

Keyu Wang

✉ keyuwang@g.harvard.edu | ☎ 617-256-1270 | 🌐 [Homepage](#) | 🎓 [Scholar](#) | 🌐 [keyuwww](#) | 📄 [keyuw](#)

Education

Harvard University

Master of Science in Data Science

January 2027 (Expected)

Cambridge, MA, USA

- Research Advisor: Dr. **Milind Tambe**
- Activities: Research Assistant at Harvard Business & Law School, Policy & Technical Fellow at AI Safety Student Team
- Relevant Courses: Advanced Practical **MLOps**, High-Performance Computing

McGill University

B.Sc in Computer Science, Minor in Geographic Info Science & Remote Sensing

December 2024

Montreal, Canada

- GPA 3.96/4.0, supervised by Dr. **Doina Precup**
- Relevant Courses: Applied Machine Learning, Natural Language Processing, Computer Vision, Database Systems, Algorithm Design, Data Structures, Discrete Mathematics, Statistics, Advanced Geospatial Science, Geospatial Web

Publications

1. **K. Wang***, J. Li*, S. Yang, Z. Zhang, D. Wang. *When Truth Is Overridden: Uncovering the Internal Origins of Sycophancy in Large Language Models*. In **AAAI Conference on Artificial Intelligence**, 2026 (acceptance rate 17.6%); and **Symposium on Model Accountability, Sustainability and Healthcare (SMASH)** (Spotlight★), 2025. [Paper](#)
2. S. Yang, S. Zhu, Z. Wu, **K. Wang**, et al. *Fraud-r1: A Multi-round Benchmark for Assessing the Robustness of LLMs Against Augmented Fraud and Phishing Inducements*. In **Findings of the ACL**, 2025. [Paper](#)
3. **K. Wang**, A.N. Iranzad, S. Schaffter, M. Risdal, D. Precup, J. Lebensold. *Mitigating Downstream Model Risks via Model Provenance*. In **NeurIPS 2024 Workshop on Socially Responsible Language Modelling Research**, 2024. [Paper](#) [Poster](#)

Experience

Research Intern

May 2026 – August 2026

Microsoft - Code AI

- Developed HydraFusion, an RL-trained model router for **GitHub Copilot**; built a 500-session evaluation **benchmark** from real VS Code telemetry with a reactive user simulator, tool-using agent, and multi-judge panel.
- Trained **routing** policies with Group Relative Policy Optimization (GRPO); Improved resolution by +9.2% relative to the previous baseline, while reducing price-sheet cost 48% (\$0.86 to \$0.45 per task) versus the best single model.

AI Workstreams Research Assistant

October 2025 – Present

Harvard Law School - Berkman Klein Center For Internet & Society

Cambridge, MA, USA

- Conducting literature review on AI self-representation and model introspection, including evaluation of LLMs' self-knowledge, reasoning accuracy, and transparency; building benchmarks for model agency and self-reporting reliability.

Machine Learning & Data Science Research Intern

May 2025 – August 2025

Okinawa Institute of Science and Technology (OIST)

Okinawa, Japan

- Led a team of 4 in designing, implementing, and analyzing experiments on understanding LLM's sycophantic behavior inside model architecture, running inference on 15 open-source models from HuggingFace on GPU clusters.
- Co-first author of a [paper](#) on LLM **mechanistic interpretability**, **accepted to AAAI 2026**.

Machine Learning Engineer

Feb 2025 – May 2025

Moonarch

Remote

- Built a document extraction pipeline using Claude API and open-source NLP tools, structuring insights from over 5,000 mixed-format startup profiles (PDFs with images and scanned docs) using **Pydantic AI**, boosting accuracy by 32%.

Visiting Student Researcher - Provable Responsible AI and Data Analytics Lab

Jan 2025 – May 2025

King Abdullah University of Science and Technology (KAUST)

Thuwal, Saudi Arabia

- Executed large-scale LLM evaluation pipelines for fraud detection research, benchmarking 15 open- and closed-source models via API on 8,500+ fraud cases self-designed.
- Co-authored a peer-reviewed [paper](#) at Findings of the Association for Computational Linguistics (ACL) 2025.

Research Intern on AI Safety

May 2024 – December 2024

Mila, AI Research Institute

Montreal, QC, Canada

- Under Dr. **Doina Precup** with collaborators from Google & Meta, illustrated model risk in healthcare, & proposed an open-source [repository](#) for tracking provenance to enhance transparency for responsible model management.
- Published a first-author [paper](#) at NeurIPS 2024 workshop.

Data Scientist Intern (2x)

May 2023 – December 2023 & May 2024 – August 2024

Bell Canada

Montreal, QC, Canada

- Built ML models for business intelligence and competitor analysis (clustering, network speed prediction across Canada), using **Teradata SQL**, **SaaS**, **Hadoop**, Python (**Scikit-learn**) in Agile teams using JIRA/Confluence.
- Enhanced cloud-based data pipelines on **GCP** using advanced RegEx, Apache Airflow, and BigQuery, achieving a 94% data reduction through automated ETL and scalable data processing.

Head Instructor Coding & Game Design

June 2022 – August 2022

Appleby College Camps

Oakville, ON, Canada

- Planned course materials for coding and video game design camps, and taught Python (basic syntax, turtle drawings, etc.) and GameMaker to 150+ children from 7 to 14 years old.

Awards & Scholarships

2x Undergraduate Student Research Award (USRA)	<i>Natural Sciences and Engineering Research Council of Canada</i>	\$17,700
Perseverance Award for Women in Technology 2022		Quebec Government \$3,000
3x Dean's Honour List		McGill University, 2021-2024

Presentations

Symposium on Model Accountability, Sustainability and Healthcare (SMASH) 2025	November 2025
Delivered a spotlight presentation on my paper "Uncovering the Internal Origins of Sycophancy in Large Language Models" in front of 120+ attendees at Mila, Quebec AI Institute.	
Socially Responsible Language Modelling Research (SoLaR) Workshop at NeurIPS 2024	December 2024
Delivered a poster presentation on "Mitigating Downstream Model Risks via Model Provenance" at a NeurIPS 2024 workshop in Vancouver, Canada.	
Undergraduate Computer Science Research Symposium (UCORE)	September 2024
Presented a poster on the "Mitigating Downstream Model Risks via Model Provenance" project at a school-level session with 60+ attendees.	
GIS (Geographic Information Science) Day 2022 <i>Historical Geography</i>	November 2022
Invited oral presentation at McGill's GIS Day showcasing a <u>story map</u> visualizing the history and impacts of immigration to California during 1848–1869.	

Academic & Professional Service

McHacks 13	January 2026
Hackathon Judge	McGill University
Symposium on Model Accountability, Sustainability and Healthcare (SMASH 2025)	November 2025
Workshop Co-organizer	Mila, Quebec AI Institute

Leadership / Extracurricular

Harvard AI Safety Student Team	June 2025 – Present
Technical & Policy Fellow	
Engaged in weekly seminars with peers and mentors on AI governance and technical safety research, focusing on model alignment, interpretability, and regulatory frameworks for frontier systems.	
East2West	April 2023 – Present
Dancer, Editor	
- Danced and edited videos for a K-pop cover crew with a YouTube channel with 1.5M subscribers; averaged 20K views. - Participated in the 2024 K-pop Cover Dance Festival; won 3rd place in the Canada division.	
K-Rave McGill	September 2021 – December 2024
Dance Leader	
Led K-pop dance projects by organizing practice logistics, teaching choreography, and providing weekly individual feedback to ensure performance quality.	
McGill University Chinese Students & Scholars Association	September 2020 – April 2024
Vice President Media – Senior Advisor	
- Directed, filmed, and edited promotional video content for sponsors (e.g., Liuyishou Hotpot, Presotea Montreal) 3 times yearly as part of discount agreements, benefiting students with 10–25% off. - Co-directed Montreal Lunar New Year Gala 2022 with 100+ performers and 15 staffs; managed venue booking, auditions, rehearsals, and stage setup. - Designed and developed WeChat Official Account webpages for 70+ sponsor promotions; reached 1,000+ student users. - Organized member photoshoots for 70 internal members, including studio bookings, creative direction, and photography.	

Projects

McGill School of Computer Science Scheduling Website <i>Web Development</i>	December 2024
As one of the 2 backend developers, built a user-friendly platform that simplifies appointment scheduling.	
You Are At Where You Tweet: GPT Prompting to Geo-locate Twitter Users <i>LLM, AI Safety</i>	April 2024
- Employed a dataset featuring user-defined locations and tweet texts; grouped tweets by user, cleaned data, and refined prompts to optimize accuracy. - Achieved top-3 accuracy of 47% for worldwide city inference and 82% for Australian city predictions.	
California Gold Rush Historical Storymap <i>JavaScript, HTML, MapBox, GeoJSON</i>	October 2022
Visualized the history and impacts of immigration to California during 1848-1869 via a storymap.	