Are We on the Right Way for Assessing Document Retrieval-Augmented Generation?

Anonymous ACL submission

Abstract

Document retrieval–augmented generation (RAG) systems have shown great promise for enhancing language models with external knowledge, but their evaluation has so far been limited to synthetic or unimodal benchmarks lacking real-world complexity. To address this gap, we introduce CDocRAG-Bench, the first large-scale, multilingual, multimodal benchmark designed specifically for document RAG. Our benchmark assembles over 62,000 pages of multilingual, multi-type documents and synthesizes 2,000 single-hop and 2,000 multi-hop queries with exhaustive evidence labels using fine-grained guidelines and a knowledge-graph-driven pipeline. All ground-truth annotations are refined through expert review to ensure high precision. We evaluate seven state-of-the-art embedding models, and three end-to-end RAG frameworks, demonstrating that multimodal embeddings yield significant retrieval gains of up to 15.48% compared to textual embeddings. Current frameworks still struggle with effective pipelines for multi-page understanding. By diagnosing key shortcomings and offering a comprehensive evaluation framework, CDOCRAG-BENCH provides a rigorous foundation for future research in multimodal document retrieval-augmented generation. The source code and dataset are publicly available at `https://anonymous.4open.science/r/DocRAG_Bench-7D34`.

1 Introduction

Retrieval-Augmented Generation (RAG)(Lewis et al., 2020) is a technique that enhances the accuracy of Large Language Models (LLMs) by integrating an information retrieval system, particularly for knowledge-intensive NLP tasks. Documents, including scanned files(Breci et al., 2024; Crosilla et al., 2025; Gervais et al., 2024), charts(Yang et al., 2025; Masry et al., 2022), and slides (Wasserman et al., 2025; Tanaka et al., 2023), are common information sources that traditionally require significant manual effort to examine. Unlike traditional text retrieval, these documents often contain multimodal information, and directly parsing them to text can be time-consuming and lead to information loss. Evaluating multimodal document RAG systems(Mortaheb et al., 2025a,b; Yu et al., 2024) is therefore essential: it uncovers modality-specific weaknesses, ensures robustness across real-world repositories, and guides the development of retrieval and fusion strategies that generalize beyond plain-text scenarios.

Nevertheless, existing evaluation benchmarks(Friel et al., 2024) for document RAG suffer from critical shortcomings in the context of large-scale, multimodal corpora. Through a pilot screening of popular datasets, we identify three pervasive issues: (1) Unrealistic prior knowledge assumptions. Many VQA-style benchmarks(Wu et al., 2025; Li et al., 2024; Adjali et al., 2024) presuppose that the target page or document is already known, making them unsuitable for global retrieval in a large corpus. (2) Ambiguous or non-unique evidence. Queries are often crafted from a single page and assume a one-to-one mapping between question and evidence, neglecting cases where multiple pages—or multiple interpretations—could satisfy the same informational need. (3) Trivial multi-hop compositions. Synthetic multi-hop queries often reduce to parallel single-hop unions, failing to test true reasoning chains across diverse document modalities. We provide a comprehensive comparison between existing document benchmarks and CDocRAG-Bench in Table 1.

To overcome these limitations, we introduce CDocRAG-Bench (Comprehensive Document Retrieval Augmented Generation Benchmark), built via a three-stage pipeline. First, we assemble and preprocess a richly diverse document corpus

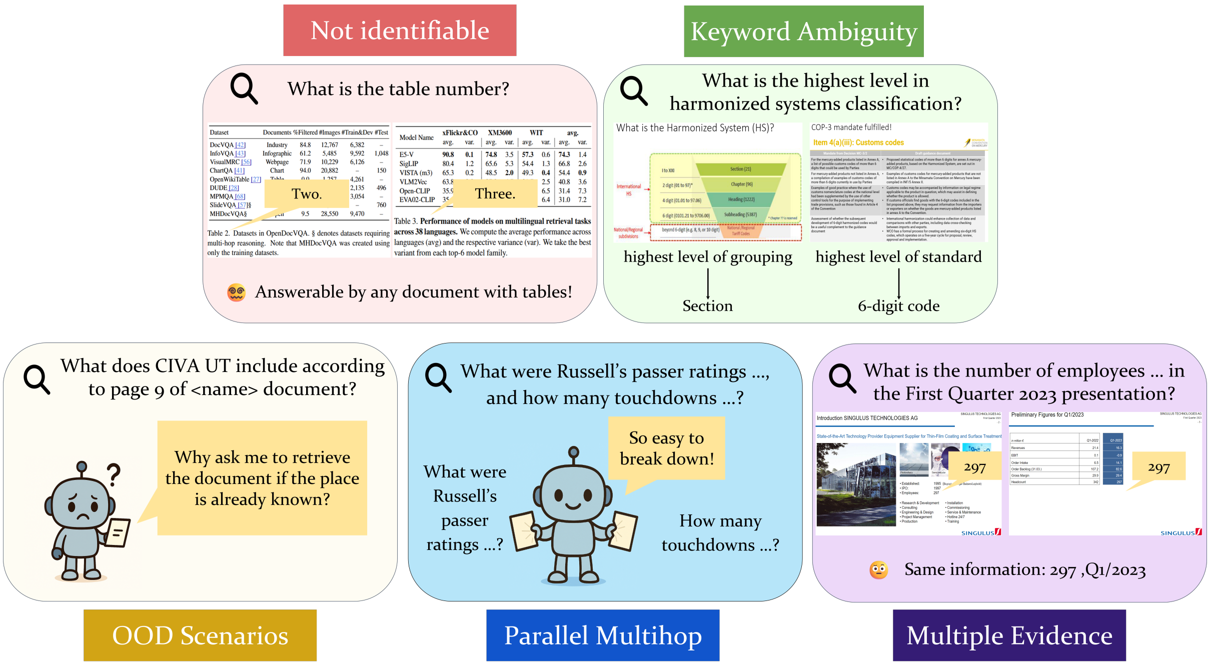

Figure 1: Limitations of Existing Document Benchmarks For RAG Evaluation

spanning printed documents and PDFs, handwritten documents, slides and web pages, by applying two-stage filtering and decomposing each page into text, chart and image segments; next, we automatically synthesize and rigorously label both single-hop and multi-hop queries(Zhang et al., 2025; Tang and Yang, 2024), using a combination of template-driven drafting, knowledge-graph-guided composition and exhaustive evidence search; finally, expert annotators review(Chiang et al., 2024; Chen et al., 2024a; Pu et al., 2025) and correct all machine-assigned evidence to ensure high-precision ground truth for realistic, large-scale multimodal retrieval evaluation.

Using CDocRAG-Bench, we perform an extensive empirical study of seven state-of-the-art embedding models and three advanced document RAG frameworks. Our findings demonstrate that vision-language embeddings markedly improve page retrieval, that end-to-end answer quality is tightly coupled to retrieval performance, and that current RAG frameworks still struggle with arranging effective pipelines for multi page information retrieval and generation. In summary, our contributions are three-fold:

- We diagnose three major limitations in existing document-RAG benchmarks, including unrealistic prior knowledge assumptions, ambiguous or non-unique evidence, and weak multi-hop query design.
- We introduce CDocRAG-Bench, the first large-scale benchmark for multilingual, multimodal RAG evaluation. It features a richly diverse document corpus, fine-grained page decomposition, and high-quality single-hop and multi-hop QA pairs with exhaustively labeled supporting evidence.
- We conduct extensive experiments with seven embedding models and three RAG frameworks, uncovering that vision embeddings enhance retrieval, and current frameworks struggle in multi-page understanding.

2 Pilot Study: Limitations of Existing Benchmarks

2.1 Task Formulation

Let C be a large corpora consisting of documents $\{d_1, d_2, ..., d_n\}$. Each document d_i is stored by page images $\{p_1^i, p_2^i, ..., p_m^i\}$. Given a query Q, the objective is to retrieve top-k possible evidence pages E_r from the entire corpora to formulate the answer A.

For single-hop queries Q_{sh}, answer A can be found if one or more evidence pages from the evidence set E_q is successfully retrieved. For multi-hop queries Q_{mh}, the requirement extends to having one or more evidence page for every evidence set $E_{q,j}$ of each query hop j.

$$(E_r \cap E_q \neq \varnothing) \implies \text{Enable}(A|Q_{sh}, E_r) \quad (1)$$

Table 1: Comparison between existing multimodal document benchmarks and the proposed CDocRAG-Bench, where each symbol represents: printed documents & PDFs handwritten documents slides HTML pages. Half-tick denotes dependent on the specific benchmark component.

Benchmarks	Queries			Labels		Document		
	Clarity	i.i.d.	Multi-Hop	GT Coverage	Multi-hop Chain	avg. # pages	Multilingual	Type
DocVQA	✗	✗	✗	partial	-	1.0	✗	PDF, handwritten
MMLongbench-Doc	✗	½✓	✓	partial	✗	47.5	✗	PDF
UDA-QA	✗	✓	½✓	partial	✗	46.3	✗	PDF, HTML
ViDoRe	½✓	✓	✗	likely complete	-	1.0	✓	PDF, handwritten, slides
ViDoSeek	✓	✗	✓	partial	✗	18.4	✗	slides
CDocRAG-Bench	✓	✓	✓	complete	✓	22.5	✓	PDF, handwritten, slides, HTML

$$\left(\bigwedge_{j=1}^{k}(E_r \cap E_{q,j} \neq \varnothing)\right) \implies \text{Enable}(A|Q_{mh}, E_r) \quad (2)$$

2.2 Three Overlooked Issues for Document RAG Evaluation

Practical document information retrieval typically involves queries that clearly state an informational need, aiming to extract direct answers from large document collections without requiring prior knowledge about the specific documents. Therefore, we start by investigating whether existing benchmarks are fully appropriate for evaluating real-world document RAG scenarios. By screening existing benchmarks with predefined concrete rules, we describe our findings in Figure 1, and report the evaluation results in Fig 6.

Benchmark design issues with prior knowledge assumptions. VQA benchmarks, such as DocVQA (Mathew et al., 2021) and MMLongbench-Doc (Ma et al., 2024), are inherently designed with a given page or document as prior knowledge, rendering them insufficient for effectively identifying the ground-truth page within a global document corpus. Similarly, manually inspected benchmarks designated for RAG, such as ViDoSeek (Wang et al., 2025), have significant gains in query information. However, such benchmarks tend to insert the exact name or page of the ground document, violating the assumption that users do not have any prior knowledge over the corpus. Such queries create gaps between evaluation and real use.

Queries with multiple interpretations and scattered evidence. Most benchmarks construct queries by selecting a ground truth page beforehand, and assume the evidence used is unique (Chen et al., 2024b; Faysse et al., 2024; Tang and Yang, 2024; Saad-Falcon et al., 2023). This generally holds true when the corpus is small enough, such as individual benchmarks in ViDoRe, but the problem grows when the corpus scales up. Some queries may also have unexpected multiple interpretations given different content in the same document. This further undermines the unique assumption.

The linearity in multi-hop query synthesis is overlooked. Some multi-hop queries are trivial conjunctions of single-hop queries. These queries do not require reasoning to dismantle, and thus could be processed in parallel (Hui et al., 2024). Such queries fail to evaluate the reasoning ability of RAG frameworks, overstating actual performance.

3 CDocRAG-Bench Benchmark

To address existing limitations, we introduce CDocRAG-Bench, a benchmark with manually verified multi-modal, multi-lingual content and an automatic benchmark construction suite. CDocRAG-Bench includes 2,776 documents (62,476 pages) in 6 languages, with 2,000 curated single-hop and multi-hop queries spanning vision-and-language content. Rigorous checklist-based inspection and page-by-page examination yield precise list-of-evidence and chain-of-evidence labels, ensuring all relevant ground truth pages are captured for accurate document retrieval system evaluation. Detailed benchmark statistics can be found in Table 5 and Figure 3. Figure 2 shows the three-stage benchmark pipeline. We also provide extensive metadata, such as queried modality, language, evidence chains/lists, parsed page chunks, to advance document RAG research.

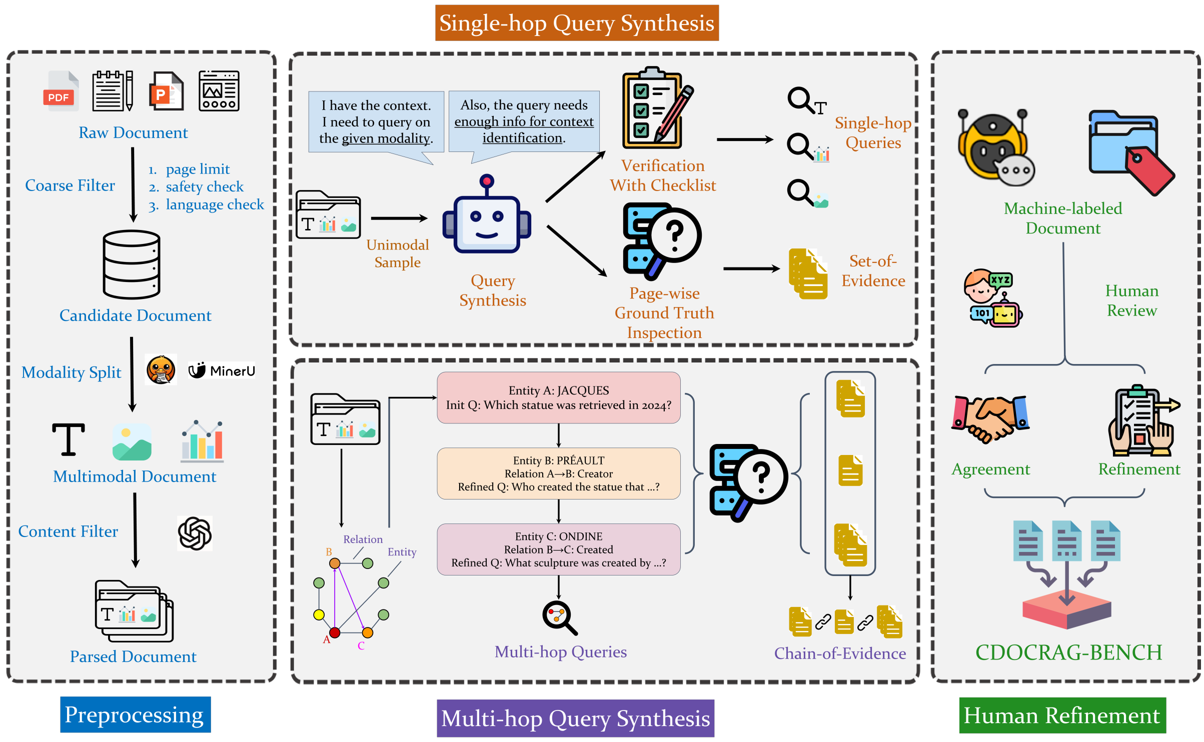

Figure 2: Overview of the construction pipeline for CDocRAG-Bench

3.1 Metadata Collection and Preprocessing

This section details the preprocessing steps applied to the raw document corpus: *(1)* Collection, *(2)* Two-stage Filtering, and *(3)* Modality Split.

Collect To ensure comprehensive evaluation, we collect a diverse range of document types and languages. The initial database comprises four popular types of documents collected from various sources.

- **Printed Documents and PDFs:** Includes high-quality PDFs from DocVQA (Mathew et al., 2021), MMLongBench-Doc (Ma et al., 2024), and recent Arxiv papers[1], all known for rich context and multimodal content.
- **Handwritten Documents:** Sourced from DocVQA (Mathew et al., 2021) and Commoncrawl[2], these present challenges due to variability in size, font, and layout.
- **Slides:** A subset from SlideVQA (Tanaka et al., 2023), augmented with multilingual slides from the Commoncrawl corpus.
- **HTML Pages:** 600 Wikipedia entries, randomly crawled across various topics, with equal non-overlapping samples per language used.

Rule Based Coarse-grained Filter. Following the initial collection, a coarse-grained, rule-based filter(Pu et al., 2025) is applied to select documents meeting basic structural requirements. This step exclusively retains documents with a page count within the range of $[10, 50]$, thereby excluding overly brief or excessively long content. For documents originating from Commoncrawl, GPT-4o (OpenAI, 2024) is utilized to analyze the initial three pages to identify and filter out harmful topics(Ngo et al., 2021) and to determine the document's primary language. We filter out languages other than English, Chinese, Japanese, French, Spanish, and Arabic, since other lanuages mostly receive limited support by existing document parsing frameworks and multilingual embedding models.

Modality Split To achieve better accuracy and queried modality control in the query formulation process, each document undergoes a modality splitting process. Using Docling (Livathinos et al., 2025) and MinerU (Wang et al., 2024), each document page is parsed and decomposed into its constituent modalities: text, charts, and images.

Fine-grained Content Filtering Parsed chunks are provided with adjacent context to GPT-4o to

[1] https://arxiv.org/
[2] https://commoncrawl.org/

determine if the semantics of the chunk fits into the context. Chunks with no actual content or differs significantly from the context will be filtered out. This step ensures that every chunk selected for query formulation has enough semantics for meaningful queries.

3.2 Single-hop Query Synthesis

The single-hop query synthesis process consists of three steps: ***(1)*** query drafting based on given modality and context ***(2)*** query quality inspection with checklist ***(3)*** pagewise ground truth search.

Principles for Single-hop Question. ingle-hop VQA queries often lack enough detail for precise document retrieval. We enhance them by adding supportive descriptions, making queries self-contained, focused on key unimodal information, and diverse in type including factual and analytical. Our method ensures queries are: 1) self-contained and globally valid; 2) focused on substantive, unimodal key information (not superficial or multimodal details); and 3) diverse in type (factual and analytical) to promote genuine understanding, especially of visual content. This produces robust queries for evaluating single-hop retrieval.

Quality Inspection With Checklist. The query drafting module strives to generate high-quality queries but requires a quality inspection to ensure all criteria are met. We carefully curate a comprehensive checklist(Pu et al., 2025) consisting of four key aspects: (a) Prohibition of Explicit Source Referencing (b) Answer Verifiability and Accuracy (c) Question Self-Containment (d) Unimodal Focus of Question Content. All queries that violate any of the aspects listed are discarded.

Pagewise Ground Truth Search. For each validated query, we locate its ground truth by thoroughly searching each document page. Pages are marked as evidence only if they directly provide or lead to the answer. To ensure all supporting information is captured, we conduct a thorough page-by-page search of the entire document. A page is marked as evidence only if its content directly provides or leads to the query's answer. This process identifies the complete set of ground truth pages for each single-hop query, facilitating the evaluation of Eq.1.

3.3 Multi-hop Query Synthesis

Principles. While multi-hop queries benefit from information across hops, mitigating the lack of information problem, their direct generation by LLMs is challenging, even with techniques like Chain-of-Thought or inference scaling. Our multi-hop query synthesis pipeline addresses this by using knowledge graphs. This approach simplifies sub-query linking by replacing key entities with new sub-queries, forming linearly combined queries. Additionally, we introduce a chain-of-evidence labeling strategy to facilitate more accurate evaluations of hop granularity.

Knowledge Graph Construction. For each filtered document, a knowledge graph $G = (E, R)$, where E denotes entities and R denotes relationships, is constructed using LightRAG(Guo et al., 2024). Visual chunks are converted to descriptive captions and combined with same-page textual content. GPT-4o then extracts important entities and relationships per page. A masking procedure ensures unique relationship identifiability, meaning a source entity and relationship uniquely determine the target. While masked relationships count as valid degrees, they are ineligible for query path extension. This unique identifiability is crucial for sub-query synthesis and combination, guaranteeing each sub-query has exactly one answer.

Query Path Selection. The generation of a multi-hop query begins with the selection of a query path through the constructed knowledge graph G. The intuition of the selection process is to provide the LLM with connectivity information of adjacent nodes to guide the iterative next node decision. Specifically, we select the initial node $e_0 \in E^n_{top}$ from entity nodes with top-n degrees. From the current node e_{cur}, the algorithm identifies candidate neighbour-relation pairs (e_{nei}, r) from the knowledge graph that have valid relationships and are unvisited by the current path P_{cur}. The model then decides the next hop e_{next} based on the content depth of the candidate pair, and the degree of the adjacent entity. This LLM-driven selection process ensures that the chosen path forms a coherent and logical reasoning chain, up to a predefined maximum number of hops H_{max}, forming the structural basis for the multi-hop question.

Iterative Synthesis and Combination. Following the selection of a query path through the knowledge graph, the algorithm proceeds to synthesize

Table 2: Answer Accuracy Results of VLMs

Query Type		Single Hop			Multi Hop		
Setting	**Model**	✓	✓✗	✗	✓	✓✗	✗
w.o. RAG	GPT-4o	**0.228**	0.301	**0.471**	0.183	**0.054**	**0.763**
	Gemini 2.5 Pro	0.157	**0.373**	0.470	**0.202**	0.053	0.745
Oracle	GPT-4o	**0.791**	0.092	0.117	0.677	**0.105**	0.218
	Gemini 2.5 pro	0.650	**0.112**	**0.238**	**0.726**	0.098	**0.176**

and combine sub-queries iteratively to form the final multi-hop question. The synthesis starts from drafting the initial query Q_0 with the name of the initial entity e_0 as the answer. For hop i in the selected path (r_i, e_i), a single-hop QA pair (Q_i, A_i) is generated by querying on the relationship r_i between the source and target entities e_{i-1} and e_i. We then instruct the LLM to seemlessly integrate the new generated sub-query Q_i into the cummulative query Q_{cum}. This involves identifying the source entity in Q_i that corresponds to the answer of Q_{cum}, substuing the entity with Q_{cum}, and rearrange the extended Q_{cum} so that it is grammatically correct, natural-sounding, and preserves the logical reasoning chain.

Multi-hop Query Inspection. Due to differences from single-hop queries, we created a separate checklist for multi-hop query quality, assessing: (a) Final question quality (clarity, specificity, no explicit final answer); (b) Logical necessity and correctness of intermediate reasoning steps; (c) Uniqueness of step answers and rigor of relations; and (d) Significance and relevance of the overall query. Queries failing any criteria are discarded.

Chain-of-Evidence Labeling. With the single-hop query splits and intermediate answers obtained, we can process each sub-query in parallel and obtain individual evidence sets. The individual sets are subsequently chained together to produce chain-of-evidence labels for each multi-hop query.

3.4 Human Refinement

To improve benchmark accuracy, we added a human refinement stage. Although automated evidence labeling is accurate, human annotators reviewed and adjusted 8% of labels with discrepancies. With 92% agreement, this step ensures precise and reliable ground truth data, enhancing CDOCRAG-BENCH 's trustworthiness for document RAG research.

4 Experiments

4.1 Evaluation Protocol

Retrieval Accuracy Following the task formulation setting in Section 2.1, we define the hit criteria for retrieval accuracy evaluation of single-hop and multi-hop queries. For single-hop queries, hit@k evaluates to 1 only if any of the top-k retrieved pages hits the list of ground truth list. For multi-hop queries, hit@k evaluates to 1 only if the processed list of pages in a framework includes at least one ground truth page for every hop of the query.

Answer Accuracy We prompt GPT-4o to evaluate the correctness of the generated answer compared to the ground truth answer on a scale of 0 to 10. Scores not lower than 7 count as correct. $[4, 6]$ count as partial. Scores not higher than 3 count as incorrect.

4.2 Evaluated Models and Frameworks

We evaluate 5 competitive open-source document page embedding models, namely colpali(Faysse et al., 2024), colqwen(Faysse et al., 2024), gme(Zhang et al., 2024), vdr-2b(LlamaIndex, 2025), mm-e5(Chen et al., 2025), and 3 advanced document RAG frameworks, namely M3DocRAG(Cho et al., 2024), MDocAgent(Han et al., 2025), VidoRAG(Wang et al., 2025). Baseline results are reported with GPT-4o directly answering the queries without RAG. The single-hopo Oracle setting directly provides GPT-4o with the exact golden page list to estimate the upper-bound performance of the generation stage.

4.3 Experiment Setups

All experiments utilized an 8-card A100 server. For embedding model evaluation, all 2000 single-hop and 2000 multi-hop queries were assessed using the entire corpus for retrieval. Text-based embedding models used parsed document chunks, with

Table 3: Retrieval Accuracy Results of Embedding Models

Query Type	Single Hop			Multi Hop		
Model	hit@1	hit@3	hit@5	hit@1	hit@3	hit@5
Text-based Model						
bge-m3	0.556	0.666	0.706	0.213	0.425	0.510
gte-Qwen2-7B-instruct	0.586	0.738	**0.796**	0.243	0.488	0.606
NV-Embed-v2	**0.645**	**0.762**	**0.796**	**0.288**	**0.542**	**0.650**
VL Embed Model						
colpali-v1.3-merged	0.646	0.737	0.768	0.268	0.514	0.618
gme-Qwen2-VL-7B-Instruct	0.687	0.815	0.850	0.251	0.498	0.606
vdr-2b-multi	0.733	0.836	0.866	0.260	0.518	0.623
colqwen2.5-3b-multilingual	**0.811**	**0.888**	**0.910**	**0.321**	**0.619**	**0.719**

Table 4: Performance of RAG Frameworks

Query Type	**Single Hop**						**Multi Hop**					
Accuracy Type	**Retrieval**			**Answer**			**Retrieval**			**Answer**		
Framework	hit@1	hit@3	hit@5	✓	✗	✗	hit@1	hit@3	hit@5	✓	✗	✗
MDocAgent	**0.768**	**0.876**	**0.904**	**0.742**	0.092	0.166	**0.292**	0.494	0.628	**0.556**	0.124	0.320
ViDoRAG	0.746	**0.876**	0.890	0.546	0.073	0.381	0.283	0.448	**0.684**	0.266	0.058	0.676
M3DOCRAG	0.706	0.806	0.830	0.458	0.082	0.460	0.272	**0.532**	0.662	0.402	0.090	0.508

graphs and figures converted to descriptive captions (generated by Qwen2.5-VL-32b-Instruct) before merging text. In RAG framework evaluation, 500 queries each for single-hop and multi-hop retrieval and answer evaluation were sampled, using involved documents as the corpus. The same page text was used by ViDoRAG and MDocAgent. ViDoRAG could invoke each component up to twice before final answer generation. All frameworks retrieved 5 pages before producing the final answer.

4.4 Main Results

VLMs contain limited information on collected documents, but can effectively identify critical information provided correct context and pipeline. As shown in Table 2, the standalone accuracy of VLMs like GPT-4o and Gemini without context is modest. This indicates limited data leakage within our curated benchmark, suggesting it appropriate for rigorous RAG evaluations. The results stand in stark contrast to their upper bound performance, providing a 57.7% and xxx accuracy improvement, respectively. Notably, the robust performance observed in the upper bound setting, which closely mirrors our benchmark curation pipeline, further suggests the robustness and effectiveness of our pipeline in identifying correct evidence pages of queries.

Performance of embedding models are clearly stratified. CDocRAG-Bench provides a clear differentiation in the retrieval performance of various embedding models, as detailed in Table 3. Vision embedding models generally outperform text embedding models, with a 15.48% average hit improvement between best performing models, colqwen2.5 and NV-Embed. This reflects the importance of visual information over textual descriptions. Furthermore, our benchmark clearly differentiates capabilities even among the visual embedding models themselves, with colqwen2.5 establishing a significant lead. This robust stratification underscores the benchmark's ability to resolve fine-grained performance tiers across different embedding architectures.

Document RAG frameworks rely heavily on retrieval accuracy, while answer synthesis pipeline is also a crucial aspect. The performance of the evaluated RAG frameworks on table 4 show strong correlations between retrieval accuracy and answer accuracy. This emphasizes that the ability of the RAG framework to generate correct answers is fundamentally gated by the quality of its retrieval component. Moreover, while three frameworks all exhibit similar accuracies in the retrieval stage, the observed significant differences in answer accuracies highlight the importance of carefully designed answer synthesis pipelines. ViDoRAG and MDocAgent frameworks trade higher computational costs of agent components for notable answer accuracy improvement.

4.5 Analysis

Time efficiency of frameworks. Model api complete times may vary, therefore Figure 4 reports the normalized time efficiency of the evaluated frameworks. Both MDocAgent and ViDoRAG employ a linear agent cooperation pattern, which considerably increases their inference time. Note ViDoRAG dynamically controls the generation process. We report the lower and upper bound theoretical time efficiency of ViDoRAG estimated by number of api calls.

Inference patterns of VLMs. We also observe interesting inference patterns in VLMs, shown by a case study in Appendix. E. When directly provided with a multi-hop query, VLMs tend not to process them hop-by-hop, contrary to what might be intuitively expected. Instead, they collect signature information – the most distinguishing or identifiable pieces of information – from the various hops. Following this, VLMs seem to perform a direct inclusion based elimination to arrive at the final answer. This observed mechanism differentiates significantly from our conventional expectation of how models might sequentially solve multi-hop queries. Furthermore, this provides a compelling point of view: merely increasing the number of hops within a query may not inherently or proportionally increase its difficulty. This hypothesis warrants thorough investigation in future works.

5 Related Work

Multimodal Document Retrieval Different from traditional text retrieval(Zhao et al., 2024; Hambarde and Proenca, 2023; Gienapp et al., 2024), documents(Masry et al., 2022; Mathew et al., 2021; Tanaka et al., 2023) often contain multimodal information, which may be time consuming and would cause information loss when directly parsed to text. Therefore, recent works have dedicated great effort to improve the accuracy and efficiency of document retrieval with VLMs. One line of work adopts high quality synthetic data (Zhang et al., 2024; Zhou et al., 2024; Chen et al., 2025), hardness aware training (Lan et al., 2025; Lee et al., 2024) and retrieval-optimized network architectures (Faysse et al., 2024) for more precise embedding models. Another line of work leverages LLM/MLLM agentic flows to process different modalities in parallel (Han et al., 2025) and perform iterative inference steps for more grounded and informative answers (Wang et al., 2025). Despite these advancements, our benchmarks indicate that even the most powerful existing frameworks struggle to achieve satisfactory accuracy, highlighting significant room for improving RAG performance within large corpora.

Document RAG Benchmarks The increasing attention on Document RAG (Ma et al., 2024) and VQA (Mathew et al., 2021) necessitates comprehensive multimodal retrieval benchmarks. Current practices often use VQA dataset queries (Friel et al., 2024; Faysse et al., 2024; Cho et al., 2024), but these are document-specific and lack information for global retrieval. Other benchmarks (Wang et al., 2025; Dong et al., 2025) craft informative queries from single pages, yet often only mark that single page as relevant, ignoring other potential matches and risking evaluation inaccuracies. We introduce a novel benchmark to overcome these identified limitations.

6 Conclusion

We introduce CDOCRAG-BENCH, a large-scale, multimodal, multilingual benchmark designed to reflect realistic retrieval-augmented generation scenarios, overcoming limitations of prior work such as ambiguous evidence and trivial reasoning. Our three-stage process—corpus assembly, automated query synthesis with exhaustive evidence labels, and human refinement—yields 4,000 high-quality single- and multi-hop queries over diverse documents. Evaluations of leading embedding models and RAG frameworks reveal vision-language embeddings' clear benefits, a strong link between retrieval accuracy and answer quality, and persistent challenges in multi-page reasoning. By providing rigorous baselines and chain-of-evidence annotations, CDOCRAG-BENCH lays a foundation for future advances in retrieval architectures, knowledge-graph integration, and dynamic query decomposition.

Limitation

We leverage LLMs to revise our paper and serving as metrics in our evaluation. We include human-annotation in Appendix B to validate the LLM-as-a-Judge process.

References

Omar Adjali, Olivier Ferret, Sahar Ghannay, and Hervé Le Borgne. 2024. Multi-level information retrieval augmented generation for knowledge-based visual question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16499–16513, Miami, Florida, USA. Association for Computational Linguistics.

Eleonora Breci, Luca Guarnera, and Sebastiano Battiato. 2024. A novel dataset for non-destructive inspection of handwritten documents. *arXiv preprint arXiv:2401.04448*.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024a. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*.

Haonan Chen, Liang Wang, Nan Yang, Yutao Zhu, Ziliang Zhao, Furu Wei, and Zhicheng Dou. 2025. mme5: Improving multimodal multilingual embeddings via high-quality synthetic data. *arXiv preprint arXiv:2502.08468*.

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.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, and 1 others. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. 2024. M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding. *arXiv preprint arXiv:2411.04952*.

Giorgia Crosilla, Lukas Klic, and Giovanni Colavizza. 2025. Benchmarking large language models for handwritten text recognition. *arXiv preprint arXiv:2503.15195*.

Kuicai Dong, Yujing Chang, Xin Deik Goh, Dexun Li, Ruiming Tang, and Yong Liu. 2025. Mmdocir: Benchmarking multi-modal retrieval for long documents. *arXiv preprint arXiv:2501.08828*.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations*.

Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2024. Ragbench: Explainable benchmark for retrieval-augmented generation systems. *arXiv preprint arXiv:2407.11005*.

Philippe Gervais, Anastasiia Fadeeva, and Andrii Maksai. 2024. Mathwriting: A dataset for handwritten mathematical expression recognition. *arXiv preprint arXiv:2404.10690*.

Lukas Gienapp, Harrisen Scells, Niklas Deckers, Janek Bevendorff, Shuai Wang, Johannes Kiesel, Shahbaz Syed, Maik Fröbe, Guido Zuccon, Benno Stein, and 1 others. 2024. Evaluating generative ad hoc information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1916–1929.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2024. Lightrag: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*.

Kailash A Hambarde and Hugo Proenca. 2023. Information retrieval: recent advances and beyond. *IEEE Access*, 11:76581–76604.

Siwei Han, Peng Xia, Ruiyi Zhang, Tong Sun, Yun Li, Hongtu Zhu, and Huaxiu Yao. 2025. Mdocagent: A multi-modal multi-agent framework for document understanding. *arXiv preprint arXiv:2503.13964*.

Yulong Hui, Yao Lu, and Huanchen Zhang. 2024. Uda: A benchmark suite for retrieval augmented generation in real-world document analysis. *arXiv preprint arXiv:2406.15187*.

Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. 2025. Llave: Large language and vision embedding models with hardness-weighted contrastive learning. *arXiv preprint arXiv:2503.04812*.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Yangning Li, Yinghui Li, Xinyu Wang, Yong Jiang, Zhen Zhang, Xinran Zheng, Hui Wang, Hai-Tao Zheng, Fei Huang, Jingren Zhou, and 1 others. 2024.

Benchmarking multimodal retrieval augmented generation with dynamic vqa dataset and self-adaptive planning agent. *arXiv preprint arXiv:2411.02937*.

Nikos Livathinos, Christoph Auer, and 1 others. 2025. Docling: An efficient open-source toolkit for AI-driven document conversion. *arXiv preprint arXiv:2501.17887*.

LlamaIndex. 2025. vdr-2b-multi-v1. Hugging Face Model Card.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, and 1 others. 2024. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *arXiv preprint arXiv:2407.01523*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Matin Mortaheb, Mohammad A Amir Khojastepour, Srimat T Chakradhar, and Sennur Ulukus. 2025a. Rag-check: Evaluating multimodal retrieval augmented generation performance. *arXiv preprint arXiv:2501.03995*.

Matin Mortaheb, Mohammad A Amir Khojastepour, Srimat T Chakradhar, and Sennur Ulukus. 2025b. Re-ranking the context for multimodal retrieval augmented generation. *arXiv preprint arXiv:2501.04695*.

Helen Ngo, Cooper Raterink, João GM Araújo, Ivan Zhang, Carol Chen, Adrien Morisot, and Nicholas Frosst. 2021. Mitigating harm in language models with conditional-likelihood filtration. *arXiv preprint arXiv:2108.07790*.

OpenAI. 2024. Gpt-4o. Accessed: 2025-05-01.

Shu Pu, Yaochen Wang, Dongping Chen, Yuhang Chen, Guohao Wang, Qi Qin, Zhongyi Zhang, Zhiyuan Zhang, Zetong Zhou, Shuang Gong, and 1 others. 2025. Judge anything: Mllm as a judge across any modality. *arXiv preprint arXiv:2503.17489*.

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*.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13636–13645.

Yixuan Tang and Yi Yang. 2024. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. *arXiv preprint arXiv:2401.15391*.

Bin Wang, Chao Xu, and Xiaomeng Zhao. 2024. MinerU: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*.

Qiuchen Wang, Ruixue Ding, Zehui Chen, Weiqi Wu, Shihang Wang, Pengjun Xie, and Feng Zhao. 2025. Vidorag: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. *arXiv preprint arXiv:2502.18017*.

Navve Wasserman, Roi Pony, Oshri Naparstek, Adi Raz Goldfarb, Eli Schwartz, Udi Barzelay, and Leonid Karlinsky. 2025. Real-mm-rag: A real-world multi-modal retrieval benchmark. *arXiv preprint arXiv:2502.12342*.

Yin Wu, Quanyu Long, Jing Li, Jianfei Yu, and Wenya Wang. 2025. Visual-rag: Benchmarking text-to-image retrieval augmented generation for visual knowledge intensive queries. *arXiv preprint arXiv:2502.16636*.

Yuming Yang, Jiang Zhong, Li Jin, Jingwang Huang, Jingpeng Gao, Qing Liu, Yang Bai, Jingyuan Zhang, Rui Jiang, and Kaiwen Wei. 2025. Benchmarking multimodal rag through a chart-based document question-answering generation framework. *arXiv preprint arXiv:2502.14864*.

Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2024. Evaluation of retrieval-augmented generation: A survey. In *CCF Conference on Big Data*, pages 102–120. Springer.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. Gme: Improving universal multimodal retrieval by multimodal llms. *arXiv preprint arXiv:2412.16855*.

Zhuocheng Zhang, Yang Feng, and Min Zhang. 2025. Levelrag: Enhancing retrieval-augmented generation with multi-hop logic planning over rewriting augmented searchers. *arXiv preprint arXiv:2502.18139*.

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2024. Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems*, 42(4):1–60.

Junjie Zhou, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, Defu Lian, and Yongping Xiong. 2024. Megapairs: Massive data synthesis for universal multimodal retrieval. *arXiv preprint arXiv:2412.14475*.

A Detailed Benchmark Construction

A.1 Step 1: Data Collection and Preprocessing

We collected rich documents from various sources, including printed PDFs, handwritten documents, slides, and HTML pages in multiple languages. We then selected documents with a length between 10-50 pages and eliminated irrelevant content. Using tools such as Docling and MinerU, we split each document into text, diagrams, and images, ensuring that each modality was processed separately. We further used GPT-4o to filter out low-quality or irrelevant data chunks and retain only important content for query generation. Finally, the document data of our benchmark is shown in the Table 5 and Figure 3

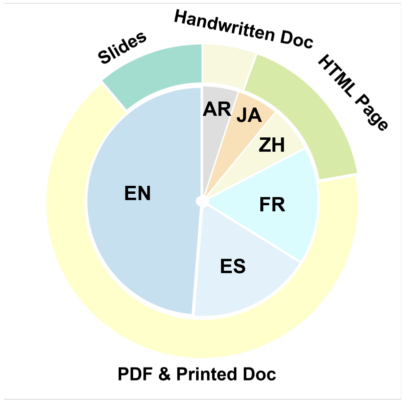

Figure 3: Statistics of the CDocRAG-Bench Dataset. The dataset contains documents across six languages with diverse page structures and visual content. The data is distributed as follows: English (1,950, 51.6%), Spanish (695, 18.4%), French (661, 17.5%), Chinese (258, 6.8%), Japanese (235, 6.2%), and Arabic (201, 5.3%). Document types are categorized as: PDF & Printed Documents (1,250, 66.0%), HTML Pages (350, 18.5%), Slides (250, 13.2%), and Handwritten Documents (150, 7.9%).

A.2 Step 2: One-Hop Query Synthesis

To generate single-hop queries, we utilized the parsed document chunks. For textual content, we concentrated on key entities and concepts, crafting questions that were self-contained and independent. For visual elements, we prioritized understanding critical visual components and patterns. Each query was meticulously reviewed to ensure clarity, relevance, and direct answerability from the document.

A.3 Step 3: Multi-Hop Query Synthesis

To address the challenge of multi-hop questions that require reasoning across multiple steps, we employed a knowledge graph-based approach. Using LightRAG, we constructed knowledge graphs for each document, extracting entities and relations from both textual and visual content. These graphs leveraged the connectivity of nodes and edges to represent the relationships between different pieces of information. We selected query paths through these graphs, starting with high-dimensional entities and expanding paths based on content depth and connectivity. Subqueries were synthesized for each hop and iteratively combined into a coherent multi-hop query. Each query was rigorously checked to ensure its logical coherence, unique answers, and overall relevance to the document.

A.4 Step 4: Post-Processing

After query synthesis, the generated queries underwent model filtering using advanced models to ensure they met quality standards, removing poorly structured or irrelevant queries. Evidence pages for both single-hop and multi-hop queries were annotated by VLMs like Qwen2.5-32B-VL, ensuring that supporting evidence was accurately identified. Finally, human annotators reviewed the filtered queries and evidence pages for accuracy and consistency. Human refinement addressed potential discrepancies, thereby enhancing the overall reliability of the benchmark.. This comprehensive post-processing approach ensured the robustness of the benchmark and its suitability for evaluating multimodal document retrieval systems.

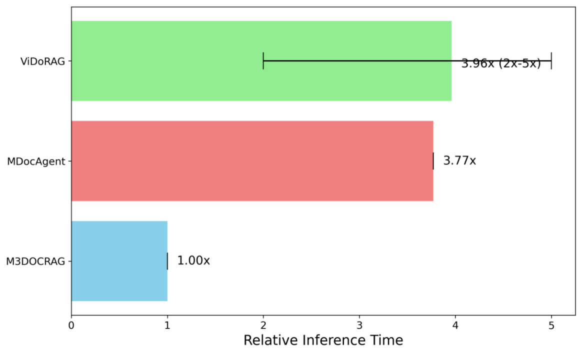

Figure 4: Time efficiency of Frameworks

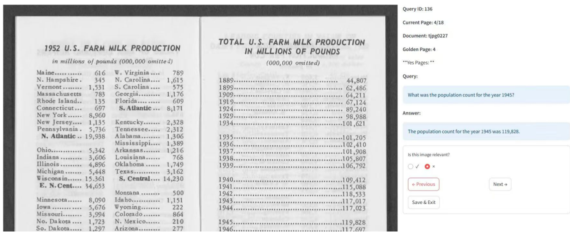

Figure 5: Human Annotation UI

Table 5: CDocRAG-Bench Dataset Statistics

Statistic	Number
Documents	2776
Languages	6
Avg. pages per doc	22.5
Avg. words per page	289.9
Avg. tables per page	0.397
Avg. figures per page	1.078
Total Questions	4000
Single-hop questions	2000
Avg. evidence pages	2.91
Multi-hop questions	2000
Avg. hops	2.82

B Human Annotation Details

The annotation is conducted by 5 authors of this paper and 1 volunteers independently. As acknowledged, the diversity of annotators plays a crucial role in reducing bias and enhancing the reliability of the benchmark. These annotators have knowledge in this domain, with different genders, ages, and educational backgrounds. To ensure the annotators can proficiently mark the data, we provide them with detailed tutorials, teaching them how to evaluate model responses more objectively. The annotation UI is showed in Figure 5

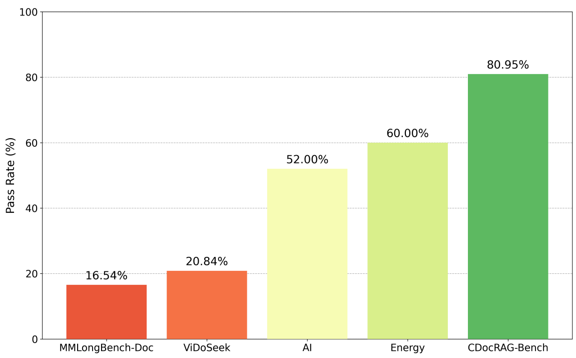

Figure 6: Ratio of queries that meet evaluation standards.

C Detailed Experiment Settings

C.1 Experimental Methodology

C.1.1 Experiment 1: Evaluation of Embedding Models

To assess the recall capabilities of state-of-the-art embedding models, we conducted experiments on both multimodal and text embedding models. For multimodal embedding models, we directly embedded the visual and textual content of the documents. In contrast, for text embedding models, we first applied OCR to extract text from images and then embedded the extracted text. Additionally, for images and tables, we used a Vision-Language Model (VLM) to generate descriptive captions, which were subsequently embedded. This approach allowed us to evaluate how effectively each type of model could capture and recall relevant information from multimodal documents.

C.1.2 Experiment 2: Performance of LLMs with and without Golden Pages

We tested the performance of advanced Large Language Models (LLMs), specifically GPT-4o and Gemini, in two different contexts. First, we evaluated their performance when provided with the exact golden pages as context, which allowed us to assess their ability to generate accurate answers with direct access to relevant information. Second, we tested their performance without the golden pages, relying solely on their inherent knowledge and reasoning capabilities. This dual evaluation provided insights into how these models perform in both ideal and more challenging conditions.

C.1.3 Experiment 3: RAG Frameworks Evaluation

To further evaluate the effectiveness of Retrieval-Augmented Generation (RAG) frameworks, we conducted experiments using MDocAgent, ViDoRAG, and M3DOCRAG. We tested both the recall capabilities of these frameworks in retrieving relevant documents and their ability to generate accurate answers based on the retrieved information. In addition, we also analyzed the Time efficiency of Frameworks,as shown in Figure 4.This comprehensive evaluation helped us understand how well these frameworks integrate retrieval and generation to address complex multimodal document retrieval tasks.

C.2 Human Refinement Detail

To ensure the accuracy and reliability of our benchmark, we incorporated a comprehensive human refinement process. This process involved meticulous human review and annotation of both the generated queries and the evidence page labels. By comparing model-generated outputs with human annotations, we aimed to eliminate biases, clarify ambiguous boundaries, and enhance the overall quality of the benchmark. This dual-layer approach was robust and suitable for rigorous evaluation of multimodal document retrieval systems.

C.2.1 Query Flitter Refinement

For the generated queries, we conducted a detailed human review on a subsample to assess the reasonableness of the model's filtering process. Human annotators carefully analyzed the queries to ensure they were clear, relevant, and answerable based on the document content. Through prompt engineering, we refined the model's prompts to eliminate potential biases and improve the quality of the generated queries. This iterative process of human review and prompt refinement helped us achieve a higher standard of query relevance and clarity.

C.2.2 Evidence Page Labeling

In addition to query refinement, we also performed a thorough human review of the evidence page labels. Human annotators individually verified each evidence page, comparing the model-generated labels with their own annotations. This process helped to resolve any ambiguities and ensured that the evidence pages were accurately labeled by Vision-Language Models, which are crucial for evaluating the performance of retrieval systems.

D Prompt Templates

Full prompt templates and examples are provided in Figure 7 to Figure 12.

E Case Study

A detailed case study is shown in Figure 13 .

Query Reviewer Prompt

System Prompt: Task

I have some QA data here, and you can observe that the questions can be divided into two categories:
The category #A: When you see this question alone without a given document, you are sure to find a unique document in a corpus to provide a unique answer. The question having some key words to help you locate the document from corpus.
The category #B: When you see this question alone without a given document, you will find hard to locate a document to give a deterministic answer for this question, because you will find multiple candidate documents in a corpus, which may lead to different answers for this question. The question do not have any special key words to help you locate the document from corpus.

Examples

The number mentioned on the right of the leftside margin? #B
What is the date mentioned in the second table? #B
What is the full form of PUF? #A
What is the number at the bottom of the page, in bold? #B
Who presented the results on cabin air quality study in commercial aircraft? #A
What is the name of the corporation? #B
Which part of Virginia is this letter sent from? #B
who were bothered by cigarette odors? #A
which cigarette would be better if offered on a thicker cigarette? #A
Cigarettes will be produced and submitted to O/C Panel for what purpose? #A
What is the heading of first table? #B
What is RIP-6 value for KOOL KS? #A
Which test is used to evaluate ART menthol levels that has been shipped? #A
How much percent had not noticed any difference in the odor of VSSS? #A
What is the cigarette code of RIP-6(W/O Filter) 21/4SE? #A
What mm Marlboro Menthol were subjectively smoked by the Richmond Panel? #A
What are the steps of Weft Preparation between Spinning bobbin and Weaving? #A
What level comes between Middle Managers and Non-managerial Employees? #A
What are the six parts of COLLABORATION MODEL of the organization where James has a role of leading the UK digital strategy? #A

User Prompt:

Query: {Query Description}

Figure 7: Prompt of Query Reviewer.

Single hop QA Generator Prompt

System Prompt You are a professional cross-document retrieval dataset assistant. Read the information on the given page and generate a high-quality QA pair"

Text-based QA Generator Prompt Generate one QA pair based on the following guideline:
Question Requirements: - Create self-contained questions requiring no contextual knowledge from other pages - Focus on explicitly mentioned key entities, concepts, or processes - Avoid page-specific references (e.g., "in this section" or "as mentioned above") - Include both factual questions (who/what/when) and explanatory questions (how/why)
Answer Specifications: - Answers may be moderately summarized but must strictly adhere to source content - Prohibit any speculation or supplementation beyond original text
Format Rules: - Response must be in JSON format containing "question" and "answer" fields
example response: { "question": "What are the clinical diagnostic criteria for Parkinson's disease?", "answer": "Diagnosis requires bradykinesia combined with either resting tremor or rigidity, often presenting with asymmetric onset." }

User Prompt:

Query: {Query Description}

Multi hop QA Generator Prompt

Candidate Nodes Selection Prompt:
You are an expert in knowledge graph reasoning. Your task is to evaluate relationship candidates and select the best one for constructing an unambiguous reasoning question.
The ideal relationship should uniquely identify the target entity. When forming a question like "What entity [relation] with [current entity]?", the answer should be specific enough that only one reasonable entity fits. It is strictly forbidden to select vague relationships such as "is related to"
For example, "is the capital of" is a strong relation for a country-city pair, while "is located in" is weaker as many cities could be located in the same country
Given the current entity '${current_node}', I have the following candidate entities and their relations to the current entity:
${candidates_json}
Please evaluate each relationship and select the ONE that would create the most unambiguous and specific reasoning question. The chosen relationship should make it possible to uniquely identify the target entity when given the current entity and the relationship.
Return your response as a JSON object with:

1. "reasoning": brief explanation of why this relationship is the most specific/unique
2. "selected_index": the index (0-based) of the chosen candidate

Example response format:
{
"reasoning": "This relationship 'is the inventor of' creates the most unique connection...", "selected_index": 2
}

Figure 8: Prompt template used for QA generating behavior.

Answer Evaluation Prompt

You are a comprehensive judge evaluating the LLM generated answer based on the reference answer. You should first collect similarities and differences between the reference answer and the generated answer. Then, give a score from 0 to 10 based on the correctness of the generated answer. Do allow the generated answer to include additional information if the correct information in the reference answer is already provided. Return your answer in strict JSON format. Your output will be directly parsed, so do not add any other text that hinders the parsing process.

Example 1:
Query: "How do educational backgrounds influence congressional vote preferences among registered voters?"
Reference Answer: "Registered voters with postgraduate degrees favor Democrats by 62
Generated Answer: "Registered voters' preference for Democrats increases with higher levels of education: 53% among those with a four-year degree and 62% among those with a postgraduate degree, contrasting with the more divided opinions of those without a college degree."
Output:
{{
"evaluation": "The generated answer accurately contains the ratio of registered voters with postgraduate degrees favoring Democrats, and the ratio of registered voters with four-year degrees supporting Democrats. The answer also mentions divided preferences among those without a college degree, which is consistent with the reference answer.",
"score": 10
}}

Example 2:
Query: "What specifications differentiate the video output capabilities across different models of Roku devices?"
Reference Answer: "Roku devices vary in their video output capabilities, with Roku 1 and Roku 2 supporting up to 1080p, Roku LT up to 720p, and Roku 3 and Roku 4 supporting up to 4K Ultra HD."
Generated Answer: "The video output capabilities differentiate across Roku devices as follows: Roku 1, Roku 2, and Roku LT support video output up to 720p, while Roku 3 and Roku 4 support video output from 1080p to 4K Ultra HD."
Output:
{{
"evaluation": "The generated answer correctly states the video output capabilities of Roku LT, Roku 3, and Roku 4. However, it incorrectly states that Roku 1 and Roku 2 support video output up to 720p instead of 1080p. The answer has 3 correct points and 2 incorrect points.",
"score": 6
}}

Input:
Query: ${query}
Reference answer: ${reference}
Generated answer: ${generated}

Figure 9: Prompt template used for answer evaluation.

Multi hop QA Generator Prompt

Candidate Nodes Selection Prompt:
You are an expert in knowledge graph reasoning. Your task is to evaluate relationship candidates and select the best one for constructing an unambiguous reasoning question.
The ideal relationship should uniquely identify the target entity. When forming a question like "What entity [relation] with [current entity]?", the answer should be specific enough that only one reasonable entity fits. It is strictly forbidden to select vague relationships such as "is related to"
For example, "is the capital of" is a strong relation for a country-city pair, while "is located in" is weaker as many cities could be located in the same country
Given the current entity '${current_node}', I have the following candidate entities and their relations to the current entity:
${candidates_json}
Please evaluate each relationship and select the ONE that would create the most unambiguous and specific reasoning question. The chosen relationship should make it possible to uniquely identify the target entity when given the current entity and the relationship.
Return your response as a JSON object with:

1. "reasoning": brief explanation of why this relationship is the most specific/unique
2. "selected_index": the index (0-based) of the chosen candidate

Example response format:
{
"reasoning": "This relationship 'is the inventor of' creates the most unique connection...", "selected_index": 2
}

Figure 10: Prompt template used for Multi hop QA Generation.

Single hop Prompt

You are an expert document-based question answering system. Your task is to provide accurate, well-structured, and comprehensive answers based solely on the provided document content.
Key Requirements:

1. **Content-Based Response**
    - Base your answer EXCLUSIVELY on the provided document content
    - Do not incorporate external knowledge or assumptions
    - If you are not exactly sure about the answer, use the answer you have most confidence on, but the answer must be based on the content
    - If the content does not contain any likely answer, explicitly state: "The provided content does not contain enough information to answer this question"
2. **Answer Quality**
    - Be precise, concise, and directly address the question
    - Structure your response in a clear, logical manner
    - Use bullet points or numbered lists when appropriate
    - Maintain academic rigor and professional tone
3. **Content Handling**
    - Consider all provided content sections equally
    - If content appears contradictory, try to identify the most reliable source as the basis for your answer
4. **Response Format**
    - Return with strict JSON format. Your output will be directly parsed, so do not add any other text that hinders the parsing process.
    - Begin with a "thought" key. Collect relevant information here, together with your reasoning process if necessary.
    - After completing the thought, add a "final_answer" key. This is the final answer to the question.
    - Do not include any other keys or information in the response.
5. **Example** Assume the given image is a document with a graph about the sales of a product over time. The question is "What was the sales trend in Q1 2023?" Your answer should be: { "thought": "The document contains a graph showing the sales trend over time. In Q1 2023, the sales were steadily increasing.", "final_answer": "The sales trend in Q1 2023 was steadily increasing." }

Figure 11: Prompt template used for instructing single-hop reasoning behavior.

Multi hop Prompt

You are an expert document-based multihop question answering system. Your task is to provide accurate, well-structured, and comprehensive answers based on a series of reasoning steps that build upon each other.
Key Requirements:

1. **Language Consistency**
   - Always respond in the same language as the input question
   - For English questions, use English in all steps and final answer
   - For Chinese questions, use Chinese in all steps and final answer
   - For other languages, maintain consistent language usage throughout
   - Never mix languages in your response
2. **Reasoning Chain Construction**
   - Break down complex questions into logical reasoning steps
   - Each step should build upon previous steps
   - Use evidence from the provided content to support each step
   - Maintain clear logical connections between steps
3. **Content-Based Response**
   - Base your reasoning EXCLUSIVELY on the provided document content
   - Do not incorporate external knowledge or assumptions
   - If you are not exactly sure about a step, use the information you have most confidence in
   - If the content does not contain enough information for a step, explicitly state this
4. **Answer Quality**
   - Each reasoning step should be precise and directly relevant
   - The final answer should integrate all reasoning steps
   - Maintain academic rigor and professional tone
   - Acknowledge any uncertainties in the reasoning chain
5. **Response Format**
   - Return with strict JSON format. Your output will be directly parsed, so do not add any other text.
   - Include a "relevant_question_list" key containing the reasoning steps
   - Include a "final_answer_prediction" key for the final answer
   - Each step in relevant_question_list should be clearly labeled (e.g., "Step 0: " for English)
6. **Examples English Example (Multihop Reasoning):** Question: "In which series does a member of the team that Freefall is part of and serves as the Geraldine, Montana appear as a supporting character?" Content: [Document sections about G.I. Joe team members and their appearances] Your answer should be: { "relevant_question_list": ["Step 0: Freefall is a member of the G.I. Joe Team.", "Step 1: Sci-Fi is a member of the G.I. Joe Team who serves as the Geraldine, Montana.", "Step 2: Sci-Fi appears as a supporting character in the Marvel Comics G.I. Joe series."], "final_answer_prediction": "Marvel Comics G.I. Joe series" }

Figure 12: Prompt template used for instructing multi-hop reasoning behavior.

Case Study:

Multi hop Question:
What disease does the entity that launched its first nationwide public service campaign in 1980 implement programs to prevent, which causes blindness in children due to vitamin A deficiency?

VLM response:
The disease that the entity likely implemented programs to prevent, which causes blindness in children due to vitamin A deficiency, is Xerophthalmia. **Note:** This discussion is missing the step of identifying the first nationwide public service campaign in 1980 that implemented such prevention programs.
While pinpointing a single entity that launched its first nationwide public service campaign specifically in 1980 in the United States focused solely on preventing xerophthalmia through vitamin A sufficiency proves challenging based on available search information, the 1980s were a crucial period for understanding and combating vitamin A deficiency and its impact on childhood blindness and mortality.
Research in the 1980s, significantly contributed to by organizations like Helen Keller Intl in collaboration with researchers such as Dr. Alfred Sommer, highlighted the profound impact of vitamin A supplementation in preventing not only blindness but also reducing childhood deaths from common illnesses. This groundbreaking evidence galvanized international health organizations and NGOs to expand their efforts in vitamin A distribution and public health initiatives in areas where deficiency was prevalent.
Therefore, while a specific US-based "first nationwide public service campaign in 1980" by a named entity for this exact issue wasn't definitively identified, organizations like Helen Keller Intl were actively working on implementing programs to prevent xerophthalmia due to vitamin A deficiency during this period, likely as part of broader public health or nutritional initiatives that may have included public awareness components. The impact of their work and the research from that era significantly contributed to the global understanding and efforts to combat this preventable cause of childhood blindness.

Figure 13: Case study showing VLM's shortcut reasoning behavior on a multi-hop query.