REAL-MM-RAG: A Real-World Multi-Modal Retrieval Benchmark

Navve Wasserman[1,2], Roi Pony[1], Oshri Naparstek[1], Adi Raz Goldfarb[1]
Eli Schwartz[1], Udi Barzelay[1], Leonid Karlinsky[1]
[1]IBM Research Israel [2]Weizmann Institute of Science
navve.wasserman@weizmann.ac.il

Abstract

Accurate multi-modal document retrieval is crucial for Retrieval-Augmented Generation (RAG), yet existing benchmarks do not fully capture real-world challenges with their current design. We introduce REAL-MM-RAG, an automatically generated benchmark designed to address four key properties essential for real-world retrieval: (i) multi-modal documents, (ii) enhanced difficulty, (iii) Realistic-RAG queries and (iv) accurate labeling. Additionally, we propose a multi-difficulty-level scheme based on query rephrasing to evaluate models' semantic understanding beyond keyword matching. Our benchmark reveals significant model weaknesses, particularly in handling table-heavy documents and robustness to query rephrasing. To mitigate these shortcomings, we curate a rephrased training set and introduce a new finance-focused, table-heavy dataset. Fine-tuning on these datasets enables models to achieve state-of-the-art retrieval performance on REAL-MM-RAG benchmark. Our work offers a better way to evaluate and improve retrieval in multi-modal RAG systems while also providing training data and models that address current limitations. Our benchmark is available at this project page.

1 Introduction

Accurate retrieval of relevant documents is a cornerstone of modern natural language processing (NLP) applications, whether used alone or in advanced pipelines like Retrieval-Augmented Generation (RAG). RAG (Lewis et al., 2020) has emerged as a powerful approach wherein models retrieve external information before generating answers or content, enabling operation over large document collections. *Multi-modal RAG* extends this to real-world scenarios involving text, figures, tables, and potentially entire page images.

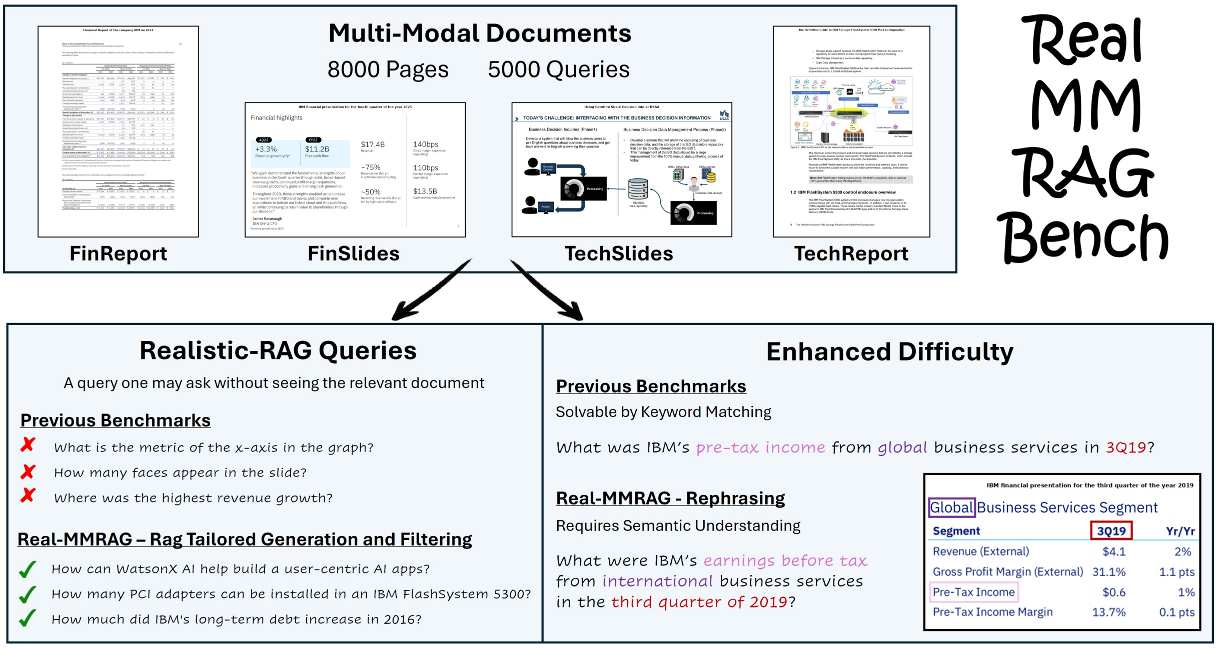
Successful retrieval is crucial for RAG, as retrieving the wrong page or document inevitably hinders the final generated response. Therefore our analysis will focus on the retrieval part. Although research on RAG is advancing, the field still lacks a complete understanding of how models perform in realistic setups, both for evaluating performance and identifying current weaknesses to overcome. This gap arises from a shortage of benchmarks that thoroughly assess real-world retrieval challenges.

We identify four essential properties for a real-world document retrieval benchmark, particularly in multi-modal contexts: ***(i) Multi-modal documents:*** The dataset should include pages with text, figures, tables, and other visual elements to reflect the complexity of real-world materials. ***(ii) Enhanced difficulty:*** Queries should require more than simple keyword matching and involve a large corpus of contextually similar pages to ensure challenging evaluations. ***(iii) Realistic-RAG queries:*** Questions must be posed naturally, without explicit references to pages—reflecting queries a person might ask when seeking information without knowing the answer's location (in contrast to generic question-answering setups). ***(iv) Accurate labeling:*** All documents relevant to a query must be correctly and exhaustively labeled to prevent underestimation of retrieval performance and to avoid false negatives.

Although a few recent benchmarks touch on some of these aspects (Faysse et al., 2024; Ma et al., 2024a,b), most fail to fully capture them, limiting their usefulness for understanding and improving multi-modal retrieval models. We introduce ***REAL-MM-RAG-Bench*** *(Real-World Multi-Modal Retrieval-Augmented Generation Benchmark)*, a benchmark designed to satisfy the properties above:

Multi-modal documents (Property i): Our dataset comprises slides and documents with text, figures, tables, and images, requiring systems to handle combined textual and visual data.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 31660–31683
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics



A query one may ask without seeing the relevant document

Previous Benchmarks

✗ What is the metric of the x-axis in the graph?

✗ How many faces appear in the slide?

✗ Where was the highest revenue growth?

Real-MMRAG – Rag Tailored Generation and Filtering

✓ How can WatsonX AI help build a user-centric AI apps?

✓ How many PCI adapters can be installed in an IBM FlashSystem 5300?

✓ How much did IBM's long-term debt increase in 2016?

Previous Benchmarks

Solvable by Keyword Matching

What was IBM's pre-tax income from global business services in 3Q19?

Real-MMRAG - Rephrasing

Requires Semantic Understanding

What were IBM's earnings before tax from international business services in the third quarter of 2019?

Figure 1: **Proposed Real-MM-RAG Benchmark**

Enhanced difficulty (Property ii): Instead of relying on isolated pages or trivial queries, we focus on long documents from specialized domains (e.g., IBM finance reports, FlashSystem technical materials). This makes retrieval significantly more challenging, as models must differentiate between highly similar content within the same domain. Additionally, we incorporate a rephrasing step to ensure that query wording and order are not identical to the page content, requiring semantic understanding rather than simple keyword matching.

Realistic-RAG queries (Property iii): We use a two-step process to generate queries: vision-language models (VLMs) create retrieval-focused queries, and large language models (LLMs) filter them to ensure natural phrasing and realistic user intent. Unlike many existing datasets, our queries avoid direct references to specific pages, reflecting authentic retrieval scenarios.

Accurate labeling (Property iv): Ensuring all pages answering a query are correctly identified is crucial, especially in benchmarks with many similar pages. Existing benchmarks often mislabel valid documents as incorrect, leading to false negatives. To address this, we employ an automated pipeline using VLMs to verify query relevance to each page. While computationally intensive, this approach enhances labeling reliability, particularly for closely related pages.

Our benchmark enables reliable evaluation of current retrieval models and uncovering some of their weaknesses. Additionally, it incorporates two essential properties that expose specific retrieval challenges:

Rephrasing Robustness Evaluation: In real-world scenarios, users rarely phrase their queries exactly as they appear in documents. However, both VLMs and human annotators tend to generate queries that closely mirror the source material, often using similar words and sentence structures (Smeaton and Kelledy, 1998; Zhu et al., 2024). This fails to reflect natural user behavior, where users are not directly exposed to the document page when forming queries. To address this, we introduce a multi-level rephrasing benchmark, modifying queries at three distinct levels—ranging from slight rewording to significant structural changes. Our experiments show that current retrieval models struggle to maintain performance across these variations, highlighting a critical weakness in their semantic understanding.

Table-Focused Scenarios: Table-heavy documents (e.g. financial reports) often contain dense tabular data, posing a major challenge for retrieval models. By incorporating table-heavy documents into our benchmark, we expose key deficiencies in table comprehension that significantly impact model performance. These properties allow us to demonstrate that all current retrieval models exhibit weaknesses in handling both rephrased queries and table-heavy financial documents.

To address these shortcomings, we leverage insights from our benchmark to enhance retrieval performance. Specifically, we introduce two targeted training strategies: (i) a *rephrased training dataset*, generated by rephrasing the ColPali training dataset (Faysse et al., 2024), and (ii) a *finance-table-heavy training set*, designed to improve retrieval in tabular contexts. Fine-tuning the current best model on these datasets achieves state-of-the-art retrieval performance on our benchmarks. This demonstrates how systematic evaluation through our benchmark can informs effective training strategies, leading to more robust and adaptable retrieval models.

The contributions of this paper are as follows:

- Defining properties of a real-world retrieval benchmark and highlighting shortcomings in existing ones.
- Introducing a high-quality multi-modal retrieval benchmark with an automated pipeline for query generation, filtering, and labeling verification.
- Establishing a rephrasing robustness evaluation framework via multi-level query rephrasing.
- Providing two specialized training datasets: (i) a rephrased dataset and (ii) a finance-table-heavy dataset, where fine-tuning on them significantly enhances retrieval performance.

2 Related Work

2.1 Text-Based Retrieval

Text-based retrieval methods identify relevant documents given a query and are widely used in RAG systems. Lexical matching techniques like BM25 (Robertson et al., 1994) and TF-IDF (Sparck Jones, 1972) are efficient but lack semantic understanding. Sparse models like SPLADE (Formal et al., 2021) improve retrieval by expanding queries into high-dimensional sparse representations but struggle with deep contextual meaning. Dense retrieval models, leveraging transformers like BERT (Devlin, 2018), T5 (Raffel et al., 2020), and DPR (Karpukhin et al., 2020), map queries and documents into a continuous vector space, enhancing recall but demanding significant computational resources for training and inference. Hybrid methods, that combine lexical and dense retrieval, such as ColBERT (Khattab and Zaharia, 2020) and ANCE (Xiong et al., 2020), often achieving state-of-the-art performance. A recent model, M3-Embedding (Chen et al., 2024), unifies dense, sparse, and multi-vector embeddings, achieving strong retrieval performance. Despite advancements, text-based retrieval struggles with multi-modal content, particularly in scenarios where visual cues enhance contextual understanding.

2.2 Multi-Modal Retrieval

Until recently, multi-modal retrieval primarily relied on Optical Character Recognition (OCR) to extract textual information from documents, including text within visual elements. More recent approaches detect visual components and process them in one of two ways: (i) Captioning-based retrieval, where a VLM generates textual descriptions of visual elements, enabling standard text-based retrieval (Ramos et al., 2023); or (ii) Direct embedding, where visual elements are embedded either using VLMs directly or through contrastive Vision-Language Models that align separate visual and text encoders via contrastive losses (Radford et al., 2021; Zhai et al., 2023).

A more recent line of work leverages the strong performance of VLMs in analyzing full document images by embedding entire document pages instead of relying on OCR-based extraction. Methods such as VISRAG (Yu et al., 2024) and DSE (Ma et al., 2024a) generate dense embeddings directly from document images. Similarly, ColPali (Faysse et al., 2024) generates multi-vector embeddings for ColBERT-style late interaction retrieval, using PaliGemma (Beyer et al., 2024) or in a newer variant, ColQwen, utilizes Qwen2-VL (Wang et al., 2024a). These methods have demonstrated significant improvements over earlier text-based and OCR-dependent retrieval approaches. Our benchmark provides a rigorous evaluation framework for both text-based and visual-based multi-modal retrieval.

2.3 Query Rephrasing

Retrieval models known to be highly vulnerable to query rephrasing (Zuccon et al., 2016; Bailey et al., 2017; Sidiropoulos and Kanoulas, 2022; Penha et al., 2022; Hagen et al., 2024), often leading to significant performance degradation. However, only few works have provided accessible and reliable evaluation frameworks for model robustness. An early study (Bailey et al., 2016) introduced a basic Query Variability dataset, while more recent works (Benham et al., 2018; Lu et al., 2019; Penha et al., 2022) focus on automatically generating query variations. Yet, no prior research has established a standardized benchmark for retrieval

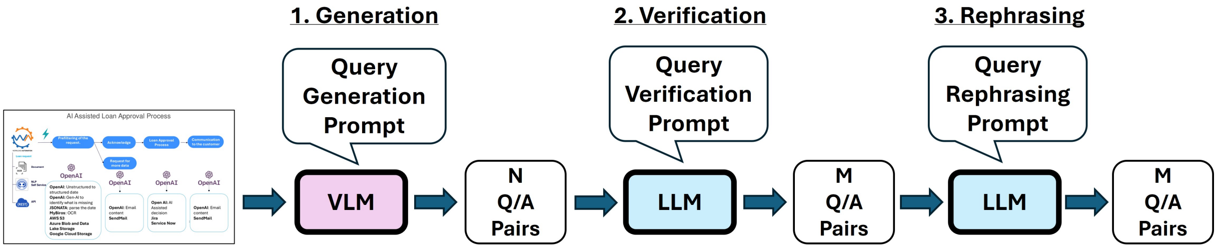


Figure 2: **Benchmark Construction Pipeline**

robustness, nor a dedicated RAG robustness benchmark—especially for multi-modal retrieval. In contrast, we leverage LLMs to generate multi-level query rephrasing, enabling structured comparative evaluation for multi-modal documents retrieval.

2.4 Multi-Modal Retrieval Benchmarks

Despite the growing importance of document retrieval in multi-modal RAG systems, only a few evaluation benchmarks exist, all of which fall short in key aspects crucial for real-world scenarios. We review recent efforts and highlight their limitations (with further comparison in Table 1 and section 4). While many question-answering benchmarks exist (Mathew et al., 2021; Zhu et al., 2022; Masry et al., 2022; Islam et al., 2023; Ding et al., 2024), they are largely unsuitable for RAG. Their queries assume exposure to a specific page (unlike RAG), and their tendency toward high variability make retrieval easier. Some benchmarks, such as MMLongBench (Ma et al., 2024b) (based on 130 lengthy PDFs) and SlideVQA (Tanaka et al., 2023), are partially relevant but not suited for RAG (see section 4).

A notable benchmark is WIKI-SS-NQ (Ma et al., 2024a), generated from Wikipedia screenshots with real human queries—the only dataset providing mostly valid retrieval queries. However, it is not multi-modal benchmark, consists of mainly text-based documents, and has narrow sub-domain coverage. The ViDoRe benchmark (Faysse et al., 2024), introduced in the ColPali paper, comprises both QA datasets and domain-specific documents with generated queries filtered by human annotators. While the QA datasets are unsuitable for RAG, the domain-specific queries are better tailored but suffer from trivial difficulty (e.g., synthetic datasets reporting an NDCG@5 > 95). This occurs because pages often differ significantly, and VLM-generated questions closely mirror the original page wording, making retrieval easy.

Our REAL-MM-RAG-Bench is the first multi-modal retrieval benchmark incorporating all essential properties for real-world RAG. It features a challenging setup with broad sub-domain coverage, long documents, RAG-tailored rephrased queries, and accurate labeling. Additionally, it is the first to offer a robustness evaluation through multi-level query rephrasing.

3 REAL-MM-RAG-Bench

Creating a high-quality benchmark manually is both exhaustive and error-prone, limiting its size and reliability. To address this, we propose an ***automated generation and verification pipeline*** tailored for Retrieval-Augmented Generation (RAG) evaluation. Our benchmark introduces robustness evaluation through ***multi-level query rephrasing***, further improving upon previous benchmarks. The benchmark construction begins with ***document collection***, followed by four key steps: (1) ***Query Generation***, (2) ***Query Verification***, (3) ***Query Rephrasing***, and (4) ***False Negative Verification***.

3.1 Document Collection

To reflect real-world retrieval challenges, we focus on ***long documents*** rather than isolated pages, and also ensuring ***many pages within the same sub-domain*** by focusing on a single company data (IBM). Our dataset consists of 8000 pages across four sub-domains, forming four specialized benchmarks (see Table S1 for details). For each page, we added the document name to the page image to provide context. **FinReport**: Financial reports (2005–2023), totaling 19 documents and 2687 pages, with a mix of text and tables. **FinSlides**: Quarterly financial presentations (2008–2024), totaling 65 presentations and 2280 pages, primarily table-heavy. **TechReport**: 17 Technical documents on FlashSystem, totaling 1674 pages, text-heavy with visual elements and tables. **TechSlides**: 62 Technical presentations on business and IT automation, totaling

	Statistics		Multi-Modal	Enhanced Difficulty			Realistic-RAG Queries		Accurate Labels
	#	#	MM	Long	Sub	Queries	RAG	RAG	False
	Pages	Queries	Pages	Docs	domain	Rephr-	Tailored	Query	Neg.
Benchmark					Cover	asing	Gen.	Verif.	Verif.
SlideVQA	52k	14.5k	✓	✓	✗	✗	✗	✗	✗
MMLONG	7k	1k	✓	✓	✗	✗	✗	✗	✗
WIKI-SS-NQ	4k	4k	✗	✗	✗	✗	✓	✓✗	✗
ViDoRe	8k	4k	✓	✗	✗	✗	✓✗	✓✗	✗
Ours	8k	5k	✓	✓	✓	✓	✓	✓	✓

Table 1: **Document Retrieval Benchmarks Comparison.**

1963 pages, with significant visual content.

3.2 Query Generation & Filtering

Generation. We aim to generate queries that are both answerable by a specific document and RAG-suitable, meaning they reflect natural user inquiries without prior knowledge of the exact page or answer location (unlike traditional Q/A datasets tied to specific pages). To achieve this, we employed a Pixtral-12B VLM (Agrawal et al., 2024), prompting it to generate RAG-specific questions (see Fig. S4). Each document page was fed into the VLM, which produced 10 query-answer pairs per page, later keeping only a subset that met the benchmark's quality criteria after filtering. Each retained query-answer pair is labeled with the corresponding page it was generated from.

Verification. Although the VLM is instructed to generate RAG-specific queries, many still do not fully align with our requirements. To systematically classify them, we use Mixtral-8x22B-v0.1 LLM (Jiang et al., 2024), which evaluates each generated query and determines whether it is suitable as retrieval query (see prompt in Fig. S5). Queries that are well-formed for RAG are those that a user might ask without prior knowledge of the document's structure, ensuring they are neither too general nor overly specific to a single page. Queries that fail this criterion fall into two categories: those with explicit page references, such as "in Figure 5" or "the title of the page", and those that are too broad, like "What is the net revenue in 2020?" instead of "What is IBM's net revenue in 2020?".

3.3 Query Rephrasing

In real-world retrieval, a user formulating a query does not have direct access to the document's content and will naturally phrase their question without mirroring the exact wording from the source. However, VLMs often generate queries by copying phrases directly, leading to an over-reliance on keyword matching rather than true semantic retrieval. To address this, we introduce a rephrasing step that preserves query meaning while reducing dependence on specific document wording. Each query is processed by Mixtral-8x22B-v0.1 with a dedicated prompt designed to alter phrasing while maintaining intent. The rephrased query is then verified by the LLM using a validation prompt (Fig. S6), along with the original query and answer, to ensure it retains the original meaning and still corresponds to the known answer in the labeled page.

To enable deeper evaluation, each query undergoes three levels of rephrasing using distinct prompts (Fig. S6). The first level introduces minor word changes while maintaining structure. The second modifies word choice and sentence order, making the phrasing more distinct. The third involves significant word rephrasing and sentence restructuring while preserving meaning. At the end of this process, each query exists in four versions: the original and three progressively rephrased forms, all linked to the same document page (see examples in Figs. S1 and S2).

3.4 Accurate Labeling

The final step in preparing our benchmark is verifying the correctness of negative labels. This is especially crucial for our challenging benchmarks, where many pages share highly similar content within the same sub-domain. Each query is systematically tested against all benchmark pages. Though computationally expensive, this step prevents false negatives and ensures reliable evaluation. Queries together with each page are processed using Pixtral-12B, which determines whether a page contains an answer to the query. Every query is then explicitly linked to all relevant pages. For simplicity, our final benchmark retains queries whose only the originally assigned page is verified to contain the correct answer. This results in a high-quality dataset of triplets: a page image, a query, and its corresponding answer. Note that our benchmark includes pages without corresponding queries. These are pages whose queries were filtered out at some stage, either because they were not suitable for RAG-style questions in general (e.g., title pages) or because the specific generated queries were not suitable for RAG.

4 Benchmarks Quality Evaluation

A high-quality benchmark for multi-modal retrieval is essential, yet few existing benchmarks are designed for this purpose, and none comprehensively define or implement the necessary properties. Table 1 compares our benchmark with other prominent ones, which suffer from limitations such as poor alignment with real-world queries, high false-negative rates, and trivial difficulty.

Accurate Labeling. Many perceived retrieval errors in existing benchmarks are actually false negatives, meaning pages that correctly answer the query but were mislabeled as irrelevant. To mitigate this, we introduce a false-negative verification process that exhaustively labels all valid pages. ***Human Evaluation.*** We sampled 50 top-1 retrieval errors of ColQwen on Vidore, MMlongbench, and our benchmark. Annotators reviewed the query and retrieved page (labeled as negative) to determine if it could answer the query (Fig. S8). A total of 234 responses from 5 annotators were collected.

	Vidore	MMLong	Ours
False Negative (%) ↓	86.9	77.8	31.9

The table shows that Vidore and MMlongbench had a high rate of false negatives, whereas our benchmark, despite its challenging design with similar sub-domain pages, had significantly fewer, proving the effectiveness of accurate labeling.

Enhanced Difficulty. A strong benchmark must pose real challenges. Existing ones fall short by offering too few relevant candidates or allowing retrieval via simple keyword matching rather than true semantic understanding. For example, having 1,000 financial pages from different companies is insufficient, as knowing the company name narrows the candidates to a few dozen. The ColQwen model achieves an NDCG@5 of around 90 on Vidore. Other sub-datasets, although reporting lower performance, contain many errors that are actually false negatives, as demonstrated by our human evaluation presented above. We address this issue through accurate labeling and by incorporating long documents and extensive sub-domain coverage. This provides many similar pages, making retrieval more challenging and better reflecting real-world scenarios. Moreover, we prevent trivial keyword-based retrieval by introducing the first rephrasing benchmark for multi-modal document RAG, ensuring robustness to query variations and promoting semantic learning.

Realistic-RAG Queries. To reflect real RAG use cases, queries must resemble natural information-seeking questions. Our benchmark ensures this through a two-step RAG-tailored pipeline: generation and filtering. ***Human Evaluation.*** We randomly sampled 500 queries from Vidore, MMLongBench, and our benchmark. Annotators, unaware of the source benchmark or study goal, evaluated whether each query could reasonably be asked by a real user (Fig. S7). A total of 2578 responses were collected from 10 annotators.

	Vidore	MMLong	Ours
Realistic-RAG Queries (%) ↑	32.9	35.8	88.0

The table shows that most Vidore/MMLongBench queries were labeled as unrealistic RAG queries (see some examples in Fig. S3), whereas 85% of ours were validated as realistic, highlighting shortcomings in existing benchmarks and the effectiveness of our query generation and filtering process.

Summary. Our benchmark enhances multi-modal retrieval evaluation by introducing non-trivial difficulty with long documents and broad sub-domain coverage. It ensures RAG-aligned queries and promotes semantic retrieval over keyword matching through query rephrasing, addressing key limitations of existing benchmarks.

5 Model Evaluation & Enhancement

Evaluation Models and Metrics. We evaluate multiple models on our benchmark, covering both text and vision-based approaches. We use the ColPali benchmark code (ViDoRe) to assess our text-based models and the vision-based models ColPali and ColQwen. For the text-based methods, we follow the framework suggested in ColPali, which employs *Unstructured* in a high-resolution configuration with OCR engine to parse PDFs. For each document, *Unstructured* produces text chunks and

Benchmark	FinReport	FinSlides	TechReport	TechSlides
Text				
BM25 (OCR)	21.7	5.9	35.1	31.2
BGE-M3 (OCR)	36.5	11.4	37.1	49.7
BM25 (Captioning)	25.3	9.9	37.2	36.1
BGE-M3 (Captioning)	35.9	13.8	37.5	51.7
Vision				
ColPali	34.5	27.6	62.0	75.8
***Rob**ColPali*	47.1 +12.6	48.4 +20.8	66.6 +4.60	82.8 +7.0
***Tab**ColPali*	50.5 +16.0	41.5 +13.9	61.3 -0.7	77.6 +1.8
***RobTab**ColPali*	**63.2** +28.7	**58.3** +30.7	**70.7** +8.7	**83.3** +7.5
ColQwen	41.8	31.1	66.9	78.1
***Rob**ColQwen*	47.5 +5.7	44.3 +13.2	69.5 +2.6	83.0 +4.9
***Tab**ColQwen*	54.0 +12.2	49.6 +18.5	65.9 -1.0	78.9 -0.8
***RobTab**ColQwen*	**67.1** +25.3	**61.6** +30.5	**73.2** +6.3	**85.0** +6.9

Table 2: **Performance of Different Models on Our Benchmark.** We evaluate various models, including text- and vision-based approaches, across our four benchmarks. Results, measured using NDCG@5, are reported on our final benchmark with queries rephrased at the highest level (Level 3). We also present results for our fine-tuned models trained on our proposed datasets: *Rob* – trained on a rephrased dataset, *Tab* – trained on a table-heavy dataset, and *RobTab* – incorporating both.

visual chunks (e.g., tables, figures, images). We consider two text-based variants: (1) *OCR*, where visual data is processed through an OCR engine, and (2) *Captioning*, where visual elements are described using a Vision Language Model (Qwen2-VL-72B-Instruct (Wang et al., 2024b)). We then evaluate two retrieval methods: *Okapi BM25* and *BGE-M3* (Chen et al., 2024) (see appendix A.1 for more details). We report NDCG@5 as our primary ranking metric, which evaluates how well relevant items are ranked within the top 5 results, giving higher importance to those appearing earlier. Additional metrics and details provided in A.1 & A.5.

5.1 Results Analysis

In Table 2, we report NDCG@5 performance for different models across our four benchmarks with high rephrasing levels, which better reflect real-world scenarios. We first present observations about the vanilla models, including text-based models, ColPali, and ColQwen: ***Visual vs. Text-Based Models.*** Vision-based models, which use VLMs on page images, significantly outperform text-based models across all benchmarks. This supports the notion that visual information is essential for our benchmark and that these models can effectively utilize it. ***Non-Trivial Difficulty.*** Performance is generally low, especially compared to Vidore, where ColQwen achieves nearly 90% on average. ***Rephrasing Effects on Performance.*** Some of the drop in performance is due to rephrasing. In Table 3, we analyze the impact of rephrasing level, showing a clear performance drop as rephrasing intensity increases. BM25 suffers the most, as expected for a lexical-based model, while dense retrieval models are more resilient. ***Table-Heavy Finance Benchmarks Are Harder.*** Our financial benchmarks (FinReport, FinSlides), which are table-heavy, are significantly more challenging than text/visual-based ones (see table vs. non-table analysis in Table S5).

5.2 Table and Finance Focused Training

To address the challenges in table-heavy datasets, we curated a table-focused finance dataset using FinTabNet (Zheng et al., 2021), which contains complex tables from S&P 500 company reports. Through the pipeline in section 3, we generated 46,000 query-answer-page triplets to enhance retrieval for table-heavy financial data (see Table S7 for a VLM ablation study). We fine-tuned ColPali and ColQwen for one epoch on this dataset while incorporating the ColPali training set, producing the ***TabCol*** models.

As shown in Fig. 3 and Table 2, TabCol models significantly improve performance on financial benchmarks, effectively addressing table-heavy dataset challenges. Importantly, this enhancement does not come at the cost of generalization, as TabCol models continue well across the two other benchmarks.

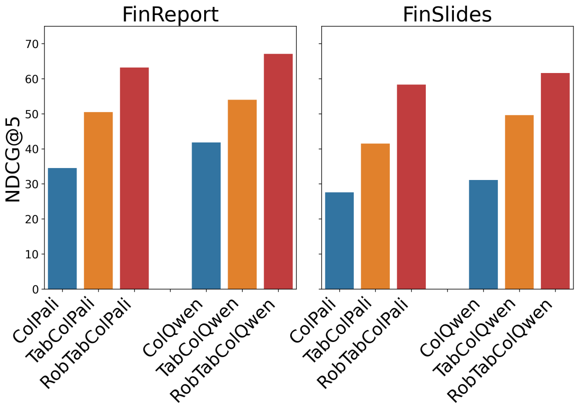


Figure 3: **Table-Focused Training Improves Financial Benchmarks.** Fine-tuning with our proposed table-heavy training set, combined with the ColPali training set (both in their original and rephrased versions) significantly enhances performance on financial benchmarks (results shown for rephrasing level 3).

Rephrasing Levels	**0**	**1**	**2**	**3**
BM25 (Captioning)	52.7	41.6	31.3	27.1
BGE-M3 (Captioning)	43.3	39.2	36.5	34.7
ColPali	71.3	66.1	56.0	50.6
RobColPali	76.7	73.9	66.0	61.2
RobTabColPali	80.8	77.8	72.5	68.9
ColQwen	73.9	67.5	59.6	54.5
RobColQwen	74.1	70.2	64.4	61.1
RobTabColQwen	83.9	80.6	75.1	71.7

Table 3: **Query Rephrasing Effect.** This table presents the NDCG@5 scores averaged across all benchmarks for different models and rephrasing levels. '0' represents no rephrasing, while '3' indicates significant rephrasing (see Table S2 for full results).

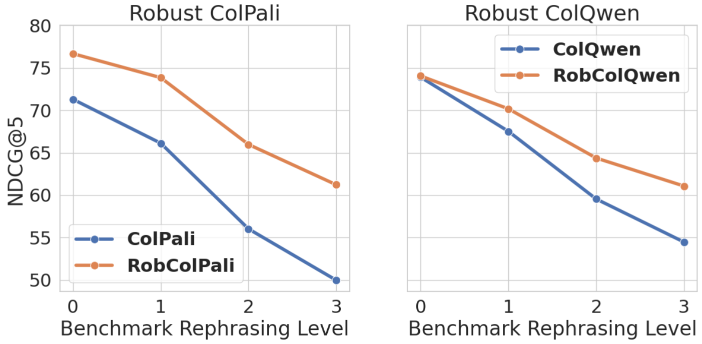


Figure 4: **Fine-Tuning on Rephrased Training Set.** We compare the NDCG@5 scores across rephrasing levels for baseline models (ColPali and ColQwen) against our fine-tuned models (RobCol). The results demonstrate that fine-tuning with our rephrased training data significantly enhances rephrasing robustness for both ColPali and ColQwen.

5.3 Rephrasing Robustness Training

Our benchmark reveals that current models struggle with rephrasing, suggesting that training and evaluation queries often closely match the phrasing of their retrieved pages. To address this, we augmented the ColPali training set by rephrasing half of its queries, randomly selecting one of three rephrasing levels. This was done using LLaMA-3-70B[1] , a different LLM than the one used for the benchmark. This dataset forces models to learn semantics rather than relying on keyword matching. We fine-tuned ColPali-v1.2 and ColQwen2-v1.0[2] (using the ColPali code) for one epoch, producing the RobCol models.

As demonstrated in Fig. 4 and Table 3 (and in Tables 2 and S2), RobCol significantly outperforms baselines on rephrased queries while maintaining comparable performance on non-rephrased cases, achieving an average NDCG@5 improvement of 11.1 at rephrasing level 3. The gains are stronger on financial benchmarks, where the task is more complex, suggesting that enhanced semantic understanding through rephrasing robustness is particularly beneficial. In Table S6, we show that fine-tuning on the original data (rather than the rephrased version) provides much lower improvement for ColPali and leads to a decrease in performance for ColQwen.

Additionally, we created a rephrased version of the tabled-focused dataset and fine-tuned both models on it, along with the rephrased ColPali training set, producing RobTabCol models. RobTabCol consistently outperforms all models across nearly all benchmarks and rephrasing levels, achieving a 25–30 NDCG@5 improvement over base models on rephrased finance benchmarks and 6–9 on non-financial ones. We further show in appendix A.6 that our RobTab models outperform the baselines on the newly released ViDoRe V2 benchmark.

6 Conclusions

We introduced *REAL-MM-RAG-Bench*, a real-world multi-modal retrieval benchmark designed to evaluate retrieval models in *reliable, challenging, and realistic* settings. Our benchmark addresses key properties essential for evaluating retrieval systems in real-world applications, which prior benchmarks often fail to capture. An important contribution of our work is the introduction of a *multi-level rephrasing evaluation*, which assesses models under increasing linguistic variation, highlighting their limitations in generalizing beyond surface-level text matching.

Our findings reveal two major weaknesses in current models: (i) retrieval over table-heavy financial documents and (ii) sensitivity to query rephrasing. To address these, we proposed dedicated training sets: a *finance-table-heavy dataset* to improve retrieval on tabular content and a *rephrased dataset* to enhance model robustness to query variations. Fine-tuning on these datasets yields significant improvements across benchmarks, demonstrating the impact of targeted training data. *REAL-MM-RAG-Bench* and our proposed solutions establish a foundation for future research, paving the way for robust and effective retrieval models in real-world multi-modal retrieval scenarios.

[1] https://huggingface.co/meta-llama/Llama-3-70B-Instruct
[2] https://huggingface.co/vidore/colqwen2-v1.0

Limitations

While REAL-MM-RAG-Bench presents a comprehensive and realistic evaluation framework for multi-modal retrieval, several limitations remain:

Query Variability: Our queries are generated using a Vision-Language Model (VLM), which, while effective, may not fully capture the full range of plausible user queries.

LLM and VLM Limitations: Despite the strength of modern Large Language Models (LLMs) and VLMs, our filtering strategies and labeling process remain subject to their limitations. While our human evaluation confirms the effectiveness of our approach, errors in labeling and query selection may still occur. As LLMs and VLMs continue to improve, future benchmarks could leverage more accurate models to refine dataset construction further.

Multi-Page Reasoning Queries: Our benchmark is designed to best evaluate the retrieval component of Retrieval-Augmented Generation (RAG). While the dataset can be used for the generation step as well, it does not explicitly assess multi-page reasoning. Future work could explore automated query generation that combines multiple pages using LLMs and/or VLMs to construct multi-page reasoning tasks, enhancing the benchmark's ability to evaluate complex retrieval scenarios.

REAL-MM-RAG-Bench provides a realistic, reliable, and challenging retrieval benchmark, helping to identify critical weaknesses in current multimodal retrieval models and paving the way for future improvements in both evaluation and model development.

References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. 2024. Pixtral 12b. *arXiv preprint arXiv:2410.07073*.

Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2016. Uqv100: A test collection with query variability. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 725–728.

Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2017. Retrieval consistency in the presence of query variations. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 395–404.

Rodger Benham, J Shane Culpepper, Luke Gallagher, Xiaolu Lu, and Joel M Mackenzie. 2018. Towards efficient and effective query variant generation. In *DESIRES*, pages 62–67.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. 2024. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yihao Ding, Kaixuan Ren, Jiabin Huang, Siwen Luo, and Soyeon Caren Han. 2024. Mvqa: A dataset for multimodal information retrieval in pdf-based visual question answering. *arXiv preprint arXiv:2404.12720*.

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*.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292.

Tim Hagen, Harrisen Scells, and Martin Potthast. 2024. Revisiting query variation robustness of transformer models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4283–4296.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

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.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Xiaolu Lu, Oren Kurland, J Shane Culpepper, Nick Craswell, and Ofri Rom. 2019. Relevance modeling with multiple query variations. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 27–34.

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, et al. 2024b. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *arXiv preprint arXiv:2407.01523*.

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

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

Gustavo Penha, Arthur Câmara, and Claudia Hauff. 2022. Evaluating the robustness of retrieval pipelines with query variation generators. In *European conference on information retrieval*, pages 397–412. Springer.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Rita Ramos, Desmond Elliott, and Bruno Martins. 2023. Retrieval-augmented image captioning. *arXiv preprint arXiv:2302.08268*.

S Robertson, Steve Walker, Susan Jones, and MHB GATFORD. 1994. Okapi at 3. In *Proceedings of the 3rd Text REtrieval Conference (-3)*, pages 109–126.

Georgios Sidiropoulos and Evangelos Kanoulas. 2022. Analysing the robustness of dual encoders for dense retrieval against misspellings. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2132–2136.

Alan F Smeaton and Fergus Kelledy. 1998. User-chosen phrases in interactive query formulation for information retrieval. In *20th Annual BCS-IRSG Colloquium on IR*. BCS Learning & Development.

Ray Smith. 2007. An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

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

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.

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

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

Xinyi Zheng, Doug Burdick, Lucian Popa, Peter Zhong, and Nancy Xin Ru Wang. 2021. Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context. *Winter Conference for Applications in Computer Vision (WACV)*.

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.

Hongyi Zhu, Jia-Hong Huang, Stevan Rudinac, and Evangelos Kanoulas. 2024. Enhancing interactive image retrieval with query rewriting using large language models and vision language models. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 978–987.

Guido Zuccon, Joao Palotti, and Allan Hanbury. 2016. Query variations and their effect on comparing information retrieval systems. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 691–700.

A Appendix

A.1 Models Evaluation

Beside the models from ColPali, we have evaluated on text based methods. Following the ColPali framework, we adopt *Unstructured* as our PDF parser in its high-resolution configuration, which relies on the Tesseract (Smith, 2007) OCR engine. Unstructured processes each document into two main types of chunks: text chunks and visual chunks (e.g., tables, figures, images). We then construct two text-based variants of our benchmark, differing in how visual chunks are converted into text:

1. **OCR**: Text is retained; tables, figures, and images undergo OCR extraction.
2. **Captioning**: Text remains unchanged; visual elements are described using Qwen2.5-VL-72B-Instruct.

Retrieval Methods. We evaluate two retrieval approaches:

- **Okapi BM25**: A sparse statistical baseline.
- **BGE-M3 (multi-vector)**: A state-of-the-art embedding model.

In line with ColPali, chunks are embedded and scored independently, and page-level scores are then obtained via maximum pooling across all chunks for a given page.

A.2 Data Generation, Filtering and Validation

To construct a high-quality benchmark, we applied a conservative filtering strategy that prioritizes precision over recall—favoring the exclusion of low-quality queries even at the cost of discarding valid ones. Setting a precise confidence threshold for VLMs/LLMs is non-trivial, so we experimented with multiple prompt templates for filtering. For each prompt, we sampled 150 queries, manually labeled them as good or bad, and selected the prompt that minimized the percentage of bad queries that passed the filter. Since the original pool of generated queries was large, our emphasis was on reducing false positives rather than preserving query volume.

We then manually reviewed queries that passed the filter and added representative errors as negative examples to the prompt context, improving robustness through iterative refinement. This process led to fewer recurring mistakes and was validated through human evaluation, where 85.0% of our queries were rated as good compared to 43.6% for ViDoRe.

Validating label correctness at scale is more challenging, as it would require exhaustively comparing each query to all documents. Instead, we qualitatively inspected queries with known answer pages to tune the labeling prompt. As shown in section 4, our benchmark yields significantly fewer false negatives than competing datasets.

We also conducted a false positive validation to assess label quality. Specifically, we randomly sampled 200 query–page pairs from each of the three benchmarks: ViDoRe, MMLong, and ours. Annotators were shown the page image alongside its corresponding query and asked whether the page contained an answer to the query. The results show 0% false positives in both our benchmark and ViDoRe, while MMLong exhibited a high false positive rate of 38.1%. These findings highlight the effectiveness of our benchmark generation pipeline in preventing false positive examples.

A.3 Proposed Training sets and Fine-tuning

Rephrasing Augmentation Training. We aimed to fine-tune a trained model with a short training phase to improve robustness to rephrasing. To achieve this, we created a rephrased training set based on the ColPali training data. Specifically, we used approximately half of the full training set (56k queries) and generated rephrased versions. The rephrasing was performed using LLaMA-3-70B, a different LLM than the one used for filtering and rephrasing in the benchmark.

Each query was randomly rephrased using one of three rephrasing levels (different prompts, see Fig. S6). To ensure semantic consistency, we used a secondary LLM verification step: the original query and its corresponding answer were fed alongside the rephrased query, and if the meaning was not preserved, the original query was retained in the rephrased training set.

For fine-tuning, we used the full ColPali training set, incorporating the rephrased queries in approximately half of it. The model was trained for one epoch with configurations similar to those used in the original ColPali/ColQwen training.

Table and Finance-Focused Training. As we observed that current models performed significantly worse on our table-heavy finance benchmark, we curated a dedicated training set to address this gap. We leveraged the publicly available FinTabNet dataset, a large-scale resource designed for financial table recognition and structure extraction. FinTabNet consists of pages from annual reports of S&P 500 companies, featuring complex tables.

We used the page images from FinTabNet to generate queries and answers using our automated pipeline, which includes a filtering process. Additionally, we explored using Qwen2-VL-72B for generating this training data, with results reported in Table S7. This process resulted in approximately 46k triplets of page images, queries, and answers.

For fine-tuning, we trained both ColPali and ColQwen on our table-focused training data, combined with the original ColPali training set, for one epoch, producing the TabCol models. For RobTab models, we incorporated LLaMA-3-70B for rephrasing, where half of the newly generated training set was randomly rephrased using three different rephrasing levels. These models were then fine-tuned for one epoch, together with the ColPali rephrased dataset.

All fine-tuning procedures followed the same configurations as the official ColPali training pipeline, utilizing ColBERT in-batch loss. The specific configurations for ColPali and ColQwen can be found at GitHub. Each model was fine-tuned starting from its respective base version: ColPali from ColPali-1.2 and ColQwen from ColQwen2-1.0. We used a batch size of 64 per GPU, resulting in a total effective batch size of 256. The training maintained the same initial learning rate, warmup steps, learning rate schedule, and LoRA configuration as the original pipeline. All training runs were conducted on four A100 80GB GPUs, with each fine-tuning session taking approximately three hours to complete.

A.4 Benchmark Document Examples

Our benchmarks include a diverse set of pages containing different types of information, including full-text pages, table-heavy pages, and slides with both tables and other visual elements. Table S1 presents the benchmark statistics for the different datasets. Additionally, for each page, we provide four types of queries: the original VLM-generated query and three levels of rephrasing. Examples of pages along with all query versions are shown in Figs. S1 and S2. In Fig. S3, we present examples from previous benchmarks, highlighting that many of their queries are not well-suited for RAG, as they often reference specific pages (i.e., QA-style queries) rather than general information-seeking queries.

A.5 Additional Results and Ablations

In our main paper, we focused on reporting the NDCG@5 metric for a subset of models and benchmarks across different rephrasing levels. In Tables S2 to S4, we provide the full results for Recall@1, Recall@5, and NDCG@5. Additionally, Table S5 presents model performance across different evidence source types. A Vision-Language Model (VLM), Pixtral-12B, was used to classify each query based on the type of evidence source from which the answer was retrieved on the corresponding page. We analyze performance across three different evidence types: Text, Table, and Visual. The reported results show NDCG@5 scores on the non-rephrased version, averaged across our four benchmarks. These results highlight a significant weakness in handling tables. However, after fine-tuning on the table-focused dataset, we observe improvements across all evidence types, with tables showing the most substantial gains.

We further provide two ablation experiments. The first, shown in Table S6, aims to demonstrate that the observed improvements after fine-tuning with our proposed datasets are due to the tailored data rather than fine-tuning itself. To verify this, we performed an additional fine-tuning run using the same training set as the RobCol models but without rephrasing, maintaining the exact number of training examples and identical training configurations.

As seen in the results, ColPali gains some improvement from fine-tuning alone, but the gains are significantly lower compared to fine-tuning with our rephrased training set. This gap is even more pronounced when evaluating on the rephrased benchmark (Level 3). For ColQwen, the baseline fine-tuning without rephrasing leads to a decrease in performance, whereas our fine-tuned models show substantial improvements on rephrased queries, as expected when training with our rephrased dataset.

The second ablation (see Table S7) aims to show that the improvements from the table-heavy training set are general and not specific to the Vision-Language Model (VLM) used for question generation. To test this, we generated an alternative version of the training set using a different VLM, Qwen2-VL-72B-Instruct, and trained models both with and without rephrasing (using LLaMA-3-3-70B). After filtering, the dataset size was slightly smaller—40k examples compared to 46k with Pixtral generation—likely due to Qwen generating more queries that were filtered out.

While results show a slight decrease in performance compared to using Pixtral-generated data, the fine-tuned models still significantly outperform the base ColPali and ColQwen models. This confirms that the effectiveness of our data is not limited to a specific VLM and that training on a table-heavy dataset remains highly beneficial.

A.6 Our Models' Results on ViDoRe V2

We report NDCG@5 performance on the newly released ViDoRe V2 benchmark, both before and after fine-tuning. As shown in Table S8, we compare base models (ColPali and ColQwen), models fine-tuned on the rephrased ColPali training set (Rob variants), and models fine-tuned with our table-focused and rephrased training set (RobTab variants).

Across 4 out 5 English benchmarks, we observe consistent improvements from fine-tuning, with particularly large gains for ColPali-based models. Among all methods, *RobTabColQwen* achieves the highest average NDCG@5 score, highlighting the effectiveness of our combined rephrasing and table-focused training strategy.

We further evaluate performance on the multilingual subset of ViDoRe V2 (Table S9). Here, fine-tuning also improves performance over the base models, with an even greater margin between *RobTabColQwen* and *ColQwen*. This suggests that rephrasing augmentation enhances the model's ability to capture semantic similarity beyond surface-level token overlap, which is especially beneficial for cross-lingual retrieval.

Overall, these results demonstrate the generalization ability of our fine-tuned models across both diverse content types and languages.

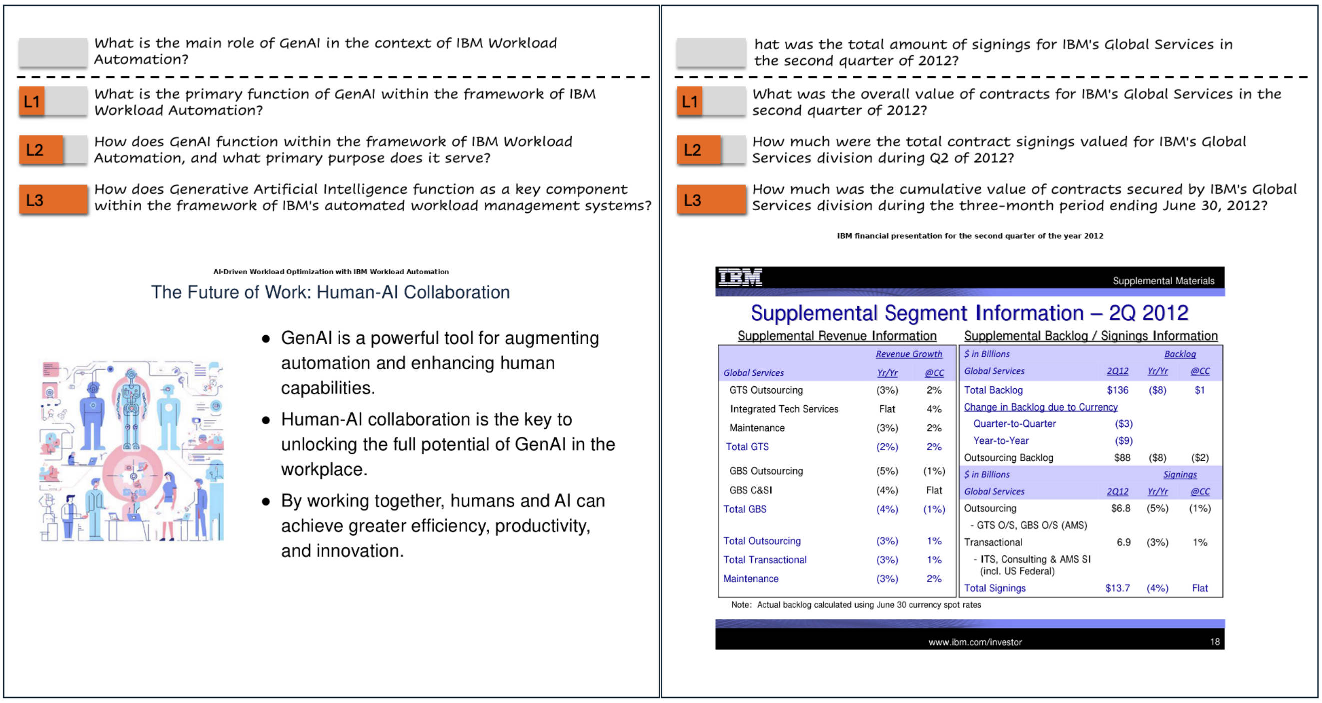
A.7 Licensing and Additional General Information

All models and datasets used in this work comply with their respective licenses. Qwen2-VL (ColQwen2) is licensed under Apache 2.0, with adapters under MIT. PaliGemma (ColPali) follows the Gemma license, with adapters under MIT. Pixtral-12B-2409 (mistralai) and Mixtral-8x22B are both under Apache 2.0, allowing unrestricted use, modification, and distribution. LLaMA 3.3 70B is licensed under the LLaMA 3.3 Community License Agreement. All datasets used are in English. The Colapli training set consists of subsampled academic datasets redistributed under their original licenses. It also includes synthetic datasets generated from publicly available internet data and VLM-generated queries, which are released without usage restrictions. Benchmark datasets are derived from openly available documents and images, with owner approval for publication. Fine-tuning data (Fintabnet) is collected from publicly available sources and processed for compatibility with our models.

For human evaluation, we published an online form alongside a request for participation in annotating queries for evaluating data intended for publication. Throughout the paper, an AI assistant (ChatGPT) was used for minor grammar and sentence structure edits.

Table S1: **Benchmark Statistics.**

Benchmark	Documents			Queries	Evidence Label		
	# Pages	# Docs	Avg. Len	# Queries	Text	Table	Visual
FinReport	2687	19	141	853	75%	24%	1%
FinSlides	2280	65	35	1052	12%	83%	5%
TechReport	1674	17	98	1294	81%	12%	7%
TechSlides	1963	62	32	1354	66%	6%	28%



What is the main role of GenAI in the context of IBM Workload Automation?

L1 What is the primary function of GenAI within the framework of IBM Workload Automation?

L2 How does GenAI function within the framework of IBM Workload Automation, and what primary purpose does it serve?

L3 How does Generative Artificial Intelligence function as a key component within the framework of IBM's automated workload management systems?

AI-Driven Workload Optimization with IBM Workload Automation

The Future of Work: Human-AI Collaboration

- GenAI is a powerful tool for augmenting automation and enhancing human capabilities.
- Human-AI collaboration is the key to unlocking the full potential of GenAI in the workplace.
- By working together, humans and AI can achieve greater efficiency, productivity, and innovation.

hat was the total amount of signings for IBM's Global Services in the second quarter of 2012?

L1 What was the overall value of contracts for IBM's Global Services in the second quarter of 2012?

L2 How much were the total contract signings valued for IBM's Global Services division during Q2 of 2012?

L3 How much was the cumulative value of contracts secured by IBM's Global Services division during the three-month period ending June 30, 2012?

IBM financial presentation for the second quarter of the year 2012

IBM Supplemental Materials

Supplemental Segment Information – 2Q 2012

Supplemental Revenue Information

	Revenue Growth	
Global Services	Yr/Yr	@CC
GTS Outsourcing	(3%)	2%
Integrated Tech Services	Flat	4%
Maintenance	(3%)	2%
Total GTS	(2%)	2%
GBS Outsourcing	(5%)	(1%)
GBS C&SI	(4%)	Flat
Total GBS	(4%)	(1%)
Total Outsourcing	(3%)	1%
Total Transactional	(3%)	1%
Maintenance	(3%)	2%

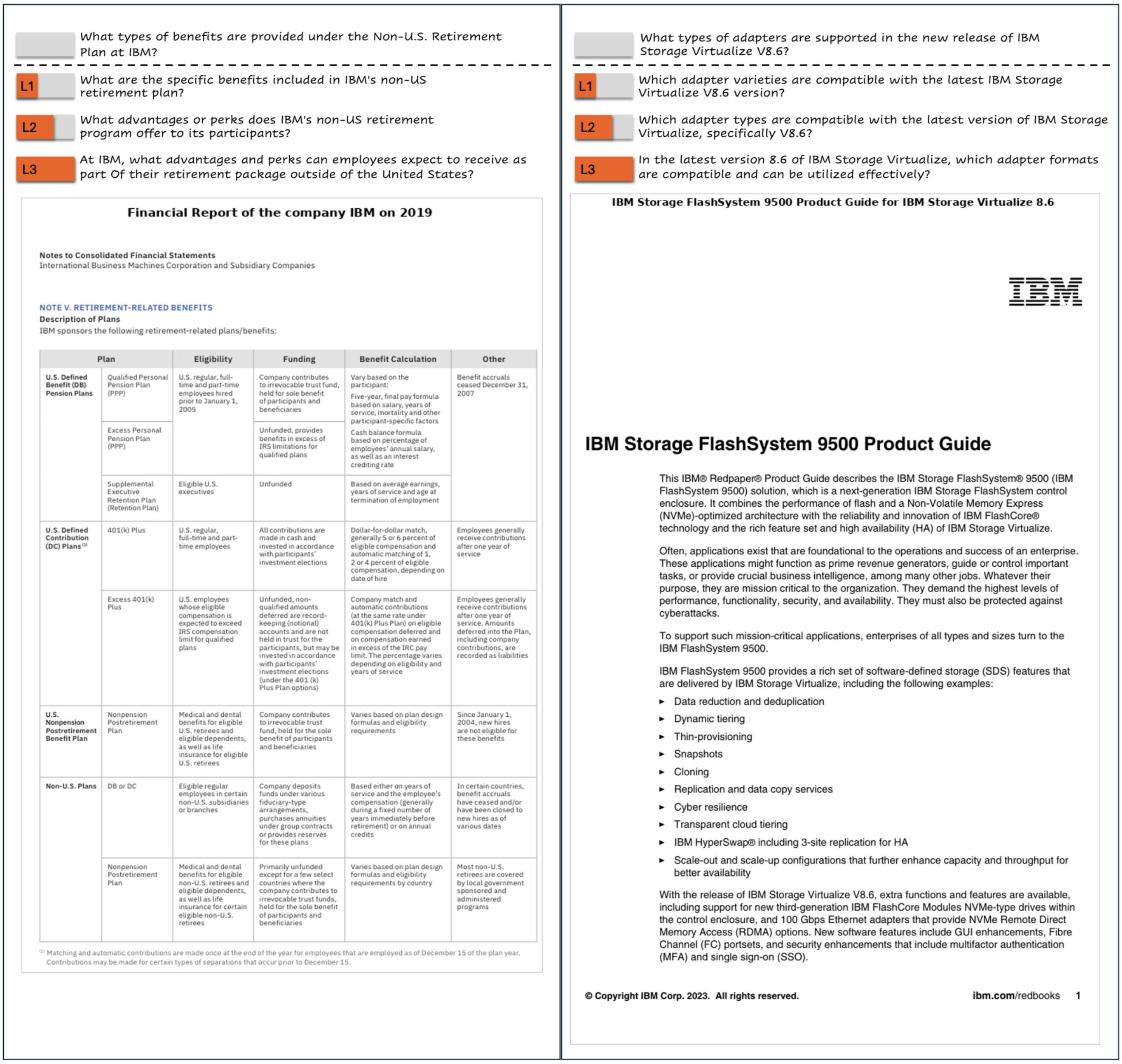
Supplemental Backlog / Signings Information

$ in Billions	Backlog		
Global Services	2Q12	Yr/Yr	@CC
Total Backlog	$136	($8)	$1
Change in Backlog due to Currency			
Quarter-to-Quarter	($3)		
Year-to-Year	($9)		
Outsourcing Backlog	$88	($8)	($2)
$ in Billions	Signings		
Global Services	2Q12	Yr/Yr	@CC
Outsourcing	$6.8	(5%)	(1%)
- GTS O/S, GBS O/S (AMS)			
Transactional	6.9	(3%)	1%
- ITS, Consulting & AMS SI (incl. US Federal)			
Total Signings	$13.7	(4%)	Flat

Note: Actual backlog calculated using June 30 currency spot rates

www.ibm.com/investor 18

Figure S1: **Real-MM-RAG Benchmark Examples with Rephrasing.** On the right: FinSlides—financial quarterly presentations. On the left: TechSlides Technical slides about business and IT automation. Questions are listed from the original query to Level 3 rephrasing.



What types of benefits are provided under the Non-U.S. Retirement Plan at IBM?

L1 What are the specific benefits included in IBM's non-US retirement plan?

L2 What advantages or perks does IBM's non-US retirement program offer to its participants?

L3 At IBM, what advantages and perks can employees expect to receive as part Of their retirement package outside of the United States?

Financial Report of the company IBM on 2019

Notes to Consolidated Financial Statements
International Business Machines Corporation and Subsidiary Companies

NOTE V. RETIREMENT-RELATED BENEFITS

Description of Plans

IBM sponsors the following retirement-related plans/benefits:

Plan		Eligibility	Funding	Benefit Calculation	Other
U.S. Defined Benefit (DB) Pension Plans	Qualified Personal Pension Plan (PPP)	U.S. regular, full-time and part-time employees hired prior to January 1, 2005	Company contributes to irrevocable trust fund, held for sole benefit of participants and beneficiaries	Vary based on the participant: Five-year, final pay formula based on salary, years of service, mortality and other participant-specific factors Cash balance formula based on percentage of employees' annual salary, as well as an interest crediting rate	Benefit accruals ceased December 31, 2007
	Excess Personal Pension Plan (PPP)		Unfunded, provides benefits in excess of IRS limitations for qualified plans		
	Supplemental Executive Retention Plan (Retention Plan)	Eligible U.S. executives	Unfunded	Based on average earnings, years of service and age at termination of employment	
U.S. Defined Contribution (DC) Plans [1]	401(k) Plus	U.S. regular, full-time and part-time employees	All contributions are made in cash and invested in accordance with participants' investment elections	Dollar-for-dollar match, generally 5 or 6 percent of eligible compensation and automatic matching of 1, 2 or 4 percent of eligible compensation, depending on date of hire	Employees generally receive contributions after one year of service
	Excess 401(k) Plus	U.S. employees whose eligible compensation is expected to exceed IRS compensation limit for qualified plans	Unfunded, non-qualified amounts deferred are record-keeping (notional) accounts and are not held in trust for the participants, but may be invested in accordance with participants' investment elections (under the 401 (k) Plus Plan options)	Company match and automatic contributions (at the same rate under 401(k) Plus Plan) on eligible compensation deferred and on compensation earned in excess of the IRC pay limit. The percentage varies depending on eligibility and years of service	Employees generally receive contributions after one year of service. Amounts deferred into the Plan, including company contributions, are recorded as liabilities
U.S. Nonpension Postretirement Benefit Plan	Nonpension Postretirement Plan	Medical and dental benefits for eligible U.S. retirees and eligible dependents, as well as life insurance for eligible U.S. retirees	Company contributes to irrevocable trust fund, held for the sole benefit of participants and beneficiaries	Varies based on plan design formulas and eligibility requirements	Since January 1, 2004, new hires are not eligible for these benefits
Non-U.S. Plans	DB or DC	Eligible regular employees in certain non-U.S. subsidiaries or branches	Company deposits funds under various fiduciary-type arrangements, purchases annuities under group contracts or provides reserves for these plans	Based either on years of service and the employee's compensation (generally during a fixed number of years immediately before retirement) or on annual credits	In certain countries, benefit accruals have ceased and/or have been closed to new hires as of various dates
	Nonpension Postretirement Plan	Medical and dental benefits for eligible non-U.S. retirees and eligible dependents, as well as life insurance for certain eligible non-U.S. retirees	Primarily unfunded except for a few select countries where the company contributes to irrevocable trust funds, held for the sole benefit of participants and beneficiaries	Varies based on plan design formulas and eligibility requirements by country	Most non-U.S. retirees are covered by local government sponsored and administered programs

[1] Matching and automatic contributions are made once at the end of the year for employees that are employed as of December 15 of the plan year. Contributions may be made for certain types of separations that occur prior to December 15.

What types of adapters are supported in the new release of IBM Storage Virtualize V8.6?

L1 Which adapter varieties are compatible with the latest IBM Storage Virtualize V8.6 version?

L2 Which adapter types are compatible with the latest version of IBM Storage Virtualize, specifically V8.6?

L3 In the latest version 8.6 of IBM Storage Virtualize, which adapter formats are compatible and can be utilized effectively?

IBM Storage FlashSystem 9500 Product Guide for IBM Storage Virtualize 8.6

IBM

IBM Storage FlashSystem 9500 Product Guide

This IBM® Redpaper® Product Guide describes the IBM Storage FlashSystem® 9500 (IBM FlashSystem 9500) solution, which is a next-generation IBM Storage FlashSystem control enclosure. It combines the performance of flash and a Non-Volatile Memory Express (NVMe)-optimized architecture with the reliability and innovation of IBM FlashCore® technology and the rich feature set and high availability (HA) of IBM Storage Virtualize.

Often, applications exist that are foundational to the operations and success of an enterprise. These applications might function as prime revenue generators, guide or control important tasks, or provide crucial business intelligence, among many other jobs. Whatever their purpose, they are mission critical to the organization. They demand the highest levels of performance, functionality, security, and availability. They must also be protected against cyberattacks.

To support such mission-critical applications, enterprises of all types and sizes turn to the IBM FlashSystem 9500.

IBM FlashSystem 9500 provides a rich set of software-defined storage (SDS) features that are delivered by IBM Storage Virtualize, including the following examples:

- Data reduction and deduplication
- Dynamic tiering
- Thin-provisioning
- Snapshots
- Cloning
- Replication and data copy services
- Cyber resilience
- Transparent cloud tiering
- IBM HyperSwap® including 3-site replication for HA
- Scale-out and scale-up configurations that further enhance capacity and throughput for better availability

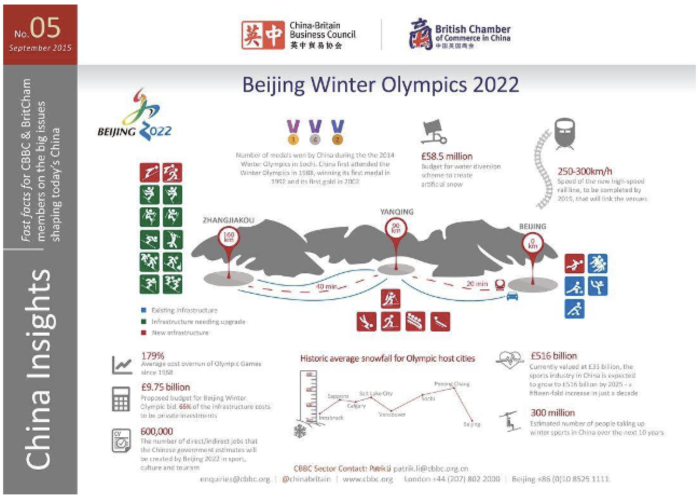
With the release of IBM Storage Virtualize V8.6, extra functions and features are available, including support for new third-generation IBM FlashCore Modules NVMe-type drives within the control enclosure, and 100 Gbps Ethernet adapters that provide NVMe Remote Direct Memory Access (RDMA) options. New software features include GUI enhancements, Fibre Channel (FC) portsets, and security enhancements that include multifactor authentication (MFA) and single sign-on (SSO).

© Copyright IBM Corp. 2023. All rights reserved.   ibm.com/redbooks   1

Figure S2: **Real-MM-RAG Benchmark Examples with Rephrasing.** Left: FinReport—financial annual reports. Right: TechReport—FlashSystem technical reports. Queries are listed from the original to Level 3 rephrasing.

ViDoRe / infovqa

How many infrastructure in this diagram are new?



Beijing Winter Olympics 2022

Number of medals won by China during the the 2014 Winter Olympics in Sochi. China first attended the Winter Olympics in 1980, winning its first medal in 1992 and its first gold in 2002

£58.5 million
Budget for water diversion scheme to create artificial snow

250-300km/h
Speed of the new high-speed rail line, to be completed by 2019, that will link the venues

■ Existing infrastructure
■ Infrastructure needing upgrade
■ New infrastructure

179%
Average cost overrun of Olympic Games since 1960

£9.75 billion
Proposed budget for Beijing Winter Olympic bid. 65% of the infrastructure costs to be private investments

600,000
The number of direct/indirect jobs that the Chinese government estimates will be created by Beijing 2022 in sport, culture and tourism

Historic average snowfall for Olympic host cities

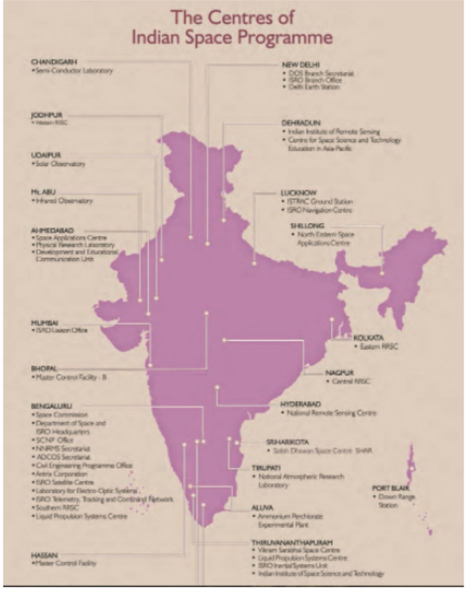
£516 billion
Currently valued at £33 billion, the sports industry in China is expected to grow to £516 billion by 2025 - a fifteen-fold increase in just a decade

300 million
Estimated number of people taking up winter sports in China over the next 10 years

CBBC Sector Contact: Patrik Li patrik.li@cbbc.org.cn
enquiries@cbbc.org | @chinabritain | www.cbbc.org London +44 (207) 802 2000 | Beijing +86 (0)10 8525 1111

MMLongBench

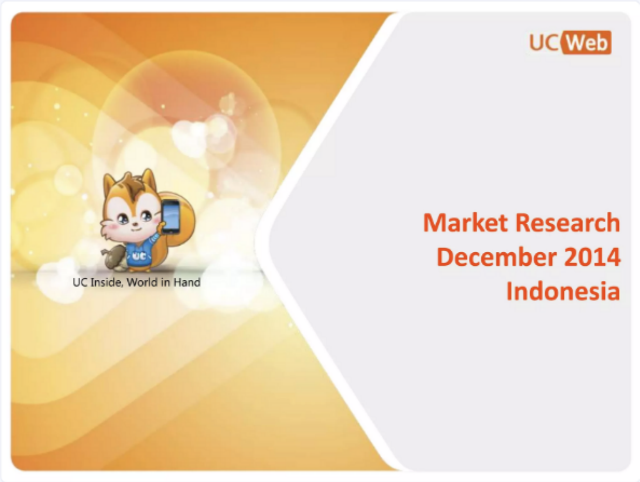
What does the map in the report shows?



The Centres of Indian Space Programme
CHANDIGARH
• Semi-Conductor Laboratory
JODHPUR
• Western RRSC
UDAIPUR
• Solar Observatory
Mt. ABU
• Infrared Observatory
AHMEDABAD
• Space Applications Centre
• Physical Research Laboratory
• Development and Educational Communication Unit
MUMBAI
• ISRO Liaison Office
BHOPAL
• Master Control Facility - B
BENGALURU
• Space Commission
• Department of Space and ISRO Headquarters
• SCNP Office
• NNRMS Secretariat
• ADCOS Secretariat
• Civil Engineering Programme Office
• Antrix Corporation
• ISRO Satellite Centre
• Laboratory for Electro-Optic Systems
• ISRO Telemetry, Tracking and Command Network
• Southern RRSC
• Liquid Propulsion Systems Centre
HASSAN
• Master Control Facility
NEW DELHI
• DOS Branch Secretariat
• ISRO Branch Office
• Delhi Earth Station
DEHRADUN
• Indian Institute of Remote Sensing
• Centre for Space Science and Technology Education in Asia-Pacific
LUCKNOW
• ISTRAC Ground Station
• ISRO Navigation Centre
SHILLONG
• North Eastern-Space Applications Centre
KOLKATA
• Eastern RRSC
NAGPUR
• Central RRSC
HYDERABAD
• National Remote Sensing Centre
SRIHARIKOTA
• Satish Dhawan Space Centre SHAR
TIRUPATI
• National Atmospheric Research Laboratory
PORT BLAIR
• Down Range Station
ALUVA
• Ammonium Perchlorate Experimental Plant
THIRUVANANTHAPURAM
• Vikram Sarabhai Space Centre
• Liquid Propulsion Systems Centre
• ISRO Inertial Systems Unit
• Indian Institute of Space Science and Technology

SlidesVQA

Where is this market research based on?



Market Research
December 2014
Indonesia

Figure S3: **Examples from Previous Benchmarks.** These examples illustrate common query types in these benchmarks. Many queries are generated for question answering and refer to a specific page rather than resembling real user queries, which are typically asked without prior knowledge of a specific page.

Table S2: **Impact of Rephrasing Levels on Document Retrieval Benchmarks.** This table shows NDCG@5 performance variations across rephrasing levels (0-3) for different benchmarks and models.

	FinReport				**FinSlides**				**TechReport**				**TechSlides**			
Rephrasing Level	**0**	**1**	**2**	**3**	**0**	**1**	**2**	**3**	**0**	**1**	**2**	**3**	**0**	**1**	**2**	**3**
BM25 (OCR)	48.8	38.4	26.6	21.7	13.6	13.0	7.1	5.9	66.6	48.0	38.7	35.1	58.7	45.7	35.7	31.2
BGE-M3 (OCR)	47.2	40.3	37.6	36.5	16.9	16.3	14.6	11.4	44.7	40.9	37.8	37.0	60.5	56.3	51.9	49.6
BM25 (Captioning)	54.4	43.0	30.6	25.3	20.9	19.0	11.0	9.9	69.0	50.8	41.8	37.2	66.4	53.4	41.7	36.1
BGE-M3 (Captioning)	46.4	39.0	36.9	35.9	19.5	18.7	16.3	13.8	44.6	40.9	38.4	37.5	62.6	58.3	54.5	51.7
ColPali	52.7	47.2	40.8	36.8	62.2	59.4	37.6	27.6	80.6	72.9	66.5	62.0	89.7	85.0	79.2	75.8
RobColPali	59.3	57.4	51.4	47.1	76.8	75.1	58.3	48.4	80.1	74.7	70.0	66.6	90.5	88.2	84.2	82.8
TabColPali	70.5	63.5	56.4	50.5	74.5	70.7	54.2	41.5	82.7	73.8	66.8	61.3	90.8	86.5	80.4	77.6
RobTabColPali	71.0	68.5	**65.0**	**63.2**	**80.9**	**79.5**	**68.0**	**58.3**	80.8	**76.2**	**72.6**	**70.7**	90.5	87.1	**84.3**	**83.3**
ColQwen	60.8	54.5	46.7	41.8	59.3	54.8	39.1	31.1	**84.2**	74.9	71.8	66.9	91.3	85.9	80.6	78.1
RobColQwen	58.4	54.5	49.7	47.5	65.8	63.0	52.0	44.3	81.9	75.4	71.8	69.5	90.1	**87.8**	84.0	83.0
TabColQwen	**78.2**	**69.0**	61.5	54.0	77.1	73.9	58.1	49.6	83.6	75.1	70.8	65.9	**92.4**	87.7	82.5	78.9
RobTabColQwen	**79.7**	**74.8**	**69.4**	**67.1**	**79.6**	**78.5**	**69.1**	**61.6**	**83.7**	**79.3**	**75.5**	**73.2**	**92.5**	**89.9**	**86.3**	**85.0**

Table S3: **Impact of Rephrasing Levels on Document Retrieval Benchmarks (Recall@1).** This table shows Recall@1 performance variations across rephrasing levels (0-3) for different benchmarks and models.

	FinReport				**FinSlides**				**TechReport**				**TechSlides**			
Rephrasing Level	**0**	**1**	**2**	**3**	**0**	**1**	**2**	**3**	**0**	**1**	**2**	**3**	**0**	**1**	**2**	**3**
BM25 (OCR)	33.9	25.4	17.1	14.0	7.7	7.5	3.7	3.0	53.4	34.1	26.1	23.1	45.9	33.4	25.4	21.3
BGE-M3 (OCR)	34.0	28.6	25.8	25.0	11.0	9.5	9.3	7.2	34.4	31.2	28.6	27.8	48.6	44.2	40.0	37.7
BM25 (Captioning)	40.1	28.0	19.8	17.5	12.2	10.9	6.2	5.7	56.2	36.8	28.8	24.6	54.0	40.8	31.0	25.2
BGE-M3 (Captioning)	33.4	27.7	25.1	25.3	12.5	11.5	9.6	8.9	34.7	31.1	28.8	27.7	50.9	45.8	41.8	40.1
ColPali	40.3	35.4	29.1	25.9	45.6	41.6	23.3	15.5	68.1	57.6	51.5	45.8	82.4	75.3	68.4	63.5
RobColPali	44.4	42.1	36.5	32.0	60.4	57.4	39.8	30.4	68.0	60.2	53.8	51.2	82.8	79.7	74.1	73.1
TabColPali	55.1	50.3	40.9	36.0	55.9	52.0	35.0	22.8	70.3	58.8	50.3	44.8	83.9	77.8	69.4	66.5
RobTabColPali	56.5	53.9	49.4	48.4	64.0	61.7	48.4	35.8	67.9	61.2	57.4	54.3	83.4	78.4	75.2	73.1
ColQwen	44.0	39.6	32.6	28.5	44.7	40.8	26.7	18.7	73.0	60.1	56.5	51.4	84.2	77.4	70.3	66.5
RobColQwen	41.4	37.7	34.8	33.1	49.1	46.6	37.2	29.3	68.5	60.4	56.1	54.2	83.3	79.5	74.6	72.1
TabColQwen	62.7	53.2	44.9	37.5	58.7	55.9	41.2	33.1	71.0	59.8	54.9	49.7	85.9	78.6	71.1	66.9
RobTabColQwen	58.1	52.5	50.2	45.8	62.2	60.7	51.4	41.8	70.3	62.2	59.0	56.3	85.0	80.9	76.0	74.9

Table S4: **Impact of Rephrasing Levels on Document Retrieval Benchmarks (Recall@5).** This table shows Recall@5 performance variations across rephrasing levels (0-3) for different benchmarks and models.

	FinReport				**FinSlides**				**TechReport**				**TechSlides**			
Rephrasing Level	**0**	**1**	**2**	**3**	**0**	**1**	**2**	**3**	**0**	**1**	**2**	**3**	**0**	**1**	**2**	**3**
BM25 (OCR)	62.0	49.9	35.3	29.0	19.7	18.6	10.4	8.7	77.6	59.8	50.0	45.9	70.3	57.1	45.2	40.6
BGE-M3 (OCR)	58.4	50.5	47.6	47.0	22.9	22.8	19.8	15.6	53.5	49.2	45.7	45.2	70.6	66.5	62.3	60.2
BM25 (Captioning)	66.9	56.3	40.4	32.5	29.4	27.2	15.5	13.8	79.7	63.5	53.5	49.0	77.2	64.8	51.3	46.1
BGE-M3 (Captioning)	57.0	48.8	46.7	45.5	25.9	25.1	22.7	18.3	52.9	49.3	46.6	46.2	72.6	68.8	65.7	62.0
ColPali	64.0	57.8	51.7	47.0	76.9	74.7	50.4	38.7	90.7	85.7	79.4	76.1	95.2	92.4	87.8	85.7
RobColPali	73.0	71.3	64.7	60.6	90.6	89.8	74.1	64.4	89.8	86.5	83.7	79.5	96.1	94.4	91.9	90.1
TabColPali	83.8	75.5	70.1	63.3	90.3	87.1	71.7	58.3	92.7	86.0	80.7	75.1	96.0	93.2	88.9	86.7
RobTabColPali	83.1	80.8	78.0	76.0	95.0	94.2	85.0	77.7	91.2	88.5	85.1	84.3	95.6	93.7	91.2	91.3
ColQwen	75.3	68.0	59.6	54.0	71.8	66.8	50.3	42.5	93.1	86.8	84.5	79.7	96.4	92.3	88.9	87.4
RobColQwen	73.0	69.5	63.0	60.7	79.8	76.5	65.6	57.8	92.6	87.9	84.8	82.7	95.1	94.1	91.3	91.4
TabColQwen	90.7	82.6	75.5	68.0	92.4	88.7	72.8	64.3	93.3	87.2	83.8	79.6	97.2	94.7	91.3	88.6
RobTabColQwen	86.3	80.1	76.8	74.4	92.5	92.1	82.0	76.2	92.6	88.8	86.4	84.7	96.9	95.3	93.3	92.6

Table S5: **Model Performance Across Different Evidence Source Types** A Vision-Language Model (VLM) was used to classify each query based on the type of evidence source from which the answer was retrieved on the corresponding page. We present the division of performance across three different source types: Text, Table, and Visual. The reported results are the NDCG@5 scores on the non-rephrased version, averaged across our four benchmarks.

Benchmark	**Text**	**Tables**	**Visual**
ColPali	75.8	58.6	84.5
ColQwen	79.2	59.9	86.8
TabColQwen	84.5	78.5	88.5

Table S6: **Ablation of Fine-Tuning Without Rephrasing.** To demonstrate that the performance improvement is not solely due to fine-tuning, we fine-tune the models on the original ColPali dataset without rephrasing, using the exact same fine-tuning configuration.

	FinReport		**FinSlides**		**TechReport**		**TechSlides**	
Rephrasing Level	**0**	**3**	**0**	**3**	**0**	**3**	**0**	**3**
ColPali	52.7	34.5	62.2	27.6	80.6	62.0	89.7	75.8
ColPali Baseline FT	57.1	39.1	69.1	35.7	80.0	61.4	90.0	77.8
RobColPali	59.3	47.1	76.8	48.4	80.1	66.6	90.5	83.3
ColQwen	60.8	41.8	59.3	31.1	84.2	66.9	91.3	78.1
ColQwen Baseline FT	58.1	39.3	53.4	27.6	79.7	62.1	90.5	77.4
RobColQwen	58.4	47.5	65.8	44.3	81.9	69.5	90.1	83.0

Table S7: **Comparison of different queries generation models.** This table compares the NDCG5 performance of the ColQwen model fine-tuned with the original data of ColPali and our generated table-focused data from the FinTabNet dataset. The evaluation is conducted for two query generation approaches: one using Pixtral and the other using Qwen. The rephrasing for the benchmarks was performed using LLaMA-3-3-70B. Results are presented across rephrasing levels (0 and 3) for our retrieval benchmarks.

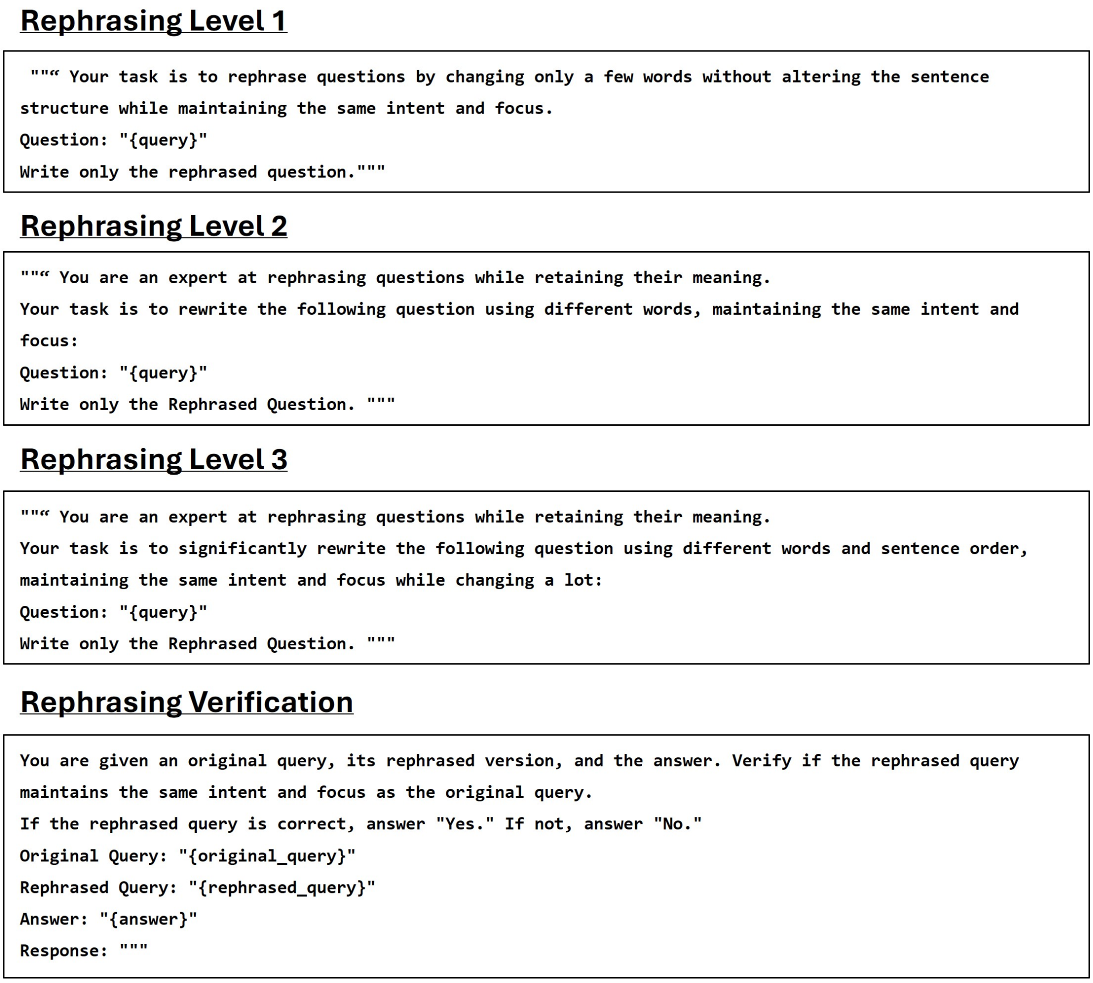
	FinReport		**FinSlides**		**TechReport**		**TechSlides**	
Rephrasing Level	**0**	**3**	**0**	**3**	**0**	**3**	**0**	**3**
ColPali	52.7	34.5	62.2	27.6	80.6	62.0	89.7	75.8
ColQwen	60.8	41.8	59.3	31.1	84.2	66.9	91.3	78.1
ColTab (Pixtral Queries Gen.)	78.2	54.0	77.1	49.6	83.6	65.9	92.4	78.9
ColTab (Qwen Queries Gen.)	74.8	49.5	74.3	41.5	83.8	66.8	92.6	79.5
ColRobTab (Pixtral Queries Gen.)	79.7	67.1	79.6	61.6	83.7	73.2	92.5	85.0
ColRobTab (Qwen Queries Gen.)	73.5	61.1	78.9	60.0	82.8	71.8	91.8	84.7

```
You are given a single page from a document. This page may contain text, figures, tables, and diagrams. The document is about
{Document_short_description}. Make sure to mention the company or product name when the question pertains to specific information about it.

Your task is to produce 10 question-answer pairs. Each question should represent a plausible inquiry that a person (who has not seen the page)
might ask about the information uniquely presented on this page. The questions should not reference this specific page directly (by page
number, pointing to a specific paragraph or figure, and never refer to the document using phrases like 'in the document'), nor should they
quote the text verbatim. They should use natural language reflecting how someone might inquire about the page’s content without direct access.

Please focus mainly on tables and figures rather than just text. Try to formulate questions that require data from multiple locations on the
page. The answer should be uniquely supported by the content of this page (i.e., the user could not find the answer elsewhere in the report or
infer it without seeing this particular page). Ensure that the questions are not too general. For example, refer to the company name instead of
using "the company" generically, as different documents may refer to different companies.

Good questions:
What types of properties does Wihlborgs specialize in?
How many women are in Wihlborgs' group management?
What was the occupancy rate of Wihlborg properties in 2021?
Was the net letting positive in the first quarter of 2022?
Which key factor contributed the most to the change in property value in 2022?
What's the difference in the number of properties in South Copenhagen between 2021 and 2022?
What is the proportion of women on the board of directors?
How many dividends were paid in 2020 in total equity?
What portion of Wihlborg’s sponsorship in 2018 was community-oriented?

Bad questions:
What is the title of the page?
What is the main purpose of this page?
What can you learn from the figure?
What is the name of the CEO mentioned in the document?
What is IBM's clear path to growth according to the document?
What is the basic earnings per share for 2023?

Make sure to include references to the company and product names when needed.

Output format:
Return as a list of objects:
[
    { "query": "...", "answer": "..." },
    { "query": "...", "answer": "..." },
    { "query": "...", "answer": "..." },
    { "query": "...", "answer": "..." },
    { "query": "...", "answer": "..." }
]
```

Figure S4: **Query Generation Prompt**

```
You have a large collection of document pages. Your goal is to determine, for each incoming question, whether that question would be
suitable for retrieving a specific page from this collection.
Classify each question into one of two categories:

Category #A:
A plausible inquiry about the unique information on a page, posed as though the reader does not have direct access to that page.
The question does NOT explicitly reference the page number, figure label, or phrases like “in the document” or “on page X.”
Instead, it uses natural language to reflect how someone would genuinely ask about the content.

Category #B:
Any question that explicitly references the page, figure, table, or paragraph (e.g., “on page 3,” “in figure 2,” “in the document”).
Or is too vague (e.g., “What is the name of the company?” without context).
Or otherwise, does not provide enough context to be meaningfully answered from a single page.

Below are examples for both categories:
“Which climate factors had the greatest impact on ABC Farming’s wheat yield in 2022?” # A
"How are topic scores calculated for each topical subset in the processing step S265?" # B
“How many corporate partners joined Apple Inc.’s technology incubator program last year?” # A
“Can you show me the figure on page 12 that references the new regulation?” # B
“Which strategic moves led to a 15% revenue increase for Toyota in Q2 2022?” # A
“List the bullet points mentioned in the second paragraph.” # B
“What percentage of employees in Google’s marketing department are working remotely?” # A
“What is the heading of Table 2 in the document?” # B
“How has the demand for Tesla’s electric vehicles changed over the past five years, based on the company’s data?” # A
“Who is labeled as the second co-author on page 7?” # B
“By how many degrees did the average global temperature rise according to NASA’s 2020-2022 research?” # A
“Which lines in the document discuss the formula for carbon emissions?” # B
“Why did the organization decide to shift to renewable energy solutions?” # B (No specific organization)
“Which historical events contributed most to the British government’s policy reforms in 1985?” # A
“How many employees does the company have?” # B (Too vague; no company mentioned)
“Did Walmart notice any major changes in consumer spending habits in Q4 2021?” # A
“What is the official job title of the lead manager in charge?” # B (Unclear which manager or context)
“Which country introduced new tax incentives for renewable energy projects in 2022, according to the IEA’s annual report?” # A
“Could you summarize the entire document for me?” # B (Overly broad; not tied to specific info)

When I give you a question, you should respond with ONLY 'A' or 'B'.
```

Figure S5: **Query Verification Prompt**



Rephrasing Level 1

```
 ""“ Your task is to rephrase questions by changing only a few words without altering the sentence
structure while maintaining the same intent and focus.
Question: "{query}"
Write only the rephrased question."""
```

Rephrasing Level 2

```
""“ You are an expert at rephrasing questions while retaining their meaning.
Your task is to rewrite the following question using different words, maintaining the same intent and
focus:
Question: "{query}"
Write only the Rephrased Question. """
```

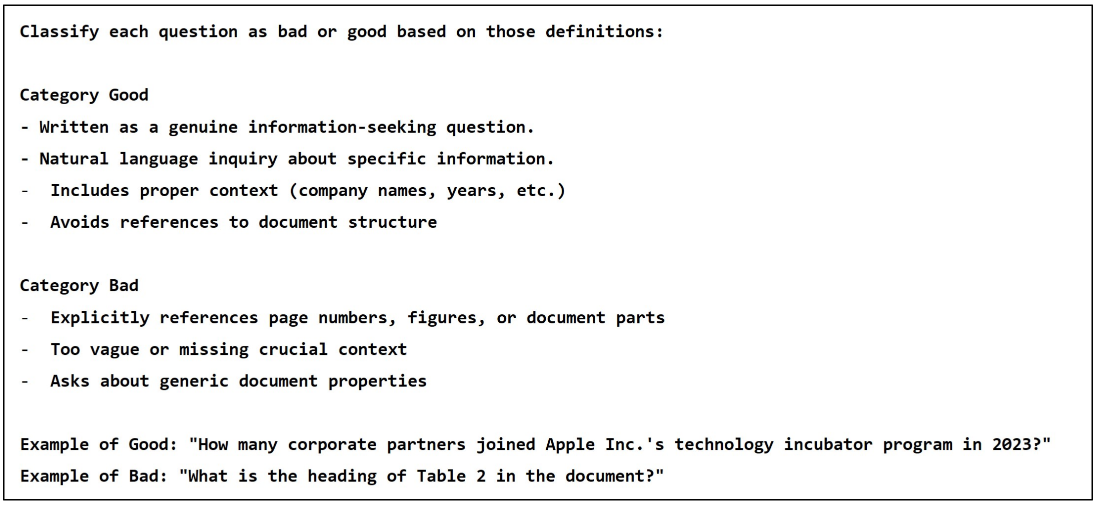
Rephrasing Level 3

```
""“ You are an expert at rephrasing questions while retaining their meaning.
Your task is to significantly rewrite the following question using different words and sentence order,
maintaining the same intent and focus while changing a lot:
Question: "{query}"
Write only the Rephrased Question. """
```

Rephrasing Verification

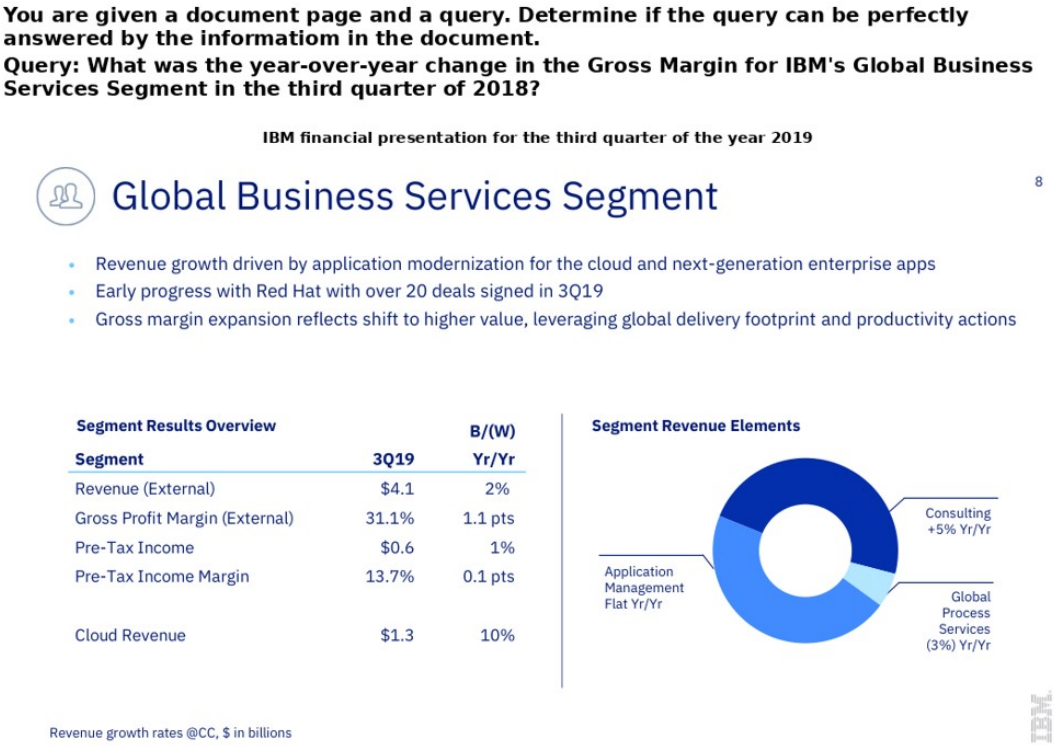
```
You are given an original query, its rephrased version, and the answer. Verify if the rephrased query
maintains the same intent and focus as the original query.
If the rephrased query is correct, answer "Yes." If not, answer "No."
Original Query: "{original_query}"
Rephrased Query: "{rephrased_query}"
Answer: "{answer}"
Response: """
```

Figure S6: **Rephrasing Generation and Verification Prompts**



```
Classify each question as bad or good based on those definitions:

Category Good
- Written as a genuine information-seeking question.
- Natural language inquiry about specific information.
-  Includes proper context (company names, years, etc.)
-  Avoids references to document structure

Category Bad
-  Explicitly references page numbers, figures, or document parts
-  Too vague or missing crucial context
-  Asks about generic document properties

Example of Good: "How many corporate partners joined Apple Inc.'s technology incubator program in 2023?"
Example of Bad: "What is the heading of Table 2 in the document?"
```

Figure S7: **Human Evaluation of Query Alignment to RAG.** This figure shows the instructions presented to human annotators along with randomly sampled queries from different benchmarks.



You are given a document page and a query. Determine if the query can be perfectly answered by the informatiom in the document.

Query: What was the year-over-year change in the Gross Margin for IBM's Global Business Services Segment in the third quarter of 2018?

IBM financial presentation for the third quarter of the year 2019

Global Business Services Segment

8

- Revenue growth driven by application modernization for the cloud and next-generation enterprise apps
- Early progress with Red Hat with over 20 deals signed in 3Q19
- Gross margin expansion reflects shift to higher value, leveraging global delivery footprint and productivity actions

Segment Results Overview

Segment	**3Q19**	**B/(W) Yr/Yr**
Revenue (External)	$4.1	2%
Gross Profit Margin (External)	31.1%	1.1 pts
Pre-Tax Income	$0.6	1%
Pre-Tax Income Margin	13.7%	0.1 pts
Cloud Revenue	$1.3	10%

Segment Revenue Elements

Revenue growth rates @CC, $ in billions

Figure S8: **Human Evaluation of False Negatives.** This figure presents an example of the image shown to human annotators, including the instructions, the query, and the negative page retrieved for the given query using ColQwen. The query and page are from our FinSlides benchmark and illustrate a case where the model incorrectly retrieved the wrong year.

Table S8: **ViDoRe V2 Results (English Benchmarks).** This table presents NDCG@5 scores for each domain and the overall average on the newly released ViDoRe V2 benchmark.

Model	Biomedical	Macroeconomics	Restaurant (RSE)	Restaurant (ESG)	Insurance (AXA)	Avg
ColPali	59.1	51.4	50.0	57.9	53.2	54.3
RobColPali	59.5	57.5	58.4	66.5	65.0	61.4
RobTabColPali	61.0	60.3	55.3	64.6	60.0	60.2
ColQwen	**61.1**	**61.5**	55.8	**63.0**	65.3	61.3
RobColQwen	61.0	**57.0**	**60.6**	62.0	**67.2**	**61.6**
RobTabColQwen	**62.1**	56.5	**61.7**	**66.0**	**69.7**	**63.2**

Table S9: **ViDoRe V2 Results (Multilingual Benchmarks).** This table presents NDCG@5 scores for each domain and the overall average on the multilingual subset of the newly released ViDoRe V2 benchmark.

Model	Biomedical	Macroeconomics	Restaurant (RSE)	Insurance (AXA)	Avg
ColPali	55.9	47.2	52.5	47.4	50.8
RobColPali	58.4	53.2	57.9	57.5	56.8
RobTabColPali	57.7	**55.0**	54.8	54.4	55.5
ColQwen	56.3	52.8	56.9	56.5	55.6
RobColQwen	**58.3**	53.0	**60.6**	**59.1**	**57.8**
RobTabColQwen	**59.1**	**55.1**	**60.2**	**61.8**	**59.1**