Sicheng Lai

+86 15820778080 122090246@link.cuhk.edu.cn https://sichenglai.com Leo-Lsc

EDUCATION

The Chinese University of Hong Kong, Shenzhen *Sept. 2022 – June 2026*
B.Sc. Computer Science and Engineering *GPA: 3.679 / 4.0*

Exchange Semester in University of California San Diego *March 2025 – July 2025*

PUBLICATIONS

Yadi Cao†, **Sicheng Lai**†, Jiahe Huang†, Yang Zhang†, Zach Lawrence†, Rohan Bhakta†, Izzy F. Thomas†, Mingyun Cao†, Chung-Hao Tsai, Zihao Zhou, Yidong Zhao, Hao Liu, Alessandro Marinoni, Alexey Arefiev, and Rose Yu*

"SimulCost: A Cost-Aware Benchmark for Automating Physics Simulations with LLMs".

arXiv preprint. *Project Website*.

Sicheng Lai†, Dingjie Song†, Mingxuan Wang, Shunian Chen, Lichao Sun, and Benyou Wang*

"Both Text and Images Leaked! A Systematic Analysis of Data Contamination in Multimodal LLM".

ICML 2025 DIG-BUGs Oral. *EMNLP 2025 Findings*.

† These authors contributed equally. * Corresponding author.

RESEARCH EXPERIENCE

SimulCost: A Cost-Aware Playground
April 2025 - Present
UC San Diego | Co-First Author
Supervisor: Prof. Rose Yu

- Developed SimulCost, a comprehensive benchmarking framework evaluating LLM cost-accuracy trade-off on scientific simulation optimization across 13 physics solvers and 40+ hyperparameter tuning tasks.
- Engineered a scalable Playground API enabling on-demand simulation generation, with caching mechanisms and pandas DataFrame-based data persistence reducing redundant computations by 50%.
- Benchmarked LLMs against brute-force (e.g., grid search) and Bayesian optimization (e.g., Gaussian process) baselines across 6 real-world scenarios, conducting statistical analysis and ablation studies.

Data Contamination in Multimodal LLMs
June 2024 - December 2024
CUHK-Shenzhen | Co-First Author
Supervisor: Prof. Benyou Wang

- Pioneered systematic multimodal data contamination detection by designing the MM-Detect framework, incorporating two novel methods: *Option Order Sensitivity Test* and *Slot Guessing for Perturbed Caption*.
- Validated the framework's effectiveness and sensitivity by training LLaVA-1.5-7B variants with controlled leakage at early, mid, and late training stages, and varying contamination ratios (10%, 50%, 100%).
- Developed a heuristic method to trace contamination origins, revealing unimodal contamination in base LLMs and cross-modal contamination through training data overlap analysis.

INTERNSHIPS

Multimodal Algorithm Engineer
December 2024 - March 2025
Tsinghua Shenzhen International Graduate School
Shenzhen, China

- Optimized MLLM configurations through systematic experimentation with various combinations of LLMs (*Qwen2.5/Vicuna*) and visual encoders (*CLIP/MLCD/SigLIP*) based on the *LLaVA-Next* framework.

- Evaluated model performance using *VLMEvalKit* and *lmms-eval* to determine the optimal configuration (*Qwen2.5 + SigLIP*), achieving top 20% ranking on the *OpenCompass* MLLM Leaderboard.
- Trained the *Qwen2.5 + SigLIP* model on agricultural datasets curated and augmented with *data-juicer*, achieving expert-level agricultural capabilities while maintaining high general-domain performance.

HONORS

Dean's List *(Outstanding academic merit)* 2022 – 2025
Undergraduate Research Awards *(Key research contributions)* 2024 – 2025

TEACHING

Undergraduate Student Teaching Fellow (STA2001: Probability and Statistics I) Spring 2024