Haoming (Neil) Li

haoming.li@utdallas.edu · +1 470 929 0691 · Homepage · Google Scholar · Github
Work authorization: CPT eligible for internships

Research Interests

My current focus is **personal AI** and **physical AI**, and the **agentic memory** that both rely on [4, 5]. Before this, I worked on **computer-use agents (CUAs)**, including error recovery in long-horizon tasks [6] and joint reinforcement learning of multiple CUAs (ongoing). Earlier, I worked on LLM planning [9–11] and scientific summarization [1].

Education

The University of Texas at Dallas, Ph.D. in Computer Science, GPA 3.83/4.00 — Aug 2025 – Present
Emory University, M.S. in Computer Science, GPA 3.97/4.00 — Aug 2022 – May 2024
Renmin University of China, B.S. in Data Science and B.Econ. in Finance — Sep 2018 – Jun 2022

Publications

†equal contribution

Published

[1] **Haoming Li**, Jessica Ouyang. *GRASP: Graph-Reasoning Aided Survey Planning for High-Fidelity Related Work Generation.* **Findings of ACL 2026**. [paper]

[2] Weili Shi, **Haoming Li**, Sheng Li. *Enhancing Dynamic Text-Attributed Graph Modeling via Instruction Tuning.* **Findings of AACL-IJCNLP 2026** (to appear).

[3] Yikun Wang, Rui Zheng, **Haoming Li**, Qi Zhang, Tao Gui, Fei Liu. *RESCUE: Ranking LLM Responses with Partial Ordering to Improve Response Generation.* **ACL 2024 Student Research Workshop**. [paper]

Under Review

[4] **Haoming Li**†, Defu Cao†, Xu Hu†, Yingtao Luo, . . . , Yuexing Hao, Xiaomin Li. *PersonaCUABench: Benchmarking Computer-Use Agents Serving Diverse Personas.* Under review at **ACL ARR October 2026**. [arXiv]

[5] **Haoming Li**, Dingyi Kang, Xu Hu, Jinxin Yang, . . . , Feng Chen, Bingzhe Li. *SAFE-Mem: Secure Agentic Memory with Fine-Grained Authority and Governance-Aware Retrieval.* Under review at **ACL ARR October 2026**. [arXiv]

[6] Xu Hu†, **Haoming Li**†, Yuxiang Wang, . . . , Bingzhe Li, Feng Chen. *CUA-Recovery: Benchmarking Error Recovery in Long-Horizon Computer-Use Agents.* Under review at **ICLR 2027**. [arXiv]

[7] Siyang Wu, Weizhi Zhang, Chen Yang, . . . , **Haoming Li**, . . . *PersonaScienceBench: Evaluating Scientific Agents with Diverse Simulated Researchers.* Under review at **ACL ARR October 2026**.

[8] Xu Hu, **Haoming Li**, Sabbir Ahmed, . . . , Feng Chen, Bingzhe Li. *Toward Trustworthy Computer-Use Agents: Risk Propagation, Evaluation Gaps, and Human Governance.* Under review at **ACM Computing Surveys**.

Preprints

[9] **Haoming Li**, Zhaoliang Chen, Jonathan Zhang, Fei Liu. *PLANET: A Collection of Benchmarks for Evaluating LLMs' Planning Capabilities.* arXiv:2504.14773, 2025.

[10] **Haoming Li**†, Zhaoliang Chen†, Songyuan Liu, Yiming Lu, Fei Liu. *Systematic Analysis of LLM Contributions to Planning: Solver, Verifier, Heuristic.* arXiv:2412.09666, 2024.

[11] **Haoming Li**†, Zhaoliang Chen†, Jonathan Zhang, Fei Liu. *LASP: Surveying the State-of-the-Art in Large Language Model-Assisted AI Planning.* arXiv:2409.01806, 2024.

Research Experience

MatrAIx, Research Collaborator (persona agents) — Sep 2026 – Present

- **PersonaCUABench** (co-first author) [4]: selected and designed 44 personas in four groups across 11 domains, and evaluated nine CUA backbones (3,330 runs). Persona-phrased requests cut mean task completion from 0.689 to 0.441 (36% relative), and even fully completed persona requests receive the top satisfaction rating in only 62.7% of cases, against 84.1% for the original requests.
- Contributed to **PersonaScienceBench** [7], which evaluates scientific agents with diverse simulated researchers, and to the public **PersonalAgentEval** leaderboard, which evaluates personal agents through user simulation.

The University of Texas at Dallas, Research Assistant (advisors: Feng Chen, Bingzhe Li) — Jun 2025 – Present

- **SAFE-Mem** (first author) [5]: designed an agentic memory that accounts for both the stored content and its information source, so each claim is used only for what its source is allowed to establish. On MPBench with GPT-4o-mini, it cuts attack success from 49.8% to 6.2% while raising recall of what users asked to remember from 79.5 to 85.9, and it raises LongMemEval accuracy from 45.2% to 58.2%.
- **CUA-Recovery** (co-first author) [6]: designed *ReRail*, the task and trajectory synthesis framework behind the benchmark (2,250 human-verified erroneous states for testing, 8.6K trajectories for training). Evaluated six agents at six takeover depths (best

error-awareness rate: 44.0%) and fine-tuned Qwen3.5-35B-A3B with SFT, raising its error-awareness rate from 19.9% to 35.2% on held-out states.

- **Poly-CUA** (in progress): a multi-agent CUA framework in which a manager agent coordinates computer-use subagents through a task graph, with the agents trained jointly by reinforcement learning.
- **Physical Mem** (in progress): agentic memory for VLM-based planning of simulated robots in changing scenes, trained with reinforcement learning to decide what to store, update, and forget.
- **Trustworthy CUA survey** (second author) [8]: reviewed and organized the literature on CUA security, privacy, reliability, controllability, and accountability.
- **GRASP** (first author, advised by Jessica Ouyang) [1]: designed a two-layer graph planner for related-work generation, pruned with a Steiner tree to keep the core relations among cited papers. It beats the strongest baseline in ROUGE-1 F1 (0.771 vs. 0.664) and in identifying the dominant cited papers (F1 0.922 vs. 0.841).

University of Virginia, Reasoning & Knowledge Discovery Lab, Research Assistant (advisor: Sheng Li) Jun 2024 – May 2025

- Instruction-tuned LLMs to encode and reason over dynamic text-attributed graphs [2].

Emory University, AI/LLM Agents Research Group, Research Assistant (advisor: Fei Liu) Mar 2023 – May 2025

- **LLMs for planning** (first author) [9–11]: evaluated GPT-4o, Claude, and DeepSeek as solver, verifier, and comparative heuristic on travel, course, and fitness planning, and built the course- and fitness-planning benchmarks (the latter tests whether LLMs learn user preferences). LLMs struggle to produce correct plans directly but rank candidate plans well (Hit@3 up to 0.96), which supports using them as heuristics in tree search. Also wrote a survey of LLM-assisted planning and a review of planning benchmarks.
- **RESCUE** [3]: trained a response ranker from partial orderings over candidate LLM outputs. With 0.4% of the e-SNLI training data, it raises accuracy from 77.5% (SFT) to 82.9%, and human raters prefer its explanations over SFT's with a 47% win rate.

Skills and Service

Research skills: LLM post-training; LLM-based user simulation; agentic memory design.
Tools: Python, PyTorch, Hugging Face Transformers, vLLM, Docker. **Languages:** English (fluent), Chinese (native).
Reviewer: NeurIPS 2026, IEEE BIBM 2026, International Journal of Data Science and Analytics (IJDSA).