

Yang Runming

(+86) 17362995487 $\diamond$ yang3121099@126.com $\diamond$ <https://rummyyang.github.io/>

Research interests: LLM post-training; Knowledge distillation; Efficient reasoning; Low-bit quantization

EDUCATION

The University of Hong Kong

Oct. 2025 - Present

Ph.D. student in Electrical and Electronic Engineering — Advisor: Prof. Ngai Wong; expected Sep. 2029

Tsinghua University

Sep. 2022 - Jun. 2025

M.Eng. in Big Data Engineering — Advisor: Prof. Yujiu Yang; GPA: 3.73/4.00

Huazhong University of Science and Technology

Sep. 2017 - Jun. 2021

B.Eng. in AI and Automation — Science and Engineering Cross-Innovation Experimental Class

PUBLICATIONS

* Equal contribution.

- [1] T. Wu*, R. Yang*, T. Liu, J. Wang, N. Wong. *Revisiting Model Interpolation for Efficient Reasoning*. **ACL 2026 (Main)** [Paper] [Code]
Interpolates fast- and slow-thinking models without training to balance reasoning quality and chain-of-thought length.
- [2] T. Liu*, T. Wu*, R. Yang, S. Sun, J. Wang, Y. Yang. *ProFit: Leveraging High-Value Signals in SFT via Probability-Guided Token Selection*. **ACL 2026 (Findings)** [Paper] [Code]
Uses current-policy information to guide token selection and relax token-level alignment during supervised fine-tuning.
- [3] T. Wu*, Y. Cheng*, C. Ding*, R. Yang, X. Feng, W. Zhou, Z. Liu, N. Wong. *Can We Trust LLMs on Memristors? Diving into Reasoning Ability under Non-Ideality*. **COLM 2026** [Paper]
Simulates how memristor non-idealities affect LLM reasoning and evaluates training-free mitigation strategies.
- [4] H. Xiao, Q. Yang, D. Xie, W. Xu, Z. Su, R. Yang, H. Liu, W. Zhou, Z. Liu, N. Wong. *Exploring Layer-wise Information Effectiveness for Post-Training Quantization in Small Language Models*. **ACL 2026 (Findings)** [Paper] [Project]
Uses layer-wise information effectiveness to allocate mixed precision across layers for low-bit quantization of small language models.
- [5] T. Wu, C. Tao, J. Wang, R. Yang, Z. Zhao, N. Wong. *Rethinking Kullback-Leibler Divergence in Knowledge Distillation for Large Language Models*. **COLING 2025 (Oral)** [Paper] [Code]
Compares forward and reverse KL divergence for LLM distillation and introduces Adaptive KL.
- [6] R. Yang, T. Wu, Y. Yang. *LoCa: Logit Calibration for Knowledge Distillation*. **ECAI 2024** [Paper]
Calibrates teacher logits for misclassified samples during knowledge distillation.

RESEARCH INTERNSHIP

Tencent AI Lab

Nov. 2023 - Jun. 2025

Research Intern, Information Security Department

- Synthesized multimodal data and conducted model post-training and evaluation for image-text content moderation in QQ group chats, Qzone, and short-video comments.
- Worked on tiered deployment and inference acceleration; trained models were deployed in internal risk-control systems, improving throughput and reducing false-positive rates over the previous solution.
- Researched large/small model collaboration, data distillation, and lightweight modeling; contributed to two academic papers (Shadow-FT and LLM-NEO) and one company patent.

ONGOING RESEARCH

CoT Data and Efficient Reasoning Post-training

Ongoing

Collaborative Ph.D. research at the University of Hong Kong

- Planned approach: compare offline CoT fine-tuning with on-policy distillation on student-generated trajectories, examining data quality and teacher-student distribution alignment.
- Research objective: improve reasoning quality and efficiency under a controlled compute budget, with planned evaluation of accuracy, reasoning length, and training cost; manuscript in preparation.

PREPRINTS AND SUBMITTED MANUSCRIPTS

* Equal contribution.

- [P1] **R. Yang***, T. Wu*, J. Li, P. Hu, Y.-C. Wu, N. Wong, Y. Yang. *Shadow-FT: Tuning Instruct Model via Training on Paired Base Model*. **Submitted to NAACL 2027** [Paper] [Code]
Transfers fine-tuning updates from a base model to its paired instruction-tuned model.
- [P2] Q. Yang, **R. Yang**, H. Xiao, W. Xu, J. Chen, H. Liu, C. Ding, R. Hu, Y.-C. Wu, N. Wong. *FluxBin: Flexible LUT-based Ultra-low-bit LLM Inference by Algorithm-Kernel Synergy*. **Submitted to NAACL 2027** [Paper] [Code]
Co-designs ultra-low-bit post-training quantization and lookup-table CUDA kernels for efficient LLM inference.
- [P3] T. Wu, **R. Yang**, T. Liu, J. Wang, Z. Xu, N. Wong. *Timber: Training-free Instruct Model Refining with Base Model via Effective Rank*. **Submitted to NAACL 2027** [Paper] [Code]
Uses effective rank to analyze post-training weight changes and refine instruction-tuned models without additional training.
- [P4] Y. Zhang, T. Wu, **R. Yang**, T. Liu, J. Wang, Y. Yang. *From Answers to Obligations: Benchmarking Financial Instruction Following with FinIF*. **Submitted to EMNLP 2027**
Studies instruction-following evaluation in financial tasks, focusing on whether model responses satisfy task requirements.
- [P5] Z. Li*, Y. Su*, S. Wang, **R. Yang**, C. Xie, A. Liu, M. Li, J. Cao, Y. Xie, N. Wong, H. Yang. *Quantization Meets Reasoning: Exploring and Mitigating Degradation of Low-Bit LLMs in Mathematical Reasoning*. arXiv:2505.11574 [Paper]
Analyzes step-level reasoning errors in quantized LLMs and uses targeted preference pairs to improve 4-bit mathematical reasoning.
- [P6] T. Liu*, T. Wu*, M. Zheng, X. Luo, **R. Yang**, X. Yang, J. Wang, Y. Yang. *Simple-OPD: Demystifying Warm-up for On-policy Distillation*. **Submitted to NAACL 2027** [Paper] [Code]
Studies warm-up for on-policy distillation and initializes students with LoRA on teacher-generated chains of thought.
- [P7] **R. Yang***, T. Wu*, J. Wang, P. Hu, Y.-C. Wu, N. Wong, Y. Yang. *LLM-NEO: Parameter Efficient Knowledge Distillation for Large Language Models*. arXiv:2411.06839 [Paper] [Code]
Combines low-rank adaptation and knowledge distillation for parameter-efficient knowledge transfer from large teachers to compact students.
- [P8] H. Xiao*, **R. Yang***, Q. Yang*, W. Xu, Z. Li, Y. Su, Z. Liu, H. Yang, N. Wong. *PTQTP: Post-Training Quantization to Trit-Planes for Large Language Models*. arXiv:2509.16989 [Paper] [Project]
Decomposes LLM weights into two ternary trit-planes with lookup-table kernels for efficient inference.