Reinforced Informativeness Optimization for Long-Form Retrieval-Augmented Generation

Anonymous ACL submission

Abstract

Long-form question answering (LFQA) poses a major challenge in natural language generation, as it requires synthesizing coherent and factually grounded responses to queries by integrating multi-source evidence. Such extended reasoning makes reward design in reinforcement learning (RL) especially critical, since the rewards must be verifiable to ensure factual grounding and stable optimization. However, conventional lexical- or rule-based rewards fail to capture the semantic and multi-evidence nature of LFQA, and existing retrieval-augmented generation (RAG) systems still lack verifiable reward mechanisms, resulting in ungrounded and inconsistent answers. We propose RioRAG, a framework for reinforced verifiable informativeness optimization. First, it defines informativeness as a measurable and externally verifiable objective for RL. Second, it introduces a nugget-centric hierarchical verification mechanism that extracts atomic facts, integrates cross-source evidence, and evaluates factual alignment between responses and references. This formulation requires no supervised data or multiple inference passes, while encouraging deliberate, slow-thinking reasoning during long-form generation. Experiments on LongFact and RAGChecker show that RioRAG achieves higher factual recall and faithfulness, establishing verifiable reward modeling as a foundation for trustworthy long-form RAG. Our codes are available at https://anonymous.4open.science/r/RioRAG/.

1 Introduction

Long-form question answering (LFQA) represents a crucial step toward enabling AI systems to deliver comprehensive and factually reliable responses by generating elaborate and multi-sentence answers, conditioning language models on input queries (Stelmakh et al., 2022). Retrieval-augmented generation (RAG) has emerged as a compelling paradigm for such knowledge-intensive tasks, as it grounds generation in factual content retrieved from external corpora (Lewis et al., 2020). However, producing reliable long-form answers remains challenging since large language models (LLMs) should synthesize information from multiple retrieved sources into coherent and factual paragraphs.

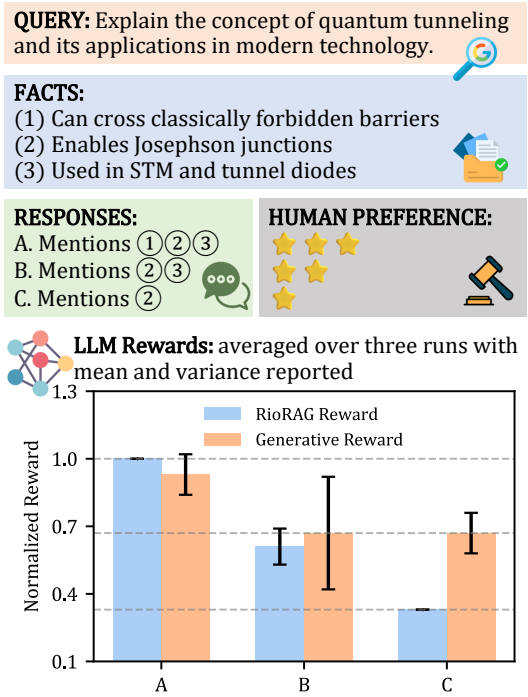


Figure 1: Existing long-form reward modeling exhibits (i) instability, with high variance when evaluating the same response multiple times, and (ii) unreliability, as their score rankings often differ from human judgments across responses.

While RAG mitigates hallucination in long-form generation by grounding models on retrieved evidence (Asai et al., 2023), several challenges remain unresolved. First, LLMs often misuse factual information (Gao et al., 2023b; Łajewska and Balog, 2025), either due to residual hallucination or failure to extract the correct evidence from lengthy retrieved documents. Besides, generated answers fre-

quently exhibit limited informativeness (Ru et al., 2024), failing to comprehensively incorporate and reuse the available evidence. Recent efforts such as prompt engineering and template-based supervised fine-tuning (SFT) partially alleviate these issues (Wei et al., 2022), yet they rely heavily on strong annotation signals and external guidance. Moreover, these methods restrict models' self-evolution and often reduce generalization and diversity (Shao et al., 2024), making them less adaptable to LFQA tasks across diverse domains.

Recently, reinforcement learning (RL) has shown strong potential to improve LLMs by providing outcome-based feedback and enabling self-exploration (Guo et al., 2025a) in reasoning tasks such as mathematics and code generation. However, applying RL to LFQA remains underexplored with several unique challenges. First, rule-based outcome reward models (ORMs) often fail in this context (Li et al., 2025). Although they have succeeded in short-form question answering with RL settings, it is difficult to directly define lexical rules for long, open-ended answers. Second, the scoring process is unstable and poorly defined (Zhang et al., 2024). Even with detailed guidelines, human-defined scoring rules can still produce inconsistent or ambiguous rewards, making it difficult to assess whether the improvement is genuine. This ambiguity also increases the difficulty of verifying the reasonableness of scores, calling for a quantifiable and verifiable reward mechanism. Third, ORM-generated scores are often inaccurate. As shown in Figure 1, when facing long reference documents, the reward model itself may hallucinate or miss key information, leading to misleading reward signals.

To address these challenges, we propose RioRAG, a Reinforced Informativeness Optimization-based RL method for long-form RAG. Our goal is to build a retrieval-augmented generation system that can generate answers with higher factual coverage and more stable reward learning. First, RioRAG introduces a verifiable informativeness reward that measures how well the generated content is supported by retrieved evidence, allowing the reward signal to be objective and reproducible rather than rule-based. Second, we design a nugget-centric hierarchical reward model that extracts atomic facts (nuggets) from retrieved documents and evaluates factual alignment in a length-adaptive way, ensuring accurate and consistent reward computation even with long references. This design activates the LLM's self-exploration ability through reinforcement learning while avoiding dependency on manual scoring or handcrafted supervision. Extensive experiments on two published benchmarks, LongFact and RAGChecker, with zero-shot evaluation show that the RioRAG achieves superior performance compared with a series of state-of-the-art methods, demonstrating the effectiveness of the proposed innovations.

2 Related Work

2.1 Long-Form Question Answering

Research on long-form question answering (LFQA) has evolved through three paradigm shifts: from fine-tuned generative language models to retrieval-augmented architectures, and more recently to human-aligned large language models (LLMs). Early foundational work established the feasibility of abstractive approaches, with the ELI5 dataset (Fan et al., 2019) demonstrating that sequence-to-sequence models leveraging retrieved evidence could generate plausible answers. Subsequent efforts leverage LLMs and human feedback, which employs reinforcement learning (RL) from human preferences to generate answers supported by explicit quotations (Menick et al., 2022). Retrieval-augmented generation has emerged as a dominant strategy to enhance factual grounding of LFQA. Several studies aim to fine-tune models in real-world text-based web-browsing interfaces (Nakano et al., 2021; Qin et al., 2023). Recent work demonstrates that explicit source citation during training significantly improves answer verifiability (Menick et al., 2022), while post-hoc attribution methods enable verification of pre-generated text (Wang et al., 2022; Gao et al., 2023a). Due to the inherent characteristics of LFQA, faithfulness has become an important consideration for ensuring that long answers are faithful to evidence (Gao et al., 2023b; Zhao et al., 2024). It has been shown that training an open-book QA model to cite sources greatly improves trustworthiness (Menick et al., 2022). Alternatively, model confidence in long-form answers can be calibrated by treating answer correctness within a probabilistic calibration framework (Huang et al., 2024). Moreover, LLM-based evaluation can be applied to assess LFQA quality (Han et al., 2024) automatically.

2.2 Reinforcement Learning based RAG

RL has shown great potential in optimizing RAG systems by dynamically adjusting retrieval and gen-

eration strategies. Central to this approach is the integration of policy-gradient methods with tailored reward functions. Early effort refines behavior-cloning pretraining of web-browsing agents via human-feedback–based rewards to improve factual alignment (Nakano et al., 2021). Fine-grained reward models are also designed to quantify coherence and information gain at multiple semantic levels (Cai et al., 2024), or fine-tuned for specific retrieval domains to align the RAG model to human performance by generating synthetic training data (Nguyen et al., 2024). In addition, composite reward ensembles balance generation quality against coverage objectives (Zhang et al., 2025). In the retrieval side, retrieval modules themselves can be tuned with group-wise relative policy optimization (Huang et al., 2025) or multi-agent coordination (Chen et al., 2025). Cost-sensitive retrieval policies further leverage learned value estimators to decide when to invoke external search, excluding latency and informational utility (Kulkarni et al., 2024). Recent advances, particularly the success of the DeepSeek-R1 (Guo et al., 2025b) model, have spurred research interest in employing RL to endow LLMs with the capability of autonomously invoking retrieval during extended reasoning chains, thereby enhancing the intelligence of the entire RAG process (Jin et al., 2025). In contrast to these methods, our proposed RioRAG framework eliminates the need for supervised training data by introducing nugget-centric hierarchical reward modeling, enabling reinforcement learning to directly optimize informativeness in long-form RAG and achieve long-CoT reasoning.

3 Task Formulation

LFQA extends conventional SFQA to generate coherent, factual, and detailed multi-paragraph responses (*e.g.,* explanations or reports). Given a user query q and a large web corpus $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$, a retriever R retrieves a subset of relevant documents $\mathcal{D}_q \subseteq \mathcal{D}$, and a generator G first produces a complete response sequence $y = G(q, \mathcal{D}_q)$ conditioned on both the query and the retrieved evidence. The response y naturally contains a reasoning part and a final answer part, which we separate by a predefined delimiter:

$$[\, r_{1:T} \parallel a_{1:M} \,] = y,$$

RioRAG follows an outcome reward model (ORM) formulation. Both evaluation and reward computation are based solely on the final answer a, while the reasoning content in y (*e.g.,* chain-of-thought tokens) is ignored by the reward model. Formally, the objective of long-form RAG is to generate an answer a that maximizes factual correctness, informativeness, and coherence with respect to $\mathcal{D}_q$. In RioRAG, we approach this objective from an RL perspective, where the model learns to maximize a verifiable informativeness reward derived from the generated outcome.

4 Method

Figure 2 shows the overall pipeline of RioRAG. Given a query, RioRAG retrieves diverse and up-to-date web documents and generates long-form answers through reinforcement learning. The framework consists of two main components: (i) *reinforced informativeness optimization*, which maximizes factual coverage reward, and (ii) *nugget-centric hierarchical reward modeling*, which provides fine-grained, verifiable feedback based on evidence-level nuggets. Together, these components enable stable and unsupervised optimization of long-form RAG systems.

4.1 Reinforced Informativeness Optimization

Previous works mainly relied on SFT, which depends on pre-defined templates or powerful teacher models and thus limits generalization and adaptability to open-domain queries (Menick et al., 2022). Inspired by the recent success of RL in enhancing LLMs through self-exploration in tasks like math and coding (Yang et al., 2025), we extend this paradigm to the LFQA setting of RAG. RioRAG aims to enable models to self-improve by interacting with real-time web knowledge and optimizing generation quality without external supervision.

Training Data Construction. Existing RAG systems often depend on labeled data from strong teacher models and static retrieval corpora, both of which limit generalization to new domains and fail to reflect real-time web knowledge. To overcome these limitations, we reconstruct the training setup using ELI5 (Fan et al., 2019), retaining only queries while removing human-written answers. For each query, we dynamically retrieve the top-K webpages to form $\mathcal{D}_q$, ensuring that the model learns from current and diverse online sources. This design allows the model to explore and adapt to evolving web information under RL optimization, improving factuality and robustness

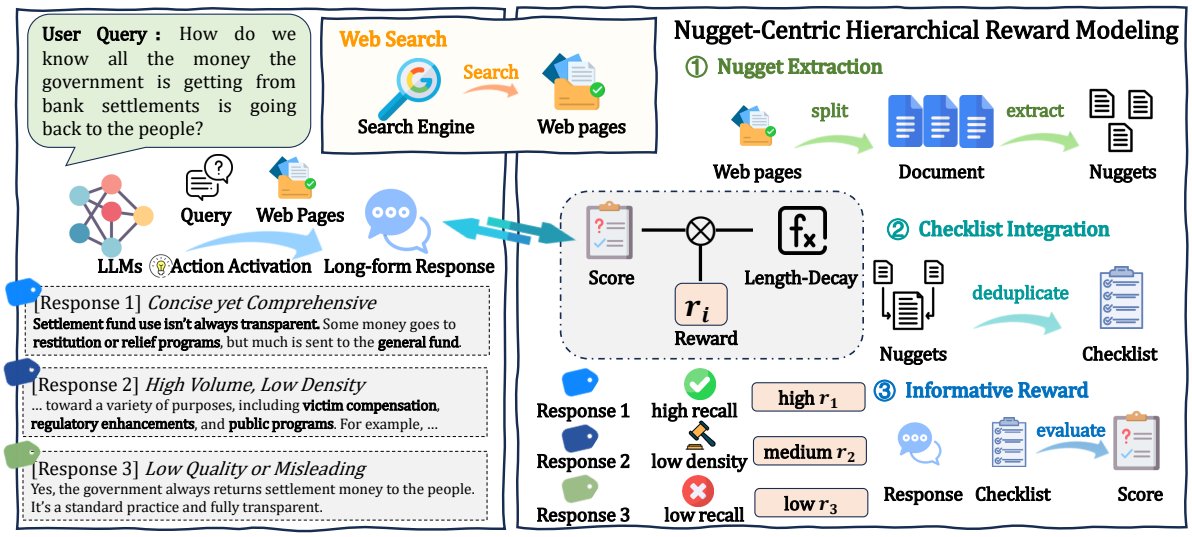


Figure 2: Overall illustration of the proposed RL-based RioRAG framework.

in long-form generation.

Stable Reinforcement Optimization. Conventional RL optimization for long-form text often suffers from unstable gradients and inconsistent reward scaling, especially when model outputs vary significantly in length or quality. Inspired by relative normalization strategies that stabilize optimization in long-horizon reasoning tasks (Guo et al., 2025a), we adopt Group-wise Relative Policy Optimization (GRPO) (Shao et al., 2024) to improve training stability. Given G sampled completions $\{o_1, \ldots, o_G\}$ with corresponding rewards $\{r_1, \ldots, r_G\}$, GRPO normalizes rewards within each sampled group:

$$A_i = \frac{r_i - \mu_r}{\sigma_r + \epsilon},$$

$$\mu_r = \frac{1}{G}\sum_{j=1}^{G} r_j, \quad \sigma_r = \sqrt{\frac{1}{G}\sum_{j=1}^{G}(r_j - \mu_r)^2}.$$

This group-wise normalization mitigates reward variance across samples, stabilizes policy updates, and leads to smoother reward improvement during long-form RL training.

Action Activation. Previous approaches introduce new delimiter tokens (*e.g.*, <think>) to structure reasoning, which hinders warm-start from instruction-tuned LLMs and biases learning toward format imitation. We observe that even base models naturally support Markdown formatting. Hence, we leverage Markdown headers to separate reasoning and answers without adding new tokens. This lightweight design preserves pretrained priors, improves optimization stability, and provides a structured foundation for our subsequent hierarchical reward modeling.

4.2 Nugget-based Verifiable Reward Modeling

Existing ORMs for long-form generation are often non-verifiable and unstable (Ru et al., 2024), as they rely on heuristic or model-based scoring that may hallucinate or fluctuate across runs. To overcome these challenges, we propose a *nugget-based hierarchical reward modeling* approach that decomposes reward computation into three verifiable stages.

(1) Factual Nugget Extraction. For each retrieved document $d_i \in \mathcal{D}_q$, atomic factual statements (nuggets) are automatically identified as minimal information units representing concrete facts. Each nugget n_{ij} is extracted as a concise factual clause from d_i, forming a nugget set:

$$\mathcal{N}(\mathcal{D}_q) = \bigcup_{d_i \in \mathcal{D}_q} \text{Extract}(d_i).$$

(2) Evidence Checklist Synthesis. The extracted nuggets are clustered and merged into a unified checklist that captures all distinct factual evidence relevant to query q. This aggregation removes redundancy and enables consistent cross-document evaluation:

$$\mathcal{C}(q, \mathcal{D}_q) = \text{MergeCluster}(\mathcal{N}(\mathcal{D}_q)).$$

(3) Informativeness Assessment. Given a generated output o, we compute the informativeness

reward by measuring the proportion of checklist nuggets covered in o:

$$\mathcal{I}(q, o, \mathcal{D}_q) = \frac{|\{n \in \mathcal{C}(q, \mathcal{D}_q) \mid n \subseteq o\}|}{|\mathcal{C}(q, \mathcal{D}_q)|}.$$

Each matched nugget can be explicitly traced to its source document, ensuring transparent and verifiable reward computation.

Length Decay. To prevent overestimation from excessively long responses, we apply a length-decay normalization on the reward computation. The decay term is applied only to the final answer segment, leaving the reasoning tokens unaffected to preserve the model's deliberative process. Formally, the length-adjusted reward is defined as:

$$r_i = \begin{cases} s_i \exp\big[-k((l - l_0)/\tau)^m\big], & l > l_0, \\ s_i, & \text{otherwise}, \end{cases} \tag{1}$$

where s_i is the base informativeness score, l denotes the answer length, l_0 is the threshold of unpenalized length, τ controls the decay rate, and k, m regulate the sharpness of the decay curve. This design maintains factual compactness without penalizing intermediate reasoning, ensuring the reward focuses on the informativeness and precision of final outputs.

This hierarchical design stabilizes reward estimation, mitigates noise from heuristic scoring, and provides a reproducible, fact-grounded supervision signal for reinforcement learning.

4.3 Discussion

The novelty and stability of RioRAG lie in three aspects. (i) **Clear reward rules:** the nugget-based design defines explicit and interpretable optimization targets, helping the model follow precise guidance during learning. (ii) **Human verifiability:** each reward can be directly checked and traced to its source document, making inspection and manual calibration simple and transparent. (iii) **Short-context reliability:** by evaluating short factual units instead of full responses, the reward leverages the LLM's accuracy on short texts and avoids long-context degradation such as the "lost in the middle" issue. These three factors together make RioRAG robust, interpretable, and easy to reproduce.

Our theoretical analysis further supports these findings. Nugget decomposition reduces reward variance as $\mathcal{O}(1/M)$, while the answer-only length decay enforces Lipschitz stability without affecting reasoning. The nugget locality also limits error accumulation and long-context degradation. These properties explain RioRAG's stable convergence and factual consistency (see Appendix A).

5 Experiment

In this section, we detail the experimental setup, present the main results, and further support our findings with ablation studies and in-depth analysis.

5.1 Experimental Setup

5.1.1 Datasets

We use the ELI5 dataset (Fan et al., 2019) as the training source, but only its question corpus is used without reference answers. This avoids overfitting to concise, single-perspective annotations and better reflects real retrieval-augmented settings. A total of 10K questions are randomly sampled for RL training.

For evaluation, we adopt two long-form QA benchmarks: **LongFact** (Wei et al., 2024) and **RAGChecker** (Ru et al., 2024). LongFact covers 38 domains consolidated into 8 major categories, with answers annotated by atomic factual units for fine-grained factual verification. RAGChecker includes 10 public datasets spanning 4K questions, designed to assess factual grounding and retrieval-based answer quality across multiple dimensions. Further dataset details are provided in Appendix B.

5.1.2 Evaluation Metrics

Following standard LFQA studies (Fan et al., 2019), we use *fact recall* (FR) and *information density* (ID) as the main metrics, measuring factual completeness and conciseness. We further report RAGChecker's multi-dimensional metrics (Ru et al., 2024), including *faithfulness*, *hallucination*, and *context utilization*, to capture both factual reliability and retrieval effectiveness. Full metric definitions are presented in Appendix C.

5.1.3 Baselines

To evaluate the performance of RioRAG, we conduct comprehensive comparisons with various classical and state-of-the-art baseline methods across different categories, ensuring a thorough understanding of the proposed approach. The baselines are categorized into three groups based on their training paradigms: prompt-based unsupervised methods, supervised fine-tuning (SFT)-based approaches, and RL-based techniques. For prompt-

Method	Science		Tech.		Medicine		Law		Culture		Events		Commun.		Lifestyle		**Average**	
	FR	**ID**	**FR**	**ID**	**FR**	**ID**	**FR**	**ID**	**FR**	**ID**	**FR**	**ID**	**FR**	**ID**	**FR**	**ID**	**FR**	**ID**
Prompt-based Methods																		
Direct-RAG	45.3	54.2	60.6	69.3	45.7	61.1	46.9	31.4	45.4	55.4	51.2	44.3	55.5	60.2	48.7	69.8	49.6	53.5
Chain-of-Thought	55.5	51.6	53.7	71.8	46.4	58.8	45.2	30.8	47.1	61.2	48.5	46.3	54.6	59.0	54.6	70.8	50.6	54.1
Chain-of-Note	45.0	52.6	58.4	71.5	43.6	58.7	42.3	27.1	40.9	56.9	47.7	46.5	47.8	56.9	48.8	71.0	46.0	52.5
GopherCite	54.4	56.6	63.8	73.9	56.2	54.1	55.6	33.7	53.3	60.6	48.9	46.4	61.5	64.0	54.1	72.5	55.9	56.1
Supervised Fine-tuning based Methods																		
Chain-of-Note	65.3	123.3	50.0	129.9	77.4	130.6	57.5	93.8	67.5	134.3	52.5	80.2	64.8	147.8	64.6	114.7	62.2	119.5
GopherCite	59.2	121.5	58.4	145.5	74.4	118.8	59.3	80.2	69.0	146.0	68.4	101.3	61.8	144.1	60.0	103.5	63.2	119.7
RL-based Methods																		
DPO	59.0	106.7	66.2	137.0	60.7	96.4	65.7	61.8	69.2	115.0	56.6	83.2	60.5	122.9	61.3	113.9	62.8	102.7
RioRAG	**69.7**	**146.7**	**63.3**	**170.4**	**77.4**	**142.1**	**77.9**	**113.4**	**78.0**	**120.4**	**71.6**	**117.7**	**75.2**	**170.7**	**61.5**	**144.9**	**72.8**	**138.8**

Table 1: The results on eight broader categories of LongFact benchmark with the average results of the eight categories, where FR denotes fact recall and ID denotes information density.

Method	**Fact-Rec↑**	**Info-Den↑**	**Cont-Util↑**	**Rel-NS↓**	**Irrel-NS↓**	**Hallu.↓**	**Self-Know↓**	**Faith.↑**
Prompt-based Methods								
Direct-RAG	38.3	91.6	22.6	8.2	7.5	37.0	8.1	45.3
Chain-of-Thought	50.4	146.5	24.3	4.6	4.2	30.2	9.7	48.0
Chain-of-Note	38.7	144.3	18.3	6.8	5.1	53.0	6.9	35.7
GopherCite	51.4	138.5	26.0	5.1	4.3	29.2	10.8	47.5
Supervised Fine-tuning based Methods								
Chain-of-Note	54.2	190.2	22.7	4.3	3.7	22.6	7.8	30.2
GopherCite	62.6	209.9	26.0	5.1	4.3	29.2	10.8	52.5
RL-based Methods								
DPO	61.2	149.6	26.0	5.2	6.0	27.8	8.0	53.1
RioRAG	**66.3**	**224.6**	**27.8**	**4.3**	**3.6**	**20.9**	**5.0**	**58.2**

Table 2: Average results across ten domains on the RAGChecker benchmark. Fact-Rec refers to fact recall, Info-Den to information density, Cont-Util to context utilization, Rel-NS and Irrel-NS to relevant and irrelevant noise sensitivity, Hallu. to hallucination, Self-Know to self-knowledge, and Faith. to faithfulness.

based methods, we select GopherCite (Menick et al., 2022), chain-of-thought (Wei et al., 2022) and chain-of-note (Yu et al., 2024). Among SFT-based approaches, we employ chain-of-note and GopherCite with the SFT setting. For RL-based methods, we adopt the Direct Preference Optimization (DPO) (Rafailov et al., 2023) framework. All baseline implementations are manually reimplemented with rigorous adherence to identical experimental configurations to ensure a fair comparison. This evaluation protocol guarantees the reliability of performance benchmarking while controlling for potential confounding factors in implementation differences.

5.2 Main Results

The results of different methods evaluated on LongFact and RAGChecker are shown in Table 1 and Table 2. It can be observed that:

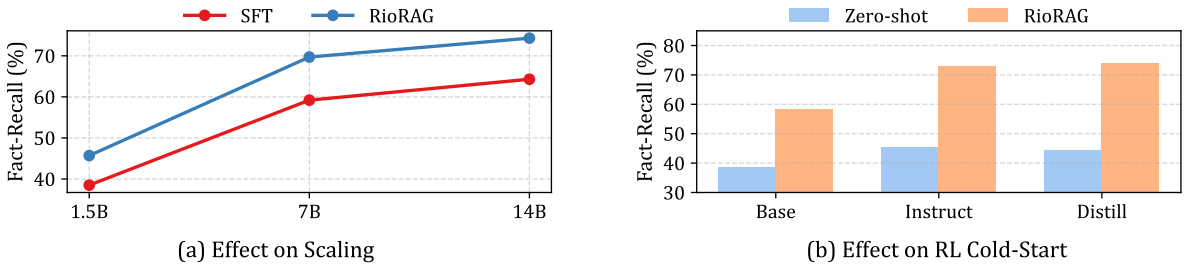
(1) Our comprehensive evaluation reveals that SFT-based baselines substantially outperform prompt-based approaches, demonstrating the inherent limitations of prompt engineering in handling complex information synthesis tasks. The proposed RioRAG framework establishes a significant improvement across all metrics. This improvement stems from the reinforced informativeness optimization paradigm, which implements a nugget-centric hierarchical reward mechanism to guide LLMs in processing long-context inputs.

(2) Comprehensive evaluation on RAGChecker demonstrates that RioRAG excels in long-form RAG tasks across multiple critical dimensions, including knowledge point coverage, information density, retrieval utilization, hallucination mitigation, and internal knowledge integration. These results underscore the multidimensional efficacy of the proposed approach.

(3) Compared to off-line RL-based methods such as DPO, RioRAG demonstrates superior performance in long-form reasoning tasks. By leveraging an enhanced on-policy GRPO algorithm, RioRAG enables more comprehensive exploration of potential reasoning strategies during generation, thereby optimizing the LLM more effectively through informativeness-driven reward feedback.

Method	Science		Tech.		Medicine		Law		Culture		Events		Commun.		Lifestyle		**Average**	
	FR	**ID**	**FR**	**ID**	**FR**	**ID**	**FR**	**ID**	**FR**	**ID**	**FR**	**ID**	**FR**	**ID**	**FR**	**ID**	**FR**	**ID**
RioRAG	**69.7**	**146.7**	**63.3**	**170.4**	**77.4**	**142.1**	**77.9**	**113.4**	**78.0**	**120.4**	**71.6**	**117.7**	**75.2**	**170.7**	**61.5**	**144.9**	**72.8**	**138.8**
w/o Info. Optim.	34.4	79.6	53.2	147.6	32.0	92.3	33.4	66.6	40.9	87.5	43.2	76.6	43.9	106.4	36.6	97.3	39.5	90.8
w/o Nugget Reward	36.0	77.6	56.6	119.3	23.8	62.5	36.0	42.4	37.2	80.7	30.7	65.4	36.8	83.2	40.1	112.7	37.0	77.3
w/o Length Decay	57.0	99.2	65.5	139.9	58.3	109.9	62.0	88.7	56.5	87.5	43.7	63.0	45.2	70.7	68.5	108.8	56.2	91.3
w/ Generative RM	58.7	104.3	63.3	159.2	36.6	62.7	55.7	70.9	63.1	112.7	54.3	70.3	52.5	103.2	61.7	99.3	54.7	97.2
w/ Off-Policy RL	41.6	41.4	61.7	65.5	34.1	35.0	44.8	25.1	46.8	52.2	46.4	39.8	56.2	53.1	61.8	60.2	49.3	45.5

Table 3: Results of the RioRAG variants on LongFact.



(a) Effect on Scaling (b) Effect on RL Cold-Start

Figure 3: In-depth analysis on scaling law and RL cold-start.

5.3 Ablation Studies

In this section, we conduct an ablation study to evaluate the effectiveness of critical strategies in RioRAG comprehensively on LongFact. Here, we consider five variants built on RioRAG for evaluation: (a) *w/o Info. Optim.* removes the informativeness-based reward optimization during RL, replaced by direct quality evaluation; (b) *w/o Nugget Reward* removes the nugget-wise information extraction and use the full webpage for checklist integration; (c) *w/o Length Decay* eliminates the length penalty in Equation (1); (d) *w/ Generative RM* denotes the setting where the policy is trained with a 32B generative reward model providing scalar feedback, rather than the verifiable reward used in RioRAG; (e) *w/ Off-Policy RL* utilizes an off-policy RL method that employs a static sampling strategy wherein all queries are pre-processed through offline rollouts to generate complete trajectories before being uniformly scored.

Table 3 presents the results for the variants of our method, from which we can observe the following findings: (a) The performance drops in *w/o Info. Optim.*, demonstrating that using informativeness as the objective for optimization enhances the performance of long-form RAG models through the guidance of reasoning. (b) The performance drops in *w/o Nugget Reward*, demonstrating incorporating nugget-wise information extraction enables the model to better capture core facts. (c) The performance drops in *w/o Length Decay*, underscoring the critical role of incorporating the length penalty in mitigating excessive response length. (d) The performance degradation observed in *w/ Generative RM* indicates that, without an appropriate pipeline design, unstable and unreliable rewards hinder effective fine-tuning. (e) The performance significantly drops in *w/ Off-Policy RL*, demonstrating that the off-policy method may contain a mismatch between the behavior policy and the target policy.

5.4 Further Analysis

5.4.1 Scaling Law of RioRAG

To investigate the scalability characteristics of RioRAG, we conduct a systematic analysis using the Qwen2.5 model with varying parameter sizes (1.5B, 7B, and 14B). As illustrated in Figure 3 (a), the experimental results demonstrate that RioRAG significantly outperforms SFT at all model scales, with performance consistently improving in accordance with scaling laws.

We can first observe that larger models exhibit improved semantic understanding for both query formulation and webpage relevance assessment. Second, RioRAG benefits from increased model capacity for learning sophisticated retrieval utilization strategies. Third, the enhanced generation capability of larger models enables more effective utilization of retrieved webpages while reducing hallucination risks through better alignment with

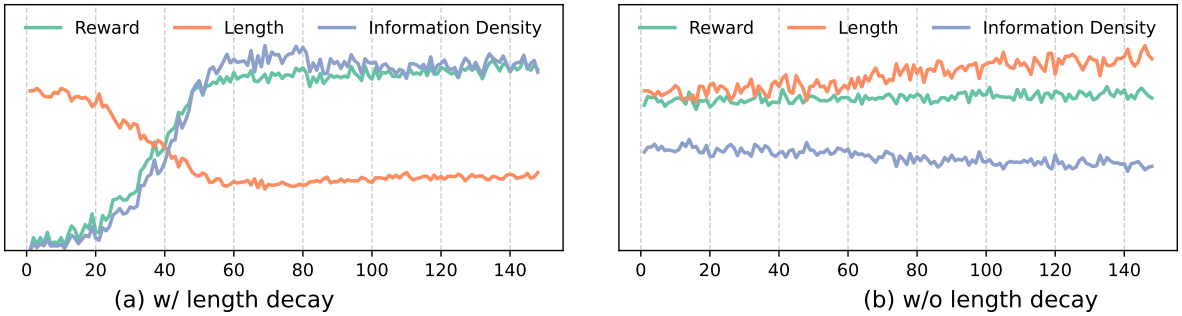


Figure 4: Analysis on co-evolution of generation length and reward during RL training.

the reward model's feedback. Notably, the performance growth curve shows a sublinear relationship between model size and metric improvements, aligning with observations from language model scaling studies (Wei et al.). This phenomenon suggests that while our RioRAG framework effectively leverages model scale, there exists an upper bound where additional parameters may not proportionally improve RAG performance, which is a critical consideration for practical system deployment.

5.4.2 Effect on RL Cold-Start

To examine how model initialization affects RL training, we compare three variants: a base model (Qwen2.5-7B-Base), an instruction-tuned model (Qwen2.5-7B-Instruct), and an R1-distilled model (R1-Distilled-Qwen2.5-7B) incorporating DeepSeek R1's slow-thinking distillation (Guo et al., 2025b). As shown in Figure 3(b), the instruction-tuned model gains 24.4%, and the R1-distilled model achieves the largest improvement of 29.6% after RL training.

The results show that initialization strongly affects training stability and reward learning. The base model, lacking alignment and reasoning priors, struggles to explore effectively. In contrast, the R1-distilled model benefits from pre-established reasoning patterns, enabling more stable reward estimation and efficient policy updates. This supports that structured reasoning provides a favorable starting point for RL.

5.4.3 Co-Evolution of Generation Length and Reward

To investigate the dynamics of generation length control during RL training, we systematically analyze the interaction between sequence length evolution and reward optimization. The experiments are conducted on the Qwen model, with length decay activated when generated sequences exceed a predefined threshold. Figure 4 illustrates the co-evolution of average generation length and reward scores across training steps.

Without length control, the model often produces overly long answers, while the reward score stays almost unchanged. This shows a form of reward hacking, where the model gains higher scores through longer outputs instead of better content. After adding the length-decay term, the model first explores longer responses and then learns to shorten them while keeping rewards stable. The results show that length regularization helps improve both clarity and information density in long-form generation.

6 Conclusion

In this work, we address long-form RAG limitations through RioRAG, an RL framework that redefines long-form RAG training via reinforced informativeness optimization with nugget-centric hierarchical reward modeling. RioRAG directly optimizes informativeness through a quantifiable reward design for factual alignment, without the need for scarce training data. Our experiments on two benchmarks demonstrate that RioRAG fundamentally improves the quality of long-form RAG. By addressing the core challenges identified in long-form RAG, RioRAG advances the development of trustworthy generative systems for real-world knowledge applications. Moreover, the success of nugget-level reward modeling suggests that future evaluation frameworks for long-form tasks should prioritize granular factual alignment over surface-level metrics. Limitations include the current focus on English corpora and reliance on automatic nugget extraction, which may inherit biases from pre-trained models. For future work, we will extend the framework to multilingual settings and investigate human-in-the-loop reward refinement.

Limitations

While RioRAG effectively improves the stability and verifiability of long-form RAG training, several limitations remain. First, our current reward primarily targets factual informativeness and does not explicitly capture other aspects of long-form quality, such as linguistic style, coherence, and readability. These factors may further influence human preference alignment and should be considered in future extensions. Second, although we examine models of different scales, computational resources restrict us from scaling beyond 32B parameters. Larger-scale experiments (*e.g.,* 72B) could provide deeper insights into the scalability and robustness of verifiable reward optimization.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Tianchi Cai, Zhiwen Tan, Xierui Song, Tao Sun, Jiyan Jiang, Yunqi Xu, Yinger Zhang, and Jinjie Gu. 2024. Forag: Factuality-optimized retrieval augmented generation for web-enhanced long-form question answering. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 199–210.

Yiqun Chen, Lingyong Yan, Weiwei Sun, Xinyu Ma, Yi Zhang, Shuaiqiang Wang, Dawei Yin, Yiming Yang, and Jiaxin Mao. 2025. Improving retrieval-augmented generation through multi-agent reinforcement learning. *arXiv preprint arXiv:2501.15228*.

LE Clarke. 1980. Probability and measure, by patrick billingsley. pp 515.£ 15· 20. 1979. sbn 0 471 03173 9 (wiley). *The Mathematical Gazette*, 64(430):293–294.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. 2023a. Rarr: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025b. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*.

Rujun Han, Yuhao Zhang, Peng Qi, Yumo Xu, Jenyuan Wang, Lan Liu, William Yang Wang, Bonan Min, and Vittorio Castelli. 2024. Rag-qa arena: Evaluating domain robustness for long-form retrieval augmented question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4354–4374.

Jerry Huang, Siddarth Madala, Risham Sidhu, Cheng Niu, Julia Hockenmaier, and Tong Zhang. 2025. Rag-rl: Advancing retrieval-augmented generation via rl and curriculum learning. *arXiv preprint arXiv:2503.12759*.

Yukun Huang, Yixin Liu, Raghuveer Thirukovalluru, Arman Cohan, and Bhuwan Dhingra. 2024. Calibrating long-form generations from large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13441–13460.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

Mandar Kulkarni, Praveen Tangarajan, Kyung Kim, and Anusua Trivedi. 2024. Reinforcement learning for optimizing rag for domain chatbots. *arXiv preprint arXiv:2401.06800*.

Weronika Łajewska and Krisztian Balog. 2025. Ginger: Grounded information nugget-based generation of responses. *arXiv preprint arXiv:2503.18174*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025. Search-o1: Agentic search-enhanced large reasoning models. *arXiv preprint arXiv:2501.05366*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: financial opinion mining and question answering. In *Companion proceedings of the the web conference 2018*, pages 1941–1942.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(2):4–es.

Thang Nguyen, Peter Chin, and Yu-Wing Tai. 2024. Reward-rag: Enhancing rag with reward driven supervision. *arXiv preprint arXiv:2410.03780*.

Yujia Qin, Zihan Cai, Dian Jin, Lan Yan, Shihao Liang, Kunlun Zhu, Yankai Lin, Xu Han, Ning Ding, Huadong Wang, et al. 2023. Webcpm: Interactive web search for chinese long-form question answering. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Sara Rosenthal, Avirup Sil, Radu Florian, and Salim Roukos. 2025. Clapnq: C ohesive l ong-form a nswers from p assages in natural questions for rag systems. *Transactions of the Association for Computational Linguistics*, 13:53–72.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, et al. 2024. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. *Advances in Neural Information Processing Systems*, 37:21999–22027.

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

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. Asqa: Factoid questions meet long-form answers. *arXiv preprint arXiv:2204.06092*.

Cunxiang Wang, Ruoxi Ning, Boqi Pan, Tonghui Wu, Qipeng Guo, Cheng Deng, Guangsheng Bao, Qian Wang, and Yue Zhang. 2024. Novelqa: A benchmark for long-range novel question answering. *arXiv e-prints*, pages arXiv–2403.

Shufan Wang, Fangyuan Xu, Laure Thompson, Eunsol Choi, and Mohit Iyyer. 2022. Modeling exemplification in long-form question answering via retrieval. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2079–2092.

Larry Wasserman. 2013. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Zixia Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, et al. 2024. Long-form factuality in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Fangyuan Xu, Kyle Lo, Luca Soldaini, Bailey Kuehl, Eunsol Choi, and David Wadden. 2024. Kiwi: A dataset of knowledge-intensive writing instructions for answering research questions. *arXiv preprint arXiv:2403.03866*.

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

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Peixin Cao, Kaixin Ma, Jian Li, Hongwei Wang, and Dong Yu. 2024. Chain-of-note: Enhancing robustness in retrieval-augmented language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14672–14685.

Hanning Zhang, Juntong Song, Juno Zhu, Yuanhao Wu, Tong Zhang, and Cheng Niu. 2025. Rag-reward: Optimizing rag with reward modeling and rlhf. *arXiv preprint arXiv:2501.13264*.

Jiajie Zhang, Zhongni Hou, Xin Lv, Shulin Cao, Zhenyu Hou, Yilin Niu, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. 2024. Longreward: Improving long-context large language models with ai feedback. *arXiv preprint arXiv:2410.21252*.

Yilun Zhao, Lyuhao Chen, Arman Cohan, and Chen Zhao. 2024. Tapera: enhancing faithfulness and interpretability in long-form table qa by content planning and execution-based reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12824–12840.

A Theoretical Analysis

This section provides formal reasoning for the stability and reliability of RioRAG's verifiable reward design.

A.1 Variance reduction by verifiable nugget decomposition

Let $\mathcal{C}(q, \mathcal{D}_q) = \{n_k\}_{k=1}^{M}$ denote the factual checklist, and define the per-nugget binary correctness as

$$Z_k(o) = \mathbf{1}\{n_k \subseteq o\}, \qquad C(o) = \frac{1}{M}\sum_{k=1}^{M} Z_k(o).$$

Assume (A1) each Z_k is an unbiased Bernoulli variable with variance $\sigma_k^2 \leq 1/4$, and pairwise covariances are bounded by $\mathrm{Cov}(Z_i, Z_j) \leq \rho$ (Clarke, 1980). Then

$$\begin{aligned}\mathrm{Var}[C(o)] &= \tfrac{1}{M^2}\left(\sum_{k=1}^{M}\sigma_k^2 + 2\sum_{1\leq i<j\leq M}\mathrm{Cov}(Z_i, Z_j)\right)\\ &\leq \tfrac{\bar{\sigma}^2}{M} + \rho\,\tfrac{M-1}{M}.\end{aligned}$$

where $\bar{\sigma}^2 = \frac{1}{M}\sum_k \sigma_k^2$. When nuggets are nearly independent ($\rho \approx 0$), the variance decays at least as $\mathcal{O}(1/M)$, yielding more stable and concentrated reward estimates than single long-form scoring (Nenkova et al., 2007).

A.2 Lipschitz stability from answer-only length decay

Let the final reward be $R(o) = C(o)\phi(l)$, where $\phi(l) = \exp[-k((l - l_0)/\tau)^m]$ for $l > l_0$ and 1 otherwise. Assume $C(o)$ is L_s-Lipschitz w.r.t. perturbations in the answer tokens, and $\phi(l)$ is L_ϕ-Lipschitz in length (Wasserman, 2013). For two outputs o, o' with answer lengths l, l',

$$\begin{aligned}|R(o) - R(o')| &= \big|C(o)\phi(l) - C(o')\phi(l')\big|\\ &\leq L_s\,\|o - o'\| \;+\; L_\phi\,|l - l'|.\end{aligned}$$

Hence $R(\cdot)$ is Lipschitz continuous and robust to small length variations. Since $\phi(l)$ is smooth and bounded in $[0, 1]$, its Lipschitz constant $L_\phi = \sup_l |\phi'(l)|$ exists under finite output lengths. Because $\phi(l)$ is applied only to the answer segment, the reasoning process remains unaffected, preventing reward drift and ensuring stable optimization.

A.3 Mitigating long-context degradation

Let $p(t)$ denote the position-dependent scoring error probability of a long-form evaluator. For monolithic evaluation, the expected error over T tokens is $E_{\text{mono}} = \frac{1}{T}\sum_{t=1}^{T} p(t)$, which increases when mid-context positions have high $p(t)$ ("lost in the middle" (Liu et al., 2023)). In RioRAG, each nugget corresponds to a short, localized window where the local scoring error $\tilde{p}$ is upper-bounded by $\sup_t p(t_{\text{nugget}})$. Thus,

$$E_{\text{nugget}} = \tfrac{1}{M}\sum_{k=1}^{M}\tilde{\xi}_k, \qquad \mathbb{E}[E_{\text{nugget}}] \leq \tilde{p} \ll \mathbb{E}[E_{\text{mono}}],$$

showing that nugget-level locality effectively suppresses long-context degradation.

Summary. Theoretical analysis shows that our verifiable reward design (i) reduces reward variance with more nuggets, (ii) ensures stable optimization through smooth answer-length decay, and (iii) alleviates long-context degradation via local nugget evaluation.

B Details on Datasets

In this section, we provide detailed descriptions of the two comprehensive benchmarks used in our experiments: **LongFact** (Wei et al., 2024) and **RAGChecker** (Ru et al., 2024). These datasets are designed to evaluate long-form retrieval-augmented generation (RAG) systems across diverse topics and multiple dimensions of factual quality. Their complementary nature enables robust assessment of both factual coverage and fine-grained answer quality in open-domain settings.

LongFact. LongFact is a manually curated benchmark focused on evaluating long-form factuality. It contains a diverse set of fact-seeking questions, where each gold answer synthesizes multiple atomic facts drawn from various evidence sources. The dataset is notable for its broad coverage across 38 fine-grained domains, which are grouped into the following 8 broader categories to support structured evaluation:

- **Science & Nature**: physics, chemistry, biology, astronomy, virology, prehistory
- **Technology & Computing**: computer science, computer security, machine learning, electrical engineering, mathematics

- **Medicine & Psychology**: medicine, clinical knowledge, psychology, psychology checkpoint
- **Law & Politics**: international law, immigration law, U.S. foreign policy, jurisprudence
- **Social Sciences & Culture**: sociology, geography, world religions, moral disputes, philosophy
- **History & Events**: history, 20th-century events, global facts, economics
- **Business & Communication**: business ethics, accounting, marketing, management, public relations
- **Entertainment & Lifestyle**: movies, music, gaming, celebrities, architecture, sports

Each example in LongFact is annotated with atomic information units, enabling precise measurement of factual recall and information density. This makes it especially well-suited for evaluating long-form answers that integrate knowledge from multiple sources.

RAGChecker. RAGChecker is a comprehensive benchmark designed to evaluate long-form Retrieval-Augmented Generation (RAG) systems across diverse domains. It repurposes examples from 10 public datasets, encompassing a total of 4,162 questions. For the 4 subsets we used, we briefly describe their characteristics below:

- **ClapNQ** (Rosenthal et al., 2025): Derived from Natural Questions (NQ), ClapNQ includes long-form answers with grounded gold passages from Wikipedia, focusing on generating cohesive long-form answers from non-contiguous text segments.
- **NovelQA** (Wang et al., 2024): NovelQA is a benchmark designed to evaluate large language models on deep narrative understanding through complex questions based on English novels.
- **FiQA** (Maia et al., 2018): A financial question answering dataset comprising 500 QA pairs, where short answers are extended to long-form using GPT-4, filtered to remove hallucinations.
- **KIWI** (Xu et al., 2024): A dataset of knowledge-intensive writing instructions for answering research questions, comprising 71 QA pairs with long-form answers validated for quality.

C Details on Evaluation Metrics

We adopt two categories of metrics to comprehensively evaluate factual quality and retrieval grounding.

Standard LFQA Metrics. Following prior work (Fan et al., 2019), we use *Fact Recall* (FR) to measure factual completeness—the ratio of atomic facts in the generated response to those in the reference answer—and *Information Density* (ID), defined as the ratio of atomic facts to total response length, reflecting conciseness and informativeness.

RAGChecker Metrics. The **RAGChecker** benchmark (Ru et al., 2024) introduces a suite of advanced metrics: *Faithfulness* (proportion of correct facts supported by retrieved pages), *Relevant Noise Sensitivity* (ratio of incorrect facts appearing in retrieved content), *Irrelevant Noise Sensitivity* (share of correct facts that are irrelevant to retrieval), *Hallucination* (incorrect facts unsupported by any retrieved page), *Self-Knowledge* (correct facts absent from all retrieved content), and *Context Utilization* (fraction of ground-truth facts covered by retrieval). Together, these metrics provide a multi-dimensional evaluation of factual grounding, reliability, and retrieval efficiency.