SPELL: Self-Play Reinforcement Learning for Evolving Long-Context Language Models

Ziyi Yang[1,2] Weizhou Shen[2] Chenliang Li[2] Ruijun Chen[1] Fanqi Wan[1] Ming Yan[2] † Xiaojun Quan[1,3] † Fei Huang[2]

[1]Sun Yat-sen University [2]Tongyi Lab, Alibaba Group [3]Shenzhen Loop Area Institute

Motivation

- **Long-Context Reasoning Bottlenecks**: The **scarcity** of reliable human annotations and programmatically verifiable reward signals.
- **Non-Verifiable RL:** String match is **brittle and unreliable** for semantic equivalence.
- **SPELL:** A scalable, **label-free** Self-Play framework where a single LLM generates, solves, and verifies its own long-context tasks.

Long-Context RL

- Maximize the KL-regularized expected reward:

$$\max_{\pi_\theta} \mathbb{E}_{c,q\sim\mathcal{D},y\sim\pi_\theta(\cdot|c,q)} \left[r_\phi(c,q,y)\right] - \beta\mathbb{D}_{\mathrm{KL}}\left[\pi_\theta(y \mid c,q) \,||\, \pi_{\mathrm{ref}}(y \mid c,q)\right]$$

Long Documents

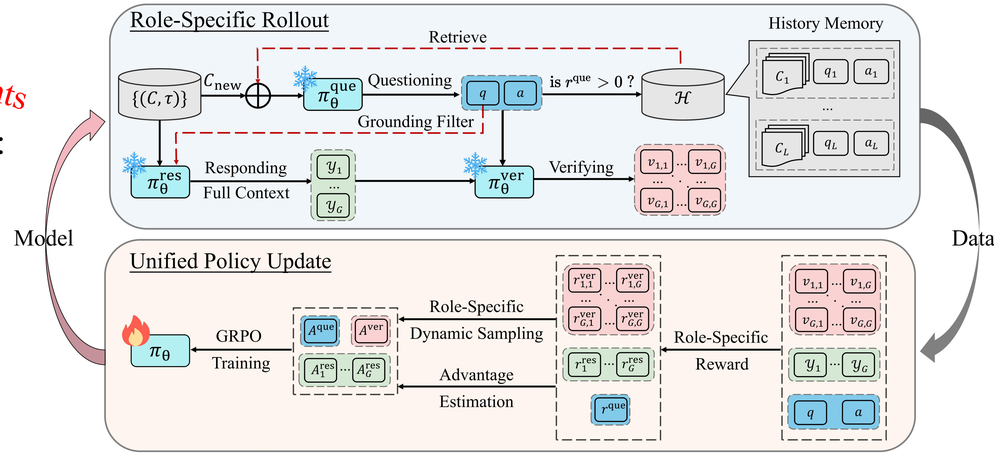
- Fully on-policy GRPO (no KL regularization):

$$\mathcal{J}_{\mathrm{GRPO}}(\theta) = \mathbb{E}_{c,q\sim\mathcal{D},\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\mathrm{old}}}} \left[\frac{1}{\sum_{j=1}^{G}|y_j|}\sum_{i=1}^{G} A_i \sum_{t=1}^{|y_i|} \frac{\pi_\theta(y_{i,t}|c,q,y_{i,t})}{\pi_{\theta_{\mathrm{old}}}(y_{i,t}|c,q,y_{i,<t})}\right]$$

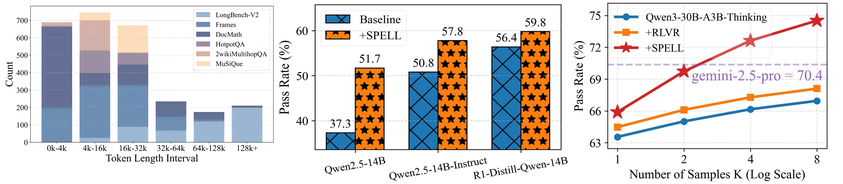
$$A_i = \frac{r_i - \mathrm{mean}(\{r_k\}_{k=1}^{G})}{\mathrm{std}(\{r_k\}_{k=1}^{G})}$$

Method

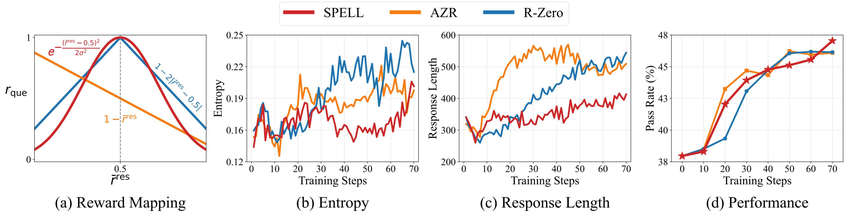
- **Role-Specific Rollout**: A single policy model (π_θ) dynamically assumes three complementary roles:
  - Questioner: Generates questions based on documents and history memory.
  - Responder: Attempts to solve the generated questions.
  - Verifier: Evaluates semantic equivalence between the responder's output and the reference answer to produce reward signals.
- **Role-Specific Reward:**
  - Verifier: Trained to improve judgment reliability through **self-consistency**.
  - Responder: Combines a rule-based check and the verifier's consensus.
  - Questioner: Incentivized to generate questions of **intermediate difficulty**.
- **Unified Policy Update**: Jointly optimizes the fully on-policy GRPO objective across all three roles.



Experiments



- **Broad Generalizability**: Consistent improvements across 12 open-source LLMs (4B to 32B parameters), including dense and MoE architectures, Llama and Qwen families.
- **Superior Exploratory Ability**: SPELL achieves an average 7.6-point gain in pass@8 on Qwen3-30B-A3B-Thinking, significantly outperforming traditional RLVR.



(a) Reward Mapping (b) Entropy (c) Response Length (d) Performance

- **Stable Training Dynamics**: SPELL outperforms prior Self-Play baselines and maintains a more stable training entropy and more controlled growth in response length.

Contact us

- **Shenzhen Area Loop Institute (深圳河套学院)** is **recruiting: Professor/AP/RAP/RA/PhD**
- Contact Email: **xiaojunquan@slai.edu.cn**