small datasets through scale-up and quality improvements

**Korean Medical LLM — Domain SFT** *2024.09 – 2024.10 | Team of 2*
*Stack:* Python, HuggingFace (Datasets, TRL), Unsloth

- Built a domain dataset via translation, web crawling, and QA conversion of medical-legal documents
- Extended KorMedMCQA with new 2025 labeling (449,500 examples)

- Ran data-mixing experiments at Korean:English = 60:40, General:Medical = 2.5:1
- Used **model merging** to retain general-domain ability while improving medical performance (google/gemma2-9b base)

**Bilingual Translation LLM (Jeju ↔ Standard Korean)** *2024.05 – 2024.06 | Team of 3*
*Stack:* Python, HuggingFace (TRL, Datasets), PyTorch

- Curated AIHub Jeju-dialect data into **453k parallel pairs**
- Built a single bidirectional model via tag prompting (`<dialect_to_standard>`, `<standard_to_dialect>`)
- Trained on Colab T4 (16GB) with **QLoRA (4-bit quantization)** for memory efficiency
- Achieved **BLEU 0.56 / ROUGE-L 0.60**, on par with full fine-tuning

**tiny-graph-RAG — Knowledge-Graph RAG** *2025.12 – 2026.01 | Solo*
*Stack:* Python, OpenAI API, PyVis, Streamlit

- Auto-built knowledge graphs by extracting entities and relations with an LLM
- Hand-implemented BFS traversal and query-relevance subgraph ranking (no external libraries)
- Used `asyncio.gather` for batched async LLM calls; reduced duplicate nodes **50 → 15**

## Technical Skills

| | |
|---|---|
| **Languages** | Python, C/C++ |
| **Deep Learning** | PyTorch, HuggingFace (Transformers, TRL, Datasets) |
| **LLM** | SFT, LoRA / QLoRA, PEFT, Unsloth, vLLM, Ollama |
| **Generative Models** | DDPM, DDIM, CFG, Stable Diffusion 3 (MMDiT, VAE, Rectified Flow) |
| **RAG / Agents** | LangChain, ChromaDB, Graph RAG, MCP, Cross-Encoder Reranking |
| **Data Engineering** | Playwright, OpenCV, web crawling, dataset synthesis |
| **Infra / Tools** | FastAPI, Streamlit, W&B, Runpod, asyncio, Git |

## Publications

- **KorMedMCQA**: Multi-Choice QA Benchmark for Korean Healthcare Licensing Examinations *(3rd author)* arxiv.org/abs/2403.01469

## Certifications & Awards

- **Orak Contest (KRAFTON)** — 8th / 117 teams
- **Unistar Scholarship** — top-entrance award, UNIST
- TOEIC 845 | Computer Proficiency Certificate, Level 1 (Korea)

## Activities

- Teaching Assistant — AI Intro and Programming 1, UNIST