Tao Luo

Agent Systems | Post-Training Infrastructure | LLM Inference

LinkedIn · github/taoluo · taoluo.net · taoluo71@seas.upenn.edu

Summary

CS Ph.D. at UPenn building **AI infrastructure** for agentic RL post-training, LLM inference, and embedding retrieval. Shipped production RL GPU orchestration at Alibaba for **100s-of-billion-parameter** models on **1000s of GPUs**; founded open-source **RLix** (270+ GitHub stars). Publications at OSDI, SOSP, SIGCOMM, SoCC.

Education

- **Ph.D. in Computer Science** *09/2021–08/2026 (expected)*

  University of Pennsylvania – Distributed Systems Lab · Philadelphia
  - Research areas: AI infra (LLM inference, agentic RL), and data systems (embedding retrieval, query optimization).
  - Mentored: Sidharth Sankhe (→ UC Berkeley Ph.D.), Zhen Ping Khor (→ UPenn Ph.D.), Lan Lu (UPenn Ph.D.), …

- **M.S. study in Computer Science** *08/2019–05/2021*

  Columbia University · New York
  - Published **3 top-tier systems papers** on privacy-preserving ML systems and measuring internet structure.

- **B.S. in Financial Mathematics** *03/2011–01/2015*

  Southern University of Science and Technology · Shenzhen
  - Founding cohort member.

Industry Experience

- **Research Engineer Intern** *06/2025–01/2026*

  Alibaba Group – DAMO Academy, Post-Training · Remote, then Beijing
  - Independently designed and shipped **Partial Overlapping**, a GPU scheduler for agentic RL, into alibaba/ROLL.
    - Pinpointed idle training GPUs as the bottleneck in Alibaba's async agentic RL pipeline.
    - Designed and built a scheduler that opportunistically places rollouts as preemptible workloads on idle training GPUs, improving rollout throughput by **3.5x**.
    - Powers production agentic RL post-training of **100s-of-billion-parameter** models on **1000s of GPUs**; first used for Alibaba's ROME launch, then across four product teams (Qoder, iFlow, Amap, Alimama).
    - Featured in the technical report as part of Alibaba's open-source Agentic Learning Ecosystem; progress reviewed by senior Alibaba leadership (skip-3).
  - Founded and lead **RLix**, a GPU-sharing orchestration framework multiplying RL post-training throughput.
    - Delivered **2.6x** rollout throughput in SWE-agent RL training via **recipe-transparent** elastic GPU sharing.
    - Designed the **priority-based scheduling algorithm**: rollout runs as the lowest-priority preemptible stage on idle GPUs, yielding to higher-priority training stages on demand.
    - Built the **selective weight-sync mechanism**: syncs latest weights only to scheduled rollout workers, keeping each RL pipeline's memory footprint minimal.
    - **270+ GitHub stars** from NVIDIA, Google, xAI, Anthropic, Fireworks; define the roadmap and build features alongside ∼10 community contributors, including early integrations with NVIDIA NeMo and MILES.
  - Introduced AI-assisted engineering on the team, raising shipping velocity.
    - Built Partial Overlapping primarily using coding agents (Claude, Codex); first high-priority alibaba/ROLL feature shipped this way.
    - Authored the team's internal playbook; featured as the case study in the team's public and internal technical blog posts.

- **Algorithm Engineer** *12/2017–04/2019*

  DataYes Inc. – R&D Center · Shanghai
  - Shipped production quant investment models across factor analysis, portfolio optimization, and asset allocation.

- **Quantitative Research Engineer** *05/2015–12/2017*

  Infore Capital Co. – Quantitative Investment Dept. · Shenzhen
  - Built quant research infrastructure: data pipelines, feature engineering, backtesting, and risk analytics.

Research Projects

- **Heterogeneous Multi-Model LLM Serving at Scale** *07/2023–07/2025*
  Published in SoCC'25
  - Designed and built a GPU-sharing serving system with **stage-aligned parallelism**, eliminating head-of-line blocking and increasing token throughput by **1.6x** while reducing median latency.
  - Extended vLLM with multi-model KV cache management, NCCL concurrency controls, and distributed execution.
  - Developed algorithms for efficient model sharding, replication, placement, and scheduling.
- **Query Optimization for Declarative Smart Contracts** *06/2022–06/2023*
  Published in FAB'24 (co-located with VLDB)
  - Modeled Datalog-compiled smart contract efficiency as view materialization under a gas-based cost model.
  - Designed and implemented a **selective view materialization** algorithm with simplification-based pruning; formally proved correctness and pruning completeness.
  - Reduced storage gas by $\sim$**78%** and total gas by $>$**50%** over naive compilation, on par with hand-tuned implementations of widely deployed contracts.
- **Privacy-Preserving Scheduling for ML Training** *08/2020–07/2021*
  Published in OSDI'21
  - Designed the first **fair-allocation** scheduling algorithm for ML training under differential-privacy constraints.
  - Improved job throughput by **2x over FCFS** under the same privacy budget, verified in large-scale simulations.
  - Proved formal efficiency and fairness guarantees.

Publications

* indicates equal contribution

Let It Flow: Agentic Crafting on Rock and Roll, Building the ROME Model within an Open Agentic Learning Ecosystem. Weixun Wang, XiaoXiao Xu, ..., **Tao Luo**, et al. *Technical Report, arXiv:2512.24873*

ScaleGANN: Scalable Vector Database Index Construction on Cost-Effective Cloud GPUs. Lan Lu, Peiqi Yin, **Tao Luo**, Isaac Yang, Hua Fan, Wenchao Zhou, Feifei Li, Boon Thau Loo. *Under submission*

Connex: Endpoint Mobility Primitives for Dynamic LLM Serving. Yanying Lin, Vincent Liu, **Tao Luo**, ChengZhong Xu, Kejiang Ye. *SIGCOMM'26: Annual Conference of the ACM Special Interest Group on Data Communication*

ParaFlex: Multiplexed Heterogeneous LLM Serving via Stage-Aligned Parallelism. **Tao Luo**, Kelvin Ng, Zhen Ping Khor, Sidharth Sankhe, Boon Thau Loo, Vincent Liu. *SoCC'25: ACM Symposium on Cloud Computing*

Practical Declarative Smart Contracts Optimization. Lan Lu, **Tao Luo**, Jingyi Li, et al. *FAB'24: Sixth International Workshop on Foundations and Applications of Blockchain (co-located with VLDB).*

Arboretum: A Planner for Large-Scale Federated Analytics with Differential Privacy. Elizabeth Margolin*, Karan Newatia*, **Tao Luo**, Edo Roth, Andreas Haeberlen. *SOSP'23: Symposium on Operating Systems Principles*

Privacy Budget Scheduling. **Tao Luo***, Mingen Pan*, Pierre Tholoniat*, Asaf Cidon, and Roxana Geambasu, Mathias Lécuyer. *OSDI'21: USENIX Symposium on Operating Systems Design and Implementation*

Towards Identifying Networks with Internet Clients Using Public Data. Weifan Jiang*, **Tao Luo***, Thomas Koch, et al. *IMC'21: Proceedings of the 21st ACM Internet Measurement Conference*

Towards an Internet Traffic Map. T. Koch, W. Jiang, **T. Luo,** et al. *HotNets'21: ACM Workshop on Hot Topics in Networks*

Honors and Awards

- **Technical Program Committee Member**, ACM Symposium on Cloud Computing (SoCC) *2025*
- **Manjushri Fellowship**, University of Pennsylvania *2021*
- **China Merchant Bank Scholarship**, Southern University of Science and Technology *2012–2014*
- **Pioneering Undergraduate Fellowship**, Southern University of Science and Technology *2011–2014*

Technical Skills

- **Agent and LLM Systems:** RL job orchestration, agent rollouts, LLM inference, LoRA fine-tuning, LLM training
- **Retrieval and Data:** vector search, vector database indexing, query optimization
- **AI-Assisted Engineering:** coding-agent workflows across design, build, test, and review
- **Frameworks:** Alibaba ROLL, vLLM, Megatron-LM, PyTorch, Ray, NCCL
- **Languages:** Python (advanced), C++

References

Boon Thau Loo, Vincent Liu (Co-Advisors, UPenn) · Asaf Cidon (M.S. Advisor, Columbia) · Shaopan Xiong (Mentor, Alibaba)