Lei Yang

Master's Student, Tianjin University | RL for LLM Reasoning and Agents

Homepage (better reading experience) | yl.shadow.yl@gmail.com | Google Scholar | GitHub

RESEARCH INTERESTS

My research goal is to make large language models **generalizable and capable** through post-training. I currently focus on **reinforcement learning for LLM reasoning** — how policies explore and why entropy collapses, where cross-domain interference arises and how models recover, and how to keep training stable — as well as **long-horizon code agents**, from data construction to high-fidelity evaluation. Looking ahead, I aim to develop RL algorithms that enable agents to plan, explore, and recover over long trajectories. This agenda is grounded in full-stack LLM development experience, including continual pre-training, dense-to-MoE upcycling, long-context extension, and low-resource multilingual modeling.

EDUCATION

Tianjin University — Tianjin, China

M.Eng. in Computer Technology — Sep 2024 – Jun 2027 (expected)
Advisor: Prof. Deyi Xiong, TJUNLP Lab

Tiangong University — Tianjin, China

B.Eng. in Network Engineering — Sep 2020 – Jun 2024

PUBLICATIONS

* denotes equal contribution.

1. **Lei Yang**, Siyu Ding, Deyi Xiong. A Local Perturbation Theory for Cross-Domain Interference and Recovery in Multi-Domain RL. Under review, 2026.
2. **Lei Yang**, Wei Bi, Chenxi Sun, Renren Jin, Deyi Xiong. Thinking Seeds: Leveraging Historical Diversity for Position-Aware RL in LLMs. Under review, 2026.
3. **Lei Yang***, Leiyu Pan*, Bojian Xiong, Renren Jin, Shaowei Zhang, Yue Chen, Ling Shi, Jiang Zhou, Junru Wu, Zhen Wang, Jianxiang Peng, Juesi Xiao, Tianyu Dong, Zhuowen Han, Zhuo Chen, Yuqi Ren, Deyi Xiong. From Curated Data to Scalable Models: Continual Pre-training of Dense and MoE Large Language Models for Tibetan. **ACL 2026 (Oral)**.
4. Zhuowen Han, **Lei Yang**, Renren Jin, Dan Shi, Chenxi Sun, Deyi Xiong. ERRV: Eliciting Efficient Reasoning through Reasoning Vectors for Policy Optimization in Large Language Models. **Findings of ACL 2026**.
5. Pengyun Zhu, Yuqi Ren, Zhen Wang, **Lei Yang**, Deyi Xiong. DVMap: Fine-Grained Pluralistic Value Alignment via High-Consensus Demographic-Value Mapping. **ACL 2026**.
6. **Lei Yang**, Renren Jin, Ling Shi, Jianxiang Peng, Yue Chen, Deyi Xiong. ProBench: Benchmarking Large Language Models in Competitive Programming. Under review, 2026.
7. **Lei Yang**, Shaoyang Xu, Jianxiang Peng, Shaolin Zhu, Deyi Xiong. DCIS: Efficient Length Extrapolation of LLMs via Divide-and-Conquer Scaling Factor Search. **EMNLP 2025**.
8. Yongqi Leng, Renren Jin, Yue Chen, Zhuowen Han, Ling Shi, Jianxiang Peng, **Lei Yang**, Juesi Xiao, Deyi Xiong. Praetor: A Fine-Grained Generative LLM Evaluator with Instance-Level Customizable Evaluation Criteria. **ACL 2025**.
9. Haoran Sun, Renren Jin, Shaoyang Xu, Leiyu Pan, Supryadi, Menglong Cui, Jiangcun Du, Yikun Lei, **Lei Yang**, Ling Shi, Juesi Xiao, Shaolin Zhu, Deyi Xiong. FuxiTranyu: A Multilingual Large Language Model Trained with Balanced Data. **EMNLP 2024 (Industry Track)**.

RESEARCH EXPERIENCE

Baidu — ERNIE Foundation Model Post-training, Large Model Algorithm Department

Research Intern — *Feb 2026 – Present*

Cross-domain interference in multi-domain reinforcement learning (first-author paper under review)

- Asked *where* interference happens when RL post-training on one domain (math, code, QA, or creative writing) degrades others, and showed that severe degradation occurs even when full-model gradients across domains are

nearly orthogonal — challenging the standard "global gradient conflict" explanation.
- Developed a **local perturbation theory** attributing interference to second-order damage that propagates along low-dimensional shared active pathways into directions sensitive for earlier domains, supported by layer/module-level gradient analysis, update-sparsity measurements, and neuron-overlap studies.
- Derived and validated a practical consequence — **selective recovery via short domain refreshes**: after sequential Code → Math → QA → Writing training dropped Math from 66.49 to 57.66, a brief Math refresh restored it to 66.04 while other domains stayed stable (66.39 average), outperforming Omni-thinker, CGPO, and joint training.

Code agents

- Designed data generation and filtering pipelines and agent skills for Code Understanding and Configuration Deployment tasks, covering repository matching, query generation, checklist annotation, agent rollouts, and diversity sampling.

Kuaishou — Foundation LLM and Applications Department

Research Intern *Jul 2025 – Jan 2026*

Exploration collapse in on-policy RL for LLM reasoning (first-author paper under review)

- Identified that on-policy RL progressively reinforces existing behaviors, causing entropy decline and loss of sampling diversity — most severely at later sequence positions — while naively mixing in off-policy trajectories destabilizes training via importance-ratio shift and gradient clipping.
- Proposed **Thinking Seeds**, a position-aware mix-policy method: prefixes generated by historical checkpoints serve as diverse off-policy "thinking seeds," while the current policy generates on-policy continuations that carry the learning signal; token-level importance ratios limit policy mismatch.
- Demonstrated consistent gains over on-policy and off-policy baselines across multiple models on mathematical reasoning benchmarks, with analysis showing higher effective entropy, less gradient loss from clipping, and a broader explorable solution space.

TJUNLP Lab, Tianjin University

Graduate Researcher (Advisor: Prof. Deyi Xiong) *Sep 2024 – Present*

Efficient long-context extension (DCIS, first author, EMNLP 2025)

- Proposed a divide-and-conquer search algorithm for RoPE scaling factors that enlarges the search space while keeping search cost tractable, addressing the high training cost and target-length degradation of prior extrapolation methods.
- Extended context windows to 64K/128K via short- or long-text fine-tuning while preserving general performance, and substantially reduced perplexity even in training-free settings, achieving state-of-the-art results.

High-fidelity code evaluation (ProBench, first author, under review)

- Proposed an online-submission evaluation paradigm for competitive programming that replaces static offline test suites with feedback from real online judges, with fine-grained problem attributes and error analyses for diagnosing advanced code-reasoning models.

SELECTED PROJECT

Tibetan Large Language Models Dec 2024 – Present

Project Lead, multi-institution collaboration

Co-first-author paper accepted as **ACL 2026 Oral**.

- Led a team building the full LLM development stack for Tibetan, a low-resource language: constructed the largest and highest-quality Tibetan corpus to date via multi-source collection (open datasets, web crawling, synthesis, private data, PDF OCR) and multi-stage cleaning (language/quality filtering, deduplication, safety and toxicity filtering).
- Trained a 7B dense model (v1) from Qwen2.5-7B-Base with Tibetan vocabulary expansion and ∼40B-token Chinese–English–Tibetan continual pre-training; **upcycled** it into a 50B-A10B MoE (v2) with 16 experts and top-2 routing, followed by SFT and DPO on curated and distilled instruction data (Megatron-LM, LLaMA-Factory).

- Iterated to v3 (35B-A3B / 122B-A10B, Qwen3.5-based) with two-stage continual pre-training (foundation modeling → long-context and mid-training), high-quality general-data anchoring, parallel-corpus alignment, and chain-of-thought distillation for post-training.
- Extended toward omni-modal Tibetan models: built vision and speech data pipelines and conducted CPT, mid-training, SFT, and DPO on Qwen3.5-35B-A3B and Qwen3-ASR-1.7B.
- Resulting models significantly outperform size-matched baselines on Tibetan adaptations of general benchmarks (HellaSwag, ARC-Challenge, etc.); the MoE model surpasses DeepSeek V3 on Tibetan question answering and translation.

HONORS & AWARDS

- **Silver Medal**, 47th ICPC Asia Regional Contest (Nanjing), 2022
- **Bronze Medal**, CCPC China Collegiate Programming Contest (Mianyang), 2022
- **Provincial First Prize**, NOIP (National Olympiad in Informatics in Provinces), Advanced Group

TECHNICAL SKILLS

- **LLM post-training:** Reinforcement learning (GRPO-style policy optimization, mix-policy training, entropy/stability analysis), SFT, DPO
- **LLM pre-training:** Continual pre-training, mid-training, dense-to-MoE upcycling, vocabulary expansion, long-context adaptation (RoPE scaling)
- **Frameworks & infrastructure:** Megatron-LM, LLaMA-Factory, large-scale distributed training
- **Data & evaluation:** Multilingual corpus curation at scale, code-agent data pipelines, online-judge and benchmark construction
- **Programming:** Strong algorithmic background (ICPC silver medalist); Python, C/C++