Ilmin Cho

Seoul, KOR | linkedin.com/in/ilmincho | ilmincho.me | github.com/ilmincho | 010-7154-7548 | choim426@gmail.com

EDUCATION

Georgia Institute of Technology — **Atlanta, GA**
MS in Computer Science — *2026-Present*

University of Massachusetts Amherst — **Amherst, MA**
BS in Computer Science (GPA: 3.78 / 4.0) | Honors: Mass Transfer Scholarship, Dean's List. — *2022-2023*

Bunker Hill CC — **Boston, MA**
Associate of Science in Computer Science (GPA: 3.97 / 4.0) | Honors: Dean's List. — *2019-2021*

EXPERIENCE

AI Scientist (Agent-Architecture Cell) — **Seoul, South Korea**
LG AI Research — *Jun 2025 – Present*

- Developing the ChatEXAONE(chatbot) Service Orchestration System and designing asynchronous workflows for LLM-tool integration.
- Applying architectural structures and fine-tuning models for Reasoning and Function Calling to optimize agentic behavior.
- Implementing the on-premise deployment of ChatEXAONE, a large-scale enterprise chatbot system, focusing on system reliability, monitoring, and scalability.

NLP Research Engineer Intern — **Seoul, South Korea**
LG AI Research — *Nov 2024 – May 2025*

- Engineered a high-throughput Topic-Modeling pipeline for 20M+ documents and performed Supervised Fine-Tuning (SFT) for 7.8B models using distillation techniques from EXAONE 32K.
- Optimized production deployment using gRPC and Triton Inference Server with Redis-based caching, significantly improving system reliability and reducing API latency and operational costs.

Generative AI Developer (Apprenticeship) — **Gyeonggi-do, South Korea**
Kakao Tech Bootcamp — *Jul 2024 – Nov 2024*

- Led two LLM-centric projects as Team Lead, focusing on SFT (GPT/BERT) and RAG pipelines.
- Built a retrieval system using ElasticSearch for semantic document search.

PUBLICATIONS & AWARDS

2026 ACL Findings (Co-Author)
LG AI Research, University of Illinois Urbana-Champaign — *Feb 2026*

- *From Documents to Segments: A Contextual Paradigm for Topic Assignment*

Grand Prize (1st Place)
MetLife AI Hackathon — *Mar 2024*

- *Developed an AI chatbot based on insurance industry data.*

PROJECTS

GitFolio (AI Resume Generator)
Python, FastAPI, GitHub API, GPT-4 — *Sep 2024 – Dec 2024*

- Developed an automated resume generation platform using GitHub repository analysis and GPT-based summarization.
- Optimized inference throughput by implementing multi-processing pipelines for concurrent API and model calls.

Review-Based Restaurant Recommender (What To Eat)
Selenium, BERT, KcELECTRA, HDBSCAN, ElasticSearch, FastAPI — *Jul 2024 – Aug 2024*

- Crawled and processed 120K+ reviews; designed an end-to-end pipeline for review clustering (UMAP + HDBSCAN) and ensemble-based ranking (KcELECTRA + BiLSTM).
- Optimized database structures and batching logic, reducing system latency from 222s to 26s (88% improvement).

Fall Detection Application (Mobile Sensing)
Python, Numpy, Scipy, Pandas, Decision Tree — *Apr 2023 – May 2023*

- Developed a real-time fall detection system by extracting key features (Max, Median, Dominant Frequency, Entropy) from mobile accelerometer sensor data.

- Implemented signal processing and data analysis pipelines to train a Decision Tree Classifier for predictive analytics.

Patient Tracker System (Healthcare Management)
FastAPI, PostgreSQL, Pytest *Apr 2023 – May 2023*

- Developed a web-based patient management system for real-time appointment orchestration, booking, and medical record tracking.
- Achieved 94% test coverage by implementing comprehensive unit and integration tests with Pytest, ensuring high system reliability.

TECHNICAL SKILLS

Languages: Python, Java, SQL, React

AI/ML: PyTorch, vLLM, TensorRT, LangChain, Triton Inference Server

Infrastructure: FastAPI, gRPC, PostgreSQL, MongoDB, Redis, ElasticSearch, Docker, AWS (EC2), GCP

Developer Tools: Git, Grafana, Jira, Argo, Slack