FinRAGCiteBench-V: A Benchmark for Vision-Based RAG with Citation in the Financial Domain

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) plays a vital role in the financial domain, with widespread applications in areas such as real-time market analysis, trend analysis, and interest rate calculation. However, most existing RAG research in finance focuses predominantly on textual data, neglecting the rich visual information embedded in financial documents, causing a significant loss of valuable insights of financial analysis. Therefore, considering the characteristics of the financial domain, where accurate and high-quality multimodal retrieval is critical, we carefully design the **FinRAGCiteBench-V**, a vision-based RAG benchmark in financial domain, including (1) a bilingual retrieval corpus with 60,780 Chinese pages and 51,219 English pages from varieties of real-world documents; (2) a diversed bilingual financial dataset for evaluating LLMs' generation, covering seven different question categories; (3) a baseline **RGenCite** covering from retrieval to generation and vision-based citation. With comprehensive experiments on RGenCite, we can validate the benchmark's robustness and diversity, providing valuable insights for multimodal RAG systems in the financial domain.

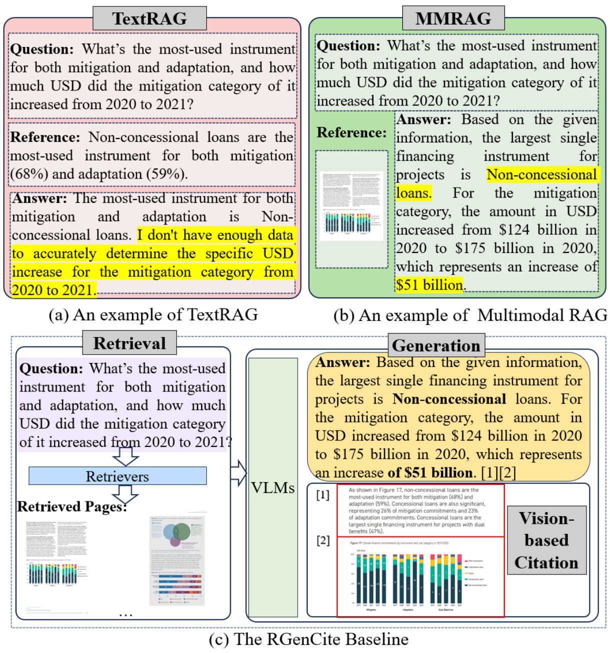


Figure 1: Comparison of TextRAG with MMRAG, and explanation of RGenCite baseline. Previous works focus on **(a) TextRAG** that losses essential graphical information, while **(b) MM RAG** retrieves both textual information and graphical information. Our **(c) RGenCite** baseline is based on MM RAG, containing both the retrieval phase and generation with vision-based citation.

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 their internal knowledge across various domains(Yang et al., 2024; Han et al., 2024; Zhang et al., 2024b). Especially in the financial domain, RAG plays a crucial role by providing LLMs with expert knowledge and time-sensitive information(Xiao et al., 2025; Shah et al., 2024). Thus, developing a comprehensive benchmark to evaluate RAG systems in the financial domain is essential. However, existing financial RAG benchmarks like Wang et al. (2024d) tend to focus primarily on textual corpora and datasets, overlooking the fact that the financial domain encompasses rich multimodal data. This includes line charts depicting price fluctuations over time and tables presenting detailed company financial statistics, which provide essential external knowledge for comprehensive financial analysis and decision-making. For example, as illustrated in Figure 1, consider the question: "*What is the most-used instrument for both mitigation and adaptation, and by how much did the USD amount for the mitigation category increase from 2020 to 2021?*" If we rely solely on the textual information in IFDC's financial report, we will lack sufficient data to accurately determine the specific USD increase for the mitigation category from 2020 to 2021. However, the bar chart at the bottom of this page provides the necessary information to answer this question.

In order to design such a benchmark, several key factors need to be taken into account:

Various Real-World Data Sources for Retrieval. In finance, it's essential to have varied data sources for accurate retrieval. By integrating text, tables, and visuals, RAG systems can gather broader information, resulting in more precise and contextually relevant answers(Zhang et al., 2024a; Suri et al., 2024). This reflects the complexity of real-world financial analysis, requiring the ability to retrieve information from diverse sources for comprehensive insights.

Diverse Types of Questions for Generation. Financial contexts require handling a variety of question types, from simple fact retrieval to complex tasks like calculations and comparisons using graphical or tabular data. RAG systems must be designed to extract insights from visual data, identify stock trends, forecast future performance, and analyze price volatility. They should generate accurate, contextually relevant answers across a wide range of financial scenarios.

Visual Citation for Reliable Attribution. In finance, answers must be supported by accurate references. Citations ensure precise attribution and answer faithfulness, crucial for RAG systems(Suri et al., 2024; Fierro et al., 2024). However, current citation methods focus on text, neglecting other formats. Therefore, it's essential to include visual data in citation techniques to improve reliability.

In light of these considerations, we propose the benchmark **FinRAGCiteBench-V**, a vision-based RAG benchmark with citations in the financial domain. The three key factors mentioned above have been carefully integrated into the design of the benchmark. First, we collect **various real-world data sources for retrieval** in both English and Chinese, including research reports, annual financial statements, prospectuses, academic papers, magazines, and news articles. In real-world scenarios, data from these sources are predominantly in PDF format, so we use PDF page images in our RAG system to capture both textual and visual information more effectively. Second, we have meticulously designed **diverse question types for generation**. This includes questions targeting text, tables, and charts, covering both single-page and multi-page queries, with answers involving either objective or subjective information, as well as requiring simple textual information extraction, or involving visual perception and complex reasoning. Finally, we implement **multimodal citation** inspired by Ma et al. (2024b). This approach requires models to generate relevant pages and identify the specific blocks within those pages, marking them as page-level and block-level citations, respectively. Additionally, we introduce automatic citation evaluation metrics to assess the recall and precision of these two types of citations, and test two types of methods to evaluate, namely box-bounding and image-cropping.

In line with these goals, the benchmark **FinRAGCiteBench-V** includes a bilingual corpus comprising 60,780 Chinese pages from 1,104 PDF files and 51,219 English pages from 1,105 PDF files. Additionally, a bilingual evaluation dataset, covering seven different categories and consisting of 855 Chinese question-answer pairs and 539 English ones, has been carefully designed. The initial data generation was done using GPT-4o, followed by meticulous manual annotation.

Based on this benchmark, we propose **RGenCite**, a simple and effective baseline, covering the retrieval, generation and citation stages in visual RAG systems. In the retrieval stage, experiments are conducted using both Optical Character Recognition (OCR) with text retrievers, such as JinaColBERT V2 (Jha et al., 2024), and multimodal retrievers, such as ColQwen2 (Faysse et al., 2024). Then, we use both proprietary multimodal LLMs, such as GPT-4o, and open-source multimodal LLMs, such as Qwen2.5-VL-72B-Instruct (Wang et al., 2024b) for the experiments on generation and citation.

Through these experiments on **RGenCite** baseline, we obtain several meaningful observations: (1) Multimodal retrieval systems outperforms the OCR-based text retrieval systems by a significant margin. This is likely due to the considerable loss of information in the OCR process of financial charts and tables, which are rich in domain-specific content, are converted into text. (2) While performing satisfactory on text-based inferences and direct information extraction from charts, numerical calculations from charts and tables present major challenges for the generation capabilities of multimodal LLMs. (3) Multimodal retrieval systems generally perform well with page-level citation, indicating their ability to correctly identify source images while generating answers. However, the model performs poorly with block-level citation. Among our two block-level citation evaluation methods, image-cropping and box-bounding, we find that image-cropping outperforms box-bounding when compared to human citation annotations. Therefore, precise attribution remains a significant challenge in multimodal RAG systems.

Our key contributions are as follows:

- We construct **FinRAGCiteBench-V**, a benchmark for vision-based RAG with citation in the financial domain, featuring diverse real-world data sources for retrieval, a variety of question types for generation, and visual citation for reliable attribution.

- We propose an automatic evaluation method for visual citation that does not rely on human-labeled ground truths, design corresponding metrics based on both of page-level citation and block-level citation, and test two types of evaluation methods: box-bounding and image-cropping.

- We propose a comprehensive baseline, **RGenCite**, for multimodal RAG systems, and conduct extensive experiments. These experiments include multimodal retrievers and textual retrievers in the retrieval stage, as well as multimodal proprietary and open-source LLMs in the generation and citation phase. Additionally, we test two types of citation methods and perform evaluations using self-designed automatic citation quality metrics.

2 Related Work

Benchmarking Retrieval Augmented Generation (RAG). Retrieval-Augmented Generation (RAG) has gained significant attention as it is an effective way leveraging external retrieval mechanisms to enhance the knowledge available to generative models.(Gao et al., 2023b; Lewis et al., 2020; Huang et al., 2023). With more and more RAG systems emerging, benchmarking and evaluating RAG models has become important in assessing their retrieval efficiency, generative performance, and factual accuracy (Chen et al., 2024b; Friel et al., 2024; Saad-Falcon et al., 2024). For domain-specific RAG benchmarks, in the financial domain, Wang et al. (2024c) proposes a benchmark including a textual dataset covering multiple financial topics and the automatic evaluation approach based on it.

Benchmarking Multimodal RAG. In the financial domain, where charts and graphs are crucial, text-only RAG benchmarks may overlook important information. Therefore, a multimodal RAG benchmark tailored to the financial domain is essential. Recently, several multimodal RAG benchmarks have been developed to ensure models can effectively handle diverse data types (Suri et al., 2024; Yu et al., 2024a). Similar multimodal RAG benchmarks have also been introduced in specialized fields, such as healthcare (Xia et al., 2024).

Citation and Its Evaluation. In specialized fields like finance, where precise domain knowledge is essential, 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 work has largely focused on textual citations, Ma et al. (2024b) introduced coordinate-based methods to enable multimodal citations—an approach particularly valuable in finance, where key insights often rely on charts, tables, and graphical data.

3 Dataset Construction

In order to construct a multimodal RAG dataset in financial domain for our benchmark, we first create a knowledge corpus from multiple real-world data sources to ensure the variety. Next, we generate the question-answer (QA) pairs based on the corpus using GPT-4o. We also implement strict quality control by manually annotating and verifying the QA pairs to ensure their accuracy.

3.1 Construction Pipeline

3.1.1 Knowledge Corpus Construction

To build the financial knowledge corpus, we collected data in PDF format from a variety of real-world sources in both Chinese and English, as demonstrated detaily in Appendix B, including:

(1) **Research reports** collected from websites like Qianzhan.com [1] , 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 FinGLM [2]dataset and official company and bank websites, which provide annual financial data in tabular form;

[1]https://Qianzhan.com/

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

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

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

(4) **Academic papers** that offer 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 and Forbus magazine, which offer reliable news, expert opinions, and financial analyses;

(6) **Financial news** from websites like China Daily and eastmoney[4].

We ultimately select 1,063 Chinese PDF files and 1,105 English PDF files from the data sources mentioned above, as shown detaily in Table 1. Each page of the PDFs was then converted into a single image, resulting in a retrieval corpus consisting of 60,780 Chinese pages and 51,219 English pages. By incorporating these diverse data types, we ensure that the knowledge corpus is both broad and reliable, providing a solid foundation for generating accurate and informative QA pairs.

3.1.2 QA pairs Generation

From the knowledge corpus, we select high-quality PDF pages and then generate a dataset of question-answer (QA) pairs using GPT-4o based on the selected pages, with predefined categories and carefully design examples provided as prompts. In terms of data scope, it includes both single-page and multi-page question answerings; Regarding data format, it covers question answering based on text, charts, and tabular data; As for answers, it contains both short and long answers; Considering the specific characteristics of the financial domain, we further categorize the QA dataset into seven main categories as follows. Please refer to Appendix C,

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

[4]https://www.eastmoney.com/

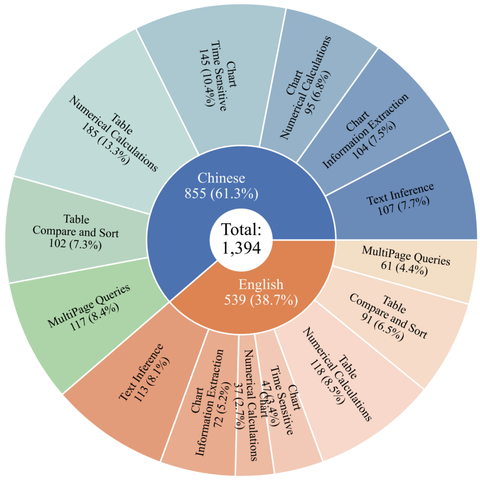


Figure 2: Statistics of Question Types in the Dataset

which provides examples for each category.

Text Inference: This includes tasks like summarization and information extraction, 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 chart data, such as calculating the changes of interest rate and summing costs.

Chart Time-Sensitive Queries: This involves 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/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.

3.2 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 the answers, and checking whether the page sources for multi-page queries were properly identified. Based on these criteria, we carefully filter and refine the from 11,328 generated QA paires, 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 2.

4 RGenCite

Based on the FinRAGCiteBench-V, we develop the baseline RGenCite, which covers the stages of retrieval, generation and vision-based citation.

4.1 Task Definition

In FinRAGCiteBench-V, we have a corpus of image pages generated from the PDF documents in the retrieval stage, defined as $\mathcal{C} = \{p_1, p_2, ..., p_i, ...\}$, where p_i represents the ith image page. Based on the corpus, we generate a dataset of QA pairs, 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 being the ground truth answer, t_i being the question type, and P_i being the set of corresponding page(s). Given a question q, we first use a retriever R to search the corpus $\mathcal{C}$ and retrieve the top-k relevant pages $\{r_1, r_2, ..., r_k\}$ as references. These top-k pages, along with the question q, are then input into a generation model M, which generates an answer a along with a set of citations. Each citation is defined as $c = (r, B)$, where r is a cited reference page, and $B = \{b_1, b_2, ..., b_n\}$ represents the exact blocks that contribute to the answer within the reference page r.

4.2 Retrieval

During the retrieval phase, we explore various multimodal retrievers alongside OCR-based text retrieval systems. We conduct a comprehensive evaluation of these two types of retrieval paradigms using multiple metrics to assess their performance from different perspectives.

Multimodal Retrievers. For the multimodal retrieval, we employ five different retrievers, namely ColQwen2 (Faysse et al., 2024), GME-Qwen2-VL-2B(Zhang et al., 2024c), GME-Qwen2-VL-7B, DSE-QWen2-2b-MRL-V1 (Ma et al., 2024a), VisRAG-Ret (Yu et al., 2024a). These retrievers are selected for their ability to handle vision-based documents, which often rely heavily on graphical and tabular content. By evaluating these retrievers, we aim to assess their effectiveness in retrieving relevant content from multimodal pages.

Text Retrievers. For the OCR-based text retrieval system, we use Marker (Paruchuri, 2024) to perform OCR recognition, converting PDF documents into JSON format. This process enables the extraction of textual information from image-based documents, which can then be used for further retrieval or analysis tasks. Subsequently, we test four different text retrievers, including BM25, JinaColBERT V2 (Jha et al., 2024), BGE-M3 (Chen et al., 2024a), and Multilingual-E5-large (Wang et al., 2024a), to evaluate their effectiveness in processing and retrieving relevant information from the extracted OCR text.

Metrics for Retrieval Evaluation. We test both the multimodal retrieval systems and the OCR-based text retrieval systems on Chinese and English datasets. The evaluation metrics include nDCG@5, nDCG@10, Recall@5, Recall@10, and MRR@10. Specifically, nDCG measures the ranking quality of retrieved results, Recall indicates the proportion of relevant documents found in the top-k results, and MRR reflects the average reciprocal rank of the first relevant document.

4.3 Generation

During the generation phase, we conduct experiments on both proprietary LLMs and open-source multimodal LLMs.

Multimodal LLMs. This includes GPT-4V, GPT-4o, GPT-4o-mini, Gemini-1.5-flash, Gemini-2.0-flash, Gemini-2.0-flash-exp, and Claude-3-5-Sonnet-20240620; while the later includes Qwen2-VL-72B-Instruct, Qwen2.5-VL-7B-Instruct, Qwen2.5-VL-72B-Instruct, Llama-3.2-90b-Vision-Instruct, Phi-3.5-vision-instruct, and MiniCPM-o-2.6. The prompt for LLMs' generation is shown in Appendix A.

Metrics for Answer Evaluation. To assess their ability to generate accurate responses based on visual elements, we use the rule-based metric ROUGE. Additionally, we employ GPT-4o to evaluate the metric Acc, assessing whether the generated responses align with the ground truth answers, thus ensuring their accuracy and consistency with the visual context. The prompt for this evaluation is shown in Appendix A.

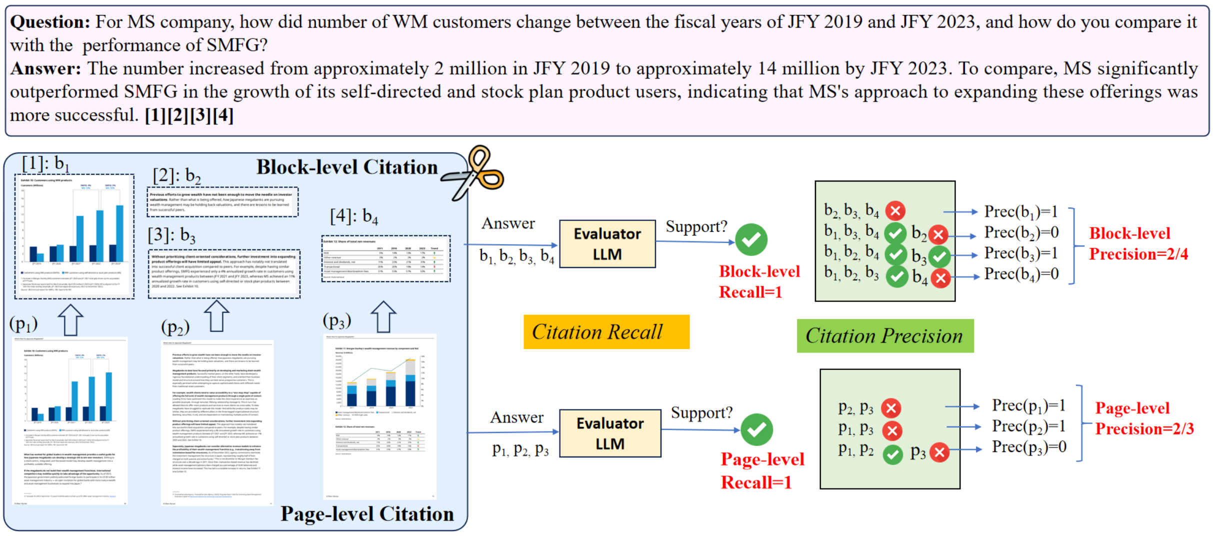


Figure 3: An example of the automatic evaluation of vision-based citation

4.4 Vision-Based Citation

When generating an answer, the model is required to specify the image pages it refers to and identify the exact regions within those pages that contribute to the response. To evaluate the ability of existing LLMs in handling vision-based citation, we use the top 10 retrieved pages to test the citation performance of LLMs listed in Section 4.3.

Citation Method. To achieve the simultaneous generation of answers and citations, we follow the vision-based citation method used in VISA (Ma et al., 2024b). Specifically, we input both the question and the reference images into the model, instructing it to generate the answer while simultaneously producing both page-level and block-level citations, denoted as $c = (r, \{b_1, b_2, ..., b_i, ...\})$. The page-level citation r refers to the reference page, while $\{b_1, b_2..., b_i, ...\}$ represents the block-level citations, indicating the specific regions of the answer within the page. Each block-level citation b_i is represented as a group of coordinates, i.e., $b_i = [x_1, y_1, x_2, y_2]$, where (x_1, y_1) denotes the coordinates of the top-left corner of the cited block b_i, and (x_2, y_2) denotes the coordinates of the bottom-right corner of b_i. The detailed citation format is displayed in Table 4.

In order to evaluate the vision-based citation quality of LLMs, we propose an automatic evaluation method that does not require ground truth and human annotation, based on two types of citation evaluation method, **box-bounding** and **image-cropping**. The first method involves drawing bounding boxes around the relevant regions, clearly marking the specific blocks of the image that inform the answer. The second method involves cropping the exact reference blocks of the image. For both methods, the corresponding bounding boxes or cropped images are automatically generated based on the coordinates model's outputs, which are then sent into the evaluator LLM to judge if they support the answer. It should be clarified that through experiments, we find that image-cropping has a higher consistency with human ratings, as explained in 5.2. Therefore, in subsequent experiments, the image-cropping method will be uniformly used for citation evaluation.

Citation Metrics. Inspired by Gao et al. (2023a), we evaluate both page-level citation and block-level citation using the two following types of metrics, and the corresponding evaluation process is illustrated using an example in Figure 3:

Recall evaluates whether the cited images are sufficient for attributing the answer. In the case of block-level citation, if the union of all cited blocks $B = \{b_1, b_2, ..., b_n\}$, called as the citation set of an answer a, is enough to support the answer a, the recall is rated 1, otherwise, it is rated 0. The evaluation of recall follows this formula:

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

The evaluation for page-level citation is similar.

Precision evaluates the proportion of citations in the cited set that are essential for supporting an answer. Specifically, in block-level citation, the cited block b_i is considered irrelevant if and

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

only if the b_i itself cannot independently support the answer, and the union of all other cited blocks $\{b_1, b_2, ..., b_{i-1}, b_{i+1}, ...\}$, in the citation set B, is sufficient to support the answer a, which can be described as:

$$\text{irrel}(B, b_i, a) = (b_i \nrightarrow a) \wedge ((B \setminus \{b_i\}) \rightarrow a)$$

The proportion of non-irrelevant blocks is defined as the citation precision of the citation set B for answer a, as illustrated in the formula:

$$\text{precision}(B, a) = \frac{|B \setminus \{b_i \mid \text{irrel}(B, b_i, a) = 1\}|}{|B|}$$

It should be noted that the precision of each citation is evaluated only when the recall of the citation set to which it belongs is judged to be 1; otherwise, the precision is 0. The evaluation for page-level citation is similar.

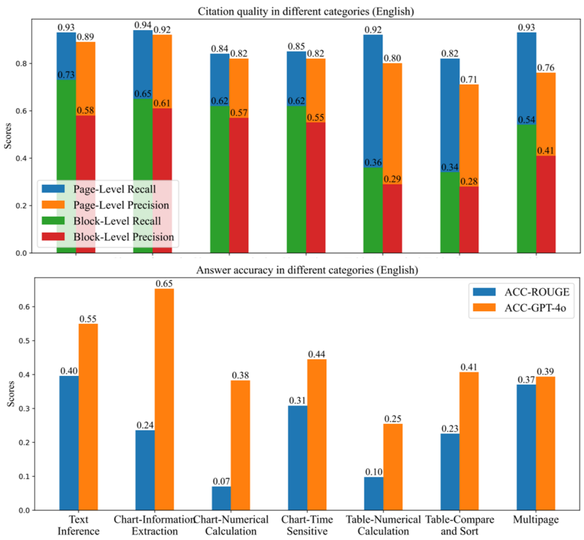


Figure 4: The comparision of answer accuracy and citation quality between different question categories.

5 Experimental Results and Analysis

We conduct primary experiments in both retrieval and generation with citations. First, in the retrieval phase, we evaluate both multimodal retrievers and OCR-based textual retrieval systems, utilizing Marker as the OCR tool. Second, for the generation and citation phases, we select the best-performing retriever, and use the top-k retrieved pages as reference inputs to the model, with $k = 10$ in the main experiments. To assess the answers, we employ ROUGE and GPT-4o evaluation metrics (Accuracy), while citation quality is measured using both page-level and block-level recall and precision, denoted as *P_Rec*, *P_Prec*, *B_Pec*, *B_Prec*, respectively. Finally, we perform detailed analysis based on the experiments.

5.1 Main Results

Retrieval. In the retrieval phase, we find that **multimodal retriever outperforms the OCR-based text retrieval system** across all evaluation metrics. As demonstrated in Table 2, the best multimodal retriever, ColQwen2, achieves recall@10 of 90.13 in Chinese tasks, and 85.86 in English ones, while the best text retriever BGE-M3 only reaches 42.71 in Chinese and 36.36 in English. This highlights the superiority of multimodal systems, which combine the strengths of different data types, especially in the financial domain where information is often conveyed through charts and tables.

Generation. In the generation phase, as shown in Table 3, we observe that **proprietary models outperform open-source models**, highlighting the challenges that open-source multimodal models face in handling complex multi-image inference tasks. To better understand the performance of LLMs on different types of tasks, we analyze the generation and citation performance for LLMs on the senven types of financial question in FinRAGCiteBench-V. The statistics, illustrated in Figure 4, show that, **LLMs excel in tasks in-**

Model	Chinese						English					
	ROUGE	ACC	P_Rec	P_Prec	B_Rec	B_Prec	ROUGE	Acc	P_Rec	P_Prec	B_Rec	B_Rec
Proprietary LLMs												
GPT-4o	**33.61**	**49.59**	**88.07**	**84.52**	**54.97**	**48.32**	24.66	43.41	**89.98**	**81.81**	54.17	**44.66**
GPT-4V	33.70	46.43	87.95	83.03	36.23	24.97	22.76	44.71	89.24	80.54	**55.43**	42.69
GPT-4o-mini	20.93	18.54	78.51	56.74	20.43	12.71	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	26.65	38.34	87.81	83.96	28.37	24.23	21.26	**48.79**	89.80	83.92	21.52	17.48
Gemini-2.0-flash-exp	28.00	44.91	86.78	82.97	34.31	29.81	21.83	46.01	89.61	85.22	20.41	17.23
Claude-3-5-Sonnet	23.57	44.80	56.73	53.31	27.01	24.31	20.92	43.41	79.78	77.99	36.73	34.49
Open-Source Multimodal LLMs												
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	65.38	62.27	9.71	8.19	19.47	36.36	51.21	49.25	18.74	15.72
Qwen2.5-VL-72B-Instruct	22.83	30.41	58.25	51.31	10.64	9.49	21.98	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
Llama-3.2-90B-V-Instruct	9.00	13.87	14.71	11.39	13.29	10.70	9.76	27.64	4.82	4.06	2.04	1.58

Table 3: Results for Generation and Citation in both languages. The best results are highlighted in **bold**

volving text inference and visual information extraction, but **struggle with numerical calculations from charts and tables**. This suggests that complex visual reasoning problems in specialized domains like finance are areas where LLMs need to make breakthroughs.

Vision-based Citation. In terms of citation, as shown in Table 3, **most LLMs perform well in page-level citations**, demonstrating their ability to accurately identify relevant pages from the provided reference documents. However, they **face significant difficulties with block-level citation, especially for open-source LLMs compared with proprietary ones**. This highlights the challenge of attributing information to specific regions within the pages, suggesting that many open-source LLMs still have notable limitations in precise citation generation. It also underscores the ongoing challenge of achieving accurate visual attribution within images, particularly when it comes to pinpointing specific regions or blocks of information.

5.2 Consistency of Citation Evaluation Methods with Human

To investigate the validity of two block-level citation evaluation methods—box-bounding and image-cropping—we compare their results with human annotations for consistency. Specifically, we sample 100 data instances and have human evaluators score the citations on a scale from 0 to 5. For the block-level recall *B_Rec* and block-level precision *B_Prec* obtained by both methods, we calculate $F1 = 2 \times \frac{\text{B\_Prec} \times \text{B\_Rec}}{\text{B\_Prec} + \text{B\_Rec}}$, as a comprehensive metric for block-level citations, facilitating the calculation of correlation with human scores. The result show that the Pearson correlation between box-bounding and human scores is 38.13%, while the correlation between image-cropping and human scores is 74.47%. These results suggest that image-cropping is more reliable for block-level citations.

5.3 Case Study

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, which is show in Appendix D. 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 11 (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 11 (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 11 (c).

6 Conclusion

In this paper, we propose FinRAGCiteBench-V, a benchmark for vision-based RAG with citations in the financial domain. Through extensive and meticulous experiments, our FinRAGCiteBench-V benchmark reveals several critical issues existing in current visual RAG systems. It serves as a powerful tool for researchers and developers to identify the weaknesses of existing models and provides clear directions for further improvement.

Limitations

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

References

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.

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.

Constanza Fierro, Reinald Kim Amplayo, Fantine Huot, Nicola De Cao, Joshua Maynez, Shashi Narayan, and Mirella Lapata. 2024. Learning to plan and generate text with citations. 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 11397–11417. Association for Computational Linguistics.

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

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.

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.

Vik Paruchuri. 2024. Marker.

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.

Shalin Shah, Srikanth Ryali, and Ramasubbu Venkatesh. 2024. Multi-document financial question answering using llms. *CoRR*, abs/2411.07264.

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. *CoRR*, abs/2409.12191.

Shuting Wang, Jiongnan Liu, Shiren Song, Jiehan Cheng, Yuqi Fu, Peidong Guo, Kun Fang, Yutao Zhu, and Zhicheng Dou. 2024c. Domainrag: A chinese benchmark for evaluating domain-specific retrieval-augmented generation. *CoRR*, abs/2406.05654.

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

Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. 2024. Mmed-rag: Versatile multimodal RAG system for medical vision language models. *CoRR*, abs/2410.13085.

Mengxi Xiao, Zihao Jiang, Lingfei Qian, Zhengyu Chen, Yueru He, Yijing Xu, Yuecheng Jiang, Dong Li, Ruey-Ling Weng, Min Peng, Jimin Huang, Sophia Ananiadou, and Qianqian Xie. 2025. Enhancing financial time-series forecasting with retrieval-augmented large language models. *Preprint*, arXiv:2502.05878.

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*.

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*.

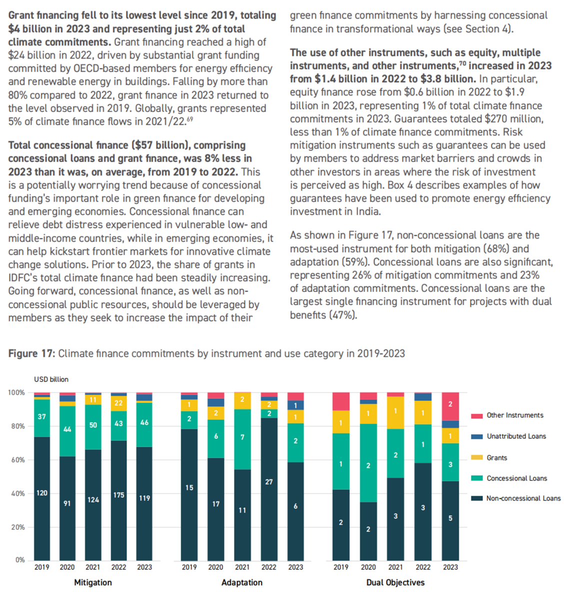
Junyuan Zhang, Qintong Zhang, Bin Wang, Linke Ouyang, Zichen Wen, Ying Li, Ka-Ho Chow, Conghui He, and Wentao Zhang. 2024a. OCR hinders RAG: evaluating the cascading impact of OCR on retrieval-augmented generation. *CoRR*, abs/2412.02592.

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. 2024b. 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. 2024c. GME: improving universal multimodal retrieval by multimodal llms. *CoRR*, abs/2412.16855.

A Prompts for Generations and Evaluations

We provide the prompts for both generating answer with visual citations, and the evaluation on the an-



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 5: An example of research report

swer and citations, shown in Table 4, 5, 6, 7, 8.

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 corpus, shown in Figure 5, 6, 7, 8, 9, 10.

C Examples of Seven Categories of QA Dataset

In this section, we provide an example for each category of questions, shown in Table 9, 10, 11, 12, 13, 14, 15.

C.1 Text Inference:

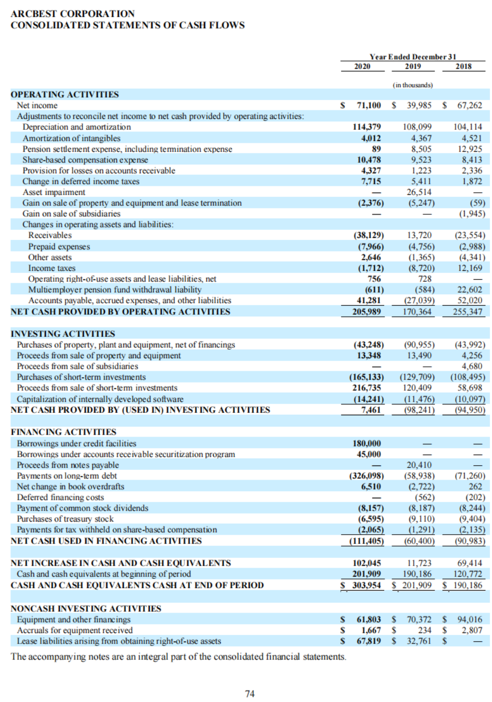
This category involves tasks such as summarization and information extraction from text. For example, deriving key insights from large volumes of text or identifying specific pieces of information, such as financial data or trends, within the content.

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.



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 6: An example of financial statements

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

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 4: 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 5: 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 6: 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 7: 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 8: Prompt for Block-Level Citation Evaluation using Image-Cropping

Query: What percent of account holders in Europe are using LinkedIn for finding job?

Category: Text Inference

Answer: Peter Ventress was appointed as the Committee Chairman, and Richard Pennycook retired.

Reference Image:

Table 9: An Example of Chart-Information Exraction 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:

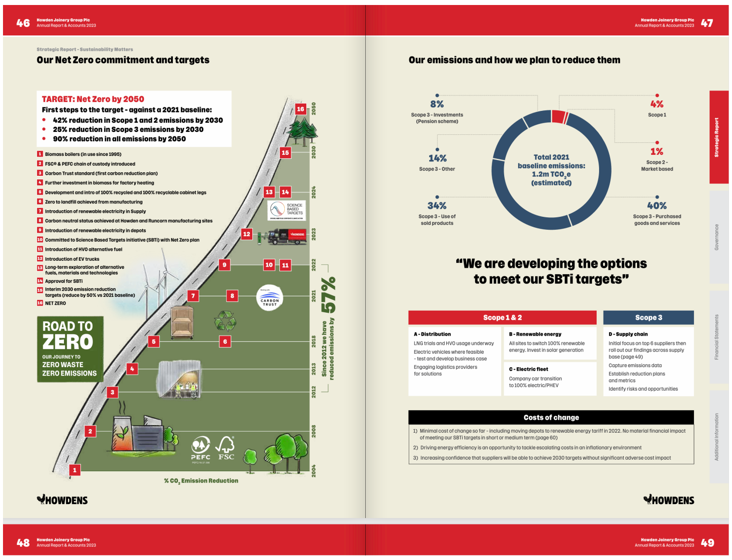


Table 10: An Example of Chart-Information Exraction Question

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:	

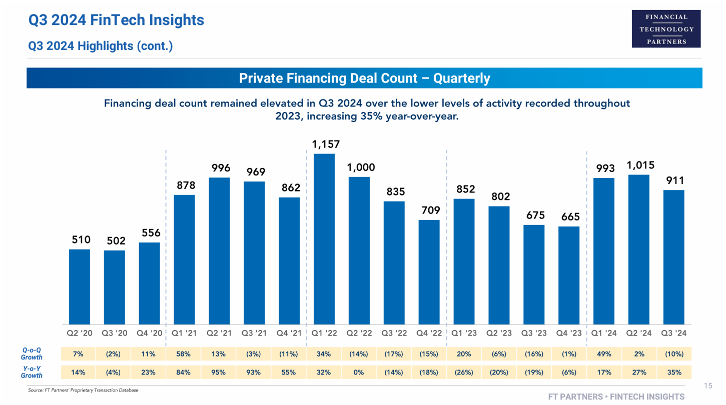


Table 11: An Example of Chart-Numerical Calculations Question

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:	

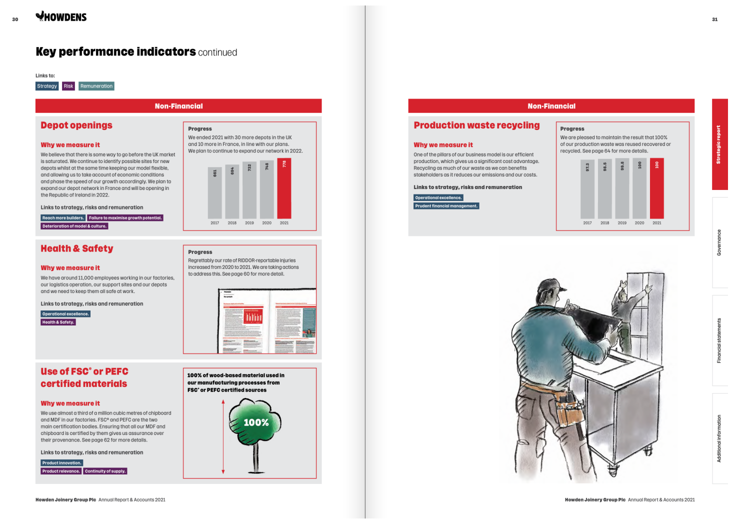


Table 12: An Example of Chart-Time Sensitive Question

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.

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:

	2023		2022	
Group	Charged $	Outstanding at year end $	Charged $	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 13: An Example of Table-Numerical Calculations Question

D Case Study

These are three case study examples to illustrate the potential errors that can occur in RGenCite during generation and citation.

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 14: 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:	

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.

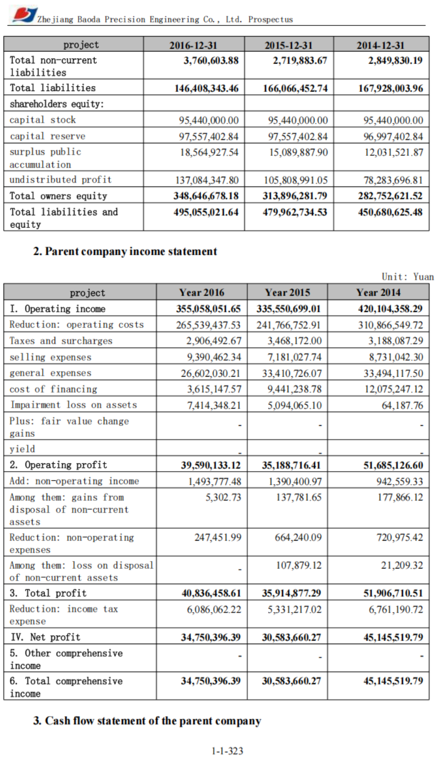
Year Ended December 31, 2023	Legacy Financial Guarantee Insurance	Specialty Property & Casualty Insurance	Insurance Distribution	Corporate & Other	Consolidated
Revenues:					
Net premiums earned	$ 26	$ 52			$ 78
Commission income			$ 51		51
Program fees		8			8
Net investment income	127	4	—	$ 9	140
Net 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

Table 15: An Example of Table-Comparison and Sorting 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

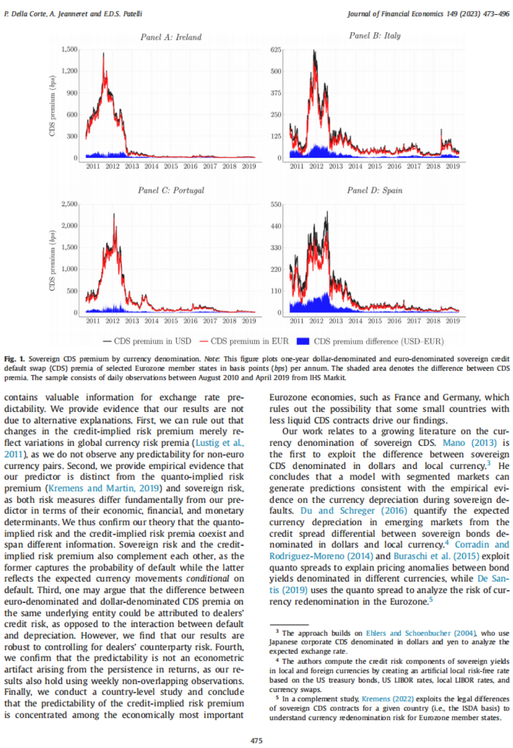
Unit: Yuan

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 7: An example of prospectus



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 8: An example of finance-related academic paper

2 · FINANCIAL TIMES · Friday 28 June 2024

INTERNATIONAL

Artificial intelligence

Gates plays down AI energy use fears

Microsoft founder argues the technology will speed transition to green power

'The tech companies are the people willing to pay a premium and to help bootstrap green energy capacity'

GLOBAL INSIGHT · UK · Alan Beattie

Tug of war looms for Labour between EU trade and US security

If US pressure increases, it could deter the UK from alignment with Brussels

OECD report

Teens lack financial literacy and maths for digital economy

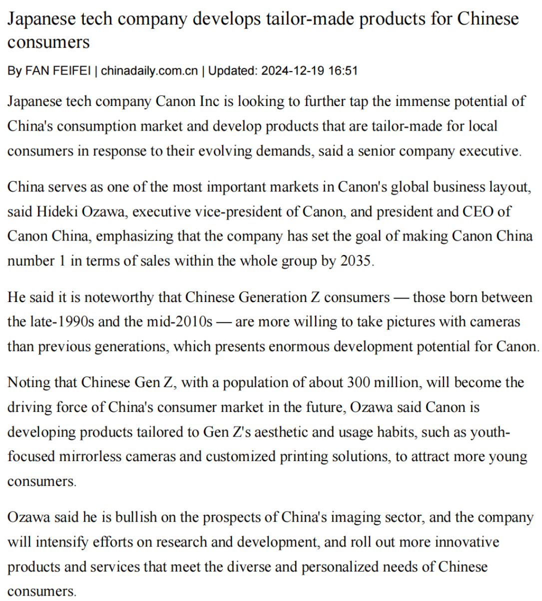
'These are uncharted territories. Ten years ago we wouldn't have talked [of] crypto or finfluencers'

Technology industry

India's manufacturing push held back by China visa bottleneck

MAKE A WISE INVESTMENT

Figure 9: An example of financial magazine



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 10: An example of financial news

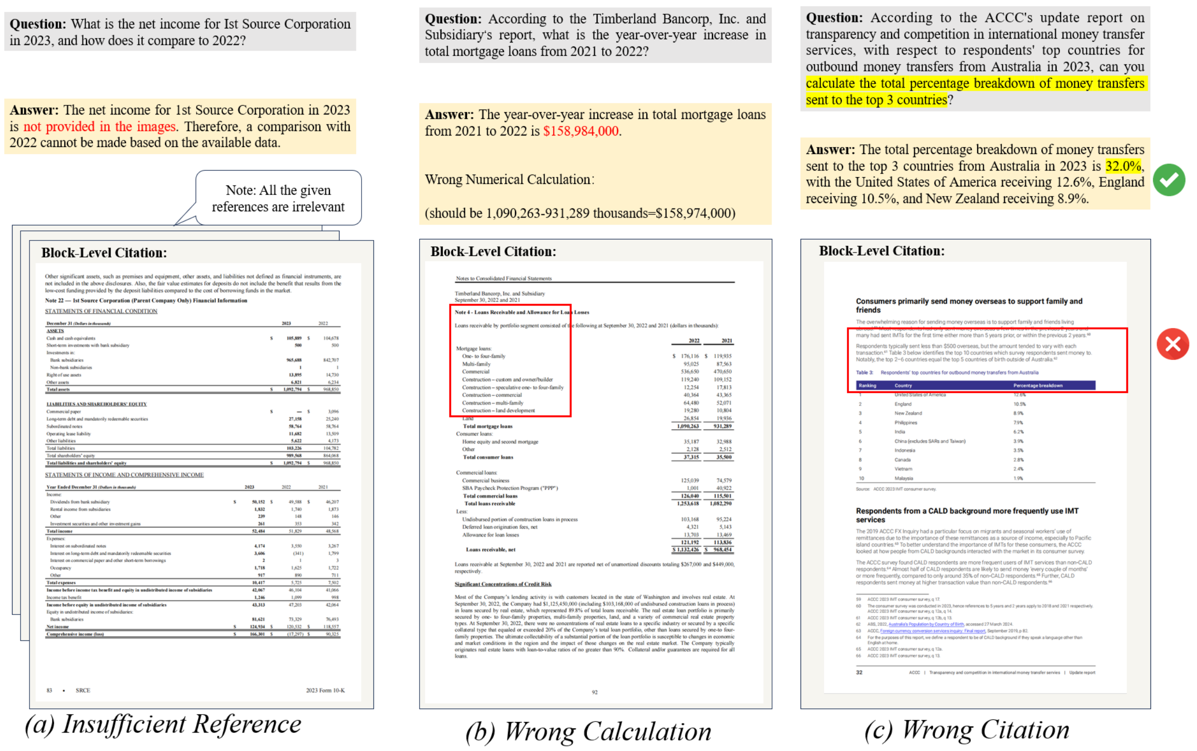


Figure 11: Three case study examples to illustrate the potential errors that can occur in RGenCite during generation and citation.