Hongli Zhou

Ph.D. Student, Computer Science & Technology, Harbin Institute of Technology

15636601311 hongli.joe@stu.hit.edu.cn Joe-Hall-Lee.github.io

Education

Harbin Institute of Technology *09.2025 – Present*

Ph.D. Student in Computer Science and Technology (Advisor: Prof. Muyun Yang)

- Research fields: LLM Evaluation, Reward Modeling

Harbin Institute of Technology *08.2021 – 06.2025*

B.Eng. in Digital Media Technology

- Thesis: Design and Implementation of an LLM Evaluation Web Application Based on LLM-as-a-Judge

Publications

[1] **Hongli Zhou***, Hui Huang*, Wei Liu*, Chenglong Wang, Xingyuan Bu, Lvyuan Han, Fuhai Song, Muyun Yang, Wenhao Jiang, Hailong Cao, and Tiejun Zhao. RM-Distiller: Exploiting Generative LLM for Reward Model Distillation. In *Proceedings of IJCAI*, 2026.

[2] **Hongli Zhou***, Hui Huang*, Ziqing Zhao, Lvyuan Han, Huicheng Wang, Kehai Chen, Muyun Yang, Wei Bao, Jian Dong, Bing Xu, Conghui Zhu, Hailong Cao, and Tiejun Zhao. Lost in Benchmarks? Rethinking Large Language Model Benchmarking with Item Response Theory. In *Proceedings of AAAI*, 2026.

[3] Hui Huang*, Yancheng He*, **Hongli Zhou***, Rui Zhang, Wei Liu, Weixun Wang, Jiaheng Liu, and Wenbo Su. Think-J: Learning to Think for Generative LLM-as-a-Judge. In *Proceedings of AAAI*, 2026.

[4] Hui Huang*, Xingyuan Bu*, **Hongli Zhou***, Yingqi Qu, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. An Empirical Study of LLM-as-a-Judge for LLM Evaluation: Fine-tuned Judge Model is not a General Substitute for GPT-4. In *Findings of ACL*, 2025.

[5] **Hongli Zhou**, Hui Huang, Yunfei Long, Bing Xu, Conghui Zhu, Hailong Cao, Muyun Yang, and Tiejun Zhao. Mitigating the Bias of Large Language Model Evaluation. In *Proceedings of CCL*, 2024.

[6] Ruolin Zheng*, **Hongli Zhou***, Wei Bao, Yang Xu, Heng Ye, Wenlin Song, Jian Dong, and Xudong Yang. LMBench: A Comprehensive and Standardized Benchmark for Evaluating Large-scale Models. *Data Intelligence*.

Internship

ModelBest Inc. *03.2026 – 09.2026*

Foundation Model Center | Research Intern

- Developed the first benchmark for evaluating legal-domain reward models.

China Electronics Standardization Institute *11.2024 – 04.2025*

Information Technology Research Center | Research Intern

- Developed an analysis framework for LLM benchmarks and systematically analyzed multiple benchmarks, revealing deficiencies in their measurement quality.

Skills

- **English:** CET-6 (528), good reading and writing competencies for English.
- **Programming:** Proficient in Python, PyTorch, LLaMA-Factory and vLLM.
- **LLM:** In-depth experience at LLM evaluation and reward modeling.