SPEAR: Self-Imitation with Progressive Exploration for Agentic Reinforcement Learning

- "Learn the ropes, then trust the wins."

Yulei Qin*, Xiaoyu Tan*, Zhengbao He*, Gang Li, Haojia Lin, Zongyi Li, Zihan Xu, Yuchen Shi, Siqi Cai, Renting Rui, Shaofei Cai, Yuzheng Cai, Xuan Zhang, Sheng Ye, Ke Li, Xing Sun

Tencent Youtu Lab

TL;DR

- SPEAR lets agentic RL **explore early** and **exploit later** by combining self-imitation, off-policy advantage recalibration, and curriculum-scheduled intrinsic reward.
- Plug-and-play with GRPO, GiGPO, and Dr.BoT; gains come with only modest theoretical overhead.

Why Agent RL Struggles

- Long-horizon tool-use tasks demand careful **exploration-exploitation trade-off**
- Entropy-only exploration is **brittle** under multi-turn distribution shift
- Naive replay **overfits early wins** and shrinks exploration

Core Contributions

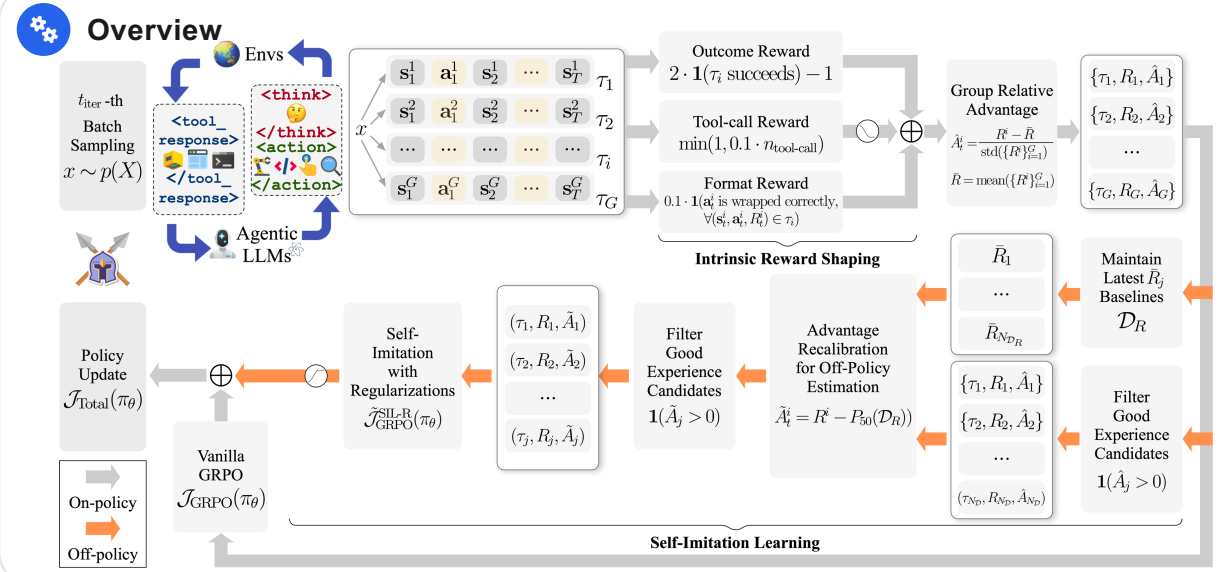
01 Experience-Guided Replay
Generalizes vanilla self-imitation learning to agentic RL and reuses rewarded experience

02 Off-policy Advantage Recalibration
Uses a rolling P50 baseline to filter stale trajectories and stabilize replay without recomputing advantages

03 Progressive Schedules
Warm up the self-imitation term and decay the tool call reward so exploration stays broad early and becomes targeted later.

04 A Strong Baseline: Dr. BoT
Consists of bag-of-tricks modifications to the GRPO to form a strong baseline for agentic RL

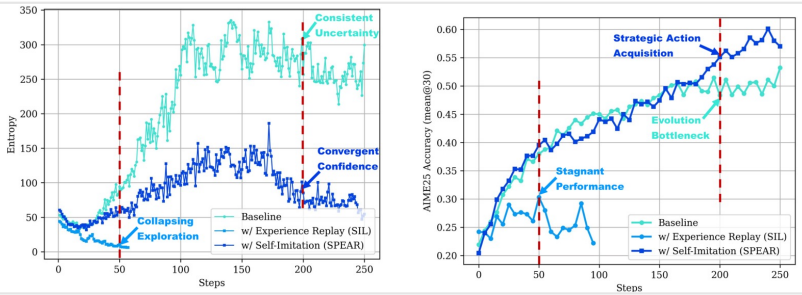
Overview



What Replay Fixes

Vanilla replay: collapses early
SPEAR replay: **entropy controlled**

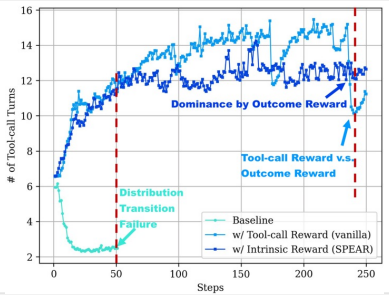
Interpretation: replay should not be a blind copy machine.

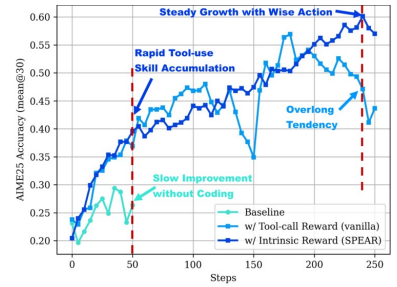


Reward Scheduling

Baseline: overlong interactions
Decay tool reward: **stable**

Interpretation: reward shaping needs training wheels, not permanent rocket fuel.





Dr. BoT (Bag of Tricks)

- Removal of KL Divergence
- Clip-Higher
- Removal of Intra-group Normalization
- Removal of Length Normalization
- Filtering of Over-long and Void-turn Samples.
- Filtering of Low-variance Groups.
- Regularization with Covariance-based Clipping

Key Results

+20.7%
WebShop
vs GRPO on Qwen2.5-1.5B

+16.1%
ALFWorld
vs GRPO on Qwen2.5-1.5B

+6.1%
AIME25
vs Dr.BoT on Qwen2.5-32B

Limitations

- Defining "good experience" is hard in noisy environments
- Entropy-control schedule is hand-designed