# Yejin Hong

+82 010-7721-8223 | lilyhong511@gmail.com

## Education

**Seoul National University** Sep. 2026 (Incoming)
M.S.-Ph.D. Integrated Program, Graduate School of Data Science

**Korea University** Mar. 2021 – Feb. 2026
Cyber Defense
Artificial Intelligence (minor)

**University of Texas at Austin** Aug. 2023 – May 2024
Exchange Student, Electrical Computer Engineering

## Research Interests

- **AI Safety & Alignment:** Enhancing the robustness and controllability of generative models (LLMs, MLLMs, Diffusion) against adversarial and Out-of-Distribution challenges.
- **Trustworthy Multimodal Systems:** Leveraging representation learning and Retrieval-Augmented Generation (RAG) to develop reliable, high-performance AI systems for complex real-world tasks.
- **AI for Security:** Advancing AI-driven vulnerability discovery, threat intelligence, and privacy-preserving techniques.

## Experiences

**SNU GSDS SKIML** **(Prof. Jay-yoon Lee)** Feb. 2026 – present

- Investigate adaptive retrieval for RAG by implementing sufficiency-score-based termination and post-retrieval pruning, while optimizing training through various loss functions and benchmark evaluation pipelines.
- Develop a hierarchical multi-agent framework with a tree-structured architecture for complex QA tasks, focusing on task decomposition and agent orchestration to enhance reasoning performance on standard benchmarks.

**KU Trustworthy AI Lab** **(Prof. Jongheon Jeong)** Sep. 2024 – Jan. 2026

- Conduct research on safety and reliability of LLMs/MLLMs, focusing on jailbreak attacks and hallucination detection in real-world generative AI systems.
- Reproduce and implement state-of-the-art defenses from recent papers, building evaluation pipelines that measure robustness with AUROC and related metrics across multiple models and datasets.
- Extend text-only methods to multimodal settings by leveraging visual features from MLLMs and running ablation studies to analyze failure modes and limitations.

**Intern** Jan 2025 – Feb 2025
Doctorsoft

- Contributed to an AI-powered reporting system that automatically extracts information from relational databases based on natural-language user requests.
- Designed and implemented end-to-end workflows in n8n to connect data retrieval, transformation, and report generation, reducing manual reporting effort and latency for internal stakeholders.

**Cardet in 42 Seoul** Jan 2022 – Nov 2022
42 Seoul

- Completed the 42 Seoul core curriculum, strengthening fundamentals in algorithms, data structures, and systems programming through peer-based learning.
- Developed individual and team projects in C/C++ on Linux, including custom libraries, shell utilities, and system-level tools under strict code review standards.
- Practiced intensive debugging and Git-based collaboration, building a strong foundation for large-scale software and AI engineering work.

## Projects

**Taxonomy-Guided Multi-Label Classification System** Sep 2025 – Dec 2025

- Engineered a multi-label classification framework for Amazon products by constructing a taxonomy-based graph structure to leverage hierarchical label dependencies.
- Integrated Graph Attention Networks (GAT) with EMA-based self-training mechanisms, establishing a unified pipeline that combines graph propagation with semi-supervised learning strategies.
- Developed automated shell scripts to streamline the training and inference processes across multiple embedding backbones (MPNet, GTE, BGE), ensuring reproducibility and operational efficiency.

**Jeju Dialect Speech-to-Text Translation** Mar 2025 – June 2025

- Proposed a modular Whisper–T5 speech translation architecture for converting Jeju dialect speech into standard Korean, comparing connector-based designs against a fine-tuned Whisper baseline.
- Implemented and evaluated multiple connector modules (MLP, Q-Former, STE) on the AI Hub Jeju Dialect Speech corpus, showing that all variants significantly underperform the fine-tuned Whisper baseline.
- Diagnosed latent-space mismatch between Whisper and T5 using UMAP visualizations, highlighting representation misalignment and tokenizer/decoder priors as key failure factors in modular STT.

**Performer-based Self-Supervised Music Source Separation** Mar 2025 – June 2025

- Replaced MixIT's ConvTasNet separator with a Performer-based Transformer to better capture long-range temporal structure in self-supervised music source separation.
- Combined FAVOR+ linear attention with Audio MAE pretraining, enabling training and inference on 30s mixtures with ˜20–30% lower training time and higher SI-SNR than ConvTasNet.
- Analyzed separation behavior using custom metrics (e.g., mask overlap, effective mask count) to study overfitting and limitations of SSL-based music separation on MOISESDB.

**Improving Military Satellite Object Detection via SR** Sep 2024 – Dec 2024

- Enhanced detection of small military vehicles in low-resolution satellite imagery by integrating deep super-resolution models with YOLO-based object detection on the xView dataset.
- Implemented and compared multiple SR models (EDSR, SRGAN, CycleGAN) and built an end-to-end pipeline from GeoTIFF/GeoJSON preprocessing to SR training, YOLO training, and mAP evaluation.
- Demonstrated that SR preprocessing—especially CycleGAN—significantly boosts mAP on small military platforms compared to training YOLO directly on low-resolution inputs.

**CAN Bus Anomaly Detection for Car Hacking** Sep 2024 – Dec 2024

- Designed an anomaly/attack detection pipeline for in-vehicle CAN bus logs, framing car-hacking scenarios as a supervised classification problem over normal and attack traffic.
- Engineered time-series features such as per-Identifier inter-arrival time (`Timedelta`), rolling-window entropy of CAN IDs, and integer-encoded identifiers, then trained a Random Forest classifier on combined normal and labeled data.
- Evaluated the model on a held-out validation split and generated attack predictions for the unlabeled test set, producing a submission file for automated grading and further analysis.

**My Cat Also Says Merry Christmas** Sep 2024 – Dec 2024

- Fine-tuned the instruction-based image editing model InstructPix2Pix for pet photos to perform object-specific, Christmas-themed edits from natural-language prompts (e.g., Santa hats, Christmas trees).
- Constructed a ˜1,300-sample triple-paired dataset (original image, edited image, instruction) using GPT-3.5-generated instructions and manual filtering with teammates.
- Developed a Gradio-based interactive demo that lets users upload pet photos, enter free-form instructions, and compare before/after Christmas-style edits in real time.

## Technical Skills

**Programming**: Python, C/C++, Java, SQL, Bash
**ML & Data**: PyTorch, TensorFlow, Hugging Face, scikit-learn, pandas, NumPy