JIAHAI FENG

https://jiahai-feng.github.io/ fengjiahai@gmail.com

EDUCATION

University of California, Berkeley *Sep 2023 -*
PhD in Computer Science

Massachussetts Institute of Technology *Sep 2019 - May 2023*
Bachelor of Science in Computation & Cognition
Bachelor of Science in Physics
Minor in Mathematics
GPA: 5.0/5.0

AWARDS AND HONORS

OpenAI Superalignment Fellowship 2024
Pi Beta Kappa National Honor Society *2023*
International Physics Olympiad Gold — Globally ranked 24th/398 *2016*
International Olympiad in Informatics Gold — Globally ranked 24th/311 *2014*

PUBLICATIONS

How do Language Models Bind Entities in Context?
J. Feng, J. Steinhardt.
In *International Conference on Learning Representations, 2024*

Learning Grounded Action Abstractions from Language
L. Wong, J. Mao, P. Sharma, Z. Siegel, **J. Feng**, N. Korneev, J. Tenenbaum, J. Andreas
In *International Conference on Learning Representations, 2024*

Structured, flexible, and robust: benchmarking and improving large language models towards more human-like behavior in out- of-distribution reasoning tasks
K. Collins, C. Wong, **J. Feng**, M. Wei, J. Tenenbaum.
In *CogSci, 2022*

AI Feynman 2.0: Pareto-optimal symbolic regression exploiting graph modularity
S. Udrescu, A. Tan, **J. Feng**, O. Neto, T. Wu, M. Tegmark.
In *Proc. NeurIPS, 2020.*

RESEARCH EXPERIENCE

Language & Intelligence @ MIT *September 2022 - May 2023*
- Advised by Prof. Jacob Andreas, Pratyusha Sharma, Catherine Wong, and Jiayuan Mao
- Explored the use of large language models as priors for abstraction learning in bilevel planning

Center for Human-Compatible AI *June 2022 - August 2022*
- Advised by Prof. Stuart Russell and Scott Emmons
- Theoretical analysis of power-seeking reinforcement learning agents

MIT Cocosci Lab *October 2021 - January 2022*
- Advised by Prof. Josh Tenenbaum, Catherine Wong and Katherine Collins
- Explored the use of large language models for reasoning in the planning domain

PRESENTATIONS

[Talk] Institute for Artificial Intelligence and Fundamental Interactions Reading Group *February 2022*
- Presented on unpublished work I did at Redwood Research

[Talk] Summer MIT Kavli Institute Undergraduate Research Forum (SMURF) *August 2020*
- Presentation on the AI Feynman project

INDUSTRY EXPERIENCE

Redwood Research *Dec 2021 - January 2022*
Machine Learning Intern Berkeley, CA
- Worked on mechanistic interpretability of transformer-based language models

Jane Street Capital *June 2021 - August 2021*
Quantitative Research Intern New York, NY
- Worked on quantitative research projects studying US equities market microstructure and robust linear regression.

SELECTED COURSEWORK

Artificial Intelligence Advanced Machine Learning (6.867) • Representation, Inference and Reasoning in AI (6.S058) • Computational Cognitive Science (9.66) • Doing Things with Words (6.884)
Statistics Information and Inference (6.437) • Stochastic Processes (18.615) • Algorithms for Inference (6.438)
Programming Programming with Categories (18.S097) • Large Scale Symbolic Systems (6.905) • Introduction to Program Synthesis (6.S081)
Mathematics Abstract Algebra (18.701) • Functional Analysis (18.102) • Eigenvalues of Random Matrices (18.338) • Market Design (14.19)
Theoretical Computer Science Advanced Algorithms (6.854) • Theory of Computation (18.404)
Physics Statistical Mechanics (8.333) • Quantum Information Science (8.371)

TEACHING & SERVICE

Teaching Assistant *Fall 2022*
- TA for Representation, inference and reasoning in AI (6.4110)

HKN Tutor *Spring 2021*
- Tutored MIT students in Design & Analysis of Algorithm (6.046) and Mathematics for Computer Science (6.042)

MIT Physics Peer Mentor *Spring 2021 & 2022*
- Mentored underclassmen in Quantum Physics and Statistical Mechanics

National Team Coach for Informatics Olympiad *Summer 2019*
- Organized the Singapore national training program for International Olympiad in Informatics 2019
- Managed logistics, planned and taught lectures, sourced training problems, and coordinated with Saudi Arabian, Malaysian, Indonesian and Vietnamese teams for joint training sessions

Developer for Notes Sharing Website *2017 - 2023*
- Developed a free notes-sharing website *https://tick.ninja* for high school students in Singapore to make education accessible to all

SKILLS

Languages:	English (fluent), Mandarin Chinese (fluent)
Programming:	**Proficient:** Python, C++, Typescript, Mathematica
	Working knowledge: Go, Haskell, Scheme
Data Analysis:	Pytorch, Jax, Pandas, Seaborn, Excel, SQL, Stan
Web Dev:	React, CSS, Flask

EXTRACURRICULARS

MIT AI Ethics Reading Group *Spring 2022*
- Managed communications for the AI Ethics reading group at MIT
- The group met biweekly with invited speakers

AI Alignment *2019 - 2023*
- Organized a reading group on *Human Compatible* in January 2020
- Organized a reading group on plausibility of existential risks from AI in January 2022
- Organized an interpretability workshop at MIT in January 2023

Traders@MIT *2019 - 2022*
- Organized the annual trading competition
- Over a hundred participants from across the country
- Designed and built electronic trading cases

MIT IEEE/ACM Student Chapter *2020 - 2022*
- As Faculty Students Relations Chair, organized and hosted fireside chats with faculty members
- As Secretary, managed and coordinated operations

MIT DanceTroupe *2021 - 2023*
- Dancer in student-run hip hop and jazz fusion choreography productions