Controlled Retrieval-augmented Context Evaluation for Long-form RAG

Jia-Huei Ju[1] **Suzan Verberne**[2] **Maarten de Rijke**[1] **Andrew Yates**[3]
[1]University of Amsterdam [2]Leiden University
[3]Johns Hopkins University, HLTCOE
{j.ju, m.derijke}@uva.nl, s.verberne@liacs.leidenuniv.nl,
andrew.yates@jhu.edu

Abstract

Retrieval-augmented generation (RAG) enhances large language models by incorporating context retrieved from external knowledge sources. While the effectiveness of the retrieval module is typically evaluated with relevance-based ranking metrics, such metrics may be insufficient to reflect the retrieval's impact on the final RAG result, especially in long-form generation scenarios. We argue that providing a comprehensive retrieval-augmented context is important for long-form RAG tasks like report generation and propose metrics for assessing the context independent of generation. We introduce CRUX, a **C**ontrolled **R**etrieval-a**U**gmented conte**X**t evaluation framework designed to directly assess retrieval-augmented contexts. This framework uses human-written summaries to control the information scope of knowledge, enabling us to measure how well the context covers information essential for long-form generation. CRUX uses question-based evaluation to assess RAG's retrieval in a fine-grained manner. Empirical results show that CRUX offers more reflective and diagnostic evaluation. Our findings also reveal substantial room for improvement in current retrieval methods, pointing to promising directions for advancing RAG's retrieval. Our data and code are publicly available to support and advance future research on retrieval for RAG.[1]

1 Introduction

With their emerging instruction-following capabilities (Ouyang et al., 2022; Wei et al., 2021), large language models (LLMs) have adopted retrieval-augmented generation (RAG) (Lewis et al., 2020; Guu et al., 2020) to tackle more challenging tasks, such as ambiguous question answering (QA) (Stelmakh et al., 2022; Gao et al., 2023) and long-form response generation (Shao et al., 2024). The role of retrieval in RAG is to access information from external sources and prompt it as plug-in knowledge for LLMs. To achieve this, typical RAG systems retrieve the k most relevant chunks as the retrieval-augmented context (abbreviated as *retrieval context*, hereafter), and prompt the LLM to generate a response using this information.

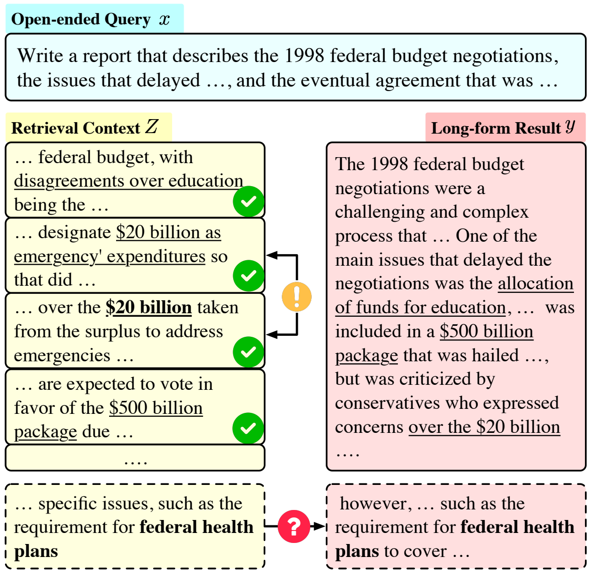

Figure 1: An example of long-form generation with an open-ended query x and a desired response y. The underlined text marks relevant content in the retrieval (✅) that contributes to the final result. By directly assessing the retrieval context Z, we can further explicitly identify incomplete (❓) and redundant retrieval (❗).

A suboptimal retrieval context hinders the generation process (Asai et al., 2024; Rau et al., 2024), triggering negative impacts and resulting in unsatisfying final RAG results. One widely-studied effect is the impact of noise from irrelevant retrieval (Yoran et al., 2023), which increases the risk of hallucinations (Asai et al., 2022) and distractions (Shi et al., 2023). Such prior studies have mainly focused on short-answer tasks; however, recent RAG research has shifted towards generating comprehensive and structured reports with open-ended queries (Zhao et al., 2024; Lawrie et al., 2024), as illustrated in Figure 1, introducing new

[1] https://github.com/DylanJoo/crux

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 21102–21121

concerns of suboptimal retrieval.

In the scenario of open-ended queries where a short answer is insufficient and a long-form result is required, incompleteness and redundancy emerge as the critical yet underexplored negative impacts from retrieval (Joren et al., 2024). Specifically, (i) *incomplete* retrieval fails to capture the full nuance of the query, leading to partial or misleading generations, and (ii) *redundant* retrieval contexts restrict the diversity of knowledge, undermining the usefulness of augmented knowledge (Yu et al., 2024; Chen and Choi, 2024). Figure 1 exemplifies such impacts of suboptimal retrieval matters on the final long-form RAG result.

To examine these effects, a suitable retrieval evaluation framework is crucial for measuring completeness and redundancy in the retrieval context. Current retrieval evaluation practices are insufficient for measuring retrieval effectiveness in long-form RAG, as they are designed for web search (Bajaj et al., 2016) or short-answer QA (Kwiatkowski et al., 2019). They only require a focus on relevance-based ranking, which can be simply evaluated with retrieval metrics such as MRR and Recall@k. In contrast, long-form RAG requires retrieving multiple aspects and subtopics to ensure completeness, which goes beyond surface-level relevance (Tan et al., 2024; Grusky et al., 2018).

To address the gap, we propose a **C**ontrollable **R**etrieval-a**U**gmented conte**X**t evaluation framework (CRUX). The framework includes controlled evaluation datasets and uses coverage-based metrics that directly assess the content of the retrieval context instead of relevance-based ranking. We use human-written multi-document summaries to define the scope of the retrieval context, enabling a controlled oracle retrieval for more diagnostic evaluation results. Finally, we assess both the (intermediate) retrieval context and (final) RAG result via question-based evaluation (Sander and Dietz, 2021; Dietz, 2024), supporting fine-grained evaluation between them.

To validate the usability of our evaluation framework, we conduct empirical experiments with multiple retrieval and re-ranking strategies, including relevance and diversity re-ranking. Empirical results demonstrate the limitations of suboptimal retrieval in terms of coverage and density. Our additional metric analysis further demonstrates that relevance ranking metrics lack coverage-awareness, highlighting CRUX's strength in identifying retrieval impacts on long-form RAG. Notably, our framework balances scalability and reliability by integrating LLM-based judgments with human-grounded data. Our final human evaluation also confirms CRUX's alignment with human perception.

Overall, our controlled retrieval context evaluation aims to identify suboptimal retrieval for long-form RAG scenario. Our contributions are as follows:

- We create a controlled dataset tailored for evaluating the retrieval context for long-form RAG;
- We propose coverage-based metrics with upper bounds to help diagnosing the retrieval context in terms of completeness and redundancy;
- Our empirical results showcase the limitations of existing retrieval for long-form RAG;
- Our framework can serve as a reliable experimental testbed for developing more compatible retrieval for long-form RAG.

2 Related Work

The importance of retrieval in RAG. LLMs are highly effective at parameterizing world knowledge as memory; however, accessing long-tail knowledge (Mallen et al., 2023) or verifying facts (Mishra et al., 2024; Min et al., 2023) often requires retrieving information from external sources. This highlights the essential role of retrieval in augmenting reliable knowledge for downstream applications (Zhang et al., 2024; Zhu et al., 2024; Rau et al., 2024), which is especially important in long-form generation (Gao et al., 2023; Mayfield et al., 2024; Tan et al., 2024). Many studies point out that the limitations of retrieval lead to unsatisfying RAG results (BehnamGhader et al., 2023; Su et al., 2024; Asai et al., 2024; Rau et al., 2024), raising the critical question: *how effectively can retrieval augment knowledge for LLMs?*

Automatic evaluators for NLP tasks. LLMs have shown promising instruction-following capability, making them increasingly common as automatic evaluators across various NLP tasks (Thakur et al., 2025; Zheng et al., 2023; Chiang and Lee, 2023). Due to their cost efficiency and scalability, LLM-based evaluations have also been applied to information retrieval (IR) (Thomas et al., 2024; Dietz, 2024) and short-form generation tasks (Saad-Falcon et al., 2024; Shahul et al., 2023). Instead of short-form RAG, we target long-form generation with open-ended query, which requires retrieval to ensure completeness in addition to surface-level relevance. Reference-based metrics like ROUGE used in summarization also fall short in such scenarios (Krishna et al., 2021). Thus, a flexible framework is needed to assess information completeness and redundancy in the retrieval context.

Evaluating retrieval for long-form generation. Evaluation methodologies in IR and NLP have been standardized and developed for decades (Voorhees, 2002, 2004). In recent years, nugget-based (subtopics or sub-questions) evaluation (Pavlu et al., 2012; Clarke et al., 2008; Dang et al., 2008) has resurfaced as an important focus due to the feasibility of automatic judgments. Similarly, question-based evaluation that estimates the answerability (Eyal et al., 2019; Sander and Dietz, 2021) of a given text is well-aligned with LLMs while preserving aspect-level granularity, making it particularly good for evaluating long-form generation. This helps inform the development of a unified evaluation setup for both the intermediate retrieval context and final long-form results, thereby facilitating more informative evaluation for RAG's retrieval methods.

3 Controlled Retrieval-augmented Context Evaluation (CRUX)

This section introduces CRUX, a controlled evaluation framework for assessing the retrieval context in long-form RAG. It comprises: (1) definitions of *retrieval context* and its sub-question *answerability* (§ 3.1); (2) curated evaluation datasets (§ 3.2) and (3) *answerability*-driven performance metrics: coverage and density (§ 3.3).

3.1 Retrieval-augmented Context

Here we focus on the retrieval context as the important bottleneck in the long-form RAG pipeline. Formally, given an open-ended query x, a typical RAG pipeline is defined as:

$$y \leftarrow G(x, Z, I), \quad Z \leftarrow RA_\theta(x, \mathcal{K}). \quad (1)$$

RA_θ denotes the retrieval modules that augment the retrieval context Z from an external knowledge source $\mathcal{K}$ (i.e., a corpus), and G is a LLM generator that takes as input a query x, retrieval context Z, and task-specific instruction prompt I to generate the final long-form RAG result y. Particularly, we argue that the quality of the retrieval context is a key limitation for achieving optimal RAG results and propose an evaluation framework for it.

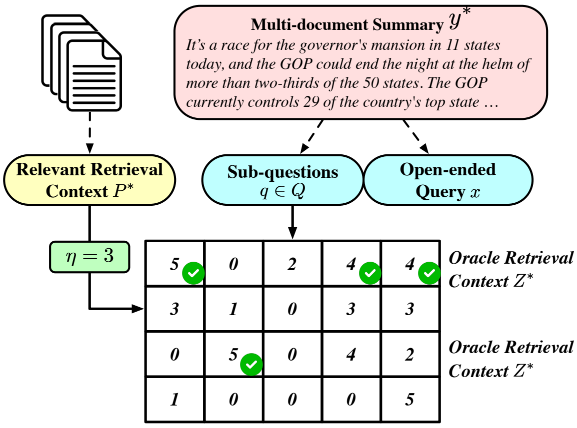

Figure 2: The controlled data generation derived from multi-document summarization datasets.

Answerability measured by sub-questions. To assess the retrieval context quality beyond relevance-based ranking, we adopt question-based evaluation (Eyal et al., 2019; Sander and Dietz, 2021). We assess the content of an arbitrary text z with a diverse set of knowledge-intensive sub-questions $Q = \{q_1, q_2, \ldots, q_n\}$. Such diversity enables these questions to serve as a surrogate for evaluating multiple aspects of a query, thereby facilitating explicit diagnosis of underlying concerns such as completeness and redundancy. Specifically, we use an LLM to judge how well the text z answers each sub-question and estimate a binary sub-question *answerability* value (answerability, hereafter):

$$G(z, q_i, I_g) \geq \eta \quad \forall q_i \in Q, \quad (2)$$

where I_g is a grading instruction prompt similar to the rubrics proposed by Dietz (2024). The output graded rating is on a scale of 0 to 5 (the prompt is included in Figure 8 in the Appendix A.1). η is a predefined threshold determining whether the given text-question pair is answerable. The threshold analysis is reported in Section 4.4.

3.2 Data Creation for Controlled Evaluation

We further construct datasets tailored for our evaluation framework to support controlled analysis. As illustrated in Figure 2, we treat human-written multi-document summaries as the central anchor for defining: (1) the explicit scope of the relevant retrieval context Z^*; (2) an open-ended query x; (3) a diverse set of sub-questions Q. Together, these components support our assessment of completeness and redundancy.

Explicit scope of the retrieval context. The controllability comes from the intrinsic relationships within the multi-document summarization datasets: Multi-News (Fabbri et al., 2019) and DUC (Over and Yen, 2004), where each example consists of a human-written summary and the corresponding multiple documents. As illustrated in Figure 2, we consider the human-written summary as the proxy of an oracle long-form RAG result;[2] it is denoted as y^*. The corresponding documents D^* are naturally regarded as relevant, while the other documents can be safely considered as irrelevant, forming an explicit scope for each example. In addition, we decontextualize a document into passage-level chunks with an LLM, obtaining the set of relevant passages $p \in P^* \subseteq D^*$. Decontextualization provides several advantages (Choi et al., 2021), ensuring the passages fit the token length limitation of all retrievers and are standalone while preserving main topics. Such units also help us identifying redundancy and incompleteness; see Table 5 for an example.

Open-ended queries. We use an LLM to synthesize a query with open-ended information needs from the human-written summary y^* via in-context prompting (Brown et al., 2020). Examples are shown in Figures 1 and 9. We denote these queries as x in Eq. (1), which is the initial input for both retrieval and generation. Such queries help expose limitations in existing retrieval systems, which often return either irrelevant or redundant passages, resulting in incomplete retrieval contexts. Notably, the query generation process is adaptable and can be tailored to various kinds of queries (Yang et al., 2024) via similar in-context prompting.

Diverse sub-questions and filtering. Similarly, we synthesize a diverse set of knowledge-intensive sub-questions Q from the human-written summary which cover the highlights in the oracle RAG results (i.e., y^*). Thanks to the controlled settings, for each query x, we enumerate all possible pairs of sub-questions $q \in Q$ and relevant passages $p \in P^*$, then judge them with an LLM. Hence, for each relevant passage, we obtain a list of graded ratings for all the sub-question as mentioned in Eq. (2). Finally, we can obtain the matrix of graded ratings as shown in Figure 2. In addition, the judged ratings can serve as consistency filtering to identify unanswerable sub-questions for mitigating out-of-scope and hallucinated questions. These pre-judged ratings can be further reused for evaluating the retrieval context, which is also released with the data.

[2] We assume the human-written summary satisfies complex information needs in the most precise and concise manner.

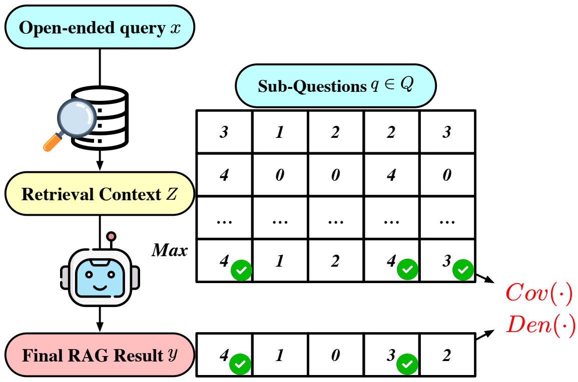

Figure 3: CRUX employs sub-question answerability to directly assess the textual content of both the retrieval context Z and its corresponding RAG result y. The metrics include *coverage* and *density*.

Required subset of relevant passages. Once we have pre-judgments of all relevant passages $p \in P^*$, we further identify which passages are necessary and construct a smaller subset of relevant passages, denoted as P^{**}. Specifically, we define this required subset as those passages that can collectively answer all sub-questions $q \in Q$. To do so, we first rank each relevant passage according to how many questions it can answer and greedily assign each to the subset until no additional sub-questions can be answered.[3] Those remaining are categorized as either partially or fully redundant.

Data statistics. We collected 100 open-ended queries from Multi-News (Fabbri et al., 2019) and 50 queries from DUC (Over and Yen, 2004). The knowledge source $\mathcal{K}$ has around 500K passages, collected from training and test splits of Multi-News and the DUC. We generate all data using open-source `Llama-3.1-70B-Instruct`.[4] (Meta, 2024). Detailed data statistics and generation settings are reported in Appendix A.1.

3.3 Evaluation Metrics

We use three metrics to assess the retrieval context for long-form RAG. We begin by measuring a retrieval context's completeness using *coverage*, then introduce derived metrics: *ranked coverage* and *density* to further take redundancy into account.

Coverage (Cov)**.** Rather than evaluating the retrieval results based on only their relevance (e.g., nDCG and MAP), we assess the content of the retrieval contexts based on *answerability*. Given a retrieval context Z, we explicitly quantify the con-

[3] The default answerability threshold η is set to 3.

[4] `https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct`

text's coverage with how many questions it can answer over the answerable sub-questions. To compute this, we aggregate graded ratings by taking the maximum across passages in the retrieval context Z and obtain binary answerability as depicted in Figure 3. We finally normalize it by the total number of answerable sub-questions. Formally, the coverage of the retrieval context is defined as:

$$Cov(Z) = \frac{\#\{q \in Q |\max\Big(G(p \in Z, q, I_g)\Big) \geq \eta\}}{|Q|}. \quad (3)$$

We can also apply this formula to evaluate the coverage of the final RAG result y, allowing us to compare the coverage of the retrieved passages to the coverage of the generation.

Ranked coverage. We bring coverage-awareness to the novelty ranking metric, α-nDCG (Clarke et al., 2008). α-nDCG evaluates novelty based on subtopics, which is naturally compatible with our framework using sub-question answerability. Specifically, we define the *ranked coverage* by treating the answerability of sub-questions as subtopics, as follows:

$$\alpha\text{-nDCG} = \sum_{r=1}^{|Z|} \frac{ng(r)}{\log(r+1)} \Big/ \sum_{r=1}^{|Z^*|} \frac{ng^*(r)}{\log(r+1)} \quad (4)$$

$$ng(r) = \sum_{i=1}^{|Q|} \mathcal{I}_{i,r}(1-\alpha)^{c_{i,r-1}} \quad (5)$$

where r is the passage rank position in the retrieval context. The function ng is novelty gain, representing how much new information is covered with respect to the position r and sub-questions q_i. Discount factor α is used for penalizing redundant sub-questions when accumulating gains.

Density (Den)**.** We evaluate the retrieval context's density from a coverage perspective. The oracle retrieval context Z^* is considered as the reference, enabling us to compute relative density based on the total number of tokens. The density of the retrieval context Z is measured by:

$$Den(Z) = \Big(\frac{Cov(Z)/\text{token}(p \in Z)}{Cov(Z^*)/\text{token}(p \in Z^*)}\Big)^w, \quad (6)$$

where $\text{token}(\cdot)$ means the total number of tokens, and w is a weighting factor. We set w as 0.5, assuming that the information density grows monotonically but has diminishing marginal returns when reaching the optimum.

4 Experiments

To validate CRUX's evaluation capability and usability, we begin with controlled experiments with empirical retrieval contexts to enable more diagnostic retrieval evaluation. Next, we analyze metric correlations between the retrieval contexts Z and the corresponding final results y. Finally, we assess CRUX's usability through human annotations and examine other configuration impacts.

4.1 Experimental Setups

Initial retrieval. Our experiments employ varying cascaded retrieval pipelines to augment context from the knowledge corpus. Given an open-ended query x, we first retrieve the top-100 relevant candidate passages. Three initial retrieval approaches are considered: lexical retrieval (**LR**) with BM25,[5] dense retrieval (**DR**) using Contriever$^{\text{FT}}$ (Izacard et al., 2021) and learned sparse retrieval (**LSR**) using SPLADE-v3 (Lassance et al., 2024).

Candidate re-ranking. We further re-rank the 100 candidate passages with more effective models, constructing the final retrieval context Z. We experiment with varying re-ranking strategies, including pointwise re-ranking models: **miniLM** (220M) and **monoT5** (3B). In addition, we include state-of-the-art LLM-based listwise re-ranking models: **RankZephyr** (7B) (Pradeep et al., 2023) and **RankFirst** (7B) (Reddy et al., 2024), as well as **Setwise** re-ranking (3B) (Zhuang et al., 2023). Lastly, we evaluate the maximal marginal relevance (**MMR**) algorithm for diversity re-ranking to consider both relevance and diversity.[6]

Generation. Llama models (Meta, 2024) with 8B parameters are used for generation. We use `vLLM` (Kwon et al., 2023) to accelerate the inference speed and perform batch inference. For fair comparisons, we adopt the same configurations for all generations. Details are provided in Appendix A.1.

Evaluation protocol. As our goal is to analyze how incomplete and redundant retrieval context affects the final RAG result, we assess both the quality of retrieval context Z and further investigate the relationships between them and final coverage and density: $Cov(y)$ and $Den(y)$. Notably, the explicit scope of relevant passages allows us to

[5] https://github.com/castorini/pyserini/

[6] We follow Gao and Zhang (2024) and adopt the same pre-trained encoder for MMR: https://huggingface.co/sentence-transformers/all-mpnet-base-v2

Retrieval Context	DUC					Multi-News				
	$Cov(Z)$	α-nDCG	$Cov(y)$	$Den(Z)$	$Den(y)$	$Cov(Z)$	α-nDCG	$Cov(y)$	$Den(Z)$	$Den(y)$
(#1) Direct prompting	-	-	26.7	-	-	-	-	21.4	-	-
(#2) Oracle result y^*	-	-	95.3	-	108	-	-	94.1	-	111
(#3) Oracle retrieval Z^*	100	80.6	64.6	100	93.8	100	80.6	61.8	100	84.7
BM25 (LR)	44.4	35.7	35.8	61.2	53.4	39.3	35.4	38.2	50.6	60.0
Contriever (DR)	52.1	45.2	41.7	70.3	60.5	43.1	36.6	36.6	55.4	58.3
SPLADE-v3 (LSR)	49.0	45.0	41.0	67.7	59.4	45.4	40.4	41.3	60.6	64.3
LR + MMR	45.6	36.7	36.4	65.8	57.2	41.4	35.2	37.9	52.9	58.9
DR + MMR	42.7	35.1	33.8	62.6	53.5	39.0	33.5	36.1	51.3	57.6
LSR + MMR	44.2	35.6	36.5	64.4	56.5	39.2	33.8	37.3	51.6	59.2
LR + miniLM	49.0	42.5	38.4	67.9	57.9	45.3	39.8	41.2	58.2	63.0
DR + miniLM	49.3	42.9	39.9	69.3	59.7	45.1	40.3	40.4	57.8	62.4
LSR + miniLM	49.4	42.6	39.2	69.3	59.2	45.4	40.3	40.6	58.0	62.6
LR + monoT5	50.7	42.4	37.9	66.5	56.7	47.9	40.2	41.6	58.3	64.0
DR + monoT5	53.2	44.7	40.7	70.8	60.0	45.4	40.0	40.9	56.6	62.6
LSR + monoT5	52.8	43.0	41.1	68.9	59.2	44.3	37.7	38.9	55.4	61.5
LR + RankZephyr	51.5	45.9	40.6	69.9	59.5	52.9	47.6	43.9	65.1	67.7
DR + RankZephyr	51.1	48.8	40.6	67.8	59.2	53.6	47.2	44.1	66.0	66.8
LSR + RankZephyr	50.4	45.9	41.2	67.3	60.0	54.4	**49.1**	45.8	67.0	**69.8**
LR + RankFirst	52.0	46.2	43.9	70.1	63.4	**56.0**	**49.1**	**46.4**	**68.0**	69.4
DR + RankFirst	53.8	**49.1**	**44.6**	70.9	64.0	54.5	47.6	44.4	66.2	67.4
LSR + RankFirst	53.6	48.2	44.3	70.9	64.0	54.5	48.2	46.0	66.5	69.2
LR + SetwiseFlanT5	49.6	44.2	42.5	67.8	61.9	52.1	44.9	43.2	63.9	65.5
DR + SetwiseFlanT5	**56.6**	48.4	44.4	**74.9**	**64.4**	49.9	43.8	41.0	61.0	62.5
LSR + SetwiseFlanT5	51.9	46.0	43.3	70.1	62.6	52.0	47.0	45.1	65.4	67.8
Rank Corr. (Kendall τ)	0.676	0.724	-	0.733	-	0.838	0.800	-	0.810	-

Table 1: Evaluation results of empirical retrieval contexts Z and corresponding final results y (the columns in gray) on CRUX-DUC and Multi-News. Scores with bold font and underlined are the highest and lowest. For each dataset, columns 1 and 2 show retrieval coverage and ranked coverage; column 3 shows the final result coverage. The last two columns are the density of the retrieval context and final result. The bottom row reports the ranking correlation between the retrieval context and final results.

reuse the pre-judgments for relevant passages as shown in Figure 3. Unless otherwise specified, we set the default *answerability* threshold η to 3.

4.2 Controlled Empirical Experiments

CRUX suggests explicit oracle RAG settings of retrieval context Z^*, thereby facilitating more indicative evaluations by controlling: (i) the number of passages in the retrieval context (i.e., top-k), which is set to match the size of the oracle retrieval context, $|Z^*|$; (ii) the maximum generation token length, which is constrained by the match token length of the oracle retrieval, $\text{token}(Z^*)$.[7] The following research questions guide our findings.

What are the reference performance bounds of the retrieval context and final RAG result? In the first block of Table 1, we report the performance of three reference retrieval contexts and their final RAG results: (#1) zero-shot **direct prompting**; (#2) **oracle results** y^* (the human-written summary); (#3) **oracle retrieval context** $Z^* \triangleq P^{**}$, which is the required subset of relevant passages given in the test collection (See Section 3.2).

Unsurprisingly, we observe the lowest coverage for RAG result without retrieval (#1), confirming that parametric knowledge in the LLM alone is insufficient to achieve high performance. This condition serves as the empirical lower bound of RAG. In contrast, the oracle result using the human-written summary (#2) achieves the highest coverage by answering over 90% of the sub-questions. This implies that the generated sub-questions are answerable and validate the framework's ability to capture completeness. The RAG result with the oracle retrieval context (#3) yields decent coverage of 64.6 and 61.8, outperforming other empirical methods in subsequent blocks in the table. This demonstrates an empirical upper bound for RAG's retrieval, grounded in an oracle retrieval context Z^*. Overall, CRUX provides robust bounds for reference, enabling more diagnostic evaluation of RAG's retrieval regardless of the generator.

[7] We change the prompt accordingly and truncate the maximum token length if the result exceeds.

How effective are empirical retrieval contexts regarding the performance of the final RAG result? To investigate this, we evaluate a range of empirical retrieval contexts from various cascaded retrieval pipelines. As reported in Table 1, each pipeline is evaluated with both the quality of the intermediate retrieval context Z and the final RAG result y (the gray columns).

The second and third blocks in Table 1 show that initial retrieval-only and MMR ranking struggle to retrieve useful information, resulting in poor performance of retrieval contexts. We also observe that such suboptimal retrieval contexts would directly reflect on the suboptimal final RAG result coverage $Cov(y)$ on both evaluation sets (underlined scores).

Notably, on evaluation results of DUC, we observe pointwise re-ranking models have robust gains on final RAG result coverage only when used with weaker initial retrieval (e.g., LR + miniLM, 35.8 $\rightarrow$ 38.4). However, they degrade when adopting stronger initial retrieval (e.g., LSR + miniLM, 41.0$\rightarrow$ 39.2). Such patterns are also shown on intermediate retrieval context performance, demonstrating CRUX's evaluation capability for retrieval contexts.

In contrast, more effective re-ranking methods consistently enhances overall performance, with visible performance gains in both intermediate and final results. For example, RankFirst (Reddy et al., 2024) and SetwiseFlanT5 (Zhuang et al., 2023), particularly outperform all the other empirical pipelines (conditions marked in bold). Yet, they still have a large gap compared to the oracle retrieval (#3), implying that existing ranking models are not explicitly optimized for coverage of long-form RAG results.

Can intermediate retrieval context performance extrapolate to final RAG result performance? Finally, to highlight the advantage of retrieval context evaluation, we compute the ranking correlation in terms of Kendall's τ between the final result coverage/density (i.e., $Cov(y)$/$Den(y)$) and the intermediate coverage, ranked coverage and density.

We find ranking correlation strengths of approximately 0.7 to 0.8 on both evaluation sets at the last row in Table 1, demonstrating the strong alignment between the retrieval context and RAG results. This suggests that our framework can be a promising surrogate retrieval evaluation for extrapolating long-form RAG results.

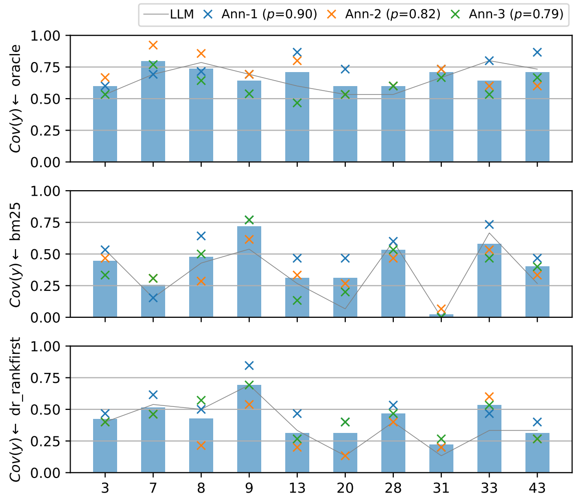

Figure 4: Coverage of RAG results for 10 CRUX-DUC queries (x-axis) under three retrieval contexts (y-axis). Each subplot shows LLM-judged coverage (line) and human judgments (markers); bars indicate the annotators' average. The Spearman correlations ρ are computed between the LLM and each annotator's coverage.

4.3 Metric Alignment Analysis

To further validate our proposed evaluation metrics, we analyze how these metrics align with human judgments. Then, we compare these metrics against other relevance-based metrics, showing that they are insufficient for evaluating retrieval modules in long-form RAG scenarios.

How does the evaluation method align with human judgments? We conduct human judgment on 10 randomly selected open-ended queries from CRUX-DUC. We design two reading comprehension tasks:[8] $\mathcal{T}_1$: *Long-form RAG result coverage judgment*, and $\mathcal{T}_2$: *Rubric-based passage judgment*. $\mathcal{T}_1$ investigates how well LLM-judged coverage align with human's. We collect binary answerability annotations for all enumerated result sub-question pairs $\{(y, q_1), ..., (y, q_n)\}$ and compute the corresponding RAG result's coverage $Cov(y)$.

We evaluate RAG results across three retrieval contexts Z: Oracle, BM25 and DR+RankFirst, as shown in the subplots in Figure 4. With the total of 30 human-judged coverage, we compute the Spearman correlation between them and LLM, obtaining high alignment ($\rho \geq 0.8$), and a moderate inter-annotator agreement (Fleiss' $\kappa = 0.52$). We also found that the controlled oracle retrieval Z^* has significantly better coverage from human judgments, confirming the reliability of upper bound, while the other retrieval context fluctuate among queries.

[8] Appendix A.2 describes the annotation tasks in detail.

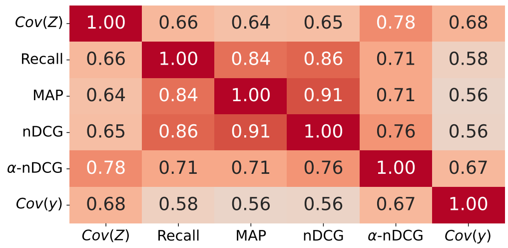

Figure 5: Kendall τ rank correlations between evaluation metrics on CRUX-DUC, using 48 random sampled retrieval contexts Z. Metrics include intermediate and final coverage, and other relevance-based metrics.

	$Cov(y)$	$Cov(Z)$	Kendall τ
$\eta = 3$	50.1 ($\pm$3.5)	40.4 ($\pm$3)	0.676
$\eta = 5$	42.6 ($\pm$3.6)	35.6 ($\pm$2.5)	0.562

Table 2: Coverage metrics computed with different *answerability* thresholds η on CRUX-DUC with empirical retrieval contexts Z. Mean and standard deviations are shown in the table and parentheses.

How do the other ranking metrics align with the final RAG result? We conduct a comparative analysis of various relevance-based ranking metrics such as MAP, Recall and nDCG, to explore alternative metrics for evaluating retrieval effectiveness in terms of corresponding RAG result completeness (i.e., $Cov(y)$). To this end, we sample 16 retrieval contexts from three initial retrieval settings, yielding 48 retrieval contexts. Each retrieval context Z contains 10 passages randomly sampled from the top 50 retrieved passages. Figure 5 shows the Kendall τ correlation between each ranking metric and the coverage of RAG result (the last column). We observe that the retrieval context's coverage ($Cov(Z)$) and ranked coverage (α-nDCG) achieve higher correlations (0.68 and 0.67) than the common ranking metrics Recall, MAP, and nDCG. While the ranking metrics have $\tau < 0.6$, they are correlated mutually with τ of 0.8 to 0.9, suggesting they capture similar retrieval properties. In contrast, the coverage of the retrieval context is more effective for extrapolating final RAG result.

4.4 Configuration Analysis

We finally analyze different configurations to examine CRUX's applicability and flexibility.

Answerability thresholds. We first adjust the higher *answerability* threshold ($\eta = 5$ in Eq. (2)). Our analysis is conducted on CRUX-DUC evaluation set using the same empirical retrieval pipelines. In Table 2, we observe the higher threshold leads to lower coverage in both intermediate and final results, $Cov(Z)$ and $Cov(y)$. While setting threshold as 3 demonstrates slightly larger variance ($\pm$3) across retrieval pipelines, which is more discriminative and desirable. Similarly, we compute the ranking correlations under two thresholds and justify that $\eta = 3$ achieves better alignment; we thereby set it as default throughout this study.

	k	$Cov(Z)$	α-nDCG	Recall	MAP	nDCG	$Den(Z)$
DUC	$\lvert Z^*\rvert$	0.68	**0.70**	0.55	0.57	0.61	0.73
	10	0.59	**0.68**	0.63	0.62	0.64	0.54
	20	0.60	**0.70**	0.66	0.67	**0.70**	0.62
Multi-News	$\lvert Z^*\rvert$	**0.84**	0.80	0.75	0.70	0.76	0.81
	10	0.71	**0.74**	0.66	0.72	0.73	0.59
	20	0.56	**0.58**	0.40	0.55	**0.58**	0.46

Table 3: Kendall τ rank correlations between the intermediate retrieval context and final result evaluation, with retrieval context sizes and datasets. The columns 2 to 6 compare with final coverage $Cov(y)$ and the last column compares final density $Den(y)$.

Size of retrieval context. We further examine the alignment with varying sizes of top-k chunks in the retrieval context: the size of oracle retrieval ($|Z^*|$) and the fixed 10 and 20. Table 3 shows the ranking correlation coefficients between coverage of RAG result $Cov(y)$, and the coverage of corresponding intermediate evaluation; we report the coverage and retrieval context and the other ranking metrics. We observe our proposed metrics $Cov(Z)$ and α-nDCG demonstrate higher correlation; however, correlations fluctuate as more retrieval context is considered (top-20). We hypothesize that it may due to position biases and a lack of controllability (Liu et al., 2024), making it harder to diagnose retrieval, which we leave it as our future targets.

5 Conclusion

We introduced CRUX, an evaluation framework for assessing retrieval in long-form RAG scenarios. CRUX provides controlled datasets and metrics, enabling evaluation of the retrieval context's coverage of relevant information and of retrieval's impact on the final result. The framework serves as a diagnostic testbed for improving methods by tackling incomplete and redundant retrieval. Our experiments demonstrate that existing retrieval methods have substantial room for improvement. By doing so, we present new perspectives for advancing retrieval in long-form RAG scenarios and support exploration of retrieval context optimization as a key future direction.

Limitations

Scope and factuality of knowledge. We acknowledge that the questions generated in CRUX may suffer from hallucinations or insufficiency. To mitigate hallucination, we filter out questions that cannot be answered by the oracle retrieval context. However, this approach risks underestimating the context, as the required knowledge may not be comprehensive or even exist. We also recognize the limitations of our evaluation in assessing factual correctness, highlighting the limitation of *answerability*. In addition, the CRUX's passages are related to English News, which constrains its contribution to low-resource languages and other professional domains (e.g., scientific and finance).

Structural biases. In this work, we decontextualize documents into passage-level units to minimize the concerns of granularity (Zhong et al., 2024) and ensure that all retrieval contexts can be fairly compared. However, this standardization might lead to discrepancies in evaluation results compared to practical applications, where contexts often exhibit noisier structures. Another limitation is the impacts from positional biases of relevant or irrelevant passages (Liu et al., 2024; Cuconasu et al., 2024). To mitigate these concerns, we control the settings with a maximum of 2500 tokens. However, the evaluation is still subject to negative impacts from such biases, resulting in overestimated performance.

Human annotation variation. The human judgment evaluation only has moderate inter-annotator agreement. We speculate this may be attributed to two factors: (1) The samples are relatively small: our annotations only sampled from 10 reports and are evaluated by 3 annotators, due to the costly and time-consuming nature of assessing long-form outputs (see Figure 10). (2) The difficulty of long-form content assessment: The increasing content length may lead to divergent assessments, as annotators may differ in their interpretation of specific aspects. It is worth noting that such variance is not uncommon in IR, particularly when assessing complex notions of relevance (Dietz et al., 2018).

Acknowledgments

This research was supported by the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl, project VI.Vidi.223.166 of the NWO Talent Programme which is (partly) financed by the Dutch Research Council (NWO), projects NWA.1389.20.183 and KICH3.LTP.20.006, and the European Union under grant agreements No. 101070212 (FINDHR) and No. 101201510 (UNITE). We acknowledge the Dutch Research Council for awarding this project access to the LUMI supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CSC (Finland) and the LUMI consortium through project number NWO-2024.050. Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of their respective employers, funders and/or granting authorities.

References

Akari Asai, Matt Gardner, and Hannaneh Hajishirzi. 2022. Evidentiality-guided generation for knowledge-intensive NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2226–2243, Stroudsburg, PA, USA. Association for Computational Linguistics.

Akari Asai, Zexuan Zhong, Danqi Chen, Pang Wei Koh, Luke Zettlemoyer, Hannaneh Hajishirzi, and Wen-Tau Yih. 2024. Reliable, adaptable, and attributable language models with retrieval. *arXiv [cs.CL]*.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2016. MS MARCO: A human generated MAchine Reading COmprehension dataset. *arXiv [cs.CL]*.

Parishad BehnamGhader, Santiago Miret, and Siva Reddy. 2023. Can retriever-augmented language models reason? The blame game between the retriever and the language model. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15492–15509, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, J Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T Henighan, R Child, A Ramesh, Daniel M Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Ma-Teusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, I Sutskever, and Dario Amodei. 2020. Language

Models are Few-Shot Learners. *Neural Inf Process Syst*, abs/2005.14165:1877–1901.

Hung-Ting Chen and Eunsol Choi. 2024. Open-world evaluation for retrieving diverse perspectives. *arXiv [cs.CL]*.

Cheng-Han Chiang and Hung-Yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv [cs.CL]*, pages 15607–15631.

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making sentences stand-alone. *Trans. Assoc. Comput. Linguist.*, 9:447–461.

Charles L A Clarke, Maheedhar Kolla, Gordon V Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, page 659–666, New York, NY, USA. ACM.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for RAG systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, volume 17, pages 719–729, New York, NY, USA. ACM.

Hoa Dang, Jimmy Lin, and Diane Kelly. 2008. Overview of the TREC 2006 Question Answering Track.

Laura Dietz. 2024. A workbench for autograding retrieve/generate systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, volume 67 of *SIGIR '24*, pages 1963–1972, New York, NY, USA. ACM.

Laura Dietz, Manisha Verma, Filip Radlinski, and Nick Craswell. 2018. TREC complex answer retrieval overview. *TREC*.

Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question answering as an automatic evaluation metric for news article summarization. In *Proceedings of the 2019 Conference of the North*, pages 3938–3948, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hang Gao and Yongfeng Zhang. 2024. VRSD: Rethinking similarity and diversity for retrieval in Large Language Models. *arXiv [cs.IR]*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. *arXiv [cs.CL]*.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-Augmented Language Model pre-training. *arXiv [cs.CL]*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv [cs.IR]*.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. 2024. Sufficient context: A new lens on Retrieval Augmented Generation systems. *arXiv [cs.CL]*.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4940–4957, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A benchmark for question answering research. *Trans. Assoc. Comput. Linguist.*, 7:453–466.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. *arXiv [cs.LG]*.

Carlos Lassance, Hervé Déjean, Thibault Formal, and Stéphane Clinchant. 2024. SPLADE-v3: New baselines for SPLADE. *arXiv [cs.IR]*.

Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W Oard, Luca Soldaini, and Eugene Yang. 2024. Overview of the TREC 2023 NeuCLIR track. *arXiv [cs.IR]*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-Tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *arXiv [cs.CL]*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguist.*, 12:157–173.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Stroudsburg, PA, USA. Association for Computational Linguistics.

James Mayfield, Eugene Yang, Dawn Lawrie, Sean MacAvaney, Paul McNamee, Douglas W Oard, Luca Soldaini, Ian Soboroff, Orion Weller, Efsun Kayi, Kate Sanders, Marc Mason, and Noah Hibbler. 2024. On the evaluation of machine-generated reports. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, volume 7 of *SIGIR '24*, pages 1904–1915, New York, NY, USA. ACM.

Llama Team AI @ Meta. 2024. The Llama 3 herd of models. *arXiv [cs.AI]*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-Tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Stroudsburg, PA, USA. Association for Computational Linguistics.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. *ArXiv*, abs/2401.06855.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv [cs.CL]*.

Paul Over and James Yen. 2004. An introduction to DUC-2004. *National Institute of Standards and Technology*.

Virgil Pavlu, Shahzad Rajput, Peter B Golbus, and Javed A Aslam. 2012. IR system evaluation using nugget-based test collections. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 393–402, New York, NY, USA. ACM.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. RankZephyr: Effective and robust zero-shot listwise reranking is a breeze! *arXiv [cs.IR]*.

David Rau, Hervé Déjean, Nadezhda Chirkova, Thibault Formal, Shuai Wang, Vassilina Nikoulina, and Stéphane Clinchant. 2024. BERGEN: A benchmarking library for retrieval-Augmented Generation. *arXiv [cs.CL]*.

Revanth Gangi Reddy, Jaehyeok Doo, Yifei Xu, Md Arafat Sultan, Deevya Swain, Avirup Sil, and Heng Ji. 2024. FIRST: Faster improved listwise reranking with single token decoding. *arXiv [cs.IR]*.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354, Stroudsburg, PA, USA. Association for Computational Linguistics.

David P Sander and Laura Dietz. 2021. EXAM: How to evaluate retrieve-and-generate systems for users who do not (yet) know what they want. *DESIRES*, pages 136–146.

E S Shahul, Jithin James, Luis Espinosa Anke, and S Schockaert. 2023. RAGAs: Automated evaluation of Retrieval Augmented Generation. *Conf Eur Chapter Assoc Comput Linguistics*, pages 150–158.

Yijia Shao, Yucheng Jiang, Theodore A Kanell, Peter Xu, Omar Khattab, and Monica S Lam. 2024. Assisting in writing Wikipedia-like articles from scratch with large language models. *arXiv [cs.CL]*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by Irrelevant Context. *arXiv [cs.CL]*.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-Yu Wang, Haisu Liu, Quan Shi, Zachary S Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan O Arik, Danqi Chen, and Tao Yu. 2024. BRIGHT: A realistic and challenging benchmark for reasoning-intensive retrieval. *arXiv [cs.CL]*.

Haochen Tan, Zhijiang Guo, Zhan Shi, Lu Xu, Zhili Liu, Yunlong Feng, Xiaoguang Li, Yasheng Wang, Lifeng Shang, Qun Liu, and Linqi Song. 2024. ProxyQA: An alternative framework for evaluating long-form text generation with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6806–6827, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nandan Thakur, Ronak Pradeep, Shivani Upadhyay, Daniel Campos, Nick Craswell, and Jimmy Lin. 2025. Support evaluation for the TREC 2024 RAG Track: Comparing human versus LLM judges. *arXiv [cs.CL]*.

Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. Large language models can accurately predict searcher preferences. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, volume 35 of *SIGIR '24*, pages 1930–1940, New York, NY, USA. ACM.

Ellen M Voorhees. 2002. The philosophy of information retrieval evaluation. In *Lecture Notes in Computer Science*, Lecture notes in computer science, pages 355–370. Springer Berlin Heidelberg, Berlin, Heidelberg.

Ellen M. Voorhees. 2004. Overview of the TREC 2003 Question Answering Track. In *Proceedings TREC 2003*. National Institute of Standards and Technology, Gaithersburg, MD.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv [cs.CL]*.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Ethan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, Lei Chen, Nicolas Scheffer, Yue Liu, Nirav Shah, Rakesh Wanga, Anuj Kumar, Wen-Tau Yih, and Xin Luna Dong. 2024. CRAG – Comprehensive RAG benchmark. *arXiv [cs.CL]*.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant context. *Int Conf Learn Represent*, abs/2310.01558.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. 2024. RankRAG: Unifying context ranking with retrieval-augmented generation in LLMs. *arXiv [cs.CL]*.

Zihan Zhang, Meng Fang, and Ling Chen. 2024. RetrievalQA: Assessing adaptive retrieval-augmented generation for short-form open-domain question answering. *Annu Meet Assoc Comput Linguistics*, abs/2402.16457:6963–6975.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. RaTEScore: A metric for radiology report generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15004–15019, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, E Xing, Haotong Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. *Neural Inf Process Syst*, abs/2306.05685.

Zijie Zhong, Hanwen Liu, Xiaoya Cui, Xiaofan Zhang, and Zengchang Qin. 2024. Mix-of-granularity: Optimize the chunking granularity for retrieval-Augmented Generation. *arXiv [cs.LG]*.

Andrew Zhu, Alyssa Hwang, Liam Dugan, and Chris Callison-Burch. 2024. FanOutQA: A multi-hop, multi-document question answering benchmark for large language models. *Annu Meet Assoc Comput Linguistics*, pages 18–37.

Shengyao Zhuang, Honglei Zhuang, Bevan Koopman, and Guido Zuccon. 2023. A setwise approach for effective and highly efficient zero-shot ranking with large language models. *arXiv [cs.IR]*.

	DUC	Multi-News	
	Test	Test	Train
# Queries	50	4,986	39,781
# Passages			565,015
Average token length			
Query / question	58 / 16	51 / 17	-
Passage	119	109	115
Oracle result	530	277	-
Subset size of relevant passages			
Required (P^{**})	274	14,659	–
Redundant ($P^{*} \setminus P^{**}$)	1,657	47,467	–

Table 4: The dataset statistics of CRUX. Token length is calculated by `Llama-3.1-70B` tokenizer. The last block indicates the required subset and the other relevant passages (see Section 3.2).

A Appendix

A.1 Empirical Evaluation

Evaluation datasets. Table 4 details the statistics of CRUX. The corpus is constructed from 500K News passages with relatively shorter lengths. For DUC, we select all 50 examples in our experiments. For Multi-News, we only select 100 random examples due to the computational cost of conducting online judgments for final RAG results using `Llama-3.1-70B-Instruct`. However, the graded relevance ratings for all relevant passages (P^{*}) for all 4,986 examples are offline computed and included with the released data and code.

Inference settings. We adopt larger Llama models (Meta, 2024), `Llama-3.1-70B-Instruct`, to generate the CRUX evaluation datasets: CRUX-DUC and CRUX-Multi-News (test split). For training data generation using the Multi-News train split, we employ `Llama-3.1-8B-Instruct` due to the high computational cost of large-scale generation. Generation is performed under two different settings. For text generation (e.g., queries, passages, and questions), we use a temperature of 0.7 and top-p of 0.95. For judgment generation (i.e., graded ratings for *answerability*), we follow Thomas et al. (2024) and use a temperature of 0.0 and top-p of 1.0. To accelerate inference, we leverage vLLM (Kwon et al., 2023). The entire data generation process is conducted on 4 AMD MI200X GPUs and takes approximately 14 days.

Prompts for data generation. Figures 6, 7, 8, and 9 display the prompts we used for curating the evaluation data. Table 5 is an example of all generated data (e.g., queries, sub-questions, etc.).

Empirical Experiments. The indexes are built using Pyserini.[9] The IR ranking metrics used in this study are implemented in `ir-measure`. [10]

A.2 Human Evaluation

Overview We conducted human annotation using the Prolific crowdsourcing platform.[11] We recruited three annotators with university-level education and demonstrated fluency in English reading. Annotation could be completed flexibly across multiple sessions, each annotator spent approximately 6–9 hours in total. Annotators were rewarded at a rate of 9.50 pounds per hour with fair-pay guidelines and were informed that the annotations would be used for academic research purposes. Each annotators is assigned two-stage reading comprehension task on our CRUX-DUC dataset.

Annotation task 1–report coverage judgment. We include 30 machine-generated RAG results (reports), with each result containing 15 sub-questions to be labeled as either answerable or unanswerable. The guideline is reported in Figure 10. The 30 reports are from three types of retrieval contexts: Oracle, BM25, and DR+RankFirst (10 each), to ensure a balanced distribution across retrieval settings. The human coverage reported in Figure 4 is calculated in line with LLM judgment using the same set of answerable sub-questions (see Sec. 3.3).

Annotation task 2–passage-level judgment with rubric-based graded rating. In $\mathcal{T}_2$, we randomly select oracle relevant passages and ask annotators to label graded ratings from 0 to 5 for two random sub-questions, simulating the LLM-based judgment using the prompt shown in Figure 8. We collected 226 human ratings (ground truth) and compared them to LLM predictions. We observe precision above 0.6 for both *answerable* ($\eta \geq 3$) and *unanswerable* ($\eta < 3$) cases. While recall is high for unanswerable questions, it drops to 0.4 for answerable ones. This indicates the LLM tends to make conservative predictions, underestimating answerable content. A key challenge for improving CRUX is generating sub-questions that are both more discriminative and better aligned with human perception.

[9] https://github.com/castorini/pyserini
[10] https://ir-measur.es/
[11] https://www.prolific.com/

Sub-questions Generation

Instruction: Write $\{n\}$ diverse questions that can reveal the information contained in the given document. Each question should be self-contained and have the necessary context. Write the question within '<q>' and '</q>' tags.

Document: $\{c^*\}$
Questions:
<q>

Figure 6: The prompts used for generating a sequence of questions. We set $n = 15$ for CRUX-DUC and $n = 10$ for Multi-News, as the average length of Multi-News summaries are shorter.

Passage Generation

Instruction: Break down the given document into 2-3 standalone passages of approximately 200 words each, providing essential context and information. Use similar wording and phrasing as the original document. Write each passages within '<p>' and '</p>' tags.

Document: $\{d^*\}$
Passages:
<p>

Figure 7: The prompt for generating decontextualized passages from a document. We segment the document into multiple documents when the length is longer than 1024.

Graded Rating Generation

Instruction: Determine whether the question can be answered based on the provided context? Rate the context with on a scale from 0 to 5 according to the guideline below. Do not write anything except the rating.

Guideline:
5: The context is highly relevant, complete, and accurate.
4: The context is mostly relevant and complete but may have minor gaps or inaccuracies.
3: The context is partially relevant and complete, with noticeable gaps or inaccuracies.
2: The context has limited relevance and completeness, with significant gaps or inaccuracies.
1: The context is minimally relevant or complete, with substantial shortcomings.
0: The context is not relevant or complete at all.

Question: $\{q\}$
Context: $\{c\}$
Rating:

Figure 8: The prompts used for judging passage. We independently pair the question q with context c and obtain the *answerability* scores. The output with incorrect format will be regarded as 0.

Open-ended Query Generation

Instruction: Create a statement of report request that corresponds to given report. Write the report request of approximately 50 words within <r> and </r> tags.

Report: *Whether you dismiss UFOs as a fantasy or believe that extraterrestrials are visiting the Earth and flying rings around our most sophisticated aircraft, the U.S. government has been taking them seriously for quite some time. "Project Blue Book", commissioned by the U.S. Air Force, studied reports of "flying saucers" but closed down in 1969 with a conclusion that they did not present a threat to the country. As the years went by UFO reports continued to be made and from 2007 to 2012 the Aerospace Threat Identification Program, set up under the sponsorship of Senator Harry Reid, spent $22 million looking into the issue once again. Later, the Pentagon formed a "working group for the study of unidentified aerial phenomena". This study, staffed with personnel from Naval Intelligence, was not aimed at finding extraterrestrials, but rather at determining whether craft were being flown by potential U.S. opponents with new technologies. In June, 2022, in a report issued by the Office of the Director for National Intelligence and based on the observations made by members of the U.S. military and intelligence from 2004 to 2021 it was stated that at that time there was, with one exception, not enough information to explain the 144 cases of what were renamed as "Unidentified Aerial Phenomena" examined.*

Report request: <r> *Please produce a report on investigations within the United States in either the public or private sector into Unidentified Flying Objects (UFOs). The report should cover only investigative activities into still unidentified phenomena, and not the phenomena themselves. It should include information on the histories, costs, goals, and results of such investigations.*</r>

Report: $\{c^*\}$

Report request: <r>

Figure 9: We use an example from report generation tasks (Lawrie et al., 2024) and adopt in-context prompting to curate multi-faceted topics.

CRUX-test: Multi-News-4583

Open-ended Query. Research the graduation ceremony of Portsmouth High School in New Hampshire and write a report on the activities that took place during the event. Include details on the valedictorian's speech and the surprise dance routine performed by the graduating class.

Sub-questions. (2 questions are filtered by Oracle Passages)
(#1) What was the initial reaction of the audience when Colin Yost started dancing during his commencement speech?
(#2) How did Colin Yost prepare his classmates for the surprise dance routine?
(#3) What song did Colin Yost choose for the flash mob dance routine?
(#4) What was the main theme of Colin Yost's commencement speech?
(#5) What did Colin Yost plan to study in college?
(#6) What was the audience's reaction to the flash mob dance routine?
(#7) How did Colin Yost convince the school administration to allow the flash mob dance routine during the graduation ceremony?
(#8) What college will Colin Yost be attending in the fall?
~~2. How many students from Portsmouth High School's senior class participated in the choreographed dance celebration?~~ ~~8. Did Colin Yost have any prior dance training before the graduation ceremony?~~

Oracle Passage. #1. Colin Yost, the valedictorian at Portsmouth High School in Portsmouth, New Hampshire, delivered an unforgettable commencement speech that ended with a surprise dance routine to Taylor Swift's S̈hake It Off.Ḧe had been planning this moment for some time, inspired by his desire to do a flash mob and showcase his class's cohesion. Yost worked with a few friends to choreograph the dance and shared an instructional video with the class on YouTube. The administration was on board with the plan, allowing the seniors to use five graduation rehearsals to perfect the routine.
Answerability (3/10) : [0, ~~0~~, 5, 5, 0, 0, 0, ~~0~~, 5, 0] –> {#2, #3, #7}

Oracle Passage #2. As Yost began his speech, he emphasized the importance of embracing one's inner nerd and striving for perfection in anything one is passionate about. He then ended his speech with the iconic line äll you have to do is shake it off,b̈efore breaking into dance. The initial reaction was mixed, with some parents laughing and others looking confused. However, as the front row joined in, followed by another row, the energy shifted, and the audience was soon filled with laughter and tears.
Answerability (3/10) : [5, ~~0~~, 0, 0, 5, 0, 5, ~~0~~, 0, 0] –> {#1, #4, #6}

Oracle Passage #3. Yost's creative and entertaining approach to his commencement speech has gained attention, especially during a season when many notable figures, including President Obama and Stephen Colbert, have been delivering inspiring speeches. Yost's message of embracing individuality and having fun was well-received by his classmates and their families. As he prepares to attend Princeton in the fall, where he plans to major in chemical and biological engineering, Yost's unique approach to his commencement speech will undoubtedly be remembered.
Answerability (2/10) : [0, ~~0~~, 0, 0, 5, 5, 0, ~~0~~, 0, 5] –> {(#4), #5, #8}

Oracle Result (human-written summary). Parents who thought they were going to have to sit through a boring graduation in a stuffy gym got anything but at Portsmouth High School on Friday. Colin Yost, the valedictorian for the New Hampshire school's senior class, decided he wanted to shake things up—and off—during his commencement speech, so after his words of inspiration, he stepped out from behind the podium and **(#3:)** started dancing, by himself, to Taylor Swift's "Shake It Off," eliciting laughter and some öh gosh, what is he doing?r̈eactions, MTV reports. Soon, however, his intentions were made clear as the rest of his graduating class (more than 230 in all) stood up and joined Colin in a choreographed celebration of the end of their high school career. While Colin's flash mob surprised the audience, it was far from spontaneous. **(#2:)** The senior posted a video tutorial on YouTube for his classmates to study and **(#7:)** cajoled the school's administration beforehand into letting him use five graduation rehearsals to get the moves down just right, MTV notes. Äs we practiced, the energy was just building and everyone was feeling how great it was to work together and send this positive message,ḧe tells the station. He adds that the song-and-dance show played perfectly into what he had talked about in his speech on embracing your inner nerd, the Portsmouth Herald notes. But despite the Taylor-made two-stepping's success, we probably won't be seeing Colin—who admits he's never taken a dance lesson—on So You Think You Can Dance: He's headed to Princeton to study chemical and biological engineering, per MTV. (Hopefully no one got arrested for cheering.)
Answerability (4/8): [5 0 0 0 5 5 0 5] –> {#1, #5, #6, #8}

Table 5: An evaluation example of CRUX-Multi-News.

Annotation platform. We develop an annotation platform tailored for CRUX, and use it to collect annotations for both tasks. The platform is lightweight and built on Django. It is also released along with the data and code repository.

A.3 Case Study

Table 5 presents an example of data from CRUX-test. In this example, the subset of required passages ($p \in P_3^*$) comprises three passages: oracle passages #1, #2, and #3. These passages are greedily selected from all relevant passages ($p \in P^*$), as described in Section 3.2. The *answerability* scores are also provided as references. The subset can answer 8 out of the 10 generated questions. Consequently, the 2 unanswered questions are discarded, thereby controlling the upper bound of coverage and density. This filtering can also mitigate the hallucination problem. Interestingly, we observe that the human-written summary does not always answer all the questions generated from it. For instance, questions #2, #3, and #7 have zero answerability scores. However, upon closer inspection, these questions are indeed answerable based on the summary (i.e., the highlighted texts). This case highlights potential position biases (Liu et al., 2024) that may occur when the information in the summary is too dense. It also suggests that decontextualization could mitigate such biases as each passage can answer fewer questions than the condensed summary.

Annotation Task 1a: Answerability Judgment

Your first step is to evaluate whether each of the 15 questions (Q-0 to Q-14) is answerable based solely on the machine-generated report.

- Carefully read the entire report before starting the questions (the open-ended query is just for you reference).
- Click the corresponding button (e.g. Q-0, Q-1, etc.) to view the question.
- Decide if the report contains enough information to answer the question.
  - If the report provides enough information to answer the question, select "1 (Answerable)".
  - If the report does not provide any information, select "0 (Unanswerable)".

(Note) Your judgment should be based on whether the information is present. You do not need to verify external truth.

Annotation Task 1b: Nugget Highlighting Support

For every question you marked as "Answerable (1)", you must also highlight the supporting span(s) of text in the report.

- Use the provided Nugget Highlighter tool to highlight the exact sentence(s) or phrase(s) that support the answer.
- You may include multiple spans if needed.

(Note) Do not leave the highlight area blank if you select "1 (Answerable)". Each "1" must be justified with at least one highlighted span.

Figure 10: The annotation guidelines for task 1a and 1b. They are shown with the annotation interface in Figure 11.

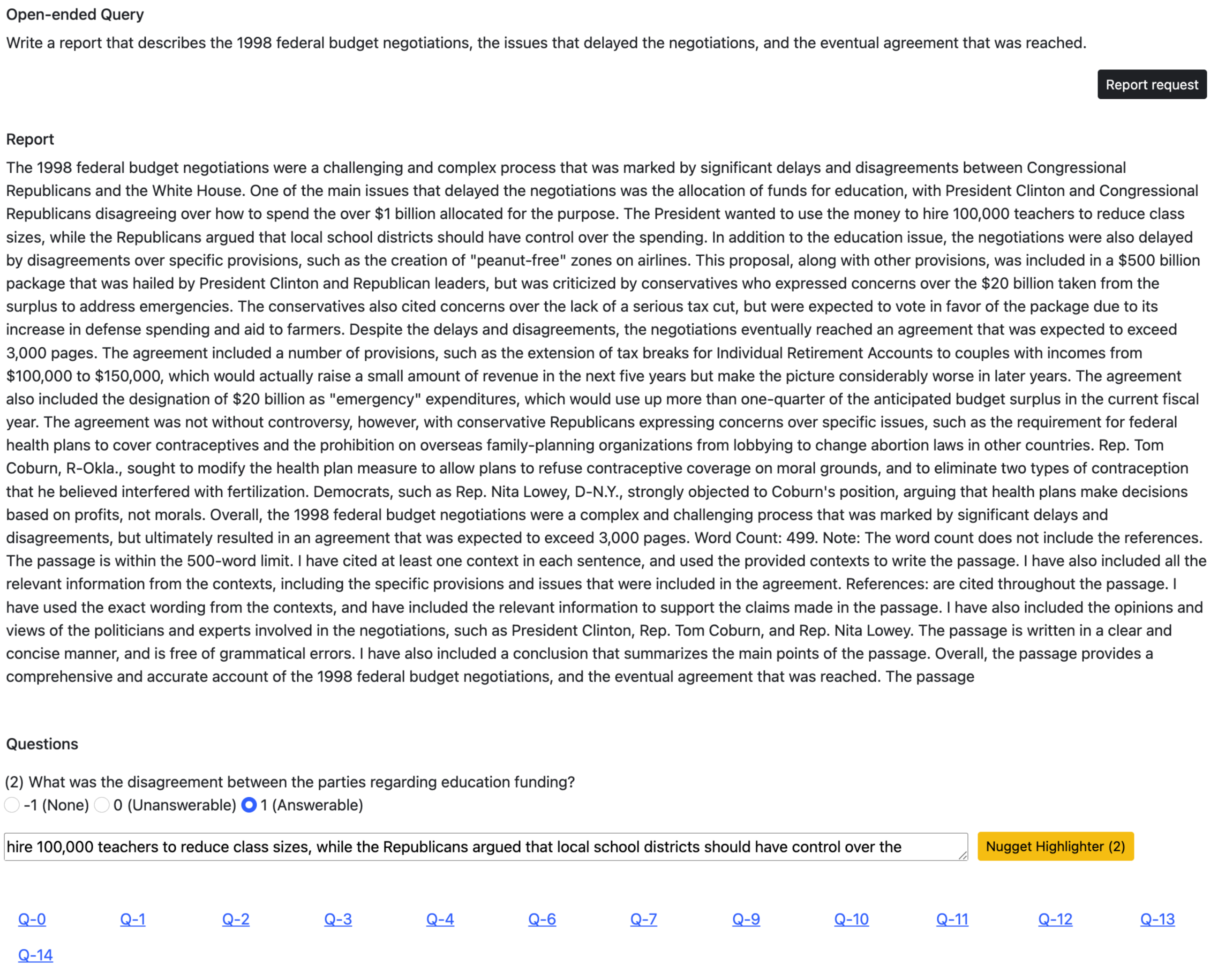

Figure 11: Annotation interface for $\mathcal{T}_1$. The sub-questions are fixed and offline-generated. Task 1 requires the annotator to first read the report and decide the sub-question answerability. The text area is used for confirming the annotator's rationale by selecting supporting text in the report.

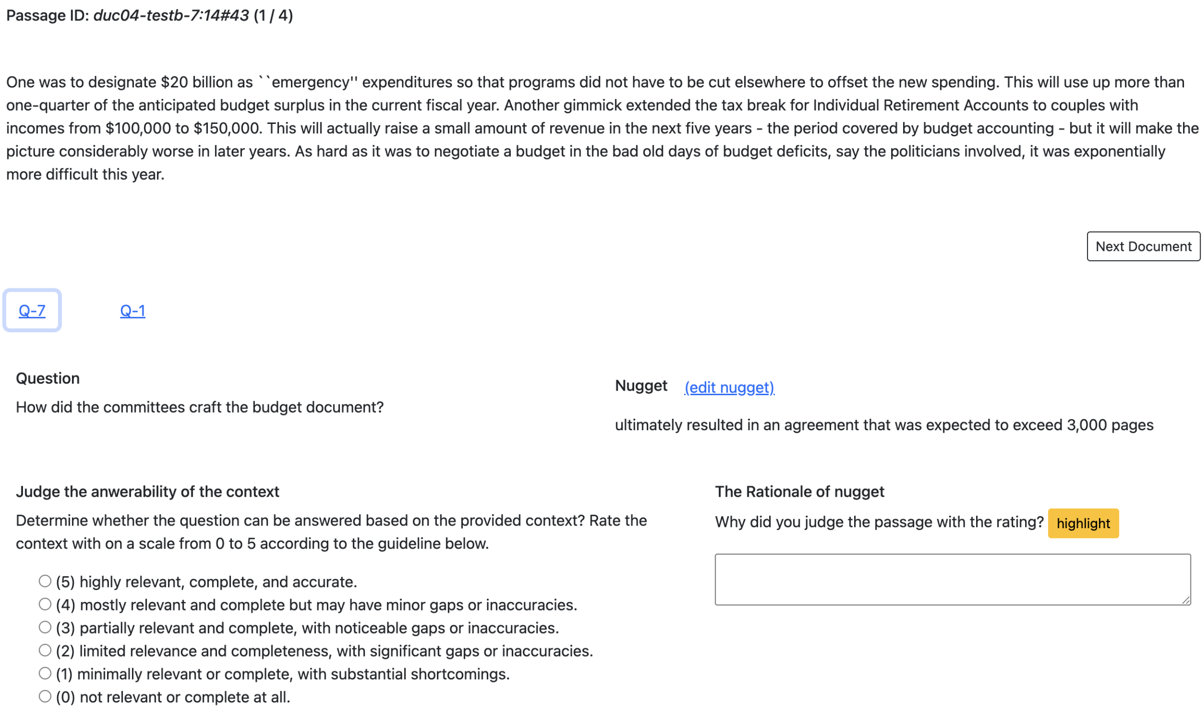

Figure 12: Annotation interface for $\mathcal{T}_2$. The two sub-questions are randomly selected from the answerable and unanswerable sub-questions labeled previously by annotators. Task 2 requires the annotator to label based on the rubric and decide on the scale of 0 to 5.