Huanhuan Ma

CONTACT INFORMATION

huanhuanma.top | Google Scholar | GitHub
hma42@uic.edu | LinkedIn
+1 773 827 4356 | Chicago, USA

EDUCATION

University of Illinois Chicago, Chicago, USA — Aug 2025 – Present
PhD in Computer Science, expected May 2029
- Advisor: Prof. Philip S. Yu (ACM/IEEE/AAAS Fellow)

University of Chinese Academy of Sciences, Beijing, China — Sep 2021 – Jul 2024
Master in Artificial Intelligence
- Advisors: Prof. Liang Wang, Prof. Shu Wu, Prof. Qiang Liu
- Thesis: Explainable Automated Fact Verification

Zhengzhou University, Zhengzhou, China — Sep 2016 – Jul 2020
Bachelor in Software Engineering

RESEARCH CONTRIBUTIONS

My research develops **evaluation frameworks and personalization methods for large language models (LLMs)** so that AI systems serve individual users in ways that are trustworthy, fair, and reliable. Published work has received over 236 citations (Google Scholar).

- **Trustworthy LLMs & Fact Verification:** built **EX-FEVER** [2], a multi-hop explainable fact-verification benchmark, and **LOGRAN** [3], a soft-logic regularization method for multimodal out-of-context detection.
- **Reliable Behavioral Evaluation:** proposed the **Core Sentiment Inventory (CSI)** [7], an implicit behavioral probe that bypasses self-report bias to test whether LLMs behave consistently.
- **Personalization & User Alignment:** developed an evaluation framework showing that anti-sycophancy methods can impair rational belief updating [1].

INTERNSHIP

Research Intern — Beijing Academy of Artificial Intelligence (BAAI)
Mentor: Dr. Li Du — *Jul 2024 – Apr 2025*
- Engineered instruction-data pipelines (collection, quality scoring, deduplication, difficulty filtering) for **Infinity-MM** [13], a **40M-sample** open multimodal instruction dataset.
- Designed tag-conditioned **synthetic data generation** targeting diagnosed failure modes; the resulting corpus trained **Aquila-VL-2B** to state-of-the-art performance at the 2B scale.

FIRST-AUTHOR PUBLICATIONS

[*] = *equal contribution.*

[1] **Huanhuan Ma**, Henry Peng Zou, Chengze Li, Enze Ma, Yunyue Su, Philip S. Yu. "Sycophancy Suppression Can Impair Rational Updating: Anti-Sycophancy Should Preserve the Ability to Update." *Findings of the Association for Computational Linguistics* **(EMNLP 2026 Findings)**. [arXiv]

[2] **Huanhuan Ma**, Weizhi Xu, Yifan Wei, et al. "EX-FEVER: A Dataset for Multi-hop Explainable Fact Verification." *Findings of the Association for Computational Linguistics* **(ACL 2024 Findings)**. [PDF], [Code/Dataset]

[3] **Huanhuan Ma**[*], Jinghao Zhang[*], Qiang Liu, Shu Wu, Liang Wang. "Interpretable Multimodal Out-of-Context Detection with Soft Logic Regularization." *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* **(ICASSP 2024, Oral)**. [PDF]

OTHER PUBLICATIONS

[4] Yifan Wei, Xiaoyan Yu, Yixuan Weng, **Huanhuan Ma**, et al. "Does Knowledge Localization Hold True? Surprising Differences Between Entity and Relation Perspectives in Language Models." *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management* **(CIKM 2024)**. [PDF]

[5] Yifan Wei, Yisong Su, **Huanhuan Ma**, et al. "MenatQA: A New Dataset for Testing the Temporal Comprehension and Reasoning Abilities of Large Language Models." *Findings of the Association for Computational Linguistics* **(EMNLP 2023 Findings)**. [PDF]

[6] Liping Wang, Qiang Liu, **Huanhuan Ma**, Shu Wu, Liang Wang. "Multi-Cause Learning for Diagnosis Prediction." *Proceedings of the International Conference on Data Mining and Big Data* **(DMBD 2022)**. [PDF]

Preprints & Under Review

The following are preprints or submitted manuscripts, not yet peer-reviewed.

[7] **Huanhuan Ma**, Haisong Gong, Xiaoyuan Yi, Xing Xie, Philip S. Yu, Dongkuan Xu. "Beyond BFI: The CSI for Enhanced Reliability and Validity in Evaluating LLM Personality Traits." *Preprint*. [arXiv]

[8] Hanrong Zhang, Yankai Chen, Shicheng Fan, Dehai Min, Shaowen Chen, **Huanhuan Ma**, et al. "Scaling LLM Agent Learning with Data Synthesis: A Comprehensive Survey." *Under review at EMNLP 2026*.

[9] Enze Ma, Yufan Zhou, Jie Yang, **Huanhuan Ma**, et al. "MEMPROBE: Probing Long-Term Agent Memory via Hidden User-State Recovery." *Under review at NeurIPS 2026*.

[10] Weizhi Zhang, Wei-Chieh Huang, Yueqing Liang, et al., **Huanhuan Ma**, et al., Kai Shu, Xue Liu, Philip S. Yu. "JARVIS-Bench: Benchmarking Personal Intelligence Agents on Long-Horizon Real-User Daily Traces." *Under review at NeurIPS 2026*.

[11] Henry Peng Zou, Chunyu Miao, Wei-Chieh Huang, et al., **Huanhuan Ma**, et al., Xue Liu, Philip S. Yu. "InterruptBench: Evaluating LLM Agents Under User Interruptions in Long-Horizon Web Tasks." *Preprint*. [arXiv]

[12] Haisong Gong, **Huanhuan Ma**, Qiang Liu, Shu Wu, Liang Wang. "Navigating the Noisy Crowd: Finding Key Information for Claim Verification." *Preprint*. [arXiv]

[13] Shuhao Gu, Jialing Zhang, et al., **Huanhuan Ma**, et al., Xinlong Wang, Guang Liu. "Infinity-MM: Scaling Multimodal Performance with Large-Scale and High-Quality Instruction Data." *Preprint*. [arXiv]

Research Experience

Personalization & User Alignment — University of Illinois Chicago, Chicago, USA
Research Assistant, Advisor: Prof. Philip S. Yu — *Aug 2025 – Present*

- Studied how LLM applications can personalize for different users while preserving rational belief updating and avoiding sycophancy from over-personalization [1]. **Findings of EMNLP 2026**.

Behavioral Evaluation for LLMs — North Carolina State University *(Remote)*
Collaborator: Dr. Dongkuan Xu — *Dec 2023 – Dec 2024*

- Proposed the Core Sentiment Inventory (CSI) [7] for reliable behavioral evaluation of LLMs.

Trustworthy LLMs & Fact Verification — University of Chinese Academy of Sciences
Master's Student, Advisors: Prof. Liang Wang, Prof. Shu Wu, Prof. Qiang Liu — *Aug 2021 – Jul 2024*

- Created **EX-FEVER** (ACL 2024) [2], a benchmark for multi-hop explainable fact verification.
- Developed **LOGRAN** for multimodal out-of-context detection (ICASSP 2024 Oral) [3].
- Contributed to **EACon** [12], an evidence-abstraction and claim-deconstruction framework for LLM-based claim verification over noisy evidence.

Awards

- **NSF ACCESS Explore Award** (**Principal Investigator**), 200,000 ACCESS computing credits 2026–2027
- **ACL ARR Outstanding Reviewer Award** (July 2025 Cycle) 2025
- Ministry of Education – Huawei "Intelligent Base" Future Star 2022

Academic Service

Reviewer:

- **ICLR** 2025, 2026 • **NeurIPS** 2026 • **ACL ARR** 2025, 2026 • **CIKM** 2024, 2025
- **AAAI** 2026 Workshop (PerFM)

Open Source

Awesome-LLM-based-Evaluators [GitHub] ★ 33

Curated collection of state-of-the-art research on **LLM-as-a-Judge** and behavioral-evaluation.