Han Luo

hanluo@uw.edu | Google Scholar | LinkedIn | X | GitHub

EDUCATION BACKGROUND

University of Washington, Information School — Mar. 2026 – Present
Visiting Student

University of Leeds & Southwest Jiaotong University — Sept. 2024 – 2028 (expected)
BSc in Computer Science (Joint Programme)

AWARDS & HONORS

- **National Scholarship** – China's top undergraduate award
- Cambridge OSRP Final Research Showcase (Top 5/97)

SELECTED PUBLICATIONS

* Equal contribution.

[1] **HERA: Harness-Environment Co-Evolution for Reliable Agentic Abstention**
Han Luo*, Bingbing Wen*, Guang Yang, Zora Zhiruo Wang, Pan Lu and Lucy Lu Wang.
Submitted to ICLR. arXiv:2610.06563, 2026.

[2] **Agentic Abstention: Do Agents Know When to Stop Instead of Act?**
Han Luo*, Bingbing Wen* and Lucy Lu Wang.
NeurIPS 2026, accepted. arXiv:2606.28733. Hugging Face #1 paper of the day.

[3] **SPASM: Stable Persona-driven Agent Simulation for Multi-turn Dialogue Generation**
Han Luo and Guy Laban.
Findings of ACL 2026.

[4] **DialogGuard: Multi-Agent Psychosocial Safety Evaluation Interface of Sensitive LLM Responses**
Han Luo and Guy Laban.
ACL 2026 System Demonstrations, accepted.

RESEARCH EXPERIENCE

Visiting Student | LARCH Lab, University of Washington — 1 Mar 2026 – Present
Supervised by Prof. Lucy Lu Wang

Agentic Abstention | *NeurIPS 2026, accepted*

- Proposed Agentic Abstention and evaluated 13 LLM-as-agent systems and two agent scaffolds across a benchmark of 28K+ instructions spanning interactive QA, WebShop, and Terminal-Bench.
- Developed CONVOLVE (Context Evolution), which distills interaction trajectories into reusable stopping rules; improved Llama-3.3-70B's timely abstention recall on WebShop from 26.7% to 57.4% using 20 training trajectories.

HERA | *Submitted to ICLR*

- Co-developed HERA for jointly evolving agent harnesses and execution environments, together with HERA-Bench, a held-out benchmark of 60 verified feasible–infeasible task pairs.
- Improved GPT-5.6-Luna's held-out abstention success from 61.7% to 83.3% and feasible-task completion from 68.3% to 76.7%; the evolved harness increased abstention success by 15.3 percentage points on average across 19 additional LLMs.

Cambridge Online Summer Research Programme (OSRP) 30 Jun 2025 – 14 Jan 2026

Supervised by Prof. Guy Laban | Research lab: LabaLab, Ben-Gurion University of the Negev

SPASM | *Findings of ACL 2026*

- Developed SPASM, a modular framework for persona-conditioned dialogue simulation, and Egocentric Context Projection (ECP) to reduce persona drift and role confusion.
- Generated and analyzed 45,000 conversations using 4,500 personas across nine client–responder pairings of three LLM backbones, combining embedding-based measurements with human validation.
- Observed no echoing under ECP in the human-validated evaluation, compared with role-adoption errors under standard history concatenation.

Research Intern | LabaLab, Ben-Gurion University of the Negev 29 Jul 2025 – 1 Dec 2025

Supervised by Prof. Guy Laban

DialogGuard | *ACL 2026 System Demonstrations, accepted*

- Designed DialogGuard for psychosocial safety evaluation across five risk dimensions, implementing four LLM-as-a-judge pipelines grounded in shared three-level rubrics.
- Evaluated judging methods against human annotations on PKU-SafeRLHF using 400 examples per risk dimension; dual-agent correction provided the strongest overall accuracy–robustness trade-off among the tested configurations.
- Developed an open-source web interface with Live Chat and Manual Input workflows, displaying per-dimension risk scores and natural-language rationales.

ACADEMIC SERVICE

ICLR 2027	Reviewer
NeurIPS 2026	Reviewer
CUI 2026	Reviewer
WWW 2026	Reviewer