Runguo Li

runguol2@illinois.edu | runguoli.com | github.com/LiRunGuo | Champaign (USA)

Research Interests

ML Systems for large-scale models: inference engines and expert streaming for MoE models, distributed training and rollout infrastructure (ZeRO-3, FSDP, Hybrid Engine), RL post-training systems, and GPU compiler and kernel-level correctness. I work between the machine and the model — how weights are read from storage, how collectives and thread teams are sized, and whether a resource plan and the runtime agree about what will actually be built — while keeping a research background in LLM reasoning, multimodal learning, and agent safety.

Education

University of Illinois Urbana-Champaign (UIUC) Aug 2026–Jul 2028
M.S. in Information Science *Urbana-Champaign, IL, USA*

Shanghai University of Finance and Economics (SUFE) Sep 2022–Jun 2026
B.S. in Business Analytics *Shanghai, China*

Open-Source Contributions (ML Systems)

Merged fixes to a pure-C MoE inference engine, DeepSpeed, Triton, verl, and LMCache, plus open pull requests across the wider AI infrastructure stack.

colibri — pure-C MoE inference engine Merged pull requests

multi-drive expert streaming for the 510 GB DeepSeek-V4.1 container, physical-core OpenMP team sizing in four engines that were 18.7× slower without it, and a plan/runtime mismatch that silently allocated 6.25 GiB of unbudgeted KV cache

DeepSpeed — distributed training Merged pull requests

fixed a ZeRO-3 rollout deadlock caused by unsynchronized generation stopping, and blocked partial Hybrid Engine policy injection for unsupported architectures (validated on H200 and MI250 GPUs)

Triton — GPU compiler Merged pull requests

stopped nested-loop fusion from trusting an `llvm.assume` outside the loop, which removed a zero-trip guard and could cause incorrect memory writes (MLIR regression test and H200 reproducer)

verl — RL post-training Merged pull requests

restored FSDP value-head critic loading after TRL relocated its value-head model classes

LMCache — KV cache management Merged pull requests

enforced lazy %-format logging (ruff G004) so new f-string logging can no longer land in already-migrated files

Open pull requests. PyTorch #196787, vLLM #56198, SGLang #39315, Megatron-LM #7287, TensorRT-LLM #18904, CUTLASS #3613, LMCache #5117, DeepSpeed #8390, DeepSpeed #8392, Miles #3266, Miles #3267, Miles #3268, Miles #3269, Cordis #148, Open Code Review #1351. Full list

Publications

(*Co-first author)

1. Zhi Yang*, **Runguo Li***, Qiqi Qiang, Jiashun Wang, Fangqi Lou, Mengping Li, Dongpo Cheng, et al. "**FinVault: Benchmarking Financial Agent Safety in Execution-Grounded Environments**. *arXiv preprint arXiv:2601.07853*, 2026.
The first execution-grounded safety benchmark for financial agents: 31 regulatory sandbox scenarios with writable state, 107 real-world vulnerabilities, and 963 test cases. Current defenses do not transfer — attack success rates reach 50.0% and stay non-negligible (6.7%) even for the most robust model.
PDF | Project page

2. **Runguo Li*** et al. "**Reasoning-Visual Critical Token Fine-Tuning for Multimodal Reasoning**. Manuscript under review at AAAI 2027.
RVCFT selects which chain-of-thought tokens receive direct supervision using reasoning-relevance and visual-sensitivity signals while retaining the full response as context; at 50% token retention it gives the best average among the evaluated multimodal reasoning benchmarks.
PDF | Project page

3. **Runguo Li*** et al. "**VeriBRT: Plan-Guided, Evidence-Based Automated Bug Reproduction**. Manuscript under review at ICSE 2027.
Represents issue-report behavior as a persistent Two-Axis Plan (a Checklist for where a test must reach, an Intentlist for what it must verify) with reliability-aware evidence and localization coverage as a validity criterion; reproduces 341 of 433 issues on SWT-Bench Verified, 9.7 points above the strongest same-model baseline.
PDF | Project page

Research Experience

University of Illinois Urbana-Champaign (UIUC) Aug 2026–Present
Research Intern, advised by Prof. Minjia Zhang *Urbana-Champaign, IL, USA*

- Research on efficient machine learning systems for large models: training and inference efficiency, model compression and sparsity, kernel- and hardware-aware optimization, and system support for LLM agents.
- Bring an upstream-first practice to this work: reproduce failures in production inference and training stacks, then contribute fixes with regression tests and measured before/after numbers.

Shanghai Jiao Tong University Mar 2026–Jul 2026
Research Intern, advised by Prof. Xiaodong Gu *Shanghai, China*

- Researched LLMs for code, with an emphasis on repository-level program understanding, automated bug reproduction, and coding-agent planning and execution.
- Developed **VeriBRT**, a plan-guided framework that represents issue-report behavior with a persistent Two-Axis Plan and uses reliability-aware evidence for bug-reproducing test generation; set up the evaluation harness on SWT-Bench Verified.

SUFE FinAI Center Jul 2025–Jul 2026
Research Intern, advised by Prof. Liwen Zhang *Shanghai, China*

- Co-developed **FinVault**, an execution-grounded benchmark for financial-agent safety with 31 regulatory scenarios, 107 real-world vulnerabilities, and 963 test cases.
- Built the executable sandboxes behind the benchmark — writable databases, permission and quota constraints, toolchains, and audit logs — so safety failures are judged from verified state transitions rather than generated text.
- Studied financial reasoning and agent safety through chain-of-thought data curation, supervised fine-tuning, GRPO training, and evaluation on FinEval, FinQA, and ConvFinQA.

Tencent Youtu AI Lab Oct 2025–May 2026
Research Intern *Shanghai, China*

- Researched LLM-driven adversarial data synthesis and low-resource, multi-label content-safety evaluation, and contributed to a unified multilingual moderation model on an XLM-RoBERTa backbone.
- Developed **RVCFT**, a selective-supervision method that identifies reasoning-relevant and visually sensitive tokens for multimodal chain-of-thought fine-tuning.

Selected Open-Source Project

ARH — AI Research Helper github.com/LiRunGuo/Arhelper (MIT, 0.2.0-alpha)

- A CLI-first research assistant runtime: tool calling, plan/execute safety with approval-gated writes, three-layer memory (preference extraction, BM25 cross-session recall, reflection summaries), skill self-learning, checkpoint/resume, and automatic context compression.
- Multi-LLM routing with cooldown-based failover across OpenAI, Anthropic, and Ollama; Hub–Spoke gateways (CLI, Telegram, FastAPI); research-native tools for arXiv, HuggingFace, GitHub, and OpenReview. Python 3.9+, Pydantic, FastAPI, SQLAlchemy; all state local under `~/.arh/`.

Technical Skills

- **ML Systems and Inference**: MoE expert streaming and disk-resident inference, KV-cache and memory budgeting, attention and quantization kernels, serving engines (vLLM, SGLang, TensorRT-LLM), CUDA/Triton kernel-level debugging, GPU compiler passes (MLIR), and performance profiling.
- **Training and Post-Training Infrastructure**: distributed data/tensor/pipeline parallelism, ZeRO-3 and FSDP, Megatron-LM, DeepSpeed Hybrid Engine, collective synchronization and deadlock debugging, RL rollout systems (verl, TRL), SFT and preference optimization (DPO/GRPO/PPO), distillation, and LoRA/PEFT.
- **Languages and Tooling**: Python, C, CUDA/Triton, SQL, Bash, LaTeX; PyTorch, Transformers, FlashAttention, FAISS; Git-based upstream contribution, regression testing, reproducible benchmarking, Linux, and Docker.
- **Systems for Agents**: agentic planning and tool use, RAG and long-term memory, execution-grounded evaluation, runtime safety gates, and multi-channel gateways (FastAPI).

Honors & Awards

National Encouragement Scholarship | People's Scholarship (2nd Class) | Silver Medal, FLTRP English Competition (Municipal)