Yuxuan (Effie) Li

liyuxuan@google.com | https://effie-li.github.io

- Research focus: machine and human cognition, mechanistic interpretability

Education

2019 – 2024 **Stanford University, PhD** in Cognitive Psychology.

2013 – 2017 **Trinity College, BS** in Computer Science and Psychology. *summa cum laude.*

Research Positions

2025 - **Research Scientist** @ Google DeepMind

2024 summer **Research Intern** @ Meta

2023 summer **Research Intern** @ Allen Institute for AI

2017 – 2019 **Research Specialist** @ UPenn

2016 summer **Research Intern** @ Columbia Business School

Projects and Publications

Visual reasoning and agentic planning

2025 Wiedemer, T., **Li, Y.**, Vicol, P., Gu, S., Matarese, N., Swersky, K., Kim, B., Jaini, P., & Geirhos, R. Video models are zero-shot learners and reasoners. *paper*, *website*

2023 **Li, Y.**, & Weihs, L. Understanding representations pretrained with auxiliary losses for embodied agent planning. *NeurIPS 2023 Generalization in Planning Workshop. paper*

Learning, generalization, and interpretability of transformers and language models

2025 **Li, Y.**, Campbell, D., Chan, S., & Lampinen, A. Just-in-time and distributed task representations in language models. *NeurIPS 2025 Mechnistic Interpretability Workshop (spotlight). paper*

2025 **Li, Y.**, & McClelland, J.L. Learning to decompose: Human-like subgoal preferences emerge in transformers learning graph traversal. *Under review.*

2023 **Li, Y.**, & McClelland, J.L. Representations and computations in transformers that support generalization on structured tasks. *Transactions on Machine Learning Research. paper*, *code*

Computational modeling of human behavior and neural signals

2024 **Li, Y.**, Pazdera, J.K., & Kahana, M.J. EEG decoders track memory dynamics. *Nature Communications. paper, code*

2023 Kahana, M.J., Lohnas, L.J., Healey, K., . . ., **Li, Y.**, . . ., & Weidemann, C.T. The Penn Electrophysiology of Encoding and Retrieval Study. *JEP: LMC*. *paper*

2022 **Li, Y.**, & McClelland, J.L. A weighted constraint satisfaction approach to human goal-directed decision making. *PLOS Computational Biology*. *paper*, *code*

2022 Katerman, B.S., **Li, Y.**, Pazdera, J.K., Keane, C., & Kahana, M.J. EEG biomarkers of free recall. *NeuroImage*. *paper*

2018 Grubb, M.A., & **Li, Y.** Assessing the role of accuracy-based feedback in value-driven attentional capture. *Attention, Perception, & Psychophysics*. *paper*

Talks and Presentations

Dec 2024 **Li, Y.** Emergent task decomposition and subgoal choices in transformers. *Mind, Brain, Computation and Technology Seminar Series, Stanford University.*

Mar 2024 **Li, Y.** Emergent structured computation from learning and its implications for cognitive science and AI. *Microsoft Research Lab, Redmond.*

Nov 2023 **Li, Y.** Systematic generalization and emergent structures in transformers trained on structured tasks. *FriSem seminar, Department of Psychology, Stanford University.*

Apr 2022 **Li, Y.** A weighted constraint satisfaction approach to human goal-directed decision making. *Cognitive Tools Lab, University of California, San Diego.*

Feb 2021 **Li, Y.** Model-based reinforcement learning and the reinforcement learning framework for human behavior. *TA Lecture in PSYCH 209, Stanford University.*

2020, 2021 **Li, Y.** Building online psychology experiments with jsPsych: a tutorial. *TA Lecture in PSYCH 251, Stanford University.*

Honors and Awards

2022 – 2024 Ric Weiland Graduate Fellowship in the Humanities & Sciences. Stanford University.

2013 – 2017 Phi Beta Kappa, Dean's Scholar (top 5%), Faculty Honors, Holland Scholar. Trinity College.

Teaching and Services

Reviewer NeurIPS, CVPR, TMLR, CogSci, CCN

TA Neural network models of cognition, brain decoding, Experimental methods, developmental psychology, introduction to computing, mathematical foundations of computing

Technical Skills

Programming **Python, R,** some experience with HTML/CSS/JavaScript

Packages **LLM** *(langchain),* **deep learning** (transformers, pytorch, pytorch-lightning, allenact, einops), **experiment management** (wandb)**, machine learning** (scikit-learn), **data analysis** (scipy, numpy, pandas), **data visualization** (matplotlib), **cognitive (neuro)science** (mne, ptsa)

Other LaTeX, statistics (linear modeling, generalized linear modeling, mixed-effects models), representation analysis, online behavioral platforms (Prolific)