LinkedIn: linkedin.com/in/jaeha-lee-323279315
GitHub: github.com/Jaeha0526
jaeha@caltech.edu

JAEHA LEE

Theoretical physics Ph.D. candidate with a strong mathematical foundation and hands-on expertise in LLM development. Experienced in building agentic systems, reinforcement learning training, and AI alignment. Passionate about bridging physics and AI applications while fostering interdisciplinary collaboration. Proven ability to drive independent, innovative research initiatives.

EDUCATION

Ph.D. in physics at Caltech ***Sep 2020 - present***
Advisor: Professor Hirosi Ooguri
B.S. in physics at Seoul National University ***Mar 2014 - Aug 2020***[*]
Overall GPA: 4.01 / 4.30, Physics: 4.11 / 4.30 (summa cum laude)

TECHNICAL SKILLS

Programming Languages: Python, Mathematica
Machine Learning: LLM training & fine-tuning (GRPO, PPO, DPO, AlphaZero-style self-play based training), Representation Engineering (RepE), Graph Neural Networks (GNN), Attention mechanism analysis, Sparse Autoencoders
Agentic System: Agentic system development (CrewAI, LangChain), LLM scaffolding for tool use (code execution, command line), SWE-bench evaluation, Automated red-teaming & jailbreak analysis

WORKING EXPERIENCE

- **Research Intern at Morph Labs** *Jan 2025 - Mar 2025*
  - Used reinforcement learning techniques to train language models for tool use and reasoning.
  - Developed training pipelines using imitation learning and policy optimization methods on coding benchmarks.

PROJECTS IN AI

- **An Agentic System for Performing Representation Engineering for Character Control** *May 2025 - present*
  - Developed an automated multi-agent system using CrewAI framework to systematically control LLM personality traits through representation engineering (RepE). (Link to the repo, paper)
  - Extracted and analyzed emotion concept vectors from Meta-Llama-3-8B across 42 distinct emotions, identifying consistent representation patterns in middle transformer layers (cosine similarity > 0.7).
  - Implemented hybrid vector library system combining fresh extractions with pre-trained vectors, demonstrating 0.3-0.5 point improvement in multi-dimensional effectiveness metrics.
  - Integrated NeMo Guardrails for safety-constrained generation, creating a production-ready framework for responsible LLM character steering.
- **Enhancing Multi-Turn Human Jailbreaks Dataset for Improved LLM Defenses** (Work in Progress) *Jan 2025 - present*
  - Extended and enhanced the Multi-Turn Human Jailbreaks (MHJ) dataset introduced by Li et al..
  - Developed automated multi-turn attacks, and conducted qualitative analysis with StrongREJECT evaluator.
  - This project is one of the AI Safety Camp 10th edition.
- **Using AlphaZero style training to bound the Ramsey number** (Work in Progress) *Mar 2025 - present*
  - Training GNN in AlphaZero setup to play clique game well to bound the Ramsey number. (link to the github)
- **Discovering Hidden Algebraic Structures via Transformers with Rank-Aware Beam GRPO** *Mar 2024 - May 2025*
  (Accepted to MOSS@ICML workshop) (link to the github)
  - Developed novel Beam Grouped Relative Policy Optimization (BGRPO) algorithm incorporating rank-aware rewards, reducing beam search width by 50% while maintaining accuracy. (75% computational savings)
  - Built synthetic data generation pipeline for multivariate polynomial decomposition (NP-hard problem), training transformers that achieved 100% accuracy on degree-4 polynomials with beam search.
  - Demonstrated rapid distribution adaptation requiring only 2% of original training data to achieve 90%+ accuracy on new coefficient ranges, validating generalizable algebraic understanding.
  - Outperformed Mathematica's FullSimplify in polynomial simplification by 14.5% average DAG vertex reduction across multiple complexity regimes.

[*] Military Service at Korean Air Force 19th-Fighter Wing (Jan 2016- Jan 2018)

SELECTED PROJECTS IN PHYSICS

- **Discovering universal formulas for high energy data in Conformal Field Theory** *Sep 2020 - May 2024*
  - Link to publications [1], [2] and [3].
  - Developed a universal formula for Conformal Field Theory (CFT) using the thermal effective action method, a novel approach that has broadened the understanding of asymptotic behavior in CFT and inspired subsequent research.
- **Suggesting novel solution for an endpoint of the Kerr-AdS superradiant instability** *2019 - 2023*
  - Link to the publication.
  - Calculated the thermodynamic properties of the Revolving Black Hole and elucidated its mechanism as a potential endpoint of the Kerr-AdS superradiant instability. This work provided new insights into hairy black hole solutions.

COURSES

- **CS 159: Large Language Models for Reasoning @ Caltech** *Spring 2025*
  - Topics: Use LLM within agentic workflow
- **AI Safety Fundamentals AI Alignment Course** *Aug 2024 - Nov 2024*
  - Topics: AI alignment, RLHF, Scalable oversight, and Mechanistic interpretability
- **EE/CNS/CS 148: Large Language and Vision Models @ Caltech** *Spring 2024*
  - Topics: MLPs, ResNets, ViTs, Diffusion Models, and Neural Image Generation

HONORS AND SCHOLARSHIPS

- James A. Cullen Memorial Fellowship Fund *Jun 2021*
- Presidential Science Scholarship, Korea Student Aid Foundation *2014-2015, 2018-2020*
- Gold medal, 44th International Physics Olympiad, Copenhagen, Denmark *Jul 2013*
- Science Gifted Student Scholarship, Woongjin Foundation *2011 - 2013*