Chunyang Li

✉ cliei@connect.ust.hk https://github.com/lcy2723 https://lcy2723.github.io/

EDUCATION

The Hong Kong University of Science and Technology Hong Kong SAR, China
MPhil in Computer Science and Engineering *Sep 2024 - Present*
HKUST-KnowComp. Advisor: Prof. Yangqiu Song

Tsinghua University Beijing, China
B.Eng. in Computer Science and Technology; GPA: 3.89 / 4.00 *Sep 2020 - Jun 2024*
THUKEG. Advisor: Prof. Juanzi Li and Prof. Lei Hou
Outstanding graduate of the Department of Computer Science and Technology

RESEARCH INTERESTS

My research interests primarily lie in the field of natural language processing, with a particular focus on language model evaluation, including cognitive capabilities (such as knowledge acquisition, adaptive evolution and reasoning) of language models and LLM-as-a-judge.

PUBLICATIONS & PREPRINTS

You can also find my latest publications on my Google Scholar.

Journals & Conference Proceedings:

[1] **Chunyang Li**, Weiqi Wang, Tianshi Zheng, and Yangqiu Song, "Patterns Over Principles: The Fragility of Inductive Reasoning in LLMs under Noisy Observations", *Findings of ACL 2025*.

(a) Robust Rule Induction: propose a 3-stage framework to evaluate LLM's rule inference capabilities and find that LLMs are inherently sensitive to noise and prone to pattern overfitting under counterfactual scenarios.

(b) Sample-steered Rule Refinement: develop a workflow which improves LLM's and LRM's inductive reasoning capability by 15% in average by incorporating code execution results as feedback.

[2] **Chunyang Li***, Hao Peng*, Xiaozhi Wang, Yunjia Qi, Lei Hou, Bin Xu, and Juanzi Li, "MAVEN-Fact: A Large-scale Event Factuality Detection Dataset", *Findings of EMNLP 2024*.

(a) Large Data Scale: propose the largest event factuality detection dataset with 112, 276 events and evidence.

(b) LLM-then-Human Annotation: develop an annotation workflow which reduces costs by 15%.

(c) Hallucination Mitigation: improve Llama 3's accuracy from 77.6% to 88.9% and GPT-4's accuracy from 83.3% to 97.8% on an event-related hallucination detection QA dataset.

[3] Tianshi Zheng*, Yixiang Chen*, Chengxi Li*, **Chunyang Li**, Qing Zong, Haochen Shi, Baixuan Xu, Yangqiu Song, Ginny Y. Wong, and Simon See, "The Curse of CoT: On the Limitations of Chain-of-Thought in In-Context Learning", *Transactions on Machine Learning Research (TMLR)*.

[4] Tianshi Zheng, Jiayang Cheng, **Chunyang Li**, Haochen Shi, Zihao Wang, Jiaxin Bai, Yangqiu Song, Ginny Y. Wong, and Simon See, "LogiDynamics: Unraveling the Dynamics of Logical Inference in Large Language Model Reasoning", *EMNLP 2025*.

[5] Weiqi Wang, Tianqing Fang, **Chunyang Li**, Haochen Shi, Wenxuan Ding, Baixuan Xu, Zhaowei Wang, Jiaxin Bai, Xin Liu, Cheng Jiayang, Chunkit Chan, and Yangqiu Song, "CANDLE: Iterative Conceptualization and Instantiation Distillation from Large Language Models for Commonsense Reasoning", *ACL 2024*.

[6] Jifan Yu*, Xiaozhi Wang*, Shangqing Tu*, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, **Chunyang Li**, Zheyuan Zhang, Yushi Bai, Yantao Liu, Amy Xin, Kaifeng Yun, Linlu GONG, Nianyi Lin, Jianhui Chen, Zhili Wu, Yunjia Qi, Weikai Li, Yong Guan, Kaisheng Zeng, Ji Qi, Hailong Jin, Jinxin Liu, Yu Gu, Yuan Yao, Ning Ding, Lei Hou, Zhiyuan Liu, Xu Bin, Jie Tang, and Juanzi Li, "KoLA: Carefully Benchmarking World Knowledge of Large Language Models", *ICLR 2024*.

[7] Shangqing Tu*, Zheyuan Zhang*, Jifan Yu, **Chunyang Li**, Siyu Zhang, Zijun Yao, Lei Hou, and Juanzi Li, "LittleMu: Deploying an Online Virtual Teaching Assistant via Heterogeneous Sources Integration and Chain of Teach Prompts", *CIKM 2023*.

Preprints:

[1] **Chunyang Li***, Yilun Zheng*, Xinting Huang, Tianqing Fang, Jiahao Xu, Yangqiu Song, Lihui Chen, Han Hu, "WebDevJudge: Evaluating (M)LLMs as Critiques for Web Development Quality", *Under Review.*

   (a) WebDevJudge is a systematic benchmark for assessing LLM-as-a-judge performance in web development, supporting with both non-interactive evaluation based on static observations and interactive evaluation with a dynamic web environment.

   (b) We also provide WebDevJudge-Unit, a diagnostic dataset specifically designed to evaluate task-level feasibility verification capabilities.

[2] Baixuan Xu*, **Chunyang Li***, Weiqi Wang*, Wei Fan, Tianshi Zheng, Haochen Shi, Tao Fan, Yangqiu Song, and Qiang Yang, "Towards Multi-Agent Reasoning Systems for Collaborative Expertise Delegation: An Exploratory Design Study", *Under Review.*

[3] Jiaxin Bai*, Wei Fan*, Qi Hu*, Qing Zong, **Chunyang Li**, Hong Ting Tsang, Hongyu Luo, Yauwai Yim, Haoyu Huang, Xiao Zhou, Feng Qin, Tianshi Zheng, Xi Peng, Xin Yao, Huiwen Yang, Leijie Wu, Yi Ji, Gong Zhang, Renhai Chen, and Yangqiu Song, "AutoSchemaKG: Autonomous Knowledge Graph Construction through Dynamic Schema Induction from Web-Scale Corpora", *Under Review.*

[4] Hao Peng*, Xiaozhi Wang*, **Chunyang Li**, Kaisheng Zeng, Jiangshan Duo, Yixin Cao, Lei Hou, and Juanzi Li, "Event-level Knowledge Editing", *Arxiv Preprint.*

[5] Shangqing Tu*, **Chunyang Li***, Jifan Yu, Xiaozhi Wang, Lei Hou, and Juanzi Li, "ChatLog: Carefully Evaluating the Evolution of ChatGPT Across Time", *Arxiv Preprint.*

* indicates equal contributions.

EXPERIENCES

Research Intern @ **Tencent AI Lab** *Jun 2025 - Present*

Mentor: Dr. Xinting Huang and Dr. Han Hu.

Worked on the evaluation of Agent/LLM-as-a-Judge.

Undergraduate Researcher @ **THUKEG, Tsinghua University** *Jun 2022 - Jun 2024*

Advised by Prof. Juanzi Li and Prof. Lei Hou.

Worked on multiple topics in natural language processing, including:

- **Cognitive Evaluation for LLM**: involved in constructing a Knowledge-oriented LLM benchmark. Proposed a dataset of responses from ChatGPT across time with extracted features to explore the evolving pattern of LLM.
- **Knowledge-driven AI in Education**: participated in the development of a Virtual Teaching Assistant, *LittleMu*, which has served more than 80,000 users with over 300,000 queries from over 500 courses.
- **Event Understanding**: introduced the largest event factuality detection dataset with an LLM-then-human annotation approach and conducted a detailed and thorough analysis of evaluation results on representative models.

Summer Intern @ **Zhipu AI** *Jun 2023 - Jul 2023*

Completed as part of the Professional Practice course requirement at Tsinghua University. Focused on enhancing visualization tools for evaluating large language model agents.

TEACHING

Teaching Assistant of COMP 4332/RMBI 4310 @ HKUST *Spring 2025*

PROFESSIONAL SERVICES

Conference Reviewer: ACL Rolling Review (2024, 2025), COLM (2025), ICLR(2026).

AWARDS & HONORS

Outstanding Graduate, Department of Computer Science and Technology, Tsinghua University	2024
Academic Excellence Scholarship, Tsinghua University	2022, 2023
Social Practice Scholarship, Tsinghua University	2022
Freshman Scholarship, Tsinghua University	2020

SKILLS

Programming Skills	Python(PyTorch), C/C++, Java, Verilog.
Language Skills	*Chinese*(native), *English*(fluent).