배수현

baeshstar@korea.ac.kr | 010-3357-4231

연구 비전: 신뢰를 설계하고 모두를 위한 AI를 만듭니다

AI의 능력을 넘어 그 능력을 '신뢰'할 수 있도록 만드는 기술에 집중하고 있습니다. LLM의 명확한 한계를 정면으로 마주하고, 기술로 그 신뢰를 쌓아가는 연구에 몰두해왔습니다. 저의 연구 여정은 크게 두 갈래의 길을 따라왔습니다.

주요 연구 경험

신뢰성 높은 AI를 위한 환각 연구 및 Agentic 시스템 설계

- **생성형 AI의 환각 유형 분석 및 평가 지표 한계 연구 (KCC 2024 발표)**
  - 질문-답변 시 발생하는 환각을 5가지 유형으로 체계화하여 분류하는 프레임워크 제시.
  - 기존 텍스트 유사도 기반 평가 지표(ROUGE, BLEU)의 환각 탐지 한계를 실험적으로 증명.
- **PFVL 프로젝트: Agentic AI 기반 수학 문제 자동 해결 시스템**
  - LLM이 생성한 코드를 외부 컴파일러로 검증하고, 에러를 피드백받아 스스로 수정하는 Agentic 자동화 루프 구현.
  - 프롬프트 엔지니어링 전략만으로 LLM의 수학적 추론 신뢰도를 향상시키는 방법론 탐구.
- **Adele Embedding: 대수적 구조를 사용해 설계한 숫자 임베딩 개발**
  - LLM이 숫자의 수학적 의미를 잃는 문제를 해결하기 위해, Adele Space 개념을 활용한 새로운 임베딩 공간 설계.
  - 임베딩 공간 자체에 대수적 구조를 보존하여 훈련 없이도 모델의 근본적인 숫자 이해 능력을 향상

On-Device AI 연구 및 개발

- **SK매직 인턴십 - On-Device AI 신제품 기술 R&D**
  - 온디바이스 AI 탑재 신제품의 기술 R&D 전략 수립 및 사업성 검토 참여.
  - 핵심 자율 주행 알고리즘 설계 및 AI 제품 데이터 인프라 기획에 기여.
- **SCPC: AI Challenge - 온디바이스 VLM 개발 (본선 진출)**
  - 스마트폰 사양의 제한된 자원만으로 고성능 VLM을 개발하여 On-Device AI의 가능성을 직접 증명
  - 이미지 캡셔닝과 언어 추론을 분리한 'Two-Stage 파이프라인' 아키텍처 설계.
  - QLoRA, Flash Attention 등 최신 경량화 기술을 적용하여 제한된 자원 내 성능 극대화.
  - '커리큘럼 러닝' 전략을 도입하여 모델 훈련 안정성 확보 및 최종 본선 진출.

논문

출판 완료

- **Suhyun Bae** and Donghun Lee "Exploring Hallucination Types in Question-Answering Generation and Limitation of Text Evaluation Metrics" *KCC 2024*.

게재 승인

- Jaeheun Jung, Bosung Jung, **Suhyun Bae** and Donghun Lee "OPC: One-Point-Contraction Unlearning Toward Deep Feature Forgetting" *ICCV 2025 Workshop U&ME*.
- **Suhyun Bae** and Donghun Lee "Numbers Already Carry Their Own Embeddings" *NeurIPS 2025 Workshop MATH-AI*.

투고 준비

- **Suhyun Bae**, Dayeon Shin, and Donghun Lee "Code to Reason in Math is Still Hard for Well-known LLMs" *(Pre-print)*.

학력

2024 – 현재 **고려대학교 수학과 MDS 석사 과정**
2018 – 2024 **고려대학교 수학과 학사**

기술 스택

사용 언어 Python, LaTeX
프레임워크 PyTorch, HuggingFace
개발 도구 Linux, Git, Docker

최종 업데이트: October 21, 2025