Writing-RL: Advancing Long-form Writing via Adaptive Curriculum Reinforcement Learning

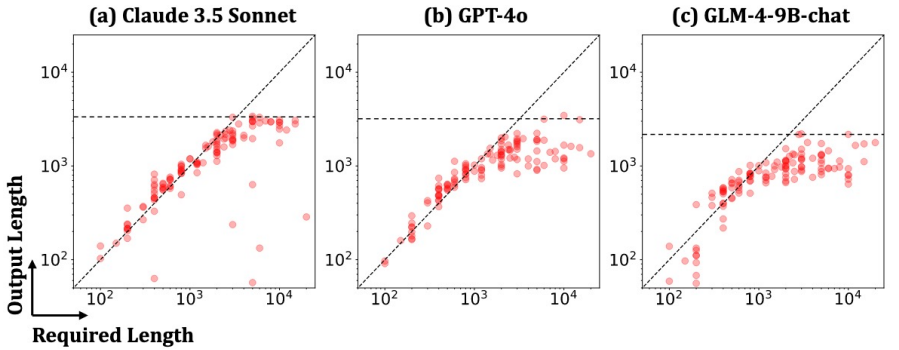
Xuanyu Lei, Chenliang Li, Yuning Wu, Kaiming Liu, Weizhou Shen,
Peng Li, Ming Yan, Fei Huang, Ya-Qin Zhang, Yang Liu

Institute for AI Industry Research (AIR), Tsinghua University, Beijing, China
Dept. of Comp. Sci. & Tech., Institute for AI, Tsinghua University, Beijing, China
Institute of Intelligent Computing, Alibaba Group

Background: Long-form Writing

- Long-form Writing: generate long and high-quality article
  - Limited output ceiling
  - Decreasing quality as length grows

	Overall			[0, 500)		[500, 2k)		[2k, 4k)		[4k, 20k)	
	$\bar{S}$	S_l	S_q	S_l	S_q	S_l	S_q	S_l	S_q	S_l	S_q
GPT-4o	78.6	65.3	91.8	91.0	94.6	91.4	93.6	65.5	93.0	5.6	85.3
+*AgentWrite*	89.1	86.6	91.6	91.0	94.6	91.4	93.6	77.3	90.2	86.8	87.5
+*Parallel*	88.5	87.2	88.9	91.0	94.6	91.4	93.6	79.2	85.6	87.3	80.9



- Existing approaches to train a good writer model
  - Data Curation + SFT
    - Effective and Relatively Cost-friendly
    - Data Saturation and Performance Ceilings
  - RL with Verifiable Rewards
    - Effective in Verifiable Tasks (e.g. math and code)
    - Can not be directly migrated to long-form writing (lack of ground truths)

Qwen2.5-7B-Instruct	69.29
Qwen2.5-7B-WritingBench-SFT (12k)	81.98
Qwen2.5-7B-WritingBench-SFT (24k)	81.94

LongWriter: Unleashing 10,000+ Word Generation from Long Context LLMs; https://arxiv.org/abs/2408.07055

Introduction: Writing-RL

- How to use RL to further advance long-form writing?

  How to extend the success of RLVR to non-verifiable domains?

- Three Challenges:
  - Data Selection: How to select samples to maximize learning efficiency?
  - Reward Design: How to provide reward signals without ground truths?
  - Curriculum Scheduling: How to arrange data to adapt the evolving model?
- Solution:
  - Margin-aware Data Selection
  - Pairwise Comparison Reward
  - Dynamic Reference Scheduling

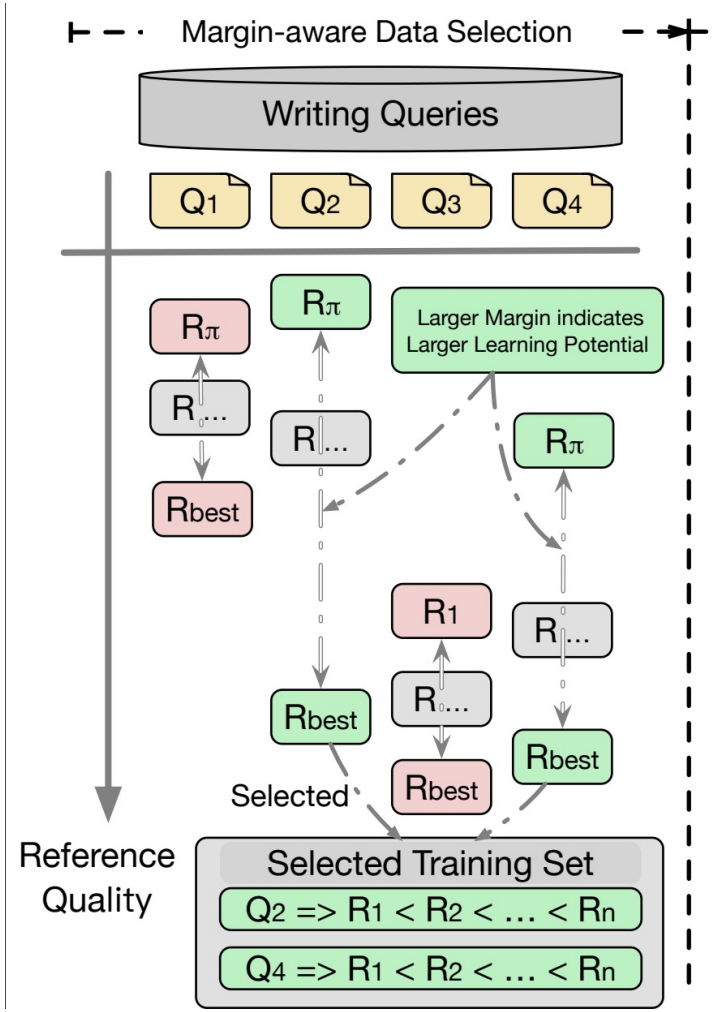
Writing-RL Framework

Methods: Margin-aware Data Selection

- Margin-aware Data Selection
  - Select samples with larger learning potential
  - Step 1: generate references with LLMs
  - Step 2: grading responses with judge model
  - Step 3: calculate *learning potential p*

$$p = \max_{j \in \mathcal{C},\, j \neq \pi} (s_j - s_\pi)$$

  - Step 4: select top-k samples with metric *p*



Methods: Pairwise Comparison Reward

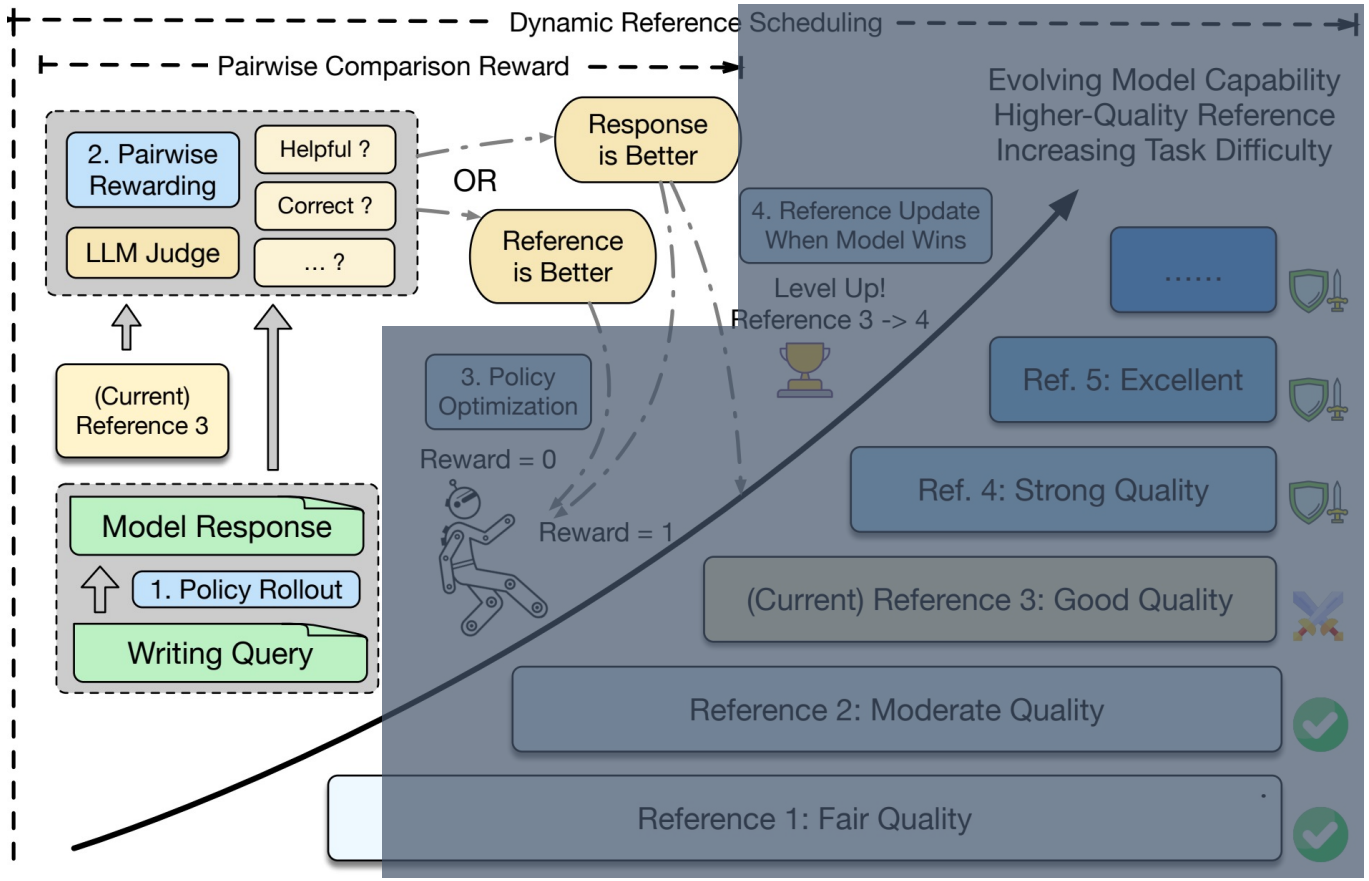
- Pairwise Comparison Reward

$$r_{\text{quality}}(\mathbf{x}) = \begin{cases} 1 & \text{if Judge}(\mathbf{ref}, \mathbf{x}) = \mathbf{x} \succ \mathbf{ref} \\ 0.5 & \text{if Judge}(\mathbf{ref}, \mathbf{x}) = \mathbf{x} \equiv \mathbf{ref} \\ 0 & \text{if Judge}(\mathbf{ref}, \mathbf{x}) = \mathbf{x} \prec \mathbf{ref} \end{cases}$$

  - Agreement with humans

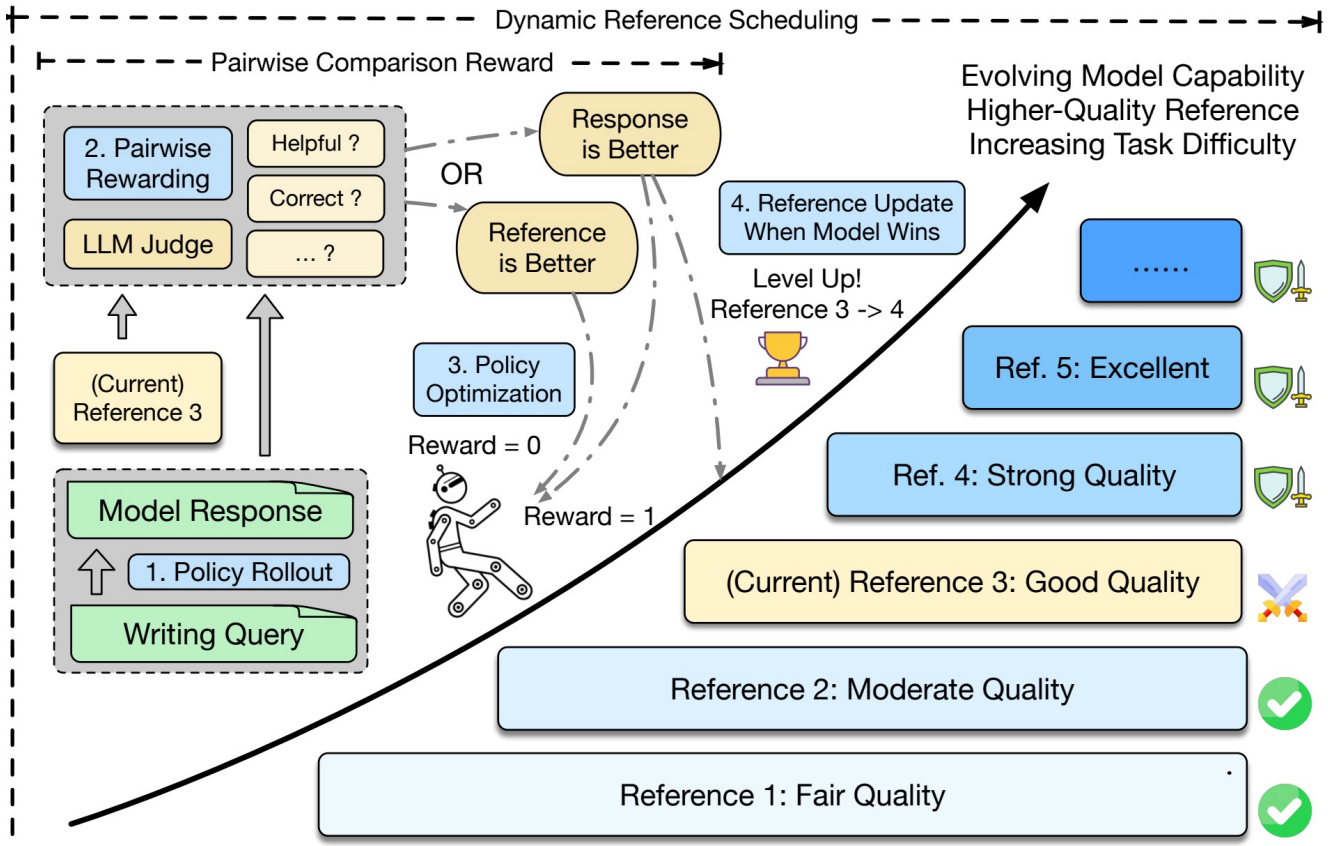
Model	Agreement
claude-3.7-sonnet	0.82
Deepseek R1	0.76
gpt-4o-2024-11-20	0.70
qwen-plus	0.75

  - Position Bias:
    - LLM Judge favors the former
    - *Positional Disadvantage*
    - Judge(reference, response)

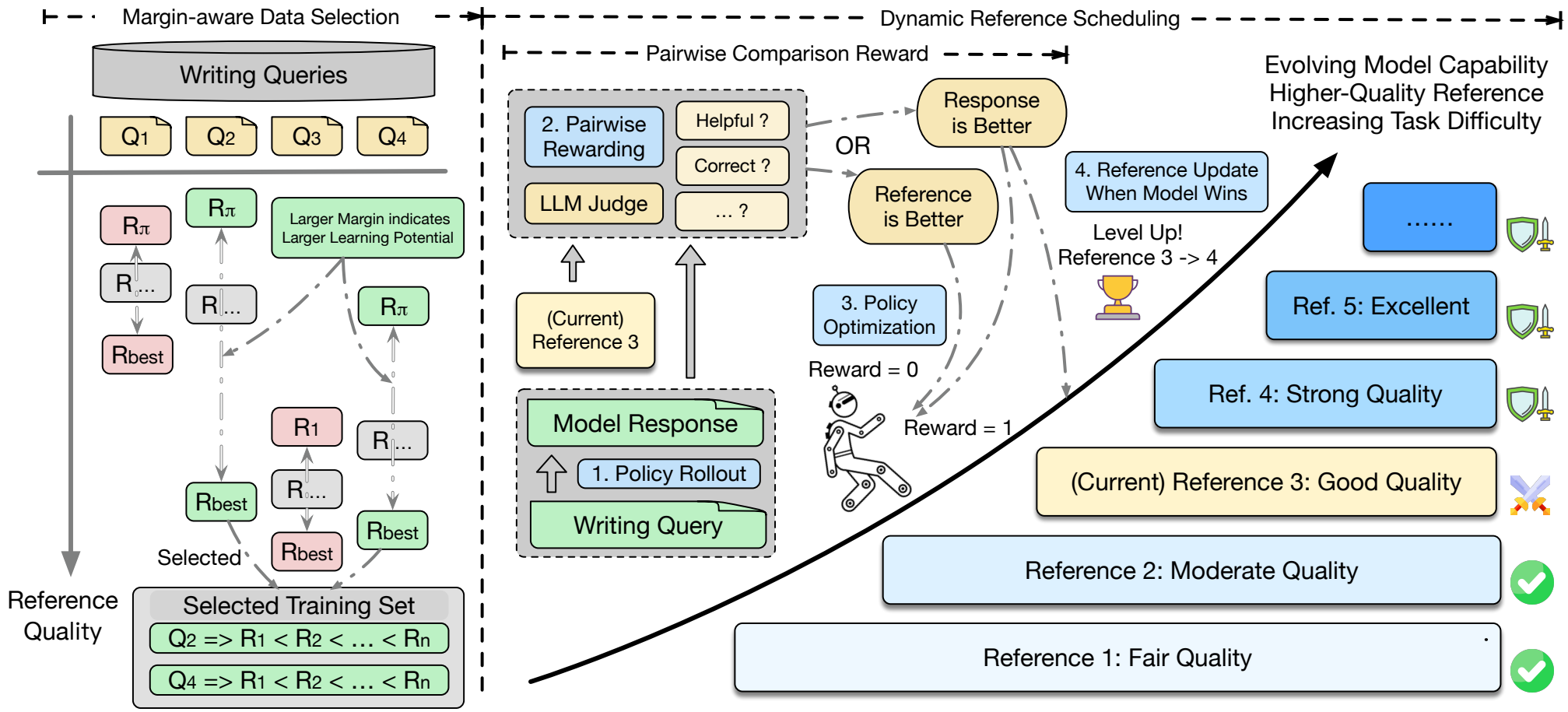


Methods: Dynamic Reference Scheduling

- Dynamic Reference Scheduling
  - Data Preparation
    - writing instruction
    - references with ascending quality
  - During Training
    - Rollout
    - Reward
    - Update Current Reference
      - if policy wins, proceed
      - if policy loses, stays the same
  - Progressively increasing the challenge
  - Adapt the evolving model capabilities



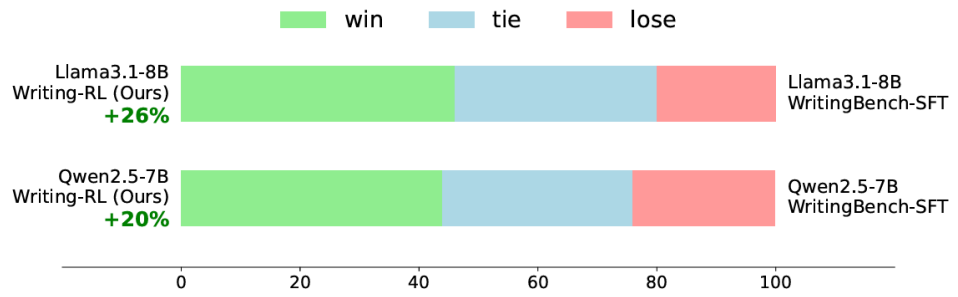
Methods: Writing-RL Framework



Results: Writing Evaluation

- Three Benchmarks:
  - WritingBench
  - Creative Writing Bench
  - LongBench-Write
- Performance Gains
  - Qwen2.5-7B-Writing-RL: +2.34
  - Llama3.1-8B-Writing-RL: +3.22
  - Qwen2.5-14B-Writing-RL: +2.54
  - Best Model *(Qwen2.5-14B-Writing-RL)* Exceeds Proprietary LLMs
- SFT Encounters Data Saturation
  - Qwen2.5-7B-WritingBench-SFT
  - 24k data = 12k data
- Robustness Across RL Algorithms
  - GRPO, PPO
- Human Evaluation
  - Beat Strong Baselines

Model	Writing-Oriented Training		Long-form Writing Evaluation			
	SFT	RL	WritingBench	Creative-W.B.	LongBench-Write	Average
(a) Proprietary LLMs						
Qwen-Plus	–	–	77.62	76.78	95.42	83.27
GPT-4o	–	–	83.42	80.45	92.92	85.60
(b) Writing-Oriented Fine-Tuned LLMs						
Suri-7B	✓	✗	49.70	18.44	33.44	33.86
Longwriter-9B	✓	DPO	79.10	44.15	80.83	68.03
Longwriter-Zero-32B	✓	GRPO	82.92	61.14	85.90	76.65
(c) Qwen2.5-7B-Instruct Model Family						
Qwen2.5-7B-Instruct	✗	✗	73.16	49.29	87.20	69.88
Qwen2.5-7B-WritingBench-SFT (12k)	✓	✗	83.71	70.24	92.56	82.17
Qwen2.5-7B-WritingBench-SFT (24k)	✓	✗	83.76	69.60	92.65	82.00
Qwen2.5-7B-Reference-SFT	✓	✗	84.23	68.89	92.88	82.00
Qwen2.5-7B-Writing-RL (Ours)	✓	PPO	**87.23**	73.19	**93.06**	84.49
Qwen2.5-7B-Writing-RL (Ours)	✓	GRPO	86.29	**75.41**	91.84	**84.51**
(d) Qwen2.5-14B-Instruct Model Family						
Qwen2.5-14B-Instruct	✗	✗	72.67	56.89	89.20	72.92
Qwen2.5-14B-WritingBench-SFT	✓	✗	84.95	79.27	93.75	85.99
Qwen2.5-14B-Writing-RL (Ours)	✓	PPO	**88.27**	**83.17**	**94.17**	**88.53**
(e) Llama3.1-8B-Instruct Model Family						
Llama3.1-8B-Instruct	✗	✗	66.56	48.83	80.79	65.39
Llama3.1-8B-WritingBench-SFT	✓	✗	83.75	77.70	90.67	84.04
Llama3.1-8B-Reference-SFT	✓	✗	83.98	76.70	91.53	84.07
Llama3.1-8B-Writing-RL (Ours)	✓	PPO	**87.11**	**82.67**	**92.09**	**87.29**

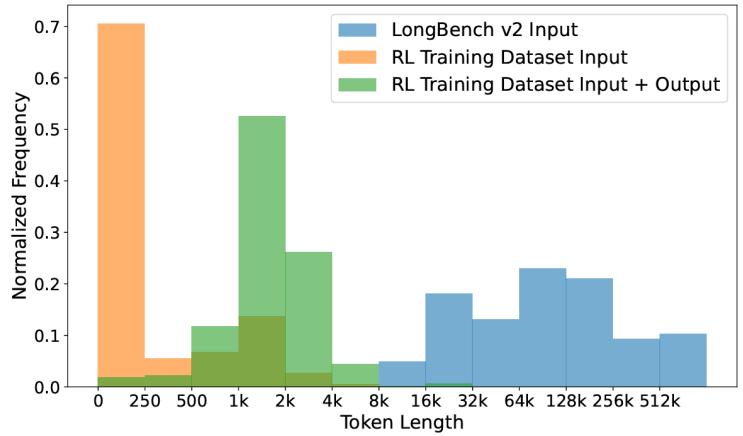


Output-to-Input Generalization

- LongBench v2: Long-input Reasoning
  - Output RL: input < 4k; Long-Input Test: 8k – 2M

Table 3: Evaluation results of the models trained with Writing-RL on LongBench v2, demonstrating the generalization potential from long-output generation to long-input reasoning.

Model	Writing-Oriented Training		Evaluation					
	SFT	RL	Easy	Hard	Short	Medium	Long	Overall
Qwen2.5-7B-Instruct	✗	✗	31.8	28.3	38.9	**26.0**	21.3	29.6
Qwen2.5-7B-WritingBench-SFT	✓	✗	27.6	27.7	35.0	25.1	20.4	27.6
Qwen2.5-7B-Writing-RL (Ours)	✓	PPO	**35.8**	**29.3**	**42.1**	25.7	**26.5**	**31.8**
Llama3.1-8B-Instruct	✗	✗	**32.3**	28.9	35.6	27.4	**26.9**	30.2
Llama3.1-8B-WritingBench-SFT	✓	✗	29.7	27.7	36.7	23.7	24.1	28.4
Llama3.1-8B-Writing-RL (Ours)	✓	PPO	31.2	**33.8**	**42.2**	**29.3**	24.1	**32.8**



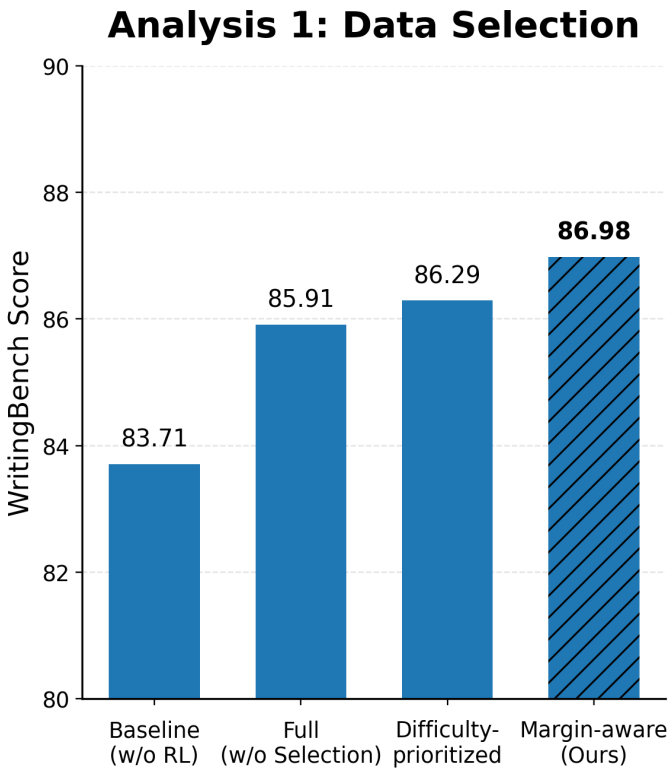
- Generalization from Long-output Generation to Long-input Reasoning
  - Our intuition is that generating high-quality long-form text requires the model to understand and organize context globally.
  - Long-output RL may also improve the ability to handle long-range dependencies.
  - Long-input understanding and long-output generation share the same fundamental capabilities.
  - May suggest a new perspective for long-context training.

Analysis: Why Writing-RL Works

Three components each contribute to the final gain

Analysis 1: **Data Selection**

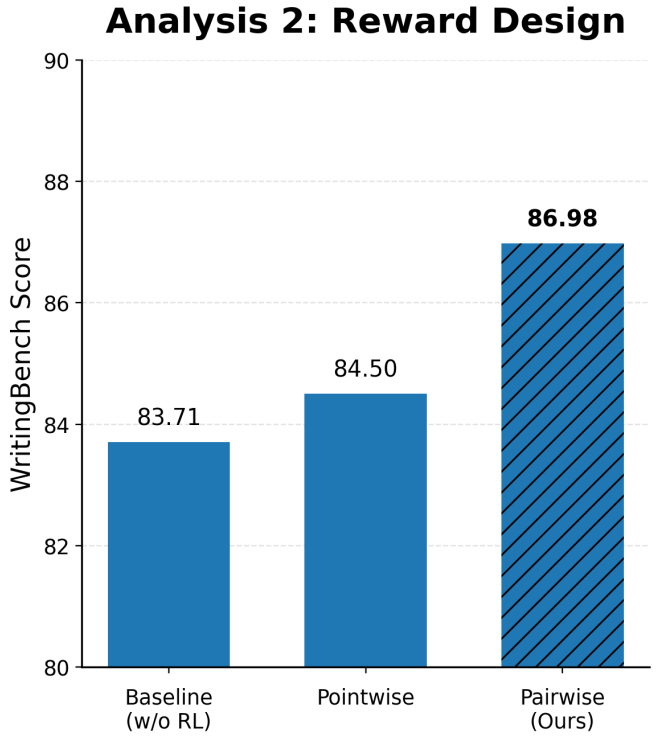
High learning potential matters.



Margin-aware > Difficulty > Full

Analysis 2: **Reward Design**

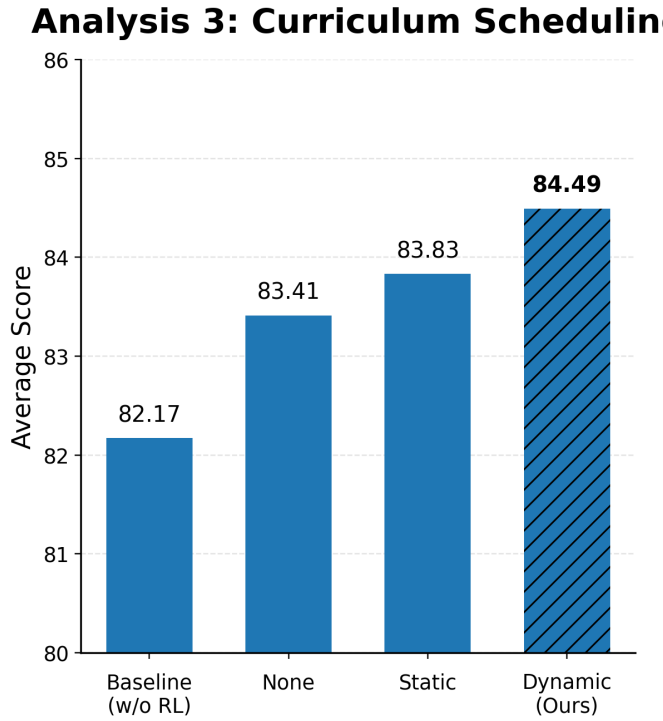
Pairwise reward is more discriminative.



Pairwise Achieves the Best Score.

Analysis 3: **Curriculum Scheduling**

Dynamic difficulty works best.



Dynamic Scheduling Yields the Best Performance.

Summary

1. We propose Writing-RL: an **Adaptive Curriculum Reinforcement Learning framework** for long-form writing, which integrates three key components: Margin-aware Data Selection, Pairwise Comparison Reward, and Dynamic Reference Scheduling.

2. Particularly, we propose **Dynamic Reference Scheduling**, which adaptively adjusts sample-level task difficulty based on the model's evolving performance. This curriculum encourages the model to continually outperform progressively stronger references.

3. Furthermore, we observe inspiring **Output-to-Input Generalization** from long-output generation to long-input reasoning, potentially offering a novel perspective to rethink long-context training.

Thanks

Paper

Code

WeChat

X