

ComDoc: A Comprehensive Benchmark for Document Retrieval-Augmented Generation

Jun Wang^{♥*} Tiancheng Gu^{♣*} Yifan Zhang^{♣*} Hengyu Shen[♣] Xiaoxing Hu[♣]
Weidong Cai[♣] Ziyong Feng[♥] Xiang An[♣] Kaicheng Yang^{♥†}

[♥]Glint Lab [♣]The University of Sydney ^{♣*}South China University of Technology

[♣]Beijing Institute of Technology [♣]Shanghai Jiao Tong University [♣]Independent Researcher

Abstract

Document Retrieval-Augmented Generation (DocRAG) is an emerging paradigm crucial for complex tasks such as financial report analysis and multi-source knowledge retrieval. However, existing benchmarks typically target isolated capabilities, such as multi-modal or multi-hop reasoning, and offer only single-perspective evaluations, revealing a need for more comprehensive assessment. In this paper, we present *ComDoc*, a large-scale multi-modal DocRAG benchmark spanning single-hop, multi-hop, and multi-page tasks. *ComDoc* includes 1,755 real-world documents from four domains, comprising 103,877 pages and 5,561 QA instances. To construct *ComDoc*, we design a three-stage pipeline: data collection, hierarchical QA pairs generation, and quality control. Using *ComDoc*, we systematically evaluate state-of-the-art encoder-based, LLM-based, and MLLM-based retrieval models, as well as the generative capabilities of leading open-source LLMs and MLLMs. Our results demonstrate substantial performance gaps across different models and task types in both retrieval and generation.

1 Introduction

Retrieval-Augmented Generation (RAG) effectively answers complex queries by integrating external information. In document-level, multimodal settings with text, images, and tables, DocRAG advances beyond traditional retrieval and generation. Consequently, DocRAG emerges as a vital paradigm for knowledge-intensive applications, such as academic document analysis (Barron et al., 2025) and financial reasoning (Wang et al., 2025c).

The increasing adoption of DocRAG drives growing interest in systematic performance evaluation. Recently, several benchmarks have emerged to assess DocRAG from diverse perspectives. REAL-MM-RAG (Wasserman et al., 2025) evaluates four key dimensions relevant to real-world retrieval: multimodal documents, task complexity, realistic RAG queries, and precise annotation. ViDoRAG (Wang et al., 2025d) introduces ViDoSeek to measure RAG performance on visually rich documents that require complex reasoning. Similarly, ViDoRe (Faysse et al., 2025) and ViDoRe V2 (Macé et al., 2025) focus on evaluating the ability of retrieval systems to align queries with relevant documents at the page level. LongDocURL (Deng et al., 2025) emphasizes the deep understanding of lengthy documents, numerical reasoning, and the synthesis of information across document elements, spanning 20 sub-tasks categorized by primary task types and evidence sources. Despite these advancements, existing benchmarks either focus primarily on a specific dimension or are limited in scale, making it difficult to adequately assess the complexity of different tasks.

In this paper, we focus on four core challenges in DocRAG: 1) *Single-Hop*: Requires extraction or reasoning based on a single retrieved element or source to generate accurate answers; 2) *Multi-Hop*: Requires reasoning across multiple retrieved elements or sources to generate accurate answers; 3) *Multi-Modal*: Involves interpreting and integrating diverse content types, such as text, image, and structured table; and 4) *Multi-Page*: Addresses scenarios

*Equal Contribution † Corresponding Author

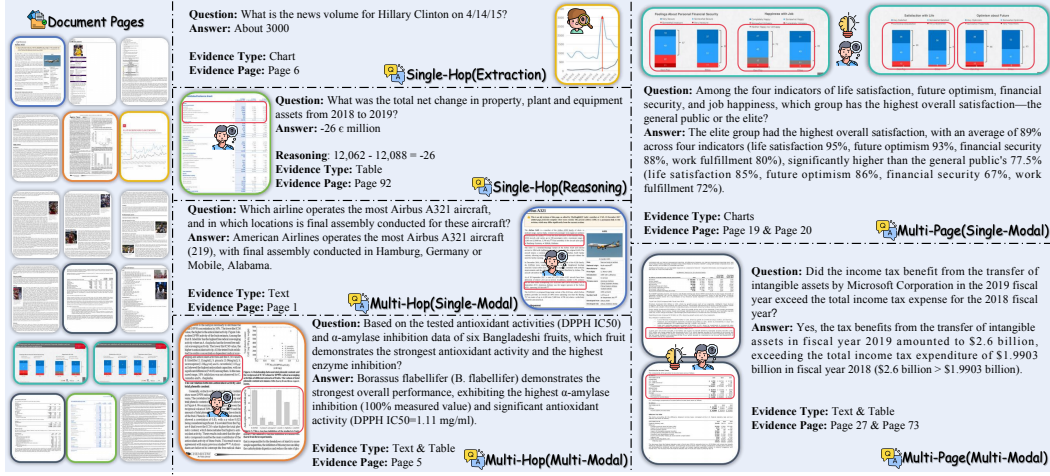


Figure 1: Overview of our proposed *ComDoc* benchmark. *ComDoc* comprises six types of QA samples: Single-Hop Extraction, Single-Hop Reasoning, Multi-Hop Single-Modal, Multi-Hop Multi-Modal, Multi-Page Single-Modal, and Multi-Page Multi-Modal.

where relevant information is distributed across multiple pages, necessitating retrieval and aggregation from non-contiguous sections.

Guided by these four dimensions, we devise a three-stage pipeline comprising data collection, hierarchical QA pair generation, and quality control. In the initial phase, we employ a domain-balanced sampling strategy to curate documents tailored for generative QA from diverse datasets across four domains. For hierarchical QA generation, we parse documents and apply task-specific selection criteria and QA generation rules for single-hop, multi-hop, and multi-page tasks. During quality control, we employ LLM-as-a-judge to filter low-quality samples and engage five human experts to verify question-answer correctness. Ultimately, we propose a large-scale multimodal DocRAG benchmark *ComDoc*. As shown in Figure 1, *ComDoc* encompasses six distinct task categories: Single-Hop (Extraction), Single-Hop (Reasoning), Multi-Hop (Single-Modal), Multi-Hop (Multi-Modal), Multi-Page (Single-Modal), and Multi-Page (Multi-Modal). *ComDoc* curates 1,755 real-world documents across four domains, totaling 103,877 pages and 5,561 QA instances. We systematically evaluate state-of-the-art encoder-based, LLM-based, and MLLM-based embedding models for document retrieval, as well as the generative capabilities of leading open-source LLMs and MLLMs. The results demonstrate substantial performance variations among different retrieval and generation paradigms across diverse tasks and domains. Our main contributions are summarized as follows:

- We **design a three-stage pipeline** consisting of data collection, hierarchical QA pair generation, and quality control for constructing a comprehensive DocRAG benchmark.
- We **propose a large-scale multi-modal document retrieval augmented generation benchmark *ComDoc*** that covers single-hop, multi-hop and multi-page tasks.
- We **perform extensive evaluations of state-of-the-art retrieval and generation models on *ComDoc***, providing actionable insights to drive advancements in DocRAG research.

2 Related Work

2.1 Benchmarks for DocRAG

While numerous benchmarks in the DocRAG domain evaluate both document retrieval and answer generation, the majority suffer from critical limitations. For retrieval, PDF-MVQA (Ding et al., 2023) offers comprehensive QA pairs but is limited to academic documents and lacks domain diversity. ViDoRe (Faysse et al., 2025) and ViDoRe-V2 (Macé et al., 2025) provide multi-domain coverage but only support single-hop reasoning. MMDocIR (Dong et al., 2025a) enables hierarchical references and domain diversity, yet its limited queries reduce evaluation robustness. All retrieval benchmarks focus solely on retrieval per-

Benchmarks	Document			Question-Answer			Evaluation		
	Domain	#Pages	#Docs	Evidence	#Queries	Human Verify	RAG Suitable	Granular Analysis	Hierarchical Structure
PDF-MVQA	Academia	5.9k	623		54.5k	✗	R	✗	SH/MH/MP
ViDoRe	Multiple	8.3k	-	✗	3.8k	✓	R	✗	SH
ViDoRe-v2	Multiple	3.3k	66	✗	0.4k	✓	R	✗	SH
MMDocIR	Multiple	65.1k	313		1.7k	✓	R	✓	SH/MH/MP
REALMM	Finance	8.6k	163		4.6k	✗	R	✗	SH
DocBench	Multiple	15.1k	229		1.1k	✓	G	✗	SH
MMLongBenchDoc	Multiple	6.4k	135		1.1k	✓	G	✗	SH/MH/MP
M-LongDoc	Multiple	37.9k	180		0.9k	✓	G	✓	SH
ViDoSeek	Multiple	-	300		1.1k	✗	R/G	✗	SH/MH
MMDocRAG	Multiple	14.9k	222		4.1k	✓	R/G	✓	SH/MH/MP
<i>ComDoc(ours)</i>	Multiple	103.9k	1,755		5.6k	✓	R/G	✓	SH/MH/MP

Table 1: Comparison of existing multimodal DocRAG benchmarks with *ComDoc*. RAG Suitable: Assesses both retrieval (R) and generation (G) capabilities. Granular Analysis: Evaluates performance across diverse sample types. Hierarchical Structure: Supports Single-Hop (SH), Multi-Hop (MH), and Multi-Page (MP) references.

formance, omitting generation assessment. For answer generation, REALMM (Wasserman et al., 2025), DocBench (Zou et al., 2025), and M-LongDoc (Ma et al., 2024) evaluate multi-evidence understanding but lack cross-modality and multi-page support. VideoSeek (Macé et al., 2025) and MMLongBenchDoc (Ma et al., 2024) include hierarchical references but offer limited scale, reducing the evaluation challenge. MMDocRAG (Dong et al., 2025b) provides extensive QA pairs and diverse capability assessment, but its small document set constrains the difficulty. Detailed information is provided in Table 1.

2.2 Retrieval and Generation Models

RAG systems consist of retrieval and generation models. In DocRAG, document retrieval is crucial as it directly impacts generation quality. Early approaches use OCR tools such as Nougat (Blecher et al., 2023) for text extraction, followed by embedding models like ColBERT (Khattab & Zaharia, 2020), BGE (Chen et al., 2024), and E5 (Wang et al., 2024b) for encoding and retrieval. Recent retrievers employ LLM-based and MLLM-based embedding models, including GTE-Qwen2-7B (Li et al., 2023), NV-Embed V2 (Lee et al., 2025), VLM2Vec (Jiang et al., 2025), UniME and UniME-V2 (Gu et al., 2025; 2026). However, these models are designed to learn unified representations and are not specifically built for document retrieval. In contrast, models such as ColQwen (Faysse et al., 2025) and ViDoRAG (Wang et al., 2025d), which are tailored for visual document retrieval, demonstrate significantly superior performance. For answer generation, RAG systems increasingly rely on pre-trained MLLMs capable of reasoning over complex, multi-page contexts (Wang et al., 2025a,b; An et al., 2025). The ecosystem features both proprietary and competitive open-source models, including DeepSeek (Liu et al., 2024), QwenVL (Qwen et al., 2025; Bai et al., 2025a), InternVL (Zhu et al., 2025; Wang et al., 2025e), and Llama (Grattafiori et al., 2024). As the cognitive core of RAG systems, these models integrate visual and textual evidence to generate accurate and coherent responses.

3 ComDoc Benchmark

To construct *ComDoc*, we design a three-stage pipeline that includes data collection (Section 3.1) to construct document datasets, hierarchical QA pairs generation (Section 3.2) to generate question-answer pairs for different tasks, and quality control (Section 3.3). We also show the statistical information of our benchmark (Section 3.4).

3.1 Data Collection

Document Source. Our benchmark integrates materials from four domains: slide decks, finance, academia, and world knowledge, comprises six document types: slide, report,

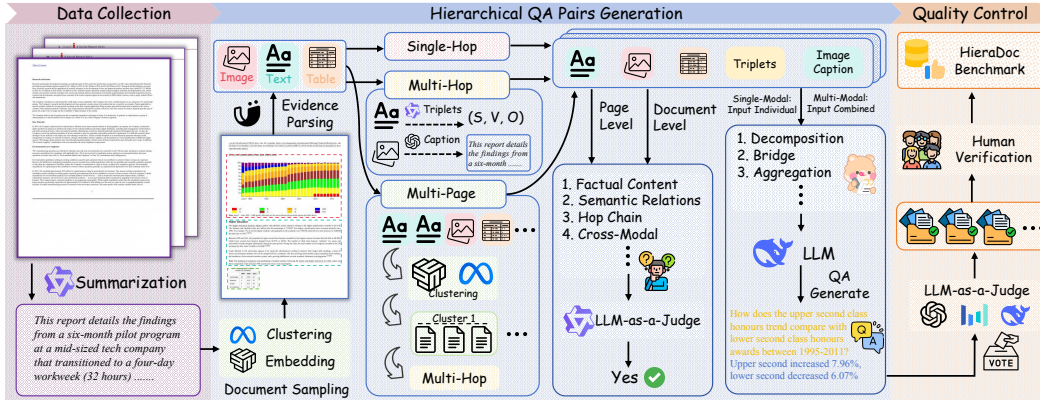


Figure 2: Overview of the three-stage pipeline for constructing *ComDoc*: data collection, hierarchical QA pair generation, and quality control.

academic paper, review, wiki article, and website. We source documents from multiple public datasets: SlideVQA (Tanaka et al., 2023), FinanceBench (Wang et al., 2025c), ArXivCap (Li et al., 2024), M2QA (Engländer et al., 2024), and FetaTab (Nan et al., 2021), and further incorporate selected segments from UDA-QA-Fin, UDA-QA-Tat, UDA-QA-Paper, NqText (Hui et al., 2024), and MINT-1T (Awadalla et al., 2024). For a more challenging evaluation, we synthesized new QA pairs instead of relying on the original data from these datasets. Table 2 provides a summary of the datasets and their key characteristics.

Document Sampling. To mitigate domain imbalance in large datasets (e.g., ArXivCap and MINT-1T, each with over 5k documents), we balance domain distribution through a clustering sampling strategy. Specifically, we first summarize the first page of each document using Qwen2.5-VL-7B (Bai et al., 2025b) to capture its primary topic and domain. Then, we encode these summaries with BGE (Chen et al., 2024) and cluster them into 1K groups using Faiss (Douze et al., 2025). From each group, we select the top 10 highest-scoring documents as representative samples, resulting in 10k documents. After integrating additional sources, we construct a 27K-document pool. To preserve task difficulty, MinerU (Wang et al., 2024a) is applied to filter out documents with fewer than three pages or lacking multiple visual elements (tables, charts, or images), distilling the pool to 10K candidates.

Domain	Type	Dataset	#Doc	#Page
Slides Decks	Slide	SlideVQA	1.1k	21.1k
		FinanceBench	368	53.9k
Finance	Report	UDA-QA-Fin	788	116.5k
		UDA-QA-Tat	170	25.2k
Academia	Paper	ArXivCap	36.5k	443.3k
		UDA-QA-Paper	1.1k	12.2k
World Knowledge	Review	M2QA	1.9k	29.0k
	Wiki	FetaTab	871	13.7k
		NqText	645	9.6k
	Website	MINT-1T	37.7k	449.4k

Table 2: Details of document collection sources.

3.2 Hierarchical QA Pairs Generation

3.2.1 Hierarchical Filtering

Inspired by the LLM/MLLM-as-judge approach (Gu et al., 2025), we introduce a hierarchical QA pairs generation pipeline for *ComDoc* (As shown in Figure 2). The pipeline first employs Qwen2.5-VL-7B (Bai et al., 2025b) to evaluate whether each document meets predefined criteria, producing a binary decision (YES or NO) to filter out noisy samples. Documents passing this stage are subsequently processed by DeepSeek-R1 (Guo et al., 2025) to generate question-answer pairs for various tasks.

Evidence Parsing. Since question-answer generation for single-hop, multi-hop, and multi-page tasks depends on identifying key evidence within documents, we first parse all documents using the MinerU (Wang et al., 2024a) toolkit. This converts PDF pages into three modalities: text (plain text and layout information), images (figures and charts), and tables. We then filter documents according to task-specific evidence requirements.

Statistic	Number
Documents	1,755
- Document Types	6
- Gold/Noisy/Total Pages:	4.1k / 99.8k / 103.9k
- Avg./Med./Max. pages per doc	59 / 20 / 815
- Avg./Med./Max. words per doc	63k / 24.2k / 585k
Total Questions	5561
- Single-Hop(Extraction/Reasoning)	1,000 / 994
- Multi-Hop(Single/Multi-Modal)	989 / 980
- Multi-Page(Single/Multi-Modal)	882 / 716
- Newly-annotated questions	5,561 (100.0%)
(Hop Number Distribution)	6
1: 1,994(35.9%)	2: 2,126(38.2%)
3: 1,235(22.2%)	≥4: 206(3.7%)
(Question Type)	5
Extraction: 1,000 (18.0%)	Reasoning: 994 (17.9%)
Aggregation: 1,384(24.9%)	Bridge: 1,324 (23.8%)
Decomposition Reasoning: 859(15.4%)	
(Evidence Modality)	7
Text: 1,480 (26.6%)	Text + Table: 753 (13.5%)
Table: 825 (14.8%)	Table + Image: 83 (1.5%)
Image: 1,596 (28.7%)	Image + Text: 605 (10.9%)
Table + Text + image: 219 (3.9%)	
Avg./Med./Max words: Question	27 / 26 / 84
Avg./Med./Max words: Answer	41 / 38 / 204

Table 3: ComDoc Statistics Overview.

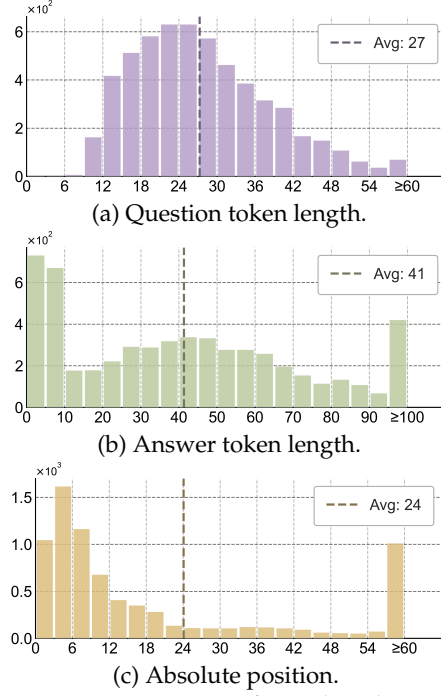


Figure 3: Overview of QA distribution.

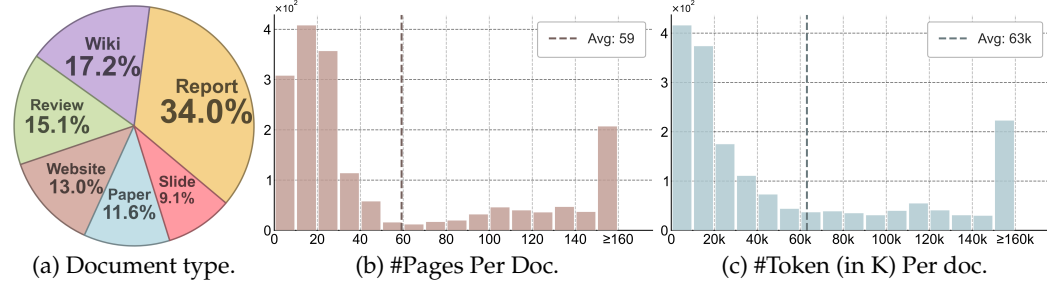


Figure 4: Overview of document distribution. (a) Distribution of document types. (b) Distribution of document page counts. (c) Distribution of document token counts.

Single-Hop Task. We perform page-level filtering based on factual content and image relevance. Initially, we use Qwen2.5-7B to assess whether the text and tables contain factual content (e.g., numbers, entities, time, or conditions). A valid page must provide at least one extractable evidence for extraction tasks and at least two for reasoning tasks. Simultaneously, we resize images and use Qwen2.5-VL-7B to filter out pages with irrelevant or noisy images (e.g., expired graphics and meaningless logos). After filtering, 8,000 documents satisfied the selection criteria, and 1,000 are ultimately selected for dataset construction.

Multi-Hop Task. The multi-hop task requires the identification of relational connections among multiple evidence items within a single page. We first extract subject-verb-object triplets (S,V,O) for each paragraph using Qwen2.5-7B for further analysis of multi-hop relationships. We then employ ChatGPT-4o to generate image captions, focusing on scenes, objects, fine-grained details, and chart metadata. Finally, we feed the triplets, raw text, tables, and captions to Qwen2.5-7B to assess whether they satisfy the following criteria: 1) *Factual Content*: Contains at least three valid facts; 2) *Semantic Relation*: Exhibits explicit semantic relations; 3) *Hop-Chain*: Supports at least one two-hop reasoning chain; and 4) *Multi-Modal*: Input incorporates multiple modalities. This process yields 3,000 eligible documents, from which 1,000 are ultimately retained for dataset construction.

Multi-Page Task. The multi-page task generalizes the multi-hop requirement to span multiple pages. We first cluster the plain text, image titles, and table titles of each document into 10 topic-centered groups using BGE (Chen et al., 2024) and Faiss (Douze et al., 2025),

Model	Single-Hop				Multi-Hop				Multi-Page				Overall	
	Extraction		Reasoning		Single-Modal		Multi-Modal		Single-Modal		Multi-Modal			
	R@3	nDCG@3	R@3	nDCG@3	R@3	nDCG@3	R@3	nDCG@3	R@3	nDCG@3	R@3	nDCG@3	R@3	nDCG@3
OCR + Text Encoders														
BGE-large (0.3B)	43.0	37.5	43.2	37.5	44.0	37.6	49.1	40.9	32.8	32.9	26.6	26.4	39.8	35.5
ColBERT (0.1B)	50.6	44.3	46.2	40.5	42.1	36.0	45.9	39.6	32.8	33.1	25.3	25.8	40.5	36.6
GTE-large (0.4B)	43.6	37.4	44.3	38.4	47.3	39.4	51.6	43.3	32.8	33.1	25.5	26.3	40.9	36.3
E5 (0.3B)	44.8	38.4	45.6	40.4	46.7	40.1	54.2	46.2	33.7	33.9	26.3	27.6	41.9	37.8
OCR + LLM-based Models														
GTE (Qwen2-7B)	36.0	29.8	43.2	36.8	49.1	41.9	54.1	46.1	32.1	31.3	26.5	26.8	40.2	35.5
NVEmbedV2 (8B)	51.6	44.3	54.7	48.1	53.9	47.1	64.5	56.3	41.0	40.6	32.7	32.9	49.7	44.9
Qwen3-Embedding(8B)	54.9	47.1	63.2	56.2	69.2	60.4	73.6	65.4	41.6	41.3	34.6	36.0	56.2	51.1
Multimodal Encoders														
CLIP-large (0.4B)	21.9	18.1	27.1	22.7	23.6	19.3	29.5	24.7	17.2	23.8	17.4	17.6	22.8	21.0
Jina-CLIP-v2 (0.9B)	30.5	25.2	36.3	30.6	37.6	31.6	46.2	39.8	25.7	25.3	22.3	23.6	33.1	29.4
MLLM-based Models														
VLM2Vec (Phi-3.5-V-4.2B)	24.3	20.7	33.4	27.7	36.3	30.5	42.5	36.1	25.3	24.8	20.9	21.6	30.5	26.9
UniME (Phi3.5-V-4.2B)	41.9	36.0	42.9	36.7	52.2	46.4	50.6	43.5	34.3	33.6	27.0	27.6	41.5	37.3
VisRAG-Ret (MiniCPM-3B)	60.7	52.5	62.7	54.8	71.6	63.9	69.8	61.3	40.6	40.7	35.3	36.0	56.8	51.5
ColPali (PaliGemma-3B)	68.5	61.2	71.7	63.2	72.9	65.6	73.3	66.0	43.9	43.5	34.7	35.5	60.8	55.8
GME (Qwen2VL-2B)	69.7	61.8	72.3	64.0	79.6	70.9	78.5	70.6	46.7	46.7	37.7	38.4	64.1	58.7
ColQwen (Qwen2VL-2B)	85.9	78.7	84.0	76.3	85.3	77.7	82.8	75.8	51.8	52.3	39.4	41.1	71.5	67.0
ViDoRAG (Qwen2VL-2B)	87.2	79.9	84.6	76.7	86.7	79.3	83.8	76.8	52.9	53.7	40.4	41.9	72.6	68.1

Table 4: Retrieval performance comparison on *ComDoc*. Models are grouped by type and sorted by overall Recall@3 in ascending order within each group. The best and second best scores are in **boldface** and underlined.

with each cluster representing to a distinct topic. For each cluster, we generate text triplets using Qwen2.5-7B and image captions using Qwen2.5-VL-7B, following the multi-hop procedure. This yields a document-level dataset that distinguishes the multi-page setting from single- and multi-hop tasks. After the clustering and filtering, 2,000 documents meet the selection criteria, of which 1,000 are ultimately retained for dataset construction.

3.2.2 Hierarchical QA Generation

After filtering, we construct three datasets corresponding to each task. We then employ DeepSeek-R1 to generate challenging QA pairs comprising five types: extraction and reasoning for single-hop tasks, decomposition reasoning, bridge, and aggregation for multi-hop and multi-page tasks. Specifically, single-hop and multi-hop tasks operate at the page level, while multi-page tasks are defined at the document level. For single-hop extraction, the LLM processes text, tables, and images separately, producing answers under 20 characters supported by evidence from a single sentence. Single-hop reasoning QA pairs emphasize arithmetic, comparative, or summarization-based reasoning. In multi-hop and multi-page, we enforce the following constraints: 1) *Decomposition Reasoning*: The reasoning chain is structured such that each subsequent question logically depends on the answer to the previous one; 2) *Bridge*: Questions referencing multiple entities require identification of implicit intermediate entities; 3) *Aggregation*: Answers synthesize information via summarization, calculation, comparison, or integration; and 4) *Subtype-Specific Constraints*: Multimodal tasks necessitate cross-modal reasoning at each step, while multi-page tasks require cross-page information retrieval. Ultimately, the generation pipeline outputs a unified document paired with diverse QA instances covering single-hop, multi-hop, and multi-page complexities.

3.3 Quality Control

LLM-as-a-Judge. To ensure benchmark quality, we employ three advanced LLMs: GPT-oss-120B (OpenAI, 2025), Seed-oss-36B-INST (Team, 2025), and DeepSeek-V3.1 (Liu et al., 2024) to evaluate document QA correspondence and overall quality. Each model assigns a score on a three-point scale: 0 (Invalid), 1 (Partially Valid), and 2 (Fully Valid). We retain only instances with an average score above 1.5 across models, filtering out approximately 40% of pages/documents and corresponding QA pairs for each task.

Human Verification. To complement LLM-based quality control, we further perform human verification with ten expert annotators, all holding master’s degrees. We provide

Model	Single-Hop				Multi-Hop				Multi-Page				Overall	
	Extraction		Reasoning		Single-Modal		Multi-Modal		Single-Modal		Multi-Modal			
	Acc	ROUGEL	Acc	ROUGEL	Acc	ROUGEL	Acc	ROUGEL	Acc	ROUGEL	Acc	ROUGEL	Acc	ROUGEL
Large Language Models														
LLaMA3.3-70B	67.6	61.9	32.2	33.0	24.0	16.2	22.8	13.3	17.1	18.3	24.9	14.8	31.4	26.3
Qwen2.5-72B	66.2	62.3	39.5	37.3	33.5	17.2	29.8	13.5	28.8	17.6	34.8	13.0	38.8	26.8
Qwen3-235B-A22B	67.8	62.4	41.6	40.0	38.4	25.3	37.3	22.4	33.6	26.0	40.0	25.5	43.1	33.6
DeepSeek-V3-685B	68.6	63.0	41.7	40.6	39.0	22.2	37.6	18.9	34.1	22.1	38.2	18.3	43.2	30.9
Multimodal Large Language Models														
InternVL3-38B	60.7	58.9	31.3	35.5	25.1	20.3	26.0	17.9	24.9	21.0	25.2	18.4	32.2	28.7
InternVL3-78B	61.2	59.3	34.9	37.6	29.1	27.3	28.3	25.1	28.7	26.4	25.6	24.5	34.6	33.4
InternVL3.5-14B	67.4	61.2	31.6	36.3	28.8	26.1	26.9	23.8	23.4	24.3	24.7	23.3	33.8	32.5
InternVL3.5-38B	69.4	67.1	35.9	39.4	27.8	21.3	29.2	19.0	23.9	20.7	27.5	18.5	35.6	31.0
Llama4-Scout-17B-16E	50.7	39.2	17.4	14.6	14.7	9.2	14.7	6.3	11.3	8.4	13.8	7.7	20.4	14.2
Llama4-Maverick-17B-128E	73.9	69.5	36.1	36.2	31.2	17.0	29.7	13.0	26.9	17.6	30.0	14.3	38.0	27.9
Qwen2.5VL-72B	76.1	73.5	40.9	39.8	31.1	18.7	31.2	15.4	29.5	19.8	31.5	16.4	40.1	30.6
Qwen3-VL-235B-A22B	78.4	73.7	45.4	41.2	37.6	19.0	37.2	15.1	33.7	18.8	38.5	14.5	45.1	30.4

Table 5: Generation performance comparison on *ComDoc*. Models are grouped by type and sorted by overall Acc in ascending order within each group. The best and second best scores are in **boldface** and underlined.

unified annotation guidelines and conduct a calibration phase before formal annotation. For each sampled instance, annotators assess the source document page, question, and reference answer in terms of answerability, faithfulness, clarity, and document quality, while identifying issues such as OCR errors, layout corruption, missing context, and poor readability. Each instance is independently reviewed by two annotators, with disagreements adjudicated by a senior annotator, and inter-annotator agreement is tracked throughout the process. This stage functions as both a high-precision noise filter and an external audit of the automatic scoring pipeline. It produces 5,561 human-verified QA pairs and nearly 0.1 million associated document pages within one week.

3.4 Statistics of *ComDoc*

Table 3 summarizes the statistical information of *ComDoc*. The *ComDoc* benchmark comprises 5.6K questions constructed from 1,755 lengthy documents spanning 6 document types, totaling 103,877 pages. As shown in Figure 4, documents are highly diverse and lengthy, with an average of 59 pages and a maximum of 815. Figure 3 details the question distribution. Each of the six tasks includes over 700 QA pairs, with most questions requiring 2–3 reasoning hops for multi-hop and multi-page tasks. Single-hop questions are evenly split between extraction and reasoning (18% and 17.9% each), while multi-hop and multi-page questions include aggregation (24.9%), bridge (23.8%), and decomposition (15.4%) reasoning. The evidence spans several modalities: single-type formats including text (26.6%), tables (14.8%), and images (28.7%), as well as composed modalities accounting for 29.8%. The average question and answer lengths are 27 and 41 tokens, respectively. The diversity of question types and evidence modalities reflects the breadth and challenge of our benchmark.

4 Experiments and Results

4.1 Evaluation Setup

For retrieval evaluation, we adopt distinct input modalities for text-only and multimodal models. For text-only models, we use the MinerU toolkit (Wang et al., 2024a) to extract text, images, and tables from document pages, with Qwen2.5-VL-7B generating captions for images and tables. We then combine all textual content for downstream processing. For multimodal models, we use raw-page images as input to preserve visual and spatial context, enabling the exploitation of layout, typography, and other visual cues. For generation evaluation, we construct evidence sets of five pages, combining golden pages and pages retrieved by ViDoRAG (Wang et al., 2025d). We apply the same modality-specific settings as in retrieval to evaluate the respective model types.

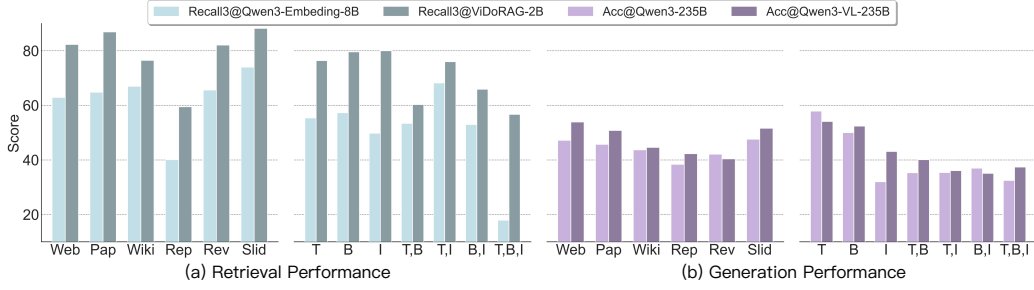


Figure 5: Fine-grained analysis for the effect of domain and evidence modality. T: Text, B: Table, I: Image.

4.2 Baseline Models

Retrieval Models. We evaluate 16 retrieval models across four types on *ComDoc*. For text encoders, we include BGE (Chen et al., 2024), GTE (Li et al., 2023), ColBERT (Khattab & Zaharia, 2020), and E5 (Wang et al., 2024b). We also evaluate LLM-based models such as GTE (Qwen2-7B) (Li et al., 2023), NV-Embed V2 (Lee et al., 2025), and Qwen3-Embedding (Zhang et al., 2025b). For multimodal evaluation, we adopt CLIP (Radford et al., 2021) and Jina-CLIP V2 (Koukounas et al., 2025). To explore the limits of multimodal understanding, we also evaluate advanced MLLM-based models, including VLM2Vec (Jiang et al., 2025), UniME (Gu et al., 2025), VisRAG-Ret (Yu et al., 2025), ColPali (Faysse et al., 2025), GME (Zhang et al., 2025a), ColQwen (Faysse et al., 2025), and ViDoRAG (Wang et al., 2025d).

Generation Models. We conduct a comprehensive evaluation of 13 generation models. For text-only models, we use the Qwen series (Yang et al., 2025), LLaMA3.3 (Grattafiori et al., 2024), and DeepSeek-V3 (Liu et al., 2024) as upper bounds. For multimodal models, we evaluate the InternVL family (InternVL3 (Zhu et al., 2025), InternVL3.5 (Wang et al., 2025e)), the QwenVL series (Qwen2.5-VL (Bai et al., 2025b), Qwen3VL (Bai et al., 2025a)), and the Llama series (Llama-4-Scout, Llama-4-Maverick).

4.3 Evaluation Metric

To evaluate page-level retrieval, we adopt Recall@3 and nDCG@3, aligning with prior works (Jin et al., 2025; Peng et al., 2025). For the generation task, we employ a dual evaluation strategy: we first report ROUGE-L to assess sentence-level quality, following MMDocRAG (Dong et al., 2025b). Additionally, inspired by ViDoRAG (Macé et al., 2025), we utilize Qwen3-32B and ChatGPT-4o (2025-12-02)¹ as evaluators to assign holistic quality scores on a 5-point scale (from 1: completely incorrect to 5: perfectly aligned). For uniformity, all metrics are scaled to percentages. More details are provided in the Appendix.

5 Results and Analysis

5.1 Retrieval Analysis

Table 4 reports retrieval performance across various model types. Compared to prior text encoders (*e.g.*, GTE, E5) and multimodal encoders (*e.g.*, CLIP, Jina-CLIP-v2), LLM and MLLM based models achieve substantial improvements. These gains primarily result from two factors: 1) The vast world knowledge embedded in LLMs further enhances semantic learning and 2) Larger model capacities enable richer semantic representations. Further analysis shows that MLLM-based models consistently outperform LLM-based (OCR) models, primarily due to their ability to jointly process textual and visual information, thus capturing more comprehensive document representations. Notably, ColQwen and ViDoRAG achieve state-of-the-art performance in visual document retrieval. Unlike VLM2Vec and UniME, these models are specifically designed for this task and are trained on large-scale document datasets, which contributes significantly to their effectiveness.

¹<https://azure.microsoft.com>

Takeaway 1: Leveraging rich semantics acquired from large-scale world knowledge and substantial model capacity, LLM-based and MLLM-based retrieval models demonstrate strong performance.

5.2 Generation Analysis

In Table 5, we present generation results using different state-of-the-art LLMs and MLLMs. As the number of model parameters increases, both InternVL-3 and InternVL3.5 show consistent improvements in accuracy. To enable a direct comparison between the text-only Qwen3-235B-A22B and the multimodal Qwen3-VL-235B-A22B, we convert document tables into markdown text and describe images with descriptive captions (Generated by using Qwen2.5-VL-7B) before feeding this unified textual input into the LLM. Notably, although Qwen3-VL-235B-A22B achieves higher overall accuracy, Qwen3-235B-A22B attains competitive performance in the multimodal setting by leveraging textualized representations of non-text modalities. At the sentence level, we observe no substantial quality gap between LLM and MLLM outputs. However, compared to the Intern and Qwen model families, Llama4 generates noticeably lower-quality answers.

Takeaway 2: By incorporating textualized multimodal information, LLMs achieve performance comparable to that of MLLMs.

5.3 Fine-Grained Analysis

Domain Effect. To analyze the impact of the six source types, we evaluate retrieval and generation performance. As illustrated in Figure 5, Qwen3-Embedding and ViDoRAG achieve their lowest retrieval recall (40.1% and 59.5%) on Reports. This difficulty stems from the distinct characteristics of reports: compared to the standardized formats and encyclopedic prose of Reviews and Wikis, reports feature dense terminology, complex hierarchies, and multi-page visual elements requiring cross-element reasoning, thereby complicating query-passage alignment. Furthermore, generation results for Qwen3-235B-A22B and Qwen3-VL-235B-A22B demonstrate a contrasting trend, indicating that Review, Report, and Wiki sources present the greatest challenges for generation tasks.

Takeaway 3: The Report source type presents significant challenges across both retrieval and generation tasks, whereas Review and Wiki sources prove particularly difficult in generation.

Evidence Modality Effect. In Figure 5, we also present retrieval and generation results under different modalities of evidence. On the retrieval task, we present Recall@3 scores for Qwen3-Embedding and ViDoRAG across seven distinct modalities and their combinations. The combined text+table+image modality proves most challenging, yielding the lowest results of 17.9% and 56.7%, respectively. Furthermore, in comparisons between two-modal combinations, text+image presents relatively lower difficulty than the other two pairings. In contrast to the retrieval task, the image modality poses greater challenges than text and table modalities in the generation task. Furthermore, combining different modalities significantly increases the difficulty of generation, as the model must integrate cross-modal information.

Takeaway 4: In retrieval tasks, difficulty increases with the number of evidence modalities. Furthermore, cross-modal integration poses even greater challenges for generation than for retrieval.

6 Conclusion

In this paper, we introduce *ComDoc*, a comprehensive multimodal DocRAG benchmark spanning single-hop, multi-hop, and multi-page settings. *ComDoc* is built from 1,755 real-world documents across four domains, covering 103,877 pages and 5,561 QA pairs, and is constructed through a three-stage pipeline of data collection, hierarchical QA generation, and quality control. Based on *ComDoc*, we conduct a systematic evaluation of representative encoder-based, LLM-based, and MLLM-based retrieval models, as well as the generation capabilities of state-of-the-art open-source LLMs and MLLMs. Our empirical results reveal substantial performance disparities across model architectures, task complexities, and domains, underscoring the unique challenges inherent in multimodal DocRAG.

References

- Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.
- Anas Awadalla, Le Xue, Oscar Lo, Manli Shu, Hannah Lee, Etash Kumar Guha, Matt Jordan, Sheng Shen, Mohamed Awadalla, Silvio Savarese, Caiming Xiong, Ran Xu, Yejin Choi, and Ludwig Schmidt. Mint-1t: Scaling open-source multimodal data by 10x: A multimodal dataset with one trillion tokens, 2024. URL <https://arxiv.org/abs/2406.11271>.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibao Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.
- Ryan C Barron, Maksim E Eren, Olga M Serafimova, Cynthia Matuszek, and Boian S Alexandrov. Bridging legal knowledge and ai: Retrieval-augmented generation with vector stores, knowledge graphs, and hierarchical non-negative matrix factorization. *arXiv preprint arXiv:2502.20364*, 2025.
- Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents, 2023. URL <https://arxiv.org/abs/2308.13418>.
- Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2024.
- Chao Deng, Jiale Yuan, Pi Bu, Peijie Wang, Zhong-Zhi Li, Jian Xu, Xiao-Hui Li, Yuan Gao, Jun Song, Bo Zheng, and Cheng-Lin Liu. Longdocurl: a comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. In *ACL*, 2025.
- Yihao Ding, Siwen Luo, Hyunsuk Chung, and Soyeon Caren Han. Pdfvqa: A new dataset for real-world vqa on pdf documents. In *ECML PKDD*, 2023.
- Kuicai Dong, Yujing Chang, Xin Deik Goh, Dexun Li, Ruiming Tang, and Yong Liu. Mmdocir: Benchmarking multi-modal retrieval for long documents. In *EMNLP*, 2025a.
- Kuicai Dong, Yujing Chang, Shijie Huang, Yasheng Wang, Ruiming Tang, and Yong Liu. Benchmarking retrieval-augmented multimodal generation for document question answering. In *NeurIPS DB*, 2025b.
- Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *IEEE Transactions on Big Data*, 2025.
- Leon Engländer, Hannah Sterz, Clifton Poth, Jonas Pfeiffer, Ilia Kuznetsov, and Iryna Gurevych. M2qa: Multi-domain multilingual question answering, 2024. URL <https://arxiv.org/abs/2407.01091>.

- Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *ICLR*, 2025.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.
- Tiancheng Gu, Kaicheng Yang, Ziyong Feng, Xingjun Wang, Yanzhao Zhang, Dingkun Long, Yingda Chen, Weidong Cai, and Jiankang Deng. Breaking the modality barrier: Universal embedding learning with multimodal llms. In *ACM MM*, 2025.
- Tiancheng Gu, Kaicheng Yang, Kaichen Zhang, Xiang An, Ziyong Feng, Yueyi Zhang, Weidong Cai, Jiankang Deng, and Lidong Bing. Unime-v2: Mllm-as-a-judge for universal multimodal embedding learning. In *AAAI*, pp. 21378–21386, 2026.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Yulong Hui, Yao Lu, and Huanchen Zhang. Uda: A benchmark suite for retrieval augmented generation in real-world document analysis, 2024. URL <https://arxiv.org/abs/2406.15187>.
- Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. Financebench: A new benchmark for financial question answering, 2023. URL <https://arxiv.org/abs/2311.11944>.
- Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhui Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks, 2025. URL <https://arxiv.org/abs/2410.05160>.
- Yiqiao Jin, Rachneet Kaur, Zhen Zeng, Sumitra Ganesh, and Srijan Kumar. Slideagent: Hierarchical agentic framework for multi-page visual document understanding. *arXiv preprint arXiv:2510.26615*, 2025.
- Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *SIGIR*, pp. 39–48, 2020.
- Andreas Koukounas, Georgios Mastrapas, Sedigheh Eslami, Bo Wang, Mohammad Kalim Akram, Michael Günther, Isabelle Mohr, Saba Sturua, Nan Wang, and Han Xiao. jina-clip-v2: Multilingual multimodal embeddings for text and images, 2025. URL <https://arxiv.org/abs/2412.08802>.
- Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoyebi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. In *ICLR*, 2025.
- Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models, 2024. URL <https://arxiv.org/abs/2403.00231>.
- Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.
- Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, et al. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *NeurIPS*, 2024.

- Quentin Macé, António Loison, and Manuel Faysse. Vidore benchmark v2: Raising the bar for visual retrieval. *arXiv preprint arXiv:2505.17166*, 2025.
- Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryściński, Nick Schoelkopf, Riley Kong, Xiangru Tang, Murori Mutuma, Ben Rosand, Isabel Trindade, Renusree Bandaru, Jacob Cunningham, Caiming Xiong, and Dragomir Radev. Fetaqa: Free-form table question answering, 2021. URL <https://arxiv.org/abs/2104.00369>.
- OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025. URL <https://arxiv.org/abs/2508.10925>.
- Xiangyu Peng, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, and Chien-Sheng Wu. Unidoc-bench: A unified benchmark for document-centric multimodal rag. *arXiv preprint arXiv:2510.03663*, 2025.
- Qwen, An Yang, Baosong Yang, Beichen Zhang, et al. Qwen2.5 technical report, 2025. URL <https://arxiv.org/abs/2412.15115>.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL <https://arxiv.org/abs/2103.00020>.
- Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: A dataset for document visual question answering on multiple images. In *AAAI*, pp. 13636–13645, 2023.
- ByteDance Seed Team. Seed-oss open-source models. <https://github.com/ByteDance-Seed/seed-oss>, 2025.
- Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and Conghui He. Mineru: An open-source solution for precise document content extraction, 2024a. URL <https://arxiv.org/abs/2409.18839>.
- Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhui Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*, 2025a.
- Haozhe Wang, Qixin Xu, Che Liu, Junhong Wu, Fangzhen Lin, and Wenhui Chen. Emergent hierarchical reasoning in llms through reinforcement learning. *arXiv preprint arXiv:2509.03646*, 2025b.
- Jingru Wang, Wen Ding, and Xiaotong Zhu. Financial analysis: Intelligent financial data analysis system based on llm-rag. *arXiv preprint arXiv:2504.06279*, 2025c.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*, 2024b.
- Qiuchen Wang, Ruixue Ding, Zehui Chen, Weiqi Wu, Shihang Wang, Pengjun Xie, and Feng Zhao. Vidorag: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. In *EMNLP*, 2025d.
- Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025e.
- Navve Wasserman, Roi Pony, Oshri Naparstek, Adi Raz Goldfarb, Eli Schwartz, Udi Barzeilay, and Leonid Karlinsky. Real-mm-rag: A real-world multi-modal retrieval benchmark. In *ACL*, 2025.

- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
- Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. In *ICLR*, 2025.
- Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Gme: Improving universal multimodal retrieval by multimodal llms. In *CVPR*, 2025a.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models, 2025b. URL <https://arxiv.org/abs/2506.05176>.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL <https://arxiv.org/abs/2504.10479>.
- Anni Zou, Wenhao Yu, Hongming Zhang, Kaixin Ma, Deng Cai, Zhuosheng Zhang, Hai Zhao, and Dong Yu. Docbench: A benchmark for evaluating llm-based document reading systems. In *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pp. 359–373, 2025.

Appendix Overview

The appendix includes the following sections:

- **Appendix A:** We present the experimental setup and outline the retrieval and generation processes.
- **Appendix B:** We provide case studies for each task type, encompassing single-hop tasks (extraction and reasoning) as well as both single-modality and multi-modality scenarios for multi-hop and multi-page tasks.
- **Appendix C:** We provide some prompts used in the benchmark construction pipeline, including four prompts for data collection, hierarchical QA generation, quality control, and final evaluation.
- **Appendix D:** We review the license agreements for *ComDoc* and all artifacts used in its construction.

A Experiment Setup

A.1 Retrieval Experiment

We evaluate 16 state-of-the-art retrieval models on 8 NVIDIA A800 GPUs. Processing times for the evaluation dataset vary significantly across models, providing insights into their practical deployment feasibility. Optimized models such as E5, GTE, and BGE process the dataset in as little as 10 seconds, while more computationally intensive models like GME and VisRAG require 5 hours and 3 hours, respectively.

A.2 Generation Experiment

During the generation stage, the retrieved context is input to an LLM to produce the final output. The hardware selection for this stage depends on the size of the generative model. Most models run on an NVIDIA L20 GPU, while larger models, such as those with 72 billion parameters, utilize the NVIDIA A800 GPU to ensure efficient inference. For the largest models in our study (e.g., Deepseek-V3), we employ commercial APIs ² due to their scale and the high cost of local deployment. Generation times vary by model architecture: text-only LLMs require approximately 7 hours per experimental run, whereas other models typically complete this stage within 4 to 5 hours.

B Case Study

Examples of *ComDoc* across different task types are shown in Figure 6 (single-hop extraction), Figure 7 (single-hop reasoning), Figure 8 (multi-hop single-modal), Figure 9 (multi-hop multi-modal), Figure 10 (multi-page single-modal), and Figure 11 (multi-page multi-modal).

C Prompts Used in Pipeline

Data Collection. As described in Section 3.1, we employ Qwen2.5-VL-7B to summarize the first page of each document for initial filtering, using the prompt shown in Figure 12.

Hierarchical QA Generation. Following the procedure in Section 3.1, we first use Qwen2.5-7B to generate triplets with the prompt in Figure 13, and ChatGPT-4o (2025-12-02) to produce image captions with the prompt in Figure 14. For the multi-page multimodal task, we then apply the prompt in Figure 15 and use DeepSeek-R1 to generate QA pairs, with the corresponding prompt shown in Figure 16.

Quality Control. As described in Section 3.3, we employ a voting mechanism with GPT-oss-120B (OpenAI, 2025), Seed-oss-36B-INST (Team, 2025), and DeepSeek-V3.1 (Liu et al.,

²<https://azure.microsoft.com>

Question:
What technology does Bell Fibe Internet use for FTTP connections?

Reasoning Chain:
None

Answer:
Fibre optic

Evidence Types:
Text

Reference Pages:
32

Json Format:

```
{
  "uid": "tat_340",
  "query": "What technology does Bell Fibe Internet use for FTTP connections?",
  "reference_answer": "fibre optic",
  "meta_info": {
    "file_name": "bce-inc_2019",
    "reference_page": [
      32
    ],
    "source_type": "tat",
    "evidence_type": [
      "text"
    ]
  }
}
```

Page: 32

Figure 6: This example shows a typical Single-Hop Extraction task.

Question:
What is the combined sales percentage of the Far East and Middle East for Cubic in 2019, and how does it compare to Australia?

Reasoning Chain:
Step 1: Calculate the combined percentage for the Far East and the Middle East: $3\% + 2\% = 5\%$
Step 2: Identify the sales percentage for Australia: 11%
Step 3: Calculate the difference: $11\% - 5\% = 6\%$

Answer:
Combined percentage is 5%, 6% lower than Australia

Evidence Types:
Image

Reference Pages:
2

Json Format:

```
{
  "uid": "tat_894",
  "query": "What is the combined sales percentage of the Far East and Middle East for Cubic in 2019, and how does it compare to Australia?",
  "reference_answer": "Combined percentage is 5%, 6% lower than Australia",
  "meta_info": {
    "file_name": "cubic-corp_2019",
    "reference_page": [
      2
    ],
    "source_type": "tat",
    "evidence_type": [
      "image"
    ]
  },
  "reasoning_chain": [
    "Step 1: Calculate the combined percentage for the Far East and the Middle East: 3% + 2% = 5%",
    "Step 2: Identify the sales percentage for Australia: 11%",
    "Step 3: Calculate the difference: 11% - 5% = 6%"
  ]
}
```

Page: 2

Figure 7: This example shows a typical Single-Hop Reasoning task.

2024) to filter samples. Evaluation templates are provided in Figure 17 for multi-hop and multi-page tasks.

LLM-as-a-Judge Evaluation. Figure 18 presents the prompt used by Qwen3-32B and ChatGPT-4o (2025-12-02) to assign holistic generation scores on a 5-point scale.

Question:

What was the total shareholder return combination (dividends per share plus earnings per share) for Caterpillar Inc. in 2018?

Reasoning Chain:**Step1**

Q: Retrieve the Earnings Per Share (EPS) for 2018 from the Profit/Loss Per Share image.

A: The Profit(Loss) Per Share image shows 2018 Earnings Per Share (EPS) of \$10.26.

Step2

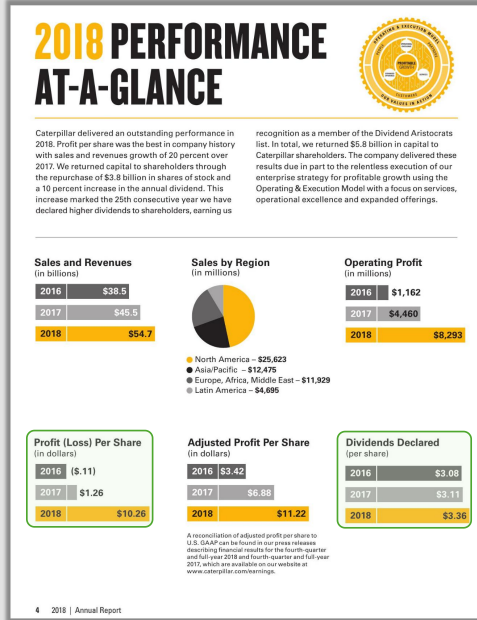
Q: Retrieve the Dividends Per Share for 2018 from the Dividends Declared image.

A: The Dividends Declared image shows 2018 Dividends Declared Per Share of \$3.36.

Step3

Q: Calculate the Total Shareholder Return Per Share for 2018.

A: The Total Shareholder Return Per Share for 2018 is $\$10.26 + \$3.36 = \$13.62$.

**Answer:**

In 2018, Caterpillar Inc.'s total shareholder return per share was \$13.62, consisting of \$10.26 earnings per share plus \$3.36 dividends per share.

Evidence Types:

Images

Reference Pages:

6

Json Format:

```
{
  "uid": "fin_1540",
  "query": "What was the total shareholder return combination (dividends per share plus earnings per share) for Caterpillar Inc. in 2018?",
  "reference_answer": "In 2018, Caterpillar Inc.'s total shareholder return per share was $13.62, consisting of $10.26 earnings per share plus $3.36 dividends per share.",
  "meta_info": {
    "file_name": "CAT_2018",
    "reference_page": [
      6
    ],
    "source_type": "fin",
    "query_type": "image_aggregation"
  },
  "evidence_type": [
    "image",
    "image"
  ],
  "reasoning_chain": [.....]
}
```

Figure 8: This example shows a typical Multi-Hop Single-Modal task.

D License Agreement

ComDoc is constructed from 8 open-source large-scale datasets: SlideVQA (Tanaka et al., 2023), FinanceBench (Islam et al., 2023), UDA (Hui et al., 2024), ArxivCap (Li et al., 2024), M2QA (Engländer et al., 2024), FeTaTab (Nan et al., 2021), NqText (Hui et al., 2024), and MINT-1T (Awadalla et al., 2024). All source datasets are released under the Apache License 2.0, which permits use, reproduction, and distribution for research purposes, subject to its terms. All new annotations (e.g., question-answer pairs, evidence annotations, and reasoning hops) introduced in this work are publicly available for research use only, and may be used, modified, and shared for academic and non-commercial research activities.

Question:

What is the difference, in US dollars, between the nominal gold price per gram in April 2025 and the inflation-adjusted peak price per ounce in 1980?

Reasoning Chain:**Step1**

Q: Retrieve the Nominal Gold Price per Gram in April 2025 from the text.

A: \$106 per gram

Step2

Q: Retrieve the Inflation-Adjusted Peak Price per Ounce in 1980 from the image.

A: The image caption mentions that the inflation-adjusted peak price in 1980 is approximately \$3,680 per troy ounce.

Step3

Q: Perform unit conversion and calculate the price difference

A: \$3,680 per troy ounce $\approx$ \$118.3 per gram; difference: \$12.3 per gram

Answer:

In April 2025, the nominal gold price stands at \$106 per gram, while the inflation-adjusted peak from 1980 equates to approximately \$118.3 per gram (from \$3,680 per troy ounce). The difference of \$12.3 per gram indicates that current prices remain below the historical inflation-adjusted peak.

Evidence Types:

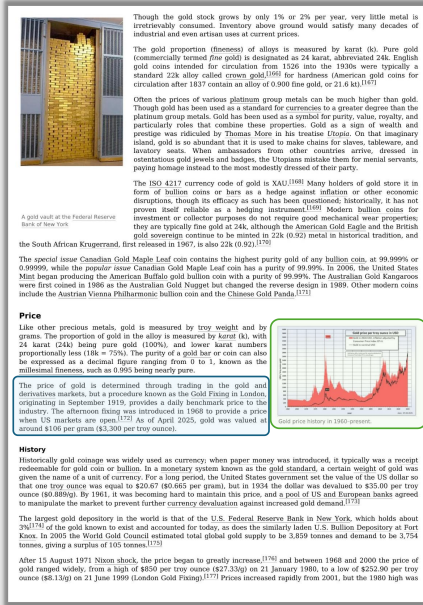
Text, Image

Reference Pages:

14

Json Format:

```
{
  "uid": "M2QA_Bench_1000",
  "query": "What is the difference, in US dollars, between the nominal gold price per gram in April 2025 and the inflation-adjusted peak price per ounce in 1980?",
  "reference_answer": "In April 2025, the nominal gold price stands at $106 per gram, while the inflation-adjusted peak from 1980 equates to approximately $118.3 per gram (from $3,680 per troy ounce). The difference of $12.3 per gram indicates that current prices remain below the historical inflation-adjusted peak.",
  "meta_info": {
    "file_name": "Gold",
    "reference_page": [
      14
    ],
    "source_type": "M2QA_Bench",
    "query_type": "aggregation"
  },
  "evidence_type": [
    "image",
    "text"
  ],
  "reasoning_chain": [
    .....
  ]
}
```



Page: 14

Figure 9: This example shows a typical Multi-Hop Multi-Modal task.

Page: 7

Page: 9

Page: 4

Question:
What was the jersey number worn by the player who became teammates with Kyrie Irving after reaching 20,000 career points?

Reasoning Chain:
Step1
Q: Identify the player who became the youngest to reach 20,000 career points in the NBA
A: LeBron James
Step2
Q: Find the team LeBron James was with when he became teammates with Kyrie Irving.
A: The images show James protecting the ball from Kyrie Irving in October 2017; the title indicates that they served together for three seasons in Cleveland
Step3
Q: Identify the jersey number LeBron James wore while playing for the Cleveland Cavaliers.
A: The images show James wearing the number 23 during his time with the Cleveland Cavaliers in 2007.

Answer:
23

Evidence Types:
Images

Reference Pages:
7, 9, 4

Json Format:

```
{
  "uid": "M2QA_Bench_565",
  "query": "What was the jersey number worn by the player who became teammates with Kyrie Irving after reaching 20,000 career points?",
  "reference_answer": "LeBron James reached 20,000 career points in 2013 and later became teammates with Kyrie Irving on the Cleveland Cavaliers, wearing jersey number 23.",
  "meta_info": {
    "file_name": "LeBron_James",
    "reference_page": [
      9,
      4,
      7
    ],
    "source_type": "M2QA_Bench",
    "query_type": "image_sequential_reasoning"
  },
  "evidence_type": [
    "image",
    "image",
    "image"
  ],
  "reasoning_chain": "[.....]"
}
```

Figure 10: This example shows a typical Multi-Page Single-Modal task.

Question:
What was JPMorgan Chase's Return on Equity (ROE) in 2004, and how does this metric relate to the company's overall operating environment at the time?

Reasoning Chain:
Step1
 Q: Extract the 2004 Return on Equity (ROE) data.
 A: 17%
Step2
 Q: Collect key macroeconomic indicators for 2004.
 A: U.S. GDP grew by 4.4%, the S&P 500 rose 9%, and the Federal Funds Rate increased from 1% to 2.25%.
Step3
 Q: Extract credit-rating evidence to substantiate financial health.
 A: Parent company long-term debt rating Aa3/A+
Step4
 Q: Analyze the contribution of investment banking performance to Return on Equity (ROE).
 A: Improved M&A ranking (No. 1 in the U.S., No. 2 globally) and enhanced equity underwriting ranking underpinned (or supported) the ROE performance.

Answer:
 In 2004, JPMorgan Chase achieved a return on equity (ROE) of 17%. This metric was closely tied to the company's operating environment: 1) Macroeconomic conditions were favorable, with U.S. GDP growing by 4.4% and the S&P 500 index rising 9%; 2) At the business level, its investment banking operations excelled (ranking 1st in U.S. M&A, 2nd globally, and climbing to 4th in IPO underwriting); 3) Financially, the firm's Aa3/A+ credit ratings reflected robust health. These factors collectively supported the 17% ROE performance.

Evidence Types:
 Texts, Table

Reference Pages:
 32, 22, 58, 33

Json Format:

```
{
  "uid": "fin_894",
  "query": "...",
  "reference_answer": "...",
  "meta_info": {
    "file_name": "JPM_2004",
    "reference_page": [32, 22, 58, 33],
    "source_type": "M2QA_Bench",
    "query_type": "image_sequential_reasoning"
  },
  "evidence_type": [
    "text", "text", "table", "text",
  ],
  "reasoning_chain": [...] }

```

Figure 11: This example shows a typical Multi-Page Multi-Modal task.



System Prompt:
You are an efficient document analysis assistant.

Your task is to quickly extract the core information from the document, with special attention to entities and key concepts in the title, as well as document type and academic domain classification.

Strictly analyze the document and output the following **6 core elements** (in English). *Pay special attention to entities and concepts mentioned in the title:*

- 1. Core Theme** (1 sentence):
 - Summarize the main content of the entire document in one precise sentence.
 - Example: "Discusses deep learning model compression techniques for mobile AI deployments."
- 2. Key Entities** (3-5):
 - List most important entities (people, organizations, locations, events) in order of importance.
 - **Key:** Prioritize entities mentioned in the document title.
 - Example: ["ResNet architecture", "TensorFlow Lite", "Stanford University"]
- 3. Main Purpose** (1 sentence):
 - Summarize the main objective or motivation.
 - Example: "Proposes a lightweight neural network pruning scheme."
- 4. Key Concepts** (3-5 keywords):
 - Extract essential technical or academic concepts.
 - **Key:** Prioritize concepts mentioned in the document title.
 - Example: ["Model pruning", "Knowledge distillation", "Computational efficiency"]

- 5. Document Type** (1 category):
 - Classify based on content/structure. Choose **ONLY** one:
 - Options: Research Paper, Economic Document, Presentation Slides, Promotional Poster, News Article, Wikipedia Document, User Manual, Other.
 - Example: "Research Paper"
- 6. Domain** (1 category):
 - Identify main academic/technical domain. Choose **ONLY** one:
 - Options: Science, Art & Design, Health & Medicine, Tech & Engineering, Business, Humanities & Social Sci., Other.
 - Example: "Tech & Engineering"

Output Requirements:

- Strictly output in JSON format, no explanatory text.
- **Special attention:** Extract entities/concepts from title.
- Document type and domain must be a single, specific category.
- Ensure content is concise and accurate.
- If undetermined, use empty array [] or string "".

Input format:
Image: IMAGE

Output Format (JSON only, no text):

```
{
  "core_themes": "",
  "key_entities": [],
  "document_purpose": "",
  "key_concepts": [],
  "document_type": "",
  "domains": ""
}
```

Figure 12: Prompt template for efficiently summarizing document topic and domain.



System Prompt:
You are a relation extraction expert, specialized in extracting relation triplets for building multi-hop question answering datasets.

Please extract relation information from the given text that can support reasoning chains.

Entity Type Restrictions (only extract these types):

- Organization:** companies, institutions, governments, schools, etc.
- Person:** names, positions, roles, etc.
- Location:** countries, cities, regions, addresses, etc.
- Time:** years, dates, periods, eras, etc.
- Product:** product names, technologies, services, etc.
- Event:** meetings, activities, incidents, achievements, etc.
- Concept:** policies, theories, methods, technologies, etc.

Multi-hop Reasoning Supported Relation Types (priority):

- Causal Reasoning** (Most Important):
 - Keywords: cause, lead to, result in, because, due to, etc.
 - Example: A causes B → (A, causes, B)
- Hierarchical Inclusion** (Important):
 - Keywords: contain, belong to, part of, manage, lead, etc.
 - Example: A contains B → (A, contains, B)
- Temporal Sequence** (Important):
 - Keywords: before, after, during, start at, occur at, etc.
 - Example: A occurs before B → (A, occurs before, B)
- Spatial Location** (Medium):
 - Keywords: located at, in, from, adjacent to, near, etc.
 - Example: A is located at B → (A, is located at, B)
- Attribute State** (Medium):
 - Keywords: be, become, serve as, have, similar to, etc.
 - Example: A is B → (A, is, B)
- Action Interaction** (Basic):
 - Keywords: create, buy, visit, cooperate, support, etc.
 - Example: A creates B → (A, creates, B)

Strict Restriction Conditions:

- Entity Quality:** Specific/clear (no pronouns); Length ≤ 20 chars; Unique/consistent; Type matching.
- Relation Quality:** Explicit (no speculation); Clear direction (A→B); Support reasoning; Standardized vocabulary.
- Triplet Quantity:** Min 3, Max 15. Prefer fewer high-quality.
- Reasoning Chain:** Min 2 hops, Max 5 hops. Complete/verifiable.
- Forbidden Relations:** Abstract ("related to"), Subjective ("good"), Distant/vague time, Missing entities.

Cross-paragraph Reasoning Requirements:

- Multi-paragraph coverage:** Chain involves ≥ 2 paragraphs.
- Paragraph distribution:** Triplets distributed across paragraphs.
- Connection:** Prioritize relations connecting different paragraphs.
- Diversity:** More paragraphs involved = better.
- Strict uniqueness:** Every step in a chain must have a different paragraph id.

Reasoning Chain Extraction Rules:

- Identification:** Identify sequences forming reasoning paths.
- Numbering:** Unique chain_id (chain_1, chain_2...).
- Lists:** List entities and relations in reasoning order.
- Path:** Use arrows (→) to represent complete paths.
- Hops:** hops = number of relations (starts from 1).
- Description:** Briefly describe logical meaning.
- Paragraph ID Uniqueness:** All steps must use different paragraph ids.

Extraction Strategy:

- Prioritize causal/hierarchical relations.
- Focus on entity repetition (key to multi-hop).
- Ensure relation completeness for reasoning.
- Avoid redundancy (extract triplet only once).
- Emphasize reasoning potential (connects multiple entities).

Quality Checklist (Check before output):

- Entities specific/clear; Length ≤ 20 chars.
- Relation vocabulary standardized; No duplicates.
- Triplet count 3-15; All have correct paragraph #.
- Forms ≥ 2-hop chains; Hops calc correct.
- Chain count 1-3; Involves ≥ 2 paragraphs (coverage ≥ 2).
- Strict:** No duplicate paragraph ids in a single chain.

Important Notes:

- Only extract explicitly existing relations.
- Return empty string if no multi-hop support found.
- Ensure paragraph numbering matches text order (starts from 1).
- Priority:** Cross-paragraph multi-hop chains.
- Forbidden:** Two steps in same chain using same paragraph id.

Input format:
Text: TEXT

Output Format (Strict JSON, no analysis):

```
{
  "triplets": [
    {
      "subject": "subject", "relation": "relation",
      "object": "object", "paragraph": "para_num",
      "page_idx": "page_index"
    }
  ],
  "reasoning_chains": [
    {
      "chain_id": "chain_1",
      "entities": ["ent1", "ent2", "ent3"],
      "relations": ["rel1", "rel2"],
      "path": "ent1 -> rel1 -> ent2 -> ...",
      "hops": 2,
      "paragraphs": [1, 2],
      "page_idxs": [1, 2],
      "paragraph_coverage": 2,
      "description": "Brief description..."
    }
  ]
}
```

Figure 13: Prompt template for triplets extraction.

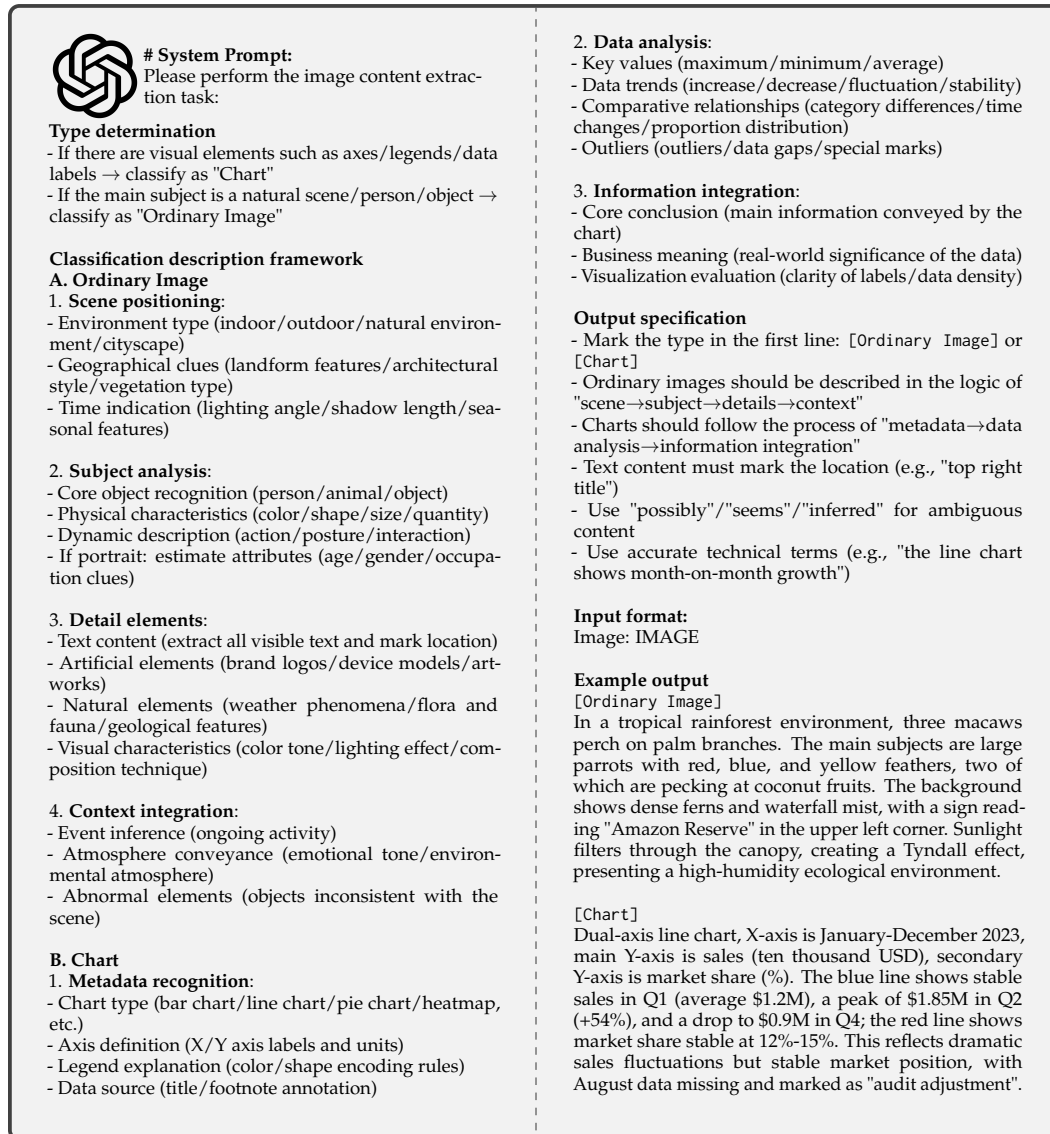


Figure 14: Prompt template for image captioning.

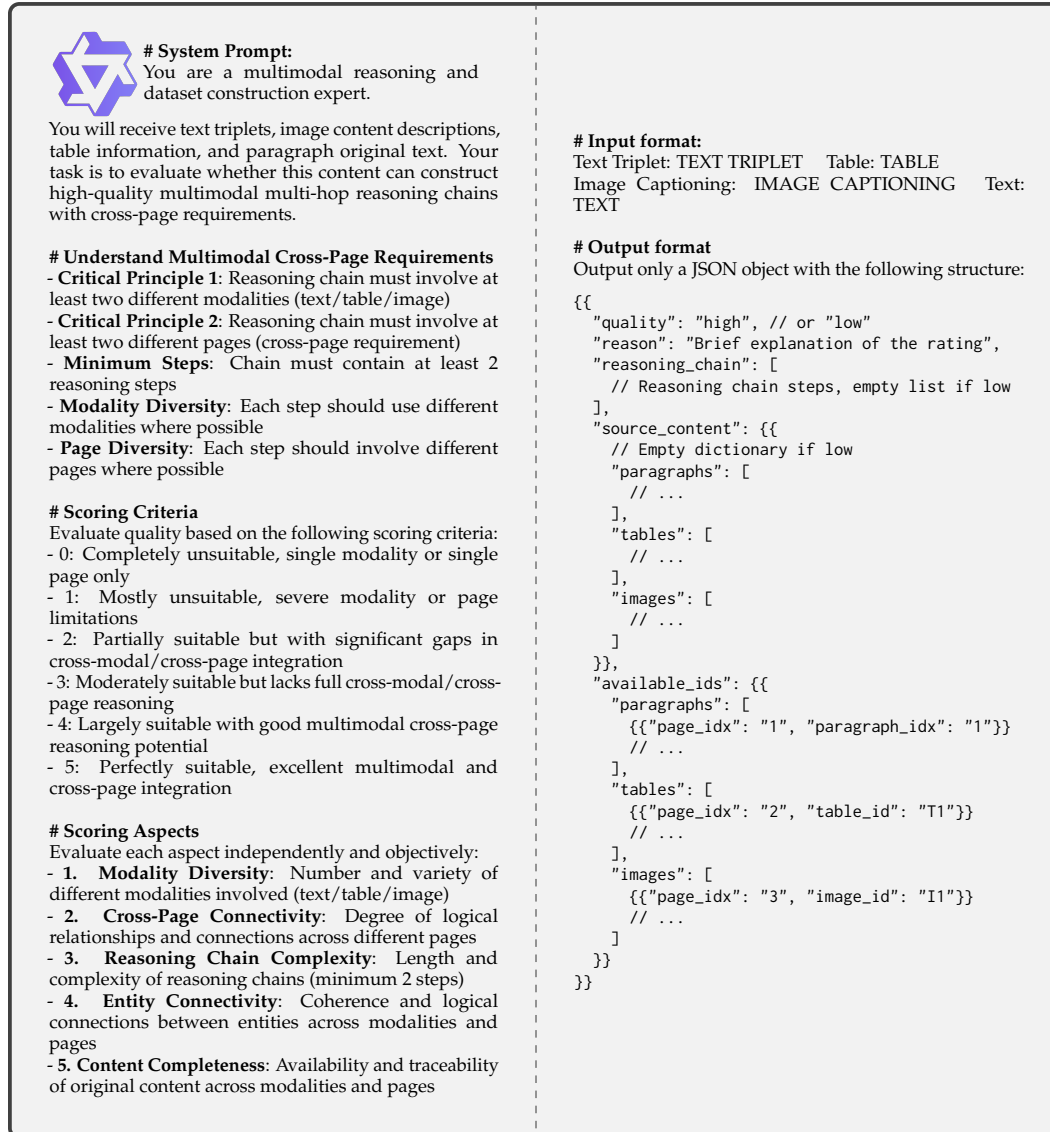


Figure 15: Prompt template for Multi-Page Multi-Modal reasoning chain assessment.

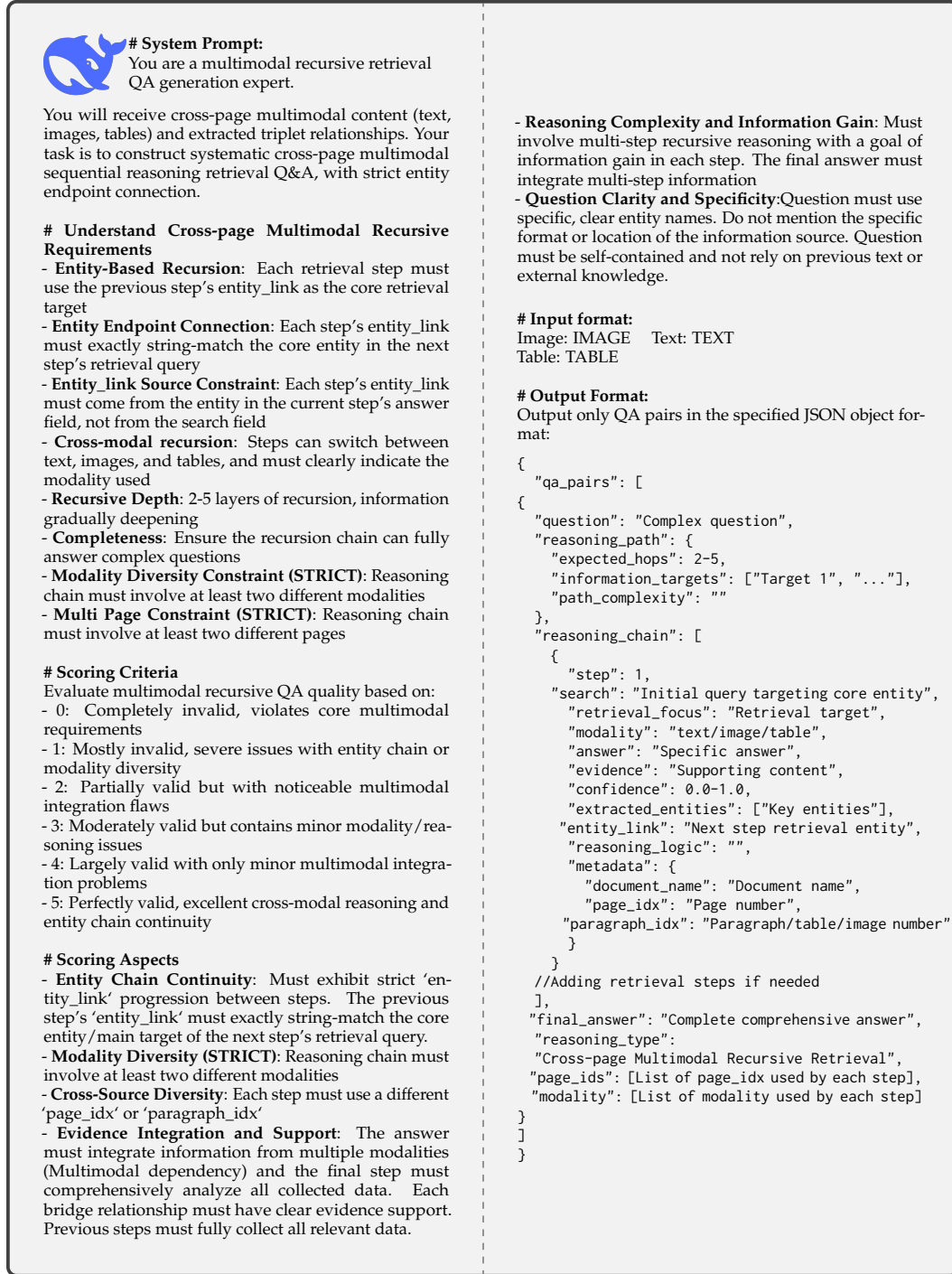


Figure 16: Prompt template for Multi-Page Multi-Modal task QA pairs generation.



You are a quality assessment expert for multi-hop question-answering sequential reasoning chain (Sequential Reasoning/bridge/aggregation) datasets. Please strictly follow the following standards to evaluate each group of sequential reasoning chain Q&A for type judgment and correction, as well as chain quality review, and output detailed assessment opinions and pass/fail conclusions.

For Sequential QA pairs

Step 1: Type Judgment and Correction
Please judge which type the Q&A sample belongs to and correct if errors exist:

- Sequential Reasoning Chain:** Each step's retrieval target depends on the answer from the previous step. The main question cannot directly decompose all retrieval targets. Must advance through step-by-step retrieval. Suitable for tracking single entity/information.
- Bridge Chain:** Contains bridge entities connecting two or more explicit entities. Answer depends on bridge relationships. Suitable for implicit entity connections.
- Aggregation Chain:** Multiple steps retrieve fragments; final step summarizes/compares/counts. Suitable for integrating multi-source information.
- Correction Rule:** If type is incorrect, modify to the correct type and output the corrected reasoning chain.

Step 2: Sequential Reasoning Chain Quality Review

- Question Standardization**
 - Single main question (no listing of sub-questions).
 - Explicit entities only (no pronouns like "that country", "he/she").
 - Self-consistent and complete without context.
- Reasoning Chain Compliance**
 - entity_link must strictly appear in next step's query and come from current step's answer.
 - entity_link cannot come from search field.
 - Length: 2-5 steps. Progressive information, no skipping/breaks.
 - Exact string matching for entity_link.
- Answer Reasonableness**
 - Comprehensive result of multi-step reasoning (not from single step).
 - Consistent with chain content; no external info.
- Error Case Detection**
 - entity_link missing in next query / not from answer / from search.
 - Vague references / Non-compliant step count.

For Bridge QA pairs

Step 1: Type Judgment and Correction
Please judge type according to definitions and correct if errors exist:

- Sequential Reasoning Chain:**
- Bridge Chain:**
- Aggregation Chain:** Multiple steps respectively retrieve information fragments, and the final step summarizes, compares, counts, or concludes to form a comprehensive answer. Suitable for reasoning that requires integrating multi-source information to reach conclusions.
- If type incorrect, modify and output corrected chain.
- If chain has problems, correct it.

Step 2: Bridge Chain Quality Review

- Question Standardization**
 - Explicit entities only (no pronouns/vague expressions).
 - Self-consistent and complete without context.
- Reasoning Chain Compliance**
 - Must have bridge entities connecting multiple steps.
 - Length: 2-5 steps, progressive, no breaks.
 - Bridge relationships must have clear evidence support.
- Answer Reasonableness**
 - Depends on bridge entity reasoning (not from single step).
 - Consistent with chain content; no external info.
- Error Case Detection**
 - Missing bridge entities / Evidence lack support.
 - Vague references / Non-compliant step count.

For Aggregation QA pairs

Step 1: Type Judgment and Correction
Please judge type according to definitions and correct if errors exist:

- Sequential Reasoning Chain:**
- Bridge Chain:**
- Aggregation Chain:** Multiple steps respectively retrieve information fragments, and the final step summarizes, compares, counts, or concludes to form a comprehensive answer. Suitable for reasoning that requires integrating multi-source information to reach conclusions.
- If type incorrect, modify and output corrected chain.
- If chain has problems, correct it.

Step 2: Aggregation Chain Quality Review

- Question Standardization**
 - Explicit entities only; self-consistent and complete.
- Reasoning Chain Compliance**
 - Steps retrieve independent fragments.
 - Final step: clear summarization/comparison/counting.
 - Length: 2-5 steps. Conclusions based on all previous results.
- Answer Reasonableness**
 - Comprehensive summarization result.
 - Consistent with chain content.
- Error Case Detection**
 - Missing conclusions / Conclusions unrelated to results.
 - Vague references / Non-compliant step count.

Input format
Original data: ORIGINAL DATA
Generated retrieval QA: GENERATED QA

Output format (Strictly JSON):

```
{
  "type_judgement": {
    "original_type": "Sequential Reasoning Chain...",
    "correct_type": "Sequential Reasoning Chain...",
    "correct_reasoning_chain": [...],
    "correct_final_answer": "Answer from correct chain",
    "reason": "Reason for type judgment and correction"
  },
  "chain_quality": {
    "pass": true/false,
    "problems": ["Problem description 1"...],
    "suggestion": "If not passed, how to modify/improve"
  }
}
```

Figure 17: Prompt template for Multi-Hop and Multi-Page QA pairs quality assessment.

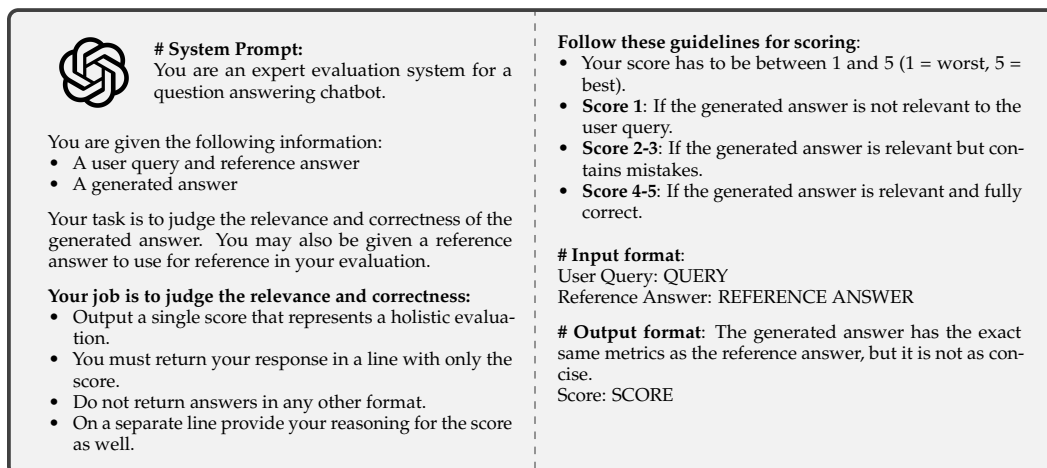


Figure 18: Prompt template for GPT-4o for answer evaluation.