

Yang Runming

Tel: (+86) 17362995487 ◊ yang3121099@126.com ◊ <https://rummyyang.github.io/>

EDUCATION

The University of Hong Kong, Ph.D. student in Electrical and Electronic Engineering Oct. 2025 - Present
Advisor: [Prof. Ngai Wong](#) — Research: LLM Efficiency and Post-Training

Tsinghua University, M.Eng. in Big Data Engineering (GPA: 3.73/4.00) Sep. 2022 - Jun. 2025
Advisor: [Prof. Yujiu Yang](#) — Research: Model Compression and Knowledge Distillation

Huazhong University of Science and Technology, B.Eng. in AI and Automation Sep. 2017 - Jun. 2021
Science and Engineering Cross-Innovation Experimental Class

INTERNSHIP

Tencent PCG Social Secure Center, Algorithm Intern Nov. 2023 - Jun. 2025
Built domain-specific multimodal datasets; conducted model post-training and evaluation; investigated cross-modal feature fusion.

PUBLICATIONS

* indicates equal contribution.

- T. Wu, **R. Yang**, T. Liu, J. Wang, N. Wong. *Revisiting Model Interpolation for Efficient Reasoning*. ACL 2026 (Main) [\[Paper\]](#) [\[Code\]](#)
A training-free, three-stage interpolation framework between fast- and slow-thinking models for shorter chains of thought.
- 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.
- 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 [\[Paper\]](#)
Studies how memristor non-idealities affect LLM reasoning and evaluates training-free mitigation strategies.
- 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.
- 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.
- **R. Yang**, T. Wu, Y. Yang. *LoCa: Logit Calibration for Knowledge Distillation*. ECAI 2024 [\[Paper\]](#)
Calibrates teacher logits for misclassified samples during knowledge distillation.

PREPRINTS

* indicates equal contribution.

- 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*. arXiv:2608.15602 (2026) [\[Paper\]](#) [\[Code\]](#)
Co-designs ultra-low-bit post-training quantization and lookup-table CUDA kernels for efficient LLM inference.
- 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*. arXiv:2608.06802 (2026) [\[Paper\]](#) [\[Code\]](#)
Studies warm-up for on-policy distillation and initializes students with LoRA on teacher-generated chains of thought.
- T. Wu*, **R. Yang***, J. Li, P. Hu, Y.-C. Wu, N. Wong, Y. Yang. *Shadow-FT: Tuning Instruct Model via Training on Paired Base Model*. arXiv:2505.12716 (2025) [\[Paper\]](#) [\[Code\]](#)
Transfers fine-tuning updates from a base model to its paired instruction-tuned model.
- T. Wu, **R. Yang**, T. Liu, J. Wang, Z. Xu, N. Wong. *Timber: Training-free Instruct Model Refining with Base via Effective Rank*. arXiv:2509.23595 (2025) [\[Paper\]](#) [\[Code\]](#)
Uses effective rank to analyze post-training weight changes and refine instruction-tuned models without additional training.

- 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 (2025) [[Paper](#)] [[Project](#)]
Decomposes LLM weights into two ternary trit-planes for multiplication-free inference.
- 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 (2025; revised 2026) [[Paper](#)]
Analyzes step-level reasoning errors in quantized LLMs and uses targeted preference pairs to improve 4-bit mathematical reasoning.
- **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 (2024; revised 2025) [[Paper](#)] [[Code](#)]
Combines low-rank adaptation and knowledge distillation for parameter-efficient knowledge transfer from large teachers to compact students.