JuHeon Ha

hjhjaueon97@gmail.com | juheon.dev | linkedin.com/in/hajuheon | github.com/hjhj97

Profile

My research focuses on **AI safety alignment** and **LLM interpretability**, with particular interests in persona, emotion, and social bias in LLMs. I study how LLMs represent emotions and social context, express empathy, and respond to persona prompts. Extending this work beyond empathy, I aim to understand how post-training, particularly reinforcement learning, shapes the internal mechanisms underlying safe and unsafe behavior. I also bring hands-on experience in LLM serving and backend development, alongside prior experience as a frontend engineer.

Education

Ajou University — **M.S. in Artificial Intelligence** Sep. 2025 – Present
LAMDA Lab · Advisor: Prof. Kyung-Ah Sohn

Ajou University — **B.S. in Software and Computer Engineering** Mar. 2017 – Aug. 2023

Publications

[1] **Empathy Is Steerable but Multi-Axial: Mechanism Geometry and Persona Effects in LLMs – [Paper]** ***EMNLP 2026*** *Main Conference*
JuHeon Ha, Byounghan Lee, Yunseo Choi, and Kyung-Ah Sohn.

[2] **Do LLMs Empathize with You, or with Your Words? Revealing How Stated Distress Makes a Friend's Loss Seem Yours** ***ACL Rolling Review*** Under review · *First author*

[3] **Nudge and Merge: Simple Post-Hoc Sparsification of Dense Adapters for Parameter-Efficient Fine-Tuning on Sparse LLMs** ***AAAI 2027*** Under review · *Second author*

Research Experience

Controlling LLM Behavior with Activation Steering
EMNLP 2026 Main Conference 2026

- Led study design, implementation, analysis, and writing; applied contrastive activation addition to Llama-3.1-8B, Qwen2.5-7B, and Mistral-7B to investigate how independently three dimensions of empathy can be controlled.
- Evaluated steering using EPITOME classifiers and an LLM judge; analyzed vector geometry to identify cross-dimension effects and persona-induced activation shifts outside the identified empathy subspace.

Controlled Evaluation of Context Attribution in LLMs
ACL Rolling Review, under review 2026

- Designed controlled scenarios separating event ownership from stated distress; measured attribution in generated responses, direct-question accuracy, and emotional-reaction activations.
- Validated response labels with lexical checks and blinded human evaluation, identifying context-attribution failures that warmth-oriented empathy metrics can miss.

Nudge and Merge: Sparse LLM Adapter Integration
AAAI 2027, under review 2026

- Co-authored Nudge and Merge, a post-hoc method for integrating pretrained dense adapters into sparse LLMs while preserving the backbone's sparsity pattern.
- The method improved average zero-shot accuracy over pruned backbones by 2.40 percentage points at 50% unstructured sparsity and 5.42 points at 2:4 sparsity, across seven benchmarks on LLaMA-7B/13B.

Projects

Patent Intelligence & M&A Recommendation Platform
Industry–University Collaboration Project Nov. 2025 – July 2026

- Built backend workflows for OCR, patent extraction, evaluation, and M&A recommendation; deployed DeepSeek-OCR-2 with vLLM on RunPod Serverless.
- Co-developed retrieval and ranking with embeddings and Pinecone, using industry and patent classification filters and quantitative signals from 124,000+ M&A records.
- Orchestrated asynchronous OCR and LLM jobs using FastAPI, Celery, Redis, S3, and Docker, with stage tracking, timeouts, failure handling, and result archiving.
- Improved OCR processing efficiency through batched inference and country-specific page selection; combined rule-based extraction with LLM interpretation for patent information extraction.

How Reinforcement Learning Shapes Safety Alignment in LLMs
Research project Oct. 2026 – Present

- Exploring how reinforcement learning shapes safety-related behaviors and internal representations in LLMs, focusing on refusal, harmful compliance, and reward hacking behavioral changes.
- Developing a framework combining RL post-training and activation interventions to investigate how safety-related behaviors emerge and identify their underlying mechanisms.

Predicting Counseling Intervention Strategies with LLMs
Research project May 2026 - Present

- Built a strategy-prediction benchmark from Korean counseling dialogues, auditing label ambiguity and overlap.
- Compared LLMs using only prior dialogue, varying context length and reasoning settings; analyzed label agreement and model-specific errors without exposing target counselor responses.

Measuring Persona-Conditioned Intergroup Gaps in LLMs
Research project June 2025 – Present

- Designed a baseline-adjusted emotion-attribution framework to quantify persona-conditioned gaps across gender, race, fandom, and political affiliation using 1,200 crowd-enVENT narratives.
- Generated and analyzed 600,000+ deterministic responses from Llama, Gemma, and Qwen; used bootstrap inference and controlled experiments to assess persona interactions and cue-order effects.

Work Experience

TAGby Company Seoul
Web Frontend Engineer Oct. 2023 – Aug. 2024

- Migrated landing pages to server-side rendering (SSR); these changes helped raise the Lighthouse performance score from 43 to 75.
- Reduced FCP and asset size by over 60% through resource cleanup and rendering optimization.
- Developed and published a browser fingerprinting module for detecting suspicious multi-account access as an open-source npm package.
- Migrated the build system to Vite, reducing build time and resolving deployment-server out-of-memory failures.

Activities

Teaching Assistant: *Computer Programming* (Fall 2025), *Compilers* (Spring 2026), and *Algorithms* (Fall 2026)

President, *A.N.S.I.* — Algorithm Problem-Solving Club (2020): Taught C programming and algorithms to members.

Problem Setter/Director: Ajou Programming Contest (*APC*), 2020 and 2021

Programming competition: *ACM-ICPC Seoul Regional* - 51st Place (2018)

Patents

희소 언어 모델의 어댑터 통합 장치 및 방법 (Apparatus and Method for Integrating an Adapter into a Sparse Language Model)

Inventors: Kyung-Ah Sohn, Hyun-woo Jeong, **JuHeon Ha**, Tae Gyu Yang, and Dongjun Kim.

KR Patent Application: 10-2026-0196669, Filed: 2026-10-13