Shuaijie She

Post-Training · RL · Reasoning · Self-Supervised

18851160388 · shesj@smail.nju.edu.cn · ricardokevins.github.io (1.1k+ Stars)

Focus on constructing **self-supervised** learning signals and exploring **general post-training strategies** across tasks, striving toward more **general** and **scalable** LLMs.

Education

Nanjing University · Ph.D. in Computer Science & Technology 2022.09 – Expected 2027.06
First/co-first author of 6 papers at ICLR, ACL, EMNLP, NAACL, and AAAI
National Scholarship · Outstanding Graduate Student
Nanjing University · B.S. in Computer Science & Technology 2018.09 – 2022.06

Internship Experience

Seed Post-training Research · Seed ALM post-training (for Doubao/Lark) · Seed-X post-training (for TikTok)

ByteDance · Seed — 2024.04 – Present — LLM / Post-Training / ALM

Seed Post-Training Research — DuPO · Self-Evolving RL

- **Metrics & Results:** Explored self-supervised training without human annotation to enable **self-evolution** across diverse tasks. Achieved average gains of **6.4** points across four math-reasoning benchmarks including AIME24, reaching SOTA. On these benchmarks, Qwen3-4B outperforms ultra-large models such as R1 and Sonnet-4. Without any training, the method also serves as a reranker for test-time scaling (**+9.3** points) and synthetic-data filtering. On generation tasks such as translation, it achieved an average gain of **2.1** COMET across 756 directions.
- **Responsibilities:** Constructed self-supervised training signals from complementary duality; collected and cleaned million-scale RL data, implemented and ran **extensive RL post-training experiments on Reasoning and Generation tasks**, and built the full training and evaluation pipeline. **RL infrastructure upgrades**: parallelized reward computation and merged dual-task sampling to improve training efficiency. Extended to **Long-Context** scenarios (LongDu, ACL 2026); further exploring **Multimodal** and **Agentic** post-training.

Seed ALM Post-Training — Launched

- **Metrics & Results:** In just two months, human-eval metrics improved substantially (bad-case rate down **9–32** pp) and supported languages grew from 8 to 22; the model is now **live in Doubao** and **Feishu (Lark)**.
- **Responsibilities: Owned the full post-training pipeline**, covering large-scale data synthesis, post-training, automatic evaluation, inference engineering, and final quantization/packaging for release: Built a multi-stage **data processing and synthesis pipeline**, producing **hundreds of millions of SFT examples**; **Built an RL post-training workflow**, synthesizing preference data; **Built a multi-dimensional automatic evaluation system**, accelerating model iteration; and developed multiple **inference strategies**. Analyzed model outputs and evaluation feedback to drive **rapid model iteration**.

Seed-X LLM Post-Training — Long-CoT Exploration

- **Metrics & Results:** An in-house model trained by our team from scratch (pre-training, SFT, and RL), covering **756** translation directions and matching ultra-large models such as Gemini-2.5, Sonnet-3.5, and GPT-4o on both automatic metrics and human evaluation. Hugging Face Trending **#6** with **200k+** downloads.
- **Responsibilities:** Owned the post-training pipeline powering the **TikTok launch**; explored long-CoT post-training, including data synthesis and extending CoT data to more languages; generated synthetic pre-training data for the foundation model; validated DuPO-based RL on Seed-X.

Other Training Work: Independently reproduced R1-style RLVR and delivered an internal talk; submitted a PR to OpenRLHF improving training efficiency.

4Paradigm · LLM Group — 2023.04 – 2024.04 — LLM / Post-Training

- **Responsibilities:** SFT, RLHF, and reasoning-capability optimization for Chinese and English LLMs; LLM pre-training data processing.
- **Engineering Practice:** Model post-training, deployment, and prompt tuning; submitted models consistently ranked first on the company's internal general-capability leaderboard.

Recent Research: RL, Reasoning & Self-Supervision

LongDu: Self-Supervised Long-Context Post-Training — ACL 2026

- **Background & Goal:** Long-input, long-output tasks suffer catastrophic information loss/errors, while human annotation is costly; the goal is to obtain stable long-context training signals from monolingual documents alone, **recovering and improving long-context capability via RL**.
- **Metrics & Results:** LMT-60-8B **+19** BLEU and Qwen3-30B-A3B **+6.5** BLEU; gains generalize beyond training lengths and to unseen target languages. Also proposed an annotation-free long-context evaluation strategy.
- **Method & Role:** Developed self-supervised rewards for information preservation in long-context input/output, and conducted RL post-training experiments: designed penalty mechanisms to improve RL stability. Designed an automatic evaluation protocol that measures information loss through changes in downstream task performance.

R-PRM: Reasoning-Driven, Self-Evolving PRM — EMNLP 2025

- **Background & Goal:** Traditional process reward models output step scores directly, with limited interpretability and data efficiency; the goal is a reward model that "thinks before judging," makes full use of existing data to explore and evolve, and scores each rollout step accurately.
- **Metrics & Results:** Outperforms strong baselines by **13.9 / 8.5 F1** on ProcessBench / PRMBench; matches SOTA PRMs with only **15%** of their data; scales flexibly with inference-time budget, improving F1 from 61.3 to **71.7**.
- **Method & Role:** Trained process reward models to reason before judging, then refined their evaluation trajectories through iterative preference optimization using existing correctness labels. Improved verification quality at inference time by aggregating multiple reasoning trajectories.

MAPO: Self-Supervised RL for Capability Transfer — ACL 2024

- **Background & Goal:** LLM reasoning ability is unevenly distributed; simply synthesizing data for tasks where the model underperforms can introduce noise and limit generalization.
- **Metrics & Results:** **No external annotation required**: **+16.2 / 6.1 / 13.3** pp accuracy on three math-reasoning datasets, with even larger gains on the OOD test set.
- **Method & Role:** Developed a preference optimization framework that uses the model's stronger capabilities to supervise its weaker ones through consistency across equivalent problem formulations. Instantiated this approach with cross-lingual reasoning alignment, without additional human-labeled reasoning traces. **RL infrastructure upgrades**: identified the performance-degradation risk of sharing the actor/value backbone and refactored the TRL training framework.

Publications

(Technical Report 2025) Seed-X: Building Strong Multilingual Translation LLM with 7B Parameters
ByteDance Seed Team

(ICLR 2026, First Author) DuPO: Enabling Reliable LLM Self-Verification via Dual Preference Optimization
***Shuaijie She**, Yu Bao, Yu Lu, Lu Xu, Tao Li, Wenhao Zhu, Jianbing Zhang, Shujian Huang, Shanbo Cheng, Lu Lu, Yuxuan Wang*

(ACL 2026, Co-First Author) Improving Long-Context Translation via Self-Supervised Dual Learning
Shanbo Cheng, **Shuaijie She***, Yu Bao, Jianbing Zhang, Jiajun Chen, Shujian Huang*

(EMNLP 2025, Co-First Author) R-PRM: Reasoning-Driven Process Reward Modeling
Shuaijie She, Junxiao Liu*, Yifeng Liu, Jiajun Chen, Xin Huang, Shujian Huang*

(ACL 2024, First Author) MAPO: Advancing Multilingual Reasoning through Multilingual Alignment-as-Preference Optimization
***Shuaijie She**, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, Jiajun Chen*

(NAACL 2024, First Author) Exploring the Factual Consistency in Dialogue Comprehension of LLMs
***Shuaijie She**, Shujian Huang, Xingyun Wang, Yanke Zhou, Jiajun Chen*

(AAAI 2023, First Author) CoP: Factual Inconsistency Detection by Controlling the Preference
***Shuaijie She**, Xiang Geng, Shujian Huang, Jiajun Chen*

(ACL 2026) A Data-Efficient Path to Multilingual LLMs: Language Expansion via Post-training PARAMΔ Integration into Upcycled MoE
*Hao Zhou, Tianhao Li, Zhijun Wang, **Shuaijie She**, et al.*

(ACL 2024) Question Translation Training for Better Multilingual Reasoning
*Wenhao Zhu, Shujian Huang, Fei Yuan, **Shuaijie She**, Jiajun Chen, Alexandra Birch*

(EMNLP 2024) Multilingual Contrastive Decoding via Language-Agnostic Layers Skipping
*Wenhao Zhu, Sizhe Liu, Shujian Huang, **Shuaijie She**, Chris Wendler, Jiajun Chen*

(EMNLP 2023) Improved Pseudo Data for Machine Translation Quality Estimation with Constrained Beam Search
*Xiang Geng, Yu Zhang, Zhejian Lai, **Shuaijie She**, et al.*