이동건

(DongGeon Lee)

NLP/AI Researcher

자연어 처리와 대규모 언어모델(LLM)에 깊은 이해를 갖고 있으며, 신뢰할 수 있고 안전한 LLM을 구축하는 것에 관심이 있습니다.

구체적으로는 언어모델의 안전성 평가를 위한 평가 방법론 및 데이터 구축, 레드 티밍(Red Teaming), 가드레일 (Guardrail) 등 언어모델의 안전 및 보안을 관리·감독하는 것에 관심이 있습니다.

E-mail: dg.lee@postech.ac.kr
Personal Site: https://donggeon.github.io
Etc: CV, Google Scholar, LinkedIn, GitHub

연구 경력

포항공과대학교 Data Intelligence Lab / Research Assistant
2024년 2월 - 현재, 경상북도 포항시 / 지도교수: 유환조

VLM (Vision-Language Model)의 안전 벤치마크 및 평가 방법론에 대한 연구.

LLM (Large Language Model)의 내외부 지식 간 지식 충돌 문제 연구.

KT 연구개발센터 / Research Intern
2025년 1월 - 2025년 2월, 서울특별시

한국어 특화 LLM의 사전 학습을 위한 수학 말뭉치 데이터 합성 연구.

인하대학교 Data Intelligence Lab / Research Assistant
2022년 11월 - 2023년 11월, 인천광역시 / 지도교수: 최원익

항공 안전 도메인 특화 언어모델 학습.

항공 사고 조사 보고서 속 사고 원인 자동 추출을 위한 언어모델 Fine-tuning.

인하대학교 Nursing Informatics Lab / Research Assistant
2021년 7월 - 2023년 5월, 인천광역시 / 지도교수: 조인숙

병원 내 낙상 사건 진술문 탐지를 위한 언어모델 Fine-tuning (F1-score ≥ 0.97).

낙상 보고서 자동 생성을 위한 개체명 인식 모델 개발.

의료 데이터 수집 및 한국어 특화 전처리.

학력

포항공과대학교 (POSTECH) / 인공지능대학원 석사과정
2024년 2월 - 현재, 경상북도 포항시

지도교수: 유환조

인하대학교 / 정보통신공학과 학사
2018년 3월 - 2024년 2월, 인천광역시

장학금 및 수상 경력

우수 논문상(국립국어원장상)
2024년 10월, 2024년도 한글 및 한국어 정보처리 & 한국코퍼스언어학회 공동 학술대회

금상(국립국어원장상) / 인공지능의 한국어 능력 평가 경진 대회 [관련 기사]
2024년 10월, 국립국어원

우수 대학원생 입학 장려 장학금
2024년 5월, 포항공과대학교

최우수 공학도상
2024년 2월, 인하대학교 공과대학

우수 학생 연구원 연구 장학금
2023년 8월, 인하대학교

국제 논문

(*: 공동 1저자)

Are Vision-Language Models Safe in the Wild? A Meme-Based Benchmark Study

DongGeon Lee*, Joonwon Jang*, Jihae Jeong, Hwanjo Yu

arXiv preprint, 2025 [paper]

REFIND at SemEval-2025 Task 3: Retrieval-Augmented Factuality Hallucination Detection in Large Language Models

DongGeon Lee, Hwanjo Yu

SemEval Workshop @ ACL, 2025 [paper]

Typed-RAG: Type-Aware Decomposition of Non-Factoid Questions for Retrieval-Augmented Generation

DongGeon Lee*, Ahjeong Park*, Hyeri Lee, Hyeonseo Nam, Yunho Maeng

XLLM Workshop @ ACL, 2025 [paper]
NAACL 2025 Student Research Workshop, 2025 (Non-Archival)

Theme-Explanation Structure for Table Summarization using Large Language Models: A Case Study on Korean Tabular Data

TaeYoon Kwack*, Jisoo Kim*, Ki Yong Jung, **DongGeon Lee**, Heesun Park

TRL Workshop @ ACL, 2025 [paper]

Enhancing Adverse Event Reporting With Clinical Language Models: Inpatient Falls

Hyunchul Park, Insook Cho, Byeong Sun Park, **DongGeon Lee**

Journal of Advanced Nursing (SCIE, IF:3.8, Q1; 97.2%), 2025 [paper]
(SCIE, IF:3.8, Q1; 97.2%)

Effects of Language Differences on Inpatient Fall Detection Using Deep Learning

Insook Cho, EunJu Lee, **DongGeon Lee**

MedInfo, 2023 [paper]

Bridging the Reporting Gap of Inpatient Falls to Improve Safety Practices Using Deep-Learning-Based Language Models and Multisite Data

DongGeon Lee, EunJu Lee, Insook Cho

AMIA Clinical Informatics Conference, 2023 [paper]
Oral Presentation

국내 논문

(*: 공동 1저자)

다중 추론 검색 증강 생성을 위한 데이터 합성과 학습 방법

이규민, 전민진, 장상환, **이동건**, 유환조

한국컴퓨터종합학술대회 (KCC), 2025

질문 유형에 따른 검색 증강 생성의 질의응답 성능 분석

이동건*, 박아정*, 이혜리, 남현서, 맹윤호

제 36 회 한글 및 한국어 정보처리 학술대회, 2024 [paper]
Oral Presentation

Tabular-TX: In-Context Learning을 통한 주제-설명 구조 기반 표 요약

곽태윤*, 김지수*, 정기용, **이동건**, 박희선

제 36 회 한글 및 한국어 정보처리 학술대회, 2024 [paper]
우수 논문상, Oral Presentation

딥 러닝 기반 영상처리를 통한 스마트 항만 주차정보시스템 설계 및 구현

구창훈*, 정윤주*, **이동건***

한국정보처리학회 추계학술발표대회 (ACK), 2021 [paper]
Oral Presentation

보유 기술

프로그래밍 언어

Python, Shell Script, C/C++

프레임워크/라이브러리

PyTorch, transformers, vLLM, LangChain, TensorFlow, Keras

시스템 및 기타

Git, Linux, Markdown, LaTeX, MySQL