HaoYang Shang

+1-778-871-2557 | hs5533@student.ubc.ca | hs5533@nyu.edu | assassin808.github.io

Education

University of British Columbia Vancouver, Canada
M.A.Sc. in Electrical & Computer Engineering (Fast-Track Ph.D.) Sept. 2026 – Present (expected graduation: 08/2028)
Advised by Prof. Xiaoxiao Li

Shanghai Jiao Tong University Shanghai, China
Bachelor of Engineering in Electrical and Computer Engineering Sept. 2021 – June 2025
GPA: 3.69 / 4.0 | Best Performance Reward | Annual TA Excellence Award

Additional Academic Experience

The Hong Kong University of Science and Technology Hong Kong SAR, China
Full-Time Research Assistant, Pervasive Intelligence Lab (PEILab) Dec. 2024 – June 2025
Advised by Prof. Song Guo

New York University New York, United States
Visiting Exchange Student Jan. 2024 – May 2024

Research

My research studies **interpretability across scales**: understanding AI systems, from a single model's internals to groups of interacting agents, well enough to change how they behave. Within a model, I study how behaviors such as memorization, factual boundaries, and vulnerabilities are represented and how they emerge from pretraining to test time, and I design targeted interventions that edit them while preserving general capabilities. Across agents, I draw on cognitive science and social psychology to explain how agents coordinate, exhibit cognitive biases, and develop social rationality in market-like environments.

Current Research

University of British Columbia, Graduate Research 2026.9 – Present
Advisor: Prof. Xiaoxiao Li
Interpretability of LLMs and omni (multimodal) models across the model lifecycle: how behaviors emerge during pretraining, how they are represented in trained models, and how to control them at test time.

Previous Research

Shanghai Jiao Tong University, HKUST & NYU, Undergraduate Research 2023.2 – 2025.6
Advisors: Prof. Song Guo (HKUST), Prof. Quanyan Zhu (NYU), Prof. Yifei Zhu (SJTU)
During my undergraduate studies, I focused on **multi-agent coordination under cognitive constraints and algorithmic privacy**. I modeled agent interactions using cognitive science frameworks (such as Cognitive Load Theory), evaluated prosocial irrationality in LLM decision-making, and investigated differential privacy mechanisms in scheduling and collaborative distributed learning systems.

Major Publications

[1] **Mitigating Memorization Where It Happens.**
HaoYang Shang[*], Xuanqi Zhang[*], Xiaoxiao Li. ([*]Co-first authors)
Advances in Neural Information Processing Systems (NeurIPS), 2026. arXiv:2602.08901. [Code]
This paper proposes Gated Subspace Steering (GSS), an inference-time intervention that uses separate closed-

form probe and steer directions to suppress memorization only on the tokens where it happens, matching or exceeding state-of-the-art mitigation with 100–1000× less compute than optimization-based methods and no weight updates.

[2] **CoBRA: Programming Cognitive Bias in Social Agents Using Classic Social Science Experiments.**
Xuan Liu, **HaoYang Shang**, Haojian Jin.
Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI), 2026.
ACM SIGCHI Best Paper Award.
This paper develops a programmable toolkit to systematically inject and calibrate human cognitive biases in LLM-based social agents, enabling faithful replication of classic psychology and behavioral economics experiments.

[3] **United Minds or Isolated Agents? Exploring Coordination of LLMs under Cognitive Load Theory.**
HaoYang Shang[*], Xuan Liu[*], Zi Liang, Jie Zhang, Haibo Hu, Song Guo. ([*]Co-first authors)
Transactions on Machine Learning Research (TMLR) / arXiv preprint, 2025. arXiv:2506.06843.
This paper explores how multiple LLM agents coordinate on complex collaborative tasks, examining whether multi-agent systems suffer from performance bottlenecks analogous to human working-memory cognitive overload.

[4] **Exploring Prosocial Irrationality for LLM Agents: A Social Cognition View.**
Xuan Liu, Jie Zhang, **HaoYang Shang**, Song Guo, Chengxu Yang, Quanyan Zhu.
Proceedings of The Thirteenth International Conference on Learning Representations (ICLR), 2025.
This paper investigates whether LLM agents exhibit altruistic, prosocial, yet economically irrational behaviors in game-theoretic settings, offering insights into aligning agents with human social norms.

Co-Led Projects

- **HumanStudy-Hub** humanstudy-hub.com
  With Xuan Liu (UCSD).
  Open infrastructure that turns published human studies into reusable datasets and runnable experiments for AI agents.

- **Social Simulation Arena** social-simulation-arena.com
  With Chance Jiajie Li (MIT) and Xuan Liu (UCSD).
  A public forecasting benchmark where AI agents submit forecasts on real social outcomes and are scored against published results, with a live leaderboard.

Other Papers & Preprints

[5] **Love First, Know Later: Persona-Based Romantic Compatibility Through LLM Text World Engines.**
HaoYang Shang[*], Zhengyang Yan[*], Xuan Liu. ([*]Equal contribution)
NeurIPS Workshop on LLM Persona Modeling (PersonaLLM), 2025. **(Oral Presentation)**.
This paper builds an LLM text-world engine to simulate interpersonal romantic interactions, studying how personality traits and persona conditioning affect conversational compatibility across critical decision moments.

[6] **HumanStudy-Bench: Towards AI Agent Design for Participant Simulation.**
Xuan Liu, **HaoYang Shang**, Zizhang Liu, Xinyan Liu, Yunze Xiao, Yiwen Tu, Haojian Jin.
arXiv preprint, 2026. arXiv:2602.00685.
This paper introduces an extensive evaluation benchmark replaying human psychology experiments with simulated LLM agents to evaluate behavioral fidelity across diverse experimental paradigms.

[7] **Diagon: A Programmable Testbed for AI-Agent Cognitive Labor Markets.**
Xuan Liu, **HaoYang Shang**, Haojian Jin.
arXiv preprint, 2026. arXiv:2604.06688. (Previously titled *"When Agent Markets Arrive"*).
This paper introduces a multi-agent simulation testbed where AI agents autonomously contract, negotiate prices, and build market reputations, studying market stability prior to real-world deployment.

[8] **Validated Hypotheses as a Lens for Human-Likeness Evaluation in AI Agents.**
Xuan Liu, **HaoYang Shang**, Zizhang Liu, Yuanjun Feng, Guankai Zhai, Yunze Xiao, Yiwen Tu, Haojian Jin.

arXiv preprint, 2026. arXiv:2605.15473.
This paper proposes a systematic evaluation protocol using validated social science hypotheses as a checklist to benchmark how closely generative agents reflect real human behavior.

[9] **Decision Potential Surface: A Theoretical and Practical Approximation of LLM's Decision Boundary.**
Zi Liang, Zhiyao Wu, **HaoYang Shang**, Yulin Jin, Qingqing Ye, Huadi Zheng, Peizhao Hu, Haibo Hu.
arXiv preprint, 2025. arXiv:2510.03271.
This paper establishes formal approximations of decision potential surfaces in LLMs, providing geometric and theoretical tools to map model decision boundaries.

[10] **FinThink: An LLM-based Multi-agent System for Financial Reasoning.**
Zeyu Liu, Jie Zhang, **HaoYang Shang**, Sing Kwong Lai, Yang Li, Song Guo.
Research Preprint / Under Review, 2025.
This paper designs a collaborative multi-agent architecture tailored for multi-step quantitative reasoning and risk analysis in financial contexts.

Scholarships & Awards

- **ACM SIGCHI Best Paper Award**, ACM CHI Conference on Human Factors in Computing Systems 2026
- **Oral Presentation Award**, NeurIPS Workshop on LLM Persona Modeling (PersonaLLM) 2025
- **Annual Excellent Teaching Assistant Award**, Shanghai Jiao Tong University 2023
- **Bronze Medal**, The University Physics Competition 2022
- **Best Performance Reward**, Engineering Design Project at Shanghai Jiao Tong University 2021

Teaching Experience

- **Teaching Assistant**, Honor Mathematics, Shanghai Jiao Tong University Sept. 2023 – Dec. 2023
Delivered recitation lectures, designed practice problems, and held weekly office hours for honors students.
- **Teaching Assistant**, Introduction to Computer Engineering & Chemistry, SJTU Mar. 2022 – Aug. 2022
Supervised laboratory hardware experiments, created assignments, and assisted students in debugging course projects.

Links

Homepage: https://assassin808.github.io
Google Scholar: https://scholar.google.com/citations?user=eAfoYDIAAAAJ
GitHub: https://github.com/assassin808
LinkedIn: https://www.linkedin.com/in/haoyang-shang-8196612b4/ *Last updated: Nov. 2, 2026*