MIA HOPMAN

hopmanma@gmail.com ◇ Google Scholar ◇ GitHub ◇ LinkedIn

EDUCATION

Master of Science, Data Science Jan 2025 - Dec 2025
Worcester Polytechnic Institute (WPI), Worcester, MA *(expected)*

- **Thesis:** Backdoor Attack on Collaborative LLM-Based Multi-Agent Systems *(proposal)*
  - Designed the first backdoor attack specifically targeting LLM-based multi-agent systems, leveraging semantic triggers (uncertainty expressions) in inter-agent communication to manipulate group consensus while maintaining stealth during single-agent evaluations

Bachelor of Science, Data Science Aug 2019 - May 2023
Worcester Polytechnic Institute (WPI), Worcester, MA

Additional: AI Safety Fundamentals - Alignment Course

AI SAFETY RESEARCH

Understanding and Evaluating Scheming Propensity in LLM Agents Jul 2025 - *present*
London AI Safety Research Labs (LASR)

- Conducting rigourous scientific analysis of AI agent behavior across frontier models, identifying key factors that influence strategic deception and self-preservation tendencies under the supervision of David Lindner (Google DeepMind)

Feedback-Based Control Protocols Jan 2025 - May 2025
Cambridge AI Safety Hub - Mentorship for Alignment Research Students, (LessWrongControlConf)

- Investigated a novel "Untrusted Editing with Trusted Feedback" protocol and discovered critical vulnerability where red team models exploited feedback signals to learn monitoring criteria, enabling stealthier backdoor attacks that reduced safety from 74% (trusted monitoring baseline) to 63%
- Developed and evaluated mitigation strategies including compliance monitoring and selective feedback application, with research conducted under mentorship from Tyler Tracy (Redwood Research)

WORK EXPERIENCE

Data Scientist May 2022 - Jan 2025
Hologic, Data & Analytics

- Architected a natural language interface for enterprise data systems using LLMs to convert text queries into SQL and Python code, implementing few-shot learning and RAG techniques to enhance reliability
- Designed and deployed a complaint classifier (80% accuracy, 15K+ cases/month), establishing new organization-wide standards for monitoring, evaluation, and production deployment of ML systems

ADDITIONAL RESEARCH PROJECTS

Generating Mathematics Explanations Using LLMs Sep 2022 - Dec 2022
WPI - ASSISTments Lab, (Conference Paper AIED 2023)

NFL Scouting Combine Predictive Analytics Aug 2022 - May 2023
WPI - Major Qualifying Project (MQP), (report)