

JuHeon Ha (하주헌)

LLM Research · Interpretability · Human-Centered AI

hjhjaeon97@gmail.com | [LinkedIn](#) | github.com/hjhj97 | juheon.dev

PROFILE

My research focuses on [LLM interpretability](#), [human-centered AI](#), [AI safety](#) and [alignment](#), with particular interests in persona, emotion, and social bias in LLMs. I explore how AI can model human emotion and behavior, and the underlying model mechanisms that shape human-like behavior. Before beginning my graduate studies, I worked as a web developer.

EDUCATION

Ajou University — M.S. in Artificial Intelligence

LAMDA Lab · Advisor: Prof. Kyung-Ah Sohn

Sep. 2025 – Present

Ajou University — B.S. in Software Engineering

Suwon, Republic of Korea

Mar. 2017 – Aug. 2023

PUBLICATIONS & MANUSCRIPTS

1. Empathy Is Steerable but Multi-Axial: Mechanism Geometry and Persona Effects in LLMs

JuHeon Ha, Byoungchan Lee, Yunseo Choi, and Kyung-Ah Sohn.

EMNLP 2026 Main Conference - [Code](#) | [Project page](#)

2. Do LLMs Empathize with You, or with Your Words? Revealing How Stated Distress Makes a Friend's Loss Seem Yours

First author · Under review — *ACL Rolling Review*.

3. Nudge and Merge: Simple Post-Hoc Sparsification of Dense Adapters for Parameter-Efficient Fine-Tuning on Sparse LLMs

Second author · Under review — *AAAI 2027*.

RESEARCH INTERESTS

LLM interpretability | Behavioral evaluation | Human-centered AI (persona, empathy, social bias)

RESEARCH EXPERIENCE

Controlling LLM Behavior with Activation Steering

First author · EMNLP 2026 Main Conference

2026

- Led experimental design, implementation, analysis, and manuscript writing to investigate whether empathy dimensions can be controlled independently and whether persona prompts use the same internal directions.
- Implemented contrastive activation addition and layer/strength sweeps on Llama-3.1-8B, Qwen2.5-7B, and Mistral-7B. Evaluated emotional reaction, interpretation, and exploration with EPITOME classifiers and a separate LLM judge.
- Analyzed vector geometry, residualization, and persona-induced activation shifts. Established cross-dimension steering effects and distinguished the recovered empathy subspace from broader persona effects.
- Released reproduction instructions, aggregate results, and an interactive project page to support independent inspection of the findings.

Controlled Evaluation of Context Attribution in LLMs

First author · ACL Rolling Review, under review 2026

- Led experimental design, implementation, analysis, and writing. Constructed controlled scenarios with known event ownership, separating who experienced an event from the speaker's stated distress.
- Built complementary measurements of ownership attribution in generated replies, direct-question accuracy, and emotional-reaction activations. Validated LLM-judge labels with an independent lexical check and blinded human evaluation.
- Separated emotional warmth from accurate situational understanding and reported prespecified tests separately from exploratory analyses, identifying failures that warmth-oriented evaluation can miss.

Persona-Conditioned Intergroup Gaps in LLMs

Research project · LAMDA Lab, Ajou University Jun. 2025 – Present

- Extended an existing research codebase and evaluation protocol with baseline-adjusted emotion attribution across gender, race, fandom, and political affiliation.
- Used vLLM batch inference to compare Llama, Gemma, and Qwen under controlled persona conditions; applied bootstrap inference to investigate relational, non-additive, and cue-order effects.

Nudge and Merge: Sparse LLM Adapter Integration

Second-author research collaboration · AAAI 2027, under review 2026

- Co-authored a manuscript on post-hoc sparsification of trained dense adapters for sparse LLMs. The method adjusts adapter updates within the orthogonal complement of calibration activations before merging, without changing the adapter training procedure.

TEACHING & ACADEMIC ACTIVITIES

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

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

ENGINEERING PROJECTS

Patent Intelligence & M&A Recommendation Platform

Backend and OCR serving owner · Analysis service co-developer Nov. 2025 – Present

Python · FastAPI · Celery · Redis · AWS S3 · Docker · vLLM · RunPod · Pinecone

- Independently built backend orchestration and OCR serving; co-developed the analysis service. Connected patent PDF ingestion, structured extraction, industry classification, evaluation, and company recommendation across three deployed services.
- Deployed DeepSeek-OCR-2 with vLLM and FlashAttention on RunPod Serverless; used batched inference and country-aware page selection to control processing of long patent documents.
- Combined deterministic INID/regex parsing and claim-dependency graphs with LLM interpretation. Built company retrieval and ranking using embeddings, Pinecone, NAICS/IPC filters, company-level aggregation, and fallback rules.
- Implemented asynchronous job orchestration with stage-level state tracking, request-ID logs, worker timeouts, error mapping, and result archiving.

Work EXPERIENCE

TAGby Company — Frontend Engineer

SaaS · Vue 3, Nuxt 3, TypeScript, Vite Oct. 2023 – Aug. 2024

- Migrated landing pages from static generation to server-side rendering and optimized script/font loading. Reduced page load time from 2.1 s to 0.8 s and improved Lighthouse Performance from 43 to 75.
- Migrated the build system from Webpack to Vite, reducing build time from 15 min to 2 min 30 s and resolving deployment-server out-of-memory build failures.
- Proposed and implemented canvas fingerprinting for shared-account abuse analysis; analyzed 12,623 logins over three months and released a related font-based identification module as the npm package font-fingerprint.