Beyond Reward Engineering: A Data Recipe for Long-Context Reinforcement Learning

Anonymous authors
Paper under double-blind review

Abstract

Long-context reasoning is an essential capability for large language models, particularly when they are deployed as autonomous agents that must reason over lengthy trajectories. Reinforcement learning (RL) has recently emerged as a dominant paradigm for improving this ability, yet existing work largely focuses on reward engineering while diverse training data remains scarce. We revisit this problem from a data-centric perspective and show that a simple yet effective data recipe alone, paired with a minimal outcome-based GRPO setup, suffices to substantially improve long-context reasoning. Our recipe targets three complementary task families – **retrieval**, **multi-evidence synthesis**, and **reasoning** – for which we construct and curate eight datasets totaling ∼14K examples. Experiments on three models (Qwen3-4B/8B/30B-A3B) yield average gains of +7.2/+3.2/+6.4 points across seven long-context benchmarks, surpassing prior RL training sets. We further demonstrate that these gains transfer to agentic tasks, where continuing RL training on an agent-tuned model with our data recipe improves GAIA by +4.8 and BrowseComp by +7.0 points. We will release our datasets to facilitate future research.

1 Introduction

Large language models (LLMs) have been increasingly deployed as autonomous agents, successfully tackling complex real-world tasks such as web searching (Mialon et al., 2024; Wei et al., 2025), software engineering (Yang et al., 2024), and GUI automation (Qin et al., 2025). These applications expose LLMs to extensive web pages, large-scale code repositories, and massive action histories, where task-relevant evidence is often scattered across lengthy inputs. LLMs must identify, integrate, and reason over this information to produce correct decisions. Long-context reasoning has therefore become an essential capability for modern LLMs.

Previous work has proven the effectiveness of reinforcement learning (RL) in enhancing the reasoning ability of LLMs (OpenAI et al., 2026; Guo et al., 2025), yet most efforts focus on short-context tasks. Extending such gains to long-context reasoning remains underexplored. Recent studies indicate that the outcome-only reward of standard Reinforcement Learning with Verifiable Rewards (RLVR) provides too sparse a signal to guide the model in locating evidence over long inputs, leading to plateaued contextual recall and shortcut reasoning that bypass grounding (Chen et al., 2026a; Guan et al., 2026). In response, most existing efforts focus on reward design of long-context RL (Ping et al., 2026b; Chen et al., 2026a; Peng et al., 2026; Guan et al., 2026), supplementing the outcome reward with auxiliary signals that explicitly score evidence grounding or intermediate reasoning. Meanwhile, high-quality training data remains scarce: while several synthetic data pipelines have been proposed (Shen et al., 2025; Wang et al., 2026; Xiao et al., 2026), the resulting datasets are either narrow in task coverage or kept closed-source.

In this paper, we revisit long-context RL from a data-centric perspective and show that a diverse data recipe alone is sufficient to substantially improve long-context reasoning, without special reward engineering. We hypothesize that long-context reasoning decomposes into a small set of complementary core abilities that need to be exercised jointly during training, a view echoed by prior empirical analyses (Li et al., 2024a; Xiao et al., 2026). For example, given the web pages gathered to answer a GAIA-like (Mialon et al., 2024) question such as *"What is the combined population of Germany's three most populous land neighbors that*

Figure 1: A long-context reasoning example requiring three abilities that define our training mixture (§3.1): **Retrieval**, **Multi-evidence Synthesis**, and **Reasoning**.

aren't EU founding members?" (Figure 1), the model must (1) locate candidate countries together with their populations, (2) integrate the border-relationship and EU-founding-member information to filter out the founding members, and (3) perform math reasoning over the three most populous remaining countries to compute the final sum.

Guided by this hypothesis, our data recipe spans three corresponding task categories: (i) **Retrieval**, where the model must locate relevant evidence from a long input, especially when the evidence is semantically indirect, or the information is scattered among abundant distractors; (ii) **Multi-evidence Synthesis**, where the model must integrate evidence from different parts of the input to derive an answer that no single piece can provide; and (iii) **Reasoning**, where the model must perform complex multi-step problem solving (e.g., math) over a long input. We accordingly collect and construct training data covering these three task categories, and conduct thorough RL experiments to evaluate the effectiveness of our data recipe on long-context benchmarks as well as its transfer to agentic tasks.

We train on three models, Qwen3-4B-Thinking-2507, Qwen3-8B, and Qwen3-30B-A3B-Thinking-2507 (Qwen Team, 2025a), and evaluate across seven long-context benchmarks. Without any special reward engineering or staged RL training, our recipe delivers consistent gains, achieving average improvements of +7.2, +3.2, and +6.4 points on the three models respectively. Under the same RL setup, it outperforms two prior long-context RL training sets, DocQA-RL-1.6K (Wan et al., 2025) and KeyChain (Wang et al., 2026). We further show that these long-context gains transfer to agentic tasks, as continuing long-context RL on an agent-tuned model improves GAIA by 4.8 and BrowseComp by 7.0 points, suggesting that strengthening core long-context abilities effectively benefits long-horizon agentic capability.

2 Related Works

Long-context Reasoning via Post-training. A growing line of work improves long-context reasoning through post-training, where a central challenge is curating diverse and high-quality long-context supervision. On the SFT side, LongAlign (Bai et al., 2024a) builds a long instruction-following dataset paired with a balanced training strategy, Li et al. (2024b) allows the model to generate its own SFT examples for self-improvement, and ACC (Su et al., 2026) compiles long agent trajectories into long-context training data. Reinforcement learning has more recently emerged as an effective paradigm for optimizing reasoning behavior under long context, which we discuss next.

Reinforcement Learning in Long-context Scenarios. Existing efforts in long-context RL can be broadly divided into two lines: *algorithm-centric* and *data-centric*. The algorithm-centric line improves the

Category	Dataset	Description	Size	Length
Retrieval	FuzzyNeedle	IS-A paraphrased needles, parent-class queries	1,500	0–32K
	MultiNeedle	Shuffled near-duplicate multiple needle retrieval	1,500	0–32K
Multi-Evidence	CrossEntity	Cross-entity aggregation over derived attributes	3,521	0–64K
	WebSearch	Bridging chains with sibling-entity distractors	684	0–64K
	MultiQuery	Embedded queries answered in order	226	0–32K
	KeyChain	UUID pointer chains hiding the question	654	0–32K
	LongDocQA	Extended from HotpotQA / MuSiQue / Qasper	3,422	0–64K
Reasoning	LongMath	Math problems with scattered variables	2,562	0–64K

Table 1: Composition of our long-context training mixture across the three task categories. "Length" denotes the estimated input length range in tokens.

RL training procedure. On the reward side, recent works supplement the outcome-only reward of RLVR with auxiliary signals that explicitly score evidence grounding (Chen et al., 2026a; Guan et al., 2026; Whitecross & Rahimi, 2026) or intermediate-step quality (Peng et al., 2026). On the optimization side, it includes modified policy optimization (Shen et al., 2025) and selective weight updates (Ping et al., 2026a). The data-centric line constructs high-quality long-context RL training sets, such as DocQA-RL (Wan et al., 2025), which contains ∼1.6K document QA problems, LoongRL (Wang et al., 2026), which synthesizes examples through a KeyChain UUID-driven pipeline, and Xiao et al. (2026), which builds a dataset via atomic-skill decomposition. Yet their training data is either limited in scale and diversity or confined to a single synthetic task format.

Our work focuses on the data-centric line. Different from these prior efforts, we contribute a more unified *recipe* that exercises three complementary core abilities of long-context reasoning, yielding consistent gains across various long-context benchmarks.

3 Methodology

Our methodology has two parts: a long-context training data mixture that jointly stresses three core abilities of long-context reasoning (§3.1), and a minimal RL setup with an outcome-based reward (§3.2).

3.1 Long-Context Training Data

Our training mixture is organized around the three task categories introduced in §1—**Retrieval**, **Multi-evidence Synthesis**, and **Reasoning**—each isolating a distinct ability that is essential for long-context reasoning. For each category, we analyze the main difficulties models face and introduce the corresponding dataset(s) we construct to address them. We then conclude with the overall data recipe and describe our data curation process.

3.1.1 Retrieval

Vanilla needle-in-a-haystack (Kamradt, 2023) is largely saturated on recent long-context models. However, information retrieval in current long-context LMs still relies heavily on lexical pattern matching (Modarressi et al., 2025; Xu et al., 2024), and degrades sharply once the relevant information becomes abundant and scattered across the context (e.g., as the number of needles grows) (Hsieh et al., 2024; Vodrahalli et al., 2024; Li et al., 2025). We therefore construct two lightweight datasets that target these two shortcomings.

FuzzyNeedle. FuzzyNeedle removes lexical shortcuts by paraphrasing the needle through a Wikidata *IS-A* relation so that the query and the evidence share no keywords, in a similar spirit to NoLiMa (Modarressi et al., 2025). We first sample a target entity (e.g., `fish`) and an instance of it (e.g., `salmon`), and attach

this instance to a fictional statement (e.g., *"Alice loves the texture of salmon"*). We then additionally sample sibling entities (e.g., `vegetable`) and draw several instances to form analogous distractors for each of them. All statements are scattered throughout an unrelated long passage, and the model is queried at the target entity level (e.g., *"Who enjoys eating fish?"*). Answering correctly requires resolving the is-instance-of relation rather than relying on keyword matching, as well as rejecting the distractor statements.

MultiNeedle. MultiNeedle requires the model to distinguish among many near-duplicate needles and retrieve a target one by its order of occurrence, following MRCR (Vodrahalli et al., 2024). We prepare a set of diverse topics, and prompt LLMs to draft multiple user–model conversations on each topic. Conversations from different topics are then shuffled and joined into a single long context, and the model is asked to retrieve the K*-th* conversation on a specified topic. Models thus need to exhaustively locate every needle, keep a count of their order and discriminate among similar candidates.

3.1.2 Multi-evidence Synthesis

Multi-evidence synthesis is a common pattern in real-world long-context tasks (Yang et al., 2018; Tang & Yang, 2024), which requires integrating information scattered across different parts of the input into a coherent answer. Long-context LMs exhibit several shortcomings here: (i) *surface-form shortcuts*, utilizing only verbatim spans rather than synthesizing the evidence; (ii) *incomplete coverage*, where the model skips parts of the long input and misses information needed for the answer (Gupta et al., 2025); and (iii) *distractor confusion*, where topically related documents are mistaken for gold evidence (Vodrahalli et al., 2024). For this kind of task, we construct new synthetic datasets targeting these shortcomings, and adapt long-context QA sources from prior work.

CrossEntity. This dataset targets both surface-form shortcuts and incomplete coverage by producing questions whose answer is an aggregation over *derived* attributes of multiple entities, which are never stated directly in the documents. Given a group of peer entities from the same Wikidata category (e.g., NBA players), we prompt LLMs to identify both *direct attributes* mentioned in the entity pages (e.g., `birth_year`) and *derived attributes* obtainable only by synthesizing over them (e.g., `winning_age` = `winning_year` − `birth_year`). We then generate cross-entity comparison questions about the derived attributes (e.g., *"Which of these players was the youngest at the time of the first championship?"*). Models are challenged to first integrate evidence within each individual document, and then perform a global aggregation across all entities without any omission.

WebSearch. Targeting distractor confusion, WebSearch draws inspiration from search-agent data construction (Liu et al., 2025; Lu et al., 2025), where referencing a target entity through a sequence of bridging relations naturally yields a multi-hop question. We first collect relation triples whose entities are linked into a single chain, and substitute most entities with letter placeholders to generate a question, e.g., *"A is the director of B; the lead actor of B is C; . . . ; who is A?"*. Web pages of the gold triples form the supporting documents. We then prompt LLMs to brainstorm similar entities satisfying the same relation (e.g., *"E is the director of F"*) and collect their pages as distractors. All documents are interleaved in a random order. The models must trace the chain hop by hop, and distinguish the documents that exactly match each bridging relation from topically similar distractors.

MultiQuery. To stress incomplete coverage, we introduce a straightforward task in which missing any piece of evidence directly causes an incorrect response. We sample documents from pre-training corpus and prompt LLMs to generate QA pairs. Queries are then randomly embedded in the concatenated documents and models are asked to answer all questions in the order they appear. In this task, each question constitutes a required piece of evidence, and failing to gather any of them makes the final response incomplete.

KeyChain. We also construct a KeyChain dataset following the pipeline of LoongRL (Wang et al., 2026). We generate QA pairs from the same corpus as MultiQuery, hide each question behind a chain of UUID pointers, and scatter the chains throughout the assembled multi-document context. The model must trace the chain from a starting UUID and recover the actual question before answering.

LongDocQA. We include the training sets of HotpotQA (Yang et al., 2018), MuSiQue (Trivedi et al., 2022), and Qasper (Dasigi et al., 2021), and extend their length to further improve data diversity. For HotpotQA, we replace the truncated supporting paragraphs with the corresponding full Wikipedia articles; for MuSiQue and Qasper, we keep the original input intact and pad it with irrelevant texts.

3.1.3 Reasoning

Previous works observe that long inputs alone degrade a model's reasoning performance (Levy et al., 2024; Du et al., 2025), which motivate us to train LLMs on long-form reasoning tasks so as to preserve deep thinking ability when inputs get lengthy. Training data of such reasoning-intensive problem is naturally scarce in long-context setting, while comparably abundant in short-context math and code domains (Hendrycks et al., 2021; Guo et al., 2025). We then address this gap by adapting difficult short-context math problems into the long-context format.

LongMath. We sample difficult problems from two challenging math datasets, DeepMath-103K and MATH-Hard (He et al., 2026; Hendrycks et al., 2021). Following the spirit of LongReason (Ling et al., 2025), we prompt an LLM to rewrite each question into a narrative scenario (e.g., science fiction, finance, or engineering), where the original variables and numerical conditions are distributed across at least five story fragments. A second LLM verifies the rewrites and filters out the unsolvable ones. Finally, the fragments are scattered throughout an irrelevant long document, and the model must locate all of them and perform complex mathematical derivation on top of this lengthy input.

3.1.4 Data Curation and Final Recipe

Across all sources, we apply a two-stage post-processing pipeline. We first discard samples exceeding 64K tokens, in order to keep rollout cost tractable while preserving sufficient long-context difficulty. We then rollout each remaining question four times with Qwen3-30B-A3B-Thinking-2507 (Qwen Team, 2025a) and retain only those that are neither all-correct nor all-wrong, isolating problems of moderate difficulty that yield informative learning signals for RL. This yields a total of 14,069 examples, as detailed in Table 1.

3.2 Long-Context Reinforcement Learning

GRPO objective. We train with Group Relative Policy Optimization (GRPO) (Shao et al., 2024). For each prompt x, GRPO samples a group of G rollouts $\{y_i\}_{i=1}^{G}$ from $\pi_{\theta_{\text{old}}}$, scores them with rewards $\{r_i\}_{i=1}^{G}$, and assigns each rollout the group-relative advantage

$$A_i = \frac{r_i - \text{mean}(\{r_j\}_{j=1}^{G})}{\text{std}(\{r_j\}_{j=1}^{G})}. \tag{1}$$

GRPO then forms the per-token surrogate loss

$$\mathcal{L}_{i,t}(\theta) = \min\big(\rho_{i,t}\, A_i, \text{clip}(\rho_{i,t}, 1-\varepsilon, 1+\varepsilon)\, A_i\big), \tag{2}$$

where

$$\rho_{i,t} = \frac{\pi_\theta(y_{i,t} \mid x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t} \mid x, y_{i,<t})}, \tag{3}$$

and maximizes the objective

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{x,\{y_i\}}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\mathcal{L}_{i,t}(\theta)\Bigg]. \tag{4}$$

Following recent practice in reasoning RL (Yu et al., 2026), we omit the KL regularization term.

Model	Avg.	Multi-hop QA		Holistic LC Reasoning			Synthetic LC Reasoning	
		LBv1-QA	FRAMES	LBv2	AA-LCR	DocFinQA	LongReason	GraphWalks
Qwen3-4B-Thinking-2507	53.42	62.60	77.55	40.60	35.25	39.80	69.50	48.67
+ DocQA-RL-1.6K	55.08	63.80	78.28	42.25	32.50	48.05	71.28	49.42
+ KeyChain-15K	57.97	66.80	82.40	43.39	41.00	47.29	76.20	48.70
+ **Ours**	**60.63**	**68.30**	**82.52**	**47.60**	**45.75**	**48.37**	**82.12**	**49.74**
Qwen3-8B-128K	49.93	62.90	74.76	33.75	28.75	38.83	68.26	**42.27**
+ DocQA-RL-1.6K	49.08	63.70	71.36	34.84	28.50	37.96	67.88	39.29
+ KeyChain-15K	52.08	66.50	**80.22**	34.14	**34.50**	35.79	72.04	41.36
+ **Ours**	**53.10**	**68.80**	77.67	**36.28**	**34.50**	**39.59**	**73.93**	40.96
Qwen3-30B-A3B-Thinking-2507	61.82	66.50	79.37	49.40	48.75	41.11	81.61	66.00
+ DocQA-RL-1.6K	64.02	68.90	81.55	53.18	43.30	51.41	80.86	**68.91**
+ KeyChain-15K	64.57	68.40	85.07	52.69	45.25	50.00	84.76	65.82
+ **Ours**	**68.23**	**77.30**	**85.80**	**54.21**	**56.00**	**51.63**	**84.89**	67.75

Table 2: Main results on seven long-context benchmarks. **Bold** denotes the best score per column within each base model. Our data recipe delivers consistent improvements across all benchmarks and base models, surpassing the two baselines.

Multi-dataset stabilization. Our training mixture combines 8 heterogeneous datasets whose reward distributions differ in both scale and variance. To stabilize training, we follow QwenLong-L1.5 (Shen et al., 2025) and apply (i) *task-balanced sampling*, which fixes the per-dataset sample ratio in each batch, and (ii) *task-level advantage normalization*, which rescales each rollout's advantage by the per-dataset reward standard deviation so that low-variance datasets are not washed out by high-variance ones.

Reward design. We use a simple reward design. For most datasets, we use a word-level recall reward $r = |\hat{y} \cap y^\star|/|y^\star|$ between the model's final answer $\hat{y}$ and the label $y^\star$. For CrossEntity and LongMath, we use DeepSeek-V3.2 as a judge, as recall reward can be hacked by enumerating candidates for CrossEntity and fails to credit equivalent forms of LongMath's numerical answers.

4 Experiments

4.1 Setup

Training Details. We conduct reinforcement learning on three base models, Qwen3-4B-Thinking-2507 (Qwen Team, 2025c), Qwen3-8B (Qwen Team, 2025d), and Qwen3-30B-A3B-Thinking-2507 (Qwen Team, 2025b), which are commonly adopted by recent long-context RL works (Ping et al., 2026b;a; Peng et al., 2026). For Qwen3-8B, we extend the context window from 32,768 to 131,072 tokens via YaRN (Peng et al., 2024), following the official recommendation of (Qwen Team, 2025d). We perform all RL training using the Miles framework (RadixArk, 2025). For each prompt we sample $G = 8$ rollouts with temperature 1.0 and top-$p = 0.95$, capping the input length at 64K tokens and the generation length at 16K tokens. The policy is updated with the Adam optimizer at a constant learning rate (1e-6 by default; see Appendix A for per-model settings), using a global batch size of 128 prompts and a PPO batch size of 1,024 (i.e., on-policy RL).

Benchmarks. We evaluate on a comprehensive suite of seven long-context benchmarks, organized into three groups according to the primary abilities they probe: (i) Multi-hop QA on real-world documents, covering LongBench v1 QA (Bai et al., 2024b), from which we follow Wan et al. (2025) and Wang et al. (2026) and use the five document-QA subsets (HotpotQA, 2WikiMultiHopQA, MuSiQue, NarrativeQA, and Qasper), and FRAMES (Krishna et al., 2025); (ii) Holistic Long-Context Reasoning, covering LongBench v2 (Bai et al., 2025), AA-LCR (Team, 2025), and DocFinQA (Reddy et al., 2024); (iii) Synthetic Long-Context Reasoning, covering LongReason (Ling et al., 2025) and GraphWalks (OpenAI, 2025). We use a unified

decoding configuration with temperature 0.6, top-p = 0.95, a maximum input length of 130K tokens, and a maximum output length of 60K tokens. For LongBench v1 QA and FRAMES, we report the maximum of exact match and LLM-judged accuracy as the final score. For LongBench v2 and AA-LCR, we report Avg@4 to reduce evaluation variance.

Baselines. We compare our data recipe against two existing long-context RL training datasets: (1) DocQA-RL-1.6K (Wan et al., 2025), the RL training set of QwenLong-L1, consisting of 1,591 long-document QA problems across three reasoning types ; (2) KeyChain-15K (Wang et al., 2026), the open-sourced LoongRL training data synthesized by the KeyChain UUID-driven pipeline, comprising ∼15K multi-hop QA examples built on HotpotQA, MuSiQue, and 2WikiMQA and uniformly preprocessed to a 16K-token context.

4.2 Main Results

Improvements across benchmarks and models. Table 2 summarizes our main results across three models. We highlight two observations. (i) Training with our data recipe yields consistent improvements : all seven benchmarks improve on Qwen3-4B-Thinking-2507 and Qwen3-30B-A3B-Thinking-2507, and six out of seven improve on Qwen3-8B-128K. In particular, significant gains are observed on LBv2, AA-LCR, and DocFinQA, three challenging benchmarks that evaluate holistic long-context reasoning (+7.0, +10.5, and +8.6 respectively for Qwen3-4B-Thinking-2507; +4.8, +7.3, and +10.5 for Qwen3-30B-A3B-Thinking-2507). (ii) Models of different sizes all benefit from our recipe, with average score gains of +7.2, +3.2, and +6.4 for Qwen3-4B-Thinking-2507, Qwen3-8B-128K, and Qwen3-30B-A3B-Thinking-2507 respectively. The consistent gains across model sizes and variants indicate the broad applicability of our training recipe.

Comparison with prior long-context RL data. Table 2 also compares our recipe against the two baseline datasets on all three models. Our data recipe achieves the highest average score on every model, outperforming the stronger KeyChain-15K baseline by +2.66, +1.02, and +3.66 on the 4B, 8B, and 30B-A3B models respectively. Moreover, our data recipe yields the largest gains on most benchmarks across three models, especially on the reasoning-heavy benchmarks (i.e., LBv2, AA-LCR, and LongReason).

Generalization to Longer Context. Although our RL training caps the input length at 64K tokens, the resulting gains transfer to substantially longer contexts. We re-evaluate Qwen3-4B-Thinking-2507 on LongBench v2 with the maximum input length extended to 230K tokens and stratify accuracy (avg@4) by example token length. As shown in the left panel of Figure 2, our model improves over the base by +8.6 on the in-distribution 0–64K bucket and retains positive gains on the longer buckets, including 64–256K (+6.9), 256–512K (+4.6), and 512K+ (+2.1), much beyond the training context. We further evaluate on HELMET (Yen et al., 2025), which supports evaluation at multiple input lengths; scores at 32K, 64K, and 128K are shown in the right panel of Figure 2. Together, these results indicate that our recipe instills a length-generic long-context reasoning skill rather than overfitting the 64K training distribution.

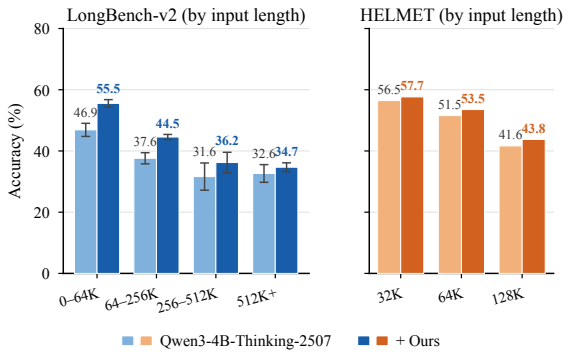

Figure 2: Training on our max-64K-token data recipe generalizes to substantially longer contexts. **Left**: LongBench-v2 accuracy stratified by input length. **Right**: HELMET accuracy at 32K, 64K, and 128K input lengths. Error bars on LongBench-v2 show the standard deviation over four sampling seeds.

5 Analysis

We present ablation studies to assess the effect of task categories, task-balancing techniques, and reward design on long-context RL performance. Each ablation uses Qwen3-4B-Thinking-2507 with the same con-

Data Category	Avg.	Multi-hop QA		Holistic LC Reasoning			Synthetic LC Reasoning	
		LBv1-QA	FRAMES	LBv2	AA-LCR	DocFinQA	LongReason	GraphWalks
Qwen3-4B-Thinking-2507	53.42	62.60	77.55	40.60	35.25	39.80	69.50	48.67
+ Retrieval	55.83	65.20	**79.98**	44.78	32.25	48.70	74.06	45.86
+ Multi-evidence	56.44	66.50	77.79	42.84	35.00	48.92	73.68	50.32
+ Reasoning	56.88	65.00	78.40	44.73	36.75	48.05	**75.06**	50.14
+ Retrieval + Multi-evidence	56.50	67.00	77.18	41.45	35.75	**49.46**	74.31	50.37
+ Retrieval + Reasoning	57.48	65.90	78.16	**45.58**	**37.25**	49.35	74.94	51.20
+ Multi-evidence + Reasoning	55.98	65.80	75.97	43.64	36.50	47.51	72.42	50.05
+ **All**	**57.58**	**67.70**	77.91	44.63	**37.25**	**49.46**	74.81	**51.33**

Table 3: Ablation on task categories of our data recipe. Each "$+X$" row trains the base model with only the data of the corresponding category, pairwise union, or the full union ("All"). Shaded cells mark improvements over the base model on each individual benchmark (the Avg. column is left unshaded since all subsets improve the average). Within each column, **bold** marks the best score across the ablation rows.

figuration as in Table 2 except for the ablated component, and train for a fixed 50 steps to ensure a fair comparison. We also apply our recipe to an agent-tuned model and show that the improved long-context reasoning benefits agentic performance as well.

5.1 Ablation Study on Datasets

To measure the contribution of each task category in our data recipe, we train Qwen3-4B-Thinking-2507 on each of the three task categories in isolation, as well as on every pairwise union, and compare them against the full mixture. The results are reported in Table 3. Every single category and pairwise combination improves both the average and most individual benchmarks (marked by shaded cells in Table 3), confirming the effectiveness of our task categorization. Among the three single-category subsets, **Reasoning** yields the largest gain (+3.46 on average), especially on LongBench v2 (+4.13) and LongReason (+5.56), suggesting that training on long-form reasoning tasks is an effective method for long-context RL. The largest pairwise union improvement comes from **Retrieval + Reasoning**, which further improves LongBench v2 and AA-LCR over Reasoning alone. In addition, no single category or pairwise combination matches the full mixture (**All**, 57.58), which indicates that the three task categories in our taxonomy are complementary, and all three are needed for the best result.

5.2 Ablation on Task-balancing Techniques

To assess the contribution of task-balanced sampling and task-level advantage normalization, we ablate both components and fall back to vanilla group-level normalization (**w/o task balancing** in Table 4). Removing task balancing leads to a moderate drop in performance, with the average score decreasing from 57.58 to 56.05 (-1.53). This indicates that multi-task long-context RL still suffers from inter-task competition, and that task balancing is necessary to mitigate this effect. Nevertheless, the ablated variant still delivers consistent improvements over the base model on most benchmarks, suggesting the robustness of our data recipe to the choice of advantage normalization.

5.3 Ablation on Reward Design

A line of long-context RL work supplements the outcome reward with auxiliary signals over the model's thinking trajectory, often relying on additional evidence annotations (Chen et al., 2026a; Guan et al., 2026). Since our training mixture contains no such annotations, we instead test a generic LLM-as-judge process reward that scores whether the thinking trajectory is internally coherent and correctly synthesizes the retrieved information into the final answer. We use DeepSeek-V3.2 as the judge and add this reward on top of our outcome-based reward. As reported in Table 4 (**w/ process reward**), this auxiliary signal does

Configuration	Avg.	Multi-hop QA		Holistic LC Reasoning			Synthetic LC Reasoning	
		LBv1-QA	FRAMES	LBv2	AA-LCR	DocFinQA	LongReason	GraphWalks
Qwen3-4B-Thinking-2507	53.42	62.60	77.55	40.60	35.25	39.80	69.50	48.67
+ **Full recipe**	**57.58**	**67.70**	**77.91**	44.63	**37.25**	**49.46**	**74.81**	**51.33**
w/o task balancing	56.05	64.80	76.58	42.15	35.00	48.59	74.06	51.16
w/ process reward	56.29	65.90	75.97	**44.83**	35.00	48.16	73.55	50.64

Table 4: Ablation on RL design choices, all trained on top of Qwen3-4B-Thinking-2507 with our full data recipe. **Full recipe** reuses the configuration from Table 2. **w/o task balancing** disables both task-balanced sampling and task-level advantage normalization (§3.2). **w/ process reward** adds a thinking-process reward on top of the outcome reward. Within each column, **bold** marks the best score across the three configurations.

Model	GAIA	BrowseComp	Avg. LC
AgentCPM-Explore	66.1	27.00	45.57
+ Ours (25 steps)	67.7	29.00	52.92
+ Ours (50 steps)	**70.9**	**34.00**	**54.64**

Table 5: We train AgentCPM-Explore with our long-context data recipe and observe improvements on GAIA and BrowseComp (Pass@3) and on the long-context average score (**Avg. LC**, over seven benchmarks). **Bold** marks the best score.

not improve over the full recipe: the average score drops from 57.58 to 56.29 (-1.29). We hypothesize two reasons. First, without ground-truth evidence, our LLM-as-judge can only score thinking coherence based on its own priors, yielding a noisier reward signal than evidence-anchored alternatives. Second, given that our diverse data recipe already provides an informative outcome signal, additional process-level supervision offers limited benefit.

5.4 Long-context Reasoning Improves Agentic Performance

Long-horizon agents naturally accumulate massive amounts of context across many turns of tool use, and recent analyses identify long-context capability as a bottleneck on their performance (Li et al., 2026). Motivated by these observations, we conduct a preliminary investigation into whether the skills acquired through our long-context RL benefit long-horizon agentic tasks.

We conduct long-context RL training on AgentCPM-Explore (Chen et al., 2026b), an agent-enhanced model built on top of Qwen3-4B-Thinking-2507 (Qwen Team, 2025a), with our proposed recipe and the same RL configuration as in the main experiments. In addition to the seven long-context benchmarks, we run evaluation on two long-horizon agent benchmarks, GAIA (Mialon et al., 2024) and BrowseComp (Wei et al., 2025), both of which involve web search and frequently require the agent to process lengthy retrieved documents within a single session. To prevent answer leakage (Han et al., 2025), we apply a rigorous filtering protocol; full evaluation details are described in Appendix C.

According to Table 5, our long-context RL data translates into stronger agentic behavior beyond long-context reasoning gains: the 50-step checkpoint improves GAIA Pass@3 by +4.8 and BrowseComp by +7.0 points over AgentCPM-Explore. These preliminary results indicate that strengthening a model's ability to reason over long contexts can yield transferable gains to agentic scenarios.

6 Conclusion

We introduce a simple yet effective data recipe for long-context RL. By constructing a training mixture that contains three complementary tasks, which are retrieval, multi-evidence synthesis, and reasoning, we

achieve consistent improvements across seven long-context reasoning benchmarks. Trained with our data recipe, Qwen3-4B-Thinking-2507, -8B, and -30B-A3B-Thinking-2507 reach average scores of 60.63, 53.10, and 68.23, gaining +7.21, +3.17, and +6.41 over their respective base models. An intriguing finding is that these long-context gains transfer to agentic tasks: continuing RL training of an agent-tuned model on our data improves GAIA by +4.8 and BrowseComp by +7.0 points, indicating that strengthening core long-context abilities is an effective pathway toward stronger agentic capability.

Broader Impact Statement

All training data in our mixture is either synthesized or sourced from publicly available datasets, and our released data will follow the original licenses of the sources.

References

Yushi Bai, Xin Lv, Jiajie Zhang, Yuze He, Ji Qi, Lei Hou, Jie Tang, Yuxiao Dong, and Juanzi Li. LongAlign: A recipe for long context alignment of large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1376–1395, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.74. URL https://aclanthology.org/2024.findings-emnlp.74/.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.172. URL https://aclanthology.org/2024.acl-long.172/.

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. LongBench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3639–3664, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.183. URL https://aclanthology.org/2025.acl-long.183/.

Guanzheng Chen, Michael Qizhe Shieh, and Lidong Bing. Longrlvr: Long-context reinforcement learning requires verifiable context rewards, 2026a. URL https://arxiv.org/abs/2603.02146.

Haotian Chen, Xin Cong, Shengda Fan, Yuyang Fu, Ziqin Gong, Yaxi Lu, Yishan Li, Boye Niu, Chengjun Pan, Zijun Song, Huadong Wang, Yesai Wu, Yueying Wu, Zihao Xie, Yukun Yan, Zhong Zhang, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Agentcpm-explore: Realizing long-horizon deep exploration for edge-scale agents, 2026b. URL https://arxiv.org/abs/2602.06485.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4599–4610, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.365. URL https://aclanthology.org/2021.naacl-main.365/.

Yufeng Du, Minyang Tian, Srikanth Ronanki, Subendhu Rongali, Sravan Babu Bodapati, Aram Galstyan, Azton Wells, Roy Schwartz, Eliu A Huerta, and Hao Peng. Context length alone hurts LLM performance despite perfect retrieval. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23281–23298, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7.

doi: 10.18653/v1/2025.findings-emnlp.1264. URL https://aclanthology.org/2025.findings-emnlp.1264/.

Xin Guan, Zijian Li, Shen Huang, Pengjun Xie, Jingren Zhou, and Jiuxin Cao. Evidence-augmented policy optimization with reward co-evolution for long-context reasoning, 2026. URL https://arxiv.org/abs/2601.10306.

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. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Abhay Gupta, Kevin Zhu, Vasu Sharma, Sean O'Brien, and Michael Lu. NovelHopQA: Diagnosing multi-hop reasoning failures in long narrative contexts. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26134–26151, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1328. URL https://aclanthology.org/2025.emnlp-main.1328/.

Ziwen Han, Meher Mankikar, Julian Michael, and Zifan Wang. Search-time data contamination. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*, 2025. URL https://openreview.net/forum?id=FBoIUyY3wX.

Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=kHB5Te5IWm.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=7Bywt2mQsCe.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=kIoBbc76Sy.

Greg Kamradt. Needle in a haystack – pressure testing LLMs. https://github.com/gkamradt/LLMTest_NeedleInAHaystack, 2023.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4745–4759, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.243. URL https://aclanthology.org/2025.naacl-long.243/.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15339–15353, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.818. URL https://aclanthology.org/2024.acl-long.818/.

Huayang Li, Pat Verga, Priyanka Sen, Bowen Yang, Vijay Viswanathan, Patrick Lewis, Taro Watanabe, and Yixuan Su. Alr^2: A retrieve-then-reason framework for long-context question answering, 2024a. URL https://arxiv.org/abs/2410.03227.

Mo Li, Songyang Zhang, Taolin Zhang, Haodong Duan, Yunxin Liu, and Kai Chen. Needlebench: Evaluating LLM retrieval and reasoning across varying information densities. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=cEvmIKsRw0.

Siheng Li, Cheng Yang, Zesen Cheng, Lemao Liu, Mo Yu, Yujiu Yang, and Wai Lam. Large language models can self-improve in long-context reasoning, 2024b. URL https://arxiv.org/abs/2411.08147.

Xiaochuan Li, Ryan Ming, Pranav Setlur, Abhijay Paladugu, Andy Tang, Hao Kang, Shuai Shao, Rong Jin, and Chenyan Xiong. Benchmark test-time scaling of general llm agents, 2026. URL https://arxiv.org/abs/2602.18998.

Zhan Ling, Kang Liu, Kai Yan, Yifan Yang, Weijian Lin, Ting-Han Fan, Lingfeng Shen, Zhengyin Du, and Jiecao Chen. Longreason: A synthetic long-context reasoning benchmark via context expansion, 2025. URL https://arxiv.org/abs/2501.15089.

Junteng Liu, Yunji Li, Chi Zhang, Jingyang Li, Aili Chen, Ke Ji, Weiyu Cheng, Zijia Wu, Chengyu Du, Qidi Xu, Jiayuan Song, Zhengmao Zhu, Wenhu Chen, Pengyu Zhao, and Junxian He. Webexplorer: Explore and evolve for training long-horizon web agents, 2025. URL https://arxiv.org/abs/2509.06501.

Rui Lu, Zhenyu Hou, Zihan Wang, Hanchen Zhang, Xiao Liu, Yujiang Li, Shi Feng, Jie Tang, and Yuxiao Dong. Deepdive: Advancing deep search agents with knowledge graphs and multi-turn rl, 2025. URL https://arxiv.org/abs/2509.10446.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: a benchmark for general AI assistants. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=fibxvahvs3.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun Yoon, and Hinrich Schuetze. Nolima: Long-context evaluation beyond literal matching. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=0OshX1hiSa.

OpenAI. GraphWalks: a multi hop reasoning long context benchmark. https://huggingface.co/datasets/openai/graphwalks, 2025. Released alongside GPT-4.1 (April 2025); see https://openai.com/index/gpt-4-1/.

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, Bohan Zhang, 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, Wenting Zhan, 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. Openai o1 system card, 2026. URL https://arxiv.org/abs/2412.16720.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=wHBfxhZu1u.

Miao Peng, Weizhou Shen, Nuo Chen, Chenliang Li, Ming Yan, and Jia Li. Incentivizing in-depth reasoning over long contexts with process advantage shaping, 2026. URL https://arxiv.org/abs/2601.12465.

Bowen Ping, Zijun Chen, Tingfeng Hui, Qize Yu, Chenxuan Li, Junchi Yan, and Baobao Chang. Longact: Harnessing intrinsic activation patterns for long-context reinforcement learning, 2026a. URL https://arxiv.org/abs/2604.14922.

Bowen Ping, Zijun Chen, Yiyao Yu, Tingfeng Hui, Junchi Yan, and Baobao Chang. Longr: Unleashing long-context reasoning via reinforcement learning with dense utility rewards, 2026b. URL https://arxiv.org/abs/2602.05758.

Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, Wanjun Zhong, Kuanye Li, Jiale Yang, Yu Miao, Woyu Lin, Longxiang Liu, Xu Jiang, Qianli Ma, Jingyu Li, Xiaojun Xiao, Kai Cai, Chuang Li, Yaowei Zheng, Chaolin Jin, Chen Li, Xiao Zhou, Minchao Wang, Haoli Chen, Zhaojian Li, Haihua Yang, Haifeng Liu, Feng Lin, Tao Peng, Xin Liu, and Guang Shi. Ui-tars: Pioneering automated gui interaction with native agents, 2025. URL https://arxiv.org/abs/2501.12326.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a. URL https://arxiv.org/abs/2505.09388.

Qwen Team. Qwen3-30B-A3B-Thinking-2507 model card. https://huggingface.co/Qwen/Qwen3-30B-A3B-Thinking-2507, 2025b. Accessed: 2026-05-25.

Qwen Team. Qwen3-4B-Thinking-2507 model card. https://huggingface.co/Qwen/Qwen3-4B-Thinking-2507, 2025c. Accessed: 2026-05-25.

Qwen Team. Qwen3-8B model card. https://huggingface.co/Qwen/Qwen3-8B, 2025d. Accessed: 2026-05-25.

RadixArk. Miles. https://github.com/radixark/miles, 2025. Miles is an enterprise-facing reinforcement learning framework for LLM and VLM post-training, forked from and co-evolving with slime.

Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, Michael Krumdick, Charles Lovering, and Chris Tanner. DocFinQA: A long-context financial reasoning dataset. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 445–458, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-short.42. URL https://aclanthology.org/2024.acl-short.42/.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Weizhou Shen, Ziyi Yang, Chenliang Li, Zhiyuan Lu, Miao Peng, Huashan Sun, Yingcheng Shi, Shengyi Liao, Shaopeng Lai, Bo Zhang, Dayiheng Liu, Fei Huang, Jingren Zhou, and Ming Yan. Qwenlong-l1.5: Post-training recipe for long-context reasoning and memory management, 2025. URL https://arxiv.org/abs/2512.12967.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism, 2020. URL https://arxiv.org/abs/1909.08053.

Qisheng Su, Zhen Fang, Shiting Huang, Yu Zeng, Yiming Zhao, Kou Shi, Ziao Zhang, Lin Chen, Zehui Chen, Lijun Wu, and Feng Zhao. Acc: Compiling agent trajectories for long-context training, 2026. URL https://arxiv.org/abs/2605.21850.

Yixuan Tang and Yi Yang. Multihop-RAG: Benchmarking retrieval-augmented generation for multi-hop queries. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=t4eB3zYWBK.

Artificial Analysis Team. Artificial analysis long context reasoning benchmark(lcr), 2025.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

Kiran Vodrahalli, Santiago Ontanon, Nilesh Tripuraneni, Kelvin Xu, Sanil Jain, Rakesh Shivanna, Jeffrey Hui, Nishanth Dikkala, Mehran Kazemi, Bahare Fatemi, Rohan Anil, Ethan Dyer, Siamak Shakeri, Roopali

Vij, Harsh Mehta, Vinay Ramasesh, Quoc Le, Ed Chi, Yifeng Lu, Orhan Firat, Angeliki Lazaridou, Jean-Baptiste Lespiau, Nithya Attaluri, and Kate Olszewska. Michelangelo: Long context evaluations beyond haystacks via latent structure queries, 2024. URL https://arxiv.org/abs/2409.12640.

Fanqi Wan, Weizhou Shen, Shengyi Liao, Yingcheng Shi, Chenliang Li, Ziyi Yang, Ji Zhang, Fei Huang, Jingren Zhou, and Ming Yan. Qwenlong-l1: Towards long-context large reasoning models with reinforcement learning, 2025. URL https://arxiv.org/abs/2505.17667.

Siyuan Wang, Gaokai Zhang, Li Lyna Zhang, Ning Shang, Fan Yang, Dongyao Chen, and Mao Yang. LoongRL: Reinforcement learning for advanced reasoning over long contexts. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=o29E01Q6bv.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. Browsecomp: A simple yet challenging benchmark for browsing agents, 2025. URL https://arxiv.org/abs/2504.12516.

Kyle Whitecross and Negin Rahimi. Recallm: Addressing the lost-in-thought phenomenon with explicit in-context retrieval, 2026. URL https://arxiv.org/abs/2604.09494.

Yanling Xiao, Huaibing Xie, Guoliang Zhao, Shihan Dou, Shaolei Wang, Yiting Liu, Nantao Zheng, Cheng Zhang, Pluto Zhou, Zhisong Zhang, and Lemao Liu. A decomposition perspective to long-context reasoning for llms, 2026. URL https://arxiv.org/abs/2604.07981.

Xiaoyue Xu, Qinyuan Ye, and Xiang Ren. Stress-testing long-context language models with lifelong ICL and task haystack. In *First Workshop on Long-Context Foundation Models @ ICML 2024*, 2024. URL https://openreview.net/forum?id=5ltgEKXnZt.

John Yang, Carlos E Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik R Narasimhan, and Ofir Press. SWE-agent: Agent-computer interfaces enable automated software engineering. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=mXpq6ut8J3.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. HELMET: How to evaluate long-context models effectively and thoroughly. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=293V3bJbmE.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, YuYue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru 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, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=2a36EMSSTp.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. SGLang: Efficient execution of structured language model programs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=VqkAKQibpq.

A Training Details

Table 6 summarizes the RL training configuration. We use Megatron-LM (Shoeybi et al., 2020) as the training backend and SGLang (Zheng et al., 2024) for rollout. All experiments are conducted on H100 GPUs. Training wall-clock time per experiment is ∼22 hours for Qwen3-4B-Thinking-2507 (100 steps, 16 GPUs), ∼50 hours for Qwen3-8B (100 steps, 16 GPUs), and ∼14 hours for Qwen3-30B-A3B-Thinking-2507 (50 steps, 32 GPUs).

During the training, we use an LLM-as-judge to score CrossEntity and LongMath (§3.2). The judge prompt is shown in Figure 4.

Hyperparameter	**Value**
Optimizer	Adam
Learning rate (4B-Thinking, 8B)	1e-6
Learning rate (30B-A3B-Thinking)	2e-6
LR schedule	constant
Global batch size (prompts)	128
PPO batch size (responses)	1,024
Rollouts per prompt (G)	8
Rollout temperature	1.0
Rollout top-p	0.95
Max input length	64K tokens
Max generation length	16K tokens
Training steps (4B-Thinking, 8B)	100
Training steps (30B-A3B-Thinking)	50

Table 6: RL training hyperparameters.

Data length distribution. Figure 3 shows the input length distribution of each dataset in our training mixture.

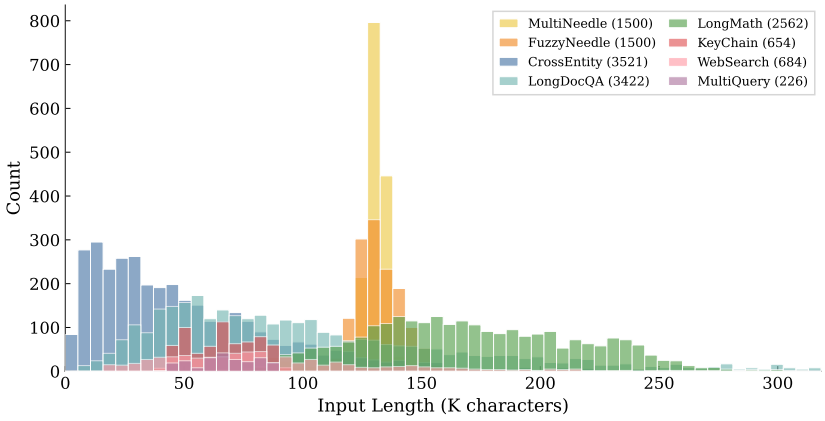

Figure 3: Input length distribution (in characters) of each dataset in our training mixture.

B Evaluation Details.

Evaluation Configurations. We use DeepSeek-V3.2 as the LLM judge throughout all evaluations. For LongBench v1 QA and FRAMES, we report the maximum of the LLM-judge score and substring exact match; the judge is queried with the prompt shown in Figure 5. For DocFinQA and AA-LCR, we similarly use the LLM judge to assess whether the model's prediction matches the gold answer. For LongBench v2, LongReason, and GraphWalks, we apply rule-based answer extraction and comparison.

Training Judge Prompt (CrossEntity & LongMath)

```
You are an intelligent judge who evaluates the correctness of a model's prediction against a standard
answer.

Rules:
1.  Ignore minor formatting differences (e.g., punctuation, case, extra whitespace).
2.  The model's prediction may contain reasoning steps.  Focus only on the final answer or conclusion.
3.  If the prediction matches the standard answer in meaning, it is correct.
4.  If the standard answer is short (e.g., "Yes"), and the prediction is "Yes, because...", it is CORRECT.

Output Format:
If the prediction is correct, output exactly [[1]].
If the prediction is incorrect, output exactly [[0]].
Do not output any other text or explanation.
```

Figure 4: LLM-as-judge prompt used during RL training to score CrossEntity and LongMath.

Eval Judge Prompt: LongBench v1 QA & FRAMES

```
You are a helpful assistant that judges the correctness of a model's prediction against a standard answer.

Question:  {question}
Standard Answer:  {answer}
Model Prediction:  {pred}

Please judge the correctness of the model prediction and output 1 if the answer is correct, 0 if the answer
is incorrect.  Be fair and objective.

Attention:
1.  The Model Prediction may contain a reasoning process (e.g., within <think> tags).  You should focus on
the final answer.
2.  If the reasoning process is incomplete (e.g., cut off) and no final answer is provided, judge it as
incorrect (0).
3.  If a final answer is provided and matches the standard answer (even if the reasoning is cut off or
lengthy), judge it as correct (1).

Output Format:
<score>1 or 0</score>
```

Figure 5: LLM-as-judge prompt used for LongBench v1 QA and FRAMES evaluation.

Eval Judge Prompt: AA-LCR & DocFinQA

```
Please act as an impartial judge and evaluate whether the predicted answer is consistent with the gold
standard answer.

<Question>
{question}
</Question>

<Gold Answer>
{gold}
</Gold Answer>

<Predicted Answer>
{model_prediction}
</Predicted Answer>

Evaluation criteria:
1.  The gold answer is definitely correct.  Do NOT re-answer the question.
2.  For numerical answers, values that are numerically equal or very close are considered correct (e.g.,
0.16 vs 0.16000, 34.8% vs 34.81%, $1.2M vs 1,200,000, -3.22% vs -3.2%).
3.  Different but equivalent representations are acceptable (e.g., 94 vs 94.0, 14% vs 0.14).
4.  For multi-part answers, all parts must match.
5.  If no clear answer can be extracted from the prediction, grade as INCORRECT.

Output ONLY your verdict:
[[A]] if the predicted answer is CORRECT
[[B]] if the predicted answer is INCORRECT
```

Figure 6: LLM-as-judge prompt used for AA-LCR and DocFinQA evaluation.

Model	Length	Recall	RAG	ICL	Cite	Re-rank	LongQA	Summ.	Avg.
Qwen3-4B-Thinking-2507	32k	99.94	67.00	51.96	35.03	69.09	41.15	30.99	56.45
	64k	99.81	69.21	41.20	25.11	49.10	42.90	33.54	51.55
	128k	90.00	62.96	10.44	7.72	36.95	49.08	34.31	41.64
+ Ours	32k	99.75	72.59	50.20	33.32	71.03	44.38	32.32	57.66
	64k	99.25	75.25	45.16	22.24	55.97	42.82	34.00	53.53
	128k	87.19	73.08	15.12	7.67	37.34	49.54	36.45	43.77

Table 7: Detailed HELMET scores of Qwen3-4B-Thinking-2507 and our RL-trained variant at different input lengths.

Model	Avg.	LBv1-QA	FRAMES	LBv2	AA-LCR	DocFinQA	LongReason	GraphWalks
AgentCPM-Explore	45.57	62.00	76.33	29.49	10.25	39.91	60.08	40.91
+ Ours (25 steps)	52.92	68.50	75.00	38.17	31.00	44.03	65.62	**48.13**
+ Ours (50 steps)	**54.64**	**69.30**	**75.85**	**41.35**	**31.50**	**47.61**	**70.28**	46.58

Table 8: Long-context benchmark results for AgentCPM-Explore and its RL-trained variants. **Bold** denotes the best score per column among the RL-trained variants.

Detailed scores on HELMET. We report detailed HELMET scores at different input lengths in Table 7.

C Agent Evaluation Details

C.1 Agent Evaluation Configurations

We adopt the same evaluation framework (AgentToLeaP) as AgentCPM-Explore (Chen et al., 2026b). The agent is equipped with three tools: (1) a web search tool backed by Google Serp API, (2) a URL fetching tool with a DeepSeek-V3.2 browser processor that summarizes retrieved web content, and (3) a sandboxed Python code executor. The inference engine (SGLang) is configured with a 240k-token context window. The maximum output tokens per interaction is set to 16,384 for both GAIA and BrowseComp, with a maximum of 200 interaction rounds per task.

For GAIA, we evaluate on the text-only subset of the validation set, totaling 127 examples. For BrowseComp, we evaluate on a 100-example subset sampled from the full test set due to the time cost of multi-round agentic interactions.

Leakage filtering. It is a known issue that GAIA test cases are frequently leaked online (Han et al., 2025), often appearing in blog posts or papers that use GAIA as an illustrative example. To prevent search tools from directly retrieving leaked answers during evaluation, we implement a real-time leakage filter: when a search or fetch results simultaneously contains a substring matching the original benchmark question and the gold answer verbatim, we filter it out and do not provide it to the agent.

C.2 Additional Benchmark Details

Model	GAIA	BrowseComp
AgentCPM-Explore	49.6	17.3
+ Ours (25 steps)	50.7	17.8
+ Ours (50 steps)	**54.9**	**19.7**

Table 9: Average Pass@3 on GAIA and BrowseComp.

Per-benchmark long-context scores. We report the per-benchmark long-context scores of AgentCPM-Explore and its RL-trained variants in Table 8, complementing the averaged **Avg. LC** column of Table 5.

Model	L1 (n=42)	L2 (n=66)	L3 (n=19)
AgentCPM-Explore	78.6	65.2	42.1
+ Ours (25 steps)	81.0	63.6	**52.6**
+ Ours (50 steps)	**85.7**	**66.7**	**52.6**

Table 10: GAIA Pass@3 broken down by official difficulty level.

Model	Easy *(n=14)*	Med *(n=9)*	Hard *(n=21)*	VHard *(n=56)*
AgentCPM-Explore	100.0	77.8	28.6	0.0
+ Ours (25 steps)	100.0	77.8	38.1	0.0
+ Ours (50 steps)	100.0	**88.9**	**57.1**	0.0

Table 11: BrowseComp Pass@3 by consensus difficulty. Difficulty is determined by the number of correct answers across 3 models $\times$ 3 rounds = 9 trials: easy (6–9), med (3–5), hard (1–2), vhard (0).

Performance by question difficulty. We further stratify Pass@3 by question difficulty. For GAIA we use the official L1–L3 levels; for BrowseComp, which provides no official labels, we apply a consensus-based bucketing across 3 models $\times$ 3 rounds = 9 trials per question, labelling questions as *easy* (6–9 correct), *med* (3–5), *hard* (1–2), and *vhard* (0). Results are reported in Tables 10 and 11. The gains of our 50-step checkpoint are most pronounced on the harder subsets—42.1 $\rightarrow$ 52.6 on GAIA L3 and 28.6 $\rightarrow$ 57.1 on BrowseComp Hard.

D Dataset Licenses

Table 12 lists the licenses of all external datasets used in this work. Our training recipe uses only the official training splits of these datasets. Remaining datasets in our mixture are constructed by us and will be released under a permissive license. The LoongRL KeyChain training data (Wang et al., 2026) does not specify an explicit license in its repository, though its underlying sources are all under permissive licenses.

Dataset	**Usage**	**License**
HotpotQA	LongDocQA	CC-BY-SA-4.0
MuSiQue	LongDocQA	CC-BY-4.0
Qasper	LongDocQA	CC-BY-4.0
DeepMath-103K	LongMath	MIT
MATH-Hard	LongMath	MIT
DocQA-RL-1.6K	Baseline	Apache-2.0
LoongRL KeyChain	Baseline	Not specified

Table 12: Licenses of external datasets.