Haochang Hao

Ph.D. Student in Computer Science | Trustworthy LLMs & Agents | Graph Learning

Chicago, IL | hhao@uic.edu | Google Scholar | ORCID | LinkedIn | GitHub

Summary

Computer Science Ph.D. student at the **University of Illinois Chicago**, advised by **Dr. Lu Cheng**, seeking an **ML/AI research or applied-science internship**. My research interests focus on two areas: **Trustworthy LLMs, Agents & Post-Training** (agent skill / security, LLM safety alignment and RL training) and **Graph Learning & Graph Foundation Models** (graph learning, foundation models and their connection to LLMs / agents).

Technical Skills

LLM & Agents: LLM evaluation, LLM-as-a-Judge, agent skill, safety alignment, RL post-training, LLM training & inference (HuggingFace, vLLM, Unsloth, TRL), agent systems (Claude Code, Codex, OpenClaw)
Graph & ML: PyTorch, graph neural networks, graph foundation models, uncertainty quantification, conformal prediction, meta-learning, domain adaptation
Programming & Tools: Python, Java, SQL, Git, Docker, Daytona, Linux, LaTeX, TensorBoard, databases (SQL, MongoDB)

Research Projects

Trustworthy LLMs, Agents & Post-Training

PUBLISHED / UNDER REVIEW

SafeCRS: Personalized Safety Alignment for LLM-Based Conversational Recommender Systems — *Co-first author | ACM SIGKDD 2026*

- Designed a personalized safety-alignment framework for LLM conversational recommenders that enforces user-specific safety constraints while preserving recommendation quality.
- Owned the core idea, model training, and writing for a (co-)first-author paper accepted to the ACM SIGKDD 2026 Research Track.

POISE: Position-Aware Undetectable Skill Injection on LLM Agents — *Co-first author (lead) | EMNLP 2026 (under review)*

- Designed a stealthy skill-poisoning attack that hides a single benign-looking trigger in a skill file, reaching 90–97% trigger rates on the Skill-Inject benchmark while evading the agent's inspection.
- Led the attack formulation, experiments, and analysis, and proposed practical hardening recommendations for agent skill ecosystems.

CorVer: Lightweight Corpus-Grounded Process Rewards for Factual QA — *Co-author (equal contribution) | EMNLP 2026 (under review)*

- Co-developed a plug-in process reward for RL on knowledge-intensive QA that replaces expensive neural verifiers with a corpus-grounded signal from Wikipedia co-occurrence statistics.
- Targeted fine-grained, step-level supervision that stays robust for rare-entity facts, where neural reward models tend to fail.

IN PROGRESS

MetaCaliJudge: Post-hoc Calibration for LLM-as-a-Judge — *First author | In progress*

- Building a post-hoc calibration pipeline that turns unreliable LLM-judge verdicts into calibrated, uncertainty-aware scores that stay reliable under dataset shift.
- Exploring how the calibrated outputs can serve as reward signals for RL-based factual grounding.

Graph Learning & Graph Foundation Models

PUBLISHED

Graph Representation Learning on Heterogeneous Information Networks — *First author | ESWA 2026; NCA 2025*

- Designed progressive alternating attribute–structure optimization for multiplex heterogeneous graphs, improving robustness under missing attributes and noisy edges (ESWA 2026).
- Developed a multi-level semantics extraction method for heterogeneous-graph node classification (NCA 2025); filed an approved patent on the technique.

IN PROGRESS

Riemannian Graph Foundation Models *First author | In progress*

- Building a graph foundation model on a product manifold of hyperbolic, spherical, and Euclidean factors so the embedding geometry matches graph structure.
- Using curvature-guided causal disentanglement to separate transferable structure from domain-specific signal for zero-shot cross-domain transfer.

ConfDock: Conformal Prediction for Molecular Docking *First author | In preparation*

- Developed a CQR-GNN framework giving distribution-free, atom-level confidence intervals for predicted binding poses across 238 protein–ligand systems.

Publications

(† equal contribution; **bold** = author.)

[1] **Hao, H.**†, Xu, Y., Li, X., Ge, Y. & Cheng, L. "SafeCRS: Personalized Safety Alignment for LLM-Based Conversational Recommender Systems." *Accepted at ACM SIGKDD 2026, Research Track, Cycle 2.* arXiv:2603.03536. arXiv

[2] **Hao, H.**†, Min, D.†, Zhang, Z.†, Zhang, Y., Xu, M., Ge, Y. & Cheng, L. "POISE: Position-Aware Undetectable Skill Injection on LLM Agents." *Under review at EMNLP 2026.* arXiv:2606.07943. arXiv

[3] Fan, S.†, **Hao, H.**†, Min, D.†, Liu, W., Yu, P. S. & Cheng, L. "Verifiable Rewards Beyond Math and Code: Lightweight Corpus-Grounded Process Supervision for Factual Question Answering." *Under review at EMNLP 2026.* arXiv:2605.29648. arXiv

[4] **Hao, H.**, Huang, J. & Rao, S. "Progressive alternating attribute-structure optimization for multiplex heterogeneous graphs." *Expert Systems with Applications.* 312, 131495, 2026. DOI

[5] **Hao, H.**, Huang, J. & Rao, S. "Heterogeneous graph multi-level semantics extraction for node classification." *Neural Computing & Applications.* 37, 11821–11841, 2025. DOI

Education

Ph.D. in Computer Science *Aug. 2025 – Present*
University of Illinois Chicago, Chicago, IL, USA
Advisor: Dr. Lu Cheng

M.Eng. in Electronic & Information Engineering *Sept. 2022 – June 2025*
Shanghai Advanced Research Institute, University of Chinese Academy of Sciences, Shanghai, China
Advisor: Prof. Jun Huang | Research focus: Data Mining on Knowledge Graphs

B.Eng. in Computer Science & Technology *Sept. 2018 – June 2022*
Soochow University, Suzhou, China

Patent & Honors

Patent: J. Huang & **H. Hao**. "Node Classification Method, Device, Terminal, and Medium Based on Multi-level Semantic Representation of Heterogeneous Knowledge Graphs." *Approved.*
Honors: Multiple "Three Good Student" awards at UCAS and Soochow University for outstanding academic and overall performance.

Last updated: June 2026