Long Document Reconstruction Unlocks Scalable Long-Context RLVR

Anonymous ACL submission

Abstract

Reinforcement Learning with Verifiable Rewards (RLVR) has become a prominent paradigm to enhance the capabilities (i.e. long-context) of Large Language Models (LLMs). However, it often relies on gold-standard answers or explicit evaluation rubrics provided by powerful teacher models or human experts, which are costly and time-consuming. In this work, we investigate unsupervised approaches to enhance the long-context capabilities of LLMs, *eliminating the need for heavy human annotations or teacher models' supervision*. Specifically, we first replace a few paragraphs with special placeholders in a long document. LLMs are then trained through reinforcement learning to reconstruct the long document by correctly identifying and sequencing missing paragraphs from a set of candidate options. This training paradigm enables the model to capture global narrative coherence, significantly boosting long-context performance. We validate the effectiveness of our method on two widely used benchmarks, RULER and LongBench v2. While acquiring noticeable gains on RULER (nearly 10 points), it can also achieve a reasonable improvement on LongBench v2 without any manually curated long-context QA data. Furthermore, we conduct extensive ablation studies to analyze the impact of reward designs, data curation strategies, training schemes, and data scaling effects on model performance. We will release our code, data, and models.

1 Introduction

Reinforcement Learning with Verifiable Rewards (RLVR) has recently achieved the state-of-the-art in Large Language Model (LLM) reasoning (Kumar et al., 2025; Yu et al., 2025; Yeo et al., 2025; Zeng et al., 2025a; Liu et al., 2025; Anthropic, PBC, 2025). Using ground-truth feedback to guide generation, RLVR enables models like DeepSeek-R1 (Guo et al., 2025) to navigate complex multi-step problem-solving paths in domains such as mathematics and programming with unprecedented precision (Yang et al., 2025a). However, as LLMs evolve into agents that must interact with expansive real-world datasets, the challenge shifts from local reasoning to global context processing (Team et al., 2025b,c; Prabhakar et al., 2025). Here, a significant gap remains: models that excel at step-by-step logic reasoning often struggle to maintain coherence when retrieving and synthesizing information across tens of thousands of tokens (Wu et al., 2025b; Zhuang et al., 2025; Wu et al., 2025c; Bai et al., 2025b). This suggests that reasoning gains do not necessarily scale with context length, leaving a gap between reasoning capacity and long-context understanding.

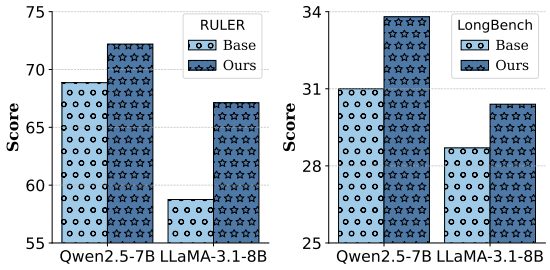


Figure 1: Average score on RULER and overall score on LongBench v2 for Qwen2.5-7B-Instruct-1M and LLaMA-3.1-8B-Instruct.

The pursuit of extending the context window has thus become a central objective in developing frontier LLMs (Zhu et al., 2024; An et al., 2024; Peng et al., 2024; Gao et al., 2025; Lu et al., 2025; Yang et al., 2025b; Wan et al., 2025; Xu et al., 2026). From analyzing massive code bases to distilling insights from an entire document, the ability to process and reason over long sequences is essential for practical applications. However, as the context length increases, the difficulty of maintaining global coherence and performing precise retrieval increases exponentially. Models frequently suffer from the well-documented "lost in the middle"

phenomenon (Liu et al., 2024), where information in the center of a long prompt is ignored, or they struggle to maintain the logical consistency of a narrative over tokens (Hsieh et al., 2024b,a; Du et al., 2025). Although RLVR offers a powerful framework for LLMs to refine their long-range dependency handling, its application is currently restricted by a heavy reliance on external supervision.

Recent RLVR approaches (Wang et al., 2025; Chen et al., 2026) to enhance the long-context capability of LLMs are constrained by the availability of gold-standard answers or evaluation rubrics, which are typically provided by expensive human experts and frontier teacher models (often close-source) (Chen et al., 2025a; Zhang et al., 2025; Huang et al., 2025). This dependency creates a significant scalability bottleneck: while the need for long-context understanding is universal, question-answering pairs from human annotations required for training are prohibitively expensive and difficult to generate at the scale needed for reinforcement learning. Furthermore, relying on teacher models can introduce biases or limit the potential of student models to the capabilities of the supervisor (Kim et al., 2025; Cheng and Amiri, 2025). To unlock the next level of long-context ability of LLMs, we explore a training mechanism that can derive objective, verifiable rewards directly from the data itself in an unsupervised manner.

In this work, we introduce a fully unsupervised RLVR framework that bypasses the need for costly human annotations or rubrics from teacher models, enabling a more scalable approach to long-context training. We hypothesize that long documents possess an inherent structure, specifically their narrative flow and logical sequence, which contains valuable internal signals. The signals can act as a natural and verifiable reward for training model through RLVR. We formalize this through a document reconstruction task. Specifically, we first mask a few random paragraphs within a long document and then require the LLM to correctly identify and sequence these missing segments from a shuffled pool of candidates. Since the original documents provide the ground truth, the resulting reward enjoys the desirable property of being both verifiable and fully automated. This reconstruction training mechanism encourages LLMs to move beyond surface-level pattern matching and instead develop a deeper, more structural understanding of the global context (Yang et al., 2019).

We evaluate our method on two of the most rigorous benchmarks in the field: RULER (Hsieh et al., 2024a) and LongBench v2 (Bai et al., 2025a). Our empirical results demonstrate that this unsupervised paradigm yields substantial gains on RULER and moderate improvement performance on LongBench v2, as shown in Figure 1. These findings suggest that the underlying long-range structural integrity of documents provides a valuable training paradigm, offering a scalable path toward more capable long-context LLMs. Our contributions are summarized as follows:

- We propose an unsupervised RLVR framework based on document reconstruction and formulate it as a sequential selection problem with verifiable rewards, encouraging models to learn global narrative coherence and long-range structural dependencies.
- We validate our approach through extensive experiments on RULER and LongBench v2, showing that it provides a scalable and effective alternative to supervised long-context training.
- We perform comprehensive ablation studies to examine the effects of reward formulation, data curation strategies, and training configurations, offering deeper insights into the characteristics of our method.

2 Background

Group Relative Policy Optimization (GRPO). Group Relative Policy Optimization (GRPO) (Shao et al., 2024; Guo et al., 2025) is a policy-gradient method that removes the need for an explicit value function by estimating advantages through relative comparisons within a group of sampled responses. Given a query q, GRPO samples a group of trajectories $\{\tau^{(i)}\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot \mid q)$ and assigns each trajectory a scalar reward $R_{\tau^{(i)}}$. The advantage for each trajectory is computed by normalizing its reward against the group:

$$A(\tau^{(i)}) = \frac{R_{\tau^{(i)}} - \text{mean}\Big(\{R_{\tau^{(j)}}\}_{j=1}^{G}\Big)}{\text{std}\Big(\{R_{\tau^{(j)}}\}_{j=1}^{G}\Big)}$$

GRPO then optimizes a PPO-style clipped surrogate objective using these group-relative advantages, enabling stable policy updates without learning a separate value function.

Using the normalized group-relative advantages, GRPO performs policy updates via a PPO-style clipped surrogate objective:

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}_{\tau}\Big[\min\big(r_\theta A, \text{clip}(r_\theta, 1-\epsilon, 1+\epsilon)A\big)\Big] \tag{1}$$

where

$$r_\theta(\tau) = \frac{\pi_\theta(\tau \mid q)}{\pi_{\theta_{\text{old}}}(\tau \mid q)}.$$

This clipped objective constrains policy updates while leveraging group-normalized rewards as a low-variance advantage estimator, eliminating the need for a learned critic.

3 Method

3.1 Task Formulation: Document Reconstruction

Our goal is to derive a verifiable reward from raw documents without labels from human annotators. Given a long document consisting of n paragraphs, denoted as $D = \{p_1, p_2, \ldots, p_n\}$, we first select a subset of paragraphs to mask. The original document is then transformed into a corrupted context, where the masked paragraphs are replaced by placeholders marked with identifiers, denoted as `<CHUNK_i>MISSING</CHUNK_i>`. i means the i-th masked paragraph of the document.

The model is presented with the context and a set of shuffled candidates labeled with options. LLMs are asked to reconstruct the original text by first reasoning and then generating a list of these options in the correct order. For example, if four paragraphs were selected, the model's output would be a formatted list such as $\{B, A, D, C\}$. This formulation transforms the long context understanding problem into a sequential decision making task, where the model must utilize global narrative flow and logical consistency to determine the correct placement of each segment. The overview of our method can be found in Figure 2.

3.2 Reward Design

Unlike open-ended generation, our reconstruction task provides an objective and verifiable answer. Since the original document provides ground truth ordering, we define a verifiable reward function that evaluates the model output against ground truth as follows:

$$R(o, g) = \begin{cases} 1, & \text{if } o = g \\ \frac{1}{K}\sum_{i=1}^{K} \mathbb{I}[o_i = g_i], & \text{if } \mathcal{V}(o) \\ & \quad \wedge\, o \neq g \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

In this formulation, K represents the total number of masked segments, while o_i and g_i denote the predicted and ground-truth options for the i-th placeholder, respectively. The function $\mathbb{I}[\cdot]$ is an indicator function that yields 1 for a correct match and 0 otherwise. A critical component of this reward structure is the $\mathcal{V}(o)$ condition. We define a predicted output as a *valid permutation* if and only if the set of options provided in the model's response is identical to the set of ground-truth options. This constraint requires the model to correctly identify and utilize the complete pool of candidates, without omitting or duplicating any options (Wu et al., 2025a; Lu et al., 2026).

This piecewise reward structure balances global accuracy with fine-grained feedback. A full reward of 1 is assigned when the reconstruction is exact ($o = g$). For outputs that are not perfectly ordered but remain structurally valid, defined as sequences whose predicted option set exactly matches the ground truth, we assign a partial reward proportional to the fraction of correctly placed segments. This design encourages progressive refinement of the global structure, even without full sequence accuracy. In contrast, any output that violates the required format or constitutes an invalid permutation receives a reward of 0. The entire training signal is fully automated and does not depend on human annotations or external teacher models.

3.3 Curriculum through Complexity Scaling

An appealing advantage of our framework is the ability to precisely calibrate the difficulty of training samples by adjusting the complexity of the reconstruction task. Our intuition is that larger K leads to more challenging samples. As K increases, the search space expands exponentially because the number of possible permutations for the candidate set is $K!$. By treating K as a tunable hyperparameter, we can implement a curriculum (Bengio et al., 2009) that allows the model to first master local coherence with fewer options before tackling the complex global structural dependencies required

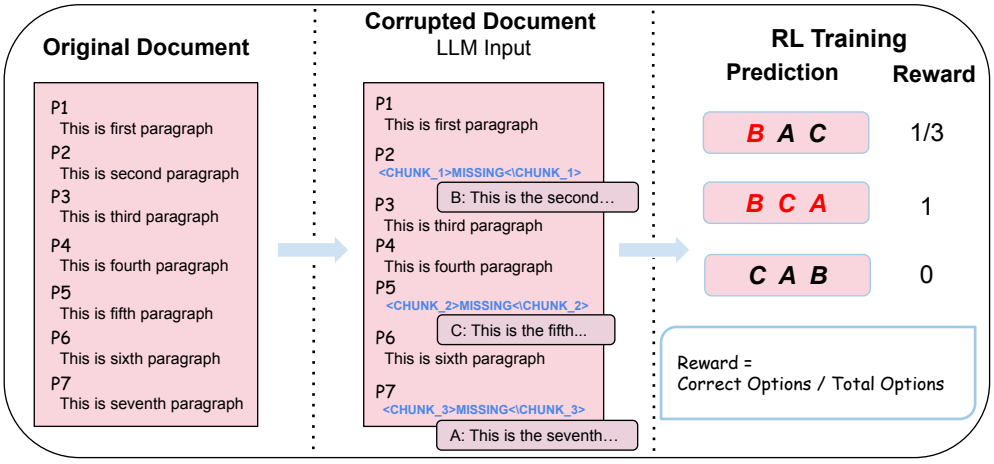


Figure 2: Overview of the document reconstruction framework. Given a long document, we corrupt it by selecting some paragraphs and shuffle them as options. We train LLMs via RLVR to reconstruct the document by generating the option sequence by order.

for extremely long contexts. This controllable difficulty (Zeng et al., 2025b; Wang et al., 2026a) ensures that the model can progressively build its long-context understanding capabilities.

4 Experimental Setup

Data Curation. We source our training documents from the corpus provided by Chen et al. (2025a), which covers three diverse domains: books, arXiv papers, and code. We curate a subset consisting of the 8,000 longest documents from the book domain, along with 3,000 longest documents each from arXiv and code (total 14,000). We apply varying levels of difficulty to these documents by setting $K \in \{2, 4, 6, 8\}$ with their corresponding number ratio being $3 : 3 : 3 : 5$, which we find useful in our preliminary experiment. The average token number of samples in the training set is 49,000. This multi-scale approach to the number of masked segments ensures that the training set provides a broad spectrum of structural challenges. In addition, we also curate 500 samples for validation set to better observe training dynamics.

Training. We adopt the curriculum schedule (Bengio et al., 2009) to progressively train models by increasing K from 2 to 8. Our implementation is built on the Verl framework (Sheng et al., 2024). To optimize the training process, we employ Group Relative Policy Optimization (GRPO) (Shao et al., 2024). For RL training, we use the AdamW optimizer with a constant learning rate of 1e-6 and a 5-step linear warmup. For rollout, we use a prompt batch size of 128 and sample 8 responses per prompt, with a maximum context length of 64K and a response length of 4096. Our reconstruction prompt can be found in Appendix A.1.

Evaluation. Benchmarks. We evaluate all models on two challenging long-context QA benchmarks: (1) RULER (Hsieh et al., 2024a): A synthetic benchmark testing multi-hop reasoning over arbitrary context length. Specifically, we evaluate on four tasks of it (Variable Tracking, Frequent Words Extraction, Common Words Extraction, Question Answering). (2) LongBench v2 (Bai et al., 2025a): A realistic multi-choice QA benchmark on documents up to 128K tokens. We evaluate the lengths of 32K, 64K, and 128K. In our analysis, we primarily focus on RULER, as it provides results of a more comprehensive and diverse task.

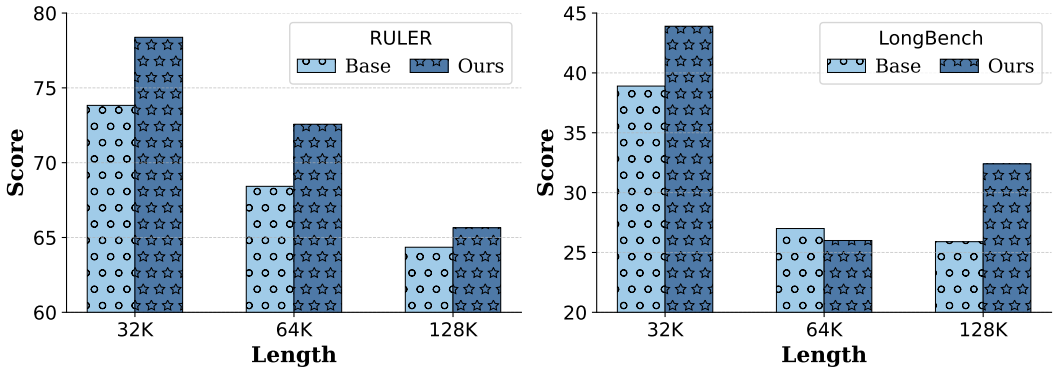
Models and Baselines. We select LLaMA-3.1-8B-Instruct and Qwen2.5-7B-Instruct-1M as our backbone models, which also serve as our baseline. A discussion about continual pretraining on long documents is also presented in Appendix A.2.

5 Results and Analysis

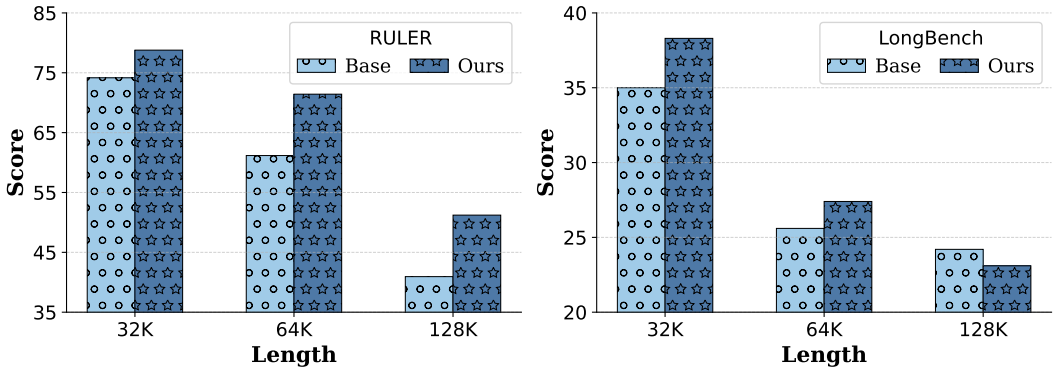
5.1 Main Results

We summarize the main results across different context lengths and backbone models in Figure 3. In addition, we also report the average score of RULER and the overall score of LongBench v2 in Figure 1.

On RULER, our method produces substantial improvements, with consistent gains as the context length increases from 32K to 128K. This indicates that our unsupervised reconstruction training effectively enhances the model's ability to maintain global coherence and retrieve relevant information



(a) Qwen2.5-7B-Instruct-1M



(b) LLaMA-3.1-8B-Instruct

Figure 3: Performance comparison across different context lengths and models.

Model	# QA	RULER-QA	LongBench
Qwen2.5-7B-Instruct-1M			
SFT	46K	64.5	33.2
Ours	0K	**68.3**	**33.8**
LLaMA-3.1-8B-Instruct			
SFT	46K	**64.7**	**31.2**
Ours	0K	61.2	30.4

Table 1: Comparison between SFT and our reconstruction training on RULER-QA (average) and LongBench v2 (overall). Note that data for SFT are QA pairs from teacher model supervision.

in long contexts. On LongBench v2, we also observe moderate improvements across most context lengths. And these gains are achieved without using any manually curated long-context QA data, highlighting the effectiveness of reconstruction training even for question answering. The improvements remain consistent across different backbone architectures, including Qwen2.5-7B-Instruct-1M and LLaMA-3.1-8B-Instruct, demonstrating that our method is not tied to a specific model family.

Taken together, these results show that unsupervised document reconstruction via RLVR provides a scalable and effective mechanism for improving long-context capabilities. The method delivers strong gains on synthetic reasoning benchmarks and meaningful improvements on realistic QA tasks, all while eliminating the need for human annotations or teacher-model supervision. Detailed scores on RULER can be found in Appendix A.4. We also record the performance on our curated validation set during the training process in Appendix A.3.

Comparison to SFT. (Chen et al., 2026) employ a powerful teacher model to curate 46K context-specific QA pairs for supervised fine-tuning. They report results on LongBench v2 and only the QA subset of RULER. In Table 1, we compare our results with theirs. With only 14K training samples and without relying on any context-specific QA pairs, our approach achieves better performance on Qwen2.5-7B-Instruct-1M, while slightly lags behind on LLaMA-3.1-8B-Instruct.

5.2 Dense vs. Sparse Reward

In our main experiments, we employ a dense reward that provides partial credit for partially correct reconstructions. To better understand the role of reward shaping in document reconstruction, we compare this design with a sparse reward formulation, defined as

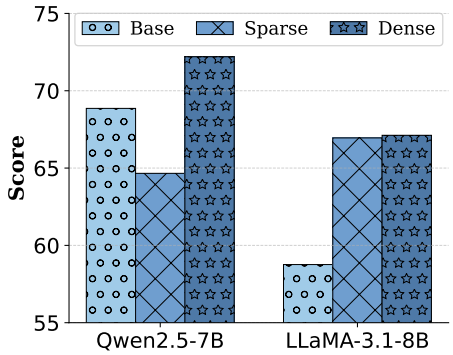


Figure 4: Average scores on RULER. We compare the performance of dense and sparse rewards.

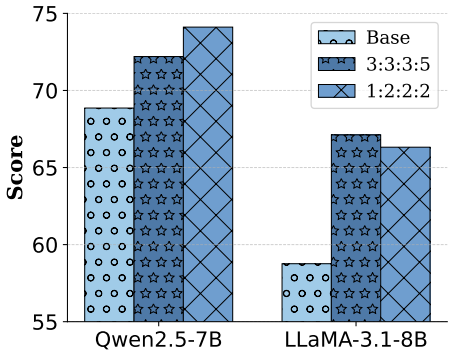


Figure 5: Average scores of RULER. We compare the performance of different option length mixture ratios.

$$R(o, g) = \begin{cases} 1, & \text{if } o = g, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

This sparse reward assigns a non-zero signal only when the predicted ordering exactly matches the ground truth, providing a stricter but less informative supervision signal. As shown in Figure 4, sparse rewards obtain performance similar to dense rewards on LLaMA-3.1-8B-Instruct. However, it causes significant performance degradation on Qwen2.5-7B-Instruct-1M. We hypothesize that sparse rewards are more likely to cause training instability due to the sparsity of positive rewards in the training process.

5.3 Robustness to Option Mixture Ratios

In our main experiments, we adopt the option mixture ratio for $K = 2, 4, 6$, and 8 as $3:3:3:5$, which assigns only a moderate portion of training samples to small values of K (e.g., K=2). To further validate robustness, we conduct an ablation study by shifting more training samples toward larger K, using the ratio $1:2:2:2$. In Figure 5, empirical results demonstrate that our method maintains high performance across varying option length distributions. For the Qwen2.5-7B-Instruct-1M model, the $1:2:2:2$ ratio yielded the highest score, surpassing baseline. In the case of LLaMA-3.1-8B-Instruct model, the $3:3:3:5$ ratio proves most effective. These findings suggest that while specific ratios can offer marginal gains on different model architecture, the overall framework remains robust. The consistency in performance across different difficulty blends confirms that the method does not rely on a brittle or overly specific data composition to succeed.

5.4 Longer Documents Bring Consistent Improvement

For data curation of main experiments, we retain the longest 8,000, 3,000, and 3,000 documents from the book, arXiv, and code domains, respectively. In this experiment, we compare this document selection strategy against two alternatives: shortest and random. The shortest counterpart selects the shortest documents while preserving the same domain ratios and total number of documents. The random counterpart samples documents randomly, keeping other factors identical.

As shown in Figure 6, although short-document training brings modest gains for LLaMA-3.1-8B-Instruct, it does not exceed the baseline performance of Qwen2.5-7B-Instruct-1M. Moreover, random document sampling leads to consistent but limited improvements on Qwen2.5-7B-Instruct-1M. In contrast, training on long documents leads to overall improvements, suggesting the importance of longer contexts for the reconstruction task. In summary, these observations indicate that document length plays a critical role in enabling effective long-context understanding through reconstruction training.

5.5 More Documents, Better Performance

Scaling training data is an important factor in understanding the effectiveness of reconstruction learning. In this part, we study the effect of training data scale on model performance by increasing the number of reconstruction training samples from 0 to 14,000. This setting allows us to examine how model performance evolves when we increase the number of reconstruction samples. For each data scale, we train models under identical optimization settings and evaluate them on RULER. When the

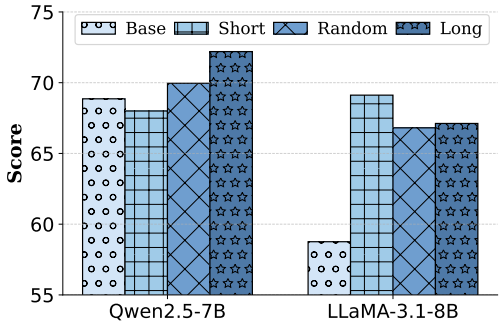


Figure 6: We report the average score of RULER. Longest documents lead to consistent improvement.

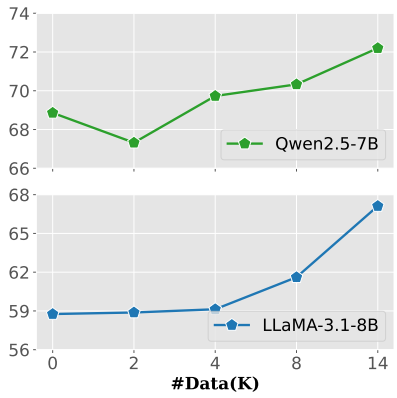


Figure 7: Average scores of RULER when we scale data from 0 to 14K.

data size is set to 0, the model corresponds to the original backbone without reconstruction training, serving as a baseline. We record the performance in Figure 7, model performance generally improves as the number of reconstruction training samples increases.

Although performance may fluctuate slightly at smaller data scales, especially when the training set is limited, the overall trend is clearly positive as more data is introduced. In particular, performance consistently increases when the data scale exceeds 4,000 samples, indicating that sufficient reconstruction data is crucial for effectively enhancing long-context understanding. These results suggest that reconstruction-based training scales well with data size and that increasing training data is an effective and stable way to improve model performance. Notably, we do not observe a clear performance plateau within the evaluated data range, suggesting that the model may continue to benefit from additional reconstruction training data. **Significant performance gains** (nearly 10 points) can be achieved on Qwen2.5-7B-Instruct-1M by scaling data to 3,0000, which is shown in Table 8. Data recipe about this can be found in Appendix A.5.

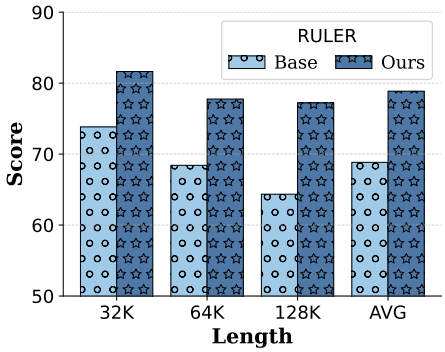


Figure 8: We report the performance gains (nearly 10 points) of RULER when scaling the data to 30,000 samples on Qwen2.5-7B-Instruct-1M.

5.6 Training Strategy: Shuffle or Not

As mentioned in Section 3, we curate training data by progressively increasing the option length K from 2 to 8, following a curriculum-style training strategy. Our intuition is that reconstruction tasks with larger K are substantially more challenging, and gradually exposing the model to harder samples can facilitate more stable optimization.

To validate this design choice, we compare the curriculum strategy with a shuffled training variant. As shown in Table 2, the curriculum-based strategy (shuffle is false) consistently outperforms shuffled training for both Qwen2.5-7B-Instruct-1M and LLaMA-3.1-8B-Instruct backbones. These results demonstrate that a curriculum of option length is an important component of our training framework.

5.7 Analysis of Option Length

In this part, we study the influence of sample difficulty on model performance. Intuitively, a sample with more candidate options is more challenging, as the model must search over a larger permutation space and rely more heavily on global contextual cues. We randomly sample 5,000 documents from the training corpus and create four distinct training sets, where each set is constructed using a fixed option length $K \in \{2, 4, 6, 8\}$, without mixing different values of K within the same set. We train an independent model for each curated training set, where each model is associated with a specific K.

As shown in Table 3, we observe that model performance is relatively stable across different option lengths K. While increasing K substantially enlarges the permutation space and task difficulty, it does not lead to a significant degradation in downstream RULER performance for either backbone.

Model	Qwen2.5-7B	LLaMA-3.1-8B
Base	68.86	58.76
Shuffle		
True	69.82	65.00
False	**72.20**	**67.12**

Table 2: Effect of training strategy during reconstruction training on RULER.

Model	Qwen2.5-7B	LLaMA-3.1-8B
Base	68.86	58.76
$K = 2$	64.79	63.96
$K = 4$	65.83	64.27
$K = 6$	64.69	64.19
$K = 8$	64.83	63.53

Table 3: Average performance of different K values on RULER.

This suggests that the reconstruction-based training objective encourages robust global structure understanding rather than overfitting to a specific difficulty level. Interestingly, intermediate option lengths ($K = 4$ and $K = 6$) slightly outperform both smaller and larger values of K, indicating a potential trade-off between task difficulty and learning efficiency. In addition, training of single K benefits weaker LLaMA-3.1-8B-Instruct, while degrades performance of stronger Qwen2.5-7B-Instruct-1M, suggesting different sensitivity to task specialization across backbones. Qwen2.5-7B-Instruct-1M may be more likely to overfit to a single pattern of reconstruction. It also highlights the importance our data curation strategy in main experiment, which mixes training samples from various K.

6 Related Work

Reinforcement Learning with Verifiable Rewards. Reinforcement Learning with Verifiable Rewards provides an objective and scalable framework for improving the reasoning capabilities of large language models by supervising them with ground-truth answers (Lambert et al., 2025; Guo et al., 2025). Previous work shows that RLVR can elevate models to expert-level reasoning performance by encouraging the discovery of correct internal reasoning trajectories through outcome-only rewards (OpenAI et al., 2024; Kim et al., 2025; Huang and Yang, 2025; Mo et al., 2025; Wang et al., 2026b). Most existing RLVR studies focus on self-contained reasoning tasks, where the central challenge is to recover a valid trajectory (Yue et al., 2025). In these settings, emergent behaviors, such as self-reflection, have been identified as important contributors to performance gains (Gandhi et al., 2025). Recent reasoning-oriented models, including OpenAI's o1 series and DeepSeek, have popularized RLVR-based training pipelines and reinforcement learning algorithms such as GRPO.

Long-Context Training of LLMs. Training large language models for long-context reasoning poses challenges that differ fundamentally from those addressed by standard reinforcement learning with verifiable rewards. The outcome reward provides limited guidance in long-context settings, where success depends on identifying and grounding relevant evidence from extensive context (Wan et al., 2025). Most existing approaches to long-context training rely on supervised fine-tuning with synthetic data rather than reinforcement learning (Li et al., 2024; Yen et al., 2025; Chen et al., 2025b). Prior work pads questions with unrelated passages, shuffles document order, or fills contexts with irrelevant text to artificially increase sequence length (Trivedi et al., 2023). Improving long-context reasoning has become increasingly important due to the rapid emergence of agent applications (Zhao et al., 2024; Team et al., 2025a,b; Prabhakar et al., 2025; Gandhi et al., 2026).

7 Conclusion

In this work, we present an unsupervised reinforcement learning framework for improving the long-context capabilities of LLMs. By formulating reconstruction as a sequential decision-making problem with verifiable rewards derived directly from raw documents, our approach eliminates the need for manually curated long-context data or teacher-model supervision. This enables a scalable and principled alternative to existing long-context training paradigms. Extensive experiments on RULER and LongBench v2 demonstrate that reconstruction-based RLVR effectively enhances long-context performance across multiple backbone models and context lengths. Beyond empirical gains, our findings highlight document structure itself as a valuable and underexplored supervision signal. We hope that this work motivates further research into unsupervised and self-supervised training objectives that leverage intrinsic structure in raw data and contributes to the development of more capable and scalable long-context language models.

Limitations

Despite its effectiveness, our approach has several limitations. Our method requires access to sufficiently long and well-structured documents, which may limit its applicability in domains where long-form data is scarce or noisy. In addition, we observe that the benefits of reconstruction training vary across backbone models. And models may have different requirements about document quality and length. Finally, while our method scales well within the evaluated data range, its behavior at substantially larger scales and with different model sizes remains an open area for future work.

References

Chenxin An, Fei Huang, Jun Zhang, Shansan Gong, Xipeng Qiu, Chang Zhou, and Lingpeng Kong. 2024. Training-free long-context scaling of large language models. In *Forty-first International Conference on Machine Learning*.

Anthropic, PBC. 2025. Claude. `https://www.claude.ai`. Accessed: 2025-10-01.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2025a. LongBench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3639–3664, Vienna, Austria. Association for Computational Linguistics.

Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2025b. Longwriter: Unleashing 10,000+ word generation from long context LLMs. In *The Thirteenth International Conference on Learning Representations*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *International Conference on Machine Learning*.

Guanzheng Chen, Xin Li, Michael Shieh, and Lidong Bing. 2025a. LongPO: Long context self-evolution of large language models through short-to-long preference optimization. In *The Thirteenth International Conference on Learning Representations*.

Guanzheng Chen, Michael Qizhe Shieh, and Lidong Bing. 2026. LongRLVR: Long-context reinforcement learning requires verifiable context rewards. In *The Fourteenth International Conference on Learning Representations*.

Hao Chen, Abdul Waheed, Xiang Li, Yidong Wang, Jindong Wang, Bhiksha Raj, and Marah I Abdin. 2025b. On the diversity of synthetic data and its impact on training large language models.

Jiali Cheng and Hadi Amiri. 2025. Do students debias like teachers? on the distillability of bias mitigation methods. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*.

Yufeng Du, Minyang Tian, Srikanth Ronanki, Subendhu Rongali, Sravan Babu Bodapati, Aram Galstyan, Azton Wells, Roy Schwartz, Eliu A Huerta, and Hao Peng. 2025. Context length alone hurts LLM performance despite perfect retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23281–23298, Suzhou, China. Association for Computational Linguistics.

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D. Goodman. 2025. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars.

Kanishk Gandhi, Shivam Garg, Noah D. Goodman, and Dimitris Papailiopoulos. 2026. Endless terminals: Scaling rl environments for terminal agents.

Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. 2025. How to train long-context language models (effectively). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7376–7399, Vienna, Austria. Association for Computational Linguistics.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q.

Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. 2024a. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*.

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. 2024b. Found in the middle: Calibrating positional attention bias improves long context utilization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14982–14995, Bangkok, Thailand. Association for Computational Linguistics.

Yichen Huang and Lin F. Yang. 2025. Winning gold at imo 2025 with a model-agnostic verification-and-refinement pipeline.

Zenan Huang, Yihong Zhuang, Guoshan Lu, Zeyu Qin, Haokai Xu, Tianyu Zhao, Ru Peng, Jiaqi Hu, Zhanming Shen, Xiaomeng Hu, Xijun Gu, Peiyi Tu, Jiaxin Liu, Wenyu Chen, Yuzhuo Fu, Zhiting Fan, Yanmei Gu, Yuanyuan Wang, Zhengkai Yang, Jianguo Li, and Junbo Zhao. 2025. Reinforcement learning with rubric anchors.

Minwu Kim, Anubhav Shrestha, Safal Shrestha, Aadim Nepal, and Keith W. Ross. 2025. RLVR vs. distillation: Understanding accuracy and capability in LLM mathematical reasoning. In *The 5th Workshop on Mathematical Reasoning and AI at NeurIPS 2025*.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, Lei M Zhang, Kay McKinney, Disha Shrivastava, Cosmin Paduraru, George Tucker, Doina Precup, Feryal Behbahani, and Aleksandra Faust. 2025. Training language models to self-correct via reinforcement learning. In *The Thirteenth International Conference on Learning Representations*.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. 2025. Tulu 3: Pushing frontiers in open language model post-training.

Yanyang Li, Shuo Liang, Michael Lyu, and Liwei Wang. 2024. Making long-context language models better multi-hop reasoners. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2462–2475, Bangkok, Thailand. Association for Computational Linguistics.

Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. 2025. ProRL: Prolonged reinforcement learning expands reasoning boundaries in large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Enzhe Lu, Zhejun Jiang, Jingyuan Liu, Yulun Du, Tao Jiang, Chao Hong, Shaowei Liu, Weiran He, Enming Yuan, Yuzhi Wang, Zhiqi Huang, Huan Yuan, Suting Xu, Xinran Xu, Guokun Lai, Yanru Chen, Huabin Zheng, Junjie Yan, Jianlin Su, Yuxin Wu, Neo Y. Zhang, Zhilin Yang, Xinyu Zhou, Mingxing Zhang, and Jiezhong Qiu. 2025. Moba: Mixture of block attention for long-context llms.

Ximing Lu, David Acuna, Jaehun Jung, Jian Hu, Di Zhang, Shizhe Diao, Yunheng Zou, Shaokun Zhang, Brandon Cui, Mingjie Liu, Hyunwoo Kim, Prithviraj Ammanabrolu, Jan Kautz, Yi Dong, and Yejin Choi. 2026. Golden goose: A simple trick to synthesize unlimited rlvr tasks from unverifiable internet text.

Zhanfeng Mo, Xingxuan Li, Yuntao Chen, and Lidong Bing. 2025. Multi-agent tool-integrated policy optimization.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher

Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. 2024. Openai o1 system card.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2024. YaRN: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*.

Akshara Prabhakar, Roshan Ram, Zixiang Chen, Silvio Savarese, Frank Wang, Caiming Xiong, Huan Wang, and Weiran Yao. 2025. Enterprise deep research: Steerable multi-agent deep research for enterprise analytics. *arXiv preprint arXiv:2510.17797*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du, Chenzhuang Du, Dikang Du, Yulun Du, Yu Fan, Yichen Feng, Kelin Fu, Bofei Gao, Hongcheng Gao, Peizhong Gao, Tong Gao, Xinran Gu, Longyu Guan, Haiqing Guo, Jianhang Guo, Hao Hu, Xiaoru Hao, Tianhong He, Weiran He, Wenyang He, Chao Hong, Yangyang Hu, Zhenxing Hu, Weixiao Huang, Zhiqi Huang, Zihao Huang, Tao Jiang, Zhejun Jiang, Xinyi Jin, Yongsheng Kang, Guokun Lai, Cheng Li, Fang Li, Haoyang Li, Ming Li, Wentao Li, Yanhao Li, Yiwei Li, Zhaowei Li, Zheming Li, Hongzhan Lin, Xiaohan Lin, Zongyu Lin, Chengyin Liu, Chenyu Liu, Hongzhang Liu, Jingyuan Liu, Junqi Liu, Liang Liu, Shaowei Liu, T. Y. Liu, Tianwei Liu, Weizhou Liu, Yangyang Liu, Yibo Liu, Yiping Liu, Yue Liu, Zhengying Liu, Enzhe Lu, Lijun Lu, Shengling Ma, Xinyu Ma, Yingwei Ma, Shaoguang Mao, Jie Mei, Xin Men, Yibo Miao, Siyuan Pan, Yebo Peng, Ruoyu Qin, Bowen Qu, Zeyu Shang, Lidong Shi, Shengyuan Shi, Feifan Song, Jianlin Su, Zhengyuan Su, Xinjie Sun, Flood Sung, Heyi Tang, Jiawen Tao, Qifeng Teng, Chensi Wang, Dinglu Wang, Feng Wang, Haiming Wang, Jianzhou Wang, Jiaxing Wang, Jinhong Wang, Shengjie Wang, Shuyi Wang, Yao Wang, Yejie Wang, Yiqin Wang, Yuxin Wang, Yuzhi Wang, Zhaoji Wang, Zhengtao Wang, Zhexu Wang, Chu Wei, Qianqian Wei, Wenhao Wu, Xingzhe Wu, Yuxin Wu, Chenjun Xiao, Xiaotong Xie, Weimin Xiong, Boyu Xu, Jing Xu, Jinjing Xu, L. H. Xu, Lin Xu, Suting Xu, Weixin Xu, Xinran Xu, Yangchuan Xu, Ziyao Xu, Junjie Yan, Yuzi Yan, Xiaofei Yang, Ying Yang, Zhen Yang, Zhilin Yang, Zonghan Yang, Haotian Yao, Xingcheng Yao, Wenjie Ye, Zhuorui Ye, Bohong Yin, Longhui Yu, Enming Yuan, Hongbang Yuan, Mengjie Yuan, Haobing Zhan, Dehao Zhang, Hao Zhang, Wanlu Zhang, Xiaobin Zhang, Yangkun Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Haotian

Zhao, Yikai Zhao, Huabin Zheng, Shaojie Zheng, Jianren Zhou, Xinyu Zhou, Zaida Zhou, Zhen Zhu, Weiyu Zhuang, and Xinxing Zu. 2025a. Kimi k2: Open agentic intelligence.

MiroMind Team, Song Bai, Lidong Bing, Carson Chen, Guanzheng Chen, Yuntao Chen, Zhe Chen, Ziyi Chen, Jifeng Dai, Xuan Dong, et al. 2025b. Mirothinker: Pushing the performance boundaries of open-source research agents via model, context, and interactive scaling. *arXiv preprint arXiv:2511.11793*.

Tongyi DeepResearch Team, Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, Kuan Li, Liangcai Su, Litu Ou, Liwen Zhang, Pengjun Xie, Rui Ye, Wenbiao Yin, Xinmiao Yu, Xinyu Wang, Xixi Wu, Xuanzhong Chen, Yida Zhao, Zhen Zhang, Zhengwei Tao, Zhongwang Zhang, Zile Qiao, Chenxi Wang, Donglei Yu, Gang Fu, Haiyang Shen, Jiayin Yang, Jun Lin, Junkai Zhang, Kui Zeng, Li Yang, Hailong Yin, Maojia Song, Ming Yan, Minpeng Liao, Peng Xia, Qian Xiao, Rui Min, Ruixue Ding, Runnan Fang, Shaowei Chen, Shen Huang, Shihang Wang, Shihao Cai, Weizhou Shen, Xiaobin Wang, Xin Guan, Xinyu Geng, Yingcheng Shi, Yuning Wu, Zhuo Chen, Zijian Li, and Yong Jiang. 2025c. Tongyi deepresearch technical report.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.

Fanqi Wan, Weizhou Shen, Shengyi Liao, Yingcheng Shi, Chenliang Li, Ziyi Yang, Ji Zhang, Fei Huang, Jingren Zhou, and Ming Yan. 2025. Qwenlong-l1: Towards long-context large reasoning models with reinforcement learning.

Siyuan Wang, Gaokai Zhang, Li Lyna Zhang, Ning Shang, Fan Yang, Dongyao Chen, and Mao Yang. 2025. Loongrl: Reinforcement learning for advanced reasoning over long contexts.

Yinjie Wang, Tianbao Xie, Ke Shen, Mengdi Wang, and Ling Yang. 2026a. Rlanything: Forge environment, policy, and reward model in completely dynamic rl system.

Zirui Wang, Junyi Zhang, Jiaxin Ge, Long Lian, Letian Fu, Lisa Dunlap, Ken Goldberg, XuDong Wang, Ion Stoica, David M. Chan, Sewon Min, and Joseph E. Gonzalez. 2026b. Visgym: Diverse, customizable, scalable environments for multimodal agents.

Penghao Wu, Yushan Zhang, Haiwen Diao, Bo Li, Lewei Lu, and Ziwei Liu. 2025a. Visual jigsaw post-training improves mllms.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. 2025b. Retrieval head mechanistically explains long-context factuality. In *The Thirteenth International Conference on Learning Representations*.

Yuhao Wu, Yushi Bai, Zhiqiang Hu, Roy Ka-Wei Lee, and Juanzi Li. 2025c. Longwriter-zero: Mastering ultra-long text generation via reinforcement learning.

Ran Xu, Tianci Liu, Zihan Dong, Tony You, Ilgee Hong, Carl Yang, Linjun Zhang, Tao Zhao, and Haoyu Wang. 2026. Alternating reinforcement learning for rubric-based reward modeling in non-verifiable llm post-training.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025a. Qwen3 technical report.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, Junyang Lin, Kai Dang, Kexin Yang, Le Yu, Mei Li, Minmin Sun, Qin Zhu, Rui Men, Tao He, Weijia Xu, Wenbiao Yin, Wenyuan Yu, Xiafei Qiu, Xingzhang Ren, Xinlong Yang, Yong Li, Zhiying Xu, and Zipeng Zhang. 2025b. Qwen2.5-1m technical report.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. 2025. HELMET: How to evaluate long-context models effectively and thoroughly. In *The Thirteenth International Conference on Learning Representations*.

Edward Yeo, Yuxuan Tong, Xinyao Niu, Graham Neubig, and Xiang Yue. 2025. Demystifying long chain-of-thought reasoning in LLMs. In *ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi

Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. 2025. Dapo: An open-source llm reinforcement learning system at scale.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model?

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun MA, and Junxian He. 2025a. SimpleRL-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. In *Second Conference on Language Modeling*.

Zhiyuan Zeng, Hamish Ivison, Yiping Wang, Lifan Yuan, Shuyue Stella Li, Zhuorui Ye, Siting Li, Jacqueline He, Runlong Zhou, Tong Chen, Chenyang Zhao, Yulia Tsvetkov, Simon Shaolei Du, Natasha Jaques, Hao Peng, Pang Wei Koh, and Hannaneh Hajishirzi. 2025b. Rlve: Scaling up reinforcement learning for language models with adaptive verifiable environments.

Jiajie Zhang, Zhongni Hou, Xin Lv, Shulin Cao, Zhenyu Hou, Yilin Niu, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. 2025. LongReward: Improving long-context large language models with AI feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3718–3739, Vienna, Austria. Association for Computational Linguistics.

Jun Zhao, Can Zu, Xu Hao, Yi Lu, Wei He, Yiwen Ding, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. LONGAGENT: Achieving question answering for 128k-token-long documents through multi-agent collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16310–16324, Miami, Florida, USA. Association for Computational Linguistics.

Dawei Zhu, Nan Yang, Liang Wang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. 2024. PoSE: Efficient context window extension of LLMs via positional skip-wise training. In *The Twelfth International Conference on Learning Representations*.

Yonghao Zhuang, Lanxiang Hu, Longfei Yun, Souvik Kundu, Zhengzhong Liu, Eric P. Xing, and Hao Zhang. 2025. Scaling long context training data by long-distance referrals. In *The Thirteenth International Conference on Learning Representations*.

A Appendix

A.1 Reconstruction Prompt

We append the reconstruction prompt below.

Reconstruction Prompt

The following document contains missing segments marked as `<C_i>MISSING</C_i>`.
Please reason about the logical and narrative structure of the document and select appropriate chunks one by one from the given options to reconstruct it.
Then, output the label for each missing chunk by order in `\boxed{}` separated by commas.
The document is as follows: `{corrupted document}`
The options are: `{options}`

A.2 On continual Pretraining

We further explore continual pretraining by training the models for one epoch on our curated long-document corpus. However, this approach results in significantly worse performance compared to the original models. We attribute this degradation to two main factors. First, the quality of our collected long documents may not exceed that of the proprietary data used during the original pretraining of the models. Second, continual pretraining on instruction-tuned models may disrupt their instruction-following capabilities. As a result, continual pretraining of LLaMA-3.1-8B-Instruct and Qwen2.5-7B-Instruct-1M does not yield performance improvements.

A.3 On Validation

During training, we monitor model performance on a held-out validation set to assess optimization stability and learning dynamics in reconstruction training. Specifically, we track three metrics: (1) the success rate of answer extraction (i.e., producing a valid permutation), (2) the dense reward, and (3) the sparse reward for Qwen2.5-7B-Instruct-1M. As shown in Figure 9, all three metrics improve smoothly over training steps, indicating stable optimization without severe oscillation or collapse. Notably, the dense reward increases earlier and more steadily than the sparse reward. Overall, these validation trends suggest that the proposed reconstruction-based RLVR framework provides a stable and effective training.

Model	Task	32k		64k		128k	
		base	ours	base	ours	base	ours
Qwen	vt	58.32	67.36	46.80	61.56	54.84	56.28
	cwe	87.92	90.66	82.46	84.38	69.78	73.42
	fwe	85.53	91.27	83.07	84.33	77.13	80.33
	qa_1	77.40	78.40	73.00	71.20	70.60	65.00
	qa_2	60.00	64.20	56.80	61.40	49.40	59.60
	avg	73.83	**78.38**	68.43	**72.57**	64.35	**65.65**
LLaMA	vt	78.48	83.76	54.04	77.20	24.72	49.48
	cwe	86.54	92.56	60.92	80.20	10.96	28.14
	fwe	81.40	87.00	73.47	73.93	61.00	67.67
	qa_1	76.60	74.40	74.20	74.40	68.00	67.80
	qa_2	47.80	56.20	43.20	51.40	40.00	43.00
	avg	74.16	**78.78**	61.17	**71.43**	40.94	**51.22**

Table 4: Performance comparison across sequence lengths.

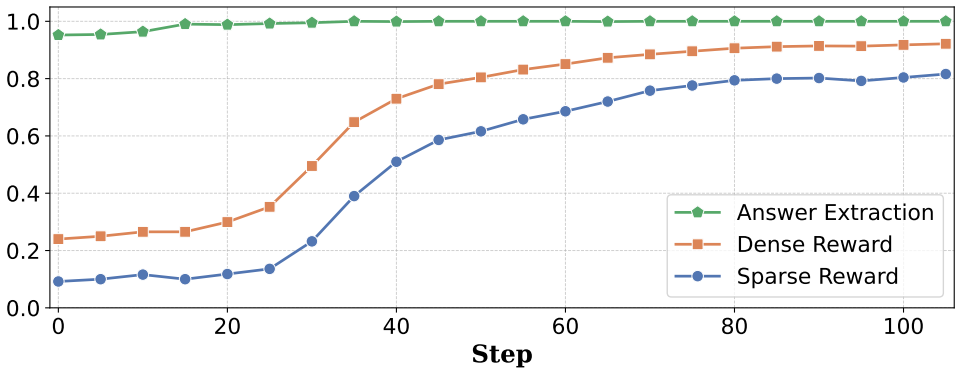


Figure 9: We report metrics on validation set during training process.

A.4 Ruler Score Details

As shown in Table 4, we report the evaluation scores for each subtask for RULER. We can see that our method can surpass base models in almost all cases.

A.5 Scaling Data to 3W

We collect 10,000 samples from each domain: books, arXiv, and code. We end up with 30,000 long document. The ratio of data with $K = 2, 4, 6, 8$ is $1 : 1 : 2 : 2$. We train on backbone model Qwen2.5-7B-Instruct-1M. The performance of our model can surpass baseline by about 10 points.