FinRAGBench-V: A Benchmark for Multimodal RAG with Visual Citation in the Financial Domain

Suifeng Zhao[1], **Zhuoran Jin**[2], **Sujian Li**[3*], **Jun Gao**[1*]
[1]Key Laboratory of High Confidence Software Technologies, CS, Peking University, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]State Key Laboratory of Multimedia Information Processing, CS, Peking University
sfzhao25@stu.pku.edu.cn, zhuoran.jin@nlpr.ia.ac.cn,
{lisujian,gaojun}@pku.edu.cn

🤗 Dataset: https://huggingface.co/datasets/zhaosuifeng/FinRAGBench-V
Code: https://github.com/zhaosuifeng/FinRAGBench-V

Abstract

Retrieval-Augmented Generation (RAG) plays a vital role in the financial domain, powering applications such as real-time market analysis, trend forecasting, and interest rate computation. However, most existing RAG research in finance focuses predominantly on textual data, overlooking the rich visual content in financial documents, resulting in the loss of key analytical insights. To bridge this gap, we present **FinRAGBench-V**, a comprehensive visual RAG benchmark tailored for finance which effectively integrates multimodal data and provides visual citation to ensure traceability. It includes a bilingual retrieval corpus with 60,780 Chinese and 51,219 English pages, along with a high-quality, human-annotated question-answering (QA) dataset spanning heterogeneous data types and seven question categories. Moreover, we introduce **RGenCite**, an RAG baseline that seamlessly integrates visual citation with generation. Furthermore, we propose an **automatic citation evaluation method** to systematically assess the visual citation capabilities of Multimodal Large Language Models (MLLMs). Extensive experiments on RGenCite underscore the challenging nature of FinRAGBench-V, providing valuable insights for the development of multimodal RAG systems in finance.

1 Introduction

Retrieval-Augmented Generation (RAG) (Izacard et al., 2023; Guu et al., 2020; Yu et al., 2024b) has become a crucial approach for enhancing the performance of Large Language Models (LLMs) by integrating external knowledge with internal knowledge (Yang et al., 2024; Han et al., 2024; Zhang et al., 2024a). This approach has been applied in a wide range of domain-specific tasks, among which, the financial domain is particularly representative due to its heavy reliance on complex multimodal data, such as line charts showing price fluctuations and tables presenting financial statistics. Therefore, it is critical to build a multimodal RAG system tailored to finance to enable reliable, explainable, and data-grounded analysis.

However, existing financial RAG efforts, such as FinQA (Chen et al., 2021) and OmniEval (Wang et al., 2024c), predominantly focus on text-only RAG, which may lose critical information when converting multimodal documents into plain text. As a result, they frequently fail to answer questions accurately, as shown in Figure 1 (a). Although MME-Finance (Gan et al., 2024) introduces a multimodal reasoning benchmark, it relies mostly on isolated screenshots and lacks retrieval support. Consequently, it falls short of reflecting the complexity of real-world financial scenarios, where answering questions often requires diverse data sources and heterogeneous data types. Furthermore, given the critical importance of precision in finance, RAG systems must ensure not only accuracy responses but also their traceability and verifiability, yet most existing benchmarks overlook these needs. Thus, designing a more comprehensive benchmark for multimodal RAG in finance is imperative.

In this work, we propose **FinRAGBench-V**, a multimodal RAG benchmark tailored for finance, featuring grounded visual citation. This benchmark effectively integrates multimodal data and provides visual citations to ensure traceability, as shown in Figure 1 (b). Specifically, we construct a large-scale retrieval corpus from diverse real-world financial sources, comprising 60,780 Chinese pages from 1,104 documents and 51,219 English pages from 1,105 documents, including research reports, financial statements, prospectuses, etc. In addition, we develop a high-quality financial question-answering (QA) dataset using GPT-4o (OpenAI, 2023) assistance with manual verification. The dataset consists of 855 Chinese and 539 English QA pairs, covering a wide range of distinctive fi-

*Corresponding authors

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 4215–4249
November 4-9, 2025

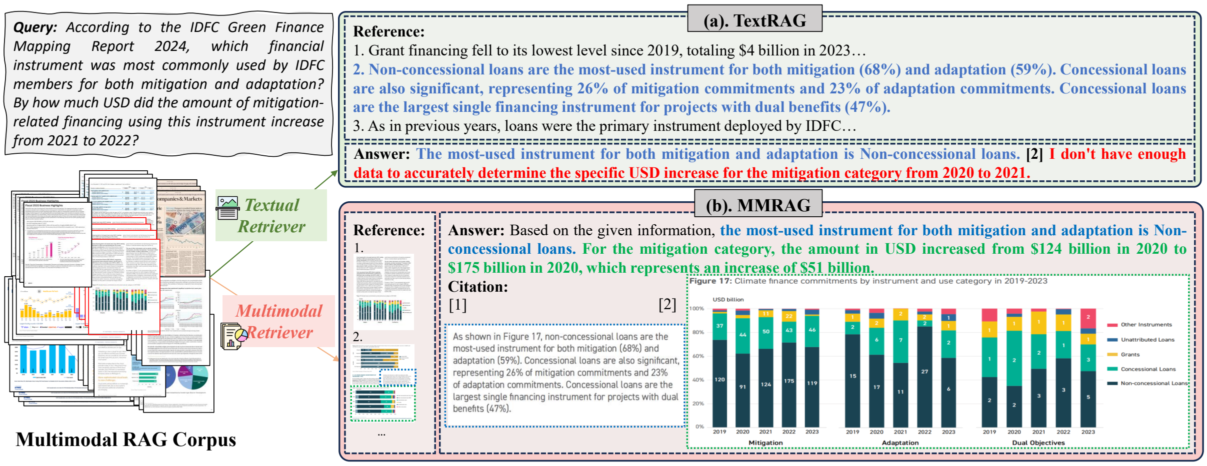


Figure 1: An example of a financial question requiring both text and visual understanding. (a) shows text-only RAG, where information loss leads to insufficient data for the model to answer the question. (b) illustrates our proposed paradigm, in which the model not only answers correctly based on retrieved information but also provides appropriate visual citations.

nancial tasks, with questions categorized by data heterogeneity, including text, charts, and tables, and reasoning type, such as time-sensitive reasoning, numerical calculations, multi-page reasoning, etc.

Based on this benchmark, we propose **RGenCite**, a simple yet effective multimodal RAG baseline that integrates retrieval, generation, and visual citation in a unified pipeline. The model is tasked with not only generating answers from retrieved contexts but also performing visual citation towards relevant document pages and specific content blocks, producing citations at both the page and block levels. To implement this, we adapt and migrate the method proposed by Ma et al. (2024b) to the multimodal RAG context to enable fine-grained block-level citation.

Although evaluation metrics for retrieval and generation are well-established, visual citation, as a novel application within RAG, still lacks dedicated evaluation methodologies. To address this gap, we propose an **automatic evaluation method for visual citation**. Specifically, we define the evaluation metrics, precision and recall, at both the page-level and block-level, and introduce two evaluation strategies: box-bounding and image-cropping.

We conduct extensive experiments and evaluations on FinRAGBench-V. For retrieval, we conduct experiments using four textual retrievers, such as Jina-ColBERT-V2 (Jha et al., 2024), and five Multilingual-E5-large (Wang et al., 2024a); and multimodal ones, such as ColQwen2 (Faysse et al., 2024), GME-Qwen2-VL-2B (Zhang et al., 2024b), and DSE-QWen2-2b-MRL-V1 (Ma et al., 2024a). For generation and citation, we employ seven proprietary Multimodal Large Language Models (MLLMs), such as GPT-4o (OpenAI, 2023), GPT-4V, and Gemini-2.0-Flash (Comanici et al., 2025), and six open-source ones, such as Qwen2.5-VL-72B-Instruct Wang et al. (2024b) and MiniCPM-o-2.6 (Yao et al., 2024).

Through the experiments, we derive several meaningful observations: (1) Multimodal retrievers outperform text-only ones by preserving information from charts and tables, avoiding information loss. (2) Current MLLMs handle text inference well but struggle with numerical reasoning on charts, tables, and multi-page inferences. (3) Multimodal RAG systems excel at page-level citation but struggle with block-level citation, highlighting challenges in precise attribution.

In summary, our contributions are as follows:

- We construct FinRAGBench-V, a benchmark for visual RAG in the financial domain, featuring diverse real-world data sources for retrieval, a wide range of question types for generation, and visual citation for attribution.

- We propose RGenCite, a comprehensive multimodal RAG baseline that combines retrieval, generation, and fine-grained visual citation. The model is required not only to generate answers from retrieved content, but also to provide page- and block-level visual citations as supporting evidence.

Benchmark	Domain	RAG Corpus	Multimodal	Multi-Task	Multi-Page	Citation
FinQA (Chen et al., 2021)	Finance	✗	✗	✗	✗	✗
OmniEval (Wang et al., 2024c)	Finance	✓	✗	✓	✗	✗
EvoChart (Huang et al., 2025)	General	✗	✗	✗	✗	✗
M3DocVQA (Cho et al., 2024)	General	✓	✓	✗	✓	✗
VisDoMBench (Suri et al., 2024)	General	✓	✓	✗	✗	✗
MME-Finance (Gan et al., 2024)	Finance	✗	✓	✓	✗	✗
FinRAGBench-V (Ours)	Finance	✓	✓	✓	✓	✓

Table 1: Comparison of our benchmark with existing benchmarks.

- We propose an automatic evaluation method for visual citation. The method incorporates precision and recall metrics for citations at different levels, with evaluation approaches including box-bounding and image-cropping.
- Extensive experiments reveal retriever differences, task-dependent model performance, and challenges in visual citation, validating FinRAGBench-V's value for evaluating multimodal RAG in finance.

2 Related Work

Benchmarking Multimodal RAG. Retrieval-Augmented Generation (RAG) has gained significant attention as an effective method of leveraging retrieval mechanisms to provide external knowledge to LLMs' generation (Gao et al., 2023b; Lewis et al., 2020; Huang et al., 2023; Chen et al., 2024b; Friel et al., 2024; Saad-Falcon et al., 2024). In the financial domain, where charts and graphs are essential, text-only RAG benchmarks often overlook critical information (Chen et al., 2021; Wang et al., 2024c), highlighting the need for a multimodal RAG benchmark. Recent efforts on financial multimodal benchmarks exhibit several limitations, as summarized in Table 1. EvoChart (Huang et al., 2025) focuses solely on chart-based questions, lacking integration with textual and tabular information. Cho et al. (2024) and Suri et al. (2024) utilize real-world PDFs but support only limited question types. MME-Finance (Gan et al., 2024) provides diverse financial questions, yet its reliance on isolated chart screenshots hinders document-level retrieval and fails to reflect the complexity of financial data.

Citation and Its Evaluation. Citations play a crucial role in enhancing the credibility and interpretability of RAG systems (Slobodkin et al., 2024; Li et al., 2023, 2024; Gao et al., 2023a). While prior works focus on textual citations, Ma et al. (2024b) introduce a coordinate-based method for multimodal citations. In specialized domains such as finance, where precise domain knowledge is essential, citation is particularly critical for RAG. Thus, we adapt this visual citation approach to the financial multimodal RAG setting and propose an automatic evaluation method for visual citation.

3 Task Definition

Our task contains two main phases: the construction of FinRAGBench-V, and the implementation of the RGenCite baseline, as shown in Figure 2.

In the first phase, given the raw documents collected from diverse sources, we first generate a retrieval corpus of pages, defined as $\mathcal{S} = \{p_1, p_2, ..., p_i, ...\}$, where p_i represents the ith page. Based on the corpus, we generate the QA dataset, defined as $\mathcal{D} = \{d_1, d_2, ..., d_i, ...\}$, where each $d_i = (q_i, a_i, t_i, P_i)$, with q_i being the question, a_i the ground truth answer, t_i the question type, and P_i the set of corresponding page(s). So far, we have constructed the retrieval corpus and QA dataset.

The second phase comprises both the retrieval stage and the generation with citation stage. Given a question q, a retriever R retrieves the top-k relevant pages $\{p_1, p_2, ..., p_k\}$ from the corpus $\mathcal{S}$. These pages, along with the question are then fed into a generator model M, which produces an answer a accompanied by a set of citations $C = \{c_1, c_2, ..., c_i\}$. Each citation $c_i = (p_i, B_i)$ consists of a cited page p_i and its corresponding supporting blocks $B_i = \{b_{i1}, b_{i2}, ..., b_{ij}\}$.

4 The Construction of FinRAGBench-V

As shown at the top of Figure 2, FinRAGBench-V consists of two components: a retrieval corpus and a QA dataset. This section outlines the construction process and provides detailed statistics.

4.1 Retrieval Corpus Collection

To build the retrieval corpus, we collect data from a variety of real-world financial document sources in

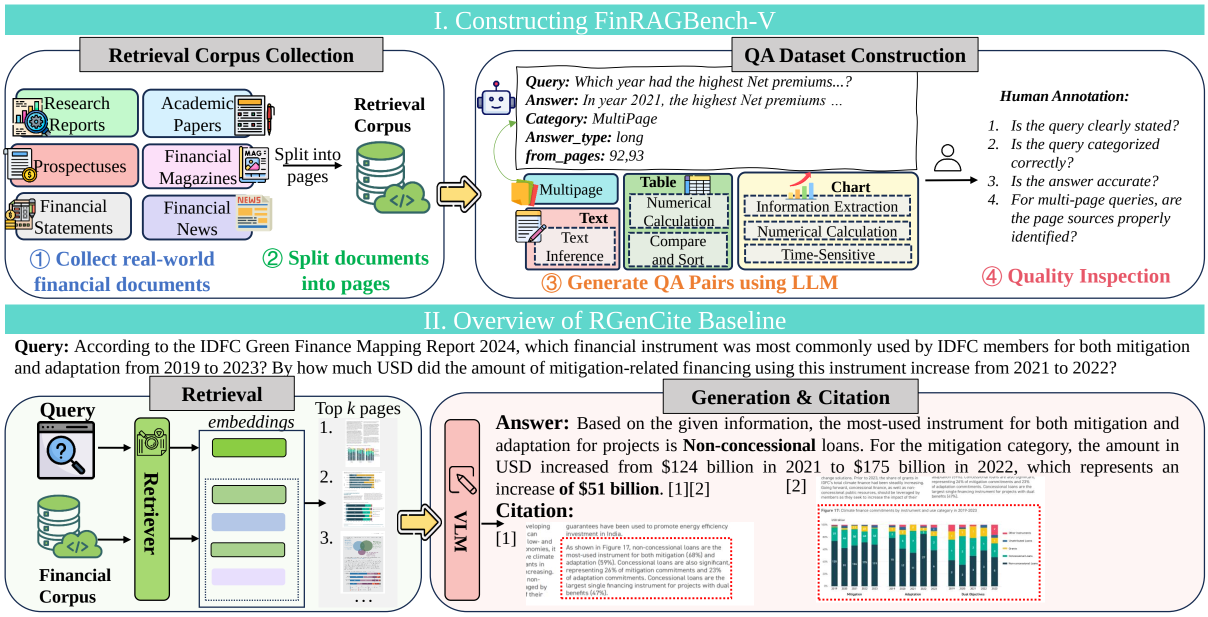


Figure 2: **I. Workflow of constructing FinRAGBench-V, including a retrieval corpus and a QA dataset**: ① collect real-world financial documents; ② split documents into pages; ③ generate data using LLM; ④ quality inspection. **II. Overview of RGenCite Baseline**: including the retrieval stage, and generation-citation stage.

both Chinese and English, as detailed in Appendix B, including:

(1) **Research reports** collected from websites like Qianzhan.com, which provide in-depth financial analyses, for example the analysis of price trends over time using line charts;

(2) **Financial statements of companies and banks** collected from the FinGLM [1]dataset and official company and bank websites, which provide annual financial data in tabular form;

(3) **Prospectuses** sourced from the BSCF [2] dataset, providing information on companies going public, including financial data and business strategies, with rich tabular information;

(4) **Academic papers** offering theoretical and empirical insights into financial markets, economic models, and financial technologies, sourced from Journal of Financial and CNKI;

(5) **Financial magazines** including respected outlets like the Financial Times, which offer reliable news, expert opinions, and financial analyses;

(6) **Financial news** from websites like China Daily and Eastmoney.

We finally select 1,104 Chinese and 1,105 English documents from the aforementioned data sources (details in Table 2). Each document page is converted into a single image, resulting in a retrieval corpus of 60,780 Chinese and 51,219 English pages. By incorporating these diverse data types, we ensure that the retrieval corpus is both broad and reliable, providing a solid foundation for generating accurate and informative QA pairs.

Data Source	Content Type	#Docs	#Pages	#Avg. Pages
Research Reports	Chart, Table, Text	219	8,583	52
Financial Statements	Table, Text	408	38,004	376
Prospectuses	Table, Text	41	539	13
Academic Papers	Chart, Table, Text	311	1,912	10
Financial Magazines	Chart, Text	191	9,958	131
Financial News	Chart, Table, Text	1,039	1,784	3

Table 2: Statistics of the corpus showing the types of document content, total document number, total pages, and average pages per document for each data source.

4.2 QA Dataset Construction

To construct the QA dataset, we follow a two-step process: first, we use a generator LLM to synthesize the QA pairs, and then conduct human annotation to ensure data quality.

4.2.1 QA Pairs Synthesis

From the retrieval corpus, we select high-quality document pages and then generate a dataset using GPT-4o based on these pages, with predefined categories and carefully designed examples provided as prompts (provided in Appendix A). In terms of data scope, the dataset includes both single-page

[1]https://tianchi.aliyun.com/competition/entrance/532164/introduction

[2]https://www.modelscope.cn/datasets/BJQW14B/bs_challenge_financial_14b_dataset/

and multi-page questions; Regarding data format, it covers text, charts, and tables; As for answers, it contains both short and long ones; Considering the characteristics of financial domain, we further categorize the QA dataset into seven main categories as follows. Appendix C shows some examples.

Text Inference: This involves tasks like information extraction and summarization, such as deriving key insights or identifying specific details (e.g., financial data or trends) from text.

Chart Information Extraction: This involves extracting key metrics or features from charts, such as the percentage of a sector in a pie chart.

Chart Numerical Calculations: This involves performing numerical calculations based on charts, such as calculating the changes of interest rate.

Chart Time-Sensitive Queries: This involves handling time-based chart queries, such as identifying event timings, analyzing trends, and pinpointing data peaks and troughs, often focusing on how indicators evolve over time.

Table Numerical Calculations: Similar to chart calculations, this involves performing numerical operations on table data, such as calculating interest rate changes and summing costs, to derive insights.

Table Comparison and Sorting: This involves comparing and sorting table data, such as comparing financial indicators between entities, ranking them, or identifying the highest or lowest values.

Multi-Page Queries: This involves queries requiring information from multiple pages, such as extracting truncated tables or combining data from multiple charts to answer a single query.

4.2.2 Quality Inspection

During the selection and annotation process, we adhere to several key principles to ensure the high quality and consistency of the dataset: examining the clarity of the questions and their correct categorization, verifying the accuracy of answers, and checking whether the page sources for multi-page queries are properly identified. The detailed annotation guideline is shown in Table 27 of Appendix F. Based on these criteria, we carefully filter and refine the original 11,328 generated QA pairs, and ultimately obtaining a total of 1,394 pairs, consisting of 855 Chinese entries and 539 English entries. The statistics of each category are shown in Figure 3, the lengths statistics of the dataset are shown in Table 3.

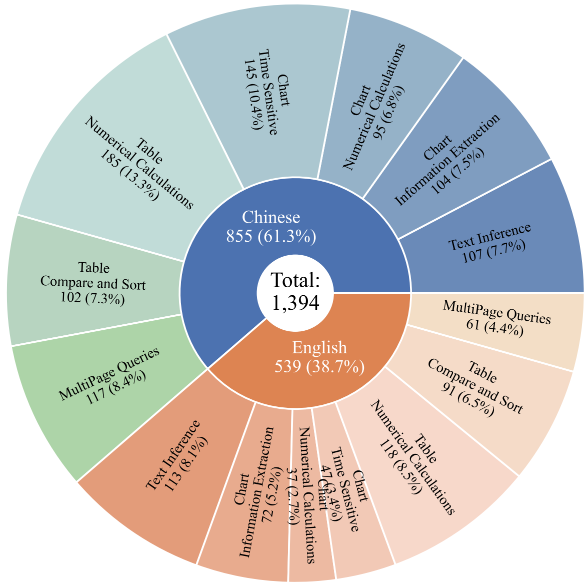


Figure 3: Statistics of Question Types in the Dataset.

Category	Question	Answer (Overall)	Short Answer	Long Answer
#Avg. Length	12.23	10.17	5.37	47.56

Table 3: Statistics of average token length of the dataset.

5 RGenCite: Retrieval, Generation, and Visual Citation

Based on our retrieval corpus and QA dataset, we develop the baseline system RGenCite, which covers both retrieval and generation, with visual citation seamlessly integrated into the generation stage, as illustrated at the bottom of Figure 2.

5.1 Retrieval

During the retrieval stage, given a query q, the retrievers aim to identify the top-k relevant pages $\{p_1, p_2, ..., p_k\}$ from the corpus S. We explore various multimodal and textual retrievers and conduct a comprehensive evaluation of these two retrieval paradigms using multiple metrics.

5.2 Generation with Visual Citation

During the generation stage, based on the retrieval result, the generator model M is tasked with producing textual answer a accompanied by visual citations C, given the query q. To enable the simultaneous generation of both answers and citations, we follow the visual citation method used in VISA (Ma et al., 2024b). Specifically, we input both the question q and the top-k relevant pages $\{p_1, p_2, ..., p_k\}$ into the generator M, instructing it to generate the answer a while simultaneously producing both

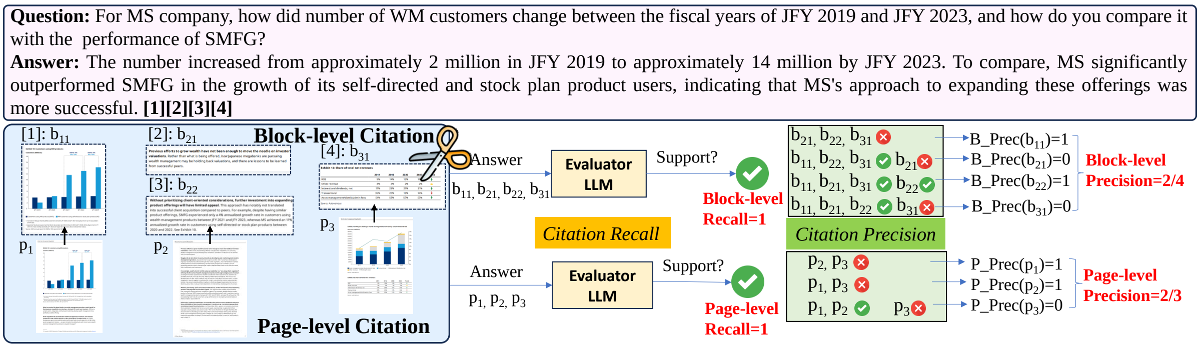


Figure 4: An example of the automatic evaluation of visual citation.

page-level and block-level citations. Each citation is denoted as $c_i = (p_i, \{b_{i1}, b_{i2}, ..., b_{ij}, ...\})$, where the page-level citation p_i refers to the reference page, $\{b_{i1}, b_{i2}, ..., b_{ij}, ...\}$ represents the block-level citations, indicating the specific regions of the answer within the page. Each block-level citation b_{ij} is represented as a set of coordinates, i.e., $b_{ij} = [x_1, y_1, x_2, y_2]$, where (x_1, y_1) and (x_2, y_2) denote the coordinates of the top-left corner and bottom-right corner of b_{ij}, respectively. Detailed output format is in Appendix A.

6 Evaluation Metrics

After implementation, we evaluate the RGenCite baseline from three perspectives: retrieval, generation, and visual citation, with citation quality assessed using our proposed evaluation method.

6.1 Retrieval Quality

To evaluate the performance of both multimodal and textual retrievers, we adopt several evaluation metrics, namely nDCG@k (for k = 5, 10), Recall@k (for k = 5, 10), and MRR@k (k = 10), which respectively capture ranking quality, retrieval coverage, and early relevance.

6.2 Answer Accuracy

To evaluate MLLMs' ability to generate accurate responses based on visual elements, we use the rule-based metric ROUGE. Additionally, we employ GPT-4o to assess the metric Acc, determining whether the generated responses align with the ground truths and are consistent with the visual context. The evaluation prompt is in Appendix A.

6.3 Citation Quality

To evaluate the visual citation quality of MLLMs, we introduce two automatic evaluation metrics: recall and precision. These metrics are applied at both the page-level and the block-level, using two distinct citation evaluation approaches: box-bounding and image-cropping. The effectiveness of our automatic citation evaluation methods is demonstrated in Section 7.3.

Citation Metrics. Inspired by Gao et al. (2023a), we evaluate both page-level and block-level citations using the following two metrics:

Recall evaluates whether the cited images are sufficient to support the answer. If the union of the citation set $C = \{c_1, c_2, ..., c_n\}$ of an answer a sufficiently support a, the recall is assigned 1; otherwise, it is assigned 0, defined in Equation 1:

$$\text{recall}(C, a) = \begin{cases} 1 & \text{if } \bigcup_{c_i \in C} c_i \text{ supports } a, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Precision evaluates the proportion of citations in the citation set C that are essential for supporting an answer. Specifically, the citation c_i is considered irrelevant if and only if c_i cannot independently support the answer, and the union of all other citations $\{c_1, c_2, ..., c_{i-1}, c_{i+1}, ...\}$ in C is sufficient to support the answer a, as described in Equation 2:

$$\text{irrel}(C, c_i, a) = (c_i \nrightarrow a) \land ((C \setminus \{c_i\}) \rightarrow a) \quad (2)$$

Thus, the citation precision of the citation set C for answer a is defined as the proportion of non-irrelevant citations in C, as shown in Equation 3:

$$\text{precision}(C, a) = \frac{|C \setminus \{c_i \mid \text{irrel}(C, c_i, a) = 1\}|}{|C|} \quad (3)$$

It should be noted that the precision of each citation is evaluated only when the recall of the citation set it belongs to is 1; otherwise, i is set to 0.

Retriever	Chinese					English				
	nDCG@5	nDCG@10	Recall@5	Recall@10	MRR@10	nDCG@5	nDCG@10	Recall@5	Recall@10	MRR@10
Multimodal Retrievers										
ColQwen2	**78.53**	**79.76**	**86.46**	**90.13**	**77.80**	**67.90**	**70.00**	**79.64**	**85.86**	**65.54**
GME-Qwen2-VL-7B	74.55	76.04	84.80	89.35	72.80	58.06	60.94	68.95	77.56	56.23
GME-Qwen2-VL-2B	63.49	79.66	73.14	79.66	64.99	53.83	56.22	64.46	71.56	52.10
DSE-Qwen2-2b-MRL-V1	61.16	63.07	69.71	75.62	60.15	62.37	64.70	74.44	81.50	60.03
VisRAG-Ret	55.17	57.81	66.40	74.47	53.60	51.56	54.99	64.93	75.40	49.48
Text Retrievers										
BGE-M3	31.49	33.09	37.92	42.71	29.93	23.90	25.87	31.17	36.36	22.21
Multilingual-E5-large	28.45	30.41	35.12	41.07	26.97	22.70	24.83	28.57	35.06	21.64
Jina-ColBERT-V2	24.61	25.93	28.82	33.02	23.68	16.72	18.56	21.52	27.27	15.88
BM25	11.39	12.65	14.70	18.67	10.79	18.26	21.63	26.35	31.54	18.52

Table 4: Retrieval results for both Chinese and English in percentage. The best results are highlighted in **bold**.

Model	Chinese						English					
	ROUGE	Acc	P_Rec	P_Prec	B_Rec	B_Prec	ROUGE	Acc	P_Rec	P_Prec	B_Rec	B_Prec
Proprietary MLLMs												
o4-mini	**38.55**	**58.13**	78.01	75.77	54.74	48.20	**40.21**	**69.20**	75.32	75.32	**60.11**	**55.75**
GPT-4o	26.82	33.26	92.15	87.27	<u>61.01</u>	<u>52.80</u>	<u>24.66</u>	43.41	**89.98**	<u>81.81</u>	<u>54.17</u>	<u>44.66</u>
GPT-4V	26.38	31.70	**93.10**	<u>88.56</u>	**61.29**	**52.88**	22.76	44.71	89.24	80.54	53.43	42.69
GPT-4o-mini	19.46	19.53	78.07	56.08	24.68	16.17	16.21	28.94	60.30	41.20	22.63	13.23
Gemini-1.5-Flash	18.18	21.34	69.58	67.10	20.62	16.80	16.24	26.72	72.17	66.71	25.97	21.05
Gemini-2.0-Flash	<u>28.00</u>	<u>41.40</u>	<u>92.87</u>	**89.58**	34.07	29.29	21.83	<u>46.01</u>	<u>89.61</u>	**85.22**	20.41	17.23
Claude-3.5-Sonnet	21.87	32.67	59.48	55.54	31.81	28.62	20.92	43.41	79.78	77.99	36.73	34.49
Open-Source MLLMs												
Qwen2-VL-72B-Instruct	22.83	30.41	58.25	51.31	10.64	9.49	**25.85**	25.97	53.80	43.68	7.42	5.91
Qwen2.5-VL-7B-Instruct	22.19	30.06	<u>65.38</u>	<u>62.27</u>	9.71	8.19	19.47	<u>36.36</u>	51.21	<u>49.25</u>	18.74	15.72
Qwen2.5-VL-32B-Instruct	**25.89**	<u>34.66</u>	**74.71**	**65.95**	**33.37**	<u>23.45</u>	21.33	30.05	<u>59.00</u>	48.03	<u>35.44</u>	<u>24.47</u>
Qwen2.5-VL-72B-Instruct	<u>25.12</u>	**36.02**	61.17	55.72	<u>32.75</u>	**28.54**	<u>21.98</u>	**38.03**	**68.09**	**63.93**	**39.52**	**35.03**
MiniCPM-o-2.6	13.15	11.58	60.94	57.68	2.81	2.48	18.32	9.83	37.29	36.30	0.74	0.46
Phi-3.5-V-Instruct	5.14	4.55	35.91	34.19	3.39	2.72	6.70	6.86	24.12	22.35	0.74	0.58

Table 5: Results for generation and citation on FinRAGBench-V in percentage. For both proprietary models and open-source models, the best result is shown in **bold**, and the second-best is <u>underlined</u>.

Citation Evaluation. The citation quality is evaluated using the aforementioned metrics at two different levels: page-level and block-level, as shown in Figure 4, denoted as: *P_Rec*, *P_Prec*, *B_Rec*, and *B_Prec*. Moreover, we use two evaluation approaches: box-bounding and image-cropping, to assess the citation quality. As shown in Appendix D, the former draws bounding boxes around relevant regions based on the citation coordinates, while the latter directly crops the cited image blocks accordingly. In both cases, we introduce an evaluator MLLM to determine citation quality. Through experiments in Section 7.3, we find that image-cropping yields higher alignment with Intersection over Union (IoU) scores and human judgments, and therefore it is used as the default approach in subsequent evaluations.

7 Experiments and Results

We evaluate both the retrieval stage and the generation stage with citation using the aforementioned metrics. For retrieval, we assess both multimodal and textual retrievers. For generation, we use the best retriever to provide the top-k pages ($k = 10$) as input, comparing the performance of proprietary and open-source MLLMs across different tasks.

7.1 Basic Settings

Retrieval. During the retrieval phase, we explore both multimodal retrievers alongside textual ones. (1) **Multimodal retrievers**: We evaluate five models, namely ColQwen2 (Faysse et al., 2024), GME-Qwen2-VL-2B (Zhang et al., 2024b), GME-Qwen2-VL-7B, DSE-QWen2-2b-MRL-V1 (Ma et al., 2024a), and VisRAG-Ret (Yu et al., 2024a), to assess their effectiveness in retrieving relevant content from multimodal pages. (2) **Text retrievers**: We use Marker (Paruchuri, 2024) for OCR-based text extraction. Subsequently, we test four text retrievers, namely BM25, Jina-ColBERT-V2 (Jha et al., 2024), BGE-M3 (Chen et al., 2024a), and Multilingual-E5-large (Wang et al., 2024a), evaluating their effectiveness in retrieving relevant information from the extracted texts.

Generation with Visual Citation In the generation phase, we conduct experiments on both proprietary and open-source MLLMs. The former consists of o4-mini, GPT-4o (OpenAI,

Eval Approach	Eval Model	Consistency with IoU			Consistency with Human Eval		
		Pearson	**Spearman**	**Kendall**	**Pearson**	**Spearman**	**Kendall**
image-cropping	GPT-4o	**65.06**	**63.08**	**54.58**	**68.01**	**64.03**	**57.37**
	GPT-4v	63.27	61.49	53.21	64.78	60.98	54.50
	GPT-4-turbo	52.44	54.66	46.87	57.56	54.82	48.70
	Gemini-1.5-Flash	53.55	50.47	43.59	50.39	47.01	41.99
	Gemini-2.0-Flash	54.18	53.89	46.17	60.09	57.86	51.42
box-bounding	GPT-4o	7.28	9.19	8.14	12.30	12.80	11.29

Table 6: Consistency of automatic citation evaluation methods with IoU and human evaluation in percentages.

2023), GPT-4V, GPT-4o-mini, Gemini-1.5-Flash (Reid et al., 2024), Gemini-2.0-Flash (Comanici et al., 2025), and Claude-3.5-Sonnet-20240620 (Anthropic, 2024); while the later comprises Qwen2-VL-72B-Instruct (Wang et al., 2024b), Qwen2.5-VL-7B-Instruct, Qwen2.5-VL-32B-Instruct, Qwen2.5-VL-72B-Instruct, Phi-3.5-vision-instruct (Abdin et al., 2024), and MiniCPM-o-2.6 (Yao et al., 2024). The prompt for generation is in Appendix A, more details are in Appendix G.

7.2 Main Results

Retrieval. In the retrieval stage, we observe that **multimodal retrievers significantly outperform textual retrievers across all metrics**. As shown in Table 4, ColQwen2 achieves a recall@10 of 90.13 (Chinese) and 85.86 (English), whereas the best textual retriever, BGE-M3, reaches only 42.71 and 36.36, respectively. This highlights the effectiveness of multimodal retrievers in handling complex financial data involving charts and tables.

Generation. From Table 5, we observe the following findings: (1) **Proprietary LLMs outperform their open-source counterparts**, underscoring the challenges that open-source MLLMs face in handling complex multimodal tasks. (2) **Different MLLMs show varying strengths on Chinese and English datasets.** Concretely, models such as GPT-4o, GPT-4V, Gemini-2.0-Flash, and Claude-3.5-Sonnet perform significantly better on English data, whereas Qwen2.5-VL-72B-Instruct and Qwen2-VL-72B-Instruct demonstrate balanced and even superior performance on Chinese data. (3) Task-wise analysis on FinRAGBench-V (Figure 5) shows that **MLLMs excel at text inference and direct information extraction, but still struggle with numerical calculations and multi-page inference**. These observations suggest that complex visual reasoning tasks in specialized domains like finance remain a key challenge for current MLLMs.

Some case studies on the typical errors are shown in Appendix E.

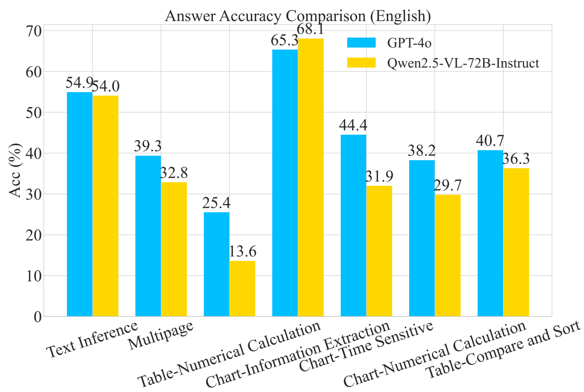


Figure 5: The comparison of answer accuracy between different question categories.

Visual Citation. In terms of citation, Table 5 shows that **most MLLMs perform well in page-level citations**, demonstrating their ability to accurately identify relevant pages from the provided references. However, **block-level citation remains difficult**, especially for open-source MLLMs. This highlights the challenge of attributing information to specific regions within a page, and suggests that many open-source MLLMs still struggle with precise citation generation. It also underscores the ongoing challenge of achieving accurate visual attribution within images, especially when pinpointing specific content blocks.

7.3 Consistency between Automatic Citation Evaluation with Human Evaluations

To validate our automatic citation evaluation method, we measure its alignment with the following two human evaluation methods.

IoU-based Human Evaluation. We employ the *labelImg*[3] tool to manually annotate citation regions, which serve as the visual ground truth. The

[3]https://github.com/HumanSignal/labelImg

Intersection over Union (IoU) between predicted and annotated boxes is computed to quantify geometric overlap. Although intuitive, this metric has notable limitations for evaluating citation grounding quality, as it can be influenced by factors such as blank space within bounding boxes or missing key information that still yields a high IoU score.

Rating-based Human Evaluation. To complement IoU, we use human ratings of the predicted citations on a 0–5 scale, considering factors such as page and block relevance, offset from ground truth, and the inclusion of redundant or irrelevant content. This provides a more nuanced and semantically meaningful assessment of citation quality. The guideline for rating is shown in Table 28 of Appendix F.

As shown in Table 6, we evaluate the citation performance of Qwen2.5-VL-72B using our automatic citation method across multiple variants, and assess its consistency with IoU scores and human ratings via Pearson, Spearman, and Kendall correlations coefficients. The image-cropping approach achieves Pearson correlations of 65.06 (with IoU) and 68.01 (with human ratings), demonstrating its effectiveness. In contrast, the box-bounding approach underperforms due to noise introduced by redundant visual content. Accordingly, we adopt GPT-4o with image-cropping in our experiments.

8 Conclusion

In this paper, we introduce FinRAGBench-V, a benchmark designed for multimodal RAG with visual citations in the financial domain, covering a retrieval corpus collected from diverse real-world financial documents and a QA dataset focusing on a wide range of financial tasks. Through extensive experiments, FinRAGBench-V exposes limitations of MLLMs and serves as a valuable resource to guide future improvements in visual RAG systems.

Limitations

Despite the comprehensive experiments conducted in FinRAGBench-V that have provided valuable insights, our work still has limitations. Specifically, we did not train a dedicated model for multimodal RAG in the financial domain. Future work should address this by developing models tailored to the unique challenges of financial multimodal RAG, thereby enhancing the applicability and effectiveness of our benchmark.

Acknowledgement

We thank the anonymous reviewers for their helpful comments on this paper. This work is supported by the National Natural Science Foundation of China (No. 62476010) and National Natural Science Foundation of China (No. 62272008).

References

Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat S. Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp A. Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *CoRR*, abs/2404.14219.

Anthropic. 2024. Claude 3.5 sonnet raises the industry bar for intelligence. `https://www.anthropic.com/news/claude-3-5-sonnet`. Accessed: 2025-09-07.

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

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024b. Benchmarking large language models in retrieval-augmented generation. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 17754–17762. AAAI Press.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Kenneth Huang, Bryan R. Routledge, and William Yang Wang. 2021. Finqa: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 3697–3711. Association for Computational Linguistics.

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

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit S. Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, Krishna Haridasan, Ahmed Omran, Nikunj Saunshi, Dara Bahri, Gaurav Mishra, Eric Chu, Toby Boyd, Brad Hekman, Aaron Parisi, Chaoyi Zhang, Kornraphop Kawintiranon, Tania Bedrax-Weiss, Oliver Wang, Ya Xu, Ollie Purkiss, Uri Mendlovic, Ilaï Deutel, Nam Nguyen, Adam Langley, Flip Korn, Lucia Rossazza, Alexandre Ramé, Sagar Waghmare, Helen Miller, Nathan Byrd, Ashrith Sheshan, Raia Hadsell Sangnie Bhardwaj, Pawel Janus, Tero Rissa, Dan Horgan, Sharon Silver, Ayzaan Wahid, Sergey Brin, Yves Raimond, Klemen Kloboves, Cindy Wang, Nitesh Bharadwaj Gundavarapu, Ilia Shumailov, Bo Wang, Mantas Pajarskas, Joe Heyward, Martin Nikoltchev, Maciej Kula, Hao Zhou, Zachary Garrett, Sushant Kafle, Sercan Arik, Ankita Goel, Mingyao Yang, Jiho Park, Koji Kojima, Parsa Mahmoudieh, Koray Kavukcuoglu, Grace Chen, Doug Fritz, Anton Bulyenov, Sudeshna Roy, Dimitris Paparas, Hadar Shemtov, Bo-Juen Chen, Robin Strudel, David Reitter, Aurko Roy, Andrey Vlasov, Changwan Ryu, Chas Leichner, Haichuan Yang, Zelda Mariet, Denis Vnukov, Tim Sohn, Amy Stuart, Wei Liang, Minmin Chen, Praynaa Rawlani, Christy Koh, JD Co-Reyes, Guangda Lai, Praseem Banzal, Dimitrios Vytiniotis, Jieru Mei, and Mu Cai. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261.

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

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

Ziliang Gan, Yu Lu, Dong Zhang, Haohan Li, Che Liu, Jian Liu, Ji Liu, Haipang Wu, Chaoyou Fu, Zenglin Xu, Rongjunchen Zhang, and Yong Dai. 2024. Mme-finance: A multimodal finance benchmark for expert-level understanding and reasoning. *CoRR*, abs/2411.03314.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023a. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6465–6488. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023b. Retrieval-augmented generation for large language models: A survey. *CoRR*, abs/2312.10997.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Rujun Han, Yuhao Zhang, Peng Qi, Yumo Xu, Jenyuan Wang, Lan Liu, William Yang Wang, Bonan Min, and Vittorio Castelli. 2024. RAG-QA arena: Evaluating domain robustness for long-form retrieval augmented question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 4354–4374. Association for Computational Linguistics.

Jie Huang, Wei Ping, Peng Xu, Mohammad Shoeybi, Kevin Chen-Chuan Chang, and Bryan Catanzaro. 2023. RAVEN: in-context learning with retrieval augmented encoder-decoder language models. *CoRR*, abs/2308.07922.

Muye Huang, Han Lai, Xinyu Zhang, Wenjun Wu, Jie Ma, Lingling Zhang, and Jun Liu. 2025. Evochart: A benchmark and a self-training approach towards real-world chart understanding. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 3680–3688. AAAI Press.

Gautier Izacard, Patrick S. H. Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *J. Mach. Learn. Res.*, 24:251:1–251:43.

Rohan Jha, Bo Wang, Michael Günther, Saba Sturua, Mohammad Kalim Akram, and Han Xiao. 2024. Jina-colbert-v2: A general-purpose multilingual late interaction retriever. *CoRR*, abs/2408.16672.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Dongfang Li, Zetian Sun, Xinshuo Hu, Zhenyu Liu, Ziyang Chen, Baotian Hu, Aiguo Wu, and Min Zhang. 2023. A survey of large language models attribution. *CoRR*, abs/2311.03731.

Xinze Li, Yixin Cao, Liangming Pan, Yubo Ma, and Aixin Sun. 2024. Towards verifiable generation: A benchmark for knowledge-aware language model attribution. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 493–516. Association for Computational Linguistics.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. 2024a. Unifying multimodal retrieval via document screenshot embedding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 6492–6505. Association for Computational Linguistics.

Xueguang Ma, Shengyao Zhuang, Bevan Koopman, Guido Zuccon, Wenhu Chen, and Jimmy Lin. 2024b. VISA: retrieval augmented generation with visual source attribution. *CoRR*, abs/2412.14457.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Vik Paruchuri. 2024. Marker.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. ARES: an automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 338–354. Association for Computational Linguistics.

Aviv Slobodkin, Eran Hirsch, Arie Cattan, Tal Schuster, and Ido Dagan. 2024. Attribute first, then generate: Locally-attributable grounded text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 3309–3344. Association for Computational Linguistics.

Manan Suri, Puneet Mathur, Franck Dernoncourt, Kanika Goswami, Ryan A. Rossi, and Dinesh Manocha. 2024. Visdom: Multi-document QA with visually rich elements using multimodal retrieval-augmented generation. *CoRR*, abs/2412.10704.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Multilingual E5 text embeddings: A technical report. *CoRR*, abs/2402.05672.

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. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Shuting Wang, Jiejun Tan, Zhicheng Dou, and Ji-Rong Wen. 2024c. Omnieval: An omnidirectional and automatic RAG evaluation benchmark in financial domain. *CoRR*, abs/2412.13018.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, Lei Chen, Nicolas Scheffer, Yue Liu, Nirav Shah, Rakesh Wanga, Anuj Kumar, Scott Yih, and Xin Dong. 2024. CRAG - comprehensive RAG benchmark. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

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

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. 2024b. Rankrag: Unifying context ranking with retrieval-augmented generation in llms. In

Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024.

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. 2024a. RAFT: adapting language model to domain specific RAG. *CoRR*, abs/2403.10131.

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. *CoRR*, abs/2412.16855.

A Prompts for QA Pairs Construction, Generation, and Evaluations

We provide the prompts for constructing QA paris, generating answer with visual citations, and the evaluation on the answer and citations, shown in Table 7, 8, 9, 10, 11, 12.

B Examples of Six Real-World Data Sources of Retrieval Corpus

In this section, we provide an example for each data source, illustrating the construction of our courpus, shown in Figure 6, 7, 8, 9, 10, 11.

C Examples of Seven Categories of QA Dataset

In this section, we provide an example for each category of questions, shown in Table 13, 14 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26.

C.1 Text Inference:

This category involves tasks such as summarization and information extraction from text. For example, deriving key insights from large volumes of text or identifying specific pieces of information, such as financial data or trends, within the content.

C.2 Chart-Information Extraction

This category focuses on extracting important metrics or features from charts. For example, it involves determining the exact percentage of a sector in a pie chart.

C.3 Chart-Numerical Calculations

In this category, the focus is on performing numerical calculations based on the data presented in charts. Tasks include calculating the change of interest rates, summing up costs, and evaluating the percentage point increase in market share, among others.

C.4 Chart-Time Sensitive

This category addresses time-based queries related to charts. It includes identifying the timing of specific events, analyzing trends over time, pinpointing the peaks and troughs in the data, etc. These queries often involve examining how certain indicators evolve and identifying key moments in time.

C.5 Table-Numerical Calculations

Similar to chart calculations, this category involves performing numerical operations on the data presented in tables. Common tasks include calculating the change of interest rates, summing up costs, etc. These calculations help derive meaningful insights from tabular data.

C.6 Table-Comparison and Sorting

This category focuses on comparing and sorting data within tables. It includes comparing financial indicators such as revenue or cost between different entities, as well as ranking them based on specific criteria. Tasks may also involve identifying the highest or lowest values among multiple entries.

C.7 Multi-page Queries

This category deals with queries that concern information from multiple pages. It includes tasks that span across text, tables, or charts split across pages. For example, it involves extracting truncated tables from different pages or interpreting information from multiple charts that need to be combined to answer a single query.

D Example for Visual Citation and the Two Evaluation Methods

Figure 12 gives an example of the MLLM's output with both answer and citations, and demonstrates two citation evaluation methods: box-bounding and image-cropping.

E Case Study

In this section, we provide several error cases based on both the different stages in the RGenCite baseline and the typical task types in finance.

E.1 Error Case Study Based on Different Stages in RGenCite

To illustrate the potential errors that can occur in RGenCite during generation and citation, we conduct a case study identifying three main types of errors. The first type occurs when the retrieved reference image provided to the model lacks relevant information, resulting in insufficient data for the model to answer the question, as shown in Figure 13 (a). The second type involves providing the correct image, but the model makes an error in graphical reasoning, often leading to incorrect numerical calculations, as shown in Figure 13 (b). The third type occurs when the model answers the question correctly but introduces bias or inaccuracies in the citation, leading to incorrect referencing, as shown in Figure 13 (c).

E.2 Error Case Study Based on Typical Task Types in the Financial Domain

Recognizing Candlestick Charts. As shown in Figure 14, for the query "Based on the report from EastMoney, what are the opening and closing prices of Zheshang Securities on October 10, 2024?" the correct analysis should recognize that red indicates an increase and green indicates a decrease in stock prices. The top of the candlestick body represents the opening price, while the bottom represents the closing price. In this case, the opening price was 14.25, and the closing price was 13.55. However, due to the lack of relevant knowledge, the models either produce incorrect results or generate responses like "The image contains news reports about Zheshang Securities' acquisition of Guodu Securities shares and some securities market data, but it does not provide the specific opening and closing prices for Zheshang Securities on October 10, 2024".

Dealing with Complex Financial Table. Figure 15 is an error case that MLLMs fail in handling complex financial tables. In this case, the model was asked to calculate the change in total global structured finance maximum exposure to loss for AMBAC Financial Group, Inc. between December 31, 2019, and December 31, 2020. Although it correctly extracted the initial value of $8,165 million, it mistakenly identified the ending value as $6,325 million instead of the correct $6,352 million. This minor misreading led to an incorrect computed decrease of $1,840 million instead of the correct $1,813 million. Such errors reveal the challenges MLLMs face in accurately interpreting numeric details from financial tables, where even small misreads can lead to significant factual inaccuracies.

Dealing with Multi-page Questions. The example in Figure 16 illustrates a typical limitation of MLLMs when dealing with lengthy financial tables that span multiple pages. The model was asked to extract and compare the quarterly GDP growth rates for the United States and Brazil in Q1 2021 from the Global Economic Prospects report. However, the relevant data was distributed across two separate pages, and the model failed to aggregate the information correctly. As a result, it misreporting the growth rate of Brazil and the U.S., leading to an inaccurate comparison. This case highlights the difficulty MLLMs face in maintaining contextual continuity across paginated tables, a common format in financial documents.

F Annotation guidelines.

This section demonstrates the annotation guidelines. The annotation guidelines for constructing the QA dataset in shown Table 27 and the guideline for rating-based human evaluation for visual citation is in Table 28.

G Resource Usage

Throughout the processes of dataset construction, response generation, and evaluation, we employed multiple proprietary language model APIs, including GPT-4o and other commercial multimodal large language models (MLLMs). The total API usage cost amounted to $3,021.47. All experiments with open-source models were conducted locally on 4×A100 80GB GPUs. The dataset was manually annotated by three experienced annotators to ensure quality and consistency.

We relied on several mainstream libraries and toolkits across retrieval, generation, and evaluation tasks, including PyTorch, Transformers, pytrec_eval, pylate.

We carefully considered the licenses and intended use cases of all third-party artifacts utilized in our study. All datasets and tools used from external sources were employed strictly within the bounds of their respective licenses and intended purposes, primarily for academic research.

H Potential Risks

Despite careful design and construction, our retrieval corpus and QA dataset may still contain potential risks. During the data collection process, some noisy, outdated, or irrelevant financial documents might not have been fully filtered. Similarly, in the QA dataset, there may be annotation errors, ambiguities, or biases due to imperfect filtering and manual oversight. These issues could affect the accuracy of model evaluation and the generalizability of experimental results. We encourage users of FinRAGBench-V to be aware of these limitations and apply additional validation where necessary.

Grant financing fell to its lowest level since 2019, totaling $4 billion in 2023 and representing just 2% of total climate commitments. Grant financing reached a high of $24 billion in 2022, driven by substantial grant funding committed by OECD-based members for energy efficiency and renewable energy in buildings. Falling by more than 80% compared to 2022, grant finance in 2023 returned to the level observed in 2019. Globally, grants represented 5% of climate finance flows in 2021/22.[69]

Total concessional finance ($57 billion), comprising concessional loans and grant finance, was 8% less in 2023 than it was, on average, from 2019 to 2022. This is a potentially worrying trend because of concessional funding's important role in green finance for developing and emerging economies. Concessional finance can relieve debt distress experienced in vulnerable low- and middle-income countries, while in emerging economies, it can help kickstart frontier markets for innovative climate change solutions. Prior to 2023, the share of grants in IDFC's total climate finance had been steadily increasing. Going forward, concessional finance, as well as non-concessional public resources, should be leveraged by members as they seek to increase the impact of their green finance commitments by harnessing concessional finance in transformational ways (see Section 4).

The use of other instruments, such as equity, multiple instruments, and other instruments,[70] increased in 2023 from $1.4 billion in 2022 to $3.8 billion. In particular, equity finance rose from $0.6 billion in 2022 to $1.9 billion in 2023, representing 1% of total climate finance commitments in 2023. Guarantees totaled $270 million, less than 1% of climate finance commitments. Risk mitigation instruments such as guarantees can be used by members to address market barriers and crowds in other investors in areas where the risk of investment is perceived as high. Box 4 describes examples of how guarantees have been used to promote energy efficiency investment in India.

As shown in Figure 17, non-concessional loans are the most-used instrument for both mitigation (68%) and adaptation (59%). Concessional loans are also significant, representing 26% of mitigation commitments and 23% of adaptation commitments. Concessional loans are the largest single financing instrument for projects with dual benefits (47%).

Figure 17: Climate finance commitments by instrument and use category in 2019-2023

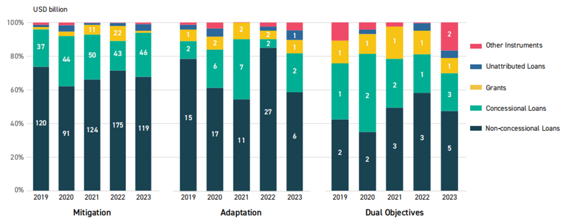


Figure 6: An example of research report

ARCBEST CORPORATION
CONSOLIDATED STATEMENTS OF CASH FLOWS

	Year Ended December 31		
	2020	**2019**	**2018**
		(in thousands)	
OPERATING ACTIVITIES			
Net income	$ **71,100**	$ 39,985	$ 67,262
Adjustments to reconcile net income to net cash provided by operating activities:			
Depreciation and amortization	**114,379**	108,099	104,114
Amortization of intangibles	**4,012**	4,367	4,521
Pension settlement expense, including termination expense	**89**	8,505	12,925
Share-based compensation expense	**10,478**	9,523	8,413
Provision for losses on accounts receivable	**4,327**	1,223	2,336
Change in deferred income taxes	**7,715**	5,411	1,872
Asset impairment	**—**	26,514	—
Gain on sale of property and equipment and lease termination	**(2,376)**	(5,247)	(59)
Gain on sale of subsidiaries	**—**	—	(1,945)
Changes in operating assets and liabilities:			
Receivables	**(38,129)**	13,720	(23,554)
Prepaid expenses	**(7,966)**	(4,756)	(2,988)
Other assets	**2,646**	(1,365)	(4,341)
Income taxes	**(1,712)**	(8,720)	12,169
Operating right-of-use assets and lease liabilities, net	**756**	728	—
Multiemployer pension fund withdrawal liability	**(611)**	(584)	22,602
Accounts payable, accrued expenses, and other liabilities	**41,281**	(27,039)	52,020
NET CASH PROVIDED BY OPERATING ACTIVITIES	**205,989**	170,364	255,347
INVESTING ACTIVITIES			
Purchases of property, plant and equipment, net of financings	**(43,248)**	(90,955)	(43,992)
Proceeds from sale of property and equipment	**13,348**	13,490	4,256
Proceeds from sale of subsidiaries	**—**	—	4,680
Purchases of short-term investments	**(165,133)**	(129,709)	(108,495)
Proceeds from sale of short-term investments	**216,735**	120,409	58,698
Capitalization of internally developed software	**(14,241)**	(11,476)	(10,097)
NET CASH PROVIDED BY (USED IN) INVESTING ACTIVITIES	**7,461**	(98,241)	(94,950)
FINANCING ACTIVITIES			
Borrowings under credit facilities	**180,000**	—	—
Borrowings under accounts receivable securitization program	**45,000**	—	—
Proceeds from notes payable	**—**	20,410	—
Payments on long-term debt	**(326,098)**	(58,938)	(71,260)
Net change in book overdrafts	**6,510**	(2,722)	262
Deferred financing costs	**—**	(562)	(202)
Payment of common stock dividends	**(8,157)**	(8,187)	(8,244)
Purchases of treasury stock	**(6,595)**	(9,110)	(9,404)
Payments for tax withheld on share-based compensation	**(2,065)**	(1,291)	(2,135)
NET CASH USED IN FINANCING ACTIVITIES	**(111,405)**	(60,400)	(90,983)
NET INCREASE IN CASH AND CASH EQUIVALENTS	**102,045**	11,723	69,414
Cash and cash equivalents at beginning of period	**201,909**	190,186	120,772
CASH AND CASH EQUIVALENTS CASH AT END OF PERIOD	$ **303,954**	$ 201,909	$ 190,186
NONCASH INVESTING ACTIVITIES			
Equipment and other financings	$ **61,803**	$ 70,372	$ 94,016
Accruals for equipment received	$ **1,667**	$ 234	$ 2,807
Lease liabilities arising from obtaining right-of-use assets	$ **67,819**	$ 32,761	$ —

The accompanying notes are an integral part of the consolidated financial statements.

74

Figure 7: An example of financial statements

Instruction: Here is an image of a document. Your task is to generate queries about this document image from various perspectives, categorize the questions (category), provide answers to the questions (answer), and specify whether the answer is a long or short answer (answer_type).
###I hope your questions are as detailed as possible. Begin by specific about which document you are referring to and describe the required text, table, or chart content without explicitly mentioning the figure or table number.
###Your questions can target the text, tables, charts, or any other elements in the image.
###Design three different queries for each document, ensuring that the question categories (category) are distinct from each other.
###The categories of questions you can include are: Text-based QA:
1. Text-Text Inference: Extraction or reasoning based on textual information.

Chart-based QA:
1. Chart-Information Extraction: Extract key metrics or features from the chart.
2. Chart-Numerical Calculation: Includes calculations such as growth rates, interest rates, total costs, etc.
3. Chart-Time-Sensitive: Includes trend descriptions, causal relationships, event sequences, frequencies, durations, etc.

Table-based QA:
1. Table-Numerical Calculation: Perform calculations such as growth rates, interest rates, total costs, etc., using table data.
2. Table-Comparison and Sorting: Compare or rank entities based on specific criteria (e.g., return rates, risks).

Here is the format of your output:

```
{
    "result":[
        {
            "query" : "",
            "category":"",
            "answer": "",
            "answer_type":""
        },
        {
            "answer": "",
            "query" : "",
            "category":"",
            "answer_type":""
        },
        {
            "answer": "",
            "query" : "",
            "category":"",
            "answer_type":""
        }
    ]
}
```

Here are some examples:
{examples}

Table 7: Prompt for Constructing QA Dataset

Instruction: Answer the following questions based on the given images, identify the images that support your answer, and further locate the source of your answer in the images by outputting coordinate pairs.
###If the answer uses more than one image, you must point out all the images used; If your answer uses information from more than one image, you must annotate all the used information.
###All your annotations must fully support your answer, and there must not be any unsupported information in your answer.
###When annotating an image, you need to annotate a full graph or text paragraph, not just a specific number.
Your replies must strictly follow the following JSON format:

```
{
    "answer":"",
    "coordinates":{
    "1":[[x1, y1, x2, y2], [x1, y1, x2, y2]],
    "2":[[x1, y1, x2, y2], [x1, y1, x2, y2]],
        ... # These are the supportive images and the coordinate pairs in them
    }
}
```

Here is the question: {query}
Here are the images:
Image 1: Width: width1, Height: height1
(Image 1 in Base64)
Image 2: Width: width2, Height: height2
(Image 2 in Base64)
.
.
.

Table 8: Prompt for Generation and Citation

Question: {query_text}
Ground_truth: {expected_answer}
Model_answer: {actual_answer}
Is the model answer correct? You only need to output 'true' for correct or 'false' for incorrect. If the model answer does not contain any information, it should be judged as 'false'.

Table 9: Prompt for Response Accuracy Evaluation

Answer: {answer} Please judge whether these pages cover the answer, your answer can only be 'yes' or 'no'.
Here are my images:
(Image 1 in Base64)
(Image 2 in Base64)
.
.
.

Table 10: Prompt for Page-Level Citation Evaluation

Answer: {answer} The following images will contain marked areas (red boxes), please judge whether these marked areas (red boxes) cover the content of the answer, your answer can only be 'yes' if it covers or 'no' if it doesn't cover.
Here are my images:
(Image 1 in Base64)
(Image 2 in Base64)
.
.
.

Table 11: Prompt for Block-Level Citation Evaluation using Box-Bounding

Answer: {answer} Below are some extracts from the images, please decide if they cover the answers given, your answer can only be ’yes’ if it covers or ’no’ if it doesn’t cover.
Here are my images:
(Image 1 in Base64)
(Image 2 in Base64)
.
.
.

Table 12: Prompt for Block-Level Citation Evaluation using Image-Cropping

Query:	In Howden Joinery Group Plc’s Annual Report & Accounts 2022, with respect to the Nominations Committee report for 2022, who is mentioned as the individual appointed to lead the Committee and who retired?
Category:	Text Inference
Answer:	Peter Ventress was appointed as the Committee Chairman, and Richard Pennycook retired.
Reference Image:	

Table 13: QA Dataset Example 1: An Example of Text Inference Question

Query:	From the document 'Independent auditors' report to the members of Craneware plc', what is the significance of revenue recognition as a key audit matter in the context of the Group's financial statement?
Category:	Text Inference
Answer:	Revenue recognition is significant because it involves determining the amount of revenue to be recognized based on contract details and conditions in contracts with customers. The risk is identified at the journal level related to the existence and occurrence of all revenue streams.
Reference Image:	

Independent auditors' report

to the members of Craneware plc [Cont'd]

The scope of our audit

As part of designing our audit, we determined materiality and assessed the risks of material misstatement in the financial statements.

Key audit matters

Key audit matters are those matters that, in the auditors' professional judgement, were of most significance in the audit of the financial statements of the current period and include the most significant assessed risks of material misstatement (whether or not due to fraud) identified by the auditors, including those which had the greatest effect on: the overall audit strategy; the allocation of resources in the audit; and directing the efforts of the engagement team. These matters, and any comments we make on the results of our procedures thereon, were addressed in the context of our audit of the financial statements as a whole, and in forming our opinion thereon, and we do not provide a separate opinion on these matters.

This is not a complete list of all risks identified by our audit.

Valuation of Purchase Price allocation related to acquisition of Sentry Data Systems Inc. (group) is a new key audit matter this year. Impact of Covid-19 (group and parent), which was a key audit matter last year, is no longer included because of no material impact of Covid-19 during the reporting period. Otherwise, the key audit matters below are consistent with last year.

Key audit matter	How our audit addressed the key audit matter
Revenue Recognition (group and parent)	
The Group has revenue of $165,544k (2021: $75,578k) and the Company has revenue of $40,863k (2021: $43,700k). These amounts are significant in the context of the Group statement of comprehensive income. The amount of revenue to be recognised is determined based on the contract details. The timing of revenue recognition is dependent on the terms contained in the contracts with customers. The risk has been identified at the journals level related to existence and occurrence of the all revenue streams.	To address significant risk at the journals level we ran unusual account combinations tests and tested journals triggered by the test to ascertain that it doesn't represent fraud. No matters arose during our testing.
Internally developed intangible assets (group and parent)	
As per note 15, the Group has net book value of development costs capitalised amounting to $40,489k (2021: $31,652k) and the Company has $37,500k (2021: 31,652k) capitalised on the balance sheet. Development costs are capitalised when the following criteria have been met: new product development costs are technically feasible; production and sale is intended; a market exists; expenditure can be measured reliably; and sufficient resources are available to complete such projects. The Directors are required to continually assess the commercial potential of each product in development in order to determine if costs can continue to be capitalised. We focus on this area as there is judgement involved in the Directors' assessment. We consider this as a key audit matter because there is a risk that the costs being capitalised are not allowable under IAS 38 and also that the intangible assets will not generate sufficient economic benefit to recover the value of the intangible asset.	On a sample basis we agreed additions to intangible assets to supporting documentation, including invoices and time records. We obtained an understanding for the proportion of employee costs being capitalised and verified these against payroll information (for example, payroll reports and employee registers) and timesheets to verify the amount of time that employees spend on the capital projects. The nature of the costs being capitalised was assessed to ensure it met the accounting requirements to capitalise and analysis was obtained from the technical team to audit time charged by employees. Discussions were held with management in order to understand how all criteria for capitalisation had been met and supporting evidence was obtained to corroborate this. Regarding recoverability of intangible assets, we had discussions with management and obtained underlying support to assess the ability of the projects to generate future economic benefit which included project road maps, sales order value generates so far as well as future pipeline and potential of sales. We also assessed the intangible assets for indications of impairment. No matters arose during our testing.

Craneware plc
90 Annual Report 2022

Table 14: QA Dataset Example 2: An Example of Text Inference Question

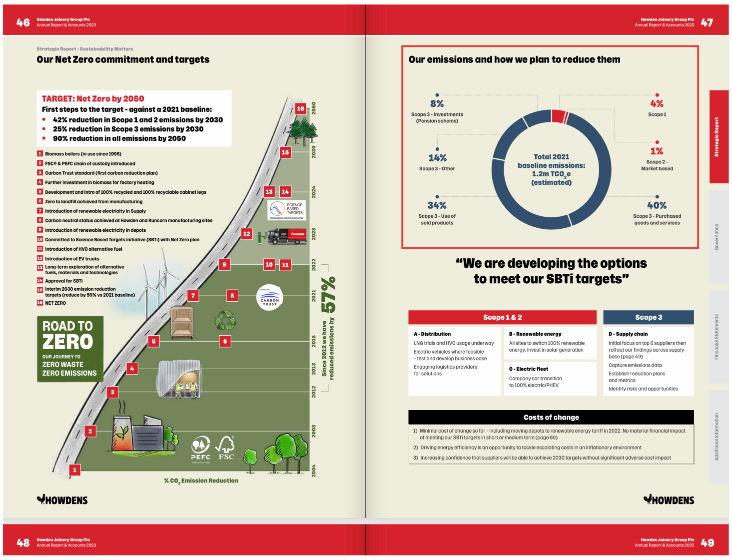
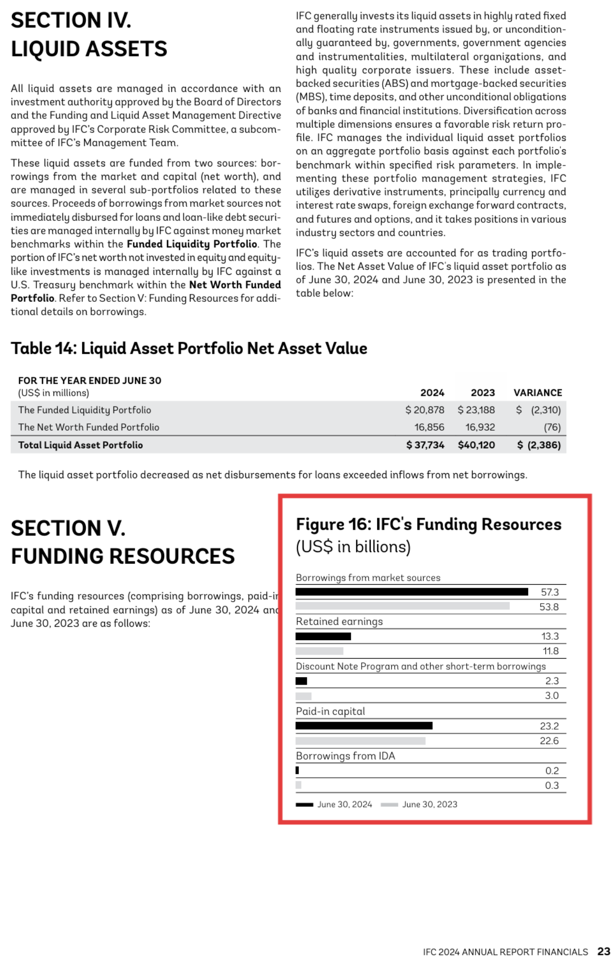
Query:	According to the Annual Report and Account for Howden Joinery Group Plc in 2023, what is the total baseline emissions estimation for 2021? How many percentage does the purchased goods and services take among them?
Category:	Chart-Information Extraction
Answer:	The total 2021 baseline emissions are estimated at 1.2m {TCO_2e}. Among them, purchased goods and services takes 40%.
Reference Image:	

Table 15: QA Dataset Example 3: An Example of Chart-Information Exraction Question

Query:	According to IFC's 2024 annual report, among all the IFC's funding resources, which one is the highest?
Category:	Chart-Information Extraction
Answer:	Borrowings from market resources.
Reference Image:	



SECTION IV.
LIQUID ASSETS

All liquid assets are managed in accordance with an investment authority approved by the Board of Directors and the Funding and Liquid Asset Management Directive approved by IFC's Corporate Risk Committee, a subcommittee of IFC's Management Team.

These liquid assets are funded from two sources: borrowings from the market and capital (net worth), and are managed in several sub-portfolios related to these sources. Proceeds of borrowings from market sources not immediately disbursed for loans and loan-like debt securities are managed internally by IFC against money market benchmarks within the **Funded Liquidity Portfolio**. The portion of IFC's net worth not invested in equity and equity-like investments is managed internally by IFC against a U.S. Treasury benchmark within the **Net Worth Funded Portfolio**. Refer to Section V: Funding Resources for additional details on borrowings.

IFC generally invests its liquid assets in highly rated fixed and floating rate instruments issued by, or unconditionally guaranteed by, governments, government agencies and instrumentalities, multilateral organizations, and high quality corporate issuers. These include asset-backed securities (ABS) and mortgage-backed securities (MBS), time deposits, and other unconditional obligations of banks and financial institutions. Diversification across multiple dimensions ensures a favorable risk return profile. IFC manages the individual liquid asset portfolios on an aggregate portfolio basis against each portfolio's benchmark within specified risk parameters. In implementing these portfolio management strategies, IFC utilizes derivative instruments, principally currency and interest rate swaps, foreign exchange forward contracts, and futures and options, and it takes positions in various industry sectors and countries.

IFC's liquid assets are accounted for as trading portfolios. The Net Asset Value of IFC's liquid asset portfolio as of June 30, 2024 and June 30, 2023 is presented in the table below:

Table 14: Liquid Asset Portfolio Net Asset Value

FOR THE YEAR ENDED JUNE 30 (US$ in millions)	2024	2023	VARIANCE
The Funded Liquidity Portfolio	$ 20,878	$ 23,188	$ (2,310)
The Net Worth Funded Portfolio	16,856	16,932	(76)
Total Liquid Asset Portfolio	**$ 37,734**	**$40,120**	**$ (2,386)**

The liquid asset portfolio decreased as net disbursements for loans exceeded inflows from net borrowings.

SECTION V.
FUNDING RESOURCES

IFC's funding resources (comprising borrowings, paid-in capital and retained earnings) as of June 30, 2024 and June 30, 2023 are as follows:

Figure 16: IFC's Funding Resources
(US$ in billions)

IFC 2024 ANNUAL REPORT FINANCIALS 23

Table 16: QA Dataset Example 4: An Example of Chart-Information Exraction Question

Query:	Analyzing the Private Financing Deal Count reported by FinTech Insights in Q3 2024, how many financing deals did it increased from Q1 2021 to Q2 2021?
Category:	Chart-Numerical Calculations
Answer:	18
Reference Image:	

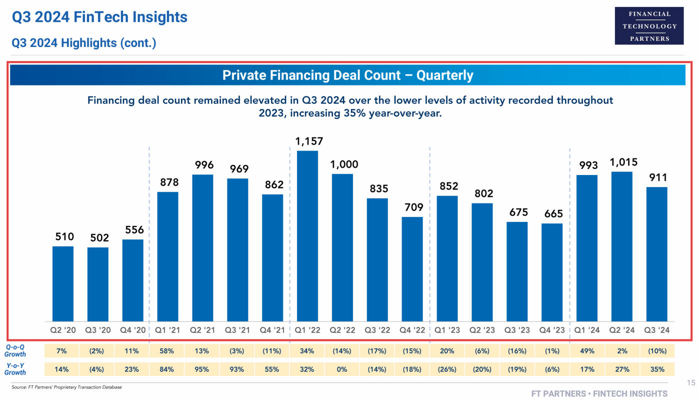
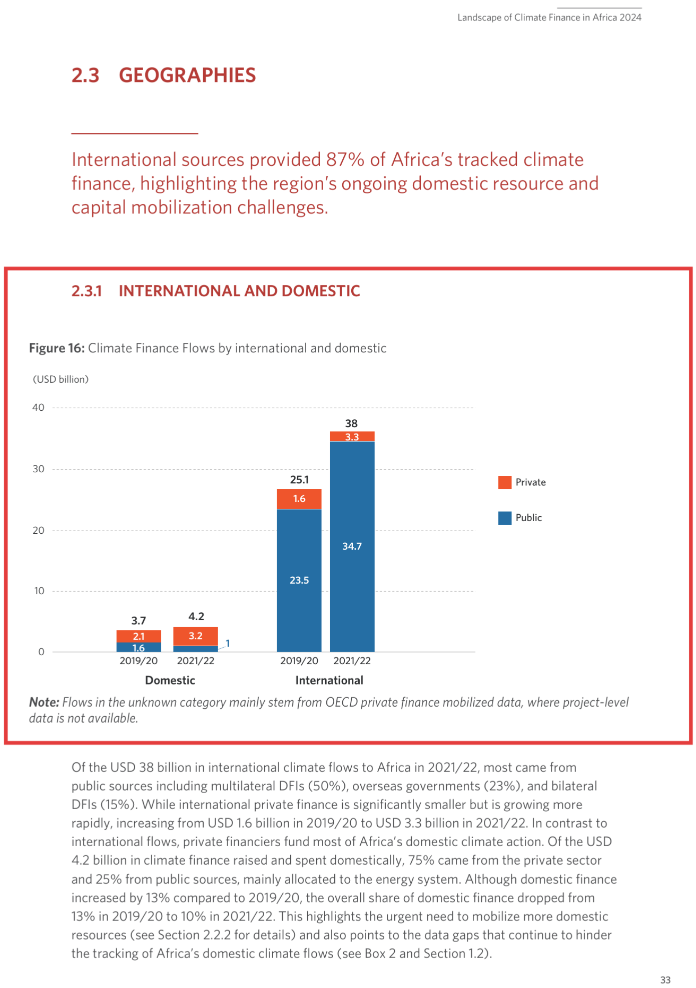


Table 17: QA Dataset Example 5: An Example of Chart-Numerical Calculations Question

Query:	Based on the statistics of climate finance flows by international and domestic, what is the growth rate of domestic public funding from 2019/20 to 2021/22?
Category:	Chart-Numerical Calculations
Answer:	-37.5%
Reference Image:	


Landscape of Climate Finance in Africa 2024

2.3 GEOGRAPHIES

International sources provided 87% of Africa's tracked climate finance, highlighting the region's ongoing domestic resource and capital mobilization challenges.

2.3.1 INTERNATIONAL AND DOMESTIC

Figure 16: Climate Finance Flows by international and domestic

Note: *Flows in the unknown category mainly stem from OECD private finance mobilized data, where project-level data is not available.*

Of the USD 38 billion in international climate flows to Africa in 2021/22, most came from public sources including multilateral DFIs (50%), overseas governments (23%), and bilateral DFIs (15%). While international private finance is significantly smaller but is growing more rapidly, increasing from USD 1.6 billion in 2019/20 to USD 3.3 billion in 2021/22. In contrast to international flows, private financiers fund most of Africa's domestic climate action. Of the USD 4.2 billion in climate finance raised and spent domestically, 75% came from the private sector and 25% from public sources, mainly allocated to the energy system. Although domestic finance increased by 13% compared to 2019/20, the overall share of domestic finance dropped from 13% in 2019/20 to 10% in 2021/22. This highlights the urgent need to mobilize more domestic resources (see Section 2.2.2 for details) and also points to the data gaps that continue to hinder the tracking of Africa's domestic climate flows (see Box 2 and Section 1.2).

33

Table 18: QA Dataset Example 6: An Example of Chart-Numerical Calculations Question

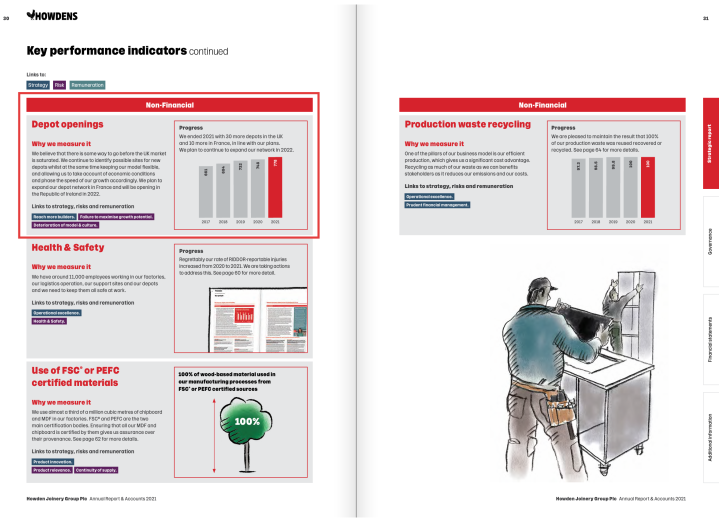
Query:	According to Howden Joinery Group Plc Annual Report & Accounts 2021, what is the trend of depot openings in the UK and France from 2017 to 2021?
Category:	Chart-Time Sensitive
Answer:	There's a consistent increase in depot openings from 2017 to 2021, with a particularly significant increase in 2021.
Reference Image:	

Table 19: QA Dataset Example 7: An Example of Chart-Time Sensitive Question

Query:	According to the Wall Street stocks data from July 31,2024 to Aug 13,2024, explain the trends of S&P 500 and Nasdaq Composite indices during that time period.
Category:	Chart-Time Sensitive
Answer:	There's a steep decline followed by a bounce back for both the S&P 500 and Nasdaq Composite indices. After an initial drop where both indices reached close to their lowest points, they recovered steadily with the Nasdaq Composite seeing a slightly stronger recovery than the S&P 500. This indicates a volatile period followed by a short-term rebound.
Reference Image:	


FINANCIAL TIMES
COMPANIES & MARKETS

Harris must not cave to threats of crypto crowd

Dennis Kelleher
Markets Insight

It remains the financial product of choice among financial predators, lawbreakers and criminals

The day in the markets

What you need to know

Markets update

Main equity markets

Biggest movers

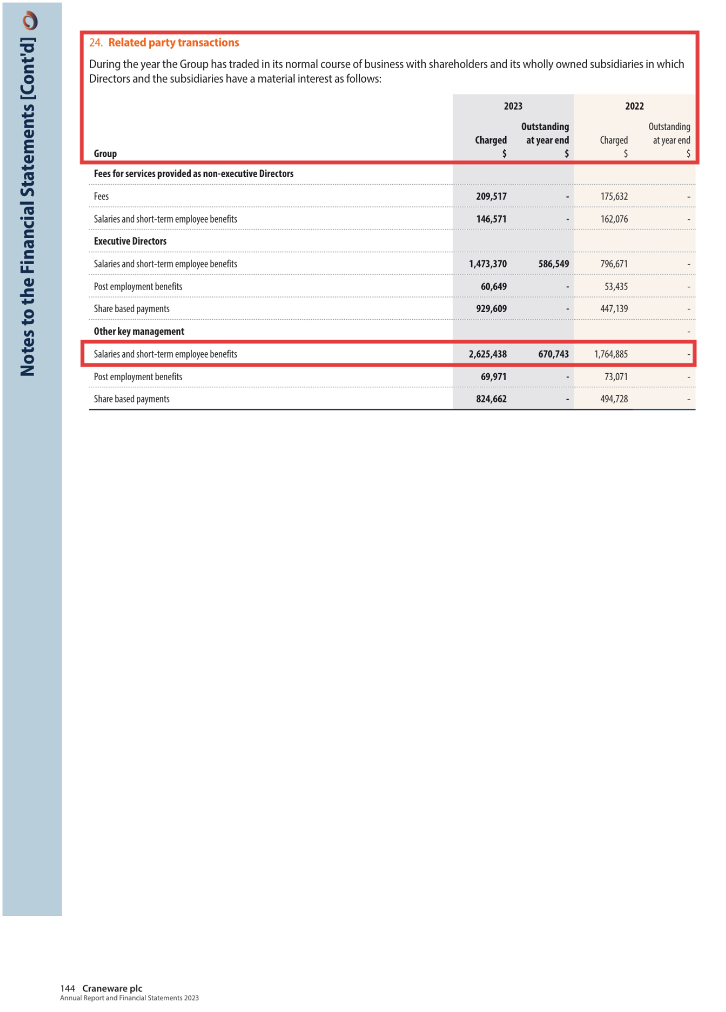
Trading infrastructure

Australian regulator sues stock exchange over botched blockchain system upgrade

'ASX's statements go to the heart of the trust in the integrity of our markets. This was a collective failure'

Table 20: QA Dataset Example 8: An Example of Chart-Time Sensitive Question

Query:	Based on the data under the 'Related party transactions' in the Craneware plc Annual Report and Financial Statements 2023, what is the percent increase in Salaries and short-term employee benefits for Executive Directors from 2022 to 2023?
Category:	Table-Numerical Calculations
Answer:	An increase of approximately 84.94%.
Reference Image:	



Notes to the Financial Statements [Cont'd]

24. **Related party transactions**

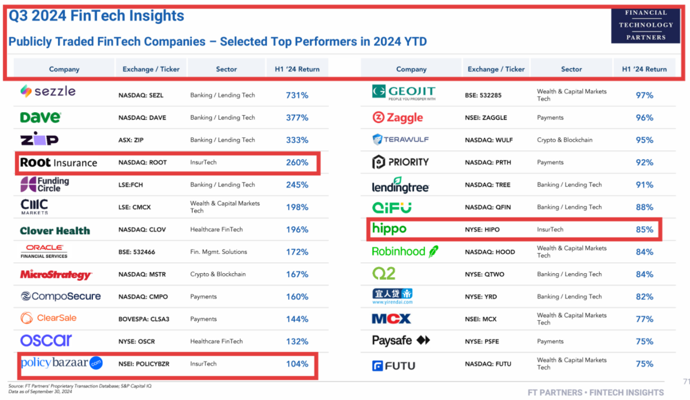
During the year the Group has traded in its normal course of business with shareholders and its wholly owned subsidiaries in which Directors and the subsidiaries have a material interest as follows:

Group	2023 Charged $	2023 Outstanding at year end $	2022 Charged $	2022 Outstanding at year end $
Fees for services provided as non-executive Directors				
Fees	**209,517**	-	175,632	-
Salaries and short-term employee benefits	**146,571**	-	162,076	-
Executive Directors				
Salaries and short-term employee benefits	**1,473,370**	**586,549**	796,671	-
Post employment benefits	**60,649**	-	53,435	-
Share based payments	**929,609**	-	447,139	-
Other key management				-
Salaries and short-term employee benefits	**2,625,438**	**670,743**	1,764,885	-
Post employment benefits	**69,971**	-	73,071	-
Share based payments	**824,662**	-	494,728	-

144 Craneware plc
Annual Report and Financial Statements 2023

Table 21: QA Dataset Example 9: An Example of Table-Numerical Calculations Question

Query:	According to the Q3 2024 FinTech Insights document, with respect to Publicly Traded FinTech Companies – Selected Top Performers in 2024 YTD, what is the combined H1 2024 Return for all companies categorized under 'InsurTech'?
Category:	Table-Numerical Calculations
Answer:	The combined H1 2024 Return for companies under 'InsurTech' is 449%. This is calculated by adding the returns of Root Insurance (260%), Hippo (85%), and Policybazaar.com (104%).
Reference Image:	


Q3 2024 FinTech Insights

Publicly Traded FinTech Companies – Selected Top Performers in 2024 YTD

FINANCIAL TECHNOLOGY PARTNERS

Company	Exchange / Ticker	Sector	H1 '24 Return
sezzle	NASDAQ: SEZL	Banking / Lending Tech	731%
Dave	NASDAQ: DAVE	Banking / Lending Tech	377%
zip	ASX: ZIP	Banking / Lending Tech	333%
Root Insurance	NASDAQ: ROOT	InsurTech	260%
Funding Circle	LSE:FCH	Banking / Lending Tech	245%
CMC Markets	LSE: CMCX	Wealth & Capital Markets Tech	198%
Clover Health	NASDAQ: CLOV	Healthcare FinTech	196%
Oracle Financial Services	BSE: 532466	Fin. Mgmt. Solutions	172%
MicroStrategy	NASDAQ: MSTR	Crypto & Blockchain	167%
CompoSecure	NASDAQ: CMPO	Payments	160%
ClearSale	BOVESPA: CLSA3	Payments	144%
oscar	NYSE: OSCR	Healthcare FinTech	132%
policybazaar.com	NSEI: POLICYBZR	InsurTech	104%

Company	Exchange / Ticker	Sector	H1 '24 Return
GEOJIT	BSE: 532285	Wealth & Capital Markets Tech	97%
Zaggle	NSEI: ZAGGLE	Payments	96%
TERAWULF	NASDAQ: WULF	Crypto & Blockchain	95%
PRIORITY	NASDAQ: PRTH	Payments	92%
lendingtree	NASDAQ: TREE	Banking / Lending Tech	91%
QIFU	NASDAQ: QFIN	Banking / Lending Tech	88%
hippo	NYSE: HIPO	InsurTech	85%
Robinhood	NASDAQ: HOOD	Wealth & Capital Markets Tech	84%
Q2	NYSE: QTWO	Banking / Lending Tech	84%
宜人贷 www.yirendai.com	NYSE: YRD	Banking / Lending Tech	82%
MCX	NSEI: MCX	Wealth & Capital Markets Tech	77%
Paysafe	NYSE: PSFE	Payments	75%
FUTU	NASDAQ: FUTU	Wealth & Capital Markets Tech	75%

Source: FT Partners' Proprietary Transaction Database; S&P Capital IQ
Data as of September 30, 2024

71

FT PARTNERS • FINTECH INSIGHTS

Table 22: QA Dataset Example 10: An Example of Table-Numerical Calculations Question

Query:	According to the 2022 annual report of Craneware plc, which plan had the larger exercise price range: the 2016 Schedule 4 Option Plan or the 2018 SAYE Option Plan?
Category:	Table-Comparison and Sorting
Answer:	2016 Schedule 4 Option Plan.
Reference Image:	

Notes to the Financial Statements [Cont'd]

8. **Share-based payments [Cont'd]**

Share option plans

Share options, granted by the Company to employees in respect of the following number of Ordinary Shares, were outstanding at 30 June 2022.

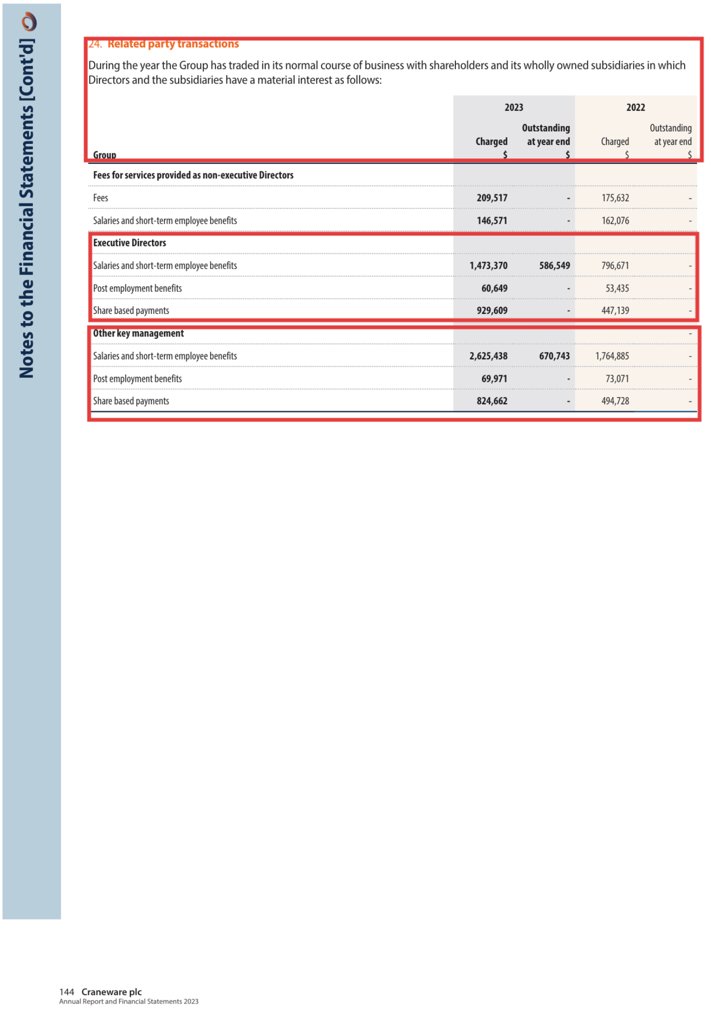
Date of grant	Exercise price (GBP)	Exercise price (USD)	Remaining life at 1 July 2021 (years)	No of options at 1 July 2021	Granted	Exercised	Lapsed	No of options at 30 June 2022	Remaining life at 30 June 2022 (years)
2007 Share Option Plan									
04 Sep 2012	£3.60	$5.72	1.2	1,725	-	(1,725)	-	-	-
21 Sep 2012	£4.00	$6.50	1.2	6,605	-	-	-	6,605	0.2
10 Sep 2013	£3.95	$6.21	2.2	47,190	-	-	-	47,190	1.2
22 Sep 2014	£5.225	$8.39	3.2	94,416	-	-	-	94,416	2.2
09 Mar 2016	£7.50	$10.66	4.7	100,756	-	-	-	100,756	3.7
12 Sep 2016	£11.775	$15.63	5.2	36,469	-	-	-	36,469	4.2
2016 Unapproved Option Plan									
24 Mar 2017	£12.375	$15.44	5.7	35,126	-	(3,838)	-	31,288	4.7
17 Jan 2018	£17.750	$24.45	6.5	48,517	-	(5,070)	-	43,447	5.5
05 Sep 2018	£27.100	$34.88	7.2	38,970	-	-	(1,615)	37,355	6.2
04 Sep 2019	£19.000	$23.01	8.2	19,456	-	-	(1,578)	17,878	7.2
02 Oct 2020	£15.050	$19.36	9.3	63,509	-	-	(6,476)	57,033	8.3
18 Nov 2021	£26.100	$35.21	-	-	168,036	-	(41,021)	127,015	9.4
2016 Schedule 4 Option Plan									
24 Mar 2017	£12.375	$15.44	5.7	15,958	-	(4,848)	-	11,110	4.7
17 Jan 2018	£17.750	$24.45	6.5	6,759	-	(845)	-	5,914	5.5
05 Sep 2018	£27.100	$34.88	7.2	3,588	-	-	(359)	3,229	6.2
04 Sep 2019	£19.000	$23.01	8.2	5,312	-	-	(1,920)	3,392	7.2
02 Oct 2020	£15.050	$19.36	9.3	11,692	-	-	(2,159)	9,533	8.3
18 Nov 2021	£26.100	$35.21	-	-	29,645	-	(5,451)	24,194	9.4
2018 Employee Stock Purchase Plan									
24 Mar 2020	£11.475	$13.34	0.7	18,498	-	(15,630)	(2,868)	-	-
23 Mar 2021	£18.360	$25.42	1.7	7,420	-	-	(1,281)	6,139	0.7
2018 SAYE Option Plan									
20 Apr 2020	£11.475	$14.32	2.3	38,726	-	-	(3,790)	34,936	1.3
19 Apr 2021	£18.360	$25.39	3.3	4,302	-	-	(1,010)	3,292	2.3
				604,994	197,681	(31,956)	(69,528)	701,191	

Craneware plc
Annual Report 2022 113

Table 23: QA Dataset Example 11: An Example of Table-Comparison and Sorting Question

Query:	In the 'Related party transactions' of the Craneware plc Annual Report and Financial Statements 2023, compare the share-based payments for Executive Directors and Other key management for 2023. Which category received higher payments?
Category:	Table-Comparison and Sorting
Answer:	For the year 2023, Executive Directors received $929,609 in share-based payments, while Other key management received $824,662. Executive Directors received higher payments.

Reference Image:



Notes to the Financial Statements [Cont'd]

24. Related party transactions

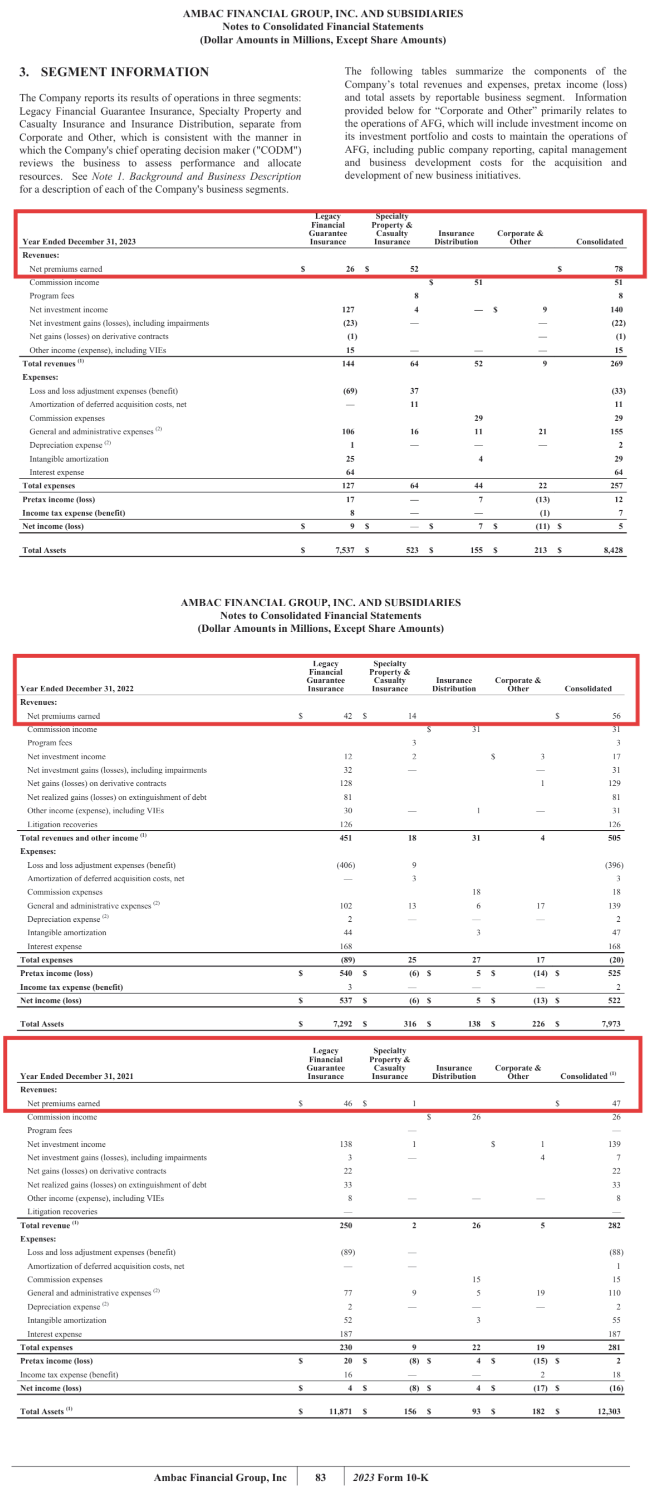
During the year the Group has traded in its normal course of business with shareholders and its wholly owned subsidiaries in which Directors and the subsidiaries have a material interest as follows:

Group	2023 Charged $	2023 Outstanding at year end $	2022 Charged $	2022 Outstanding at year end $
Fees for services provided as non-executive Directors				
Fees	209,517	-	175,632	-
Salaries and short-term employee benefits	146,571	-	162,076	-
Executive Directors				
Salaries and short-term employee benefits	1,473,370	586,549	796,671	-
Post employment benefits	60,649	-	53,435	-
Share based payments	929,609	-	447,139	-
Other key management				-
Salaries and short-term employee benefits	2,625,438	670,743	1,764,885	-
Post employment benefits	69,971	-	73,071	-
Share based payments	824,662	-	494,728	-

144 Craneware plc
Annual Report and Financial Statements 2023

Table 24: QA Dataset Example 12: An Example of Table-Comparison and Sorting Question

Query:	According to Ambac Financial Group, Inc' 2023 Form 10-K, during the years 2021 to 2023, which year had the highest Net premiums earned under Legacy Financial Guarantee Insurance?
Category:	Multi-page
Answer:	During the years 2021 to 2023, the highest net premiums earned by Legacy Financial Guarantee Insurance were in 2021, amounting to 46 million US dollars.
Reference Image:	



AMBAC FINANCIAL GROUP, INC. AND SUBSIDIARIES
Notes to Consolidated Financial Statements
(Dollar Amounts in Millions, Except Share Amounts)

3. SEGMENT INFORMATION

The Company reports its results of operations in three segments: Legacy Financial Guarantee Insurance, Specialty Property and Casualty Insurance and Insurance Distribution, separate from Corporate and Other, which is consistent with the manner in which the Company's chief operating decision maker ("CODM") reviews the business to assess performance and allocate resources. See *Note 1. Background and Business Description* for a description of each of the Company's business segments.

The following tables summarize the components of the Company's total revenues and expenses, pretax income (loss) and total assets by reportable business segment. Information provided below for "Corporate and Other" primarily relates to the operations of AFG, which will include investment income on its investment portfolio and costs to maintain the operations of AFG, including public company reporting, capital management and business development costs for the acquisition and development of new business initiatives.

Year Ended December 31, 2023	Legacy Financial Guarantee Insurance	Specialty Property & Casualty Insurance	Insurance Distribution	Corporate & Other	Consolidated
Revenues:					
Net premiums earned	$ 26	$ 52			$ 78
Commission income			$ 51		51
Program fees		8			8
Net investment income	127	4	—	$ 9	140
Net investment gains (losses), including impairments	(23)	—		—	(22)
Net gains (losses) on derivative contracts	(1)			—	(1)
Other income (expense), including VIEs	15	—	—	—	15
Total revenues [1]	144	64	52	9	269
Expenses:					
Loss and loss adjustment expenses (benefit)	(69)	37			(33)
Amortization of deferred acquisition costs, net	—	11			11
Commission expenses			29		29
General and administrative expenses [2]	106	16	11	21	155
Depreciation expense [2]	1	—	—	—	2
Intangible amortization	25		4		29
Interest expense	64				64
Total expenses	127	64	44	22	257
Pretax income (loss)	17	—	7	(13)	12
Income tax expense (benefit)	8	—	—	(1)	7
Net income (loss)	$ 9	$ —	$ 7	$ (11)	$ 5
Total Assets	$ 7,537	$ 523	$ 155	$ 213	$ 8,428

AMBAC FINANCIAL GROUP, INC. AND SUBSIDIARIES
Notes to Consolidated Financial Statements
(Dollar Amounts in Millions, Except Share Amounts)

Year Ended December 31, 2022	Legacy Financial Guarantee Insurance	Specialty Property & Casualty Insurance	Insurance Distribution	Corporate & Other	Consolidated
Revenues:					
Net premiums earned	$ 42	$ 14			$ 56
Commission income			$ 31		31
Program fees		3			3
Net investment income	12	2		$ 3	17
Net investment gains (losses), including impairments	32	—		—	31
Net gains (losses) on derivative contracts	128			1	129
Net realized gains (losses) on extinguishment of debt	81				81
Other income (expense), including VIEs	30	—	1	—	31
Litigation recoveries	126				126
Total revenues and other income [1]	451	18	31	4	505
Expenses:					
Loss and loss adjustment expenses (benefit)	(406)	9			(396)
Amortization of deferred acquisition costs, net	—	3			3
Commission expenses			18		18
General and administrative expenses [2]	102	13	6	17	139
Depreciation expense [2]	2	—	—	—	2
Intangible amortization	44		3		47
Interest expense	168				168
Total expenses	(89)	25	27	17	(20)
Pretax income (loss)	$ 540	$ (6)	$ 5	$ (14)	$ 525
Income tax expense (benefit)	3	—	—	—	2
Net income (loss)	$ 537	$ (6)	$ 5	$ (13)	$ 522
Total Assets	$ 7,292	$ 316	$ 138	$ 226	$ 7,973

Year Ended December 31, 2021	Legacy Financial Guarantee Insurance	Specialty Property & Casualty Insurance	Insurance Distribution	Corporate & Other	Consolidated [1]
Revenues:					
Net premiums earned	$ 46	$ 1			$ 47
Commission income			$ 26		26
Program fees		—			—
Net investment income	138	1		$ 1	139
Net investment gains (losses), including impairments	3	—		4	7
Net gains (losses) on derivative contracts	22				22
Net realized gains (losses) on extinguishment of debt	33				33
Other income (expense), including VIEs	8	—	—	—	8
Litigation recoveries	—				—
Total revenue [1]	250	2	26	5	282
Expenses:					
Loss and loss adjustment expenses (benefit)	(89)	—			(88)
Amortization of deferred acquisition costs, net	—	—			1
Commission expenses			15		15
General and administrative expenses [2]	77	9	5	19	110
Depreciation expense [2]	2	—	—	—	2
Intangible amortization	52		3		55
Interest expense	187				187
Total expenses	230	9	22	19	281
Pretax income (loss)	$ 20	$ (8)	$ 4	$ (15)	$ 2
Income tax expense (benefit)	16	—	—	2	18
Net income (loss)	$ 4	$ (8)	$ 4	$ (17)	$ (16)
Total Assets [1]	$ 11,871	$ 156	$ 93	$ 182	$ 12,303

Ambac Financial Group, Inc | 83 | *2023* Form 10-K

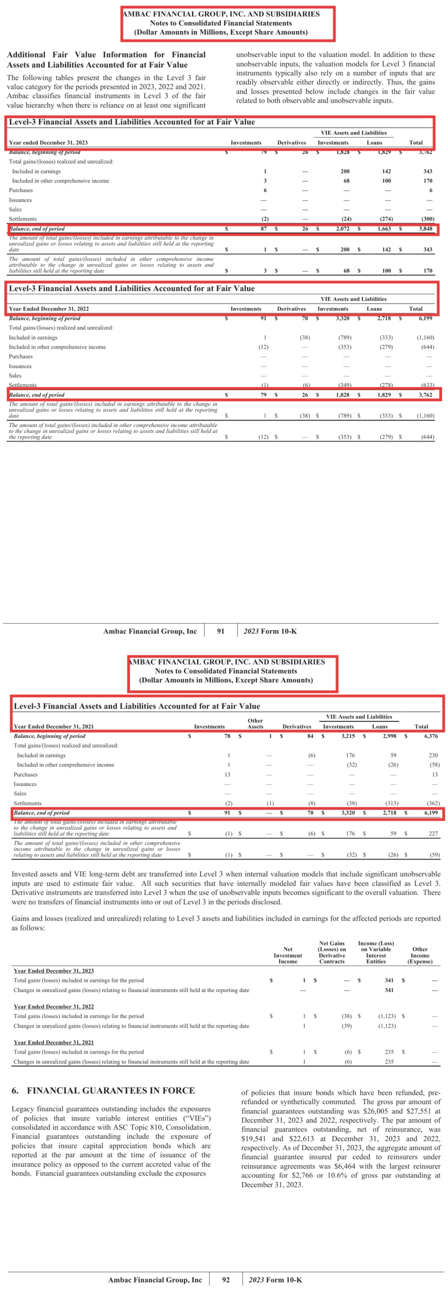
Table 25: QA Dataset Example 13: An Example of Multi-page Question

Query: According to Ambac Financial Group, Inc. 2023 Form 10-K, how did the total value of Level-3 Financial Assets and Liabilities change for AMBAC Financial Group, Inc. and its subsidiaries for each end of period from 2021 to 2023??

Category: Multi-page

Answer: The total value of Level-3 Financial Assets and Liabilities for AMBAC Financial Group, Inc. and its subsidiaries at the end of each period from 2021 to 2023 changed as follows: At the end of December 31, 2021, the total value was $6,199 million; At the end of December 31, 2022, the total value was $3,762 million; At the end of December 31, 2023, the total value was $3,848 million. This shows a decrease in the total value from 2021 to 2022, followed by a slight increase from 2022 to 2023.

Reference Image:



AMBAC FINANCIAL GROUP, INC. AND SUBSIDIARIES
Notes to Consolidated Financial Statements
(Dollar Amounts in Millions, Except Share Amounts)

Additional Fair Value Information for Financial Assets and Liabilities Accounted for at Fair Value

The following tables present the changes in the Level 3 fair value category for the periods presented in 2023, 2022 and 2021. Ambac classifies financial instruments in Level 3 of the fair value hierarchy when there is reliance on at least one significant unobservable input to the valuation model. In addition to these unobservable inputs, the valuation models for Level 3 financial instruments typically also rely on a number of inputs that are readily observable either directly or indirectly. Thus, the gains and losses presented below include changes in the fair value related to both observable and unobservable inputs.

Level-3 Financial Assets and Liabilities Accounted for at Fair Value

			VIE Assets and Liabilities		
Year ended December 31, 2023	Investments	Derivatives	Investments	Loans	Total
Balance, beginning of period	$ 79	$ 26	$ 1,828	$ 1,829	$ 3,762
Total gains/(losses) realized and unrealized:					
Included in earnings	1	—	200	142	343
Included in other comprehensive income	3	—	68	100	170
Purchases	6	—	—	—	6
Issuances	—	—	—	—	—
Sales	—	—	—	—	—
Settlements	(2)	—	(24)	(274)	(300)
Balance, end of period	$ 87	$ 26	$ 2,072	$ 1,663	$ 3,848
The amount of total gains/(losses) included in earnings attributable to the change in unrealized gains or losses relating to assets and liabilities still held at the reporting date	$ 1	$ —	$ 200	$ 142	$ 343
The amount of total gains/(losses) included in other comprehensive income attributable to the change in unrealized gains or losses relating to assets and liabilities still held at the reporting date	$ 3	$ —	$ 68	$ 100	$ 170

Level-3 Financial Assets and Liabilities Accounted for at Fair Value

			VIE Assets and Liabilities		
Year Ended December 31, 2022	Investments	Derivatives	Investments	Loans	Total
Balance, beginning of period	$ 91	$ 70	$ 3,320	$ 2,718	$ 6,199
Total gains/(losses) realized and unrealized:					
Included in earnings	1	(38)	(789)	(333)	(1,160)
Included in other comprehensive income	(12)	—	(353)	(279)	(644)
Purchases	—	—	—	—	—
Issuances	—	—	—	—	—
Sales	—	—	—	—	—
Settlements	(1)	(6)	(349)	(278)	(633)
Balance, end of period	$ 79	$ 26	$ 1,828	$ 1,829	$ 3,762
The amount of total gains/(losses) included in earnings attributable to the change in unrealized gains or losses relating to assets and liabilities still held at the reporting date	$ 1	$ (38)	$ (789)	$ (333)	$ (1,160)
The amount of total gains/(losses) included in other comprehensive income attributable to the change in unrealized gains or losses relating to assets and liabilities still held at the reporting date	$ (12)	$ —	$ (353)	$ (279)	$ (644)

Ambac Financial Group, Inc | 91 | *2023* Form 10-K

AMBAC FINANCIAL GROUP, INC. AND SUBSIDIARIES
Notes to Consolidated Financial Statements
(Dollar Amounts in Millions, Except Share Amounts)

Level-3 Financial Assets and Liabilities Accounted for at Fair Value

				VIE Assets and Liabilities		
Year Ended December 31, 2021	Investments	Other Assets	Derivatives	Investments	Loans	Total
Balance, beginning of period	$ 78	$ 1	$ 84	$ 3,215	$ 2,998	$ 6,376
Total gains/(losses) realized and unrealized:						
Included in earnings	1	—	(6)	176	59	230
Included in other comprehensive income	1	—	—	(32)	(26)	(58)
Purchases	13	—	—	—	—	13
Issuances	—	—	—	—	—	—
Sales	—	—	—	—	—	—
Settlements	(2)	(1)	(8)	(38)	(313)	(362)
Balance, end of period	$ 91	$ —	$ 70	$ 3,320	$ 2,718	$ 6,199
The amount of total gains/(losses) included in earnings attributable to the change in unrealized gains or losses relating to assets and liabilities still held at the reporting date	$ (1)	$ —	$ (6)	$ 176	$ 59	$ 227
The amount of total gains/(losses) included in other comprehensive income attributable to the change in unrealized gains or losses relating to assets and liabilities still held at the reporting date	$ (1)	$ —	$ —	$ (32)	$ (26)	$ (59)

Invested assets and VIE long-term debt are transferred into Level 3 when internal valuation models that include significant unobservable inputs are used to estimate fair value. All such securities that have internally modeled fair values have been classified as Level 3. Derivative instruments are transferred into Level 3 when the use of unobservable inputs becomes significant to the overall valuation. There were no transfers of financial instruments into or out of Level 3 in the periods disclosed.

Gains and losses (realized and unrealized) relating to Level 3 assets and liabilities included in earnings for the affected periods are reported as follows:

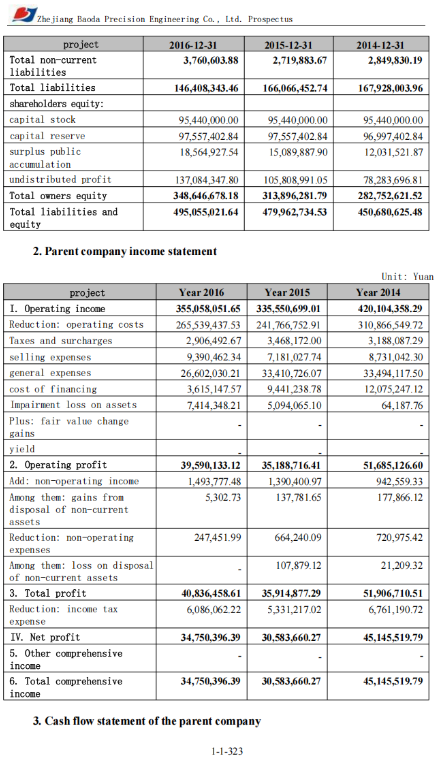
	Net Investment Income	Net Gains (Losses) on Derivative Contracts	Income (Loss) on Variable Interest Entities	Other Income (Expense)
Year Ended December 31, 2023				
Total gains (losses) included in earnings for the period	$ 1	$ —	$ 341	$ —
Changes in unrealized gains (losses) relating to financial instruments still held at the reporting date	—	—	341	—
Year Ended December 31, 2022				
Total gains (losses) included in earnings for the period	$ 1	$ (38)	$ (1,123)	$ —
Changes in unrealized gains (losses) relating to financial instruments still held at the reporting date	1	(39)	(1,123)	—
Year Ended December 31, 2021				
Total gains (losses) included in earnings for the period	$ 1	$ (6)	$ 235	$ —
Changes in unrealized gains (losses) relating to financial instruments still held at the reporting date	1	(6)	235	—

6. FINANCIAL GUARANTEES IN FORCE

Legacy financial guarantees outstanding includes the exposures of policies that insure variable interest entities ("VIEs") consolidated in accordance with ASC Topic 810, Consolidation. Financial guarantees outstanding include the exposure of policies that insure capital appreciation bonds which are reported at the par amount at the time of issuance of the insurance policy as opposed to the current accreted value of the bonds. Financial guarantees outstanding exclude the exposures of policies that insure bonds which have been refunded, pre-refunded or synthetically commuted. The gross par amount of financial guarantees outstanding was $26,005 and $27,551 at December 31, 2023 and 2022, respectively. The par amount of financial guarantees outstanding, net of reinsurance, was $19,541 and $22,613 at December 31, 2023 and 2022, respectively. As of December 31, 2023, the aggregate amount of financial guarantee insured par ceded to reinsurers under reinsurance agreements was $6,464 with the largest reinsurer accounting for $2,766 or 10.6% of gross par outstanding at December 31, 2023.

Ambac Financial Group, Inc | 92 | *2023* Form 10-K

Table 26: QA Dataset Example 14: An Example of Multi-page Question



Zhejiang Baoda Precision Engineering Co., Ltd. Prospectus

project	2016-12-31	2015-12-31	2014-12-31
Total non-current liabilities	**3,760,603.88**	**2,719,883.67**	**2,849,830.19**
Total liabilities	**146,408,343.46**	**166,066,452.74**	**167,928,003.96**
shareholders equity:			
capital stock	95,440,000.00	95,440,000.00	95,440,000.00
capital reserve	97,557,402.84	97,557,402.84	96,997,402.84
surplus public accumulation	18,564,927.54	15,089,887.90	12,031,521.87
undistributed profit	137,084,347.80	105,808,991.05	78,283,696.81
Total owners equity	**348,646,678.18**	**313,896,281.79**	**282,752,621.52**
Total liabilities and equity	**495,055,021.64**	**479,962,734.53**	**450,680,625.48**

2. Parent company income statement

Unit: Yuan

project	Year 2016	Year 2015	Year 2014
I. Operating income	**355,058,051.65**	**335,550,699.01**	**420,104,358.29**
Reduction: operating costs	265,539,437.53	241,766,752.91	310,866,549.72
Taxes and surcharges	2,906,492.67	3,468,172.00	3,188,087.29
selling expenses	9,390,462.34	7,181,027.74	8,731,042.30
general expenses	26,602,030.21	33,410,726.07	33,494,117.50
cost of financing	3,615,147.57	9,441,238.78	12,075,247.12
Impairment loss on assets	7,414,348.21	5,094,065.10	64,187.76
Plus: fair value change gains	-	-	-
yield	-	-	-
2. Operating profit	**39,590,133.12**	**35,188,716.41**	**51,685,126.60**
Add: non-operating income	1,493,777.48	1,390,400.97	942,559.33
Among them: gains from disposal of non-current assets	5,302.73	137,781.65	177,866.12
Reduction: non-operating expenses	247,451.99	664,240.09	720,975.42
Among them: loss on disposal of non-current assets	-	107,879.12	21,209.32
3. Total profit	**40,836,458.61**	**35,914,877.29**	**51,906,710.51**
Reduction: income tax expense	6,086,062.22	5,331,217.02	6,761,190.72
IV. Net profit	**34,750,396.39**	**30,583,660.27**	**45,145,519.79**
5. Other comprehensive income	-	-	-
6. Total comprehensive income	**34,750,396.39**	**30,583,660.27**	**45,145,519.79**

3. Cash flow statement of the parent company

1-1-323

Figure 8: An example of prospectus

FINANCIAL TIMES

INTERNATIONAL

Artificial intelligence

Gates plays down AI energy use fears

Microsoft founder argues the technology will speed transition to green power

GLOBAL INSIGHT

Alan Beattie

Tug of war looms for Labour between EU and US security

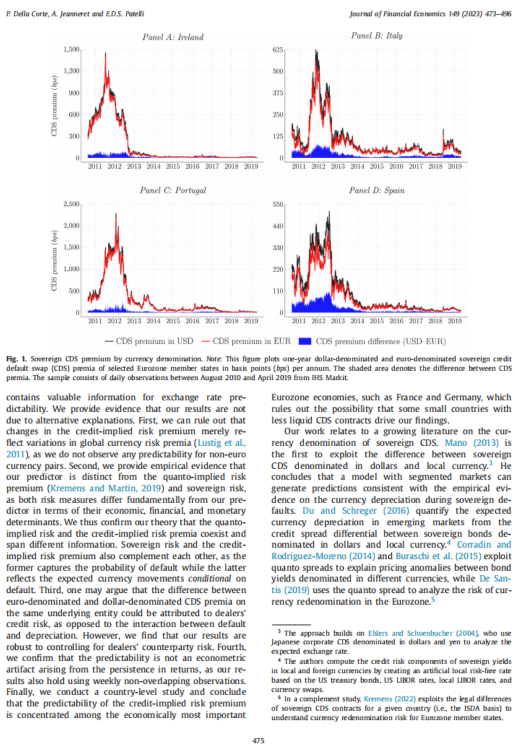
OECD report

Teens lack financial literacy and maths for digital economy

Technology industry

India's manufacturing push held back by China visa bottleneck

Figure 10: An example of financial magazine



P. Della Corte, A. Jeanneret and E.D.S. Patelli — Journal of Financial Economics 149 (2023) 473–496

Fig. 1. Sovereign CDS premium by currency denomination. *Note:* This figure plots one-year dollar-denominated and euro-denominated sovereign credit default swap (CDS) premia of selected Eurozone member states in basis points (*bps*) per annum. The shaded area denotes the difference between CDS premia. The sample consists of daily observations between August 2010 and April 2019 from IHS Markit.

contains valuable information for exchange rate predictability. We provide evidence that our results are not due to alternative explanations. First, we can rule out that changes in the credit-implied risk premium merely reflect variations in global currency risk premia (Lustig et al., 2011), as we do not observe any predictability for non-euro currency pairs. Second, we provide empirical evidence that our predictor is distinct from the quanto-implied risk premium (Kremens and Martin, 2019) and sovereign risk, as both risk measures differ fundamentally from our predictor in terms of their economic, financial, and monetary determinants. We thus confirm our theory that the quanto-implied risk and the credit-implied risk premia coexist and span different information. Sovereign risk and the credit-implied risk premium also complement each other, as the former captures the probability of default while the latter reflects the expected currency movements *conditional* on default. Third, one may argue that the difference between euro-denominated and dollar-denominated CDS premia on the same underlying entity could be attributed to dealers' credit risk, as opposed to the interaction between default and depreciation. However, we find that our results are robust to controlling for dealers' counterparty risk. Fourth, we confirm that the predictability is not an econometric artifact arising from the persistence in returns, as our results also hold using weekly non-overlapping observations. Finally, we conduct a country-level study and conclude that the predictability of the credit-implied risk premium is concentrated among the economically most important Eurozone economies, such as France and Germany, which rules out the possibility that some small countries with less liquid CDS contracts drive our findings.

Our work relates to a growing literature on the currency denomination of sovereign CDS. Mano (2013) is the first to exploit the difference between sovereign CDS denominated in dollars and local currency.[3] He concludes that a model with segmented markets can generate predictions consistent with the empirical evidence on the currency depreciation during sovereign defaults. Du and Schreger (2016) quantify the expected currency depreciation in emerging markets from the credit spread differential between sovereign bonds denominated in dollars and local currency.[4] Corradin and Rodriguez-Moreno (2014) and Buraschi et al. (2015) exploit quanto spreads to explain pricing anomalies between bond yields denominated in different currencies, while De Santis (2019) uses the quanto spread to analyze the risk of currency redenomination in the Eurozone.[5]

[3] The approach builds on Ehlers and Schoenbucher (2004), who use Japanese corporate CDS denominated in dollars and yen to analyze the expected exchange rate.

[4] The authors compute the credit risk components of sovereign yields in local and foreign currencies by creating an artificial local risk-free rate based on the US treasury bonds, US LIBOR rates, local LIBOR rates, and currency swaps.

[5] In a complement study, Kremens (2022) exploits the legal differences of sovereign CDS contracts for a given country (i.e., the ISDA basis) to understand currency redenomination risk for Eurozone member states.

475

Figure 9: An example of finance-related academic paper

Japanese tech company develops tailor-made products for Chinese consumers

By FAN FEIFEI | chinadaily.com.cn | Updated: 2024-12-19 16:51

Japanese tech company Canon Inc is looking to further tap the immense potential of China's consumption market and develop products that are tailor-made for local consumers in response to their evolving demands, said a senior company executive.

China serves as one of the most important markets in Canon's global business layout, said Hideki Ozawa, executive vice-president of Canon, and president and CEO of Canon China, emphasizing that the company has set the goal of making Canon China number 1 in terms of sales within the whole group by 2035.

He said it is noteworthy that Chinese Generation Z consumers — those born between the late-1990s and the mid-2010s — are more willing to take pictures with cameras than previous generations, which presents enormous development potential for Canon.

Noting that Chinese Gen Z, with a population of about 300 million, will become the driving force of China's consumer market in the future, Ozawa said Canon is developing products tailored to Gen Z's aesthetic and usage habits, such as youth-focused mirrorless cameras and customized printing solutions, to attract more young consumers.

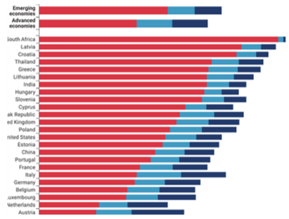
Ozawa said he is bullish on the prospects of China's imaging sector, and the company will intensify efforts on research and development, and roll out more innovative products and services that meet the diverse and personalized needs of Chinese consumers.

Figure 11: An example of financial news

```
{
    "query": "Which of the emerging economies has the highest proportion of households able to sustain one year of living expenses in the face of a loss of income?",
    "answer": "South Africa has the highest proportion of households among the emerging economies that can sustain one year of living expenses in the face of a loss of income.",
    "coordinates": {
        "1": [
            [
                93,
                103,
                539,
                436
            ]
        ]
    }
}
```

(a) Model Response with Answer and Citation

(b) Box-bounding

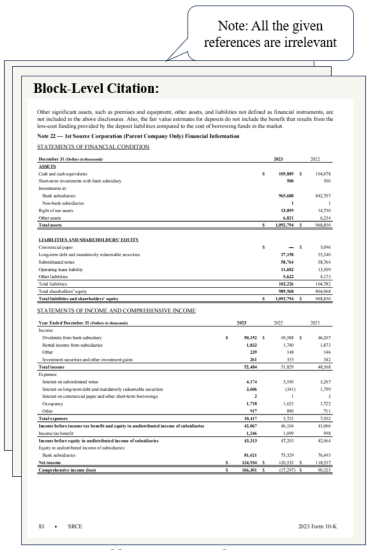


(c) Image-cropping

Figure 12: An example of MLLM's output of answer with visual citations, and the two evaluation methods: box-bounding and image-cropping.

Question: What is the net income for Ist Source Corporation in 2023, and how does it compare to 2022?

Answer: The net income for 1st Source Corporation in 2023 is not provided in the images. Therefore, a comparison with 2022 cannot be made based on the available data.



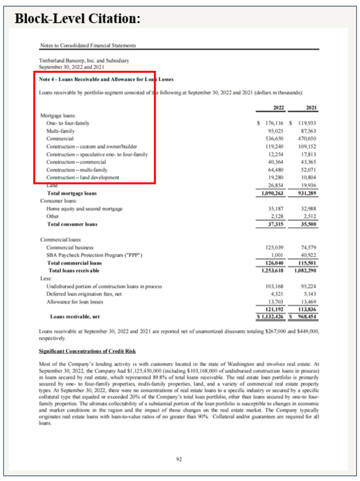
(a) Insufficient Reference

Question: According to the Timberland Bancorp, Inc. and Subsidiary's report, what is the year-over-year increase in total mortgage loans from 2021 to 2022?

Answer: The year-over-year increase in total mortgage loans from 2021 to 2022 is $158,984,000.

Wrong Numerical Calculation:

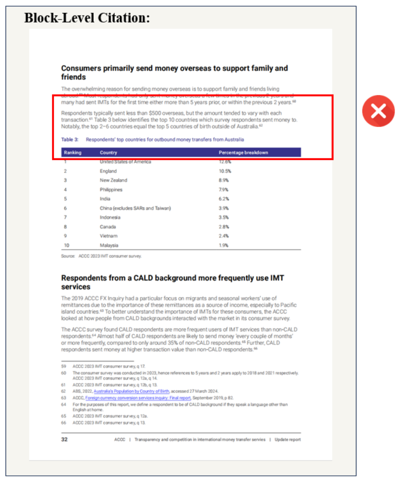
(should be 1,090,263-931,289 thousands=$158,974,000)



(b) Wrong Calculation

Question: According to the ACCC's update report on transparency and competition in international money transfer services, with respect to respondents' top countries for outbound money transfers from Australia in 2023, can you calculate the total percentage breakdown of money transfers sent to the top 3 countries?

Answer: The total percentage breakdown of money transfers sent to the top 3 countries from Australia in 2023 is 32.0%, with the United States of America receiving 12.6%, England receiving 10.5%, and New Zealand receiving 8.9%.



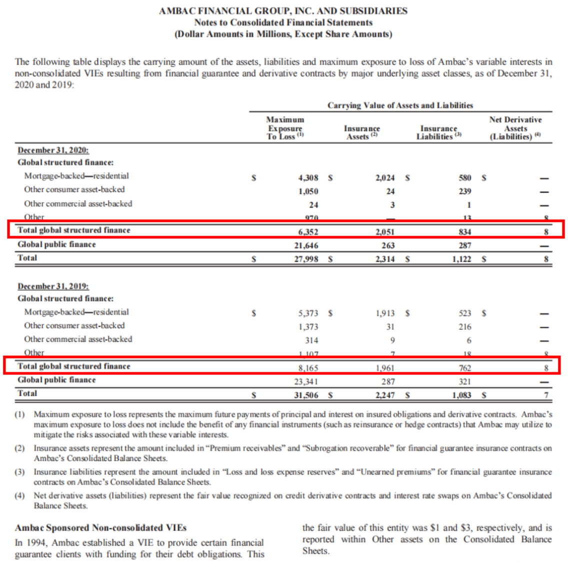
(c) Wrong Citation

Figure 13: Three case study examples to illustrate the potential errors that can occur in RGenCite during generation and citation.

{
"query_text": "According to Eastmoney, on December 28, 2024, Zheshang Securities completed the transfer of a 34.25% equity stake in Guodu Securities. Please analyze the opening and closing prices of Zheshang Securities on October 10.",
"expected_answer": "Based on the characteristics of candlestick charts, red indicates an upward trend and green indicates a downward trend. Therefore, Zheshang Securities' stock fell on October 10. The top of the candlestick body represents the opening price, and the bottom represents the closing price. Hence, the opening price on October 10 was 14.25, and the closing price was 13.55.",
"actual_answer": "According to the candlestick chart in Image 1, Zheshang Securities' opening price on October 10 was 13.58 yuan, and the closing price was 13.72 yuan."
}

Figure 14: An Error Case of Information Extraction from Candlestick Chart

{
'query_text': "According to AMBAC Financial Group, Inc. and Subsidiaries's Financial Statements, how much did the total global structured finance maximum exposure to loss change from December 31, 2019, to December 31, 2020?",
'expected_answer': 'decreased **$1813 millions**',
'actual_answer': 'The total global structured finance maximum exposure to loss decreased from $8,165 million on December 31, 2019, to $6,325 million on December 31, 2020, which is a decrease of $1,840 million.'
}



AMBAC FINANCIAL GROUP, INC. AND SUBSIDIARIES
Notes to Consolidated Financial Statements
(Dollar Amounts in Millions, Except Share Amounts)

The following table displays the carrying amount of the assets, liabilities and maximum exposure to loss of Ambac's variable interests in non-consolidated VIEs resulting from financial guarantee and derivative contracts by major underlying asset classes, as of December 31, 2020 and 2019:

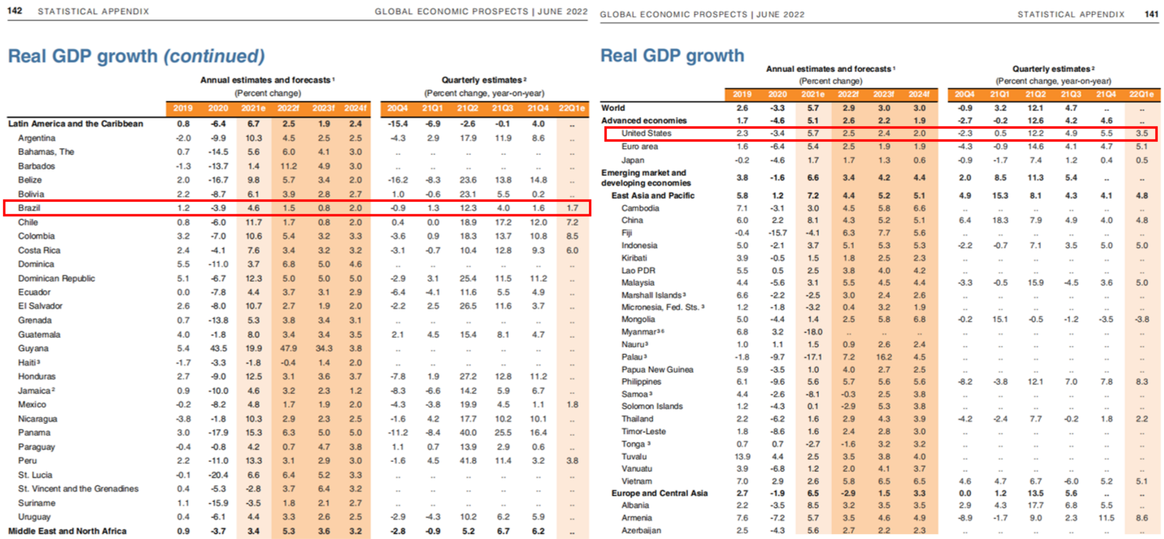
		Carrying Value of Assets and Liabilities		
	Maximum Exposure To Loss (1)	Insurance Assets (2)	Insurance Liabilities (3)	Net Derivative Assets (Liabilities) (4)
December 31, 2020:				
Global structured finance:				
Mortgage-backed—residential	$ 4,308	$ 2,024	$ 580	$ —
Other consumer asset-backed	1,050	24	239	—
Other commercial asset-backed	24	3	1	—
Other	970	—	13	8
Total global structured finance	6,352	2,051	834	8
Global public finance	21,646	263	287	—
Total	$ 27,998	$ 2,314	$ 1,122	$ 8
December 31, 2019:				
Global structured finance:				
Mortgage-backed—residential	$ 5,373	$ 1,913	$ 523	$ —
Other consumer asset-backed	1,373	31	216	—
Other commercial asset-backed	314	9	6	—
Other	1,107	7	18	8
Total global structured finance	8,165	1,961	762	8
Global public finance	23,341	287	321	—
Total	$ 31,506	$ 2,247	$ 1,083	$ 7

(1) Maximum exposure to loss represents the maximum future payments of principal and interest on insured obligations and derivative contracts. Ambac's maximum exposure to loss does not include the benefit of any financial instruments (such as reinsurance or hedge contracts) that Ambac may utilize to mitigate the risks associated with these variable interests.
(2) Insurance assets represent the amount included in "Premium receivables" and "Subrogation recoverable" for financial guarantee insurance contracts on Ambac's Consolidated Balance Sheets.
(3) Insurance liabilities represent the amount included in "Loss and loss expense reserves" and "Unearned premiums" for financial guarantee insurance contracts on Ambac's Consolidated Balance Sheets.
(4) Net derivative assets (liabilities) represent the fair value recognized on credit derivative contracts and interest rate swaps on Ambac's Consolidated Balance Sheets.

Ambac Sponsored Non-consolidated VIEs
In 1994, Ambac established a VIE to provide certain financial guarantee clients with funding for their debt obligations. This the fair value of this entity was $1 and $3, respectively, and is reported within Other assets on the Consolidated Balance Sheets.

Figure 15: An Error Case of Numerical Calculation on Financial Table

{
'query_text': 'What was the estimated quarterly GDP growth rate for the United States and Brazil in the first quarter of 2021 according to the global economic prospects, which country is has a higher GDP growth rate?',
'expected_answer': 'United States: 0.5%, Brazil:1.3%, Brazil is higher',
'actual_answer': 'The estimated quarterly GDP growth rate for the United States in the first quarter of 2021 was -0.9 percent. The estimated quarterly GDP growth rate for Brazil in the first quarter of 2021 was 0.8 percent. The United States had a lower GDP growth rate compared to Brazil in the first quarter of 2021.
}



142 STATISTICAL APPENDIX GLOBAL ECONOMIC PROSPECTS | JUNE 2022

Real GDP growth *(continued)*

	Annual estimates and forecasts [1] (Percent change)						Quarterly estimates [2] (Percent change, year-on-year)					
	2019	2020	2021e	2022f	2023f	2024f	20Q4	21Q1	21Q2	21Q3	21Q4	22Q1e
Latin America and the Caribbean	**0.8**	**-6.4**	**6.7**	**2.5**	**1.9**	**2.4**	**-15.4**	**-6.9**	**-2.6**	**-0.1**	**4.0**	**..**
Argentina	-2.0	-9.9	10.3	4.5	2.5	2.5	-4.3	2.9	17.9	11.9	8.6	..
Bahamas, The	0.7	-14.5	5.6	6.0	4.1	3.0	..	..	..	..	..	..
Barbados	-1.3	-13.7	1.4	11.2	4.9	3.0	..	..	..	..	..	..
Belize	2.0	-16.7	9.8	5.7	3.4	2.0	-16.2	-8.3	23.6	13.8	14.8	..
Bolivia	2.2	-8.7	6.1	3.9	2.8	2.7	1.0	-0.6	23.1	5.5	0.2	..
Brazil	1.2	-3.9	4.6	1.5	0.8	2.0	-0.9	1.3	12.3	4.0	1.6	1.7
Chile	0.8	-6.0	11.7	1.7	0.8	2.0	0.4	0.0	18.9	17.2	12.0	7.2
Colombia	3.2	-7.0	10.6	5.4	3.2	3.3	-3.6	0.9	18.3	13.7	10.8	8.5
Costa Rica	2.4	-4.1	7.6	3.4	3.2	3.2	-3.1	-0.7	10.4	12.8	9.3	6.0
Dominica	5.5	-11.0	3.7	6.8	5.0	4.6	..	..	..	..	..	..
Dominican Republic	5.1	-6.7	12.3	5.0	5.0	5.0	-2.9	3.1	25.4	11.5	11.2	..
Ecuador	0.0	-7.8	4.4	3.7	3.1	2.9	-6.4	-4.1	11.6	5.5	4.9	..
El Salvador	2.6	-8.0	10.7	2.7	1.9	2.0	-2.2	2.5	26.5	11.6	3.7	..
Grenada	0.7	-13.8	5.3	3.8	3.4	3.1	..	..	..	..	..	..
Guatemala	4.0	-1.8	8.0	3.4	3.4	3.5	2.1	4.5	15.4	8.1	4.7	..
Guyana	5.4	43.5	19.9	47.9	34.3	3.8	..	..	..	..	..	..
Haiti [3]	-1.7	-3.3	-1.8	-0.4	1.4	2.0	..	..	..	..	..	..
Honduras	2.7	-9.0	12.5	3.1	3.6	3.7	-7.8	1.9	27.2	12.8	11.2	..
Jamaica [2]	0.9	-10.0	4.6	3.2	2.3	1.2	-8.3	-6.6	14.2	5.9	6.7	..
Mexico	-0.2	-8.2	4.8	1.7	1.9	2.0	-4.3	-3.8	19.9	4.5	1.1	1.8
Nicaragua	-3.8	-1.8	10.3	2.9	2.3	2.5	-1.6	4.2	17.7	10.2	10.1	..
Panama	3.0	-17.9	15.3	6.3	5.0	5.0	-11.2	-8.4	40.0	25.5	16.4	..
Paraguay	-0.4	-0.8	4.2	0.7	4.7	3.8	1.1	0.7	13.9	2.9	0.6	..
Peru	2.2	-11.0	13.3	3.1	2.9	3.0	-1.6	4.5	41.8	11.4	3.2	3.8
St. Lucia	-0.1	-20.4	6.6	6.4	5.2	3.3	..	..	..	..	..	..
St. Vincent and the Grenadines	0.4	-5.3	-2.8	3.7	6.4	3.2	..	..	..	..	..	..
Suriname	1.1	-15.9	-3.5	1.8	2.1	2.7	..	..	..	..	..	..
Uruguay	0.4	-6.1	4.4	3.3	2.6	2.5	-2.9	-4.3	10.2	6.2	5.9	..
Middle East and North Africa	**0.9**	**-3.7**	**3.4**	**5.3**	**3.6**	**3.2**	**-2.8**	**-0.9**	**5.2**	**6.7**	**6.2**	**..**

GLOBAL ECONOMIC PROSPECTS | JUNE 2022 STATISTICAL APPENDIX 141

Real GDP growth

	Annual estimates and forecasts [1] (Percent change)						Quarterly estimates [2] (Percent change, year-on-year)					
	2019	2020	2021e	2022f	2023f	2024f	20Q4	21Q1	21Q2	21Q3	21Q4	22Q1e
World	**2.6**	**-3.3**	**5.7**	**2.9**	**3.0**	**3.0**	**-0.9**	**3.2**	**12.1**	**4.7**	**..**	**..**
Advanced economies	**1.7**	**-4.6**	**5.1**	**2.6**	**2.2**	**1.9**	**-2.7**	**-0.2**	**12.6**	**4.2**	**4.6**	**..**
United States	2.3	-3.4	5.7	2.5	2.4	2.0	-2.3	0.5	12.2	4.9	5.5	3.5
Euro area	1.6	-6.4	5.4	2.5	1.9	1.9	-4.3	-0.9	14.6	4.1	4.7	5.1
Japan	-0.2	-4.6	1.7	1.7	1.3	0.6	-0.9	-1.7	7.4	1.2	0.4	0.5
Emerging market and developing economies	**3.8**	**-1.6**	**6.6**	**3.4**	**4.2**	**4.4**	**2.0**	**8.5**	**11.3**	**5.4**	**..**	**..**
East Asia and Pacific	**5.8**	**1.2**	**7.2**	**4.4**	**5.2**	**5.1**	**4.9**	**15.3**	**8.1**	**4.3**	**4.1**	**4.8**
Cambodia	7.1	-3.1	3.0	4.5	5.8	6.6	..	..	..	..	..	..
China	6.0	2.2	8.1	4.3	5.2	5.1	6.4	18.3	7.9	4.9	4.0	4.8
Fiji	-0.4	-15.7	-4.1	6.3	7.7	5.6	..	..	..	..	..	..
Indonesia	5.0	-2.1	3.7	5.1	5.3	5.3	-2.2	-0.7	7.1	3.5	5.0	5.0
Kiribati	3.9	-0.5	1.5	1.8	2.5	2.3	..	..	..	..	..	..
Lao PDR	5.5	0.5	2.5	3.8	4.0	4.2	..	..	..	..	..	..
Malaysia	4.4	-5.6	3.1	5.5	4.5	4.4	-3.3	-0.5	15.9	-4.5	3.6	5.0
Marshall Islands [3]	6.6	-2.2	-2.5	3.0	2.4	2.6	..	..	..	..	..	..
Micronesia, Fed. Sts. [3]	1.2	-1.8	-3.2	0.4	3.2	1.9	..	..	..	..	..	..
Mongolia	5.0	-4.4	1.4	2.5	5.8	6.8	-0.2	15.1	-0.5	-1.2	-3.5	-3.8
Myanmar [3,6]	6.8	3.2	-18.0	..	..	..	..	..	..	..	..	..
Nauru [3]	1.0	1.1	1.5	0.9	2.6	2.4	..	..	..	..	..	..
Palau [3]	-1.8	-9.7	-17.1	7.2	16.2	4.5	..	..	..	..	..	..
Papua New Guinea	5.9	-3.5	1.0	4.0	2.7	2.5	..	..	..	..	..	..
Philippines	6.1	-9.6	5.6	5.7	5.6	5.6	-8.2	-3.8	12.1	7.0	7.8	8.3
Samoa [3]	4.4	-2.6	-8.1	-0.3	2.5	3.8	..	..	..	..	..	..
Solomon Islands	1.2	-4.3	0.1	-2.9	5.3	3.8	..	..	..	..	..	..
Thailand	2.2	-6.2	1.6	2.9	4.3	3.9	-4.2	-2.4	7.7	-0.2	1.8	2.2
Timor-Leste	1.8	-8.6	1.6	2.4	2.8	3.0	..	..	..	..	..	..
Tonga [3]	0.7	0.7	-2.7	-1.6	3.2	3.2	..	..	..	..	..	..
Tuvalu	13.9	4.4	2.5	3.5	3.8	4.0	..	..	..	..	..	..
Vanuatu	3.9	-6.8	1.2	2.0	4.1	3.7	..	..	..	..	..	..
Vietnam	7.0	2.9	2.6	5.8	6.5	6.5	4.6	4.7	6.7	-6.0	5.2	5.1
Europe and Central Asia	**2.7**	**-1.9**	**6.5**	**-2.9**	**1.5**	**3.3**	**0.0**	**1.2**	**13.5**	**5.6**	**..**	**..**
Albania	2.2	-3.5	8.5	3.2	3.5	3.5	2.9	4.3	17.7	6.8	5.5	..
Armenia	7.6	-7.2	5.7	3.5	4.6	4.9	-8.9	-1.7	9.0	2.3	11.5	8.6
Azerbaijan	2.5	-4.3	5.6	2.7	2.2	2.3	..	..	..	..	..	..

Figure 16: An Error Case of Multi-page Question

Annotation Guideline for QA Pairs Verification
GUIDELINE: Please verify the QA pairs produced by GPT-4o. For each sample, you may choose to retain, revise, or discard it. Your decision should be based on the following four criteria:
Verification Criteria:
1. Query Clarity: The query should be specific and unambiguous, targeting a particular topic in a document, and avoiding vague or overly general questions.
2. Answer Correctness: The answer must be factually correct and directly supported by the visual content. It should not include hallucinations or inferred information beyond what is presented. If calculation is involved, the answer should be accurate.
3. Category Appropriateness: The question should match its assigned category (e.g., table-numerical calculations). Mislabelled or ambiguous categories should lead to revision or rejection.
4. Correctness of Multi-Page Sources: For multi-page queries, if the answer is derived from multiple pages, all referenced page sources must be accurately identified.
Decision Rule: Retain the data if all criteria are met, revise it if there are minor issues (e.g., unclear query, incorrect category), and discard it if there are major errors or cannot be fixed reliably.

Table 27: Annotation guideline for QA Pairs Verification

Annotation guideline for the Rating-based Human Evaluation
GUIDELINE: Please evaluate the quality of the visual citation produced by the Retrieval-Augmented Generation system, rating it from score 0 to 5. Your rating should adhere to the following criteria:
Scoring Criteria:
0: Error image, or no reference/empty reference box.
1: Correct image, but selected the wrong area, containing no readable information or completely unrelated to the referenced content.
2: Correct image, area roughly related, but significantly offset, causing key information to be missing.
3: Correct image and roughly correct area, with offset or incomplete capture, information discernible but affecting reading experience.
4: Correct image and area, referenced information complete and accurate, with minor offset, or includes some redundant content (e.g., extra paragraphs, whitespace), but does not affect reading.
5: Perfect match. Image and area completely accurate, no offset, no redundancy, precise boundaries, referenced content clear and complete.

Table 28: Annotation guideline for the Rating-based Human Evaluation