Jing-Jing Li

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

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 Experiences

Amazon Web Services Agentic AI — Seattle, WA
Applied Scientist Intern — May 2025 - August 2025

- Led a research project on the adversarial robustness of tool-enabled LLM agents.
- Performed context engineering, synthetic data generation, and evaluation on diverse LLM agents.

Allen Institute for Artificial Intelligence — Seattle, WA
Research Intern — May 2024 - August 2024

- Developed a framework to improve the interpretability, transparency, and steerability of AI safety moderation.
- Performed prompt engineering, taxonomy development, crowdsourcing, synthetic data generation, model distillation, supervised fine-tuning, and evaluation on LLMs.

Skills

- **Large Language Models:** Prompt Engineering, Supervised Fine-Tuning, Crowdsourcing
- **Machine Learning:** Pytorch, TensorFlow, scikit-learn, CUDA, Hugging Face
- **Programming:** Python, Java, C, C++, Bash, Shell, HTML, CSS, JavaScript, GitHub
- **Data Science:** Numpy, SciPy, pandas, Matplotlib, R, MATLAB, SQL
- **Other:** LaTeX, Adobe Illustrator, Adobe Photoshop, Linux, Microsoft Excel

Relevant Courses

- **Machine Learning:** Deep Unsupervised Learning, LLMs and Cognition, Deep Reinforcement Learning, Computer Vision, Large-Scale Machine Learning, Intro to Machine Learning, Computational Genetics
- **Software Engineering:** Data Structures (Honors), Algorithms, Operating Systems, Database Systems
- **Mathematics and Statistics:** Numerical Analysis, Biological Statistics, Probability Theory, Abstract Algebra, Linear Algebra (Honors), Multi-variable Calculus

Grants and Fellowships

- UC Berkeley ICBS Grant ($5,000; *Co-recipient with Eve Fleisig*) — 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

Publications

[1] **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.

[2] **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.

[3] **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.

[4] 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.

[5] J. Chase, **J.-J. Li**, W. C. Lin, L.-H. Tai, A. G. Collins, and L. Wilbrecht, "Genetic changes linked to two different syndromic forms of autism enhance reinforcement learning in adolescent male but not female mice", *bioRxiv*, pp. 2025–01, 2025.

[6] **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.

[7] D. S. Jin, O. Agdali, T. Yadav, S. I. Kronemer, S. Kunkler, S. Majumder, M. Khurana, M. C. McCusker, I. Fu, A. Khalaf, K. L. Christison-Lagay, S. L. Aerts, Q. Xin, **J.-J. Li**, S. H. McGill, M. J. Crowley, and H. Blumenfeld, "Neural mechanisms of awareness of action", *bioRxiv*, 2024.

[8] **J.-J. Li**, C. Shi, L. Li, and A. G. Collins, "A generalized method for dynamic noise inference in modeling sequential decision-making", in *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2023.

[9] C. McCafferty, B. F. Gruenbaum, R. Tung, **J.-J. Li**, X. Zheng, P. Salvino, P. Vincent, Z. Kratochvil, J. H. Ryu, A. Khalaf, K. Swift, R. Akbari, W. Islam, P. Antwi, E. A. Johnson, P. Vitkovskiy, J. Sampognaro, I. G. Freedman, A. Kundishora, A. Depaulis, F. David, V. Crunelli, B. G. Sanganahalli, P. Herman, F. Hyder, and H. Blumenfeld, "Decreased but diverse activity of cortical and thalamic neurons in consciousness-impairing rodent absence seizures", *Nature Communications*, vol. 14, no. 1, pp. 1–19, 2023.

[10] **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.

[11] J. Ding, **J.-J. Li**, and M. Xu, "Classification of murmurs in pcg using combined frequency domain and physician inspired features", in *2022 Computing in Cardiology (CinC)*, IEEE, vol. 498, 2022, pp. 1–4.

Presentations

Invited talks

AI Agent Safety Social Panel — Vancouver, Canada
ICML 2025 — July 2025

Dynamic noise modeling in decision-making — Uniklinikum Würzburg
Cognitive and Computational Neuroscience in Development Psychiatry Research Group — June 2024

Conference talks

Humans integrate heuristics and Bayesian inference to efficiently explore — San Francisco, CA
CogSci Conference — July 2025

Dynamic noise modeling in decision-making — Tahoe, CA
Berkeley Neuroscience Conference — October 2023

A generalized method for dynamic noise inference — Sydney, Australia
CogSci Conference — July 2023

Credit assignment in the transfer of hierarchical options — Toronto, Canada
CogSci Conference — July 2022

Conference posters

Interpretable, transparent, and steerable LLM safety moderation — Vancouver, Canada
ICML — July 2025

🔦 **Humans Integrate heuristics and Bayesian inference to efficiently explore** — Dublin, Ireland
RLDM Conference (spotlight) — June 2025

Interpretable, transparent, and steerable LLM safety moderation — Vancouver, Canada
NeurIPS SoLaR Workshop — December 2024

Modeling how humans learn, transfer, and compose hierarchical policies — Chicago, IL
Society for Neuroscience Conference — October 2024

Modeling the emergence of instrumental learning in an odor-based 2AFC task — Boston, MA
Cognitive Computational Neuroscience Conference — August 2024

Modeling how humans learn, transfer, and compose hierarchical policies — Boston, MA
Cognitive Computational Neuroscience Conference — August 2024

Credit assignment in the learning and transfer of hierarchical options — San Francisco, CA
Cognitive Neuroscience Society Conference — April 2022