RAGEval: Scenario Specific RAG Evaluation Dataset Generation Framework

Kunlun Zhu[16*], **Yifan Luo**[1*], **Dingling Xu**[2*], **Yukun Yan**[1†], **Zhenghao Liu**[3], **Shi Yu**[1], **Ruobing Wang**[4], **Shuo Wang**[1], **Yishan Li**[5], **Nan Zhang**[5], **Xu Han**[1], **Zhiyuan Liu**[1†], **Maosong Sun**[1]
[1]Tsinghua University, [2]Beijing Normal University,
[3]Northeastern University, [4]University of Chinese Academy of Sciences
[5]ModelBest, [6]University of Illinois Urbana-Champaign
yanyk.thu@gmail.com

Abstract

Retrieval-Augmented Generation (RAG) is a powerful approach that enables large language models (LLMs) to incorporate external knowledge. However, evaluating the effectiveness of RAG systems in specialized scenarios remains challenging due to the high costs of data construction and the lack of suitable evaluation metrics. This paper introduces **RAGEval**, a framework designed to assess RAG systems across diverse scenarios by generating high-quality documents, questions, answers, and references through a schema-based pipeline. With a focus on factual accuracy, we propose three novel metrics—Completeness, Hallucination, and Irrelevance—to evaluate LLM-generated responses rigorously. Experimental results show that RAGEval outperforms zero-shot and one-shot methods in terms of clarity, safety, conformity, and richness of generated samples. Furthermore, the use of LLMs for scoring the proposed metrics demonstrates a high level of consistency with human evaluations. RAGEval establishes a new paradigm for evaluating RAG systems in real-world applications. The code and dataset are released at https://github.com/OpenBMB/RAGEval.

1 Introduction

Retrieval-Augmented Generation (RAG) systems are increasingly gaining attention (Gao et al., 2023; Asai et al., 2024) due to their ability to integrate external knowledge into large language models (LLMs). This ability is crucial in fields such as medicine, finance, and law, where factual accuracy is crucial in decision-making. However, RAG systems are still prone to hallucination, mainly due to noise introduced during retrieval and LLMs' limited capacity to exploit retrieved information fully.

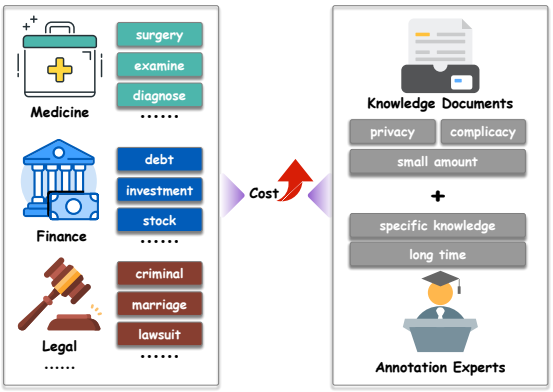

Figure 1: The challenges of building scenario-specific RAG evaluation datasets: scenario coverage and annotation costs.

Although various benchmarks for measuring the capabilities of existing RAG systems have been proposed (Joshi et al., 2017; Nguyen et al., 2017; Kwiatkowski et al., 2019; Chen et al., 2024b; Lyu et al., 2024), they often lack sufficient coverage of diverse, domain-specific scenarios and fail to incorporate comprehensive metrics for assessing factual accuracy, which limits their applicability in real-world contexts that require precise and reliable information (Bruckhaus, 2024). Furthermore, the challenges of building scenario-specific evaluation datasets—such as dynamic real-world conditions, privacy concerns, and the need for expert annotation—further exacerbate the issue.

To address these challenges, we propose **RAGEval**, a novel framework designed to automatically generate scenario-specific RAG evaluation datasets. By summarizing essential knowledge from seed documents, RAGEval creates a schema that forms the basis for generating questions, answers, and references for evaluation. Additionally, factual key points are extracted from each answer, enabling a more accurate assessment of the RAG system predictions.

In RAG system assessments, evaluation met-

*Equal contribution; in random order; each reserves the right to be listed first.

† Corresponding authors.

rics, like data, also play a pivotal role. Traditional metrics such as F1, ROUGE-L (Lin, 2004), and BLEU (Papineni et al., 2002) are often inadequate for complex or long-form responses. They mainly focus on the lexical overlap of the responses with the gold reference and overlook the semantic similarity. Some novel approaches, such as those relying on LLMs to evaluate responses directly (Es et al., 2024; Saad-Falcon et al., 2023), suffer from issues of stability and comparability. To address these limitations, we introduce three novel metrics—Completeness, Hallucination, and Irrelevance—that are grounded in factual key points and provide a more stable and comparable scoring method.

Our main contributions are: (1) We propose `RAGEval`, a novel framework for automatically generating scenario-specific RAG evaluation datasets. (2) We introduce three novel evaluation metrics to assess the factual accuracy of generated answers more effectively than existing metrics like ROUGE-L and BLEU. (3) We develop a new RAG benchmark, DragonBall, and conduct comprehensive experiments to analyze the impact of RAG systems' retrieval and generation components on the performance results.

2 Related Work

The evaluation of question-answering (QA) and RAG systems has seen significant advancements in recent years. Traditional open-domain QA benchmarks, such as HotpotQA (Yang et al., 2018), TriviaQA (Joshi et al., 2017), MS Marco (Nguyen et al., 2017), and Natural Questions (Kwiatkowski et al., 2019), have long served as foundational datasets for general QA tasks. However, these benchmarks face limitations in evaluating modern RAG systems, particularly their ability to assess domain-specific knowledge, nuanced outputs, and retrieval accuracy. For instance, potential data leakage in these datasets and a lack of fine-grained metrics hinder their effectiveness in evaluating the nuanced behaviour of RAG systems.

In response, several RAG-specific benchmarks have emerged. RGB (Chen et al., 2024b) focuses on assessing LLMs' ability to integrate retrieved information, emphasizing noise robustness. CRUD-RAG (Lyu et al., 2024) categorizes RAG tasks into Create, Read, Update, and Delete operations to evaluate different aspects of information retrieval. CRAG (Yang et al., 2024) extends domain coverage by introducing mock APIs to simulate real-world retrieval tasks, while MultiHop-RAG (Tang and Yang, 2024) challenges systems with multi-hop reasoning across multiple documents. These benchmarks, while valuable, remain constrained by predefined domains and fixed task structures, limiting their adaptability to dynamic, real-world applications.

Traditional evaluation metrics, such as F1, ROUGE-L, and BLEU, have been widely used in various benchmarks to assess the quality of generated answers in RAG systems. However, these metrics focusing on lexical often fail to capture the full complexity of generative tasks, especially in the case of long-form responses where factual accuracy and contextual relevance are critical. Moreover, metrics like Hit Rate, MRR, and NDCG are commonly used for retrieval evaluation but cannot assess generative capabilities (Liu, 2023; Nguyen, 2023).

In recent years, newer approaches have integrated LLMs into the evaluation process, trying to solve the problems of traditional metrics. RAGAS (Es et al., 2024) and ARES (Saad-Falcon et al., 2023) use LLM-generated data to evaluate contextual relevance and informativeness without relying on ground truth references. While these methods provide valuable insights, they fail to address the complexities of scenario-specific evaluations. RGB (Chen et al., 2024b) introduces task-oriented metrics that assess noise robustness and information integration, but it does not offer the flexibility required for dynamic, application-specific tasks. RAGTruth (Niu et al., 2024) proposes a corpus for evaluating hallucinations, a critical issue in RAG systems.

While the aforementioned benchmarks and evaluation methods have made significant strides, they still face challenges in addressing the diversity of real-world application scenarios, which often require domain-specific data generation and context-sensitive evaluation. To solve the problem of scenario diversity in RAG evaluation, our method builds upon these advancements by introducing a novel framework for automatically generating evaluation datasets. Unlike existing frameworks that rely on predefined datasets and fixed benchmarks, our method offers higher contextual agility, enabling the design of scenario-specific factual queries tailored to different applications.

Furthermore, we introduce three novel keypoint-based evaluation metrics—Completeness, Halluci-

nation, and Irrelevance—designed to assess factual accuracy and relevance in these dynamically generated, scenario-specific contexts. These metrics stand in contrast to traditional benchmarks that assess RAG systems using a single, static set of evaluation criteria. Our framework enables the automatic generation of diverse datasets and provides more adaptable evaluation metrics, making it better suited for evolving application domains.

3 Method

In this section, we introduce the proposed `RAGEval` method. To provide an overview, we summarize the overall generation process as follows:

$\mathcal{S} \rightarrow \mathcal{C} \rightarrow \mathcal{D} \rightarrow (\mathcal{Q}, \mathcal{A}) \rightarrow \mathcal{R} \rightarrow$ Keypoints

This sequence outlines how the schema summary ($\mathcal{S}$) leads to configuration generation ($\mathcal{C}$), followed by document generation ($\mathcal{D}$). From there, question-answer pairs ($\mathcal{Q}, \mathcal{A}$) are derived, and supporting references ($\mathcal{R}$) are identified. Finally, keypoints are extracted, serving as concise representations of the critical information in the answers.

3.1 Schema Summary

In scenario-specific text generation, a schema ($\mathcal{S}$) is an abstract representation of key elements, encapsulating the aspects of essential factual knowledge from input documents. This schema serves as the backbone that ensures content diversity and reliability while standardizing outputs across various scenarios to maintain alignment with professional standards.

The schema defines a structural framework of key elements for domain-specific documents without containing actual data. In medicine, it may outline categories for symptoms and treatments; in finance, it could establish classifications for sectors, organizations, and metrics. Specific data is later populated into this predefined framework during configuration generation. For example, in legal contexts, the schema might encompass fundamental legal concepts—such as case law, statutes, and court rulings—ensuring broad applicability without relying on specific legal instances. This approach allows the schema to remain versatile and scalable across various legal scenarios. A concrete example of a legal domain schema illustrating these principles is provided in figure 11 in Appendix E.

The schema is initially generated using GPTs[1] based on a curated set of seed documents, which establish the foundational domain-specific knowledge. Following this, the schema undergoes a series of iterative refinements guided by human intuition and contextual understanding. This process ensures that the schema maintains a balance between comprehensiveness, accuracy, and generalizability, effectively supporting content generation across diverse sub-scenarios. The refinement process[2] is designed to prevent over-specialization, thereby enhancing the schema's scalability and adaptability.

3.2 Configuration and Document Generation

Generating scenario-specific documents with rich factual information and internal consistency is crucial for creating high-quality datasets, ensuring the generated content can be evaluated accurately and applied effectively in downstream tasks. To achieve this, we first generate configurations $\mathcal{C}$, derived from the previously established schema $\mathcal{S}$. These configurations act as references and constraints for text generation, ensuring consistency across the document.

We adopt a hybrid approach to generate configurations $\mathcal{C}$, combining rule-based methods with LLMs to assign values to schema elements. Rule-based methods (e.g., selecting values randomly from predefined scenario-specific options) ensure high accuracy and factual consistency for structured data. Meanwhile, LLMs generate more complex or diverse content, balancing consistency and creativity. For instance, in financial reports, configurations may include various sectors such as agriculture, aviation, and construction, each covering multiple aspects of its respective domain. An illustrative configuration for the legal scenario is provided in figure 11 in Appendix E, demonstrating how different elements can be combined within this domain.

We then use GPT-4o to convert the factual information from the configuration $\mathcal{C}$ into a structured narrative format tailored to a specific scenario. For example, in medical records, the generated document may include categories such as patient information, medical history, and treatment plan to ensure accuracy and relevance. Similarly, we include a company summary in financial reports to maintain continuity and distinct sections such as Financial Report, Corporate Governance, and Environmental and Social Responsibility.

[1] https://chatgpt.com/gpts

[2] See the Appendix A for details on the refining process.

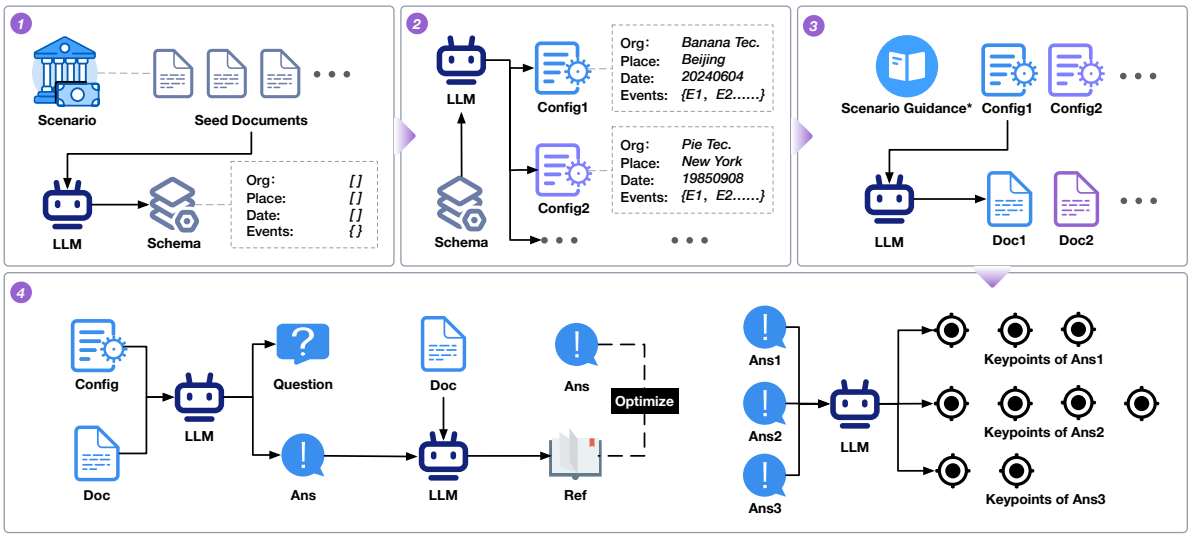

Figure 2: RAGEval Progress: ① summarizing a schema containing specific knowledge from seed documents. ② filling in factual information based on this schema to generate diverse configurations. ③ generating documents according to the configurations. ④ creating evaluation data composed of questions, answers, and references derived from the configurations and documents.

3.3 QRA Generation

In this subsection, we describe the process of generating Question-Reference-Answer (QRA) triples using the documents $\mathcal{D}$ and configurations $\mathcal{C}$ to establish a robust evaluation framework for information retrieval and reasoning. The goal is to ensure that generated content can be evaluated comprehensively across multiple aspects of information understanding.

We utilize configurations $\mathcal{C}$ to guide the generation of questions and initial answers, ensuring the generated content is aligned with the schema elements. These configurations are embedded within prompts to ensure that the generated questions are specific and that the answers are precise and grounded in the schema elements. We address different types of questions, such as factual, multi-hop reasoning, summarization, and multi-document questions, each designed to evaluate specific facets of language understanding. To ensure the diversity and controllability of the questions generated by the model, we have designed 7 question types, as detailed in Table 13 in Appendix E. The GPT-4o model is provided with detailed instructions and examples for each question type, generating targeted questions $\mathcal{Q}$ and initial answers $\mathcal{A}$.

Specific prompts and examples are detailed in the Appendix E. Using the generated questions $\mathcal{Q}$ and initial answers $\mathcal{A}$, we extract relevant information fragments (references) $\mathcal{R}$ from the documents $\mathcal{D}$. This is accomplished using an extraction prompt, ensuring that the generated answers are grounded in the source material for reliability and traceability. Extracting these references enhances the comprehensiveness and consistency of the generated content.

To ensure alignment between answers $\mathcal{A}$ and references $\mathcal{R}$, we iteratively refine the answers to improve coherence and accuracy. If references contain content missing from the answers, we supplement them accordingly. Conversely, if the answers contain unsupported content, we either locate the relevant references or remove the unsupported sections. This step reduces hallucinations and ensures that the final answers are accurate and well-supported by $\mathcal{R}$.

Keypoints are generated from answers $\mathcal{A}$ for each question $\mathcal{Q}$ to highlight the critical information in the responses. We employ a predefined prompt with in-context learning, including examples across different scenarios and question types. Typically, each response is distilled into 3-5 keypoints, encompassing essential factual details, relevant inferences, and conclusions. This keypoint extraction supports a precise and reliable evaluation of generated content.

3.4 DragonBall Dataset

We construct the DragonBall dataset, which stands for **D**iverse **RAG** **O**mn**i**-**B**enchmark for **All** scenarios, by leveraging the generation method described above. This dataset encompasses a range

of texts and RAG questions across three critical domains—finance, law, and medical—chosen for their real-world importance. In addition, the dataset features both Chinese and English texts, serving as a comprehensive resource for multilingual, scenario-specific research. Overall, the dataset contains 6,711 questions, reflecting its extensive scale and diversity. Additional details on the generated DragonBall dataset, including human evaluations of data quality, are provided in Appendix C and D.

3.5 Evaluation Metrics for RAG Systems

In this work, we propose a comprehensive evaluation framework for RAG systems, considering both retrieval and generation components.

We define multiple metrics to evaluate the model's effectiveness and efficiency in the retrieval phase. These metrics are designed explicitly for RAG systems, considering the situations when generating answers with incomplete and noisy information.

3.5.1 Retrieval Metrics

Recall. We introduce the RAG Retrieval Recall metric to evaluate the effectiveness of the retrieval process in matching ground truth references. The Recall is formally defined as

$$\text{Recall} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(M(G_i, \mathcal{R})), \tag{1}$$

where n is the total number of ground truth references, G_i denotes the i-th ground truth reference, $\mathcal{R} = \{R_1, R_2, \ldots, R_k\}$ represents the set of retrieved references, $M(G_i, \mathcal{R})$ is a boolean function that returns true if all sentences in G_i are found in at least one reference in $\mathcal{R}$, and false otherwise, and $\mathbb{1}(\cdot)$ is the indicator function, returning 1 if the condition is true and 0 otherwise.

This metric assesses the alignment between retrieved and ground truth references at the sentence level. A ground truth reference is considered successfully recalled if all its constituent sentences are present in at least one retrieved reference.

Effective Information Rate (EIR). This metric quantifies the proportion of relevant information within the retrieved passages, ensuring that the retrieval process is accurate and efficient regarding information content. It is calculated as

$$\text{EIR} = \frac{\sum_{i=1}^{m} |G_i \cap R_t|}{\sum_{j=1}^{k} |R_j|}, \tag{2}$$

where G_i is the i-th ground truth reference, R_t is the set of total retrieved passages, m is the number of ground truth references successfully matched, $|G_i \cap R_t|$ represents the number of words in the intersection of the i-th ground truth reference and the concatenated retrieved passages R_t, calculated only if G_i is matched in R_t, $|R_j|$ represents the total number of words in the j-th retrieved passage, and k is the total number of retrieved passages.

To calculate $|G_i \cap R_t|$ at the sentence level, follow these steps: 1) divide G_i into individual sentences, 2) for each sentence in G_i, check if it matches any sentence in R_t, 3) calculate the number of words in the matched sentences, and 4) sum the number of words from all matched sentences to get $|G_i \cap R_t|$. These steps ensure the overlap is calculated based on sentence-level matches, providing a more granular and accurate measure of relevant information within the retrieved passages.

3.5.2 Generation Metrics

For the generation component, we introduce novel metrics tailored for RAG evaluation. These metrics comprehensively evaluate the quality and reliability of generated answers.

Completeness. Completeness measures how well the generated answer captures the key information from the ground truth. We employ LLM to generate a set of key points $K = \{k_1, k_2, \ldots, k_n\}$ from the ground truth. The Completeness score is then calculated as the proportion of key points semantically covered by the generated answer A:

$$\text{Comp}(A, K) = \frac{1}{|K|}\sum_{i=1}^{n} \mathbb{1}[A \text{ covers } k_i], \tag{3}$$

where $\mathbb{1}[\cdot]$ is an indicator function that evaluates to 1 if the generated answer A semantically covers the key point k_i, and 0 otherwise. Here, "covers" means that the generated answer contains information consistent with and correctly representing the key point. Specifically, for a key point to be considered covered, the generated answer must include the relevant information and present it accurately without contradictions or factual errors.

Hallucination. Hallucination identifies instances where the content contradicts key points, highlighting potential inaccuracies. The Hallucination score is calculated as

$$\text{Hallu}(A, K) = \frac{1}{|K|}\sum_{i=1}^{n} \mathbb{1}[A \text{ contradicts } k_i], \tag{4}$$

where $\mathbb{1}[\cdot]$ is an indicator function that evaluates to 1 if the generated answer A contradicts the key point k_i, and 0 otherwise.

Irrelevancy. Irrelevancy assesses the proportion of key points from the ground truth that are neither covered nor contradicted by the generated answer. Irrelevancy quantifies the proportion of key points neither covered nor contradicted, indicating areas where the answer fails to engage with relevant information. The Irrelevancy score is calculated as

$$\text{Irr}(A, K) = 1 - \text{Comp}(A, K) - \text{Hallu}(A, K). \quad (5)$$

Completeness, Hallucination, and Relevance pinpoint specific RAG models' strengths and weaknesses. They ensure that generated answers are informative, accurate, and relevant, enhancing their quality and trustworthiness. More details about the prompt for evaluation and keypoints generation, and the comparison with human evaluation can refer to the Appendix C.

4 Experiments

4.1 Setup

In our experiments (Table 1), the BGE-M3 (Chen et al., 2024a) model is used both for Chinese and English, with the following hyperparameters: the TopK retrieved documents are set to 5, the retrieval batch size is 256. The maximum length for the retrieval query is capped at 128 tokens. The default chunk size is set to 512, and meta-information (e.g., company name, patient details) is added to enhance retrieval.

For generation, the maximum input length for the query generator is set to 4096 tokens, and batches of 5 are processed. The generation parameters include a maximum of 512 new tokens per output.

We use the model's default generation configurations (e.g., temperature, Top-P). If not available, the default settings from Hugging Face will be applied. For ChatGPT models, temperature is set to 0.2 and TopP to 1.0, generating one response per query.

We use FlashRAG (Jin et al., 2024) as the RAG inference pipeline with vLLM (Kwon et al., 2023) as the backend.

4.2 Generation Performance Comparison

In this experiment, we compare the performance of 9 popular open/close-sourced generation models with different parameter sizes, including MiniCPM-2B-sft and MiniCPM3-4B (Hu et al., 2024), Baichuan-2-7B-chat (Yang et al., 2023), Llama3-8B-Instruct (AI@Meta, 2024), Qwen1.5-7B/14B-chat (Bai et al., 2023), Qwen2-7B-Instruct (Bai et al., 2023), GPT-3.5-Turbo-0125, and GPT-4o-2024-0806 [3]. We use the same input prompt to compare the outputs of the different generation models. We chose 50 random questions of all question types for each scenario and language for evaluation. The overall experimental results of the different generation models are shown in Table 1.

GPT-4o and MiniCPM3-4B Show Superior Generation Performance. According to our proposed keypoint-based evaluation shown in Table 1, GPT-4o achieves the highest Completeness scores of 79.13% (CN) and 69.36% (EN) and the lowest Hallucination scores in Chinese at 12.10%. What's more, the best-performing small-to-medium open-source model is MiniCPM3-4B, which highlights significant room for improvement among open-source alternatives.

Findings on Model Size. Our experimental results in Table 1 further validate the scaling law (Kaplan et al., 2020) within the same model family. For instance, Qwen1.5-14B-chat outperforms Qwen1.5-7B-chat and other open-source models except for MiniCPM3-4B, achieving better scores in both Completeness and Hallucination.

The Effectiveness of Keypoint-Based Metrics. Our analyses reveal notable discrepancies between traditional evaluation metrics, such as Rouge-L and BLEU—and keypoint-based metrics that assess deep semantic alignment. For instance, in the Chinese setting, Baichuan-2-7B-chat achieves the highest Rouge-L (38.30%) and BLEU (21.55%) scores, yet its Completeness score is relatively low at only 60.25%. Conversely, GPT-4o performs the best on Completeness, scoring 79.13% in Chinese, while it exhibits both the low Rouge-L (21.30%) and BLEU (8.70%) in Chinese. These results suggest that while Rouge-L and BLEU primarily measure surface-level language similarity, keypoint-based metrics capture deeper semantic correspondence, thereby offering a more nuanced reflection of a model's true performance in RAG tasks.

[3] https://platform.openai.com/docs/models

Model	Completeness (↑)		Hallucination (↓)		Irrelevance (↓)		Rouge-L (↑)		BLEU (↑)	
	CN	EN	CN	EN	CN	EN	CN	EN	CN	EN
MiniCPM-2B-sft	54.59	57.88	28.82	19.49	16.58	22.63	31.11	26.94	15.19	6.38
MiniCPM3-4B	75.74	64.09	13.78	16.42	10.48	19.49	32.06	27.99	16.34	6.82
Baichuan-2-7B-chat	60.25	57.40	23.97	19.60	15.77	22.99	**38.30**	**30.68**	**21.55**	**8.84**
Qwen1.5-7B-chat	69.50	62.76	19.25	17.65	11.25	19.60	32.49	21.62	17.11	4.06
Qwen2-7B-Instruct	70.83	65.38	16.93	16.41	12.24	18.21	24.55	22.99	10.26	4.69
Llama3-8B-Instruct	69.26	63.61	18.29	15.12	12.45	21.27	21.54	25.22	9.15	5.12
Qwen1.5-14B-chat	73.17	64.41	14.40	15.50	12.43	20.09	31.93	23.99	15.25	4.84
GPT-3.5-Turbo	75.40	68.37	13.10	15.72	11.50	**15.91**	18.92	19.84	6.45	3.35
GPT-4o	**79.13**	**69.36**	**12.10**	**13.79**	**8.77**	16.85	21.30	23.25	8.70	4.80

Table 1: Overall model performance results (%) of nine language models in generation across Chinese (CN) and English (EN) datasets. The evaluation covers both open-source and proprietary models, with open-source models ranging from 2B to 14B parameters.

Model	Retrieval				Generation					
	Recall (↑)		EIR (↑)		Completeness (↑)		Hallucination (↓)		Irrelevance (↓)	
	CN	EN	CN	EN	CN	EN	CN	EN	CN	EN
BM25	**74.21**	**58.08**	**4.11**	7.05	**71.89**	63.80	**17.34**	**16.61**	10.77	19.60
GTE-multilingual-Base	52.55	41.61	2.94	5.72	55.17	54.30	28.35	23.01	16.48	22.69
MiniCPM-Embedding	71.67	55.29	4.02	**7.56**	69.89	63.08	20.02	18.79	**10.09**	**18.13**
BGE-M3	72.94	55.10	4.03	6.84	70.24	**64.08**	18.62	17.03	11.14	18.89

Table 2: The performance results (%) of various retrieval models on Chinese (CN) and English (EN) datasets. Metrics include Recall, EIR, Completeness, Hallucination and Irrelevance. We sample 50 queries for each query type in each domain randomly, 2100 queries in total. List of query types can be found at Figure 13

4.3 Hyperparameter Comparison

In the RAG system, various hyperparameters—such as the chunk size, the number of retrieved items (Top-K), and the selection of retrieval models—play a pivotal role in determining overall performance. To examine their impact on our dataset, we first explore different retrieval models, including BM25, GTE-multilingual-Base (Zhang et al., 2024), MiniCPM-Embedding[4], and BGE-M3 (Chen et al., 2024a). Next, with the chunk size fixed at 512, we investigate how varying the Top-K retrieval value affects the results. Finally, we assess the impact of 3 distinct chunk Top-K selection strategies on completeness under different scenarios. In these experiments, we randomly select 50 samples from all query types, consistent with the data proportions described in Section 4.2. All other parameters remain identical to those in the main experimental setup, and we employ the Llama3-8B-Instruct model for testing. Through these hyperparameter evaluations, we aim to develop a more comprehensive understanding of our DragonBall dataset.

4.3.1 Retrieval Model observation

Our experiments shown in Table 2 demonstrate a strong correlation between retrieval metrics (Recall, EIR) and downstream generation quality. For instance, BM25 achieves the highest Recall (74.21% CN) and simultaneously attains the best Completeness (71.89% CN), aligning with expectations. High EIR score also indicates normally low Hallucination, for instance BM25 has the lowest Hallucination and the highest EIR. However, retrieval superiority alone does not guarantee optimal generation performance—BGE-M3 exhibits marginally lower Recall (55.10% EN) yet best Completeness (64.08% EN). We hypothesize that while BM25 effectively retrieves keyword-matching chunks, it may miss contextually nuanced passages requiring deeper reasoning, which are also critical for keypoint coverage in generation tasks.

Notably, the strong performance of BM25 (surpassing dense retrievers like BGE-M3 in generation metrics) can be attributed to two factors: 1) Queries in our benchmark often contain explicit

[4] https://huggingface.co/openbmb/MiniCPM-Embedding

TopK	Retrieval		Generation					
	Recall (↑)		Completeness (↑)		Hallucination (↓)		Irrelevance (↓)	
	CN	EN	CN	EN	CN	EN	CN	EN
2	49.18	38.16	55.04	51.29	24.52	22.29	20.45	26.42
5	72.94	55.10	70.38	63.96	18.63	16.80	10.99	19.24
8	**78.94**	**64.05**	**72.32**	**69.41**	**17.10**	**13.96**	**10.58**	**16.63**

Table 3: TopK Performance Results (%).

keywords that align with document chunks, and 2) The limited number of relevant references per query allows simple methods to dominate when retrieving top-5 passages. This contrasts with GTE-multilingual-Base, which underperforms in both retrieval (52.55% CN Recall) and generation (55.17% CN Completeness), likely due to its suboptimal cross-lingual alignment.

4.3.2 TopK Retrieval Observations.

As shown in Table 3, increasing TopK from 2 to 8 improves Recall by 60.51% (CN) and 67.8% (EN) relative to Recall at TopK = 2, confirming that broader retrieval enhances coverage of critical information. Due the no-equivalent of the total tokens retrieved, we don't include EIR metric in Table 3 since it would be useless.

Notably, generation quality exhibits diminishing returns: expanding TopK from 2 to 5 boosts Completeness by 27.87% (CN) and 24.70% (EN), whereas further increasing to TopK=8 yields only marginal gains (2.76% CN, 8.52% EN). This suggests that while initial retrieval expansion (2→5) addresses core information gaps, subsequent additions (5→8) primarily refine minor details.

Balancing Retrieval Breadth and Noise. While increasing TopK generally improves generation robustness, excessive expansion risks introducing irrelevant passages that may overwhelm the LLM's processing capacity. Although our current results show reduced hallucination with larger TopK, this trend could reverse in scenarios with lower retrieval precision, where noisy inputs mislead the generator. Thus, selecting an optimal TopK—sufficiently large to capture key information yet within the LLM's context window constraints—is critical.

4.3.3 Different hyper-parameter trend across three Question Types

As shown in Fig 3, the 3 query types (FQ, MRQ, and NCQ) respond differently to changes in the Chunk-TopK configuration. FQ achieves its highest Completeness at 256-10, MRQ peaks at 512-5, and NCQ performs best at 128-20, demonstrating that each query type requires a distinct configuration for optimal performance.

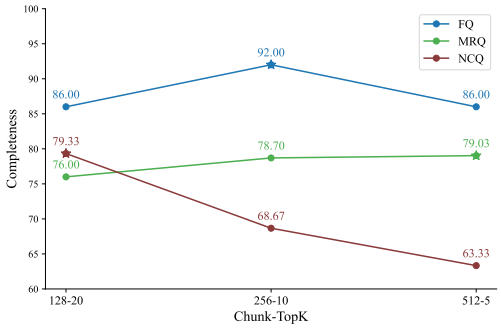

Figure 3: Results (%) of Completeness of different query types under different Chunk-TopK settings on finance scenario in English dataset. We test three query types: Factual Question (FQ), Multi-hop Reasoning Question (MRQ), Numerical Comparison Question (NCQ).

These results highlight that no single configuration uniformly benefits all query types. Instead, each query type demands a tailored Chunk-TopK setting. This underscores our core insight: adapting retrieval-augmented generation to the specific characteristics of each query type leads to more robust performance across different scenarios.

5 Conclusion

This paper introduces `RAGEval`, a framework for rapidly generating scenario-specific datasets to evaluate RAG systems. Our approach addresses the limitations of existing benchmarks by prioritizing factual accuracy and scenario-specific knowledge, which are critical across industries. Experimental results show that our metrics offer a more comprehensive and accurate RAG assessment in specific scenarios compared to conventional ones. GPT-4o

outperforms overall, but the performance gap with top open-source models is small, showing potential for improvement. Our experiments also demonstrate that scenario-specific settings are crucial for RAG assessment. Future work could explore extending the framework to diverse scenarios and further close the performance gap in RAG systems.

Limitations

We highlight two primary limitations of our framework. First, the text generation component heavily relies on large language models, which may produce hallucinations despite our careful prompt design and validation steps. Second, using advanced closed-source models can be costly, although open-source alternatives can help mitigate expenses.

References

AI@Meta. 2024. Llama 3 model card.

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 preprint arXiv:2403.03187*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Tilmann Bruckhaus. 2024. Rag does not work for enterprises. *Preprint*, arXiv:2406.04369.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2318–2335, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024b. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.

Jiajie Jin, Yutao Zhu, Xinyu Yang, Chenghao Zhang, and Zhicheng Dou. 2024. Flashrag: A modular toolkit for efficient retrieval-augmented generation research. *CoRR*, abs/2405.13576.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 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. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Jerry Liu. 2023. Building production-ready rag applications.

Yuanjie Lyu, Zhiyu Li, Simin Niu, Feiyu Xiong, Bo Tang, Wenjin Wang, Hao Wu, Huanyong Liu, Tong Xu, and Enhong Chen. 2024. Crud-rag:

A comprehensive chinese benchmark for retrieval-augmented generation of large language models. *arXiv preprint arXiv:2401.17043*.

Isabelle Nguyen. 2023. Evaluating rag part i: How to evaluate document retrieval.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2017. MS MARCO: A human-generated MAchine reading COmprehension dataset.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878, Bangkok, Thailand. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2023. Ares: An automated evaluation framework for retrieval-augmented generation systems. *arXiv preprint arXiv:2311.09476*.

Yixuan Tang and Yi Yang. 2024. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, et al. 2024. Crag–comprehensive rag benchmark. *arXiv preprint arXiv:2406.04744*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, et al. 2024. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412.

A Schema Refinement

Our refinement constitutes only a small portion of the overall schema and is primarily focused on optimizing its format to better align with the requirements for generating configurations and documents. Specifically, the schema is represented in JSON format. In the original schema, the keys for events were often direct descriptions of the events themselves (e.g., "Major Asset Acquisition"), while the corresponding values were dictionaries composed of fields such as "time", "description", and "impact". However, this structure is not conducive to generating configurations or handling the schema universally in code.

To address this issue, we implemented a manual refinement process, transforming most schema keys into more generic names. For instance, "Major Asset Acquisition" was converted from a specific key into a value in a general dictionary structure comprising the following fields: "event", "time", "description", and "impact". This refinement not only standardizes and unifies the schema structure but also facilitates universal handling and extensibility in subsequent configuration generation and code processing. See figure 9 and 10 for example. Due to the rapid advancement and remarkable progress in model capabilities, current models can now meet such requirements with some constraints in the prompt, potentially eliminating the need for manual refinement in the future.

B Document Generation

To ensure content consistency in complex document generation, we have developed a hierarchical configuration generation mechanism. The implementation involves three key phases: First, constructing fundamental event schemas that may contain multiple sub-events. The configuration parameters of these base events are then fed back as secondary inputs to drive sub-event generation. For financial reports (particularly multi-sectional filings), we employ a modular approach: generating structured outline configurations for each section first, then producing and integrating content based on these outlines. Additionally, pre-generated standardized company profiles are dynamically embedded into documents to maintain consistent corporate descriptions throughout the report.

5: The response is completely correct and fluent.
4: The response is correct but includes redundant information.
3: Most of the response is correct.
2: About half of the response is correct.
1: A small part of the response is correct, or there are logical errors.
0: The response is irrelevant or completely incorrect.

Figure 4: QRA quality scoring criteria.

Safety: Avoidance of real-world sensitive information.
Clarity: Clear and specific information.
Conformity: Resemblance to real documents like financial reports or medical records.
Richness: Depth and breadth of information.

Figure 5: Document quality comparison criteria.

C Quality Assessment

In this section, we introduce the human verification process used to assess the quality of the generated dataset and the evaluation. The assessment is divided into three main tasks: QRA validation, generated documents quality assessment, and automated evaluation validation.

QRA Quality Assessment. We ask 8 annotators to assess the quality of the QRAs by scoring the correctness of the QRAs generated under different configurations according to the standards listed in Figure 4. Those annotators are highly educated students or researchers with enough background knowledge for certain annotated fields and are adequately paid for after the annotations. We randomly select ten samples per question type for every language and scenario, resulting in 420 samples in total for annotation. When scoring, annotators are provided with the document, question, question type, generated response, and references. The results from Table 4 indicate that the QRA quality scores are consistently high across different scenarios, with slight variations between languages. Specifically, the combined proportion of scores 4 and 5 for all scenarios is approximately 95% or higher. This suggests that our approach maintains a high standard of accuracy and fluency in QRAs.

Document Quality Assessment. We evaluate the quality of the documents generated using `RAGEval` by comparing them with documents generated using baseline methods, which include zero-shot prompting (to ask the LLM to generate the document given only a scenario prompt) and one-shot prompting (to ask the LLM to generate the document given a scenario prompt and a sample document). We randomly select 20, 20, and 19 generated documents for finance, legal, and medical scenarios for both languages, respectively, and pack each document with 2 baseline documents generated by zero- and one-shot prompting into one group for comparison. Annotators are asked to rank the documents in each group in terms of clarity, safety, richness, and conformity, as defined in Figure 5, with ties allowed. Results shown in Figure 6 demonstrate that our method consistently outperforms zero-shot and one-shot methods across all criteria, particularly in safety, clarity, conformity, and richness. Specifically, for the Chinese and English datasets across the three aspects of richness, clarity, and safety, our method ranks first in over 85% of the cases. This demonstrates the effectiveness of our approach in generating high-quality articles with diverse and rich content without compromising safety and clarity.

Validation of Automated Evaluation. To validate the consistency between LLM evaluations and human assessments, we compare the LLM-reported metrics—completeness, hallucination, and irrelevance—with those provided by human evaluators. Specifically, we selected the top five results for each question type across various scenarios, covering both Chinese and English, from the Baichuan-2-7B-chat model. This process yielded a total of 210 annotated questions, with each question evaluated by three independent annotators. The annotations were then aggregated using a voting mechanism, classifying each keypoint as either "relevant to the answer," "irrelevant," or "contradictory to the answer." We then calculate the three metrics and compare them with LLM-annotated results. Results in Figure 7 show that the machine and human evaluations show a high degree of alignment in all metrics. The final absolute difference between the human evaluation and the machine evaluation is 1.67%. The Fleiss' Kappa value between the three annotators is 0.7686. This validates the reliability of our automated evaluation metrics and confirms their consistency with human judgment.

In summary, the human evaluation results highlight the robustness and effectiveness of our method in generating accurate, safe, and rich content across

	Finance	Law	Medical
CN	4.94	4.81	4.76
EN	4.84	4.79	4.87

Table 4: QAR quality human review scores by domain.

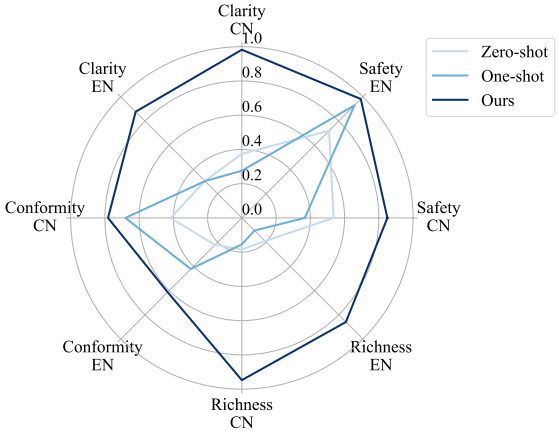

Figure 6: Document generation comparison by scenario.

various scenarios, as well as the reliability of our automated evaluation metrics in reflecting human judgment.

D DragonBall Dateset Details

For document generation, the dataset includes texts from 20 different corporate scenarios in finance, with one randomly selected text per scenario; 10 different legal scenarios, with two randomly selected texts per scenario; and 19 major medical categories, each with two subcategories and one randomly selected text per major category. This ensures a balanced number of human-evaluated documents across finance, law, and medical scenarios.

scenario	Language	Document Count
Finance	CN & EN	40 & 40
Legal	CN & EN	30 & 30
Medical	CN & EN	38 & 38

Table 5: Distribution of Documents in the DRAGONBALL Dataset, in total, we have 6711 questions.

In Table 5, we present a detailed breakdown of the DRAGONBall dataset. The first section of the table shows the distribution of documents across the three scenarios (finance, legal, and medical)

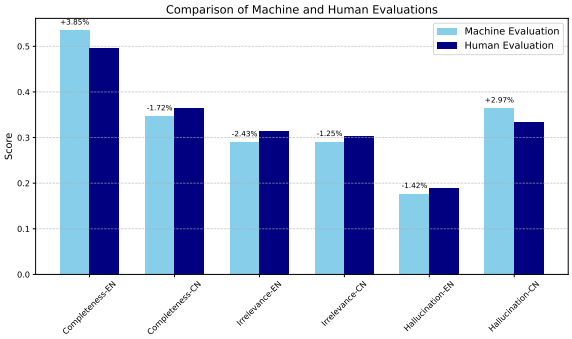

Figure 7: Automated metric validation results. We show the absolute differences between the two evaluations.

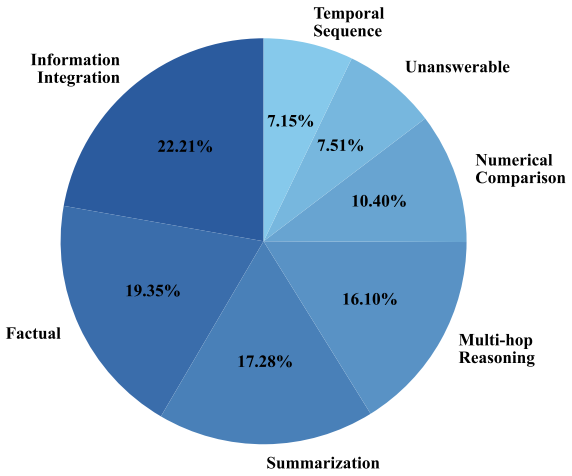

Figure 8: Questions type ratios of DragonBall.

in both Chinese (CN) and English (EN), with an equal number of documents for each language. The second section categorizes the types of questions included in the dataset, providing percentages for each type. The third section details the distribution of the number of reference documents used in answering the questions, reflecting the complexity and variability of the dataset. In total, the dataset comprises 6711 questions.

To ensure the high quality of the QRA triples, we first consider the balance and diversity among the different question types, and then we remove homogeneous and meaningless questions. For example, if the number of unanswerable questions is insufficient, we supplement them according to the article. Second, we eliminate redundant references and answer statements and correct logical reasoning errors in the answers to ensure the dataset quality.

The dataset and the framework will be released under a CC-BY-NC license to ensure its safe use.

E Examples

Model	Retrieval			
	Recall (↑)		EIR (↑)	
	CN	EN	CN	EN
BM25	80.69	**66.17**	4.46	8.80
GTE-multilingual-Base	65.24	52.64	3.69	7.83
MiniCPM-Embedding	81.23	65.85	4.62	**9.47**
BGE-M3	**82.54**	64.98	**4.64**	8.60

Table 6: The performance results (%) of various retrieval models on DragonBall dataset. The primary metrics evaluated include Recall and EIR. We test all queries in Dragonball dataset.


```
{
 "courtAndProcuratorate": {
  "court": "",
  "procuratorate": ""
 },
 "chiefJudge": "",
 "judge": "",
 "clerk": "",
 "defendant": {
  "name": "",
  "gender": "",
  "birthdate": "",
  "residence": "",
  "ethnicity": "",
  "occupation": ""
 },
 "defenseLawyer": {
    "name": "",
    "lawFirm": ""
 },
 "caseProcess": {
   "Case Filing and Investigation": {
    "date": ""
   },
   "Detention Measures Taken": {
    "date": ""
   },
   "Criminal Detention": {
    "date": ""
   },
   "Arrest": {
    "date": ""
   }
 },
 "criminalFacts": {
  "Crime Name": {
   "details": [
    {
     "timePeriod": "",
     "behavior": "",
     "evidence": ""
    }
   ]
  }
 },
 "legalProcedure": {
  "judgmentDate": "",
  "judgmentResult": {
   "Crime Name": {
    "sentence": "",
    "sentencingConsiderations": ""
   }
  }
 }
}
```


Figure 9: A schema example of Law scenario before refinement.

```
{
 "courtAndProcuratorate": {
  "court": "",
  "procuratorate": ""
 },
 "chiefJudge": "",
 "judge": "",
 "clerk": "",
 "defendant": {
  "name": "",
  "gender": "",
  "birthdate": "",
  "residence": "",
  "ethnicity": "",
  "occupation": ""
 },
 "defenseLawyer": {
    "name": "",
    "lawFirm": ""
 },
 "caseProcess": [
   {
    "event": "Case Filing and Investigation",
    "date": ""
   },
   {
    "event": "Detention Measures Taken",
    "date": ""
   },
   {
    "event": "Criminal Detention",
    "date": ""
   },
   {
    "event": "Arrest",
    "date": ""
   }
 ],
 "criminalFacts": [
  {
   "crimeName": "",
   "details": [
    {
     "timePeriod": "",
     "behavior": "",
     "evidence": ""
    }
   ]
  }
 ],
 "legalProcedure": {
  "judgmentDate": "",
  "judgmentResult": [
   {
    "crimeName": "",
    "sentence": "",
    "sentencingConsiderations": ""
   }
  ]
 }
}
```


Figure 10: A schema example of Law scenario after refinement.

```
{
    "courtAndProcuratorate": {
        "court": "Ashton, Clarksville, Court",
        "procuratorate": "Ashton, Clarksville, Procuratorate"
    },
    "chiefJudge": "M. Gray",
    "judge": "H. Torres",
    "clerk": "A. Brown",
    "defendant": {
        "name": "J. Gonzalez",
        "gender": "female",
        "birthdate": "15th, June, 1999",
        "residence": "53, Bayside street, Clarksville",
        "ethnicity": "Hispanic",
        "occupation": "Senior Inspector, Clarksville Tax Department"
    },
    "defenseLawyer": {
        "name": "M. Smith",
        "lawFirm": "Clarksville Legal Associates"
    },
    "caseProcess": [
        {
            "event": "Case Filing and Investigation",
            "date": "1st March 2023"
        },
        {
            "event": "Detention Measures Taken",
            "date": "5th March 2023"
        },
        {
            "event": "Criminal Detention",
            "date": "10th March 2023"
        },
        {
            "event": "Arrest",
            "date": "12th March 2023"
        }
    ],
    "criminalFacts": [
        {
            "crimeName": "Crime of Bending the Law for Personal Gain",
            "details": [
                {
                    "timePeriod": "January 2022 - December 2022",
                    "behavior": "J. Gonzalez utilized her position as Senior Inspector in ...",
                    "evidence": "Email correspondences between J. Gonzalez and ..."
                }
                ...
            ]
        }
    ],
    "legalProcedure": {
        "judgmentDate": "15th May 2023",
        "judgmentResult": [
            {
                "crimeName": "Crime of Bending the Law for Personal Gain",
                "sentence": "5 years of fixed-term imprisonment",
                "sentencingConsiderations": "The defendant's position of trust ..."
            }
        ]
    }
}
```


Figure 11: A config example of Law scenario.

```
"content":"Ashton, Clarksville Court
Case No. XYZ12345

                                **IN THE COURT OF ASHTON, CLARKSVILLE**
                                    **CRIMINAL JUDGMENT**

**COURT AND PROSECUTOR INFORMATION**

**Court:**
Ashton, Clarksville Court

**Prosecutorate:**
Ashton, Clarksville Procuratorate

**Chief Judge:**
M. Gray

**Judge:**
H. Torres

**Clerk:**
A. Brown

**DEFENDANT AND DEFENSE LAWYER INFORMATION**

**Defendant:**
Name: J. Gonzalez
Gender: Female
Birthdate: 15th June 1999
Residence: 53 Bayside Street, Clarksville
Ethnicity: Hispanic
Occupation: Senior Inspector, Clarksville Tax Department

**Defense Lawyer:**
Name: M. Smith
Law Firm: Clarksville Legal Associates

**CASE PROCEDURES**

The case against J. Gonzalez commenced with an investigation following a suspicious tip received by
     the Ashton, Clarksville Procuratorate on 1st March 2023. The investigation revealed substantial
     evidence implicating the defendant in the Crime of Bending the Law for Personal Gain.
     Consequently, J. Gonzalez was taken into detention on 5th March 2023. Criminal detention was
     applied on 10th March 2023, and the defendant was formally arrested on 12th March 2023.

**CASE STATEMENT**

The Crime of Bending the Law for Personal Gain by the defendant, J. Gonzalez, occurred over a span
     of one year, from January 2022 to December 2022. During this period, J. Gonzalez exploited her
     position as a Senior Inspector within the Clarksville Tax Department to manipulate tax audits,
     reports, and reduce penalty fees for several conspiring local businesses in exchange for
     substantial financial bribes. This court will detail the pertinent events chronologically to
     provide a comprehensive understanding of the criminal activities committed.

**Charge:**
Crime of Bending the Law for Personal Gain as per Article 397 of the applicable law.

**EVIDENCE DESCRIPTION**

**1. January 2022 - December 2022: Manipulation of Tax Audits in Exchange for Bribes**

During the year of 2022, J. Gonzalez engaged in illicit activities using her privileged position.
     Emails confirmed numerous correspondences between J. Gonzalez and various local business owners.
      These emails explicitly outlined her agreement to manipulate tax audits and financial reports
     for monetary compensation. Bank statements revealed a series of significant transactions
     amounting to $125,000 deposited into an account owned by J. Gonzalez from suspicious sources.
     Testimonies from several business owners corroborated these findings, revealing a consistent
     pattern of bribery and exploitation.

...

**Date of Judgment:**
15th May 2023

**___**
M. Gray, Chief Judge
**___**
H. Torres, Judge
**___**
A. Brown, Clerk"
```

Figure 12: A document example of Law scenario.

```
{
    "qa_fact_based": [
        {
            "Question Type": "Factual Question",
            "Question": "According to the court judgment of Ashton, Clarksville, Court, what was the
      judgment date?",
            "ref": [
                "Date of Judgment: 15th May 2023"
            ],
            "Answer": "15th May 2023."
        }
    ],
    "qa_multi_hop": [
        {
            "Question Type": "Multi-hop Reasoning Question",
            "Question": "According to the judgment of Ashton, Clarksville, Court, how many instances
      of bending the law for personal gain did J. Gonzalez commit?",
            "ref": [
                "The Crime of Bending the Law for Personal Gain by the defendant, J. Gonzalez,
     occurred over a span of one year, from January 2022 to December 2022.",
                "During this period, J. Gonzalez exploited her position as a Senior Inspector within
      the Clarksville Tax Department to manipulate tax audits, reports, and reduce penalty fees for
     several conspiring local businesses in exchange for substantial financial bribes.",
                "In March 2022, J. Gonzalez revised the tax records for Sunrise Construction Inc.,
     drastically reducing their tax liability after receiving a bribe of $50,000.",
                "In exchange for $30,000, J. Gonzalez facilitated the undue reduction of penalty
     fees levied on Downtown Boutique Ltd. for late tax submissions.",
                "The most egregious of the offenses occurred in November 2022, when J. Gonzalez
     disclosed sensitive and confidential information about ongoing tax investigations to executives
     at Riven Pharmaceuticals, securing a bribe of $45,000."
            ],
            "Answer": "According to the judgment, J. Gonzalez committed four instances of bending
     the law for personal gain: manipulating tax audits and reports, altering tax records, reducing
     penalty fees, and providing confidential information."
        }
    ],
    "qa_summary": [
        {
            "Question Type": "Summary Question",
            "Summary Content": "Facts of the crime",
            "Question": "According to the judgment of Ashton, Clarksville, Court, summarize the
     facts of J. Gonzalez's crimes.",
            "ref": [
                "The Crime of Bending the Law for Personal Gain by the defendant, J. Gonzalez,
     occurred over a span of one year, from January 2022 to December 2022.",
                "During this period, J. Gonzalez exploited her position as a Senior Inspector within
      the Clarksville Tax Department to manipulate tax audits, reports, and reduce penalty fees for
     several conspiring local businesses in exchange for substantial financial bribes.",
                "In March 2022, J. Gonzalez revised the tax records for Sunrise Construction Inc.,
     drastically reducing their tax liability after receiving a bribe of $50,000.",
                "In exchange for $30,000, J. Gonzalez facilitated the undue reduction of penalty
     fees levied on Downtown Boutique Ltd. for late tax submissions.",
                "The most egregious of the offenses occurred in November 2022, when J. Gonzalez
     disclosed sensitive and confidential information about ongoing tax investigations to executives
     at Riven Pharmaceuticals, securing a bribe of $45,000."
            ],
            "Answer": "J. Gonzalez, a Senior Inspector at the Clarksville Tax Department, committed
     the crime of bending the law for personal gain. From January 2022 to December 2022, she
     manipulated tax audits and reports in exchange for bribes from multiple local businesses. In
     March 2022, she altered tax records to reduce the tax liability for Sunrise Construction Inc.
     after receiving $50,000. In August 2022, she reduced penalty fees for late tax submission of
     Downtown Boutique Ltd. in exchange for $30,000. In November 2022, she provided confidential
     information about ongoing tax investigations to Riven Pharmaceuticals in exchange for $45,000."
        }
    ]
}
```


Figure 13: A QRA example of Law scenario.

```
{
"prompt":"In this task, you will be given a question and a standard answer. Based on the standard
     answer, you need to summarize the key points necessary to answer the question. List them as
     follows:

1. ...
2. ...
   and so on, as needed.

Example:
Question: What are the significant changes in the newly amended Company Law?
Standard Answer: The 2023 amendment to the Company Law introduced several significant changes.
     Firstly, the amendment strengthens the regulation of corporate governance, specifically
     detailing the responsibilities of the board of directors and the supervisory board [1]. Secondly
     , it introduces mandatory disclosure requirements for Environmental, Social, and Governance (ESG
     ) reports [2]. Additionally, the amendment adjusts the corporate capital system, lowering the
     minimum registered capital requirements [3]. Finally, the amendment introduces special support
     measures for small and medium-sized enterprises to promote their development [4].
Key Points:

1. The amendment strengthens the regulation of corporate governance, detailing the responsibilities
     of the board of directors and the supervisory board.
2. It introduces mandatory disclosure requirements for ESG reports.
3. It adjusts the corporate capital system, lowering the minimum registered capital requirements.
4. It introduces special support measures for small and medium-sized enterprises.

Question: Comparing the major asset acquisitions of Huaxia Entertainment Co., Ltd. in 2017 and Top
     Shopping Mall in 2018, which company's acquisition amount was larger?
Standard Answer: Huaxia Entertainment Co., Ltd.'s asset acquisition amount in 2017 was larger [1],
     amounting to 120 million yuan [2], whereas Top Shopping Mall's asset acquisition amount in 2018
     was 50 million yuan [3].
Key Points:

1. Huaxia Entertainment Co., Ltd.'s asset acquisition amount in 2017 was larger.
2. Huaxia Entertainment Co., Ltd.'s asset acquisition amount was 120 million yuan in 2017.
3. Top Shopping Mall's asset acquisition amount was 50 million yuan in 2018.

Question: Comparing the timing of sustainability and social responsibility initiatives by Meihome
     Housekeeping Services Co., Ltd. and Cultural Media Co., Ltd., which company initiated these
     efforts earlier?
Standard Answer: Meihome Housekeeping Services Co., Ltd. initiated its sustainability and social
     responsibility efforts earlier [1], in December 2018 [2], whereas Cultural Media Co., Ltd.
     initiated its efforts in December 2019 [3].
Key Points:

1. Meihome Housekeeping Services Co., Ltd. initiated its sustainability and social responsibility
     efforts earlier.
2. Meihome Housekeeping Services Co., Ltd. initiated its efforts in December 2018.
3. Cultural Media Co., Ltd. initiated its efforts in December 2019.

Question: Based on the 2017 Environmental and Social Responsibility Report of Green Source
     Environmental Protection Co., Ltd., how did the company improve community relations through
     participation in charitable activities, community support and development projects, and public
     service projects?
Standard Answer: Green Source Environmental Protection Co., Ltd. improved community relations
     through several social responsibility activities. Firstly, in March 2017, the company
     participated in or funded charitable activities and institutions to support education, health,
     and poverty alleviation, enhancing the company's social image and brand recognition [1].
     Secondly, in June 2017, the company invested in the local community, supporting education,
     health, and social development projects, deepening its connection with the community and
     promoting overall community well-being and development [2]. Finally, in August 2017, the company
      participated in public service projects such as urban greening and public health improvement
     projects, enhancing the quality of life in the community and promoting sustainable development
     [3]. These measures enhanced public perception of the company and improved community relations
     [4].
Key Points:

1. In March 2017, the company participated in or funded charitable activities and institutions to
     support education, health, and poverty alleviation, enhancing the company's social image and
     brand recognition.
2. In June 2017, the company invested in the local community, supporting education, health, and
     social development projects, deepening its connection with the community and promoting overall
     community well-being and development.
3. In August 2017, the company participated in public service projects such as urban greening and
     public health improvement projects, enhancing the quality of life in the community and promoting
      sustainable development.
4. These measures enhanced public perception of the company and improved community relations.

Test Case:
Question: {question}
Standard Answer: {ground_truth}
Key Points:"
```

Figure 14: Key points generation prompt.

```
{"prompt":"""In this task, you will receive a question, a generated answer, and multiple key points
     from a standard answer. Please categorize each key point by determining whether it is Relevant,
     Irrelevant, or Wrong based on the generated answer. For each key point, provide a brief analysis
      and conclude with one of the following classifications:

[[[Relevant]]] indicates that the generated answer contains key information that is related to and
     consistent with the key point described in the standard answer.
[[[Irrelevant]]] indicates that the generated answer does not contain or involve information related
      to the key point in the standard answer.
[[[Wrong]]] indicates that the generated answer contains information related to the key point but it
      is incorrect or contradicts the standard answer keypoints.

Ensure that each key point is categorized into only one of the three categories. Provide your
     analysis and conclusion for each key point sequentially.

Example 1:

Question: What is the current situation and future outlook for wealth management fee income?

Generated Answer: In the first three quarters, the bank achieved wealth management fee income of
     5.105 billion yuan, down 20.7% year-on-year. This decline was mainly due to the active reduction
      of non-standard product scales and the decline in fund sales caused by fluctuations in the
     securities market. To address this situation, the bank plans to continue upgrading its products,
      teams, professionalism, and technological capabilities. Specific measures include improving
     asset selection and product screening capabilities, enriching and optimizing product shelves,
     continuously strengthening the construction of professional teams, and focusing on enhancing
     team professional skills and middle-office empowerment levels to enhance the endogenous
     development capabilities of the teams. The future outlook includes achieving more stable and
     higher-quality development through these measures to cope with the challenges of declining
     wealth management fee income.

Standard Answer Key Points:
Here are 2 key points
1. Current situation of wealth management fee income: Income in the first three quarters decreased
     by 21.6% year-on-year.
2. Future measures: Plans to upgrade products and teams to achieve stable development.

Key Point Evaluation:

Key Point 1:
Comparable fragment of the generated answer: "wealth management fee income of 5.105 billion yuan,
     down 20.7% year-on-year."
Analysis: The generated answer mentions a decrease of 20.7% year-on-year, whereas the standard key
     point specifies a decrease of 21.6%. The percentage difference makes this information incorrect.
Conclusion: [[[Wrong]]]
Key Point 2:
Comparable fragment of the generated answer: "plans to continue upgrading its products, teams,
     professionalism, and technological capabilities."
Analysis: The generated answer aligns with the standard key point by detailing plans to upgrade
     products and teams to achieve more stable development.
Conclusion: [[[Relevant]]]
.... omit three example here

Before you begin the evaluation, please pay attention to the following points:

1. [[[Wrong]]] should only be assigned when there is a specific factual or logical conflict between
     the key point and the generated answer. If important content is missing, it should be
     categorized as [[[Irrelevant]]], not [[[Wrong]]]. More special cases should refer to point 5
     below.
2. [[[Relevant]]] does not require the generated answer to include all the details. It only needs to
      contain the key information necessary to answer the question. Not all details are required. We
     ensure that each key point in the standard answer is typically necessary, although some details
     might not be important for answering the question. When making judgments, focus only on whether
     the most important information is included and consistent. Also, identical content in different
     forms can be considered relevant as long as the core key information is present.
3. Please ensure that the number of key points evaluated matches the number of key points in the
     standard answer. Each key point must be evaluated; do not skip or over-evaluate any key point.
4. After evaluating the key points, do not repeat your conclusions. Ensure that the total number of
     classifications - [[[Relevant]]], [[[Wrong]]], and [[[Irrelevant]]] - matches the number of key
     points in the standard answer.
.... omit three more instruction

Test cases:
Question: {question}
Generated Answer: {prediction}
Standard Answer Key Points:
Here are {key_points_num} key points
{key_points}
Key Point Evaluation:"""
}
```

Figure 15: Key points evaluation prompt.

Query Type	Retrieval						Generation		
	Recall (↑)			EIR (↑)			Completeness (↑)		
	128-20	256-10	512-5	128-20	256-10	512-5	128-20	256-10	512-5
FQ	98.00	**100.00**	94.00	1.02	2.66	**4.89**	86.00	**92.00**	86.00
IIQ	59.26	68.39	**73.31**	1.43	5.79	**13.62**	77.77	**82.77**	71.60
NCQ	**86.33**	77.33	68.33	1.50	5.45	**9.01**	**79.33**	68.67	63.33
TSQ	77.00	**78.47**	74.93	1.94	6.36	**12.47**	**82.00**	79.67	67.67
MRQ	79.95	**84.74**	84.54	4.56	7.92	**15.21**	76.00	78.70	**79.03**
SQ	55.94	50.21	**57.89**	3.73	7.74	**21.17**	58.92	54.26	**60.12**
UQ	13.00	13.00	**16.00**	0.10	0.39	**1.34**	48.00	48.67	**54.00**
Avg.	67.07	**67.45**	67.00	2.04	5.18	**11.10**	**72.57**	72.10	68.82

Table 7: Chunk-TopK results (%) of various query types on finance scenario in English. Seven query types are evaluated: Factual Question (FQ), Information Integration Question (IIQ), Numerical Comparison Question (NCQ), Temporal Sequence Question (TSQ), Multi-hop Reasoning Question (MRQ), Summarization Question (SQ), Unanswerable Question (UQ).

Query Type	Retrieval						Generation		
	Recall (↑)			EIR (↑)			Completeness (↑)		
	128-20	256-10	512-5	128-20	256-10	512-5	128-20	256-10	512-5
FQ	**100.00**	**100.00**	98.00	1.71	**2.07**	1.90	**98.00**	88.00	90.00
IIQ	**96.00**	91.00	88.00	3.26	**3.69**	3.28	78.40	**78.80**	72.60
NCQ	**92.00**	86.67	90.67	2.84	3.16	**3.22**	84.00	82.67	**85.67**
TSQ	**94.00**	89.00	75.00	3.10	**3.41**	2.73	**90.67**	79.33	71.00
MRQ	**99.00**	98.00	**99.00**	6.73	**8.33**	7.84	**85.79**	80.63	84.37
SQ	**94.04**	90.86	93.25	10.05	12.87	**12.98**	**71.85**	69.33	68.85
UQ	**16.00**	**16.00**	12.00	0.25	**0.30**	**0.30**	21.00	**30.00**	22.00
Avg.	**84.43**	81.65	79.42	3.99	**4.83**	4.61	**75.67**	72.68	70.64

Table 8: Chunk-TopK results (%) of various query types on finance scenario in Chinese. Seven query types are evaluated: Factual Question (FQ), Information Integration Question (IIQ), Numerical Comparison Question (NCQ), Temporal Sequence Question (TSQ), Multi-hop Reasoning Question (MRQ), Summarization Question (SQ), Unanswerable Question (UQ).

Query Type	Retrieval						Generation		
	Recall (↑)			EIR (↑)			Completeness (↑)		
	128-20	256-10	512-5	128-20	256-10	512-5	128-20	256-10	512-5
FQ	**98.00**	94.00	88.83	2.24	**2.52**	2.07	**99.00**	88.00	91.00
IIQ	**84.63**	81.92	79.15	2.77	**2.97**	2.72	78.00	**79.00**	**79.00**
NCQ	**93.00**	78.33	68.67	**2.83**	2.58	2.12	**84.00**	71.67	65.33
TSQ	**81.33**	76.33	67.67	2.58	**2.89**	2.38	**61.33**	60.67	56.00
MRQ	**91.41**	81.93	82.47	11.95	**12.10**	11.23	49.37	46.41	**49.41**
SQ	**74.63**	73.68	74.23	13.98	**14.68**	14.09	58.89	**64.97**	60.80
UQ	**10.00**	7.00	3.00	**0.14**	0.08	0.08	44.00	**57.00**	44.00
Avg.	**76.14**	70.46	66.29	5.21	**5.40**	4.96	**67.85**	66.83	63.67

Table 9: Chunk-TopK results (%) of various query types on law scenario in Chinese. Seven query types are evaluated: Factual Question (FQ), Information Integration Question (IIQ), Numerical Comparison Question (NCQ), Temporal Sequence Question (TSQ), Multi-hop Reasoning Question (MRQ), Summarization Question (SQ), Unanswerable Question (UQ).

Query Type	Retrieval						Generation		
	Recall (↑)			EIR (↑)			Completeness (↑)		
	128-20	256-10	512-5	128-20	256-10	512-5	128-20	256-10	512-5
FQ	85.00	87.00	**90.00**	1.58	2.22	**5.15**	88.00	88.00	**91.00**
IIQ	**75.07**	72.23	64.03	3.33	4.08	**5.20**	**77.50**	77.00	76.17
NCQ	52.00	**64.50**	45.17	2.34	4.17	**5.25**	**54.83**	**54.83**	39.33
TSQ	**79.33**	59.17	52.00	**5.44**	4.40	4.54	**58.67**	45.33	40.00
MRQ	29.42	**30.30**	19.71	3.19	7.50	**8.93**	**41.74**	28.28	23.85
SQ	18.51	25.75	**25.90**	4.23	9.03	**12.86**	33.38	**36.29**	34.89
UQ	**2.00**	**2.00**	**2.00**	**0.12**	**0.12**	0.11	79.37	83.50	**85.83**
Avg.	**48.76**	48.71	42.69	2.89	4.50	**6.01**	**62.01**	59.10	55.93

Table 10: Chunk-TopK results (%) of various query types on law scenario in English. Seven query types are evaluated: Factual Question (FQ), Information Integration Question (IIQ), Numerical Comparison Question (NCQ), Temporal Sequence Question (TSQ), Multi-hop Reasoning Question (MRQ), Summarization Question (SQ), Unanswerable Question (UQ).

Query Type	Retrieval						Generation		
	Recall (↑)			EIR (↑)			Completeness (↑)		
	128-20	256-10	512-5	128-20	256-10	512-5	128-20	256-10	512-5
FQ	85.00	**99.00**	95.00	1.04	**1.31**	0.97	80.00	94.00	**96.00**
IIQ	**86.83**	83.83	83.50	1.95	**2.13**	1.61	**87.00**	83.00	79.00
NCQ	**83.39**	70.20	70.61	**2.90**	2.87	2.39	**84.33**	71.33	73.00
TSQ	**94.00**	**94.00**	87.67	**2.72**	3.21	2.35	87.33	**88.00**	80.33
MRQ	83.36	80.90	**89.66**	4.46	**4.64**	4.44	67.55	62.50	**67.73**
SQ	**85.10**	77.96	83.42	6.27	**6.32**	5.98	**67.63**	62.52	63.15
UQ	**2.00**	0.00	**2.00**	**0.03**	0.00	**0.03**	**64.67**	60.67	63.33
Avg.	**74.24**	72.27	73.12	2.77	**2.93**	2.54	**76.93**	74.57	74.65

Table 11: Chunk-TopK results (%) of various query types on medical scenario in Chinese. Seven query types are evaluated: Factual Question (FQ), Information Integration Question (IIQ), Numerical Comparison Question (NCQ), Temporal Sequence Question (TSQ), Multi-hop Reasoning Question (MRQ), Summarization Question (SQ), Unanswerable Question (UQ).

Query Type	Retrieval						Generation		
	Recall (↑)			EIR (↑)			Completeness (↑)		
	128-20	256-10	512-5	128-20	256-10	512-5	128-20	256-10	512-5
FQ	**98.67**	91.67	80.00	**2.61**	2.24	1.70	**90.00**	**90.00**	88.00
IIQ	**93.27**	80.60	81.77	**4.93**	3.27	4.28	83.00	76.00	**88.00**
NCQ	**90.50**	80.50	44.67	**2.22**	1.49	0.98	**81.00**	74.33	40.67
TSQ	81.00	**97.00**	92.00	**2.82**	2.53	2.81	66.00	**73.33**	72.67
MRQ	**66.17**	64.16	40.46	**5.83**	5.23	2.64	51.22	**55.38**	43.20
SQ	57.42	**64.35**	50.29	13.74	**14.00**	11.47	52.96	**59.79**	52.37
UQ	0.00	0.00	0.00	0.00	0.00	0.00	96.00	90.00	**98.00**
Avg.	**69.57**	68.33	55.60	**4.59**	4.11	3.41	**74.31**	74.12	68.99

Table 12: Chunk-TopK results (%) of various query types on medical scenario in English. Seven query types are evaluated: Factual Question (FQ), Information Integration Question (IIQ), Numerical Comparison Question (NCQ), Temporal Sequence Question (TSQ), Multi-hop Reasoning Question (MRQ), Summarization Question (SQ), Unanswerable Question (UQ).

Question Type	Definition
	Single-document QA
Factual	Questions targeting specific details within a reference (e.g., a company's profit in a report, a verdict in a legal case, or symptoms in a medical record) to test RAG's retrieval accuracy.
Summarization	Questions that require comprehensive answers, covering all relevant information, to mainly evaluate the recall rate of RAG retrieval.
Multi-hop Reasoning	Questions involve logical relationships among events and details within a document, forming a reasoning chain to assess RAG's logical reasoning ability.
	Multi-document QA
Information Integration	Questions that need information from two documents combined, typically containing distinct information fragments, to test cross-document retrieval accuracy.
Numerical Comparison	Questions requiring RAG to find and compare data fragments to draw conclusions, focusing on the model's summarizing ability.
Temporal Sequence	Questions requiring RAG to determine the chronological order of events from information fragments, testing the model's temporal reasoning skills.
	Unanswerable Questions
Unanswerable	Questions arise from potential information loss during the schema-to-article generation, where no corresponding information fragment exists, or the information is insufficient for an answer.

Table 13: DragonBall Dataset question types and their definitions