香港浸會大學
HONG KONG BAPTIST UNIVERSITY

DEPARTMENT OF
COMPUTER SCIENCE
計算機科學系

TMLR Young Scientist SEMINAR

2026 SERIES

Trustworthy Machine Learning and Reasoning Group

Dr. Lin Li

Postdoctoral Researcher,
OATML Group,
Oxford University

Date: 13 October 2026 (Tuesday)
Time: 16:30 – 18:30 (HKT)
Meeting: https://hkbu.zoom.us/j/6603117755

Gaming AI– Assisted Peer Reviews Poses New Risk to the Scientific Community

ABSTRACT

AI-assisted peer review promises to reduce reviewer burden, yet its robustness to strategic manipulation remains poorly understood. In this talk, we show that AI-mediated review is vulnerable to a simple, low-cost manipulation: superficial rephrasing of the manuscript abstract. Without changing underlying scientific content, adversarially rewritten abstracts substantially improve AI-generated outcomes, increasing acceptance ratings by up to +1.31 for Gemini 3 flash and +0.88 for GPT 5.4mini.

These findings reveal a systemic vulnerability where inflated AI reviews can bias downstream human decision-making. More broadly, our results suggest that authors may be incentivized to optimize manuscripts for AI judgment rather than scientific merit, highlighting the urgent need for systematic robustness testing and careful human oversight in AI-integrated evaluation.
arXiv link: https://arxiv.org/pdf/2606.10159

BIOGRAPHY

Lin Li is a Postdoctoral Researcher in the Department of Computer Science at the University of Oxford, where he works with Prof. Yarin Gal, and a Junior Research Fellow at Jesus College, Oxford. He leads a research stream on multimodal foundation model safety within the Horizon Europe project, DVPS. He is a Co-Principal Investigator on a research grant from Coefficient Giving to develop large language models with realistic worst-case robustness guarantees. His research aims to build safe AI systems through advances in robustness, uncertainty quantification, and mechanistic interpretability. His work has been published in top AI/ML venues such as ICLR, ICML, ACL, EMNLP, CVPR, IJCV. Web page: https://treelli.github.io.

ENQUIRY

Email: bhanml@comp.hkbu.edu.hk