

Sicheng Lai

+86 15820778080

122090246@link.cuhk.edu.cn

<https://sichenglai.com>

Leo-Lsc

EDUCATION

The Hong Kong University of Science and Technology

Sept. 2026 – Expected June 2027

M.Sc. in Big Data Technology

The Chinese University of Hong Kong, Shenzhen

Sept. 2022 – June 2026

B.Sc. in Computer Science and Engineering

Exchange Semester in University of California San Diego

March 2025 – July 2025

PUBLICATIONS

† Equal Contribution. * Corresponding author.

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](#).

RESEARCH EXPERIENCE

SimulCost: A Cost-Aware Playground

April 2025 - March 2026

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
Tsinghua Shenzhen International Graduate School

December 2024 - March 2025
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.

PROJECTS

Transformer Language Modeling Stack from Scratch
Stanford University

March 2026 - Present
Spring 2026

- Built an end-to-end GPT-style language modeling stack from first principles, covering byte-level BPE tokenization, decoder-only Transformer modeling, optimization, training, autoregressive generation.
- Designed an efficient tokenizer pipeline with GPT-2 regex pre-tokenization, multiprocessing chunking around special tokens, and incremental pair-count updates for faster BPE merges.
- Implemented core Transformer modules manually in PyTorch, including Linear/Embedding, RMSNorm, SwiGLU, RoPE-based causal self-attention, and AdamW.
- Engineered a configurable pre-training pipeline for TinyStories and OpenWebText with memory-mapped data loading, warmup + cosine LR decay, gradient clipping, and WandB tracking.

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

CORE COMPETENCIES

- **Programming & Systems:** Python, Java, C/C++ , RISC-V
- **LLM:** PyTorch, Transformers, LangChain, Accelerate, Triton, DeepSpeed, bitsandbytes, vLLM, SGLang
- **Data Science:** NumPy, pandas, SciPy, Dask, xarray, Matplotlib, Seaborn, Weights & Biases
- **Developer Tools & Deployment:** Hydra, Git, Conda, Poetry, uv, Docker, Kubernetes, SLURM, Linux
- **Backend & Web:** Flask, REST APIs, SQLite, HTML/CSS/JavaScript, TypeScript, Astro, Tailwind CSS