Jing-Jing Li

jl3676@berkeley.edu
http://jl3676.github.io
Google Scholar

Research Summary

AI researcher and computational cognitive scientist bridging alignment, human cognition, and AI safety. Develops **interpretable, pluralistic, and steerable frameworks** for **LLM safety and agentic robustness**, integrating **reasoning, synthetic data, and human-grounded evaluation**.

Education

University of California, Berkeley — Berkeley, CA
Ph.D. in Neuroscience with concentrations in Computation and Cognition, GPA: 3.97/4.00 — 2021–2026

Cornell University — Ithaca, NY
B.A. in Computer Science and Mathematics, Minor in Cognitive Science, GPA: 4.01/4.30 — 2017–2020

Work and Research Experiences

Anthropic — Berkeley, CA
Research Fellow — November 2025 – May 2026

- Leading a research project on alignment and safety.

Amazon Web Services (AWS) — Agentic AI — Seattle, WA
Applied Scientist Intern — May 2025 – August 2025

- Led research on **adversarial robustness** of tool-enabled LLM agents via multi-turn attack–defense evaluation.
- Built scalable pipelines for **synthetic data generation and context engineering** in agent benchmarking.
- Produced a **first-authored publication** and open-source benchmark on agentic AI safety.

Allen Institute for Artificial Intelligence (Ai2) — Seattle, WA
Research Intern — May 2024 – August 2024

- Built **SafetyAnalyst**, an interpretable and steerable LLM safety moderation framework.
- Designed multi-stage pipeline combining **reasoning, synthetic data, and model distillation.**
- Resulted in a first-author **ICML 2025 paper** and open-sourced models, data, and code suite.

University of California, Berkeley — Berkeley, CA
Ph.D. Researcher, Computational Cognitive Neuroscience Lab (Advisor: Anne Collins) — 2021 – 2026 (expected)

- Combines **reinforcement learning, Bayesian inference, and neural data** to model adaptive behavior.
- Leads behavioral and modeling studies on **hierarchy, compositionality, exploration, and transfer**.
- Created **latent state estimation frameworks** for cognitive model fitting.
- Published in top journals including **Cognition** and **Journal of Mathematical Psychology**.

Technical Skills

- **Large Language Models:** Prompt Engineering, Synthetic Data Generation, Fine-Tuning, Evaluation, Red-Teaming, Model Behavior Analysis, Human-Centered Alignment, Crowdsourcing Pipelines
- **Machine Learning:** PyTorch, Hugging Face, TensorFlow, CUDA, Reinforcement Learning, Bayesian Modeling
- **Programming & Infrastructure:** Python, R, C/C++, Java, Bash, Git, Linux, HPC/Cluster Environments
- **Data Analysis:** NumPy, pandas, SciPy, Matplotlib, Regression Modeling, SQL, MATLAB

Grants and Fellowships

- UC Berkeley ICBS Grant ($5,000; *Co-recipient*) 2024–2025
- Society for Neuroscience Trainee Professional Development Award 2024
- CogSci Conference Travel Grant 2023
- Milton I. and Florence Mack Neurology Research Fund 2021–2022
- Summer Undergraduate Research Fellowship, Caltech 2018

Selected Publications

[1] **J.-J. Li**, J. Mire, E. Fleisig, V. Pyatkin, A. Collins, M. Sap, and S. Levine, *Pluriharms: Benchmarking the full spectrum of human judgments on ai harm*, 2026. arXiv: 2601.08951 [cs.CY].

[2] **J.-J. Li**, J. He, C. Shang, D. Kulshreshtha, X. Xian, Y. Zhang, H. Su, S. Swamy, and Y. Qi, *STAC: When innocent tools form dangerous chains to jailbreak LLM agents*, 2025. arXiv: 2509.25624 [cs.CR].

[3] **J.-J. Li**, V. Pyatkin, M. Kleiman-Weiner, L. Jiang, N. Dziri, A. G. E. Collins, J. S. Borg, M. Sap, Y. Choi, and S. Levine, "SafetyAnalyst: Interpretable, transparent, and steerable safety moderation for AI behavior", in *ICML*, 2025.

[4] **J.-J. Li** and A. G. Collins, "An algorithmic account for how humans efficiently learn, transfer, and compose hierarchically structured decision policies", *Cognition*, vol. 254, p. 105 967, 2025.

[5] **J.-J. Li**, C. Chen, and A. G. Collins, "Humans integrate heuristics and bayesian inference to efficiently explore under uncertainty", in *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2025.

[6] T.-F. Pan, **J.-J. Li**, B. Thompson, and A. GE Collins, "Latent variable sequence identification for cognitive models with neural network estimators", *Behavior Research Methods*, vol. 57, no. 10, p. 272, 2025.

[7] **J.-J. Li**, C. Shi, L. Li, and A. G. Collins, "Dynamic noise estimation: A generalized method for modeling noise fluctuations in decision-making", *Journal of Mathematical Psychology*, vol. 119, p. 102 842, 2024.

[8] **J.-J. Li**, L. Xia, F. Dong, and A. G. Collins, "Credit assignment in hierarchical option transfer", in *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2022.

Selected Talks & Presentations

- **AI Agent Safety Social Panel**, ICML 2025 (Invited Panel), Vancouver, Canada
- **Interpretable LLM Safety Moderation**, ICML 2025 (Poster), Vancouver, Canada
- **Interpreting Human Judgments on AI Harm**, NeurIPS CogInterp Workshop 2025 (Poster), San Diego, CA
- **Heuristics and Bayesian Inference for Efficient Exploration**, CogSci 2025 (Talk), San Francisco, CA
- **Dynamic Noise Modeling in Decision-Making**, Cognitive & Computational Neuroscience in Development Group 2024 (Invited Talk), Würzburg, Germany
- **A Generalized Method for Dynamic Noise Inference**, CogSci Conference 2023 (Talk), Sydney, Australia
- **Credit Assignment in the Transfer of Hierarchical Options**, CogSci Conference 2022 (Talk), Toronto, Canada