UniDoc-Bench: A Unified Benchmark for Document-Centric Multimodal RAG

Anonymous ACL submission

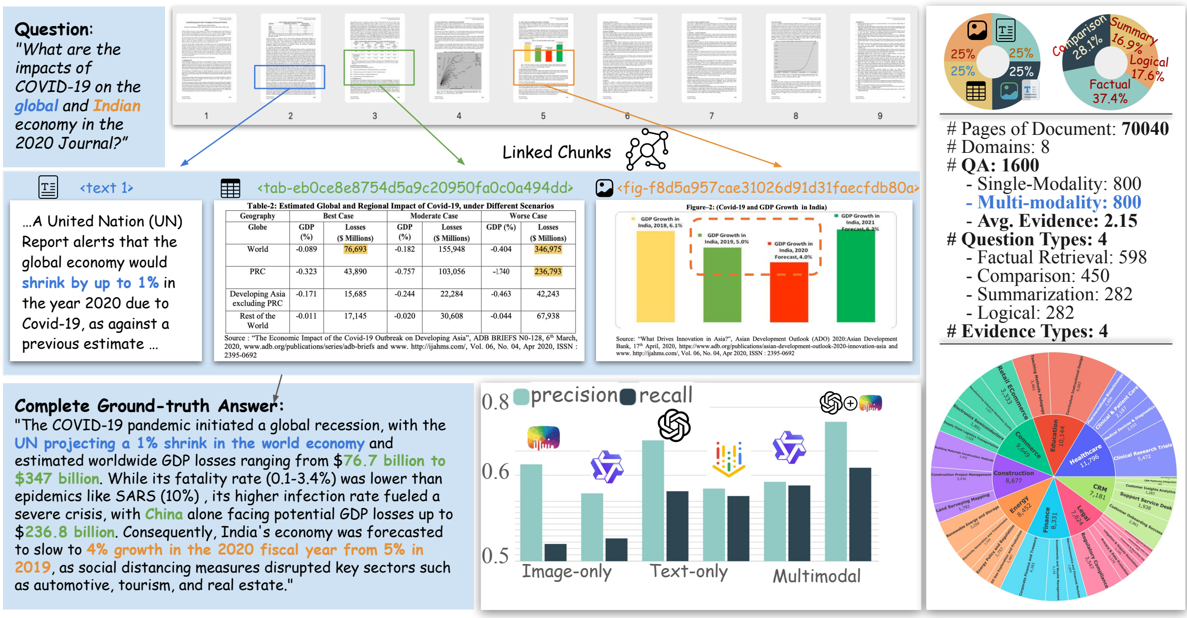
Abstract

Multimodal retrieval-augmented Generation (MM-RAG) is a key approach for applying large language models and agents to real-world knowledge bases, yet current evaluations are fragmented—focusing on either text or images in isolation, or simplified multimodal setup, failing to capture document-centric multimodal use cases. In this paper, we introduce `UniDoc-Bench`, the first large-scale, realistic benchmark for MM-RAG built from 70k real-world PDF pages across 8 domains. Our pipeline extracts and links evidence from text, tables, and figures, then generates $1,600$ multimodal QA pairs spanning factual retrieval, comparison, summarization, and logical reasoning queries. To ensure reliability, all of QA pairs are validated and rewritten by multiple human annotators and expert adjudication. `UniDoc-Bench` supports apples-to-apples comparison across four paradigms — 1) text-only, 2) image-only, 3) *multimodal* text–image fusion and 4) *multimodal* joint retrieval — under a unified protocol with standardized candidate pools, prompts, and evaluation metrics. `UniDoc-Bench` can also be used to evaluate Visual Question Answering tasks. Our experiments show that multimodal text–image fusion RAG systems outperform both unimodal and jointly multimodal embedding–based retrieval, indicating that neither text nor images alone are sufficient and that current multimodal embeddings remain inadequate. Beyond benchmarking, our analysis reveals when and how visual context complements textual evidence, uncovers systematic failure modes, and offers actionable guidance for developing more robust MM-RAG pipelines.

1 Introduction

Retrieval-augmented generation (RAG) has become a widely used approach for applying large language models (LLMs) and agents to real-world knowledge bases (Gao et al., 2023; Fan et al., 2024). The dominant text-only pipeline applies Optical Character Recognition (OCR) (Li et al., 2022; Xue et al., 2024; Poznanski et al., 2025) to flatten document pages into text, indexes them as chunks, retrieves top-k text passages, and feeds them to a generator. However, many answers depend on information embedded in figures, charts, tables, and complex layouts, where OCR often discards crucial spatial and visual semantics (e.g., map, axes, bar lengths, color encodings) (Ma et al., 2024a; Faysse et al., 2024). These limitations have driven the rapid development of multimodal RAG (MM-RAG), which embeds documents across modalities (text, tables, and images) and retrieves and reasons over them jointly, emerging as a key paradigm for document intelligence.

Current MM-RAG evaluation benchmarks exhibit substantial limitations, as summarized in Table 1. Many are restricted to a single image or a single document page as reference (Mathew et al., 2021, 2022; Zhu et al., 2022; Li et al., 2024; Ma et al., 2024b), cover narrow domains (Mathew et al., 2021, 2022; Zhu et al., 2022; Li et al., 2024), under-represent modalities (Li et al., 2024; Mathew et al., 2022), operate at limited scale (few queries/pages) (Ma et al., 2024b; Wang et al., 2025a) or lack a highly relevant database for RAG evaluation (Ma et al., 2024b). These gaps hinder fair and comprehensive comparison across methods. Moreover, debatable claims have emerged — such as that “image retrieval is all you need" (Faysse et al., 2024; Su et al., 2025) or that multimodal retrieval is inherently superior (Zhang et al., 2024a; Yu et al., 2024a)— without enough fair and unified evaluation. In response, we introduce `UniDoc-Bench`, a human-verified benchmark spanning 8 domains and covering text, chart, and table content, explicitly designed for cross-modality grounding with examples shown in Figure 1. Crucially, `UniDoc-Bench` enables apples-to-apples evaluation of text-retrieval, image-retrieval,



(a) Dataset example (b) Performance overview (c) Dataset statistics

Figure 1: UniDoc-Bench overview.

multimodal text-image-fusion retrieval, and multimodal joint retrieval pipelines using highly relevant large document database and multi-type, cross-modality-grounding queries under a unified protocol. This provides an unbiased view of when multimodal retrieval offers advantages beyond single modalities. In practice, UniDoc-Bench quantifies multimodal gains, guides system design choices, and accelerates the development of effective MM-RAG systems for real-world document intelligence.

We curate a high-quality multimodal RAG evaluation benchmark by designing and applying a classification-based filtering scheme to unlabeled, real-world PDF documents (PDFA (Montalvo and Wightman, 2024)), yielding 70k highly relevant pages across eight widely used domains —*Finance, Legal, Healthcare, Commerce and Manufacturing, CRM, Energy, Education, and Construction*— containing rich cross-modality content, including text, tables, and images. We construct a knowledge graph that links cross-modality contents across documents via overlapping entities, and leverage these connections to synthesize 1,600 QA pairs spanning four question types: *factual retrieval*, *comparison*, *summarization*, and *logical reasoning*, enabling multi-modality grounding and reflecting realistic retrieval scenarios. To ensure quality, all of the QA pairs are evaluated and rewritten by three independent annotators for faithfulness, completeness, self-containment, human intent, and evidence usability, with disagreements resolved through expert adjudication. Figure 2 illustrates the full pipeline from PDF segmentation to dataset creation and evaluation.

In this paper, we compare text-only, image-only, multimodal joint, and text-image-fusion retrieval augmented generation pipelines under a unified setup, using identical candidate pools, fixed top-k, consistent prompts, and standardized evaluation criteria. We report retrieval metrics (Recall@10, Precision@10), answer completeness and faithfulness defined at Section 4.2. We observe consistent gains for text–image-fusion RAG systems (completeness = 68.4%) over multimodal joint retrieval systems (64.1%), text-retrieval systems (65.3%), and image-retrieval systems (54.5%). This indicates that retrieving text and images separately using dedicated embeddings, then combining them in the final LLM query, outperforms unified embeddings or single-modality retrieval. Moreover, visual evidence improves answer completeness and enhances faithfulness when paired with textual context, though image-only retrieval cannot fully capture the textual information contained in images. Questions requiring images to answer remain challenging for all systems, suggesting that future RAG improvements should prioritize image-dependent queries. In contrast, performance differences across question types, such as comparison or factual retrieval, are minimal.

We make the following contributions:

Benchmarks	Domain	Evidence	# Queries	# Pages of Doc	RAG Suitable	Unified Evaluation	Multiple Reference	Human Verif
ArxivQA (Li et al., 2024)	single	🖼	100k	-	✗	✗	✗	✗
TAT-DQA (Zhu et al., 2022)	single	🖹 ▦	17k	3k	✗	✗	✗	✓
InfoVQA (Mathew et al., 2022)	single	🖼	6k	-	✗	✗	✗	✓
DocVQA (Mathew et al., 2021)	single	🖼 ▦	11k	-	✗	✗	✗	✓
MMLONG (Ma et al., 2024b)	multiple	🖹 🖼 ▦	1.1k	5k	✗	✗	✓	✓
REALMM (Wasserman et al., 2025)	multiple	🖹 🖼 ▦	5k	8k	✓	✗	✗	✗
ViDoSeek (Wang et al., 2025a)	multiple	🖹 🖼 ▦	1.2k	10k	✓	✗	✗	✗
`UniDoc-Bench` (ours)	multiple	🖹 🖼 ▦	1.6k	70k	✓	✓	✓	✓

RAG Suitable: The dataset provides RAG-style data: queries are self-contained and reflect realistic human questions, with each paired to a grounding corpus (text, images, tables) for retrieval-conditioned answering, supported by a large, highly relevant knowledge base to evaluate retrieval. **Unified Evaluation**: Apples-to-apples comparison across different baseline RAG systems. **Multiple Reference**: Supports multi-hop, multi-modality, multi-source grounding. **Human Verif**: Introduce human experts to review and verify the correctness and quality of all the QA pairs, or to annotate the entire dataset.

Table 1: Comparison of existing document QA datasets with `UniDoc-Bench`.

- We introduce a new multimodal RAG benchmark built from real-world PDF documents, comprising 70k pages across 8 domains, with 1,600 human-verified QA pairs referencing text, figures, and tables, spanning 4 question types.
- We present a high-quality data synthesizing pipeline for creating MM-RAG evaluation datasets, designed to be compatible with any document database.
- We propose a fair and reproducible evaluation framework by fixing candidate pools across modalities and measuring retrieval effectiveness, answer faithfulness, and completeness end-to-end across different RAG systems.
- We compare text retrieval, image retrieval, text–image fusion, and multimodal joint retrieval pipelines, evaluating which strategy performs best across question types, evidence modalities, and document characteristics. We also show `UniDoc-Bench`'s use for evaluating Visual Question Answering (VQA) tasks, highlighting its versatility for MM-RAG research.

2 Related Works

2.1 Multimodal Retrieval-augmented Generation (MM-RAG)

Recent advances in multimodal understanding underscore the importance of MM-RAG for reducing hallucinations. VLM2Vec (Jiang et al., 2024; Meng et al., 2025) shows that instruction-tuning vision-language models improves embeddings for robust text–image alignment. SeBe (Chen et al., 2025) adapts LLaVA-1.5 (Liu et al., 2024) into a retrieval-oriented model that aligns user queries with external knowledge. GME (Zhang et al., 2024b) proposes a unified multimodal embedding capable of text-to-image, image-to-text, and text-to-text retrieval. Uni-Retrieval (Jia et al., 2025) combines VLMs with prompt-tuning to flexibly handle heterogeneous queries and modalities. Routing-based methods like UniversalRAG (Yeo et al., 2025) and UniRAG (Sharifymoghaddam et al., 2025) use adaptive query routing to select the best modality and level of granularity.

2.2 Visual Document Evaluation

Document understanding with interleaved text and visuals has led to specialized vision-based RAG pipelines (Yu et al., 2024b; Wang et al., 2025b,c) that process document screenshots directly. For example, ColPali (Faysse et al., 2024) uses VLMs to jointly encode textual queries and visual documents via MaxSim (Khattab and Zaharia, 2020), while ViDoRAG (Wang et al., 2025b) employs multi-agent reasoning for iterative cross-modal queries. Optimization-focused methods like VRAG (Wang et al., 2025c) use GRPO (Shao et al., 2024), to adapt VLMs for end-to-end document understanding. However, comparisons with text-only baselines are often unfair, as these baselines ignore non-text modalities. Existing evaluations are also limited: MMLongBench-Doc (Ma et al., 2024c) covers long-context multimodal documents but is poorly suited for retrieval; REAL-MM (Wasserman et al., 2025) and VidoSeek (Wang et al., 2025a) lack cross-page and cross-modal evidence; other benchmarks (Mathew et al., 2021, 2022; Zhu et al., 2022; Li et al., 2024) are narrow in scope, covering single images or pages, limited domains, or small scales (Table 1). To fill these gaps, we introduce `UniDoc-Bench`, a benchmark designed for practical MM-RAG use cases with multi-page, cross-modal evidence and scalable evaluation.

3 Dataset Curation

First, a large-scale, high-quality multi-modal database is needed for evaluating RAG systems, where each document contains content-rich figures, tables and corresponding textual information. Documents should be domain-specific and exhibit high inter-document similarity to evaluate effective retrieval. The construction of this database is detailed in Section 3.1. Then, we require high-quality query–answer pairs to evaluate the RAG system. Each query is designed to reflect realistic human intent and is written as a self-contained question. The corresponding ground-truth answer must be retrievable solely from the curated database and supported by evidence across multiple modalities. In Section 3.2, we describe our synthetic QA pipeline, and in Section 3.3, we validate dataset quality through human annotation.

3.1 Source Document Collection

We use PDFA (Montalvo and Wightman, 2024) as our data source, containing diverse formats (e.g., reports, slides, posters) and covering broad domains, but it lacks tags or labels. Therefore, our first step is data filtering to collect a high-quality database. We design a field scheme (Appendix B.1) that captures key metadata, including domain, subdomain, language, modality (e.g., text, tables, figures), image quality (whether the resolution is clear), and text proportion. This allows us to standardize the data and build a high-quality cross-modality database. As shown in Figure 1 (c), we select 8 domains across industries and define subdomains within each, grouping similar documents. To ensure high inter-document similarity, we retain only documents from $3-5$ related subdomains containing multiple modalities, yielding on $\sim 8,000$ pages per domain. The final dataset spans *Legal, Commerce and Manufacturing, Education, Energy, Construction, Finance, Healthcare, and CRM*, with detailed subdomain descriptions in Appendix B.2.

3.2 Question and Answer Synthesis Pipeline

As shown in Figure 2, we introduce a data-synthesis pipeline for building multimodal RAG evaluation datasets with high-quality QA pairs, compatible with various document databases.

3.2.1 Evidence Collection

PDF Parsing. We parse our curated PDF document database[1] by extracting text chunks, tables, and figures, with the latter two stored separately as image files. Within the parsed text chunk, each image and table is replaced with a unique placeholder tag (e.g., «`fig-XXX`» or «`tab-YYY`»), along with its corresponding caption and parsed content

[1] https://unstructured.io/

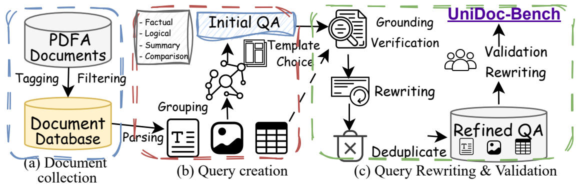


Figure 2: Data Construction pipeline. (a) We filter and tag PDFA documents to curate a high-quality database of 70k pages spanning 8 domains. (b) We parse documents into text, figures, and tables, then synthesize initial QA pairs covering four question types and three modalities using adapted templates. (c) We ground answers in supporting evidence, refine questions for human-intent and self-containment, and verify responses for factuality and completeness, yielding $1,600$ QA pairs. To ensure quality, the entire dataset is validated and rewritten by human annotators.

to fully represent interleaved multimodal content. An example is provided in Appendix B.3.

Chunks Grouping. To support multimodal evidence QA, we construct a knowledge graph ($\mathcal{G}_i$) (ExplodingGradients, 2024; Peng et al., 2024) over the parsed chunks for domain i, where nodes ($N_i = \{n_{i1}, n_{i2}, ...\}$) represent chunks and edges (E_i) denote overlapping entities (e.g., "AI Agent Platform"). Chunks across three modalities (text, tables, figures), from within or across documents, are linked to form ground-truth evidence, which are then used for QA synthesis in the next step.

3.2.2 Question and Answer Generation

Template Choice. First, we ensure the synthesized questions are **diverse** and span multiple categories, since focusing on a single category or using only the same few-shot example questions can introduce bias and limit the comprehensiveness of RAG evaluation. We designed 4 RAG question types: 1) `factual retrieval`, 2) `comparison`, 3) `summarization`, and 4) `logical reasoning`. For each type and database domain, we design 10–15 templates (Appendix B.4). We then sample linked chunks (n_{ij}, e_{ij}, n_{ik}) and prompt the LLM to select 1–3 templates (T_{ij}) that best match the provided chunks and are most likely to produce QA pairs that humans would naturally ask, thereby improving both the diversity and coverage of the questions.

Evidence Grounding. To ensure comprehensive evaluation of MM-RAG, we design 4 *answer types* with distinct evidence requirements, each supported by specialized prompts:

- Text-only: The question can be fully answered using natural language text from the documents.
- Image-only: The question requires information

exclusively from an image, such as numerical values shown only in a figure.
- Image-plus-text: Answering the question requires both text and images, testing the model's ability to reason across modalities.
- Table-required: The question required tabular information to answer, requiring the system to understand table structure and content.

To construct QA pairs, we prompt `GPT-4.1` with parsed text chunks and extracted figures/tables (PNG format), guided by prompts P_n corresponding to the above answer types (see details in Appendix B.5) and templates T_{ij}. We then employ `Gemini-Pro-2.5` — to mitigate single-LLM bias — to verify that the ground-truth answers are correctly grounded in the referenced text, tables, or images, ensuring factual correctness and re-classifying question types when necessary.

Rewriting. To ensure that questions are **self-contained** and reflect realistic **human intent**, we refine the initially synthesized QA pairs. In the first stage, many synthesized questions follow a long-context QA style and may include vague references such as "in this report" or "in Figure 8". To make them suitable for RAG evaluation, we rewrite these questions to ensure they are self-contained and understandable without external context (Appendix B.6). Also, many QA pairs are grounded in images, leading to VQA-style questions (e.g., "How many logos are in Apple Inc.'s 2023 report?"), which do not reflect natural human queries in a RAG context, so we filter and rewrite them to better align with realistic human intent. To ensure comprehensive evaluation, ground-truth answers must be **complete** and **diverse**. In the final step, we revise answers to cover all relevant aspects of their corresponding questions (see Appendix B.7).

3.3 Dataset Quality

We evaluate whether our `UniDoc-Bench` is of sufficient quality to support reliable evaluation of different RAG systems by recruiting 5 human annotators to evaluate the $1,600$ question–response pairs against the provided source documents. The annotation process involved assessing each question-response pair across five dimensions (Appendix C):

- **Factuality**: evaluates whether the claims made in the question (`Factuality-Question`) and the response (`Factuality-Response`) were factually supported by the source documents.
- **Completeness**: assesses whether the response incorporates all necessary information from the retrieved sources to fully answer the question.
- **Grounding**: assesses whether each source chunk (text, image, or table) used to generate the ground-truth response is required to answer the question, by labeling it as either `required` or `not required`, and these labels serve as the ground truth. We then compare the labels produced by our pipeline against the human-annotated ground truth to compute accuracy.
- **Self-Contained**: assesses whether the question was understandable and answerable on its own, without needing external context beyond the provided documents.
- **Human-like Intent**: evaluates whether the question reflected a natural, meaningful query that a human would ask to retrieve information.

	Fact.–Q	**Fact.–R**	**Complete.**
(%)	98.61	94.20	93.63
	Self-Cont.	**Human-like**	**Grounding**
(%)	98.25	96.25	84.38

Table 2: Human evaluation quality on the $1{,}600$ QAs.

As shown in Table 2, the sample shows near-perfect question factuality and self-containment, with strong response factuality and completeness. Human-like intent remains very high (96.25%). Grounding label accuracy is also solid (84.38%). Any questions or responses that do not receive uniformly positive labels are revised by human annotators. These results demonstrate the high quality of `UniDoc-Bench` for evaluating MM-RAG systems, as well as the robustness of our synthesis pipeline, which can be readily used to generate reliable QA pairs for new databases.

Dataset Statistics. `UniDoc-Bench` consists of 200 QA pairs for each domain, in total 1600 **human-verified and revised** QAs. Within each domain, we have an equal distribution of 50 text-only, image-only, text-plus-image, and table-required questions. In total, the dataset contains 800 single-modality and 800 multi-modality questions. On average, each question requires 2.15 evidence items (text chunks, images, or tables) for a complete answer. More details can be found in Figure 1(b).

4 Experiments

To fairly evaluate different RAG systems, we focus on two aspects: retrieval and end-to-end performance. In this section, we first evaluate the retrieval performance of 4 embedding and retrieval models, including text-only, image-only, and two multimodal approaches (§ 4.1). Next, we evaluate the end-to-end response quality of nine RAG systems that differ in their embeddings, retrieval strategies, and underlying LLMs (Section 4.2). Finally, we demonstrate how our dataset can be used for VQA tasks (Section 4.3). Together, these experiments highlight the usefulness of our dataset and provide practical guidance for selecting RAG components and evaluating VQA systems.

4.1 Retrieval Performance

Baselines. We use the curated PDF documents as the knowledge base and the synthesized $1,600$ QA pairs to evaluate the following 4 embedding–retrieval models. For all methods, we retrieve the top-$k = 10$ candidates.

- **Text:** PDFs are parsed[1] into text chunks, each embedded with OpenAI's `text-embedding-3-small`, and retrieved via vector search.
- **Image:** Each PDF page is converted to an image, which is embedded using `ColQwen2.5-v0.2` (Faysse et al., 2024) for image retrieval.
- **MM:** Both text chunks and page-level images are embedded.
  - **MM (GME)**: Text and images are jointly embedded using `GME-Qwen2-VL-7B-Instruct` (Zhang et al., 2024b), enabling multimodal retrieval.
  - **MM (T+I)**: A fusion baseline that selects the top-5 candidates from Text and the top-5 from Image retrieval.

Metrics. We report `Precision@10` and `Recall@10` as the retrieval metrics. Since no re-ranker is applied, recall is more informative than `nDCG` for evaluation. Since we need to evaluate both image and text retrieval, each retrieved text chunk or PDF image-page is mapped back to its original PDF page, and the ground-truth contexts are mapped in the same way. Consequently, a retrieved chunk may span multiple consecutive pages of the source document (e.g., pages 2–3 of document A). A retrieval is considered a true positive if the retrieved text chunk or image-page matches the ground-truth context in both page number and file. This criterion may slightly inflate `Recall@10`, since partial overlaps (e.g., retrieved pages 1–3 vs. ground-truth pages 3–5, with the answer on page 5) are still treated as correct. However, this approach offers the most practical and fair basis for comparing text and image retrieval. Thus, absolute scores should not be overinterpreted; the key is the relative performance differences across methods.

Domain	Text		Image		MM (GME)		T+I	
	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
Com.	.286	.829	.179	.843	.437	.882	**.449**	**.914**
Cons.	.246	.762	.159	.792	**.429**	**.864**	.422	.853
CRM	.271	.783	.175	.830	**.437**	.860	.426	**.863**
Edu	.278	.855	.160	.851	**.432**	.878	.427	**.896**
Energy	.239	.706	.148	.718	.366	.723	**.374**	**.746**
Fin.	.254	.781	.177	.818	**.434**	.891	.427	**.912**
HC	.297	.746	.151	.856	**.455**	**.859**	.455	.851
Legal	.312	.861	.178	.861	**.462**	.883	.458	**.903**
Avg.	.273	.790	.166	.821	**.431**	.855	.430	**.864**
By Question Type								
F.R.	.205	.747	.140	.825	**.226**	.859	.219	**.869**
Comp.	.283	.820	.163	.835	**.313**	.901	.309	**.909**
Summary	.336	.828	.200	.800	.355	.880	**.360**	**.895**
Logical	.365	.820	.201	.813	**.386**	.870	.382	**.882**
By Answer Type								
Text-only	.383	.839	.217	.790	**.406**	.847	.400	**.849**
Img-only	.081	.724	.092	.878	.097	.909	**.097**	**.920**
Text + Img	.336	.847	.190	.824	.350	.888	**.351**	**.908**
Table-req.	.291	.752	.163	.791	**.326**	.851	.316	**.861**

Table 3: Retrieval performance (`Precision@10` / `Recall@10`) of four RAG systems on 1,600 QA pairs across eight domains (top) and broken down by question and answer types (bottom).

Table 3 summarizes the retrieval performance of four RAG embedding–retrieval models across eight domains, four question types, and four answer types We observe that **image-based retrieval achieves consistently higher recall but lower precision than text-based retrieval**, as page-image chunks cover more information than individual text chunks. Combining text and image retrieval (T+I) further improves both recall and precision, effectively leveraging the strengths of both modalities. In contrast, multimodal embeddings (`GME-Qwen2-VL-7B-Instruct`), which encode text and images jointly rather than separately, achieve comparable precision but lower recall, suggesting that current multimodal embeddings still lag behind fusion of unimodal embeddings.

4.2 End-to-End Performance

Baselines. We have following six baselines:

- **Image-only RAG:** Each PDF page is converted to a JPEG and retrieved via image embeddings.
  - **Image-only RAG** (IMG): Uses LlamaIndex with `colqwen2.5-v0.2` (Faysse et al., 2024) for image retrieval. After retrieval, the question and retrieved images are provided to `GPT-4.1` to obtain the final response.
  - **VRAG** (Wang et al., 2025d): a multimodal RAG agent that uses a vision-specific action

space — cropping and scaling — to iteratively extract information from image-formatted PDF pages in a coarse-to-fine manner. The embedding model is `colqwen2.5-v0.2`, and the final LLM is `GPT-4.1`.

- **Text-only RAG (TEXT):** Most multimodal RAG studies (Wang et al., 2025a; Faysse et al., 2024) compare only against text-only baselines. For a fairer comparison, PDF pages are parsed into text chunks, embedded for retrieval, with associated images/tables linked back for final responses. In this baseline, each text chunk is embedded using `text-embedding-3-small` and retrieved. The retrieved text chunks, along with their associated images, are then fed into `GPT-4.1` to generate the final response.
- **MM-RAG:** Both parsed text and image-format page images are embedded and retrieved.
  - **Multimodal Text-Image-Fusion RAG (T+I):** Retrieves text and images separately using `text-embedding-3-small` and `colqwen2.5-v0.2`, then combines them for generation with `GPT-4.1`. We also evaluate multiple state-of-the-art LLMs, including `Gemini-pro-2.5`, `Claude 4.5`, and `GPT-5`.
  - **Multimodal-joint-Retrieval RAG** (MM): Uses `GME-Qwen2-VL-7B-Instruct` (Zhang et al., 2024b) (MM(G)) or `voyage -multimodal-3` (MM(V)) as a multimodal embedding model for both text and images.Unlike T+I, where text and images are embedded and retrieved separately, the text chunks and image-formatted PDF pages are embedded together, retrieved jointly, and then fed into `GPT-4.1` for the final response.

Metrics. For **end-to-end** performance, we use an LLM-based judge to measure faithfulness and completeness. Specifically, we first ask the LLM to extract the facts required to answer each question and then verify whether these facts are grounded in the ground-truth chunks; this is measured as `faithfulness` (↑). Next, we ask the LLM to extract the facts required to answer the question from the ground-truth answer and then check whether each fact appears in the system's response; this is measured as `completeness` (↑).

Table 4 (red background) reports the completeness of responses generated by the six RAG systems. **Text-only RAG** (0.619) **substantially outperforms Image-only RAG systems** (IMG: 0.527, VRAG: 0.536), highlighting the significant performance gap between text-based and image-based retrieval in current RAG architectures. Although image retrieval achieves higher recall at the retrieval stage, this advantage does not translate into better end-to-end performance, since multimodal LLMs (GPT-4.1) are more effective when processing text and image chunks together rather than page-level image PDFs alone. In addition, the low precision of image retrieval makes it harder for the model to identify the correct information. The text-image-fusion RAG (T+I) achieves the best overall performance (0.654) across eight domains, demonstrating that image-based PDF representations can effectively complement text retrieval. Although VRAG leverages cropping and scaling to enhance image-based retrieval (0.536 for VRAG vs.0.527 for IMG), it still lags behind the combined T+I approach, underscoring the advantage of explicitly integrating both modalities. Multimodal joint-retrieval RAG systems (MM (`voyage-multimodal-3`): 0.637; MM (`GME-Qwen2-VL-7B-Instruct`): 0.639) also fall short of the simple combination of the best text and image embeddings. This indicates that current multimodal embedding approaches still have substantial room for improvement, and that explicitly **combining separate text and image embeddings remains the most effective strategy** for leveraging multimodal documents. More notably, **in some domains—CRM, Education and Legal—multimodal joint RAG performs worse than text-only RAG**, indicating that current multimodal models still lag behind strong unimodal baselines in certain domains. These results highlight the importance of establishing fair baselines and the value of `UniDoc-Bench`: multimodal RAG systems should be benchmarked against strong, balanced baselines on diverse and high-quality datasets rather than against overly weak text-only settings.

Table 4 (column T+I) compares different state-of-the-art LLMs used in the Text&Image Retrieval setting. `Claude-4.5-sonnet` achieves the best performance across all domains, question types, and answer types. The table also shows that questions requiring only text are most effectively handled by RAG systems with text-embedding. **Questions requiring tables are also relatively easy for RAG systems**, as tables can be accurately parsed as text, which is a straightforward step before embedding documents for text-based retrieval. In contrast, questions requiring images remain challenging across all embedding types — text, image, or multimodal — highlighting that future **RAG improve-**

Domain	Image-only		Text-only	Multimodal						VQA			GT
	IMG	VRAG	TEXT	MM (V)	MM (G)	T+I							
	GPT-4.1					GPT-4.1	Gemini	Claude	GPT-5	Gemini	Claude	GPT-5	Claude
Com.	.545	.547	.633	.663	.657	**.693**	.707	**.789**	.746	.613/.665	**.670/.805**	.629/.706	**.883**
Cons.	.502	.536	.561	.600	.592	**.607**	.662	**.737**	.648	.566/.610	**.669/.706**	.597/.627	**.776**
CRM	.524	.523	.643	.635	.640	**.647**	.689	**.771**	.696	.612/.679	**.756/.774**	.614/.666	**.848**
Edu	.569	.517	**.692**	.660	.673	.688	.672	**.765**	.637	.612/.636	**.720/.741**	.612/.651	**.845**
Energy	.535	.558	.607	**.675**	.661	.649	.682	**.768**	.721	.584/.710	**.750/.799**	.646/.679	**.830**
Fin.	.500	.529	.584	**.641**	.638	.638	.672	**.788**	.693	.585/.635	**.727/.808**	.631/.670	**.835**
HC	.481	.481	.602	.628	**.651**	.621	.689	**.767**	.665	.580/.673	**.723/.735**	.604/.647	**.849**
Legal	.558	.599	.629	.597	.600	**.689**	.705	**.770**	.714	.636/.654	.647/**.740**	**.671**/.680	**.858**
Avg.	.527	.536	.619	.637	.639	**.654**	.685	**.770**	.690	.599/.658	**.708/.763**	.625/.666	**.840**
By Question Type													
F.R.	.557	.344	.648	.619	.612	**.677**	.687	**.739**	.694	.569/.685	**.709/.756**	.645/.698	**.829**
Comp.	.542	.418	.633	.638	**.646**	.641	.700	**.792**	.683	.516/.660	**.722/.768**	.638/.662	**.825**
Summary	.536	.407	.626	**.652**	.649	.640	.666	**.759**	.689	.530/.638	**.670/.777**	.596/.627	**.867**
Logical	.548	.513	.637	.664	**.681**	.630	.681	**.813**	.706	.514/.602	**.719/.774**	.607/.639	**.864**
By Answer Type													
Text-only	.588	.464	.700	**.777**	.771	.695	.767	**.863**	.773	.582/.680	**.752/.823**	.676/.691	**.923**
Img-only	.486	.336	.616	.465	.462	**.619**	.588	**.644**	.629	.510/.613	**.577/.651**	.546/.611	**.756**
Text+Img	.502	.453	.600	.584	.580	**.617**	.611	**.719**	.617	.441/.587	**.677/.729**	.578/.609	**.813**
Table-req.	.610	.392	.633	.723	**.742**	.683	.773	**.853**	.741	.609/.752	**.812/.851**	.700/.752	**.870**

Table 4: Completeness of systems evaluated on 1,600 QA pairs across 8 domains. Average recall is reported over all domains, with similarity top-k set to 10. Gemini refers to Gemini-2.5-pro. Claude refers to Claude-4.5-sonnet. For VQA, the first value uses the entire document as image input, while the second value uses ground-truth images only. GT is the performance of Claude-4.5-sonnet on the ground-truth text chunks, images and tables.

ments should prioritize image-required questions. We further observe that **multimodal joint RAG achieves stronger performance on text-dominant questions, whereas the T+I RAG is more effective for image-dominant queries.** We also provide detailed case studies in Appendix D.

4.3 Visual Question Answering Performance

UniDoc-Bench can also be used to evaluate Visual Question Answering (VQA) tasks. Table 4 (gray background) reports the performance of state-of-the-art LLMs — Gemini-pro-2.5, Claude-4.5-Sonnet, and GPT-5 — when applied to entire image-format PDFs and to ground-truth pages only. The results show that Claude-4.5-Sonnet consistently achieves the highest completeness scores across all domains and question types in the VQA setting. All models exhibit a performance gap between the two settings, confirming that reasoning over entire documents is more challenging than over isolated ground-truth images. Gemini-pro-2.5 is the most sensitive to this noise. In contrast, Claude-4.5-Sonnet and GPT-5 are more robust to full-document inputs, showing smaller performance drops.

Additional Findings. We show the best performance of Claude-4.5-sonnet on the ground-truth chunks in "GT" column of Table 4. Cost and latency are reported in Appendix E.1. Case studies on the impact of content-rich images are presented in Appendix F.1. Analyses of how question type affects difficulty are provided in Appendix D.1, D.2, and F.2. Finally, Appendix F shows that the number of pages and document formats do not significantly affect MM-RAG performance.

5 Conclusion

In this paper, we introduced UniDoc-Bench, a large-scale benchmark for document-centric multimodal RAG, built from 70k real-world PDF pages across 8 domains with $1,600$ human-verified QA pairs. Our experiments establish a clear performance hierarchy, showing that **text-image fusion RAG performs the best**, consistently outperforming both joint multimodal (MM) RAG and single-modality RAG systems. This key finding demonstrates that fusing separate, strong retrievers for text and images is currently a more effective strategy than relying on a single joint multimodal embedding or a single modality alone. Our analysis further pinpoints image-dependent queries as the primary challenge for all systems. By providing a standardized platform for fair comparison, UniDoc-Bench serves as a crucial resource to guide the development of more robust and faithful document intelligence systems.

Limitations

UniDoc-Bench relies on LLM-synthesized, template-based queries which, despite human verification, may lack the linguistic diversity and conversational dependency (e.g., multi-turn follow-ups) characteristic of organic user interactions. The evaluation protocol relies on assumptions, such as treating page-level retrieval matches as correct—potentially inflating recall for dense documents—and explicitly excluding uncaptioned figures under the assumption they are non-informative. Furthermore, the benchmark is currently limited to English-centric documents across eight specific domains and employs LLM-based judges for end-to-end metrics, suggesting that findings may not generalize to low-resource languages or remain robust against inherent judge model biases.

References

Boqi Chen, Anuj Khare, Gaurav Kumar, Arjun Akula, and Pradyumna Narayana. 2025. Seeing beyond: Enhancing visual question answering with multi-modal retrieval. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 410–421, Abu Dhabi, UAE. Association for Computational Linguistics.

ExplodingGradients. 2024. Ragas: Supercharge your llm application evaluations. https://github.com/explodinggradients/ragas.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 6491–6501.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).

Yanhao Jia, Xinyi Wu, Hao Li, Qinglin Zhang, Yuxiao Hu, Shuai Zhao, and Wenqi Fan. 2025. Uni-retrieval: A multi-style retrieval framework for stem's education. *arXiv preprint arXiv:2502.05863*.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. 2024. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Chenxia Li, Weiwei Liu, Ruoyu Guo, Xiaoting Yin, Kaitao Jiang, Yongkun Du, Yuning Du, Lingfeng Zhu, Baohua Lai, Xiaoguang Hu, and 1 others. 2022. Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system. *arXiv preprint arXiv:2206.03001*.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint arXiv:2403.00231*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. 2024a. Unifying multimodal retrieval via document screenshot embedding. *arXiv preprint arXiv:2406.11251*.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, and 1 others. 2024b. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *Advances in Neural Information Processing Systems*, 37:95963–96010.

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

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

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.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, and 1 others. 2025. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *arXiv preprint arXiv:2507.04590*.

Pablo Montalvo and Ross Wightman. 2024. pixparse/pdfa-eng-wds [dataset]. Hugging Face Datasets. Accessed August 2025.

Xiangyu Peng, Prafulla Kumar Choubey, Caiming Xiong, and Chien-Sheng Wu. 2024. Unanswerability evaluation for retrieval augmented generation. *arXiv preprint arXiv:2412.12300*.

Jake Poznanski, Aman Rangapur, Jon Borchardt, Jason Dunkelberger, Regan Huff, Daniel Lin, Christopher Wilhelm, Kyle Lo, and Luca Soldaini. 2025. olmocr: Unlocking trillions of tokens in pdfs with vision language models. *arXiv preprint arXiv:2502.18443*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Sahel Sharifymoghaddam, Shivani Upadhyay, Wenhu Chen, and Jimmy Lin. 2025. Unirag: Universal retrieval augmentation for large vision language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2026–2039.

Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, and 1 others. 2025. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*.

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

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

Qiuchen Wang, Ruixue Ding, Yu Zeng, Zehui Chen, Lin Chen, Shihang Wang, Pengjun Xie, Fei Huang, and Feng Zhao. 2025c. Vrag-rl: Empower vision-perception-based rag for visually rich information understanding via iterative reasoning with reinforcement learning. *arXiv preprint arXiv:2505.22019*.

Qiuchen Wang, Ruixue Ding, Yu Zeng, Zehui Chen, Lin Chen, Shihang Wang, Pengjun Xie, Fei Huang, and Feng Zhao. 2025d. Vrag-rl: Empower vision-perception-based rag for visually rich information understanding via iterative reasoning with reinforcement learning. *Preprint*, arXiv:2505.22019.

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*.

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, and 1 others. 2024. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*.

Woongyeong Yeo, Kangsan Kim, Soyeong Jeong, Jinheon Baek, and Sung Ju Hwang. 2025. Universalrag: Retrieval-augmented generation over multiple corpora with diverse modalities and granularities. *arXiv preprint arXiv:2504.20734*.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and 1 others. 2024a. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and 1 others. 2024b. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*.

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

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

Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. 2022. Towards complex document understanding by discrete reasoning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4857–4866.

A The Use of Large Language Models (LLMs)

We used LLMs for three purposes: (i) polishing grammar and improving readability, and (ii) assisting in the evaluation of RAG outputs (iii) synthesizing the QA pairs. All research ideas and analyses were conducted by the authors, who take full responsibility for the content.

B Dataset Creation Details

B.1 Document fields

We classify each PDF document into the following fields:

- *domain*: one or more from {Healthcare, Finance, Technology and Software, Commerce and Manufacturing, Marketing, Arts and Entertainment, Government, Legal, Education, Scientific Research and Development, Customer Relationship Management (CRM). others}
- *subdomain*: optional finer-grained categories
- *date*: year or estimated year (e.g., 2005)
- *language*: language of the document (e.g., `en`)
- *modality*: possible values include {text, table, figure, formula, image, drawing}
- *quality*: parsing confidence, values {easy-parse, hard-parse}
- *format*: one or more from {form, report, notice, paper, slide, poster, book, newspaper, article, textbook, note, webpage, document, record}
- *text_proportion*: percentage of textual content (e.g., 25%)

As described in Section 3.1, we do not include every domain or subdomain in our benchmark. Instead, we filter the source data and retain eight highly representative domains.

B.2 Domain Definitions

We classify documents into domains and subdomains, each with a brief description for clarity. These labels are used for tagging. As detailed in Section 3.1, we filter the source data and retain eight highly representative domains rather than including all possible ones.

Domain	Subdomain	Description
Healthcare	Clinical & Patient Care	Direct provider-patient interaction: diagnosis, treatment, and care management.
Healthcare	Pharmaceuticals & Biotechnology	Development and regulation of drugs, vaccines, and biotechnological products (no patient records).
Healthcare	Medical Devices & Diagnostics	Design, production, and regulation of medical equipment and diagnostic tools (no patient records).
Healthcare	Clinical Research & Trials	Controlled studies testing treatments, drugs, or therapies.
Healthcare	Public Health & Policy	Population-level promotion, disease prevention, accessibility (not individual records).
Healthcare	Other Healthcare Topics	Healthcare economics, law, and alternative medicine.
Finance	Investments & Wealth Management	Stock portfolios, retirement planning, mutual funds, hedge funds.
Finance	Insurance & Risk Management	Health, life, auto, property insurance; actuarial analysis.
Finance	Corporate Finance & Treasury	Budgeting, fundraising, M&A, investor relations, corporate structure.
Finance	Personal Finance & FinTech	Budgeting apps, personal loans, P2P lending, digital wallets.
Finance	Real Estate Finance	Mortgages, REITs, valuations, market dynamics.
Finance	Macroeconomics & Financial Markets	Markets, currency, fiscal/monetary policy, global economics.
Finance	Other Finance Topics	Microfinance, Islamic banking, niche financial products.
Technology & Software	Software Engineering & DevOps	Coding, testing, deployment, CI/CD, APIs.
Technology & Software	Cybersecurity & Information Security	Risk management, encryption, compliance, network defense.
Technology & Software	Data Science, AI & Analytics	ML, pipelines, visualization, BI tools.
Technology & Software	HCI & UX	Design, prototyping, accessibility, usability studies.
Technology & Software	Emerging Technologies	AR/VR, quantum computing, IoT, blockchain.
Technology & Software	Other Tech Topics	Legacy systems, databases, systems architecture.
Commerce & Manufacturing	Supply Chain & Logistics	Procurement, warehousing, transportation, inventory.

Domain	**Subdomain**	**Description**
Commerce & Manufacturing	Industrial Engineering & Production	Process optimization, quality control, Lean/Six Sigma.
Commerce & Manufacturing	Retail & E-Commerce	Marketplaces, POS systems, consumer engagement.
Commerce & Manufacturing	Trade Policy & Global Commerce	Tariffs, export-import regulation, global trade.
Commerce & Manufacturing	Other Commerce Topics	Business operations, sales, distribution.
Marketing	Digital Marketing & Advertising	Social media, SEO/SEM, online campaigns.
Marketing	Consumer Behavior & Market Research	Surveys, focus groups, data-driven insights.
Marketing	Branding & Corporate Identity	Logo, image, brand value, messaging.
Marketing	Marketing Analytics & Metrics	ROI, attribution models, dashboards.
Marketing	Other Marketing Topics	Public relations, sponsorships, offline campaigns.
Arts & Entertainment	Performing Arts	Music, theater, dance, performance reviews.
Arts & Entertainment	Visual Arts & Design	Painting, sculpture, illustration, graphic design.
Arts & Entertainment	Film, TV & Media Studies	Criticism, production, audience reception.
Arts & Entertainment	Literature & Writing	Fiction, non-fiction, literary analysis.
Arts & Entertainment	Games & Interactive Media	Video games, role-playing, esports.
Arts & Entertainment	Other Arts Topics	Fashion, photography, cultural heritage.
Government	Public Administration & Policy	Bureaucracy, policymaking, implementation.
Government	Law Enforcement & Security	Policing, intelligence, defense, military studies.
Government	International Relations & Diplomacy	Foreign policy, treaties, global governance.
Government	Elections & Governance	Voting, political systems, representation.
Government	Other Government Topics	Civil rights, immigration, taxation.
Legal	Corporate & Business Law	Contracts, mergers, compliance.

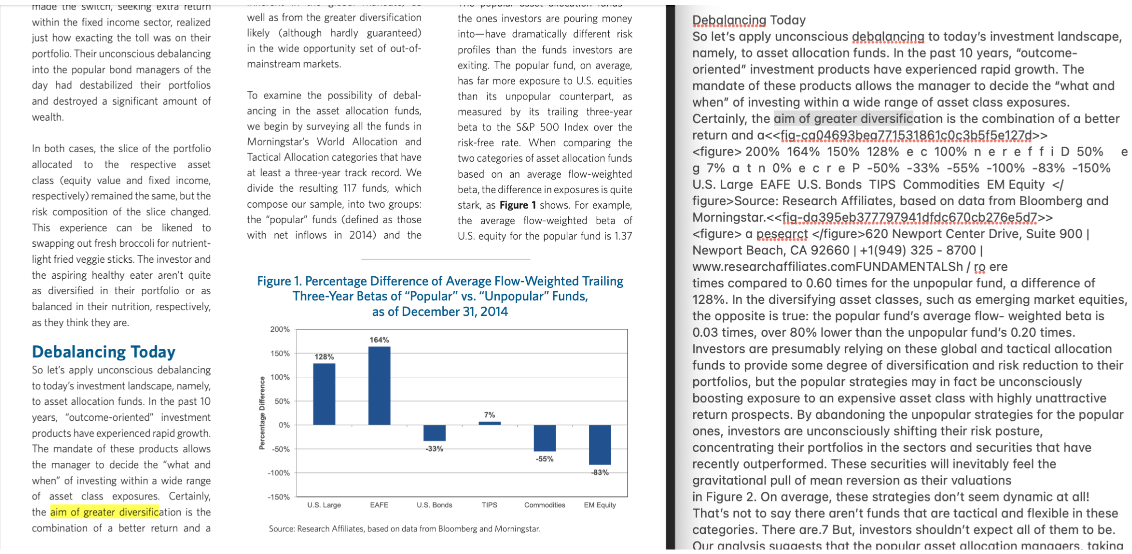
Domain	Subdomain	Description
Legal	Criminal & Civil Law	Courts, trials, disputes, legal rights.
Legal	Intellectual Property Law	Copyrights, patents, trademarks.
Legal	International & Comparative Law	Cross-border legal systems, treaties.
Legal	Legal Theory & Jurisprudence	Philosophy of law, frameworks.
Legal	Other Legal Topics	Niche legal issues, regulatory law.
Education	K-12 Education	Curriculum, pedagogy, assessments.
Education	Higher Education & Academia	Universities, research, accreditation.
Education	Online & Distance Learning	MOOCs, e-learning, virtual platforms.
Education	Education Policy & Reform	Accessibility, standards, funding.
Education	Other Education Topics	Lifelong learning, teacher training.
Scientific R&D	Natural Sciences	Physics, chemistry, biology, earth science.
Scientific R&D	Engineering & Applied Sciences	Electrical, mechanical, civil, aerospace.
Scientific R&D	Medical & Life Sciences	Biomedical, genetics, ecology.
Scientific R&D	Computer Science & Computational Fields	Algorithms, theory, AI, networks.
Scientific R&D	Other Science Topics	Interdisciplinary, niche fields.
CRM	Customer Support & Helpdesk	Call centers, chatbots, support tickets.
CRM	Sales & Lead Management	CRM tools, customer tracking, pipelines.
CRM	Customer Analytics & Insights	Segmentation, lifetime value, churn analysis.
CRM	Customer Experience (CX) & Engagement	Feedback, personalization, loyalty programs.
CRM	Other CRM Topics	Partnerships, integrations, omni-channel strategies.

B.3 Parsing Examples

We use `unstructured` to parse each PDF into three components: text chunks, images of figures, and images of tables. Since many figures (e.g., signatures or logos) are not informative, we only retain figures that include captions. Figure 3 shows an example of the parsing output, where figures are represented by placeholders such as «`fig-XXX`» and the parsed text from the figures.

B.4 Dataset Templates

This is the templates for the domain: `finance`. We create different templates for different domains, which can be found in our code files in the supplementary materials.



made the switch, seeking extra return within the fixed income sector, realized just how exacting the toll was on their portfolio. Their unconscious debalancing into the popular bond managers of the day had destabilized their portfolios and destroyed a significant amount of wealth.

In both cases, the slice of the portfolio allocated to the respective asset class (equity value and fixed income, respectively) remained the same, but the risk composition of the slice changed. This experience can be likened to swapping out fresh broccoli for nutrient-light fried veggie sticks. The investor and the aspiring healthy eater aren't quite as diversified in their portfolio or as balanced in their nutrition, respectively, as they think they are.

Debalancing Today

So let's apply unconscious debalancing to today's investment landscape, namely, to asset allocation funds. In the past 10 years, "outcome-oriented" investment products have experienced rapid growth. The mandate of these products allows the manager to decide the "what and when" of investing within a wide range of asset class exposures. Certainly, the aim of greater diversification is the combination of a better return and a

well as from the greater diversification likely (although hardly guaranteed) in the wide opportunity set of out-of-mainstream markets.

To examine the possibility of debal-ancing in the asset allocation funds, we begin by surveying all the funds in Morningstar's World Allocation and Tactical Allocation categories that have at least a three-year track record. We divide the resulting 117 funds, which compose our sample, into two groups: the "popular" funds (defined as those with net inflows in 2014) and the

the ones investors are pouring money into—have dramatically different risk profiles than the funds investors are exiting. The popular fund, on average, has far more exposure to U.S. equities than its unpopular counterpart, as measured by its trailing three-year beta to the S&P 500 Index over the risk-free rate. When comparing the two categories of asset allocation funds based on an average flow-weighted beta, the difference in exposures is quite stark, as **Figure 1** shows. For example, the average flow-weighted beta of U.S. equity for the popular fund is 1.37

Figure 1. Percentage Difference of Average Flow-Weighted Trailing Three-Year Betas of "Popular" vs. "Unpopular" Funds, as of December 31, 2014

Source: Research Affiliates, based on data from Bloomberg and Morningstar.

Debalancing Today
So let's apply unconscious debalancing to today's investment landscape, namely, to asset allocation funds. In the past 10 years, "outcome-oriented" investment products have experienced rapid growth. The mandate of these products allows the manager to decide the "what and when" of investing within a wide range of asset class exposures. Certainly, the aim of greater diversification is the combination of a better return and a<<fig-ca04693bea771531861c0c3b5f5e127d>> <figure> 200% 164% 150% 128% e c 100% n e r e f f i D 50% e g 7% a t n 0% e c r e P -50% -33% -55% -100% -83% -150% U.S. Large EAFE U.S. Bonds TIPS Commodities EM Equity </figure>Source: Research Affiliates, based on data from Bloomberg and Morningstar.<<fig-da395eb377797941dfdc670cb276e5d7>> <figure> a pesearct </figure>620 Newport Center Drive, Suite 900 | Newport Beach, CA 92660 | +1(949) 325 - 8700 | www.researchaffiliates.comFUNDAMENTALSh / ro ere
times compared to 0.60 times for the unpopular fund, a difference of 128%. In the diversifying asset classes, such as emerging market equities, the opposite is true: the popular fund's average flow- weighted beta is 0.03 times, over 80% lower than the unpopular fund's 0.20 times. Investors are presumably relying on these global and tactical allocation funds to provide some degree of diversification and risk reduction to their portfolios, but the popular strategies may in fact be unconsciously boosting exposure to an expensive asset class with highly unattractive return prospects. By abandoning the unpopular strategies for the popular ones, investors are unconsciously shifting their risk posture, concentrating their portfolios in the sectors and securities that have recently outperformed. These securities will inevitably feel the gravitational pull of mean reversion as their valuations
in Figure 2. On average, these strategies don't seem dynamic at all! That's not to say there aren't funds that are tactical and flexible in these categories. There are.7 But, investors shouldn't expect all of them to be.

Figure 3: Example of PDF parsing with figure placeholders («fig-XXX»).

Factual Retrieval

Template	Example
What indicators, policies, or tools are described in the discussion of [Economic Topic/-Financial Strategy]?	What inflation indicators are cited in the ECB's policy blog from June?
Which markets, sectors, or instruments are emphasized in relation to [Trend/Event/Goal]?	Which sectors are favored in the 2025 sustainable investing outlook?
What key positions or exposures are taken by [Investor/Desk/Division] in response to [Condition/Event]?	What position changes did the multi-asset team make in response to rising real yields?
What assumptions, constraints, or parameters are specified in [Scenario/Strategy/Model]?	What assumptions are used in the stress testing scenario for oil price shocks?
When was [Policy/Event/Adjustment] implemented, and what immediate actions followed?	When did the Bank of Japan change its yield curve control stance?
Who oversees or initiates [Financial Decision/Policy/Investment Move] in the described context?	Who approves short-term borrowing requests in the global treasury function?
How is [Strategy/Instrument/Term] defined or operationalized in this context?	How is "duration-neutral tilt" defined in the Q3 fixed income note?
How do you carry out or execute [Action/-Transaction/Plan] in [Financial Context]?	How do you implement a covered call overlay in an income-focused portfolio?
What are the procedural steps or controls listed for [Financial Task/Compliance/Change]?	What steps are required to evaluate bond ladder rollovers in rising rates?

Comparison

Template	**Example**
How do [Strategies/Regions/Instruments] compare in terms of [Risk/Performance/Conditions]?	How do TIPS and gold compare for inflation protection in the current macro setup?
Which asset class, sector, or product is better suited for [Objective/Environment]?	Which is better for income stability in retirement: dividend ETFs or bond ladders?
What are the structural or tactical differences between [Financial Approaches]?	What are the key differences between liability-driven investment and balanced allocation strategies?
How did [Metric/Position/Exposure] change between [Period A] and [Period B]?	How did corporate cash allocation to floating-rate debt shift over 2023?
How do regulatory or monetary responses differ between [Jurisdictions]?	How does Fed liquidity provision compare to ECB emergency facilities post-crisis?

Summarization

Template	Example
What are the key findings or takeaways from [Brief/Update/Policy/Strategy]?	What are the key points in the tactical asset allocation update from July?
Summarize the main market movements, themes, or risks discussed in [Note/Newsletter/Memo].	Summarize the interest rate risk themes highlighted in the October bond outlook.
What portfolio, liquidity, or policy adjustments are recommended or implemented?	What rebalancing steps were taken in the client model portfolios in Q1?
List the major economic risks or opportunities discussed in [Period/Event/Note].	What macro risks are cited ahead of the U.S. election cycle?
What are the key operational or structural features of [Product/Plan/Tool]?	What are the structural features of the new drawdown facility described in the treasury toolkit?

Causal / Reasoning / Why Questions

Template	Example
Why did [Entity/Desk/Advisor] make [Move/Shift/Decision] in response to [Condition/Event]?	Why did the balanced portfolio reduce international equity in Q2?
How did [Macro Event/Regulatory Shift] influence [Positioning/Allocation/Operations]?	How did the Basel III revisions alter corporate liquidity buffers?
What drove the shift from [Approach A] to [Approach B] in [Context]?	What drove the shift from risk-parity to volatility-targeting in multi-asset allocation?
Why was [Instrument/Policy/Vehicle] introduced or phased out?	Why was the internal netting structure retired in the 2024 treasury overhaul?
What sequence of factors or events led to [Market Reaction/Portfolio Impact/Policy Result]?	What sequence of events led to capital outflows from EM debt in late 2023?

B.5 QA Synthesizing Prompts

B.5.1 Text-only

Prompt P.1: Text-only RAG Question Generation

Prompt: You are an assistant specialized in creating Multimodal RAG tasks. The task is the following: Given some natural language contexts and images inside these contexts, you will generate questions that can be asked by a user to retrieve information from a large documentary corpus.

Requirements:

- The 2-hop synthesized question must be a single, self-contained question and must not use "and" to connect multiple questions.
- The answer of the synthesized question will only be found in the contexts.
- The answer of the synthesized question cannot be found in the images.
- The synthesized question must require all the chunks in the contexts to be answered.
- The synthesized question must be specific enough to locate the contexts in a large documentary corpus.
- You must also provide an explanation why the answer can only be found in the provided contexts.

Question Template:

- Use the following template to generate the QA:

```
{{TEMPLATES}}
```

Output Format:

```
{
    "questions": [
        {
            "question": "<synthesized-question>",
            "answer": "<answer-of-the-question>",
            "question_type":
            <choose from "factual_retrieval", "comparison",
                "summarization", "causal_reasoning">,
            "explanation-chunks": "<explanation-chunks>",
            "sentences-chunks-used": {"Chunk1": "sentences-chunk1",
                "Chunk2": "sentences-chunk2", ...}
        }
    ]
}
```

Input Data:

- Contexts: “{{contexts}}”
- Images: The image is as follows:

Notes:

- If the image can only be used for visualization or illustration, return an empty list for ‘sentences-chunks-used’.
- If you cannot use all the chunks in the answer, return an empty list for ‘sentences-chunks-used’.

B.5.2 Image-only

Prompt P.2: Image-only RAG Question Generation

Prompt: You are an assistant specialized in creating Multimodal RAG tasks. The task is the following: Given some natural language contexts and images inside these contexts, you will generate questions that can be asked by a user to retrieve information from a large documentary corpus.

Requirements:

1. The synthesized question must be a single, self-contained question and must not use “and" to connect multiple questions.
2. The answer of the synthesized question will only be found in the image and cannot be found in any sentences in the chunks of the provided contexts.
3. The synthesized question must require chunks/contexts to locate the image and cannot mention the image directly.
4. The synthesized question must be specific enough to locate the contexts in a large documentary corpus.
5. Do not ask “what XYZ in the graph/image/figure"; the question must be general enough to be asked in a large corpus.
6. If you cannot synthesize a question which can only be answered in the image based on the above requirements, do not synthesize anything.
7. Provide an explanation why the answer can only be found in the image and cannot be found in the provided chunks/contexts.
8. Avoid phrasing like “what is shown in the image,” e.g., "what color/logo/name in the image."
9. Emphasize reasoning, aggregation, temporal comparison, or retrieval from source data. Imagine the question being asked without the image still making partial sense.

Question Template:

- Use the following template to generate the QA:

```
{{TEMPLATES}}
```

Output Format:

```
{
    "questions": [
        {
            "question": "<synthesized-question>",
            "answer": "<answer-of-the-question>",
            "question_type":
            <choose from "factual_retrieval", "comparison",
                "summarization", "causal_reasoning">,
            "image": "<<fig-aaaaa>>",
            "explanation-image": "<explanation-image>",
            "explanation-chunks": "<explanation-chunks>",
            "sentences-chunks-used":
            {"Chunk1": "sentences-chunk1",
                "Chunk2": "sentences-chunk2", ...}
        }
    ]
}
```

Input Data:

- Contexts: “{{contexts}}”
- Images: The image is as follows:

Notes:

- If the image can only be used for visualization or illustration, return an empty list for ‘sentences-chunks-used’.
- If you cannot use all the chunks in the answer, return an empty list for ‘sentences-chunks-used’.

B.5.3 Text-plus-Image

Prompt P.3: Text-plus-image RAG Question Generation

Prompt: You are an assistant specialized in creating Multimodal RAG tasks. The task is the following: Given some natural language contexts and images inside these contexts, you will generate questions that can be asked by a user to retrieve information from a large documentary corpus.

Requirements:

1. The 2-hop synthesized question must require both the provided contexts and images to answer.
2. The concise answer of the synthesized question will directly require information in the image to answer.
3. The concise answer of the synthesized question will also require information in the natural language contexts to answer.
4. The synthesized question must require contexts to locate the image and cannot mention the image directly.
5. The synthesized question must be specific enough to locate the contexts in a large documentary corpus.
6. Provide an explanation indicating which part of the image is used to answer and which sentence in the contexts is used to answer the question.
7. Do not ask “what XYZ in the graph"; the question must be general enough to be asked in a large corpus.
8. If you cannot synthesize a question based on these requirements or directly use the information in the images, do not synthesize anything.
9. If the image can only be used for visualization or illustration, do not synthesize anything. If you cannot use all the chunks in the answer, do not synthesize the question.
10. The synthesized question must be a single, self-contained question and must not use “and" to connect multiple questions.

Question Template:

- Use the following template to generate the QA:

```
{{TEMPLATES}}
```

Output Format:

```
{
    "questions": [
        {
            "question": "<synthesized-question>",
            "answer": "<answer-of-the-question>",
            "question_type": <choose from "factual_retrieval",
                "comparison", "summarization", "causal_reasoning">,
            "image": "<<fig-aaaaa>>",
            "explanation-image": "<explanation-image>",
            "explanation-chunks": "<explanation-chunks>",
            "sentences-chunks-used":
            {"Chunk1": "sentences-chunk1",
                "Chunk2": "sentences-chunk2", ...}
        },...
    ]
}
```

Input Data:

- Contexts: “{{contexts}}”
- Images: The image is as follows:

Notes:

- If the image can only be used for visualization or illustration, return an empty list for ‘sentences-chunks-used’.
- If you cannot use all the chunks in the answer, return an empty list for ‘sentences-chunks-used’.

B.5.4 Table-required

Prompt P.4: Table-required RAG Question Generation

Prompt: You are an assistant specialized in creating Multimodal RAG tasks. The task is the following: Given some natural language contexts containing tables, you will generate questions that can be asked by a user to retrieve information from a large documentary corpus.

Requirements:

1. The synthesized question must be a single, self-contained question and must not use “and" to connect multiple questions.
2. The answer of the synthesized question will only be found in the table (within ⟨table⟩ and ⟨/table⟩) and cannot be found in any sentences outside the ⟨table⟩ and ⟨/table⟩ in the chunks of the provided contexts.
3. The synthesized question must require chunks/contexts to locate the table and cannot mention the ‘table’ directly.
4. The synthesized question must be specific enough to locate the contexts in a large documentary corpus.
5. Do not ask “what XYZ in the table"; the question must be general enough to be asked in a large corpus.
6. If you cannot synthesize a question which can only be answered in the table based on the above requirements, do not synthesize anything.
7. Provide an explanation why the answer can only be found in the table and cannot be found in other parts of the chunks/contexts.
8. Emphasize reasoning, aggregation, temporal comparison, or retrieval from source data. Imagine the question being asked without the table still making partial sense.

Question Template:

- Use the following template to generate the QA:

```
{{TEMPLATES}}
```

Output Format:

```
{
    "questions": [
        {
            "question": "<synthesized-question>",
            "answer": "<answer-of-the-question>",
            "question_type": <choose from "factual_retrieval", "comparison",
                "summarization", "causal_reasoning">,
            "image": "<<tab-aaaaa>>",
            "explanation-table": "<explanation-table>",
            "explanation-chunks": "<explanation-chunks>",
            "sentences-chunks-used":
            {"Chunk1": "sentences-chunk1",
                "Chunk2": "sentences-chunk2", ...}
        },...
    ]
}
```

Input Data:

- Contexts: “{{contexts}}”
- Table: The table is included as ‘⟨table⟩... ⟨/table⟩’ in the context.

Notes:

- If the table can be used only for visualization or illustration, return an empty list for ‘sentences-chunks-used’.
- If you cannot use all the chunks in the answer, return an empty list for ‘sentences-chunks-used’.

B.6 Rewriting prompts

Prompt P.5: Question Rewriting

Prompt: You are tasked with rewriting the following question in two different ways, using only the provided Contexts and without hallucinating any information.
Date {{current_date}}
Tasks:

1. **Specific Rewrite**: Add or substitute minimal keywords to tie the question to the Contexts, making retrieval unique while preserving meaning.
2. **Obscured Rewrite**: Paraphrase the specific version to reduce keyword overlap while keeping all needed details intact.

Requirements:

- No hallucinated facts.
- Do not remove critical content.
- Avoid source-referencing phrases (“in figure”, “in table”, etc.).
- Rewrites must be standalone, fluent, faithful to Contexts.
- Only add essential keywords (avoid over-specification).

Check if the original answer remains fully correct for both rewrites. If not, set `"answer_wrong" = "True"`, else `"False"`.
Output Format:

```
{
  "specific_question":
  "More specific version with essential keywords.",
  "obscured_question":
  "Paraphrased version with reduced keyword overlap.",
  "answer_wrong": "True/False"
}
```

Example 1: Original: “What is the revenue growth shown in Figure 3 in 2024’s report?"

```
{
  "specific_question":
  "What is the revenue growth for Company XYZ in 2024?",
  "obscured_question":
  "How did XYZ's financial outcomes change in 2024?",
  "answer_wrong": "False"
}
```

Example 2: Original: “What is the median differential rate between hurdle rates and costs of capital for cyclical and non-cyclical firms?"

```
{
  "specific_question":
  "What is the median differential between hurdle
  rates and costs of capital for cyclical vs. non-cyclical firms in
  the S&P 500 according to the Corporate Finance Advisory?",
  "obscured_question":
  "Within the Corporate Finance Advisory, what is the
  median gap between
  required returns and capital costs for S&P 500 firms
  sensitive to the economy vs. stable sectors?",
  "answer_wrong": "False"
}
```

B.7 Answer Rewriting Prompts

Prompt P.6: Answer Rewriting

Prompt: You are tasked with rewriting the following answer so that it contains all the facts for answering the question, given the contexts and the image.
Instruction:

- Do not hallucinate any additional information. Use only the provided contexts and images.
- The rewritten answer must include the **old correct answer**, if it is correct.
- If the answer is already complete, you may leave it unchanged.
- Make the answer as concise as possible.
- If the **old correct answer** is incomplete, expand it so that the `"complete_answer"` fully addresses the question.

Output Format:

```
{
  "complete_answer": "Final rewritten answer that is concise,
  faithful to contexts and images, and fully answers the question."
}
```

Input Data:

- Question: "{{rewritten_question_obscured}}"
- Contexts: "{{contexts}}"
- Old Correct Answer: "{{answer}}"
- Images: The image is as follows:

C Human Annotation

Annotators were provided with the following instructions to evaluate the quality of synthesized questions and responses against source documents.

C.1 Task Overview

The primary task is to read a synthesized question and response, then evaluate their quality based on the provided PDF pages and images. The core evaluation criterion is factuality.

C.2 Factuality Evaluation

Annotators must determine whether the question and response are factually supported by the source material.

C.2.1 Procedure

Annotators were instructed to follow these steps:

1. Open the folder corresponding to the given ID.

2. Read the text from the PDF pages located in the `chunk_X` subfolder. Annotators were told to read all text, including tables and image captions, but to ignore the content of the images themselves.

3. Review the images in the `img_X` subfolder to understand which image is being referenced, then locate that image within the source PDF to read its context and caption.

4. Read the provided Question and Response pair.

5. Assign a factuality label to both the question and the response.

C.2.2 Label Definitions

Factuality-Question: Factual All facts and claims in the question are directly supported by the source material. There are no hallucinations or unsupported statements.

Factuality-Question: Not Factual One or more facts or claims in the question are not supported by the source (i.e., contain hallucinated or fabricated content).

Factuality-Response: Factual All facts and claims in the response are directly supported by the source material. There are no hallucinations or unsupported statements.

Factuality-Response: Not Factual One or more facts or claims in the response are not supported by the source (i.e., contain hallucinated or fabricated content).

Note: The original instructions included a rule stating, "If a question or response is not factual, it should be labeled as 'Incomplete'." However, the provided examples use the "Not Factual" label, which was the standard followed during annotation.

C.2.3 Examples

The following examples were provided to the annotators for guidance.

```
{
  "id": 0,
  "question": "What is the logo of a
    major telecommunications company
    mentioned in the context related to
    personalization strategies?",
  "response": "AT&T",
}

# Steps:
# 1. I open folder "0", read all the
    chunks and images.
# 2. The question seems factual from one
    of the chunk.
# 3. The response seems to NOT be the
    correct answer.

# Then, I label Factual-Question as `
    Factual`
# Then, I label Factual-Response as `Not
     Factual`
```

Listing 1: Example of a factual question with a non-factual response.

```
{
  "id": 4,
  "question": "What businesses are
    located near the proposed
    development area in the Project
    Catalyst?",
  "response": "AT&T",
}

# Steps:
# 1. I open folder "4", read all the
    chunks and images.
# 2. The question seems to be NOT
    factual because I did not see
    Project Catalyst in the pdf or
    images.
# 3. The response seems to be incorrect
    because the question is not factual.

# Then, I label Factual-Question as `Not
     Factual`
# Then, I label Factual-Response as `Not
     Factual`
```

Listing 2: Example of a non-factual question and response.

C.3 Completeness Evaluation

This task assesses whether the response provides all the necessary information to fully answer the question, based on the provided source material.

C.3.1 Procedure

The procedure for evaluating completeness is identical to the factuality task: annotators must review all provided PDF chunks and images before making a judgment.

C.3.2 Label Definitions

Complete: The response includes all the required facts and details present in the source material needed to comprehensively answer the question.

Incomplete: The response omits one or more facts or claims that are present in the source and are necessary to fully answer the question.

Example 1: Incomplete Response

```
{
  "id": 2,
  "question": "What businesses are
    located near the proposed
    development area in the Project
    Catalyst?",
  "response": "AutoZone Auto Parts,
    Pizza Hut, Sonic Drive In, Joe's
    Pizza Italian",
}

# Steps:
# 1. I open folder "2", read all the
    chunks and images.
# 2. The response seems to miss: "Mr Jim
    's Pizza, Justin Spirits, Allsup's
    Convenience Store."

# Then, I label Completeness as `
    Incomplete`
```

Listing 3: Example of a response that is missing information available in the source document.

Example 2: Complete Response

```
{
  "id": 0,
  "question": "What is the logo of a
    major telecommunications company
    mentioned in the context related to
    personalization strategies?",
  "response": "AT&T",
}

# Steps:
# 1. I open folder "0", read all the
    chunks and images.
# 2. The response seems to be complete.
    AT&T is the only answer.

# Then, I label Completeness as `
    Complete`
```

Listing 4: Example of a response that contains all necessary information.

C.4 Grounding Verification

For each question, annotators were required to verify which specific source materials (PDF text chunks or images) were necessary to formulate the answer.

C.4.1 Procedure and Label Definitions

Grounding Verification-chunk-X: After reading the question, the annotator must determine if the text content of `chunk_X.pdf` contains any information used in, or required for, the answer.

- **Required:** The chunk's text contains information needed to answer the question.
- **Not Required:** The chunk's text does not contain any relevant information.

Grounding Verification-img-X: The annotator must determine if `img_X` (including its caption and context within the PDF) contains any information used in, or required for, the answer.

- **Required:** The image or its caption contains information needed to answer the question.
- **Not Required:** The image and its caption do not contain any relevant information.

Example: Grounding Verification

```
{
  "id": 0,
  "question": "What businesses are
    located near the proposed
    development area in the Project
    Catalyst?",
  "response": "AutoZone Auto Parts,
    Pizza Hut, Sonic Drive In, Joe's
    Pizza Italian",
}

# Steps for chunk-0:
# 1. I open folder "0" and then the sub-
    folder chunk_0.
# 2. I read the text within pages.pdf.
# 3. I find part of the answer to the
    question in the text.
# 4. I label `Grounding Verification-
    chunk-0` as `Required`.
```

```
# Steps for chunk-1:
# 1. I check for a sub-folder named
    chunk_1 in folder "0".
# 2. No chunk_1 sub-folder exists, so I
    skip this label.

# Steps for img-0:
# 1. I open folder "0" and then the sub-
    folder img_0.
# 2. I view img_0.jpg and locate it in
    the original PDF to check its
    context.
# 3. I find part of the answer to the
    question in the image.
# 4. I label `Grounding Verification-img
    -0` as `Required`.

# Steps for img-1:
# 1. I open folder "0" and then the sub-
    folder img_1.
# 2. I view img_1.jpg and its context.
# 3. I do NOT find any part of the
    answer in this image.
# 4. I label `Grounding Verification-img
    -1` as `Not Required`.
```

Listing 5: Example demonstrating how to label individual source chunks and images as required or not required.

C.5 Self-Contained Evaluation

This task assesses whether a question is understandable and complete on its own, without needing external context or references to specific, unnamed documents.

C.5.1 Procedure

Annotators were instructed to read only the question and determine if it could be understood and answered without ambiguity, assuming one had access to a large database of documents.

C.5.2 Label Definitions

True: The question is self-contained. It is clearly phrased, makes sense on its own, and provides enough specific detail (e.g., names, topics, concepts) to be answerable. It does not rely on vague document references. For example, "What are the key benefits of solar energy mentioned in the 2022 Department of Energy report?" is self-contained.

False: The question depends on external or implicit context to be meaningful. It may contain vague deictic references (e.g., "in the image above," "according to this chart," "what does this mean?") without clarifying what the reference points to. For example, "What is the logo in the image?" is not self-contained as it requires seeing a specific, un-referenced image.

Example 1: Not Self-Contained

```
{
  "id": 1,
  "question": "What is the logo in the
    image?",
  "response": "AT&T",
}

# Steps:
# 1. I read the question.
# 2. I find it is NOT clear; "what image
    ?" is an unanswered prerequisite.
# 3. I label `Self-Contained` as `False
    `.
```

Listing 6: Example of a question that is not self-contained due to a vague reference ("the image").

Example 2: Self-Contained

```
{
  "id": 0,
  "question": "What is the logo of a
    major telecommunications company
    mentioned in the context related to
    personalization strategies?",
  "response": "AT&T",
}

# Steps:
# 1. I read the question.
# 2. I find it is clear. I can use the
    information within the question to
    search for a relevant document.
# 3. I label `Self-Contained` as `True`.
```

Listing 7: Example of a question that is self-contained because it provides sufficient context ("personalization strategies," "telecommunications company").

C.6 Human-like Intent Evaluation

This task assesses whether a question reflects a natural and meaningful information-seeking intent, typical of a human user interacting with a document or database.

C.6.1 Procedure

Annotators were instructed to read the question and judge its authenticity as a genuine human query. The focus was on the nature of the question's intent rather than its grammatical perfection.

C.6.2 Label Definitions

True: The question represents a reasonable and natural query a human would make. It seeks meaningful information such as facts, summaries, comparisons, or explanations, and is phrased in a way that reflects a real information need. For example: "What were the company's main revenue streams in the last fiscal year?"

False: The question is unnatural, trivial, or does not reflect a plausible human intent. This includes questions that are overly literal (e.g., counting word occurrences), focus on formatting (e.g., font sizes), are phrased robotically, or seek bizarrely specific details that a human would be unlikely to ask.

Example 1: Not Human-like

```
{
  "id": 1,
  "question": "How many logos in the
    Figure one of the major
    telecommunications company?",
  "response": "13",
}

# Steps:
# 1. I read the question.
# 2. I do not think a person using an
    information retrieval system would
    ask this style of question.
# 3. I label `Human-like` as `False`.
```

Listing 8: Example of a question that is not human-like due to its trivial, count-based nature.

Example 2: Human-like

```
{
  "id": 3,
  "question": "What were the top two
    revenues for the EMS division in
    2012?",
  "response": "In 2012, the revenues
    were approximately HK$493,208,000
    and HK$391,677,000.",
}

# Steps:
# 1. I read the question.
# 2. I find it is clear and reflects a
    specific, meaningful financial
    inquiry.
# 3. I label `Human-like` as `True`.
```

Listing 9: Example of a question that reflects a clear, natural, and meaningful information need.

D Examples

D.1 Examples for text-retrieval better than image-retrieval

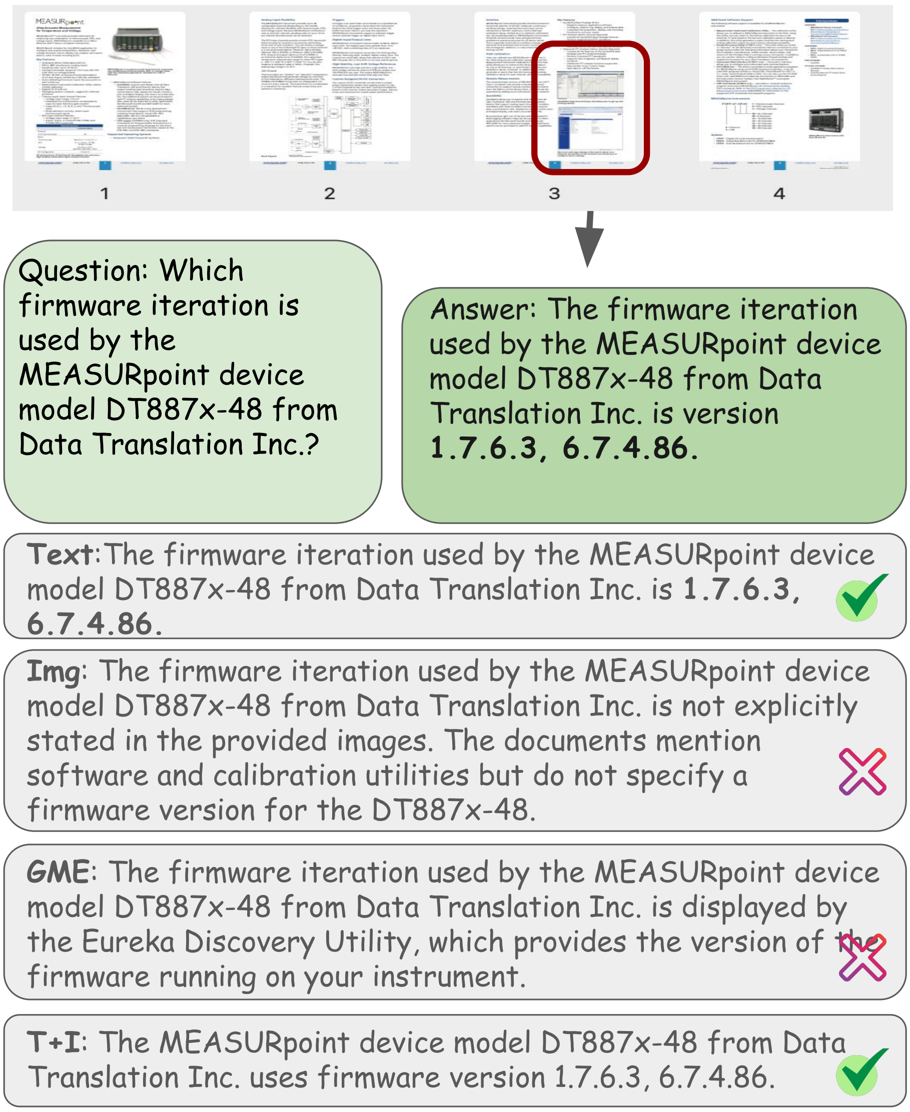


Figure 4: Image-retrieval system fails to extract factual facts and details.

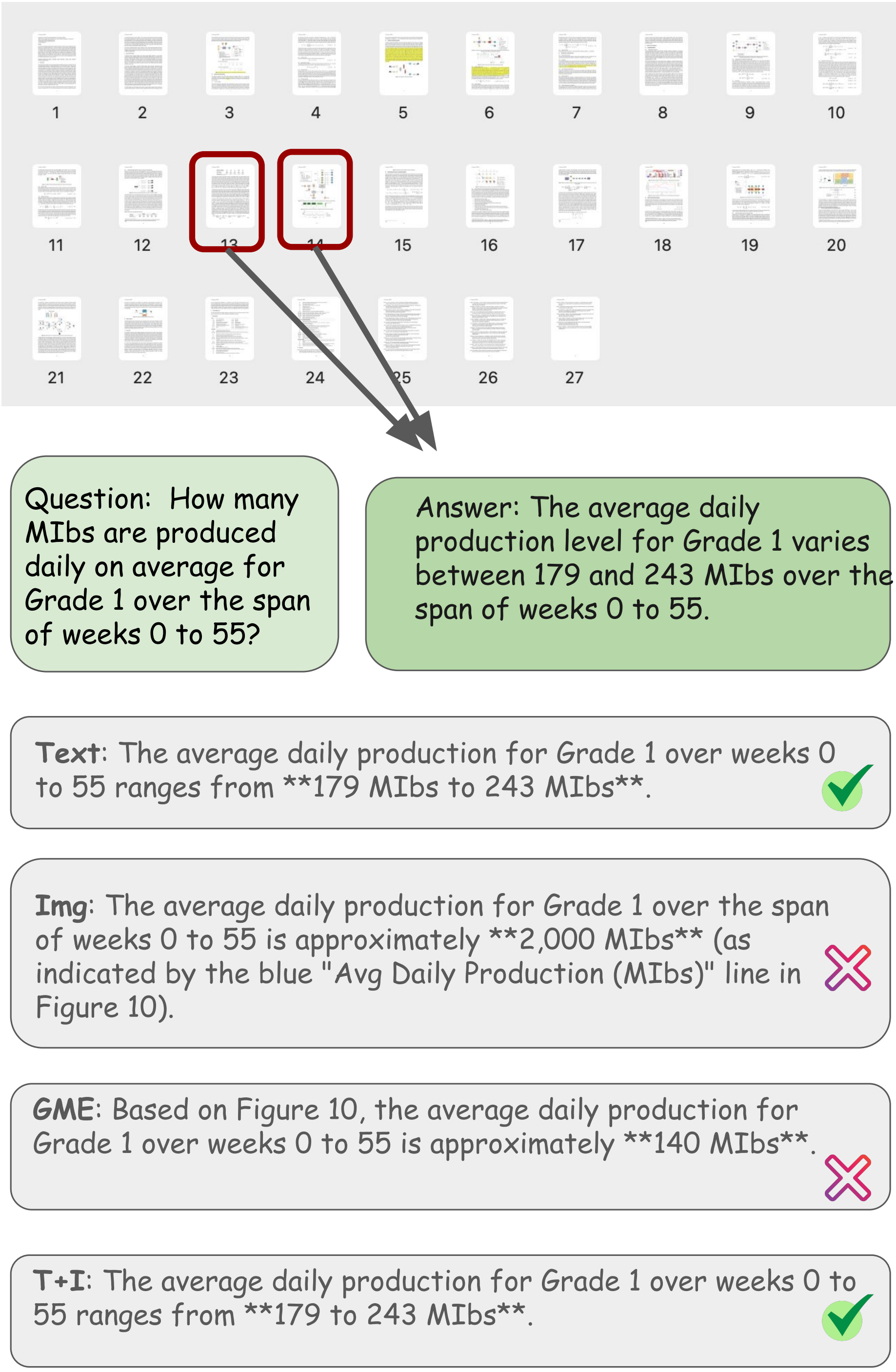


Figure 5: Image-retrieval system fails to extract factual facts and details in the image.

D.2 Examples for image-retrieval better than image-retrieval

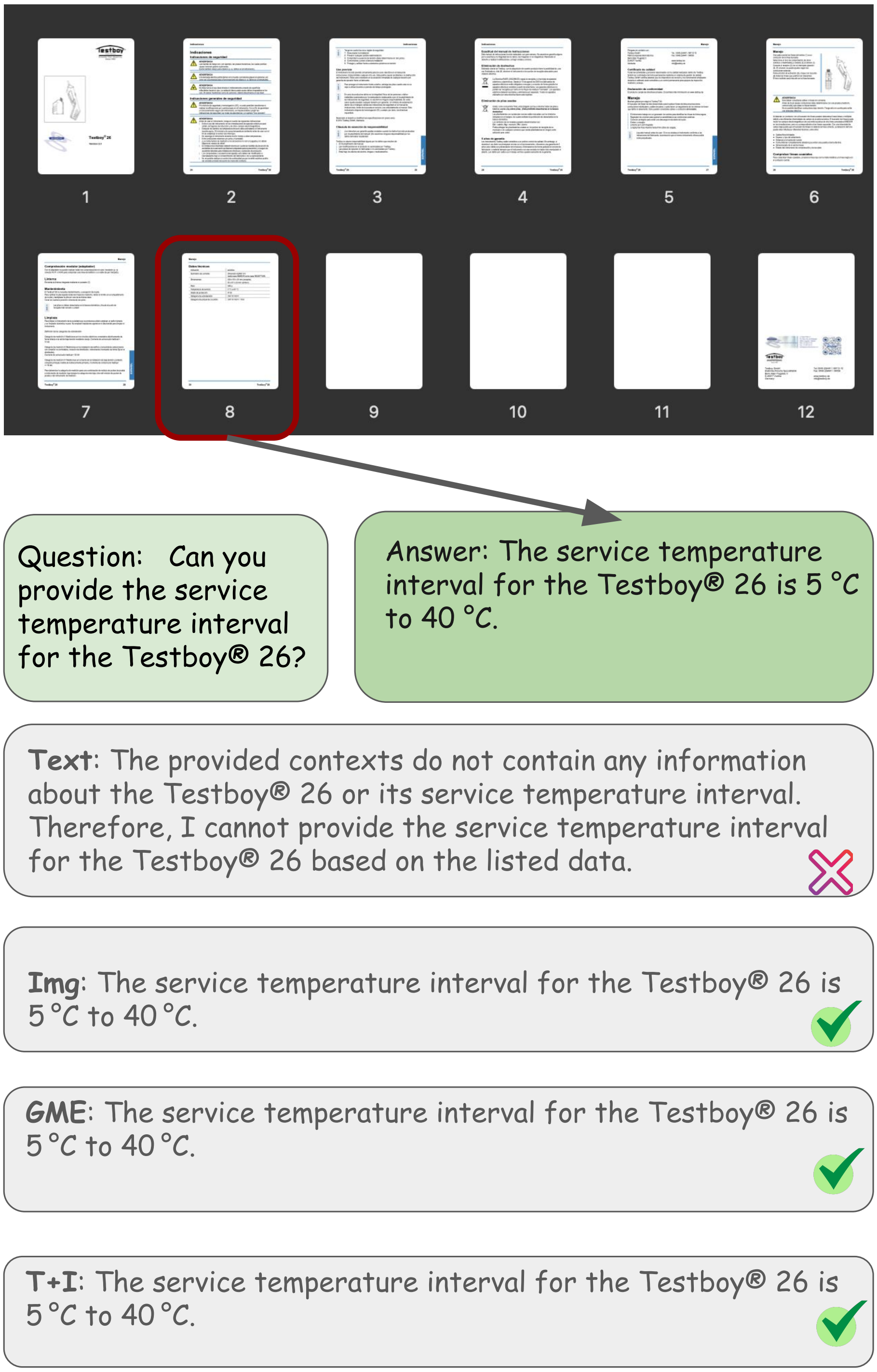


Figure 6: Text-retrieval system fails to extract factual facts and details in the table.

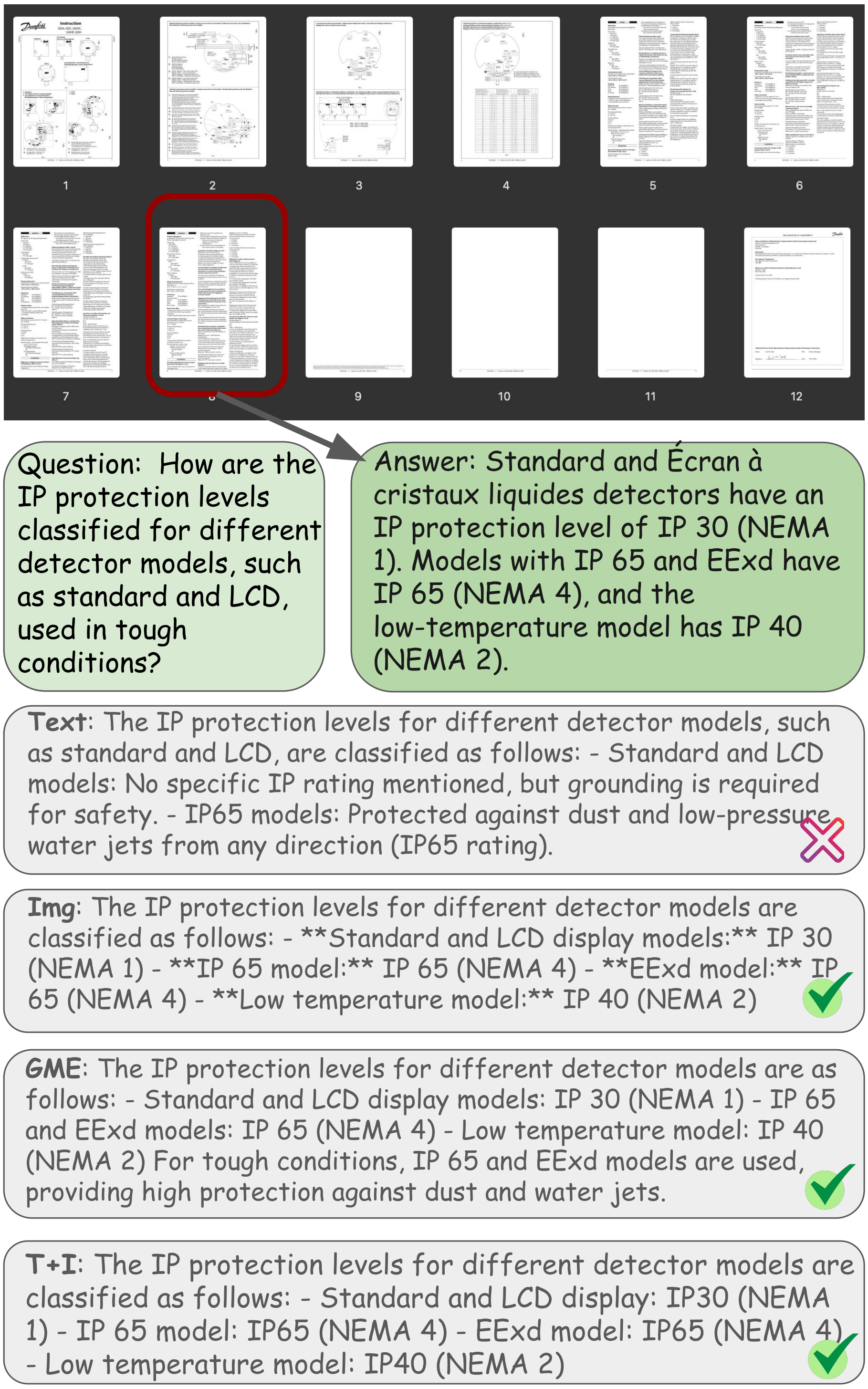


Figure 7: Text-retrieval system fails to extract factual facts and details in the table.

D.3 Examples for multimodal-retrieval better than single-modality-retrieval

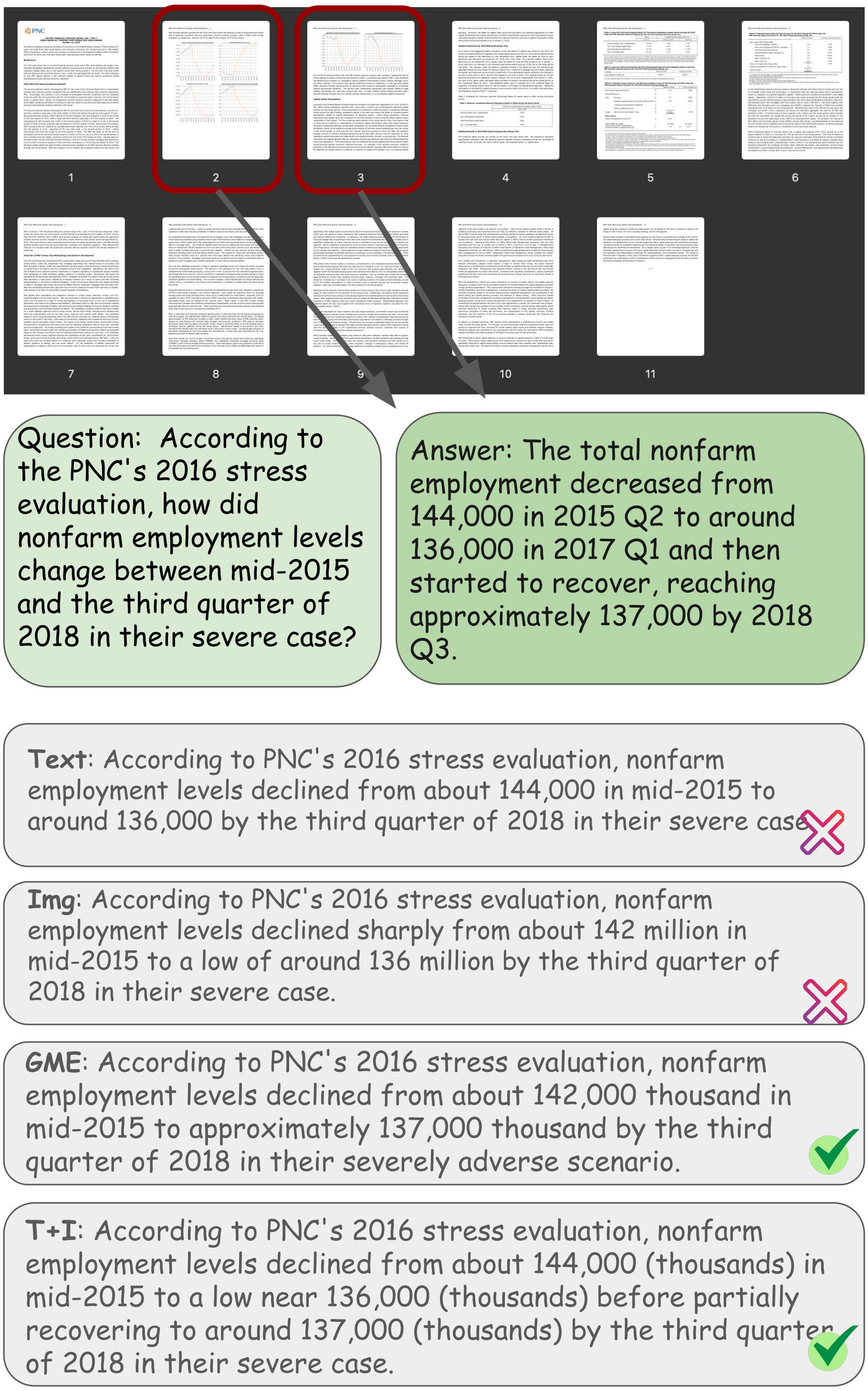


Figure 8: MM RAG system handles multi-modality-evidence questions better.

E Additional Experiments

E.1 Cost Comparison

MM-RAG systems can achieve both better end-to-end performance and lower cost than text-only RAG. Text-only RAG is the most expensive due to high token consumption, while image-only RAG has the lowest cost and latency. Multimodal RAG offers lower cost than text-only RAG with comparable latency.

We report the average inference cost and latency of different RAG systems in Table 6. Image-only systems (IMG) are the most efficient, whereas multimodal systems (MM) are the slowest, reflecting the trade-off between model complexity and capability. The T+I fusion RAG incurs additional latency because it retrieves text chunks before images. Overall, these results show that modern MM-RAG systems can provide improved performance at lower cost than text-only RAG.

F Additional Analysis

F.1 Content-rich images increase difficulty

We analyze images from the easiest domains (*commerce manufacturing* and *legal*) and the most challenging domains (*finance* and *construction*). Using `gemini-2.5-pro`, we classify images as *content-rich* (containing information not present in the text) or *illustrative*. Content-rich images are substantially more common in finance (62.8%) and construction (69.3%) than in commerce manufacturing (40.0%) and legal (49.5%). This suggests that domains with a higher proportion of content-rich images pose greater challenges for RAG, as they require effective multimodal understanding beyond text, consistent with the results in Table 4.

F.2 Question type affects difficulty

As shown in Section 4.2, the type of context required to answer a question is the most significant factor influencing RAG performance. Different categories of questions contribute unevenly to the advantage of either text- or image-retrieval RAG systems. By carefully analyzing questions that can only be answered correctly by one of the two systems, we summarize the key distinguishing features:

Text-Retrieval Advantages:

- *Entity Recognition* (e.g., brands, organizations; 53.9% of text advantage): Strong at identifying specific people, companies, or organizations.
- *Comparative Analysis* (37.6%): Ranking, evaluating differences, or determining which option is preferable.
- *Contextual Numerical Reasoning* (34.8%): Numbers requiring understanding of surrounding context.
- *Quantity Estimation* (29.1%): Questions asking about amounts, counts, or measurements.
- *Domain-Specific Terminology* (16.3%): Technical, scientific, or specialized terms and standards.

Image-Retrieval Advantages:

- *Visual Chart Data Interpretation* (64.2% of image wins): Charts and tables make numerical information more accessible. *Example:* How much of the auto ABS senior tranches in Europe were rated AAA in early 2018?
- *Temporal / Chronological Data* (40.0%): Timeline visualizations clarify temporal relationships. *Example:* When did U.S. petroleum imports drop under $20 billion?
- *Technical / Measurement Information* (19.2%): Diagrams often contain measurements or specifications not in text. *Example:* What is the service temperature interval for Testboy® 26 based on the listed data?
- *Spatial / Geographic Reasoning* (13.3%): Maps and layouts convey location context and spatial relationships. *Example:* What is the impact of delivery time on scheduling at 22 Bishopsgate?

F.3 Document formats do not affect performance.

As discussed in Section 3.1, documents span formats such as newspapers, textbooks, webpages, forms, reports, papers, slides, and posters. In the best-performing domain, *commerce manufacturing*, the distribution is diverse, with reports (45.2%), textbooks (23.6%), papers (18.7%), and webpages (10.5%). In contrast, the worst-performing domain, *finance*, is dominated by reports (80.8%), with only small shares of papers (12.2%), textbooks (2.9%), and webpages (2.3%). Yet this trend is not consistent: the second-worst domain, *construction*, is also diverse, with reports (53.9%), papers (30.4%), and textbooks (11.3%). Therefore, format distribution alone cannot explain performance differences.

Document layouts do not affect performance. In the best-performing domain, *commerce manufacturing*, documents are composed of text (73.9%), tables (4.0%), and figures (22.1%), while the worst-performing domain, *finance*, shows a nearly identical distribution (72.9% text, 3.7% tables, 23.4%

	IMG	TEXT	MM (GME)	MM (T+I)
Avg. Cost ($)	0.012	0.036	0.022	0.029
Avg. Latency (s)	5.606	7.290	7.897	9.383

Table 6: Average cost of different RAG systems.

figures). Since all domains exhibit similar layout patterns, layout does not appear to be a key factor in RAG performance.

F.4 Document page numbers do not affect performance.

In the best-performing domains (*commerce manufacturing*, *education*, and *legal*), the average lengths are 13.1, 14.6, and 12.6 pages, respectively. In contrast, the worst-performing domains (*finance*, *construction*, and *healthcare*) average 15.4, 12.9, and 12.1 pages. These small differences suggest that document length is not a major factor in RAG performance.