

## EDUCATION

<b>University of California, Berkeley</b> Ph.D. in Neuroscience with concentrations in Computation and Cognition, GPA: 3.96/4.00	Berkeley, CA 2021–2026
<b>Cornell University</b> B.A. in Computer Science and Mathematics, Minor in Cognitive Science, GPA: 4.01/4.30	Ithaca, NY 2017–2020

## WORK EXPERIENCE

<b>Allen Institute for Artificial Intelligence</b> PhD Research Intern	Seattle, WA May 2024 - August 2024
---	---------------------------------------

- Developed a system to improve the interpretability, transparency, and controllability of LLM safety moderation.
- Performed prompt engineering, taxonomy development, batched inference, crowdsourcing, symbolic knowledge distillation, supervised fine-tuning, and evaluation on large language models (LLMs).

## PUBLICATIONS

- [1] **J.-J. Li**, V. Pyatkin, M. Kleiman-Weiner, L. Jiang, N. Dziri, A. Collins, J. S. Borg, M. Sap, Y. Choi, and S. Levine, “SafetyAnalyst: Interpretable, transparent, and steerable LLM safety moderation”, *In submission to ICLR*, 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. 105967, 2025.
- [3] **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. 102842, 2024.
- [4] T.-F. Pan, **J.-J. Li**, B. Thompson, and A. Collins, *Latent variable sequence identification for cognitive models with neural bayes estimation*, 2024. arXiv: 2406.14742 [cs.LG].
- [5] 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.
- [6] **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.
- [7] 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.
- [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.
- [9] 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.

## SKILLS

---

- **Programming:** Python, Java, Julia, C, C++, Swift, Bash, Shell, OCaml
- **Data Science:** R, Numpy, SciPy, pandas, Matplotlib, Seaborn, MATLAB, SPM, FSL
- **Machine Learning:** TensorFlow, scikit-learn, PyTorch, OpenAI Gym, MuJoCo, CUDA, Kaggle, Google Colab
- **Natural Language Processing:** Large Language Models, Prompt Engineering, Fine-Tuning, Crowdsourcing
- **Experimental Design:** PsychoPy, Psychtoolbox, jsPsych, Amazon MTurk, EEGLAB, Persyst, EyeLink
- **Operating Systems:** Linux, Unix, Windows
- **Web Development:** HTML, CSS, JavaScript, Heroku
- **Database Management:** SQL, Microsoft Excel, RAID
- **Other:** LaTeX, Adobe Illustrator, Adobe Photoshop, GitHub

## RELEVANT COURSES

---

- **Machine Learning:** Deep Unsupervised Learning, LLMs and Cognition, Deep Reinforcement Learning, Computer Vision, Intro to Machine Learning, Large-Scale Machine Learning, Computational Genetics
- **Software Engineering:** Object-Oriented Design and Data Structures (Honors), Algorithms, Computational Problem Solving, Operating Systems, Database Systems, Database Systems Practicum
- **Mathematics and Statistics:** Numerical Analysis, Biological Statistics, Basic Probability, Applicable Abstract Algebra, Linear Algebra (Honors), Multi-variable Calculus
- **Neuroscience:** Methods in Computational Modeling for Cognitive Science, Computational Psychology, Clinical Neuroscience, Developmental Psychology, Biopsychology, Cellular and Developmental Neuroscience

## SCHOLARSHIPS AND AWARDS

---

- |   |           |
|---|-----------|
| • 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      |

## PRESENTATIONS

---

### Invited talks

<b>Dynamic noise modeling in decision-making</b>	Uniklinikum Würzburg
Cognitive and Computational Neuroscience in Development Psychiatry Research Group	June 2024

### Conference talks

<b>Dynamic noise modeling in decision-making</b>	Tahoe, CA
Berkeley Neuroscience Conference	October 2023
<b>A generalized method for dynamic noise inference</b>	Sydney, Australia
CogSci Conference	July 2023
<b>Credit assignment in the transfer of hierarchical options</b>	Toronto, Canada
CogSci Conference	July 2022

## Conference posters

<b>Modeling how humans learn, transfer, and compose hierarchical policies</b> Society for Neuroscience Conference	Chicago, IL October 2024
<b>Modeling the emergence of instrumental learning in an odor-based 2AFC task</b> Cognitive Computational Neuroscience Conference	Boston, MA August 2024
<b>Modeling how humans learn, transfer, and compose hierarchical policies</b> Cognitive Computational Neuroscience Conference	Boston, MA August 2024
<b>Credit assignment in the learning and transfer of hierarchical options</b> Cognitive Neuroscience Society Conference	San Francisco, CA April 2022