Seoyeon Kim

Republic of Korea

seoyeon.riley.kim@gmail.com | GitHub | LinkedIn



EXECUTIVE SUMMARY

- **Agent Applications and Agentic RAG:** Design and build enterprise AI applications with multi-agent orchestration, dynamic routing, parallel retrieval, and tool calling. Develop document parsing and indexing pipelines, hybrid search, and hierarchical knowledge-base filtering to connect enterprise documents and structured data with context-aware, source-grounded responses.
- **Multimodal LLM / Multimodal RAG:** Research and develop vision–language systems for document understanding and evidence retrieval across text, images, and tables. Combine GraphRAG, late-interaction retrieval, and verifier-guided workflows to ground responses in visual and cross-page evidence; fine-tune multimodal models for multilingual document understanding.
- **On-Device AI and Model Adaptation:** Build offline inference pipelines integrating language models, speech recognition, OCR, speech synthesis, and visual language detection. Apply model benchmarking and LoRA/PEFT fine-tuning to assess quality, latency, and CPU/GPU deployment constraints in restricted-network environments.

EDUCATION

- **Korea University** September 2019 – February 2023 (Graduated)
  Seoul, South Korea
  Master of Engineering – Big Data Convergence
  Machine Learning & Vision Lab (Advisor: Prof. Hyunwoo Kim)
  Research: Computer Vision, Vision Transformer, Deep Learning.
- **Kookmin University** March 2016 – February 2019 (Graduated)
  Seoul, South Korea
  Bachelor of Business Administration – International Business

WORK EXPERIENCE

- **Mazelone** June 2025 – October 2026
  Seoul, South Korea · AI Lab
  AI Research Engineer

  – **Enterprise AI Agents for Internal Documents & Hierarchical Knowledge Bases**
  Project Lead / AI Agent Development — Korean Register
  Designed and implemented query routing and parallel evidence aggregation across specialized agents for internal policies, quality documents, bulletin boards, official correspondence, and international classification requirements, plus a department-lookup sub-agent.
  Improved retrieval over maritime regulatory documents by linking vessel-specific conditions, convention applicability, document text, and revision history; exposed execution progress and answer sources through tool hooks and SSE.
  Used OpenTelemetry and LangSmith to diagnose bottlenecks; narrowed retrieval scope, limited execution turns, introduced early stopping and query-rewrite caching, and separated list retrieval from full-text fetching.
  Improved response speed by optimizing agent execution and reducing redundant model calls and searches.

- **Evaluation Studio – Agent Quality Evaluation & Feedback Loops**

  Planned, designed, and developed dataset/split/run management and batch evaluation for agent retrieval and response quality.
  Incorporated answer judgments, reference documents, and user feedback into a ground-truth loop, linking failure analysis to evaluation-set improvements.

- **Knowledge-Base Filtering & Personalization**

  Designed and implemented hierarchical knowledge-base filters and catalog synchronization across approximately 20,000 document-filter nodes, covering enterprise and regulatory documents.
  Built personal filter presets and a prompt library for storing, managing, and reusing document scopes and instructions.

- **Survey Guidance AI & Multi-Agent Hybrid RAG (Korean Register)**

  Contributed to knowledge retrieval and response services over approximately 540,000 documents and 4 million chunks. Connected unstructured retrieval and structured lookup using Orchestrator–Agent–Sub-agent workflows, hybrid search, and NL-to-OData tool calling.
  Linked maritime conventions, classification rules, and survey guidance for inspection-related queries with domain-specific retrieval conditions.
  Evaluated retrieval and responses with NDCG, MRR, Recall@K, and LLM-as-a-Judge; achieved 59/60 E2E passes (98.3%) on a development evaluation set.
  Built FastAPI/Azure AI Search/SSE serving and source attribution; managed the scope of existing index operations and new chunking/indexing transitions.

- **Enterprise AI Agent Hub PoC – Automotive-Parts Manufacturing**

  Designed requirements and sub-agent collaboration for PLM, BOM, SCM, quality, purchasing, and production workflows.
  Connected tool-calling schemas, API connectors, and RAG retrieval through reusable enterprise-system interfaces.

- **LLM Email & Official-Document Generation**

  Built an email/document auto-generation system that retrieves user- and context-specific information from databases and generates emails and official documents aligned with standard forms, organizational guidelines, and user context — a draft-then-review flow, not unattended send.

- **Elasticsearch/OpenSearch-based On-Premises Hybrid Search**

  Designed mapping, index schema, and vector fields, plus BM25, dense, and multivector search pipelines, reflecting closed-network and security constraints.
  Designed LLM–DB-linked document generation that reduces repetitive document search-and-write work against standard forms and guidelines, and a layout-aware OCR and GraphRAG document-understanding flow.

- **Multimodal GraphRAG & Late-Interaction Retrieval Research**

  Selected for Korea's 2026 HPC Support Program (2× NVIDIA A100 GPUs); GraphRAG and late-interaction (MaxSim) re-ranking for industrial drawings and manuals.
  Secured 78.6M KRW GPU resources; reduced infrastructure cost by 30.4M KRW (76%) over four months.

- **Multimodal Parser Research**

  Research on multimodal information extraction from scanned, unstructured, and low-quality PDFs using multimodal LLMs / VLMs; focus on visually grounded parsing and document understanding for on-device deployment.

- **National Police Agency – On-Device Multilingual Translator**

  Built offline multimodal translation (text, speech, image, document) for police/investigation field use under limited connectivity; benchmarked 11 languages and 10 models.
  Applied PEFT/LoRA and full fine-tuning to Gemma SLM and Whisper STT with domain data; CPU/GPU lightweight inference.
  Extended PATIMT-Bench and fine-tuned LLaVA-NeXT to improve document understanding beyond OCR-centric pipelines.
  VLM image language detection; contributor to patent application 10-2025-0139821.

- **Kakao Brain** September 2021 – August 2024
  Seongnam, South Korea · Healthcare Service Planning
  AI Product Manager

  – **KARA-CXR: Medical Imaging Draft-Report VLM**
    Led research-to-product work on a VLM that turns chest X-rays into structured draft reports; turned research, clinical, and hospital needs into features, data requirements, metrics, and an expert review–and-edit workflow so clinicians approve and correct model output before it is used.
    Supported 9-hospital joint research and CXR–report pairing; SQL/BigQuery/Python analysis for model/clinical metrics and release scope; used evaluation results to prioritize product improvements.

  – **KakaoTalk AI Shopping Mate**
    Participated in generative AI and conversational search planning for style commerce; defined scenarios, requirements, and UX hypotheses from product/search intent.

  – **Today Recovery TFT**
    Planned a digital-health app with video-based exercise routines and conversational AI feedback; validated via FGI/Sprint (weekly exercise +1.8, satisfaction 4.0/5.0, AI feedback intent 4.3/5.0).

  – **Galux AI Drug-Design Strategic Investment & Joint Research**
    Researched AI drug-discovery market and global protein-structure prediction trends; contributed to Kakao Brain's KRW 5B strategic investment and Galux joint-research partnership.

- **GC Care** December 2018 – September 2021
  Seoul, South Korea · Service Planning
  Product Manager

  – **HowCare Health Screening App**
    Led health screening result data integration from public insurance and ∼150 institutions; standardized item names, units, formats, and exceptions into searchable data for downstream visualization and personalized analysis/recommendation (100,000+ download app).
    Contributed to App Award Korea 2022 Digital Healthcare – App of the Year (Grand Prize).

PUBLICATIONS & RESEARCH

- **ColGraphRAG: Late-Interaction Evidence Retrieval for Multimodal GraphRAG**
  arXiv
  Formulates **multimodal evidence grounding** and graph-constrained **vision–language retrieval** as **late-interaction** (ColEmbed/MaxSim) matching over graph-linked visual nodes, **in lieu of** pooled bi-encoder similarity, while **holding constant** upstream graph induction, textual and tabular retrieval, and decoding so that improvements accrue primarily to the retrieval stage.
  On MultimodalQA (paired vs. QD-MMGraphRAG), aggregate F1/EM increase to **58.3 / 55.0** (from **54.7 / 53.0**), with graph-linked image retrieval improving to Hit@1 **57.0** (vs. 38.0) and nDCG@3 **39.5** (vs. 32.6); gains concentrate on image- and table-centric questions.
- **LITTA: Late-Interaction and Test-Time Alignment for Visually-Grounded Multimodal Retrieval**
  arXiv
  MaxSim-style **late-interaction** and **multi-vector retrieval** with frozen encoder, LLM query expansion, and reciprocal rank fusion for **document understanding** and **vision–language retrieval** over figures, tables, and captions.
  Multi-query expansion (q=3) reaches Recall@10 of **78.45%**; on industrial manuals, Recall@5 improves from 59.81% to 60.37% (q=5) without retriever retraining.

- **VLD-RAG: Agentic Vision–Language Retrieval-Augmented Generation for Long, Visually-Rich Multi-Page Documents**
  arXiv
  Agentic **vision–language RAG** for long, visually rich documents: page-preserving indexing, keyword–semantic **hybrid retrieval**, and verifier-guided agents for cross-page **multimodal evidence grounding** and **document understanding**; evaluated on LongDocURL and MMLongBench-Doc.
- **MedBioRAG: Semantic Search and RAG for Medical and Biological QA**
  arXiv — December 2025
  Built biomedical QA pipeline with semantic search, document **reranking**, and RAG; reported benchmark improvements on medical and biological QA tasks.
- **TransMLLM: Multimodal Multilingual Multitask Document Understanding for Low-Resource Languages**
  arXiv — November 2025
  End-to-end **vision–language document understanding** from document images to target-language text without OCR preprocessing (11 languages); **MLLM fine-tuning** (LoRA and full fine-tuning) of LLaVA-NeXT-7B.
  Full fine-tuning improves average BLEU by **6.5** and COMET by **6.6** across 10 language pairs vs. baseline.
- **MedBioLM: Fine-Tuned LLMs and RAG for Medical and Biological QA**
  arXiv — February 2025
  Domain-adapted biomedical QA combining fine-tuning and RAG for short- and long-form queries.
- **FracFormer: Semi-supervised Learning for Vertebrae and Fracture Classification on 3D Radiographs with Transformers**
  Master's Thesis, Korea University — February 2023
- **3DYoga90: Hierarchical Video Dataset for Yoga Pose Understanding**
  arXiv — October 2023

AWARDS

- **Microsoft AI School Final Project – Excellence Award (2nd Place)**
  Microsoft — February 2025
  Built medical consultation system integrating STT, summarization, and RAG chatbot; fine-tuned and deployed domain-specific LLM.
- **Dongjak-gu Big Data Competition – Excellence Award (3rd Place)**
  Dongjak-gu Office — May 2020
  Built ML regression model for district aging index analysis with policy visualization.

PATENTS

- **On-Device AI Translation Service (Method, User Terminal, Computer Program)**
  Patent Application No. **10-2025-0139821** — September 26, 2025
  On-device SLM translation with OCR, STT, and TTS; CPU/GPU inference; VLM image language detection and glossary-based customization.

ACTIVITIES

- **Instructor, 2026 Advanced GPU Training Program for Physical AI**
  Korea Polytechnics University — May 2026
- **Assistant Instructor, AI for Impact Program: AI Capacity-Building Education Initiative**
  SK hynix & Microsoft — May 2025

CERTIFICATIONS

- **AI-900: Microsoft Azure AI Fundamentals** January 13, 2025
  Microsoft
- **SQLD: SQL Developer** June 25, 2021
  KOREA Data Agency
- **ADsP Data Analysis Semi-Professional** December 22, 2020
  KOREA Data Agency

TECHNICAL SKILLS & TOOLS

- **LLM Application / Agent Orchestration:** Python; LangChain, LangGraph, LangSmith; ReAct, Plan-and-Execute, Orchestrator–Agent–Sub-agent; tool/function calling; NL-to-OData tool agents; prompting, context engineering, structured outputs; HITL review/approval; multiturn dialogue and conversation history.
- **Evaluation / Observability:** LangSmith tracing; OpenTelemetry E2E latency tracing; Evaluation Studio dataset/split/run management and ground-truth feedback loops; LLM-as-a-Judge; A/B tests; NDCG, MRR, Recall@K, Precision@K; failure-case analysis; quality / latency / cost optimization.
- **Multimodal AI / Vision–Language Models:** Vision–language models (CLIP, LLaVA) and multimodal LLMs; vision–language retrieval; VLM/MLLM fine-tuning; classification, summarization, translation, and QA.
- **Document Understanding / Information Extraction:** Document understanding, document information extraction, and multimodal information extraction over documents, images, and sequences; OCR and visually grounded parsing; layout-aware indexing and structured report generation.
- **Retrieval / Memory / Search Infrastructure:** Agentic RAG, Hybrid Search, GraphRAG; BM25 + dense + late-interaction (MaxSim) / multi-vector retrieval; query expansion; hierarchical indexing; conversation history as context; Azure AI Search, Elasticsearch, OpenSearch, Qdrant, PostgreSQL (PGVector).
- **Backend / MLOps:** FastAPI, REST/Inference APIs, SSE streaming; PostgreSQL; Docker, Docker Compose, Kubernetes, Airflow; batch ingestion; Spark; Pandas, NumPy; TensorBoard, W&B.
- **ML / DL:** PyTorch; Hugging Face Transformers, Accelerate, PEFT, LoRA; LLM/VLM fine-tuning; embedding and representation learning; multi-GPU training.