송준영 (Junyeong Song)

LLM · Generative AI · RAG Engineer

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

Summary

LLM Fine-tuning, Diffusion, RAG 시스템을 중심으로 생성형 AI의 End-to-End 파이프라인을 구축해 온 개발자입니다. 데이터 수집·합성부터 모델 학습, 평가 자동화까지 독립적으로 수행하며, 한국어 도메인에서의 성능 최적화에 집중하고 있습니다.

Education

UNIST (울산과학기술원) *2020.03 – 2026.07*
Computer Science and Engineering, B.S. | GPA 3.83 / 4.3 | Unistar 장학금 (수석입학)

Work Experience

국내 대기업 AI 부서 (CJ AI Division) — Intern *2025.02 – 2025.10*

- 드라마 대본 분석 Multi-Agent 파이프라인 프로젝트 참여 (Parser / Scene Analysis Agent 개발)
- Qwen2.5-7B-VL 기반 Multi-Format Document Parser 설계, 한국어 OCR **WER 20% → 7%** 개선
- Thinking Model 도입 및 워크플로우 단순화로 Scene 분석 **F1-score 0.2 → 0.5 (2.5배)** 향상

EmbeddedAI LAB (UNIST) — Lab Intern *2024.07 – 2024.09*

- Diffusion 모델 기반 생성 연구 수행 (DDPM / DDIM / CFG 논문 재현)

Key Projects

Drama Scripts Analysis Pipeline with Multi-Agents *2025.02 – 2025.10 · 4인*
기술: Python, FastAPI, LangChain, Transformers, PyTorch

- PDF/HWP/DOCX 대본을 TXT로 변환하는 Multi-Format Document Parser 개발
- VLM(Qwen2.5-7B-VL) 도입으로 기존 OCR 대비 한국어 인식 WER **20%** → **7%** 달성
- Chain-of-Thought 기반 Scene 단위 강/약점 분류 에이전트 설계, F1-score **0.2** → **0.5** 향상
- Parser/Analyzer/Evaluator 에이전트 LangChain 워크플로우 통합

Synthetic OCR Image Generator *2025.12 – 2026.02 · 1인*
기술: Python, Transformers, OpenCV

- 5개 서브셋(Document, Table, Sentence, Markdown, KIE) 폰트 기반 합성 데이터 생성 파이프라인 구축
- Wikipedia corpus 기반 유사 문자 DB 구축 시 SSIM $O(N^2)$ 병목을 **2-Stage 파이프라인**(임베딩 고속 필터링 + SSIM 정밀 비교)으로 최적화
- Background texture, blur, contrast, shadow 등 시각적 변형을 config로 제어하는 엣지 케이스 평가 구조 설계

tiny-chatbot-agents — Dual-Stage RAG 챗봇 *2026.01 – 2026.02 · 1인*
기술: Python, Streamlit, vLLM, Ollama, Playwright, ChromaDB

- Playwright 기반 금융사 이용약관·QnA 동적 크롤링 파이프라인 구축
- QnA 우선 검색(Speed) → 약관 정밀 검색(Precision) 하이브리드 Dual-Stage RAG 설계
- Vector + Rule + Triplet Search 결합 및 Cross-Encoder Reranking 적용, **Top-5 Recall 약 20% 향상**
- Hallucination Verifier 에이전트 및 LLM Judge 자동 평가(Accuracy / Faithfulness / Completeness) 구축
- MCP(Model Context Protocol) 표준 기반 에이전틱 워크플로우 구현

tiny-stable-diffusion — SD3 From Scratch *2025.12 – 2026.02 · 1인*
기술: Python, PyTorch, HuggingFace Datasets

- Stable Diffusion 3의 **MMDiT + VAE** 아키텍처를 바닥부터 구현
- Joint Attention, adaLN-zero, Rectified Flow, Logit-normal sampling, Min-SNR 등 최신 기법 적용
- CC3M-wds(3M) 스트리밍 모드 학습 및 Continual Training 시스템 구성
- 소규모 데이터셋 프롬프트 Overfitting 문제를 데이터 규모 확장·품질 개선으로 해결

Korean Medical LLM — Domain SFT *2024.09 – 2024.10 · 2인*
기술: Python, HuggingFace (Datasets, TRL), Unsloth

- 영어 의료 데이터 번역, 웹 크롤링, 법률 문서 QA 변환 등 다각적 데이터 구축 전략 실행
- KorMedMCQA 2025년도 확장 라벨링 수행 (449,500 examples)
- 한국어:영어 = 60:40, General:Medical = 2.5:1 비율의 데이터 배합 실험

- **Model Merge** 기법으로 일반 도메인 성능 하락 최소화, 의료 도메인 성능 안정적 향상 (google/gemma2-9b 기반)

Bilingual Translation LLM (Jeju ↔ Standard Korean) *2024.05 – 2024.06 · 3인*
기술: Python, HuggingFace (TRL, Datasets), PyTorch
- AIHub 제주도 방언 데이터를 정제해 **453k 병렬 쌍** 학습 데이터 구축
- 태그 기반 프롬프팅으로 단일 모델 양방향 번역 구현
- Colab T4(16GB) 환경에서 **QLoRA(4-bit)** 로 메모리 최적화
- **BLEU 0.56 / ROUGE-L 0.60** 달성, Full Fine-tuning 대비 동등 수준 검증

tiny-graph-RAG — 지식 그래프 기반 RAG *2025.12 – 2026.01 · 1인*
기술: Python, OpenAI API, PyVis, Streamlit
- LLM 기반 엔티티·관계 추출로 지식 그래프 자동 구축
- BFS 탐색 및 질의 연관성 기반 서브그래프 랭킹 알고리즘 직접 구현 (외부 라이브러리 미사용)
- `asyncio.gather` 기반 비동기 배치 처리로 LLM 호출 병목 해소, 중복 노드 **50** → **15** 감소

Technical Skills

Language	Python, C/C++
Deep Learning	PyTorch, HuggingFace (Transformers, TRL, Datasets)
LLM	SFT, LoRA / QLoRA, PEFT, Unsloth, vLLM, Ollama
Generative Model	DDPM, DDIM, CFG, Stable Diffusion 3 (MMDiT, VAE, Rectified Flow)
RAG / Agent	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 (*3*저자)
  arxiv.org/abs/2403.01469

Certifications & Awards

- **Orak Contest (KRAFTON)** — 8위 / 117팀
- **Unistar 장학금 (UNIST 수석입학)**
- TOEIC 845 · 컴퓨터활용능력 1급

Activities

- Teaching Assistant — AI Intro and Programming 1 (UNIST)