Kenneth Li

Website: https://likenneth.github.io
Contact: ke_li@g.harvard.edu

Education

Harvard University — *Cambridge, MA*
Ph.D. candidate in Computer Science — *Sept. 2021 - May 2025 (tentative)*
Advisor: Prof. Martin Wattenberg and Prof. Hanspeter Pfister

University of Chinese Academy of Sciences — *Beijing, China*
Bachelor of Engeering in Computer Science — *Sept. 2017 - May 2021*
Supervisor: Prof. Xilin Chen

Publication

○ **What Does it Mean for a Neural Network to Learn a "World Model"?**
Kenneth Li, Fernanda Viégas, Martin Wattenberg.
preprint

○ **Communicating Activations Between Language Model Agents**
Vignav Ramesh, **Kenneth Li**.
preprint

○ **When Bad Data Leads to Good Models**
Kenneth Li, Yida Chen, Fernanda Viégas, Martin Wattenberg.
preprint

○ **Dialogue Action Tokens: Steering Language Models in Goal-Directed Dialogue with a Multi-Turn Planner**
Kenneth Li*, Yiming Wang*, Fernanda Viégas, Martin Wattenberg. (*: equal contribution)
preprint

○ **Designing a Dashboard for Transparency and Control of Conversational AI**
Yida Chen, Aoyu Wu, Trevor DePodesta, Catherine Yeh, **Kenneth Li**, Nicholas Castillo Marin, Oam Patel, Jan Riecke, Shivam Raval, Olivia Seow, Fernanda Viégas, Martin Wattenberg.
preprint

○ **Measuring and Controlling Instruction (In)Stability in Language Model Dialogs**
Kenneth Li, Tyler Liu, Naomi Bashkansky, Fernanda Viégas, Hanspeter Pfister, Martin Wattenberg.
Conference on Language Modeling (**COLM 2024**) (**oral**)

○ **Q-Probe: A Lightweight Approach to Reward Maximization for Language Models.**
Kenneth Li, Samy Jelassi, Sham Kakade, Martin Wattenberg, David Brandfonbrener.
International Conference on Machine Learning (**ICML 2024**).

○ **An AI-Resilient Text Rendering Technique for Reading and Skimming Documents.**
Ziwei Gu, Ian Arawjo, **Kenneth Li**, Jonathan K. Kummerfeld, Elena L. Glassman.
International Conference of Human-Computer Interaction (**CHI 2024**).

○ **Inference-Time Intervention: Eliciting Truthful Answers from a Language Model.**
Kenneth Li*, Oam Patel*, Fernanda Viégas, Hanspeter Pfister, Martin Wattenberg. (*: equal contribution)
International Conference on Neural Information Processing Systems (**NeurIPS 2023**) (**spotlight**).

○ **Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task.**
Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, Martin Wattenberg.
International Conference on Learning Representations (**ICLR 2023**) (**oral/notable-top-5%**).

○ **Pose Recognition with Cascade Transformers.**
Ke Li*, Shijie Wang*, Xiang Zhang*, Yifan Xu, Weijian Xu, Zhuowen Tu. (*: equal contribution)
Conference on Computer Vision and Pattern Recognition (**CVPR 2021**).

○ **Multi-Modal Graph Neural Network for Joint Reasoning on Vision and Scene Text.**
Difei Gao*, **Ke Li***, Ruiping Wang, Shiguang Shan, Xilin Chen. (*: equal contribution)
Conference on Computer Vision and Pattern Recognition (**CVPR 2020**).

Research Experiences

Meta AI Research — *Menlo Park, CA*
Research Intern mentored by Du Tran and Matt Feiszli — *May 2022 - Aug. 2022*

Microsoft Research Asia — *Beijing, China*
Research Intern mentored by Zhirong Wu and Stephen Lin — *Dec. 2020 - June 2021*

Awards & Grants

OpenAI Superalignment Fast Grant (50 recipients out of 2700 applicants)	2024 - present
Kempner Institute Graduate Fellowship (22 recipients out of all Harvard grads)	2023 - present
William A. and Fay L. Shutzer Innovation Fund	2022
Harvard SEAS Fellowship	2021

Professional Services & Affiliation

Teaching Assistent, Introduction to Data Science (Harvard COMPSCI109A/STAT109A/APCOMP209A)	2022
Conference reviewer for	2021 - present
- ICML, NeurIPS, ICLR, CVPR, ECCV, ICCV, etc.	
Journal reviewer for	2021 - present
- IEEE Transactions on Pattern Analysis and Machine Intelligence	
- Trends in Cognitive Sciences	