# Reinforced Informativeness Optimization for Long-Form Retrieval-Augmented Generation

**Yuhao Wang**[1,3*] **Ruiyang Ren**[1,3†] **Yucheng Wang**[2] **Wayne Xin Zhao**[1,3†]
**Jing Liu**[2†] **Hua Wu**[2] **Haifeng Wang**[2]
[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]Baidu Inc.
[3]Beijing Key Laboratory of Research on Large Models and Intelligent Governance
{yh.wang500, reyon_ren}@outlook.com, batmanfly@gmail.com

## Abstract

Long-form question answering (LFQA) requires open-ended long-form responses that synthesize coherent, factually grounded content from multi-source evidence. This makes reinforcement learning (RL) reward design critical. The reward must be verifiable for faithful grounding and stable optimization. However, many standard rewards assume a unique target with an exact-match notion of correctness, which fits short-form QA and math but breaks in LFQA. As a result, current RAG systems still lack verifiable reward mechanisms, yielding unstable feedback signals and suboptimal optimization outcomes. We propose RioRAG, a framework for reinforced verifiable informativeness optimization. First, it defines informativeness as a measurable and externally verifiable objective for RL. Second, RioRAG uses nugget-centric verification with cross-source checks to enable self-evolution of smaller LLMs and to provide denser, action-discriminative rewards that mitigate reward sparsity and stabilize optimization. This formulation avoids handcrafted supervision for the policy model and strong teacher-model distillation, relying instead on externally verifiable feedback. 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://github.com/RUCAIBox/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 (Ren et al., 2023; Wang et al., 2024b). 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 (Zhao et al., 2023).

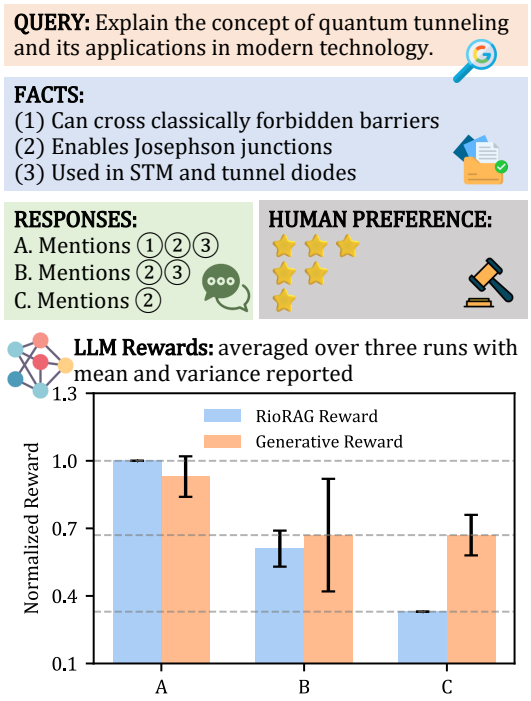



Figure 1: Existing long-form reward modeling exhibits (i) instability, with high variance when evaluating the same response multiple times, and (ii) unreliability, as their score rankings often differ from human judgments across responses.

While RAG mitigates hallucination in long-form generation by grounding models on retrieved evidence (Ren et al., 2025b; Li et al., 2025a; Wang et al., 2025), several challenges remain unresolved. First, LLMs often misuse factual information (Gao et al., 2023; Wang et al., 2026), either due to resid-

[*] The work was done during the internship at Baidu.
[†] Corresponding authors.



*Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13393–13406
July 2-7, 2026 

ual hallucination or failure to extract the correct evidence from lengthy retrieved documents. Besides, generated answers frequently exhibit limited informativeness (Ren et al., 2025a), 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 (Tang et al., 2025a), 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 improved LLMs in math and short-form QA by optimizing outcome-based feedback (Li et al., 2025b; Tang et al., 2025b; Zhan et al., 2026). However, RL for LFQA is bottlenecked by reward design. Many standard rewards assume an exact-match notion of correctness, which does not apply to open-ended long-form answers. Rule-based outcome reward modelings (ORMs) are difficult to specify for diverse responses (Guo et al., 2025; Li et al., 2025a). Moreover, long-form scoring is often unstable and hard to verify, even under well-guided evaluation protocols (Zhang et al., 2024). Reward models can also miss key evidence under long references, yielding misleading signals (Figure 1). Such instability can effectively sparsify the reward signal, further suppressing self-exploration and self-evolution during RL (Guo et al., 2025).

To address these challenges, we propose RioRAG, a Reinforced Informativeness Optimization-based RL method for long-form RAG. RioRAG aims to improve factual coverage with stable, verifiable reward learning. First, it defines informativeness as an externally verifiable objective and derives rewards from evidence support rather than heuristic lexical rules. Second, it employs a nugget-centric hierarchical verifier with length-adaptive scoring to provide fine-grained and action-discriminative feedback. This design mitigates reward sparsity and enables self-evolution of smaller policy models, without relying on handcrafted policy supervision or strong teacher-model distillation (*e.g.,* sequential SFT). 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 Task Formulation

In this work, we define *informativeness* not as response length or stylistic richness, but as **information coverage under the RAG setting**: how completely an answer covers the distilled, evidence-grounded key points supported by retrieved sources. We refer to these fine-grained factual units as *nuggets*, following prior nugget-based grounded evaluation (Łajewska and Balog, 2025). Concretely, a nugget is a concise, reusable unit of information, such as a short fact, claim, or QA-style statement, that can be explicitly verified against source content and used to assess whether key information is covered.

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 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.

## 3 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 pro-

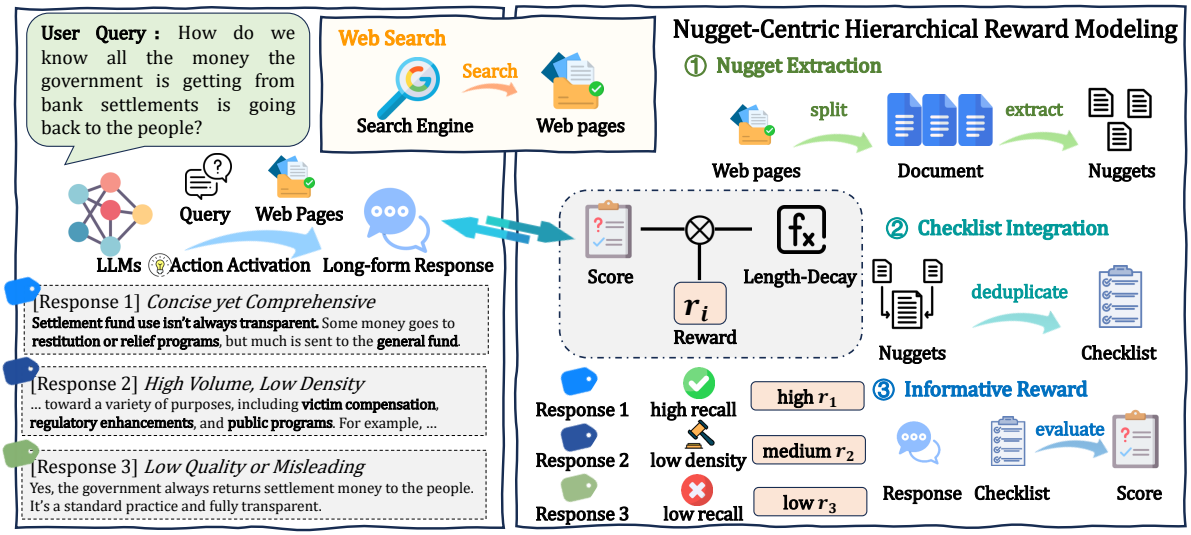



Figure 2: Overall illustration of the proposed RL-based RioRAG framework.

vides fine-grained, verifiable feedback based on evidence-level nuggets. Together, these components enable stable and unsupervised optimization of long-form RAG systems.

### 3.1 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; Li et al., 2026; Liu et al., 2026), 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 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., 2025), 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.

### 3.2 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 $d_i \in \mathcal{D}_q$, we identify *factual nuggets* as concise evidence statements. Each nugget $n_{ij}$ is extracted as a short 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:

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

| Setting | Runtime | Complexity |
|---|---|---|
| Generative Reward | 1.19 | $\mathcal{O}((q+nd)^2)$ |
| RioRAG sequential | 2.32 | $\mathcal{O}(n(q+d)^2 + q^2)$ |
| RioRAG w/ parallel | 0.98 | – |
| RioRAG w/ async | 0.72 | – |

Table 1: Reward computation cost comparison. Runtime is measured in seconds per training sample. Here $q$ is the query length and $d$ is document length.

where $s_i$ is the base informativeness score, $l$ denotes the answer length, $l_0$ is the threshold of unpenalized length, $\tau$ 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.

### 3.3 Discussion

**Novelty.** RioRAG's novelty comes from turning open-ended LFQA evaluation into explicit, verifiable reward computation. (1) From vague holistic scoring to checklist credit assignment. Instead of asking a judge model to output a coarse overall score under ambiguous rules, RioRAG converts informativeness into nugget-aligned checklist items. Each item becomes a concrete scoring point with clear credit. (2) From long-context judging to short-context verification. RioRAG first condenses retrieved evidence into nugget-aligned checklists. It then compares the model output against the checklists and aggregates item-wise scores into the final reward. This avoids long-context degradation, yields denser feedback, and supports self-evolution without strong teacher-model distillation (*e.g.,* sequential SFT).

**Efficiency.** Although RioRAG adds checklist extraction and nugget-based evaluation, these steps operate on compact inputs (single documents or short checklists), keeping token counts low. As a result, RioRAG is comparable to generative reward baselines in the sequential setting, and becomes faster with parallel execution. Reward computation can also be executed asynchronously and pipelined with RL training, so it does not block gradient updates in practice.

**Relation to Recent Reward Design.** Recent studies have explored rubric- or checklist-based reward design for long-form generation and related RL settings (Gunjal et al.). RioRAG is complementary to this line of work, but is specifically designed for long-form RAG, where reward reliability is challenged by lengthy retrieved contexts and multi-source evidence integration. Our reward construction first extracts factual nuggets from retrieved documents in a decomposed manner and then merges them into a unified checklist, reducing the effective context length for verification and improving reward reliability under long contexts. We also distinguish RioRAG from work that uses retrieval mainly as a verifier during training rather than as the inference-time generation setting (Chen et al., 2025b,a). In contrast, our setting focuses on long-form QA with RAG at inference time, where the model must synthesize answers from retrieved multi-source evidence during generation.

# 4 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.

## 4.1 Experimental Setup

### 4.1.1 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.

### 4.1.2 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. Importantly, these RAGChecker metrics are used only for evaluation and are not provided to the model in any form during training, ensuring an objective assessment. Full metric definitions are presented in Appendix C.

### 4.1.3 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-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.

## 4.2 Main Results

The results of different methods evaluated on LongFact and RAGChecker are shown in Table 2 and Table 3. 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

| 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 2: 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 3: 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.

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.

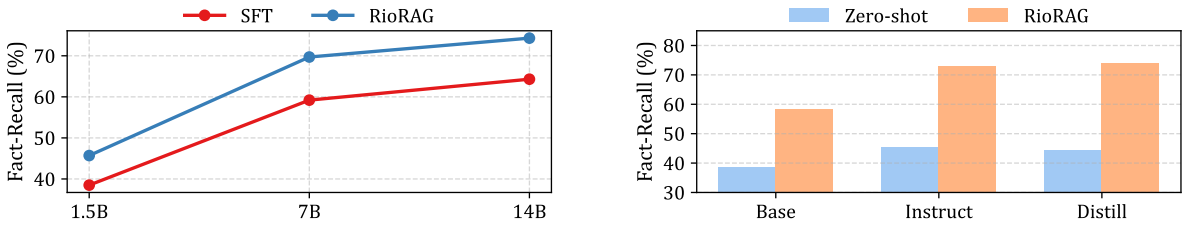
### 4.3 Ablation Studies

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 GRPO* 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 4 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

| 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 GRPO | 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 4: 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.

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.

## 4.4 Further Analysis

### 4.4.1 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 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.

### 4.4.2 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., 2025). 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.

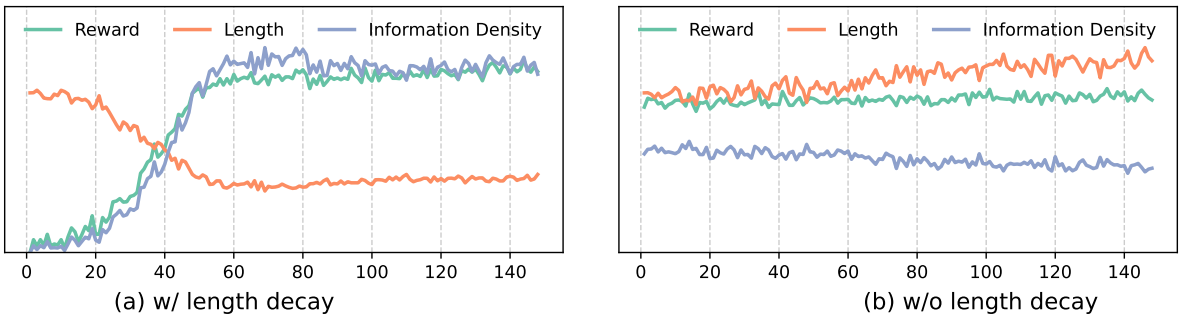



Figure 4: Analysis on co-evolution of generation length and reward during RL training.

#### 4.4.3 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. 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.

## 5 Related Work

### 5.1 Long-Form Question Answering

Research on LFQA has progressed through three shifts: fine-tuned generative models, retrieval-augmented architectures, and human-aligned LLMs. Early abstractive LFQA was enabled by datasets such as ELI5 (Fan et al., 2019), showing that seq2seq models can generate plausible long answers with retrieved evidence. More recent systems leverage LLMs with human feedback. Often via RL from human preferences. To produce answers grounded in explicit quotations (Menick et al., 2022). RAG has since become the dominant approach for factual grounding, including training models in web-browsing interfaces (Nakano et al., 2021; Qin et al., 2023). Verifiability can be strengthened by training with explicit source citation (Menick et al., 2022), while post-hoc attribution verifies pre-generated text (Wang et al., 2022). Given LFQA's open-ended nature, faithfulness to evidence is a central objective (Gao et al., 2023; Zhao et al., 2024). It can be improved through open-book training with citations (Menick et al., 2022) or modeled via probabilistic calibration of answer correctness (Huang et al., 2024). Complementarily, LLM-based evaluators provide automated quality assessment for LFQA (Han et al., 2024).

### 5.2 Reinforcement Learning based RAG

RL has emerged as a promising tool for improving RAG by optimizing retrieval and generation with reward-driven policy updates. Early work strengthens behavior-cloned web-browsing agents with human-feedback rewards to improve factual alignment (Nakano et al., 2021). Subsequent studies design fine-grained rewards for coherence and information gain (Cai et al., 2024), or build domain-specific reward models with synthetic supervision to align RAG with human performance (Nguyen et al., 2024). Others use composite reward ensembles to balance answer quality and coverage (Zhang et al., 2025). On the retrieval side, modules can be optimized via group-wise relative policy optimization (Huang et al., 2025) or multi-agent coordination (Chen et al., 2025c). Cost-sensitive retrieval further uses value estimation to decide when to invoke external search under latency–utility trade-offs (Kulkarni et al., 2024). Recent progress, exemplified by DeepSeek-R1 (Guo et al., 2025), has also motivated RL for autonomous retrieval invocation within long reasoning chains (Jin et al., 2025; Tang et al., 2026). In contrast, RioRAG introduces nugget-centric hierarchical reward modeling to optimize verifiable informativeness for long-form RAG, without handcrafted policy supervision or strong teacher-model distillation.

## 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.

We used AI assistants for minor language polishing; all technical content and results were produced and verified by the authors. Potential risks may arise from imperfect automatic nugget extraction and verification, especially in high-stakes domains.

## Acknowledgments

This work was partially supported by the National Natural Science Foundation of China No. 92470205 and Beijing Major Science and Technology Project under Contract No. Z251100008425002.

## References

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.

Tong Chen, Akari Asai, Luke Zettlemoyer, Hannaneh Hajishirzi, and Faeze Brahman. 2025a. Train for truth, keep the skills: Binary retrieval-augmented reward mitigates hallucinations. *arXiv preprint arXiv:2510.17733*.

Xilun Chen, Ilia Kulikov, Vincent-Pierre Berges, Barlas Oğuz, Rulin Shao, Gargi Ghosh, Jason Weston, and Wen-tau Yih. 2025b. Learning to reason for factuality. *arXiv preprint arXiv:2508.05618*.

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

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.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. 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.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains, 2025. *URL https://arxiv. org/abs/2507.17746*.

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

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. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2723–2727.

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

Yifan Li, Zhenghao Chen, Ziheng Wu, Kun Zhou, Ruipu Luo, Can Zhang, Zhentao He, Yufei Zhan, Wayne Xin Zhao, and Minghui Qiu. 2025b. Unleashing perception-time scaling to multimodal reasoning models. *CoRR*, abs/2510.08964.

Yifan Li, Yukai Gu, Yingqian Min, Zikang Liu, Yifan Du, Kun Zhou, Min Yang, Wayne Xin Zhao, and Minghui Qiu. 2025c. Beyond the last frame: Process-aware evaluation for generative video reasoning. *arXiv preprint arXiv:2512.24952*.

Yifan Li, Kun Zhou, Xin Zhao, Lei Fang, and Jirong Wen. 2026. Analyzing and mitigating object hallucination: A training bias perspective. In *AAAI*, pages 6636–6643. AAAI Press.

Yurou Liu, Mingyang Li, Xinyuan Zhu, Rui Jiao, Yiming Dong, Xinyu Tang, Yang Liu, Jieping Ye, Bing Su, and Zheng Wang. 2026. Drugtrail: Explainable drug discovery via structured reasoning and druggability-tailored preference optimization. In *The Fourteenth International Conference on Learning Representations*.

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*.

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.

Ruiyang Ren, Yuhao Wang, Junyi Li, Jinhao Jiang, Wayne Xin Zhao, Wenjie Wang, and Tat-Seng Chua. 2025a. Llm-based search assistant with holistically guided mcts for intricate information seeking. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1098–1108.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*.

Ruiyang Ren, Yuhao Wang, Kun Zhou, Wayne Xin Zhao, Wenjie Wang, Jing Liu, Ji-Rong Wen, and Tat-Seng Chua. 2025b. Self-calibrated listwise reranking with large language models. In *Proceedings of the ACM on Web Conference 2025*, pages 3692–3701.

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*.

Xinyu Tang, Xiaolei Wang, Zhihao Lv, Yingqian Min, Xin Zhao, Binbin Hu, Ziqi Liu, and Zhiqiang Zhang. 2025a. Unlocking general long chain-of-thought reasoning capabilities of large language models via representation engineering. In *ACL (1)*, pages 6832–6849. Association for Computational Linguistics.

Xinyu Tang, Yuliang Zhan, Zhixun Li, Wayne Xin Zhao, Zhenduo Zhang, Zujie Wen, Zhiqiang Zhang, and Jun Zhou. 2025b. Rethinking sample polarity in reinforcement learning with verifiable rewards. *CoRR*, abs/2512.21625.

Xinyu Tang, Zhenduo Zhang, Yurou Liu, Xin Zhao, zujie wen, Zhiqiang Zhang, and JUN ZHOU. 2026. Towards high data efficiency in reinforcement learning with verifiable reward. In *The Fourteenth International Conference on Learning Representations*.

Cunxiang Wang, Ruoxi Ning, Boqi Pan, Tonghui Wu, Qipeng Guo, Cheng Deng, Guangsheng Bao, Qian Wang, and Yue Zhang. 2024a. 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.

Yuhao Wang, Ruiyang Ren, Junyi Li, Wayne Xin Zhao, Jing Liu, and Ji-Rong Wen. 2024b. Rear: A relevance-aware retrieval-augmented framework for open-domain question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5613–5626.

Yuhao Wang, Ruiyang Ren, Yucheng Wang, Jing Liu, Xin Zhao, Hua Wu, and Haifeng Wang. 2026. Beerag: Balanced entropy engineering for retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 33737–33745.

Yuhao Wang, Ruiyang Ren, Yucheng Wang, Wayne Xin Zhao, Jing Liu, Hua Wu, and Haifeng Wang. 2025. Unveiling knowledge utilization mechanisms in llm-based retrieval-augmented generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1262–1271.

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.

Yu-Liang Zhan, Zhong-Yi Lu, Hao Sun, and Ze-Feng Gao. 2024. Over-parameterized student model via tensor decomposition boosted knowledge distillation. In *NeurIPS*.

Yu-Liang Zhan, Xinyu Tang, Han Wan, Jian Li, Jirong Wen, and Hao Sun. 2026. L2v-cot: Cross-modal transfer of chain-of-thought reasoning via latent intervention. In *AAAI*, pages 12358–12366. AAAI Press.

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*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2):1–124.

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 Implementation Details

To improve reproducibility, we provide the key training configuration of RioRAG in this appendix, and include the full prompt templates used in the "Extract" and "MergeCluster" stages of our reward construction pipeline.

## A.1 Training Hyperparameters

Following previous approaches (Li et al., 2025c; Zhan et al., 2024), We use a maximum prompt length of 8192 and a maximum response length of 4096. The training batch size is set to 64 with dynamic batch sizing enabled, and the maximum PPO token length per GPU is 12288. For optimization, we apply gradient clipping with a threshold of 1.0, use a clip ratio of 0.2, and set the entropy coefficient to 0.001. We enable KL regularization with `use_kl_loss=True`, using `low_var_kl` as the KL loss type and a KL coefficient of 0.04. PPO is run for one epoch without shuffling. We set the Ulysses sequence parallel size to 2. The learning rate is $1 \times 10^{-6}$, with a warmup ratio of 0.1 and cosine warmup scheduling. We will also release the full configuration file in our codebase. In the final version, we additionally report the hardware setup and random seed configuration explicitly.

## A.2 Details of Prompts

In this appendix, we provide the exact prompt templates used in the "Extract" and "MergeCluster" stages of our reward construction pipeline. These prompts are used to convert retrieved web documents into concise factual nuggets and then merge them into a compact checklist for subsequent verification.

**Extract Prompt.** The "Extract" stage processes each retrieved web document independently. Given a user query and a retrieved web content, it outputs several highly relevant key points grounded only in the given content.

**MergeCluster Prompt.** After applying "Extract" to all retrieved documents, we obtain a list of candidate key points. The "MergeCluster" stage filters irrelevant items, removes exact duplicates, preserves complementary points with different focuses, and merges them into a final checklist.

**Usage in Reward Construction.** The output of "Extract" is a document-level set of candidate nuggets, while the output of "MergeCluster" is a query-level checklist used for informativeness assessment. This decomposition reduces the effective verification context length and provides a more compact and structured basis for reward computation.

**Extract Prompt**

You are given a user query and a retrieved web content.

Your task:
- Output several highly relevant key points from the web content.
- Each key point must be one concise sentence and separated by a newline.
- The key points must be wrapped inside \boxed{
Key Point 1
Key Point 2
...
}.
- Only use the given web content. - If no relevant information is found, output \boxed{No relevant information}.
- Do not create or assume content not present in the web content.

Input:
- Query: {query}
- Web Content:
{web_content}

- Query: {query}

# 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:

**MergeCluster Prompt**

You are given a user question and a list of candidate key points.

Your task:
- Keep only the key points that are highly relevant to the question.
- Merge exact duplicates; if two points have slightly different focuses, keep both.
- Each item = one single idea, in one concise sentence.
- If there are conflicting views, use wording like: "Some studies suggest [...], others indicate [...]."
- Output format: \boxed{
key point 1
key point 2
key point 3
...
}
- Only output inside \boxed{}.

Output Format:
\boxed{
Final Key Point 1
Final Key Point 2
...
}

Input:
- Query: {query}
- Key Points: {key_points}

- Query: {query}

- *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., 2024a): 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.