Enhancing Long-Context Reasoning via Verifiable Atomic Skills

Anonymous ACL submission

Abstract

Long-context reasoning remains a critical challenge for Large Language Models (LLMs). Current research typically curates specialized training data by treating long-context reasoning as a monolithic task, yet it is challenging to verify its quality due to its internal cognitive complexity. In this paper, we propose a paradigm shift: viewing long-context reasoning as a hierarchical spectrum of core atomic skills. To operationalize this, we present an Anchor-based Reasoning (AbR) pipeline that automatically synthesizes strictly verifiable training data for each skill, thereby enabling reliable quality. Diagnostic analyses confirm that proficiency in these atomic skills strongly predicts general long-context reasoning performance. By training exclusively on a highly compact set of merely four thousand synthesized samples with Reinforcement Learning (RL), we explicitly sharpen these foundational capabilities. Extensive experiments across six challenging benchmarks demonstrate that this targeted RL causally delivers robust and generalizable performance gains, outperforming DeepSeek-R1-distill-32B by an average margin of 7.7% (improving from 46.3% to 54.0%). Furthermore, in-depth ablations reveal that these skills operate synergistically, establishing a scalable, data-efficient, and verifiable path to advancing long-context reasoning.

1 Introduction

The rapid evolution of Large Language Models (LLMs) (Guo et al., 2025; Liu et al., 2025; Comanici et al., 2025) has made the ability to process extensive context windows increasingly critical. Real-world applications, such as analyzing multi-document repositories, synthesizing legal contracts, and processing financial reports (Meyur et al., 2025; Reddy et al., 2024), demand that models not only "comprehend" massive texts but also reason over them effectively. However, although

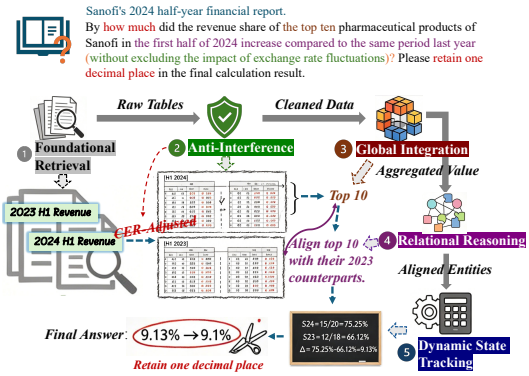

Figure 1: Solving a complex financial query (LongBench-v2) into atomic skills. The solution relies on a five-step synergistic pipeline: locating reports (*Foundational Retrieval*), filtering out exchange-rate distractors (*Anti-Interference*), aggregating data for the 2024 Top 10 products (*Global Integration*), aligning these targets with 2023 figures (*Relational Reasoning*), and computing the final numerical share (*Dynamic State Tracking*).

recent advancements have expanded context windows to up to 1 million tokens (Team et al., 2024; GLM et al., 2024), a pronounced chasm persists between the scale of context models can accommodate and the efficacy of reasoning they deliver (Paulsen, 2025).

To bridge this critical gap, conventional approaches typically entail curating specialized training datasets tailored for complex long-context reasoning tasks (Chen et al., 2023; Zhang et al., 2025; Bai et al., 2024), followed by fine-tuning (Li et al., 2024a) or reinforcement learning (RL) (Wan et al., 2025; Wang et al., 2025) to boost model performance. Nevertheless, long-context reasoning is conventionally treated as a monolithic, indivisible task. Directly curating high-quality data for such complex tasks is fraught with challenges (Yang et al., 2025b), including latent knowledge conflicts (Xu et al., 2024) and severe hallucinations. Crucially, this monolithic complexity makes it diffi-

cult to provide strict and verifiable supervision for RL(Li et al., 2024c).

In this paper, we propose a paradigm shift: rather than treating long-context reasoning as a single indivisible task, we investigate whether it can be systematically improved by explicitly enhancing a core set of *atomic skills*. Inspired by classic theories on discourse comprehension and multi-step problem solving (Kintsch, 1991; Simon and Newell, 1971), we posit that resolving complex queries over extensive contexts is inherently a multi-stage process. As illustrated by the real-world financial analysis case in Figure 1, solving complex queries fundamentally decomposes into simpler, recurring capabilities. We abstract these requirements into five intuitive atomic skills ubiquitous in long-context workloads: *Foundational Retrieval*, *Anti-Interference* , *Global Integration*, *Relational Reasoning*, and *Dynamic State Tracking* (§2.1). Crucially, unlike inherently complex real-world tasks, each atomic skill is conceptually straightforward in isolation. This insight motivates our Anchor-based Reasoning (AbR) pipeline, which automatically synthesizes skill-specific training data equipped with algorithmically generated and strictly verifiable ground-truth answers (§2.2).

We conduct intensive experiments to demonstrate the effectiveness of the AbR framework. We utilize these atomic skills as both diagnostic probes and training targets and empirical analysis reveals that performance on these probes closely aligns with standard long-context benchmarks, validating them as accurate proxies (§3.1). Building upon this, we enhance these skills via RL (Shao et al., 2024; Yu et al., 2025) using merely 4,000 synthetic AbR samples. Extensive experiments across six challenging benchmarks demonstrate consistent and significant gains. Notably, DeepSeek-R1-distill-32B improves from 46.3% to 54.0% on average (§3.2). Further in-depth ablations reveal the underlying mechanics: these atomic skills function synergistically and are hierarchically coupled, confirming that explicit atomic enhancement scales effectively to general long-context reasoning (§3.3).

Our contributions are summarized as follows:

- **Atomic Skills for Long-context Reasoning:** We identify five intuitive atomic skills that frequently arise in complex long-context tasks and statistically validate them as highly accurate diagnostic proxies for real-world long-context capabilities.

- **Verifiable AbR Data Construction:** We introduce the Anchor-based Reasoning (AbR) pipeline, an automated and scalable approach to synthesize strictly verifiable training data tailored to each atomic skill, effectively mitigating the challenges of monolithic long-context data curation and RL supervision.

- **Highly Efficient Enhancement & Synergy Insights:** We demonstrate that applying RL on merely 4,000 AbR samples yields substantial generalization gains across leading long-context benchmarks. Furthermore, our extensive ablations reveal the internal hierarchical dependency of these skills, offering new insights into how LLMs process extended contexts.

2 Atomic Skills and Anchor-based Reasoning

As motivated in the introduction, solving complex real-world tasks (e.g., Figure 1: financial analysis over lengthy reports) requires a cluster of sub-skills, from locating scattered figures to performing multi-step calculations. We view long-context reasoning as a hierarchy of cognitive skills rather than a single capability. Based on this view, we identify five common skills under two main design constraints:

- **(C1) Pipeline Coverage**: Collectively, the skills should cover the fundamental pipeline that any robust model must execute in complex scenarios. (e.g., retrieve→filter noise→integrate→deduce relations→compute).

- **(C2) Verifiable supervision**: Each skill must admit an algorithmic, strictly verifiable answer (e.g., numeric values, symbolic expressions) to enable exact-match rewards for RL, avoiding subjective evaluation flaws.

2.1 Atomic Skill Taxonomy

Ordered by increasing cognitive complexity and directly mapping to the fundamental pipeline (C1), our five targeted atomic skills are defined as follows. To ensure consistency with our subsequent experimental analyses (Figures 3, 5, and 6), we provide the corresponding shorthand notations in parentheses.

Foundational Retrieval(*NIAH*): The Prerequisite Before any complex reasoning can occur, a model must first prove it can reliably locate isolated facts ("the needle") anywhere within a vast sea of text ("the haystack")(NIAH), overcoming the well-documented "lost-in-the-middle" phenomenon. This is the bedrock of all long-context capabilities.

Anti-Interference (*Anti-interfere*): Robustness to Noise Building on simple retrieval, Robustness to Noise addresses a more realistic challenge. It's not enough to just find information; a model must distinguish the correct target from similar-looking but incorrect "distractors". This skill measures the ability to maintain focus and factual accuracy when faced with deceptive or confusing information.

Global Integration(*Multi-source*): Multi-Source Synthesis Moving beyond localized extraction, this skill requires the model to break free from local attention traps. It must actively scan across document boundaries to identify, extract, and synthesize multiple scattered data points, proving its capability to process and unify parallel streams of information.

Relational Reasoning (*Logic*): Structure Understanding and Logic The next level of complexity, Relational Reasoning, requires more than just gathering facts; it demands an understanding of the logical relationships between them. A model must recognize the text's underlying structure to perform operations like filtering, joining, or comparing different sets of information, much like executing a database query on unstructured text.

Dynamic State Tracking (*Calc-reason*): Long-Range Computational Reasoning At the peak of this hierarchy, Dynamic State Tracking tests a model's ability to perform multi-step computational reasoning. Here, a model must not only retrieve and relate information but also use it to perform intermediate calculations. It must derive new values from the text, hold these "states" in its working memory, and then execute a final computation using these derived results, completing a full "retrieve-solve-then-compute" workflow.

2.2 Automated Dataset Construction Pipeline

To instantiate this conceptual taxonomy into scalable and strictly verifiable training data, we introduce the **Anchor-based Reasoning(AbR)** framework. The core idea of AbR is to embed algorithmically generated anchors—unique strings paired with specific, verifiable questions—into extensive, noise-laden documents. By strategically distributing these "anchor-question" pairs, AbR enables us to construct meta-questions that require a model to locate, disambiguate, aggregate, or reason over the answers to the embedded queries. In this way, AbR turns abstract atomic skills into concrete training and evaluation instances with controllable structure and verifiable ground truth. As shown in figure 2, we establish a three-stage automated pipeline to construct scalable, controllable, and verifiable datasets.

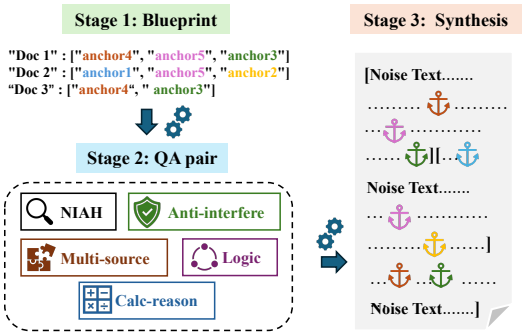

Figure 2: The Automated Dataset Construction Pipeline of the Anchor-based Reasoning (AbR) Framework.

Stage 1: Logical Blueprint Generation. We first programmatically generate a `doc_mapping` JSON file. This blueprint defines the "ground truth" by specifying which anchors (`anc_id`) are contained within each virtual document (`doc_id`).

```
// 5 anchors across 3 docs.
{
  "doc1": ["anc4", "anc5", "anc3"],
  "doc2": ["anc1", "anc5", "anc2"],
  "doc3": ["anc4", "anc3"]
}
```

Stage 2: Question-Answer Pair Generation. Based on `doc_mapping`, we then generate diverse meta-questions requiring complex reasoning via rule-based templates or LLMs. The answers are generated as an *executable expression*, representing the procedural steps needed to arrive at the correct solution. $Solve(Q(anc))$ denotes solving the question associated with a specific anchor.

- *Question (Relational):* "In the document that contains both 'anchor5' and 'anchor3', what is the answer to the question associated with 'anchor4'?"

- solve $(Q\,(\text{anc}_4, d \in (\text{docs}(\text{anc}_5) \cap \text{docs}(\text{anc}_3))))$
- *Question (Computational):* "Calculate the sum of the answers to the questions for 'anchor3' in all documents that contain it."
- $\sum[\text{solve}(Q(\text{doc}_1, \text{anc}_3)), \text{solve}(Q(\text{doc}_3, \text{anc}_3))]$

Stage 3: Multi-Document Context Synthesis. We assemble the final sample by: 1) selecting unrelated background texts for each `doc_id`; 2) inserting anchor-question pairs at random positions based on the `doc_mapping`; and 3) pairing the meta-question with the full synthesized context.

Skill-Specific Task Construction We tailor AbR tasks for each atomic skill: (1) *Foundational Retrieval* inserts specific pairs requiring the model to locate distributed anchors for objective answers. (2) *Robustness to Noise* employs two interference patterns: *Similarity Discrimination* uses highly similar anchors for fine-grained distinction, while *Conflicting Information Resolution* distributes identical anchors to enforce contradiction resolution. (3) *Global Integration* fragments clues across separate documents, compelling the model to aggregate dispersed data points into coherent logical chains. (4) *Relational Reasoning* imposes logical constraints on structural positions, requiring set operations (e.g., intersection, union) on document locations to identify targets. (5) *Dynamic State Tracking* necessitates a multi-stage process deriving numerical values from distributed anchors to execute sequential mathematical operations. Detailed showcases for each skill are provided in the Appendix B.

Controllable Difficulty A key advantage of our methodology is the precise control of training difficulty across a continuous complexity spectrum. By systematically tuning parameters such as context length, anchor density, noise similarity, and reasoning depth, we establish a controlled environment for generating diverse challenges. This granularity facilitates the seamless blending of different difficulty levels during training, while also laying a solid groundwork for future research into long-context curriculum learning.

2.3 RL Training Protocols

To translate these atomic proxies into capability gains, we perform RL on a compact 4000 sample AbR dataset of verifiable atomic skill instances. We post-train base reasoning models via Group Relative Policy Optimization (GRPO) (Shao et al., 2024) augmented with Dynamic Sampling (Yu et al., 2025) for efficient policy optimization. Rewards are binary, computed by `gpt-oss-120B` (Agarwal et al., 2025) judging whether outputs match ground-truth answers. Notably, `gpt-oss-120B` underperforms our base models on long-context reasoning(Table 1, §3.3), so any RL gains stem from the AbR dataset rather than the checker's capabilities.

3 Empirical Evidence for Atomic Skill Enhancement

Our empirical evaluation has three core goals: first, validate that our atomic skills reliably proxy real-world long-context capabilities; second, demonstrate the effectiveness of enhancing these skills via RL on standard benchmarks; third, conduct ablations to show these skills function synergistically within a hierarchical cognitive structure.

3.1 Experimental Setup

Diagnostic Validation Setup To verify that our atomic skills serve as effective proxies for general long-context capability, we compute Spearman rank correlations (ρ) between performance on our atomic probes and six real-world benchmarks. We evaluate 11 diverse open-source models (7B to 32B), using context lengths up to 128K tokens for consistency. Besides, we further analyze absolute skill scores to characterize the strengths and weaknesses of current models.

RL Training Configuration For the RL enhancement phase, we utilize three backbone models: `Qwen2.5-14B-Instruct`, `Qwen2.5-32B-Instruct` (Alibaba, 2024), and `DeepSeek-R1-Distill-32B` (DeepSeek, 2025a). All models are cold-started in the RL phase without prior long-context fine-tuning. To ensure optimal training difficulty, the dataset is filtered using DeepSeek-V3.1 (DeepSeek, 2025b) (retaining pass rates between 0.3 and 0.6) and blended across skills. Exhaustive details regarding hyperparameter configurations, data mixture ratios, and reward mechanism are provided in Appendix C.

Evaluation Framework and Baselines We assess the final models on a comprehensive suite of benchmarks: `LongBench-v2` (Bai et al., 2025), `Loong` (Wang et al., 2024), `MRCR` (OpenAI, 2025c), `BrowsCompLong` (OpenAI, 2025a),

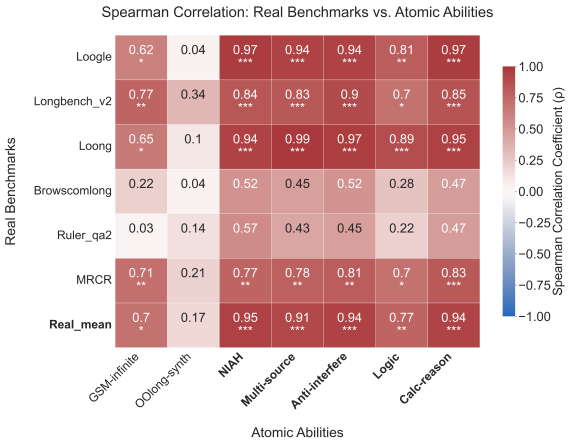

Figure 3: **Spearman Correlation Analysis.** The heatmap compares the correlation of our proposed atomic capabilities against real-world long-context benchmarks.

the `qa_2` subset from `Ruler` (Hsieh et al., 2024), and the real-prompt subset of `Loogle` (Li et al., 2024b). We compare against leading closed-source (`Gemini-3-Pro`) and open-source models (`Kimi-K2-Thinking` (Team et al., 2025), `DeepSeek-V3.1`, `Qwen3-235B` (Yang et al., 2025a)). To isolate the impact of our data, we also construct direct baselines by training our primary backbone on existing synthetic datasets such as LongReason (Ling et al., 2025), QwenDocqa (Wan et al., 2025), and LoongRl (Wang et al., 2025). Detailed configurations are in Appendix D.

3.2 Diagnostic Validation: Are Atomic Skills Good Proxies?

As shown in Figure 3, the results provide strong statistical evidence (ρ-values) validating our core premise: performance on these atomic probes tightly mirrors real-world long-context capabilities.

High Predictive Validity. Our atomic probes demonstrate exceptional predictive power regarding the average performance on real-world benchmarks (*Real_mean*). For instance, our *NIAH* and *Anti-interfere* probes achieve exceptional alignment with *Real_mean* ($\rho = 0.95$ and $\rho = 0.94$, respectively). This superiority is particularly evident in challenging benchmarks like *Loong*, where *Multi-source* reaches a correlation of $\rho = 0.99$. These consistently high correlations (all significant at $p < 0.001$) validate our taxonomy: rather than being an arbitrary collection of tasks, these probes serve as accurate "proxies" that effectively decompose the complexity of long-context reasoning into measurable atomic units.

Beyond Generic Reasoning Baselines. While *GSM-infinite* (Zhou et al., 2025) demonstrates a moderate correlation ($\rho = 0.70$), indicating it is an important indicator of general multi-step reasoning, it consistently lags behind our *Calc-reason* ($\rho = 0.94$). This suggests that our probes capture specific context-handling and integration mechanisms beyond standard mathematical reasoning, making them complementary rather than contradictory.

Diagnosing the capability gap. We further observe an "importance-proficiency" mismatch: skills highly correlated with real-world performance, such as *Anti-interfere* and *Multi-source*, remain weak in current models. In contrast, clean *NIAH* is relatively better solved, suggesting diminishing returns from further improving simple retrieval. Detailed analysis is provided in Appendix E.

3.3 Main Results: Elevating Long-Context Reasoning

Having established the validity of our atomic proxies and identified critical capability gaps in current models, we evaluate the real-world performance of models enhanced via RL on our AbR-generated dataset.

Outperforming General Baselines. Table 1 presents the comprehensive evaluation results across six challenging long-context benchmarks. We can see that our approach yields an absolute gain of 7.7% in average over the backbone `DeepSeek-R1-distill-32B`. Our explicitly enhanced model demonstrates highly competitive performance against state-of-the-art models, outperforming the remarkable `Kimi-K2-Thinking`, `DeepSeek-V3.1`, and `Qwen3-235B` by a large margin. This establishes the effectiveness of atomic skill enhancement in pushing the upper bounds of long-context reasoning. Furthermore, rather than compromising the model's inherent capabilities, our approach actively enhances them: evaluations on standard short-context multi-hop benchmarks (detailed in Appendix F) reveal an 8.8% average absolute gain over the base model.

Superiority over Existing Long-context Training Data. To rigorously isolate the impact of our data construction, we compare against direct baselines fine-tuned on the same `DeepSeek-R1-distill-32B` backbone. As

Table 1: Performance comparison on Open Long Context Benchmarks. The top block shows general baselines and the bottom block illustrates the direct baselines with the same backbone.

Model	Loogle	Longbench-v2	Loong	Browscomplong	Ruler-qa2	MRCR	Average
Gemini-3-pro	52.86%	**69.38%**	**65.43%**	**88.07%**	**83.01%**	**75.30%**	**72.34%**
Kimi-K2-Thinking	51.50%	49.30%	58.01%	58.10%	49.98%	51.77%	53.11%
DeepSeek-V3.1	**55.77%**	52.88%	50.55%	56.27%	42.97%	46.62%	50.84%
Qwen3-235B-A22B-thinking-2507	52.77%	49.30%	53.70%	50.76%	46.81%	44.61%	49.66%
QwenLong-L1-32B	49.32%	43.74%	44.68%	69.93%	47.70%	27.70%	47.18%
GPT-OSS-120B	40.15%	15.31%	22.83%	71.87%	39.92%	22.83%	35.49%
DeepSeek-R1-distill-32B	42.31%	43.94%	38.17%	64.22%	57.23%	31.94%	46.30%
DeepSeek-R1-distill-32B+LongReason	43.23%	44.14%	38.57%	57.90%	58.36%	32.85%	45.84%
DeepSeek-R1-distill-32B+QwenDocqa	47.96%	46.32%	39.98%	71.56%	67.45%	34.59%	51.31%
DeepSeek-R1-distill-32B+LoongRL	48.41%	45.73%	40.29%	70.74%	61.84%	**37.23%**	50.71%
DeepSeek-R1-distill-32B+Ours	50.59%	49.70%	**44.68%**	**73.09%**	**69.38%**	36.74%	54.03%
DeepSeek-R1-distill-32B+Ours+LoongRL	**55.59%**	**51.29%**	44.45%	72.27%	67.01%	36.12%	**54.46%**

Table 2: Performance of our approach applied to more backbone models. The best result in each column with its block is highlighted in **bold**.

Model	Loogle	Longbench-v2	Loong	Browscomplong	Ruler-qa2	MRCR	Average
Qwen2.5-14B-instruct	33.06%	36.38%	24.65%	47.30%	40.28%	31.87%	35.59%
Qwen2.5-14B-instruct+Ours	**40.51%**	**41.15%**	**28.63%**	**69.72%**	**63.21%**	31.73%	**45.83%**
Qwen2.5-32B-instruct	35.97%	39.76%	34.27%	60.65%	45.25%	33.33%	41.54%
Qwen2.5-32B-instruct+Ours	**45.78%**	**43.94%**	**38.34%**	**60.86%**	**47.74%**	**36.33%**	**45.50%**

presented in Table 1, while existing strategies like QwenDocQA and LoongRL effectively improve the baseline (raising the average score from 46.30% to 51.31% and 50.71%, respectively), our method achieves a significantly higher average score of **54.03%**. This significant margin indicates that our construction strategy captures critical long-context dependencies more effectively than previous approaches.

Scalability Across Model Families. Crucially, this performance superiority is not limited to a single model family. As detailed in Table 2, applying our approach to the Qwen2.5-14B and 32B models yields substantial absolute gains (+10.24% and +3.9% respectively), demonstrating broad generalization across model scales.

Synergistic Effect and Extensibility. Experimental results demonstrate a strong synergistic effect when combining our method with 1,000 LoongRL samples, yielding the highest overall performance of **54.46%**. This "stacking" effect suggests that our atomic skill enhancement and existing datasets address different, complementary aspects of long-context capabilities. Furthermore, our proposed AbR framework is inherently extensible: other unexplored atomic skills can be seamlessly defined, synthesized, and optimized using the exact same methodology to continually push the boundaries of long-context reasoning.

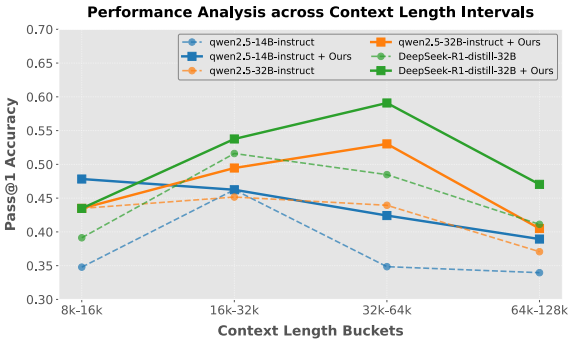

Figure 4: **Performance Comparison across Context Length Intervals on LongBench-v2.** The Pass@1 accuracy of baseline models (dashed lines) versus our method (solid lines) across different length buckets.

Length-Invariant Robustness To investigate the robustness of our method under varying input lengths, we analyze performance on LongBench-v2 across four distinct intervals ranging from 8k to 128k. As illustrated in Figure 4, our method consistently shifts the performance curve upward, maintaining superiority regardless of context length. The results confirm that our approach effectively mitigates the performance degradation typically observed in extended contexts. Specifically, for the `DeepSeek-R1-distill-32B` model, we achieve a substantial gain in the 32k-64k interval, boost-

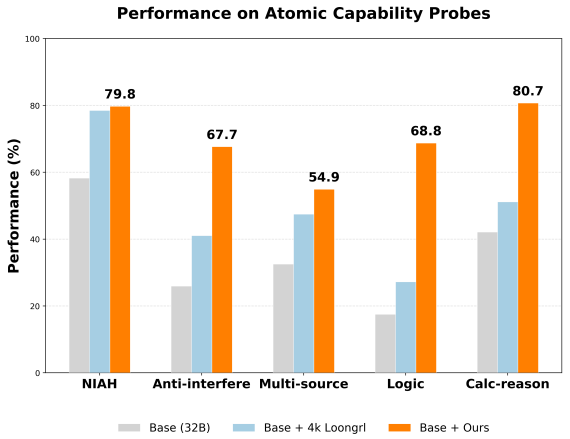

Figure 5: Performance comparison on Atomic Capability Probes. We compare the DeepSeek-R1-distill-32B base model (Grey), the model trained with 4k LoongRL (Blue), and our proposed method (Orange).

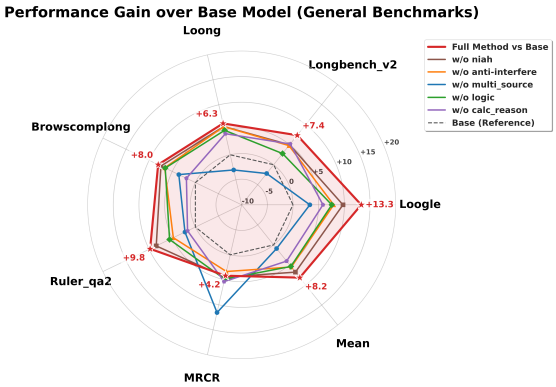

Figure 6: **Performance Gain over Base Model.** The radar chart compares the performance improvements of our full method (red, with stars) against various ablation variants across six real-world long-context benchmarks.

ing accuracy from 48.48% to **59.09%**. Crucially, this advantage persists even in the challenging 64k-128k bucket, demonstrating that our method sustains high-quality reasoning capabilities across the entire long-context spectrum without suffering from significant information loss.

3.4 In-depth Analyses: Why Atomic Skill Enhancement Works?

Why does explicitly targeting isolated atomic skills lead to such significant gains on complex tasks? In this section, we dissect the mechanics of our framework, proving that our method directly enhances targeted skills, relies on their synergy, and reflects a fundamental cognitive hierarchy.

3.4.1 Direct Atomic Enhancement: Beyond Simple Retrieval

Before assessing general generalization, we first verify whether our RL training directly enhances the targeted atomic skills. As shown in Figure 5, our method yields substantial improvements over the Base model across all dimensions, particularly in tasks requiring deep cognitive processing. Notably, *Logic* surges from $\sim$ 18% to **68.8%**, and *Calc_reason* nearly doubles to **80.7%**. Compared with LoongRL, our method achieves similar performance on simple retrieval (*NIAH*: $\sim$ 78% vs. **79.8%**), but substantially outperforms it on harder skills such as *Anti-interfere* (+26.7%) and *Logic* (+41.8%), showing that AbR pushes models beyond the retrieval ceiling toward robust discrimination and structured reasoning.

3.4.2 Ablation Study: Are All Atomic Skills Necessary?

If the model has acquired these skills, are they all necessary for real-world tasks? To answer this, we ablated specific components from our training mixture.

Full Skill Mixture Performs Best. The **Full Method** (red line) consistently yields the highest improvements, completely enveloping all other ablated variants in Figure 6. It achieves remarkable gains across diverse benchmarks, such as **+13.3** on *Loogle* and **+9.8** on *Ruler-qa2*. This comprehensive superiority demonstrates that the synergy of all atomic capabilities is essential for maximizing robust long-context performance.

Global Integration: The Foundational Stabilizer. Removing **Multi-source** data (blue line) reveals a critical phenomenon: performance on general benchmarks like *Loong* and *LongBench-v2* drops below the Base model, despite remaining positive on MRCR. This suggests that Multi-source data acts as a foundational stabilizer; without it, the model over-specializes in single-document retrieval (yielding the highest MRCR score), creating a skill imbalance that degrades its fundamental ability to process complex, general long contexts.

Relational Reasoning: The Structural Backbone. The removal of **Logic** (green line) leads to substantial performance degradation on *Browscomplong* and *Ruler_qa2*. The wide gap between the green and red lines on these axes suggests that tasks involving long-document browsing or complex QA rules rely heavily on the model's logical structure and reasoning chain, rather than simple retrieval.

Dynamic State Tracking: The Universal Precision Enhancer. The **Calc_reason** capability (purple line) proves to be a global performance enhancer, not limited to the numerical tasks in *Loogle*. The consistent drops across general benchmarks like *Ruler_qa2* and *Loong* indicate that training on calculation data instills a rigorous reasoning mindset, improving the model's generalized ability to track complex dependencies and maintain precision over long contexts.

3.4.3 Skill Interaction: Distinct but Hierarchically Coupled

Having established that all skills are necessary for external real-world tasks, we now investigate how these skills interact internally. We cross-analyzed the performance drops on individual atomic probes when specific training components were ablated (Figure 7). The results perfectly validate that long-context capability is not a flat list of tasks, but a hierarchical cognitive spectrum.

Diagonal Dominance Indicates Distinctness. The heatmap exhibits a strong diagonal pattern, particularly for **Logic** (-29) and **Anti-interfere** (-21.1). This confirms that "Logical Structuring" and "Robust Discrimination" are specialized skills requiring dedicated training, as they cannot be implicitly learned solely through simple retrieval tasks.

Off-diagonal Drops Reveal Hierarchical Dependencies. The off-diagonal values reveal a clear cognitive hierarchy. We observe an asymmetric dependency where removing **Logic** significantly impairs *Calc_reason* (-12.3), whereas removing *Calc_reason* has a much smaller impact on *Logic* (-6.0). This supports the hypothesis that dynamic state manipulation relies on underlying logical structuring. Furthermore, the removal of **Multi_source** causes consistent degradation across all atomic capabilities (e.g., -10.6 on *Anti-interfere*, -10.5 on *Calc_reason*). This corroborates the "skill imbalance" observed in 3.4.2, confirming that *Global Integration* acts as a foundational stabilizer, which is essential for maintaining the general distribution alignment required to support specialized cognitive skills.

4 Related Works

Enhancing long-context reasoning in LLMs has attracted extensive efforts (Chen et al., 2023; Li et al., 2024d; Bai et al., 2024). Conventional paradigms typically curate task-specific datasets (Chen et al., 2023; Bai et al., 2024) to optimize models via fine-tuning (Li et al., 2024a; Zhang et al., 2025) or reinforcement learning (Wan et al., 2025; Shen et al., 2025; Wang et al., 2025). However, treating long-context reasoning as a monolithic task complicates high-quality data construction (Yang et al., 2025b), often introducing inadequate verification (Li et al., 2024c) and latent knowledge conflicts (Xu et al., 2024). To address these limitations, our AbR framework shifts the paradigm from holistic task-matching to atomic skill decomposition. This approach enables the automatic synthesis of strictly verifiable training data, effectively overcoming the quality and scalability bottlenecks of conventional monolithic approaches.

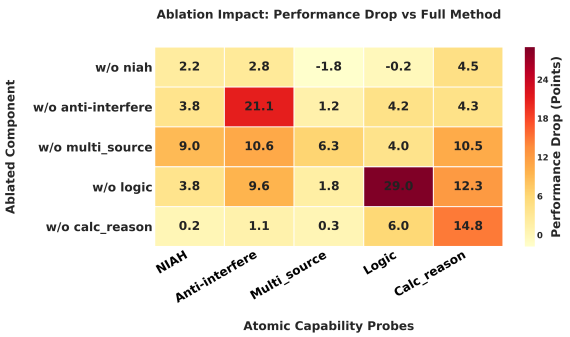

Figure 7: **Skill Interaction Analysis.** The heatmap shows performance drops on atomic capability probes after removing each skill-specific training component, revealing both skill distinctness and cross-skill dependencies.

5 Conclusions

In this paper, we propose to enhance long-context reasoning by improving a hierarchical spectrum of five atomic skills. We present the Anchor-based Reasoning (AbR) framework, which automatically synthesizes verifiable training data for each atomic skill. Then, we employ the reinforcement learning to post-train a base model on the obtained AbR dataset. Our empirical results demonstrate that explicitly targeting these foundational skills is highly effective: applying RL on a compact dataset of merely 4,000 AbR samples yields substantial and consistent improvements across six challenging real-world benchmarks. Furthermore, our in-depth ablations reveal that these atomic skills operate synergistically within a hierarchical cognitive structure rather than in isolation. Ultimately, this work provides a scalable, data-efficient, and strictly verifiable path to advancing long-context capabilities in LLMs.

Limitations

While our atomic-skill enhancement paradigm demonstrates significant improvements in long-context reasoning, it also opens several exciting avenues for future exploration. Because our current RL framework relies on strictly verifiable exact-match rewards, we intentionally focused on deterministic extraction and QA tasks; future research should extend this approach to open-ended generative tasks, such as repository-level code generation and complex summarization, by developing robust evaluation metrics for long-form faithfulness. Additionally, our proposed five-skill taxonomy serves as a flexible scaffold rather than an exhaustive cognitive model, inviting further expansion into dimensions like temporal reasoning and cross-tabular planning. To further maximize training efficiency, future studies could also transition from our empirically derived static data mixtures to dynamic curriculum learning during RL.

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Alibaba. 2024. Qwen 2.5 technical report. https://arxiv.org/abs/2409.13586. *arXiv preprint*.

Yushi Bai, Xin Lv, Jiajie Zhang, Yuze He, Ji Qi, Lei Hou, Jie Tang, Yuxiao Dong, and Juanzi Li. 2024. Longalign: A recipe for long context alignment of large language models. *arXiv preprint arXiv:2401.18058*.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, and 1 others. 2025. 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.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2023. Longlora: Efficient fine-tuning of long-context large language models. *arXiv preprint arXiv:2309.12307*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

DeepSeek. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning (distillation supplement). https://arxiv.org/abs/2501.12948. *arXiv preprint*.

DeepSeek. 2025b. Deepseek-v3 technical report (v3.1 update). https://stardust108.github.io/DeepSeek-V3/DeepSeek_V3.pdf.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*.

Walter Kintsch. 1991. The role of knowledge in discourse comprehension: A construction-integration model. In G.E. Stelmach and P.A. Vroon, editors, *Text and Text Processing*, volume 79 of *Advances in Psychology*, pages 107–153. North-Holland.

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

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2024b. Loogle: Can long-context language models understand long contexts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16304–16333.

Siheng Li, Cheng Yang, Zesen Cheng, Lemao Liu, Mo Yu, Yujiu Yang, and Wai Lam. 2024c. Large language models can self-improve in long-context reasoning. *arXiv preprint arXiv:2411.08147*.

Yanyang Li, Shuo Liang, Michael Lyu, and Liwei Wang. 2024d. 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.

Zhan Ling, Kang Liu, Kai Yan, Yifan Yang, Weijian Lin, Ting-Han Fan, Lingfeng Shen, Zhengyin Du, and Jiecao Chen. 2025. Longreason: A synthetic long-context reasoning benchmark via context expansion. *arXiv preprint arXiv:2501.15089*.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Rounak Meyur, Hung D Phan, Koby B Hayashi, Ian Stewart, Shivam Sharma, Sarthak Chaturvedi, Mike Parker, Daniel M Nally, Sadie A Montgomery, Karl Pazdernik, and 1 others. 2025. Benchmarking llms for environmental review and permitting. In *Large Language Models for Scientific and Societal Advances*.

OpenAI. 2025a. Browscomp long hugging face dataset. https://huggingface.co/datasets/openai/BrowseCompLongContext.

OpenAI. 2025b. Gpt-oss model card: Open-weight reasoning models (120b parameters). https://cdn.openai.com/pdf/419b6906-9da6-406c-a19d-1bb078ac7637/oai_gpt-oss_model_card.pdf.

OpenAI. 2025c. Openai mrcr: Long context multiple needle in a haystack benchmark. https://huggingface.co/datasets/openai/mrcr.

Norman Paulsen. 2025. Context is what you need: The maximum effective context window for real world limits of llms. *arXiv preprint arXiv:2509.21361*.

Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, Michael Krumdick, Charles Lovering, and Chris Tanner. 2024. Docfinqa: A long-context financial reasoning dataset. *arXiv preprint arXiv:2401.06915*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Weizhou Shen, Ziyi Yang, Chenliang Li, Zhiyuan Lu, Miao Peng, Huashan Sun, Yingcheng Shi, Shengyi Liao, Shaopeng Lai, Bo Zhang, and 1 others. 2025. Qwenlong-l1. 5: Post-training recipe for long-context reasoning and memory management. *arXiv preprint arXiv:2512.12967*.

Herbert Simon and Allen Newell. 1971. Human problem solving: The state of the theory in 1970. *American Psychologist*, 26:145–159.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. *URL https://arxiv.org/abs/2403.05530*.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and 1 others. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

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. *arXiv preprint arXiv:2505.17667*.

Minzheng Wang, Longze Chen, Fu Cheng, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, and 1 others. 2024. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5627–5646.

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. *arXiv preprint arXiv:2510.19363*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:2403.08319*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Cehao Yang, Xueyuan Lin, Chengjin Xu, Xuhui Jiang, Shengjie Ma, Aofan Liu, Hui Xiong, and Jian Guo. 2025b. Longfaith: Enhancing long-context reasoning in llms with faithful synthetic data. *arXiv preprint arXiv:2502.12583*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

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.

Yang Zhou, Hongyi Liu, Zhuoming Chen, Yuandong Tian, and Beidi Chen. 2025. Gsm-infinite: How do your llms behave over infinitely increasing context length and reasoning complexity? *arXiv preprint arXiv:2502.05252*.

A The Use of LLMs

During the preparation of this manuscript, we utilized Large Language Models (LLMs) to assist with the writing process. Specifically, LLMs were employed to correct grammatical errors, improve phrasing, and enhance the overall clarity and readability of the text.

B Showcases for 5 Atomic Skills

B.1 Foundational Retrieval: NIAH

Multiple specific anchor-question pairs are distributed across a long context. The model is tested on its ability to precisely locate a specific anchor and other similar anchors, and answer associated objective questions.

Case ID: Distributed-Retrieval-NIAH
Category: NIAH
Key Mechanism: *Distributed Anchors*

Context Overview:

- **The Haystack (Background Context):** A long sequence of unrelated text segments (e.g., financial reports, historical essays, or technical logs) serving as noise.
- **Inserted Needles (Distributed Pairs):** Specific Anchor-Question pairs are inserted at random intervals throughout the context.

  [Segment 0-10%] ... The industrial revolution marked a turning point ... **[ID: A-105]** → **{Question 1}** ...
  [Segment 40-50%] ... regarding the molecular structure of polymers ... **[ID: B-292]** → **{Question 2}** ...
  [Segment 80-90%] ... market volatility observed in the last quarter ... **[ID: C-345]** → **{Question 3}** ...
- **Target Needle:** The specific pair required by the instruction (e.g., the pair located at the 50% depth).

Instruction:
Please answer the question following the anchor `[ID: B-292]` in the text above.

Target Answer:
`3398` *(The model must retrieve the exact objective question associated with the specific key, and answer the question correctly.)*

B.2 Robustness to Noise: Anti-Interference

The anchor string 1760536321726-5em0 appears in Doc 1 (twice) and Doc 3 (once). The prompt explicitly requests the content from Doc 3. Tests if the model respects document boundaries despite conflicted retrieval anchors.

Case ID: Anchor-Conflict
Category: Anti-interfere
Key Mechanism: *Anchor Duplication & Document Scope Constraint*

Context Overview:

- **Document 1 (Distractor):** Contains the anchor string twice, associated with a Geometry problem and a Matrix Trace problem.

  ... [Noise] ... **1760536321726-5em0**: Given a set of points $P = \{P_1, \dots\}$ on a plane... [Noise] ... **1760536321726-5em0**: Calculate the trace of the following matrix...
- **Document 2 (Noise):** Irrelevant content (Programming Q&A).
- **Document 3 (Target):** Contains the anchor string once, associated with a Complex Analysis problem.

  ... [Noise] ... **1760536321726-5em0**: Evaluate the integral $\int_C \frac{\exp(z^2)}{z^2(z-1-i)} dz$, where C consists of the contour...

Instruction:
Please answer the question following '`1760536321726-5em0`' in **document 3**.

Target Answer:
πe^{2i} *(The model must ignore the questions in Document 1 and solve the integral in Document 3).*

B.3 Global Integration: Multi-Source Information Processing

A single mathematical problem is split into three parts (Setup, Question 1, Question 2) across three different documents. The model must perform cross-document retrieval to reconstruct the full problem context before solving it.

Case ID: Global-Integration
Category: Global Integration
Key Mechanism: *Fragmented Information Aggregation*

Context Overview:

- **Document 1 (Problem Setup):** Contains the initial conditions of the geometry problem embedded within medical text.

  ... [Medical Q&A Noise] ... **h6qKmUOmz2m**: Given circle $C : x^2 + y^2 + 2x - 2y - 6 = 0$, line

l passes through point $P(1,2)$ and intersects circle C at points A and B. ... [Noise] ...

- **Document 2 (Sub-question 1):** Contains the first part of the specific question embedded within different medical/pharmaceutical text.

  ... [Medical Q&A Noise] ... **h6qKmUOmz2m**: (1) If $\triangle ABC$ is an isosceles right-angled triangle, find the equation of line l; ... [Noise] ...

- **Document 3 (Sub-question 2):** Contains the second part of the question embedded within a financial project report.

  ... [Financial Report Noise] ... **h6qKmUOmz2m**: (2) When $PC \perp l$, find the equation of the circumcircle of $\triangle ABC$. ... [Noise] ...

Instruction:
Please assemble the question corresponding to '`h6qKmUOmz2m`' and answer it.

Target Answer:
(1) $x = 1$ or $3x + 4y - 11 = 0$
(2) $5x^2 + 5y^2 - 6x - 18y + 2 = 0$

(The model must retrieve the setup from Doc 1, combine it with conditions from Doc 2 and Doc 3, and solve the aggregated geometry problem).

B.4 Relational Reasoning: Structure Understanding and Logic

The model must perform a global scan to determine anchor uniqueness frequencies, select the correct document based on key density, and apply strict positional logic to locate the target question, filtering out "trap" anchors (duplicates) during the process.

Case ID: Logic-Constrained
Category: Relational Reasoning
Key Mechanism: Global Frequency Analysis & Relative Positional Logic

---

Context Overview:

- **Document 1 (E-Commerce Report):** Contains multiple unique keys associated with Math and Logic problems.

  ... Consumer Expectations in 2018 ... **OSXVANVP**: Sequence problem $a_{n+1} = a_n + 2^n$ ... **BEKAHJXW**: Statistics problem ...

- **Document 2 (Environmental Report):** Plain text with no embedded keys (Distractor).

- **Document 3 (Biography):** Contains a mix of unique keys and a *duplicated* key.

  ... General Walker commanded the Eighth Army ... **KNGUKM**: Tennis tournament logic ... **ABNKRKRH**: Inequality problem $\frac{\ln a}{e^a} = \dots$ ... **GIEDWE**: Physics wave calculation ... **RJTGAYG**: House logic puzzle ...

- **Document 4 (Linear Algebra):** Contains the *duplicated* key found in Document 3.

  ... Problem 5: Decide if range of map ... **GIEDWE**: Geometry point set problem ...

Instruction:
First, identify anchors that appear **only once** across all documents. Find the document with the highest total count of anchors. In that document, locate the **last unique anchor** and answer the question associated with the **unique anchor immediately preceding it**.

Target Answer (for ABNKRKRH):
C *(Based on the analysis of the inequality* $\frac{\ln a}{e^a} = \frac{\ln b}{b} = -\frac{\ln c}{c} < 0$*, implying* $a < b < c$*).*

B.5 Dynamic State Tracking: Long-Range Computational Reasoning

The model cannot simply retrieve a value; it must first determine the state of the context (counting specific key occurrences), evaluate a logical condition based on that state, and then perform a specific sequence of mathematical operations on values retrieved from distributed anchors.

Case ID: Dynamic-State-Tracking-Math
Category: Dynamic State Tracking / Computational Reasoning
Key Mechanism: *Conditional Logic & Multi-Stage Aggregation*

Context Overview:

- **Document 1 (Survey Analysis):** Contains a key embedded in statistical text.

  ... excluding incomplete questionnaires ... **LTUCRHGAXK**: 82×67 ...

- **Document 2 (Game Design Manual):** Contains a key embedded in UI/UX instructions.

  ... visual balance lines ... **NNQABR**: $6 - (8 - 4) + 16$ ...

- **Document 3 (Stock Market Report):** Contains the target keys for the "False" branch.

  ... A-share defense battle ... **ROX-**

BXXYMYU: $2 - 14.9/6 + 7$ ... market leverage ... **MPEOEZHO**: $19 \times 19.7 + 20/2$ ...

- **Document 4 (Electricity Report):** Contains the target keys for the "True" branch (Distractors).

  ... 16-candle lamps installed ... **TBWYKIE**: $(20 \times 4) \times (3 - 16)$ ... **KYPNFYC**: $\sin(135°) \times \cos(135°) + 3$ ...

Instruction:
If the number of documents containing `LTUCRHGAXK` is greater than the number of documents containing `NNQABR`, calculate the sum of the answers for `TBWYKIE` and `KYPNFYC`. Otherwise, calculate the product of the answers for `ROXBXXYMYU` and `MPEOEZHO`. (Round intermediate steps to 2 decimal places).

Target Answer:
2505.64

C Training Details

This section details all training configurations utilized for enhancing atomic skills, including the optimization objective, sampling strategies, training hyperparameters, reward mechanisms, and data filtering and mixture strategy.

C.1 Group Relative Policy Optimization (GRPO)

Group Relative Policy Optimization (GRPO) (Shao et al., 2024) is designed to reduce computational overhead in training large-scale models by eschewing a trained value function. The core innovation lies in its advantage estimation mechanism.

Instead of relying on a critic network to provide a value baseline, GRPO computes the advantage for each sample relative to the average reward of a group of samples generated from the same prompt. For a given prompt, a group of $\mathcal{G}$ outputs $\{o_1, o_2, \ldots, o_\mathcal{G}\}$ is sampled from the policy π_{old}. Upon obtaining a reward r_i for each output o_i, the advantage A_i is calculated as:

$$A_i = r_i - \frac{1}{|\mathcal{G}|} \sum_{j=1}^{|\mathcal{G}|} r_j \tag{1}$$

This formulation utilizes the group mean reward as a dynamic baseline. The policy is updated using a clipped surrogate objective augmented with a KL-divergence penalty. Formally, the objective is defined as follows.

$$\mathbb{E}\left[\min\left(\frac{\pi_\theta(a|s)}{\pi_{\text{old}}(a|s)} A_i, \text{clip}\left(\frac{\pi_\theta(a|s)}{\pi_{\text{old}}(a|s)}, 1-\epsilon, 1+\epsilon\right) A_i\right)\right] - \beta \cdot D_{\text{KL}}(\pi_\theta \| \pi_{\text{old}}) \tag{2}$$

where θ represents the parameters, ϵ is the clipping hyperparameter, and β controls the strength of the KL regularization.

C.2 Dynamic Sampling

To address the issue of reward homogenization—where similar rewards within a group lead to near-zero advantages and vanishing gradients—Dynamic Sampling (Yu et al., 2025) is used. This strategy dynamically prunes the samples with redundant rewards during training. By ensuring that the training batches are composed of diverse and informative trajectories, this method strengthens the gradient signal and accelerates convergence.

C.3 Training hyperparameters

Hyperparameter	**Value**
Computational Setup	
Hardware	64 × NVIDIA H20
Parallelism Strategy	TP=8
Offload	Disabled
Training Parameters	
Max Sequence Capacity	32,768
Max Input Length	28,576
Max Output Length	8,092
Global Train Batch Size	128
Mini-Batch Size	32
Learning Rate	2×10^{-6}
KL Coefficient (β)	0.001
Rollout Batch Size (Gen BS)	256
Group Size (G)	16
Temperature	0.85
Top-p	0.6 (Disabled)
Clip Ratio High (ϵ_{high})	0.28
Clip Ratio Low (ϵ_{low})	0.20

Table 3: Detailed hyperparameters for RL training.

To ensure reproducibility and provide a comprehensive view of our experimental setup, we detail the complete set of hyperparameters used during the Reinforcement Learning (GRPO) training phase in Table 3. The models were trained using 64 NVIDIA H20 GPUs with Tensor Parallelism (TP) set to 8. To accommodate the extensive context

lengths required for our atomic skill datasets, the maximum sequence capacity was set to 32,768 tokens (allocating 24k tokens for input and 8k tokens for output generation). We utilized a KL divergence coefficient of 0.001 to constrain the policy drift and prevent catastrophic forgetting of the base model's inherent capabilities. During the rollout generation, the group size was configured to 16 with a temperature of 0.85 to maintain a balance between exploration diversity and response quality.

C.4 Reward Mechanisms

Prompt Template for Answer Verification

You are a rigorous answer verification expert. Please verify the following model's answer to ensure it is correct.

Please carefully compare the model's answer with the standard answer. Verification Criteria:

1. **Numerical Answers:** Numerical values must be essentially identical (two decimal places are allowed; e.g., 3.92 and 3.94 are considered correct).
2. **Textual Answers:** Core information must be consistent; the expression style can differ.
3. **Tabular Answers:** Data content must be consistent; the format can differ.
4. Answers that cannot be judged or are unclear are considered incorrect.

Output Format Requirements
Please strictly adhere to the following format and do not add any other content:

[Scoring Criteria]: [Based on the verification criteria and standard answer, provide the scoring criteria]
[Scoring]: ['0' or '1']

The following content requires scoring:

[Standard Answer]: {ground_truth}
[Student Answer]: {model_answer}

Instead of relying on a subjective LLM-as-a-Judge paradigm, our reward modeling leverages the strictly verifiable nature of the Anchor-based Reasoning (AbR) datasets. Since the ground-truth answers for all atomic skills are algorithmically generated and deterministic, we utilize gpt-oss-120B exclusively as a robust answer extractor and equivalence checker. Specifically, it parses the final answer from the model's output, normalizes it, and compares it against the deterministic ground truth. The binary reward (1 for a match, 0 otherwise) is then assigned. This exact-match design effectively mitigates subjective judgment bias and classic reward hacking.

To ensure full reproducibility, we detail the exact prompt templates utilized during our Reinforcement Learning pipeline.

To facilitate complex long-context reasoning, different strategies are applied depending on the backbone model's native capabilities:

- **Instruct Models:** To guide standard instruction-tuned models (e.g., Qwen2.5-Instruct) toward a deliberate reasoning mode, a specialized "chain-of-thought" (CoT) system prompt is introduced. This is coupled with a format-compliance reward signal to ensure the model explicitly outputs its reasoning steps before yielding the final answer.
- **Models with CoT:** For models that inherently incorporate a reasoning process (e.g., DeepSeek-R1-Distill), standard training procedures are followed without additional reasoning-inducing prompts.

C.5 Data Filtering and Mixture Strategy

To ensure optimal training efficiency and avoid both reward sparsity and weak reinforcement signals, we implemented a pass-rate-based data filtering strategy. This mechanism induces an implicit training curriculum designed to target the model's "zone of proximal development." The detailed data curation pipeline proceeds as follows:

Initial Pool Generation: We first generate a vast and balanced pool of AbR tasks spanning all five atomic skills. During this stage, we systematically vary key complexity parameters—such as context length, document count, anchor density, and reasoning depth—to create a broad and continuous spectrum of difficulty.

Pass-rate Computation: To quantify the inherent difficulty of each task, we employ DeepSeek-V3.1 as a reference model. For every query in the initial pool, we sample 16 independent rollouts. We then compute a specific pass-rate for each query, defined as the proportion of correct exact-match answers out of the 16 rollouts.

Difficulty Filtering: We empirically observe that queries with a pass-rate greater than 0.6 are "too

easy" and yield weak gradient signals for reinforcement learning. Conversely, queries with a pass-rate below 0.3 are "too hard" and lead to severe reward sparsity, which destabilizes training. Consequently, we discard items outside this optimal range, retaining only those queries with a pass-rate in the interval [0.3, 0.6].

Emergent Mixture Ratio: It is important to note that the final dataset composition is not based on arbitrary heuristics. Instead, the final training mixture ratio (Anti-interfere : Multi-hop : Multi-source : Logic : Calc-reason : NIAH = 5 : 3 : 2 : 2 : 2 : 1) organically emerges from the aforementioned pass-rate filtering process. This ratio parametrizes a difficulty-aware data distribution, accurately reflecting the relative proportion of skill tasks that naturally fall within the optimal difficulty window. A comprehensive grid search over explicitly defined mixture ratios is left for future work.

D Evaluation and Baseline Details

Evaluation Our evaluation framework incorporates both standard open-source benchmarks and a custom atomic capability dataset. For open-source benchmarks, we select `LongBench-v2` (Bai et al., 2025), `Loong` (Wang et al., 2024), `MRCR` (OpenAI, 2025c), `BrowsCompLong` (OpenAI, 2025a), the `qa_2` subset from `Ruler` (Hsieh et al., 2024), and the real-prompt subset of `Loogle` (Li et al., 2024b). To ensure consistency, we filter these datasets to include only samples with context lengths less than or equal to 128k tokens. The evaluation metrics for these open-source benchmarks align strictly with the protocols defined in their respective original papers. Additionally, we assess atomic capabilities using a dataset constructed via our proposed methodology, employing `gpt-oss-120B` (OpenAI, 2025b) to conduct consistency-based evaluation.

Baselines We compare our approach against some general baselines including the closed-source (`Gemini-3-Pro`) as well as open-source baselines including `Kimi-K2-Thinking` (Team et al., 2025), `DeepSeek-V3.1` (DeepSeek, 2025b), `QwenLong-L1-32B` (Wan et al., 2025), and `Qwen3-235B` (Yang et al., 2025a). In addition, to further show the superiority of our approach, we compare it with some direct baselines, which are obtained by training DeepSeek-R1-distill-32B on the synthetic long-context reasoning datasets named LongReason (Ling et al., 2025), QwenDocqa (Wan et al., 2025), LoonngRl (Wang et al., 2025). In details, we utilize the officially released datasets for LongReason and QwenDocQA. Additionally, we reproduce the data construction pipeline of Loonngrl to generate a dataset of $4,000$ samples for comparison.

To accommodate the long-context requirements of our evaluation, for any baseline model with a native context window smaller than 128k, we apply YaRN for length extrapolation.

E Diagnosing the Capability Gap

By cross-referencing the correlation data (Figure 3) with the absolute performance scores (Table 4), we further identify critical structural flaws in current models.

The "Importance-Proficiency" Mismatch. Our analysis reveals a specific pattern of High Correlation, Low Performance. For example, *Anti-interfere* and *Multi-source* are strong predictors of real-world success ($\rho = 0.94$ and $\rho = 0.91$, respectively). However, most models struggle on these tasks: while `Qwen2.5-32b-instruct` achieves 37.00% on basic NIAH, it drops to **22.13%** on Anti-interfere and **23.88%** on Multi-source. Similarly, *Logic* proves to be the most difficult task (e.g., 13.00% for the same model), though its correlation with real-world performance is moderate ($\rho = 0.77$). Moreover, there exists a Retrieval Ceiling, that is, most models perform relatively well on clean *NIAH* tasks. However, since NIAH is merely a prerequisite, improving it further yields diminishing returns for complex real-world tasks.

The Robustness Bottleneck. The perform difference between *NIAH* and *Anti-Interference* (e.g., 37.00% v.s. 22.13% for `Qwen2.5-32b-instruct`) highlights that models lack discrimination capabilities. They can retrieve information but tend to be easily distracted by "lure" noise. Since *Anti-Interference* correlates highly with real-world benchmarks ($\rho = 0.94$), this fragility can be an important factor for general performance.

F Impact on General Capabilities and Short-Context Reasoning

A common concern when fine-tuning or applying reinforcement learning to Large Language Models on specialized tasks is the potential for catastrophic forgetting, where the model loses its general capabilities or proficiency in shorter contexts. To mitigate this risk, our RL setup is deliberately

Model	NIAH	Anti-interfere	Multi_source	Logic	Calc_reason
QwenLong-L1-32B	66.50%	33.83%	38.81%	28.50%	47.37%
Qwen2.5-32b-instruct	37.00%	22.13%	23.88%	13.00%	36.59%
DeepSeek-R1-Distill-Qwen-32B	58.25%	25.96%	32.54%	17.50%	42.11%
Qwen3-32B	23.00%	12.55%	23.28%	19.00%	19.80%
Qwen3-32B	69.50%	29.15%	37.01%	27.25%	52.13%
Qwen3-30B-A3B-thinking-2507	74.25%	41.70%	47.46%	31.50%	60.15%
Qwen3-14B	46.50%	21.06%	29.25%	19.50%	37.09%
Deepseek-R1-Distill-Qwen-14B	37.50%	13.40%	19.10%	9.00%	24.31%
Qwen2.5-14b-instruct	27.75%	14.47%	19.70%	14.25%	31.58%
Qwen3-8B	42.00%	17.66%	25.37%	15.00%	37.09%
Qwen2.5-7b-instruct	16.50%	8.30%	11.64%	9.25%	19.30%
Deepseek-R1-Distill-Qwen-7B	5.25%	2.13%	3.28%	3.75%	4.51%

Table 4: Performance comparison on atomic skills.

Model	2WikiMultihopQA	MuSiQue	HotpotQA	Average
DeepSeek-R1-Distill-32B	74.0%	51.5%	78.0%	67.8%
DeepSeek-R1-Distill-32B + Ours	**82.5%**	**64.5%**	**83.0%**	**76.7%**

Table 5: Performance comparison on short/medium-context multi-hop reasoning benchmarks.

conservative: we train on a highly targeted, small-scale dataset consisting of only synthetic atomic-skill samples. Furthermore, the GRPO algorithm incorporates an explicit KL divergence penalty to strictly constrain the policy drift, keeping the optimized model close to the initial base model (`DeepSeek-R1-Distill-32B`).

Empirically, we observe no evidence of performance degradation on standard short-to-medium-context reasoning tasks. On the contrary, evaluating the models on widely used multi-hop reasoning benchmarks—including 2WikiMultihopQA, MuSiQue, and HotpotQA reveals consistent and significant improvements. As shown in Table 5, our approach substantially boosts performance across all three datasets compared to the strong base model.

These results suggest that explicitly sharpening atomic skills (e.g., global integration and logical structuring) for long-context reasoning does not erode the model's fundamental abilities. Instead, this atomic-level alignment effectively enhances general multi-hop reasoning capabilities even within shorter contexts.