When Tables Go Crazy: Evaluating Multimodal Models on French Financial Documents

Virginie Mouilleron[1] **Théo Lasnier**[1,2] **Anna Mosolova**[1] **Djamé Seddah**[1]
[1] Inria Paris, France
[2] Sorbonne Université, Paris, France
{virginie.a.mouilleron, theo.lasnier, anna.mosolova, djame.seddah} @inria.fr

Abstract

Vision-language models (VLMs) perform well on many document understanding tasks, yet their reliability in specialized, non-English domains remains underexplored. This gap is especially critical in finance, where documents mix dense regulatory text, numerical tables, and visual charts, and where extraction errors can have real-world consequences. We introduce SCRIBE FINANCE, the first multimodal benchmark for evaluating French financial document understanding. The dataset contains 1,204 expert-validated questions spanning text extraction, table comprehension, chart interpretation, and multi-turn conversational reasoning, drawn from real investment prospectuses, KIDs, and PRIIPs. We evaluate six open-weight VLMs (8B–124B parameters) using an LLM-as-judge protocol. While models achieve strong performance on text and table tasks (85–90% accuracy), they struggle with chart interpretation (34–62%). Most notably, multi-turn dialogue reveals a sharp failure mode: early mistakes propagate across turns, driving accuracy down to roughly 50% regardless of model size.

These results show that current VLMs are effective for well-defined extraction tasks but remain brittle in interactive, multi-step financial analysis. SCRIBE FINANCE offers a challenging benchmark to measure and drive progress in this high-stakes setting.

Keywords: Financial documents, Multimodal evaluation, Vision Language Models

1. Introduction

The 2008–2009 global financial crisis exposed major failures in transparency and regulatory oversight across financial markets, prompting a coordinated international response to strengthen disclosure and investor protection requirements (G20, 2009). In the European Union, these reforms were subsequently formalized through regulations such as *Markets in Financial Instruments Directive (MiFID II)* and *Packaged Retail and Insurance-based Investment Products (PRIIPs)*, requiring asset management companies to issue standardized prospectuses at the beginning of the fiscal year and subjecting them to end-of-year review by regulatory authorities (European Commission, 2014; European Union, 2014). Following the implementation of these regulations, the volume of regulated disclosure documents has become substantial in all major financial jurisdictions. In the European Union and the United States alone, tens of thousands of prospectuses and prospectus-like documents, including amendments and standardized investor disclosures, are produced each year, illustrating the scale of financial documentation that must be reviewed and interpreted.

Because of their unprecedented level of performance in many text understanding tasks (Grattafiori et al., 2024), Large Language Models (LLMs) have become the central component of the modern Natural Language Processing (NLP) arsenal. Despite this progress, their evaluation in certain specialized domains remains uneven. Finance, for example, presents several particular challenges: documents are long, terminology is technical, and information is often distributed across text, tables, and charts. Moreover, most evaluation resources focus on English, leaving models' capabilities in other languages, particularly in domain-specific contexts, largely untested. This gap is especially problematic for regulatory and advisory applications where extraction accuracy is critical: a financial advisor querying a prospectus for the entry fee of a specific share class cannot tolerate hallucinated percentages.

French financial documents exemplify these challenges. Investment prospectuses, which describe potential returns and risks of financial products, can span 10 to 600+ pages and combine dense legal prose with complex tabular data and visual elements. Deploying LLMs in such challenging scenarios requires rigorous evaluation of their ability to locate and extract precise information, a prerequisite for higher-level tasks like summarization or compliance verification.

Despite several efforts to build specialized benchmarks in French for the financial domain (Faysse et al., 2025; Xue et al., 2025), the proposed datasets remain small in scale and limited in coverage ($\approx$ 200 examples; see Table 1), making it hard to assess whether state-of-the-art models are ready for these high-stakes applications.

To address this gap, we introduce SCRIBE FINANCE, a multimodal benchmark dataset of

Proceedings of the 7th Financial Narrative Processing Workshop @ LREC 2026, pages 12–27

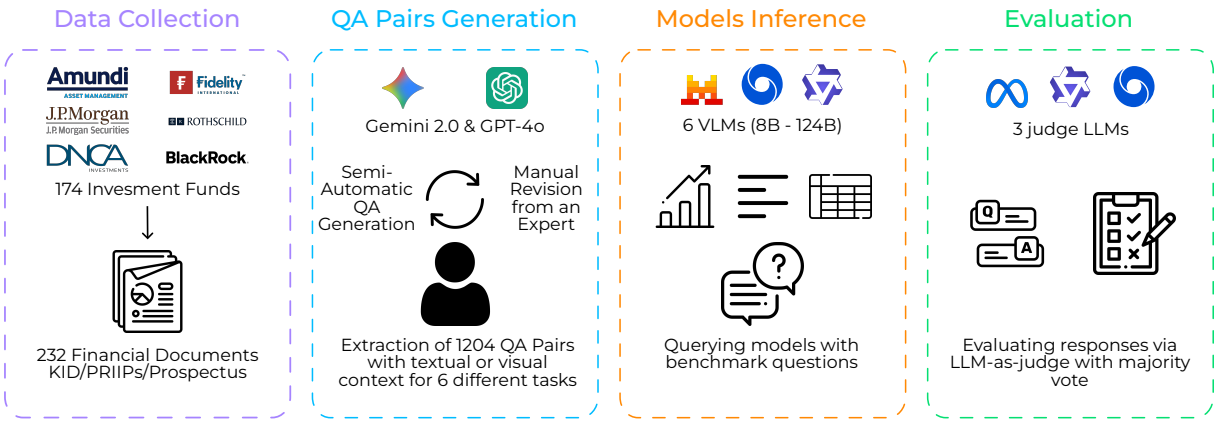


Figure 1: Overview of the SCRIBE FINANCE benchmark construction and evaluation pipeline. French financial documents (prospectuses, KIDs, PRIIPs) are collected from asset management companies, then processed to generate question-answer pairs spanning text, tables, and charts. Six Vision-Language Models are evaluated on these tasks, with responses assessed using a majority-vote LLM-as-judge protocol.

1,204 questions designed to evaluate VLMs on French financial document understanding. The dataset spans multiple question types (open-ended, Yes/No, True/False, and multi-turn conversational) and input modalities (text, tables, and charts). Questions range from named entity extraction to complex reasoning requiring integration of information across document sections. Figure 1 provides an overview of the benchmark construction and evaluation pipeline.

We evaluate six open-weight, state-of-the-art Vision-Language Models (VLMs) from three model families, spanning scales from 8B to 124B parameters, using an LLM-as-judge evaluation protocol. Results show that while models perform well on text-based questions (∼88–90%) and achieve moderate to strong performance on table comprehension (∼52–86%), chart interpretation remains challenging across all models (∼34–62%). More critically, the multi-turn conversational task reveals a systematic failure mode: errors propagate across dialogue turns, causing accuracy to collapse to approximately 50% (∼46–59%) regardless of model size. This behavior raises concerns about the reliability of current VLMs in interactive financial analysis settings.

Together, these results suggest that while VLMs are effective for well-scoped information extraction, they remain fragile when reasoning must be maintained across visual modalities and conversational context. SCRIBE FINANCE provides a benchmark for quantifying these limitations and tracking progress on French financial document understanding. Our main contributions are:

- SCRIBE FINANCE, the first multimodal benchmark for French financial document understanding, consists of 1,204 expert-validated questions spanning text extraction, table comprehension, chart interpretation, and multi-turn dialogue.[1]
- A systematic evaluation of six state-of-the-art VLMs, showing strong performance on text and tables but persistent weaknesses in chart interpretation and conversational settings.
- Empirical evidence that error propagation in multi-turn dialogue negates scaling benefits, with model accuracy converging to approximately 50% regardless of parameter count.

2. Related Works

2.1. French Evaluation Resources

Most NLP evaluation benchmarks target English, but some efforts have introduced resources for French-language evaluation as well. General-purpose question answering benchmarks, such as FQuAD (d'Hoffschmidt et al., 2020) and PIAF (Keraron et al., 2020), largely derived from Wikipedia and inspired by their English counterpart SQuAD (Rajpurkar et al., 2016), have played an important role in enabling French-language QA evaluation. More recent efforts have extended evaluation beyond general domains, including FrenchMedMCQA (Labrak et al., 2023) for medical reasoning and French CrowS-Pairs (Névéol et al., 2022; Nangia et al., 2020) for bias assessment. Additional datasets such as Newsquadfr[2] further explore model performance on journalistic and informal French content.

[1] The dataset and its accompanying resources can be accessed here https://github.com/dseddah/Scribe_finance/

[2] https://huggingface.co/datasets/lincoln/newsquadfr

This section does not attempt to provide an exhaustive survey of French evaluation datasets. Instead, these resources illustrate that, despite growing coverage of French language understanding, existing benchmarks largely focus on short, text-only inputs and general or domain-specific knowledge. In contrast, our work targets multimodal, long-form financial documents and evaluates model behavior in high-stakes, document-centric settings.

2.2. NLP Work in the Finance Domain

General-purpose Multilingual Benchmarks Including French While English-centric evaluation remains the norm, several multilingual benchmarks provide partial coverage of French. Datasets such as MKQA (Longpre et al., 2021), XQA (Liu et al., 2019), and MIRACL (Zhang et al., 2023) provide cross-lingual question-answering benchmarks primarily based on Wikipedia, enabling evaluation of multilingual transfer across a range of languages, including French. These resources have played an important role in advancing multilingual evaluation, but they focus on short, text-only inputs and do not address the challenges posed by long, structured, or domain-specific documents.

Specialized Financial Benchmarks in English The financial domain has also motivated the development of specialized benchmarks targeting numerical and document-level reasoning. TAT-QA (Zhu et al., 2021) and FinQA (Chen et al., 2022a) evaluate reasoning over financial reports by combining textual passages with tabular data, requiring models to perform arithmetic and logical operations rather than simple extraction. More recent datasets such as ConvFinQA (Chen et al., 2022b) and PACIFIC (Deng et al., 2023) extend this setting to multi-turn conversational scenarios, exposing the challenges of numerical reasoning and context tracking in dialogue-based interactions. Very recently, Lithgow-Serrano et al. (2025) introduced a banking-domain retrieval-augmented generation benchmark focusing on full documents comprising approximately 600 question-answer pairs.

Specialized Financial Benchmarks in French Recently, Faysse et al. (2025) shared a small dataset focusing on answering questions partly about financial tables in French documents. In parallel, FAMMA (Xue et al., 2025) presents a multilingual containing 9% of french content, multimodal financial benchmark derived from university-level instructional and assessment materials across eight core finance areas, requiring joint reasoning over text, tables, and charts, and proving challenging even for strong models. See Table 1 for a detailed comparison of these datasets.

Dataset	Size	Question type	Context type
Faysse et al. (2025)	210	Retrieval	Table
Xue et al. (2025)	190	Open, MCQ	Table, None
Ours	1,204	Open, MCQ, TFQ, Yes/No	Table, Chart, Text

Table 1: Comparison between existing financial benchmarks and our newly proposed SCRIBE FINANCE (see Section 3). *MCQ* = multiple-choice questions, *TFQ* = True/False questions, *Yes/No* = Yes/No questions.

Despite these advances, existing financial benchmarks remain limited in several respects: they are predominantly English-only, focus on relatively short excerpts rather than full-length regulatory documents, and largely exclude multimodal inputs such as charts. In contrast, our work targets French financial prospectuses, which are long, multimodal, and legally constrained, and evaluates model behavior in high-stakes document understanding and conversational settings.

3. Designing SCRIBE FINANCE

Building SCRIBE FINANCE required balancing realism, scale, and annotation reliability. French financial prospectuses are long, highly structured, and repetitive documents that combine dense legal text with tables and charts, frequently spanning hundreds of pages. Rather than treating these documents as monolithic inputs, we extract excerpts of varying lengths (0.5-30 pages) and focus on evaluating a model's ability to accurately locate and extract specific, document-grounded information, which is a prerequisite for reliable downstream reasoning in financial settings.

The dataset was constructed from publicly available French financial documents collected from multiple asset management companies, including *prospectuses, Key Information Documents (KIDs)*, and *Packaged Retail and Insurance-based Investment Products (PRIIPs)* published over the past 15 years.[3]

In the next section, the approach to generate questions for text-based and image-based tasks is described in detail.

[3]The documents were collected from asset management companies and are publicly available under EU and U.S. financial regulations (European Commission, 2014; European Union, 2014), which require publication for investor protection and public transparency. They are accessible without authentication or paywalls on official issuer or regulator websites and contain no private or personally identifiable information.

Question/Context Type ↓		Task type ↓				
		Text-Based	Image-Based			
		Text	Tables	Charts	Conv.	Special Case
Question Type	Open	501	248	94	0	19
	Yes/No	0	213	28	0	3
	True/False	0	27	6	0	0
	MCQ	0	0	0	65	0
Total (per question type)		501	488	128	65	22
Context Type	Small text	38	0	0	0	0
	Medium text	146	0	0	0	6
	Large text	108	0	0	30	16
	Very large text	14	0	0	0	0
	Document-wise (KID)	184	0	0	0	0
	Table	11	442	0	15	0
	Table & Small text	0	35	0	20	0
	Table & Medium text	0	11	0	0	0
	Chart	0	0	73	0	0
	Chart & Small text	0	0	55	0	0
Total (per context type)		501	488	128	65	22

Table 2: Distribution of the SCRIBE FINANCE benchmark (1,204 questions) across question types and context modalities. The dataset spans text-based questions (501 questions) and image-based questions including tables, charts, multi-turn conversations, and special cases, for a total of 703 questions with an associated image. Open-ended questions dominate (862 questions), with binary (Yes/No, True/False) and conversational MCQ formats targeting specific reasoning challenges. *Conv.* = multi-turn conversation questions.

3.1. Question Generation

Question construction followed an iterative, semi-automatic process designed to identify salient, extractable financial information. Two LLMs (GPT-4o and Gemini-2.0) assisted in generating candidate question and answers, after which all outputs were reviewed and revised by a human annotator. When necessary, input contexts were expanded to ensure completeness and faithful grounding in the source documents. The specific design choices for each question type are described below.

Text-Based Task Text-based questions were derived from PDF documents converted semi-automatically into text. During this process, tables were preserved and transformed into tabulated textual format, resulting in contexts combining plain text and structured tables. This subset primarily focuses on extracting key financial information, such as applicable taxes and minimum investment durations.

To assess the suitability of LLMs for question generation in this task, we first conducted a preliminary analysis to determine whether salient and informative content could be reliably extracted from the source documents. As the result were satisfactory, we proceeded with the creation of question-answers pairs via prompting. The model was asked to identify twenty key informational items that could form the basis for potential questions, link each item to its textual extract, and generate an open-ended question grounded in that excerpt. All outputs were subsequently validated and, when necessary, rewritten by a human annotator.

Image-Based Tasks: Tables and Charts For table- and chart-centered tasks, questions and answers were generated directly from visual inputs. This subset includes open-ended, Yes/No, and True/False questions. The objective is to evaluate structured and graphical data interpretation in financial documents.

Image-Based Task: Conversational Setting The conversation-based subset (referred to as *Conv.* in the tables) targets multi-step reasoning over financial content. Unlike the rest of the dataset, where answers are directly extractable, these questions involve mathematical reasoning and are formatted as multiple-choice questions. This subset was generated using a dedicated prompt specifying both the structure of the conversational turns and the syntactic diversity, as well as the nature of the references to be included in the dialogue. This design enables controlled evaluation of error propagation in interactive settings.

Question Type	Example
Text Question	*« Ce fonds est-il géré activement ou suit-il un indice de manière passive ? »*
Table Comprehension	*« À combien s'élèvent les frais courants annuels prélevés par le FCPE ? »*
Chart Interpretation	*« Combien de périodes consécutives sans cristallisation sont visibles sur le graphique ? »*
Special Cases	*« Quels instruments dérivés spécifiques peuvent être utilisés par le compartiment ? »*
Conversational	1st turn: *« Si je place 25 000 € sur la part A, combien me coûteraient les frais d'entrée maximum ? »* 2nd turn: *« Et si je prends cette même somme pour la part I ou R, j'aurais une différence au niveau des frais ? »*

Table 3: Examples of each question type in the SCRIBE FINANCE dataset. Visual examples are provided in Appendix 12.1 and English translations in Appendix 5.

3.2. Manual Validation and Refinement

All question–answer pairs were validated by a French financial domain expert.[4] Each question was assigned a gold-standard answer confirmed by the expert.

The validation process covered question formulation, answer correctness, and manual verification of all document excerpts used as inputs, with particular attention to open-ended questions. Instances were removed if answers were incorrect, questions were overly generic, insufficiently grounded in the source table, too short, or repetitive. To increase linguistic and structural diversity, a substantial subset of the remaining questions was rewritten by the same annotator. In total, 75% of questions were reformulated or removed by the annotator.

3.3. Task Overview and Dataset Composition

The benchmark comprises six task categories reflecting realistic financial document understanding scenarios. Except for text-only questions, all tasks involve multimodal inputs, where a relevant image (e.g., a table, chart, or document page) is provided alongside the textual context.

The task categories (examples in Table 3) are:

- **Text Question**, focusing on extraction from purely textual contexts;
- **Table Comprehension**, requiring reasoning over structured tabular data;
- **Chart Interpretation**, based on graphical financial representations;
- **Special Cases**, involving nuanced terminology or implicit reasoning;
- **Conversational (Gold Context)**, with a dialogue with oracle previous answers;
- **Conversational (Model Context)**, with a dialogue with model-generated previous answers to study error propagation.

To capture a range of retrieval and reasoning challenges, tasks vary along two axes: **context length**, ranging from short excerpts to document-level inputs, and **context modality**, including plain text, tables, charts, and mixed formats. Questions are formulated as open-ended, binary (Yes/No, True/False), or multiple-choice depending on the task.

3.4. Dataset Statistics

Table 2 summarizes the distribution of questions across task categories, question types, and context modalities. The dataset includes both text-based and image-based questions, with open-ended formats dominating overall, while binary and conversational formats target more constrained reasoning settings.

Context lengths range from short passages (1–2 sentences) to multi-page documents. The conversational subset consists of 5–10 turn dialogues, explicitly designed to probe error propagation and robustness in interactive scenarios.

4. Experimental Setup

Models We evaluated six state-of-the-art Vision-Language Models spanning different scales and architectures: Qwen/Qwen3-VL-8B-Instruct and Qwen/Qwen3-VL-32B-Instruct (Qwen Team, 2025), google/gemma-3-12b-it and google/gemma-3-27b-it (Gemma Team et al., 2025), and mistralai/Pixtral-12B-2409 and mistralai/Pixtral-Large-Instruct-2411 (Agrawal et al., 2024). Model sizes range from 8B to 124B parameters, enabling analysis of scaling effects on financial document understanding.

Answer Generation For each task, models received a prompt and were instructed to respond concisely without explanations. Single turn image-based tasks (Table, Charts, Special Cases) included the image followed by the question. The Text Question task provided textual context instead

[4]The expert annotator has two years of experience as a financial data scientist leading a financial data annotation team. As each generated question had a single directly verifiable answer, one expert was deemed sufficient for dataset verification and rewriting.

of an image. Conversational tasks built a multi-turn dialogue incrementally, with the image provided only at the first turn and subsequent model responses appended to the history. In all cases, the assistant turn was prefilled with "Answer:" to constrain the response format. We used greedy decoding to ensure reproducibility. Complete prompt templates are provided in Appendix 12.2.

Evaluation Protocol Given that more than a half of the proposed dataset consists of open-ended questions (Table 2), and to ensure a unified evaluation process across all tasks, we adopt an LLM-as-judge approach[5] to avoid the high cost of human validation (Zheng et al., 2023). We used three open-source judge models[6] independently to assess each response: meta-llama/Llama-3.3-70B-Instruct (Grattafiori et al., 2024), Qwen/Qwen3-32B (Qwen Team, 2025), and google/gemma-3-27b-it (Gemma Team et al., 2025)[7]. An answer was considered correct if a majority of judges determined it to be correct. Scores reported in Table 4 represent the percentage of questions answered correctly under this majority-vote criterion.

5. Results and Analysis

Table 4 presents results across all task categories. Qwen3-VL-32B achieved the strongest overall performance with an average score of 75.6%, obtaining top scores across all six categories. Qwen3-VL-8B followed with 67.8%, Gemma-3-27B reached 66.2%, Gemma-3-12B scored 63.8%, Pixtral-Large-124B achieved 55.2%, and Pixtral-12B showed the lowest performance at 53.4%.

For most tasks, models achieved strong performance in the 70-90% range. Text Question scores approached 90% across all models, and table comprehension reached 85.8% for the best performers, indicating that current Vision-Language Models handle both textual entity extraction and structured visual information effectively when the task is well-defined.

Figure 3 shows that text-based accuracy remains high across short and medium contexts, with only moderate degradation as context length increases.

Two tasks exposed notable limitations. First, chart interpretation is challenging for all tested models, with scores ranging from 34.4% (Pixtral-12B) to 61.7% (Qwen3-VL-32B). The second best-performing model does not reach 50%. Though charts are designed to render complex information more accessible to human understanding, this visual simplification paradoxically challenges our tested models, which struggle to extract trends, comparisons, and proportions from graphical elements rather than explicit text or tabular image.

Figure 2 provides a fine-grained breakdown of image-based performance, showing strong results on table comprehension but a substantial drop on chart-based questions across all models.

Second, the conversational task evaluation showed how error propagation affects multi-turn reasoning. In the gold context condition, where correct previous answers are provided, models achieved 63.1–86.2% accuracy with clear differentiation by model capacity. In the standard condition, where models must build on their own previous responses, performance dropped sharply and converged to a narrow 46.2–58.5% range. The comparison between the Conversational Gold and Conversational Standard task suggests that the bottleneck is not reasoning capacity per se, but rather the accumulation of errors across turns: once a model makes an early mistake, subsequent answers are compromised by incorrect context, and larger models offer no protection against this cascade. These findings question the reliability of VLMs in interactive, multi-turn financial analysis scenarios where accumulated errors cannot be corrected.

6. Discussion

This work evaluates the capabilities of current Vision-Language Models on French financial document understanding through the SCRIBE FINANCE benchmark. Beyond reporting performance scores, our results reveal several structural limitations that are particularly relevant for high-stakes, real-world deployment.

6.1. Model Performance and Limitations

First, the strong performance observed on text-based and table-based tasks suggests that contemporary VLMs are generally reliable when the task is well-scoped and the relevant information is explicitly present in the input. Extraction of named entities, numerical values, and clearly localized facts appears largely solved under these conditions. This aligns with prior findings on document understanding benchmarks (Clark et al., 2026) and

[5]We initially attempted to extract answers automatically using regular expressions for certain question types, however this approach proved error-prone, so we switched to the LLM-as-judge method.

[6]Only open-source models were used, in compliance with the restrictions established by our institution.

[7]Although Qwen and Gemma family models are used both for answer generation and evaluation, which may raise concerns about self-preference bias, Chen et al. (2025) show that models larger than 7B parameters exhibit limited self-bias and that the strongest self-preference effects are observed in the Llama family, which in our setup is used only for the evaluation.

Model	Task						Avg.
	Text	Tables	Charts	Conv. Gold	Conv.	Special Cases	
Qwen3-VL-8B	89.4	80.0	45.3	73.8	52.3	63.6	67.8
Gemma-3-12B	88.0	**85.8**	46.1	70.8	50.8	40.9	63.8
Pixtral-12B	88.4	51.7	34.4	63.1	52.3	27.3	53.4
Gemma-3-27B	89.0	85.0	48.4	73.8	49.2	54.5	66.2
Qwen3-VL-32B	**89.8**	**85.8**	**61.7**	**86.2**	**58.5**	**72.7**	**75.6**
Pixtral-Large-124B	87.8	71.7	46.1	70.8	46.2	63.6	55.2

Table 4: Model performance on SCRIBE FINANCE (accuracy %). Text and table tasks achieve strong results (80–90%), while chart interpretation (34–62%) and multi-turn conversation (46–59%) display significant weaknesses. **Bold** indicates best performance; underline indicates second best.

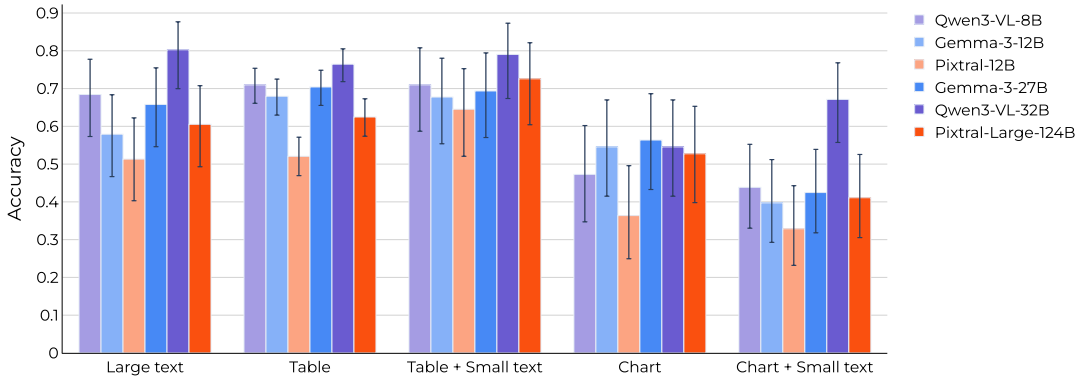


Figure 2: Model accuracy on image-based question subcategories. Performance remains strong on table comprehension tasks (70–86%) but degrades substantially on chart interpretation (34–62%). Qwen3-VL-32B consistently outperforms other models across all visual modalities.

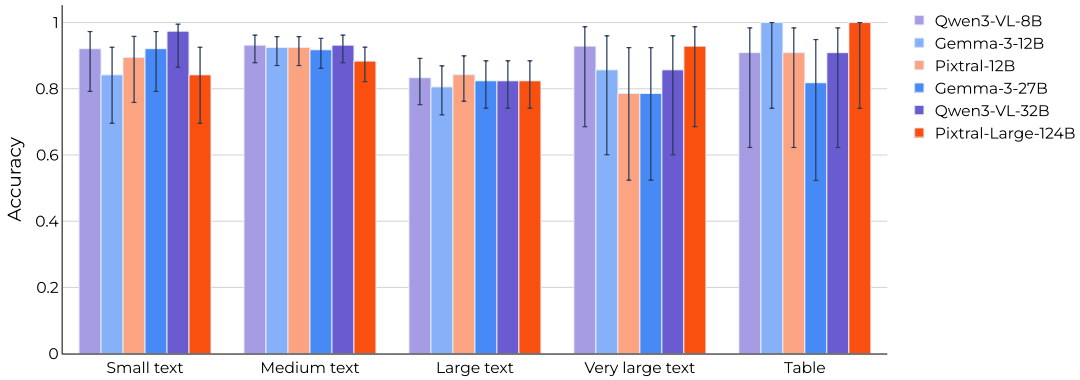


Figure 3: Model accuracy on text-based question subcategories by context length. All models achieve high performance (85–95%) on short and medium text contexts, with moderate degradation on larger contexts. Performance on tabular text (rightmost) remains competitive, indicating that text-based table comprehension is less challenging than image-based table interpretation.

indicates that scaling and multimodal pretraining have effectively addressed many single-step retrieval problems.

However, this apparent robustness does not extend to more visually or temporally complex settings. Chart interpretation remains a consistent weakness across all evaluated models, with large performance gaps relative to text and table tasks. Unlike tables, charts require models to infer trends, relative comparisons, and implicit values that are not directly encoded as text. The persistent difficulty observed here suggests that current VLMs rely heavily on surface-level pattern matching rather than deeper visual abstraction, limiting their ability to reason over graphical representations commonly used in financial reporting.

The most striking finding concerns multi-turn conversational evaluation. When models are required to build on their own previous answers, performance collapses to approximately 50% regard-

less of model size. This behavior exposes a failure mode that is not visible in single-turn benchmarks: early mistakes introduce incorrect context that subsequent reasoning cannot recover from. Importantly, the comparison with the Conversational Gold setting indicates that this degradation is not primarily due to a lack of reasoning capacity, but rather to error accumulation and context contamination. Scaling the model does not mitigate this effect, suggesting that architectural or training-level changes may be required to support reliable multi-step financial reasoning.

6.2. Generation Biases and Implications for Dataset Construction

Our analysis also reveals several systematic tendencies in the semi-automatic question generation process that have implications for both dataset composition and evaluation outcomes. In the open-ended setting, generated questions disproportionately focused on short, easily identifiable facts, such as entry fee percentages or single numerical values. In contrast, more complex information—particularly investment rules or constraints distributed across multiple sections of a document—was less frequently captured. This suggests that current generation pipelines favor information that is locally salient, which may underrepresent questions requiring broader contextual integration.

We further observed limited lexical diversity and originality in a subset of the generated questions. Similar formulations were often reused across documents, resulting in questions that were syntactically correct but insufficiently specific to the source material. A comparable pattern emerged for table-based inputs: even when tables contained structurally rich or nuanced information, generated questions tended to target straightforward value extraction rather than higher-level relationships or constraints. These tendencies required manual revision (described in Section 3) to ensure adequate coverage of more challenging reasoning scenarios.

In the multiple-choice setting, additional artifacts emerged. Although the model was able to generate candidate distractors, incorrect answer options were frequently implausible, often falling well outside the range of values or concepts presented in the document. Moreover, the correct answer was repeatedly assigned to the same option label, introducing a positional bias that could be exploited during evaluation. Addressing these issues required manual correction of both distractor content and label assignment. Together, these observations underscore current limitations of automated generation methods and reinforce the importance of human oversight when constructing evaluation benchmarks in high-stakes domains such as finance.

These observations have practical implications. Financial analysis often involves iterative questioning, clarification, and dependency on prior answers. The inability of current models to correct or contain earlier errors raises concerns about their suitability for interactive advisory or compliance-related applications, where even small inaccuracies can propagate into significant downstream risks. Our results therefore caution against over-reliance on conversational interfaces for complex financial document analysis without additional safeguards.

Finally, the use of an LLM-as-judge evaluation protocol reflects a trade-off between scalability and human validation. While this approach enables consistent and reproducible assessment across a large benchmark, it may inherit biases or blind spots from the judge models themselves. Although majority voting across multiple judges mitigates some of these concerns, future work should further investigate alignment between automated judgments and expert human evaluation, particularly for nuanced or ambiguous financial questions.

Overall, our dataset exposes a clear gap between strong single-step extraction performance and fragile multi-step reasoning in financial contexts. Addressing this gap will likely require advances beyond model scaling, including improved training objectives, explicit uncertainty modeling, and mechanisms for error detection and correction in multi-turn interactions.

7. Conclusion

We introduced SCRIBE FINANCE, a multimodal benchmark for evaluating Vision-Language Models on French financial document understanding. The benchmark targets realistic, high-stakes scenarios involving long, heterogeneous documents that combine legal text, numerical tables, and charts, and includes both single-turn and multi-turn conversational tasks. By focusing on excerpt-grounded information extraction rather than full-document access, SCRIBE FINANCE emphasizes precise retrieval as a prerequisite for reliable financial reasoning.

Our evaluation of six state-of-the-art VLMs presents a clear contrast between strong performance on well-scoped text and table extraction tasks and persistent weaknesses in chart interpretation and conversational settings. In particular, we show that error propagation across dialogue turns causes model performance to collapse regardless of scale, exposing a failure mode that is largely invisible in standard single-turn benchmarks.

Together, these findings suggest that progress in financial document understanding will require advances beyond model scaling alone. Future

work should explore training objectives and architectural mechanisms that explicitly support uncertainty awareness, error correction, and robust multi-step reasoning over multimodal inputs. We hope that SCRIBE FINANCE will serve as a useful testbed for measuring such progress and for guiding the development of more reliable models for real-world financial analysis.

8. Limitations

Our benchmark focuses exclusively on French-language investment documents, which limits the direct generalizability of our findings to other languages or regulatory settings. While this choice addresses a clear gap, financial disclosure practices may differ across jurisdictions. The benchmark primarily evaluates information extraction and reasoning grounded in explicit document content. It does not cover more speculative or advisory use cases, such as portfolio recommendation or forward-looking decision-making, and should therefore be viewed as assessing foundational document understanding rather than full financial expertise. Dataset construction relies in part on semi-automatic question generation using large language models, followed by expert revision. We observed occasional references to information outside the provided input context, as well as limited originality and lexical diversity in generated questions, suggesting potential memorization effects and a bias toward easily extractable facts. These observations stress the continued necessity of human expert validation to ensure proper grounding and question quality. Finally, our evaluation relies on an LLM-as-judge protocol rather than exhaustive human annotation. While majority voting across multiple judges improves robustness, subtle numerical or legal errors may still be missed. In addition, our conversational evaluation highlights indeed error propagation but does not explicitly model uncertainty awareness or error correction.

9. Ethics

All documents used in this study are publicly available financial disclosures released by asset management companies. No private, sensitive, or personally identifiable information was collected or processed. The expert reviewer involved in dataset construction and evaluation was fairly compensated for their contributions. While the benchmark is designed for document understanding and evaluation purposes, financial applications are inherently high-stakes. Our results identify failure modes such as error propagation in conversational settings, exposing the risk of over-reliance on automated systems for financial analysis or advisory tasks. The benchmark is intended to support research and evaluation, not to replace professional judgment in real-world financial decision-making.

10. Acknowledgments

We thank the reviewers for their valuable feedback on our work. We are grateful for all the comments and feedback from Iacopo Poli, Oskar Hallström, and Adrien Cavaillès from LightOn on an earlier version of the SCRIBE FINANCE dataset.

This work was performed using HPC resources from GENCI–IDRIS (Grant 2025-AD011016564) on the supercomputer Jean Zay's CSL, A100, and H100 partitions. This project was supported by the BPI Code Common and Scribe projects as well as by Djamé Seddah's PRAIRIE-PSAI chair, funded by the French national agency ANR, as part of the "France 2030" strategy under the reference ANR-23-IACL-0008.

11. Bibliographical References

References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, et al. 2024. Pixtral 12b.

Zhi-Yuan Chen, Hao Wang, Xinyu Zhang, Enrui Hu, and Yankai Lin. 2025. Beyond the surface: Measuring self-preference in LLM judgments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1653–1672, Suzhou, China. Association for Computational Linguistics.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2022a. FinQA: A Dataset of Numerical Reasoning over Financial Data. ArXiv:2109.00122 [cs].

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022b. ConvFinQA: Exploring the Chain of Numerical Reasoning in Conversational Finance Question Answering. ArXiv:2210.03849 [cs].

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, et al. 2026. Molmo2: Open weights and data for vision-language models with video understanding and grounding. *arXiv preprint arXiv:2601.10611*.

Yang Deng, Wenqiang Lei, Wenxuan Zhang, Wai Lam, and Tat-Seng Chua. 2023. PACIFIC: Towards Proactive Conversational Question Answering over Tabular and Textual Data in Finance. ArXiv:2210.08817 [cs].

Martin d'Hoffschmidt, Wacim Belblidia, Quentin Heinrich, Tom Brendlé, and Maxime Vidal. 2020. FQuAD: French question answering dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1193–1208, Online. Association for Computational Linguistics.

European Commission. 2014. Directive 2014/65/eu of the european parliament and of the council of 15 may 2014 on markets in financial instruments (mifid ii). Official Journal of the European Union.

European Union. 2014. Regulation (eu) no 1286/2014 of the european parliament and of the council of 26 november 2014 on key information documents for packaged retail and insurance-based investment products (priips). Official Journal of the European Union.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2025. Colpali: Efficient document retrieval with vision language models.

G20. 2009. Leaders' statement: The global plan for recovery and reform. G20 London Summit, April 2009.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, et al. 2025. Gemma 3 technical report.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat,

Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin

Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Rachel Keraron, Guillaume Lancrenon, Mathilde Bras, Frédéric Allary, Gilles Moyse, Thomas Scialom, Edmundo-Pavel Soriano-Morales, and Jacopo Staiano. 2020. Project PIAF: Building a Native French Question-Answering Dataset. ArXiv:2007.00968 [cs].

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. FrenchMedMCQA: A French Multiple-Choice Question Answering Dataset for Medical domain. ArXiv:2304.04280 [cs].

Oscar Lithgow-Serrano, David Kletz, Vani Kanjirangat, David Adametz, Marzio Lunghi, Claudio Bonesana, Matilde Tristany-Farinha, Yuntao Li, Detlef Repplinger, Marco Pierbattista, Stefania Stan, and Oleg Szehr. 2025. Assessing RAG system capabilities on financial documents. In *Proceedings of The 10th Workshop on Financial Technology and Natural Language Processing*, pages 124–147, Suzhou, China. Association for Computational Linguistics.

Jiahua Liu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2019. XQA: A Cross-lingual Open-domain Question Answering Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2358–2368, Florence, Italy. Association for Computational Linguistics.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A Linguistically Diverse Benchmark for Multilingual Open Domain Question Answering. ArXiv:2007.15207 [cs].

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. ArXiv:2010.00133 [cs].

Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. 2022. French CrowS-Pairs: Extension à une langue autre que l'anglais d'un corpus de mesure des biais sociétaux dans les modèles de langue masqués. *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles*.

Qwen Team. 2025. Qwen3 technical report.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Siqiao Xue, Xiaojing Li, Fan Zhou, Qingyang Dai, Zhixuan Chu, and Hongyuan Mei. 2025. Famma: A benchmark for financial domain multilingual multimodal question answering.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. **MIRACL** : A Multilingual Retrieval Dataset Covering 18 Diverse Languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A Question Answering Benchmark on a Hybrid of Tabular and Textual Content in Finance. ArXiv:2105.07624 [cs].

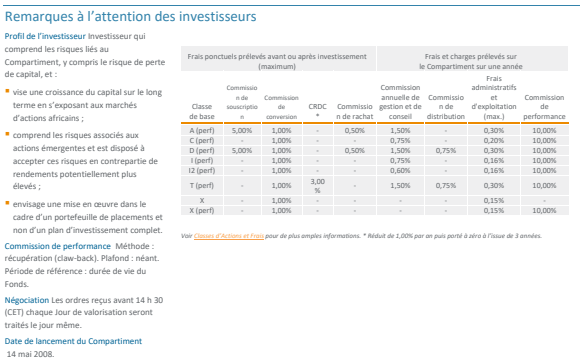
12. Appendix

12.1. Dataset Examples

This section provides representative examples from the SCRIBE FINANCE benchmark across different task categories. For convenience, all examples have been translated into English, the original French prompts are available in the paper's accompanying repository.

12.1.1. Table Comprehension Example

Figure 5 presents a typical table comprehension task from the dataset.



Remarques à l'attention des investisseurs

Profil de l'investisseur Investisseur qui comprend les risques liés au Compartiment, y compris le risque de perte de capital, et :

- vise une croissance du capital sur le long terme en s'exposant aux marchés d'actions africains ;
- comprend les risques associés aux actions émergentes et est disposé à accepter ces risques en contrepartie de rendements potentiellement plus élevés ;
- envisage une mise en œuvre dans le cadre d'un portefeuille de placements et non d'un plan d'investissement complet.

Commission de performance Méthode : récupération (claw-back). Plafond : néant. Période de référence : durée de vie du Fonds.

Négociation Les ordres reçus avant 14 h 30 (CET) chaque Jour de valorisation seront traités le jour même.

Date de lancement du Compartiment 14 mai 2008.

	Frais ponctuels prélevés avant ou après investissement (maximum)				Frais et charges prélevés sur le Compartiment sur une année			
Classe de base	Commission de souscription	Commission de conversion	CRDC *	Commission de rachat	Commission annuelle de gestion et de conseil	Commission de distribution	Frais administratifs et d'exploitation (max.)	Commission de performance
A (perf)	5,00%	1,00%	-	0,50%	1,50%	-	0,30%	10,00%
C (perf)	-	1,00%	-	-	0,75%	-	0,20%	10,00%
D (perf)	5,00%	1,00%	-	0,50%	1,50%	0,75%	0,30%	10,00%
I (perf)	-	1,00%	-	-	0,75%	-	0,16%	10,00%
I2 (perf)	-	1,00%	-	-	0,60%	-	0,16%	10,00%
T (perf)	-	1,00%	3,00 %	-	1,50%	0,75%	0,30%	10,00%
X	-	1,00%	-	-	-	-	0,15%	-
X (perf)	-	1,00%	-	-	-	-	0,15%	10,00%

*Voir Classes d'Actions et Frais pour de plus amples informations. * Réduit de 1,00% par an puis porté à zéro à l'issue de 3 années.*

Figure 4: Example table from a financial document.

Question: Which class applies a redemption fee?

Answer: A (perf), D (perf)

Figure 5: Table comprehension example based on Figure 4.

12.1.2. Chart Interpretation Example

Figure 7 illustrates a chart interpretation task.

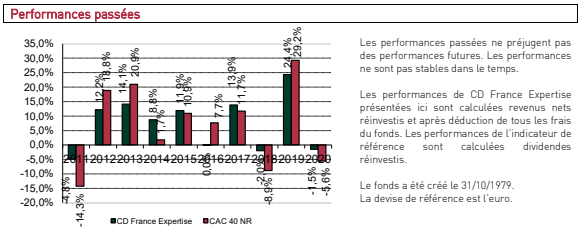


Figure 6: Example financial chart requiring visual interpretation.

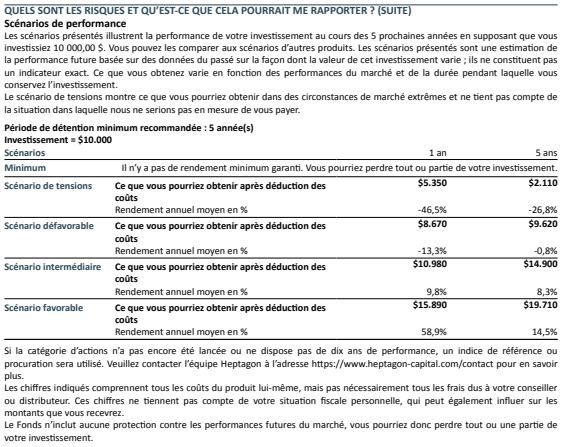
12.1.3. Conversational Task Example

Figure 8 shows the financial projection table used as context for the multi-turn dialogue presented in Figure 9. This example demonstrates how questions require maintaining context across turns and performing numerical reasoning based on tabular financial data.

Question: What was the performance of the CD France Expertise fund in 2018?

Answer: -2.0%

Figure 7: Chart interpretation example based on Figure 6.



QUELS SONT LES RISQUES ET QU'EST-CE QUE CELA POURRAIT ME RAPPORTER ? (SUITE)

Scénarios de performance

Les scénarios présentés illustrent la performance de votre investissement au cours des 5 prochaines années en supposant que vous investissiez 10 000,00 $. Vous pouvez les comparer aux scénarios d'autres produits. Les scénarios présentés sont une estimation de la performance future basée sur des données du passé sur la façon dont la valeur de cet investissement varie ; ils ne constituent pas un indicateur exact. Ce que vous obtenez varie en fonction des performances du marché et de la durée pendant laquelle vous conservez l'investissement.

Le scénario de tensions montre ce que vous pourriez obtenir dans des circonstances de marché extrêmes et ne tient pas compte de la situation dans laquelle nous ne serions pas en mesure de vous payer.

Période de détention minimum recommandée : 5 année(s)

Investissement = $10.000

Scénarios		1 an	5 ans
Minimum	Il n'y a pas de rendement minimum garanti. Vous pourriez perdre tout ou partie de votre investissement.		
Scénario de tensions	**Ce que vous pourriez obtenir après déduction des coûts**	**$5.350**	**$2.110**
	Rendement annuel moyen en %	-46,5%	-26,8%
Scénario défavorable	**Ce que vous pourriez obtenir après déduction des coûts**	**$8.670**	**$9.620**
	Rendement annuel moyen en %	-13,3%	-0,8%
Scénario intermédiaire	**Ce que vous pourriez obtenir après déduction des coûts**	**$10.980**	**$14.900**
	Rendement annuel moyen en %	9,8%	8,3%
Scénario favorable	**Ce que vous pourriez obtenir après déduction des coûts**	**$15.890**	**$19.710**
	Rendement annuel moyen en %	58,9%	14,5%

Si la catégorie d'actions n'a pas encore été lancée ou ne dispose pas de dix ans de performance, un indice de référence ou procuration sera utilisé. Veuillez contacter l'équipe Heptagon à l'adresse https://www.heptagon-capital.com/contact pour en savoir plus.

Les chiffres indiqués comprennent tous les coûts du produit lui-même, mais pas nécessairement tous les frais dus à votre conseiller ou distributeur. Ces chiffres ne tiennent pas compte de votre situation fiscale personnelle, qui peut également influer sur les montants que vous recevrez.

Le Fonds n'inclut aucune protection contre les performances futures du marché, vous pourriez donc perdre tout ou une partie de votre investissement.

Figure 8: Financial projection table showing investment scenarios over different time periods. This image serves as the visual context for the conversational dialogue in Figure 9.

12.1.4. Translation of Questions from Table 3

Table 5 presents the English translations of the example questions shown in Table 3.

12.2. Prompt Templates

We used three prompt templates to obtain answers from the LLM during evaluation, depending on the task type. In all cases, models completed an assistant turn prefilled with "Answer:". The original prompts were written in French, here we provide their English versions for convenience.

12.2.1. Image-Based Tasks

Figure 10 presents the template used for Table, Table Yes/No and True/False, Charts, and Special Cases tasks. The model receives an image followed by the question and must complete the assistant turn.

12.2.2. Text-Based Task

Figure 11 presents the template used for the Text Question task, which operates on textual context

Question Type	Example
Text Question	*« Is this fund actively managed, or does it passively follow an index? »*
Table Comprehension	*« What are the annual ongoing charges applied by the FCPE? »*
Chart Interpretation	*« How many consecutive periods without crystallization are visible in this chart? »*
Special Cases	*« Which specific derivative instruments may be used by the sub-fund? »*
Conversational	1st turn: *« If I invest €25,000 in share class A, what would be the maximum entry fees? »* 2nd turn: *« And if I invest the same amount in share class I or R, would there be any difference in the fees? »*

Table 5: English translations of the example questions from the SCRIBE FINANCE dataset, presented in Table 3.

rather than images.

12.2.3. Conversational Tasks

Figure 12 illustrates the template used for conversational tasks. The dialogue is built incrementally over 5–10 turns: the image is provided only in the first turn, and each subsequent question is appended as a new user message. In the *Conv.* setting, the model's own previous answers are included in the conversation history. In the *Conv. Gold* setting, ground-truth answers replace model completions.

12.2.4. LLM-as-judge Evaluation Template

Figure 13 presents the prompt template used for LLM-as-judge evaluation. Each judge receives the question, reference answer, and prediction, then outputs "Correct" or "Incorrect".

Context: Financial projection table (Figure 8)

Q1: If I invest $25,000 and the favorable scenario occurs after one year, approximately how much will I get at the end of the year?

A. $29,500
B. $34,725
C. $39,725
D. $40,000

Answer: C

Q2: And out of curiosity, by how much does that value exceed the intermediate scenario over the same period?

A. Approximately $7,625
B. A little over $5,000
C. $12,275
D. They are equivalent

Answer: C

Q3: OK, but if I look ahead 5 years with this same scenario, what final amount do I reach?

A. $32,000
B. $37,250
C. $39,750
D. $43,750

Answer: B

Q4: Oh right, and with the stress scenario over 1 year with the amount I invested... roughly how much do I lose?

A. $9,250
B. $12,750
C. $11,625
D. $19,650

Answer: C

Figure 9: Example of a multi-turn conversational question sequence requiring numerical reasoning across different investment scenarios. Each question builds on previous context, testing the model's ability to maintain coherence and perform calculations based on the tabular financial data shown in Figure 8.

User:

`<image>`

Question: {question}
Answer the question concisely based on the image provided. Don't include any explanations.

Assistant:

Answer: *[model completion]*

Figure 10: Prompt template for Image-based tasks.

User:

Context: {context}

Question: {question}
Answer the question concisely based on the context provided. Don't include any explanations.

Assistant:

Answer: *[model completion]*

Figure 11: Prompt template for Text-based task.

Turn 1 — User:

`<image>`

Answer all questions concisely based on the image provided. Don't include any explanations.
{question_1}

Turn 1 — Assistant:

Answer: *[model completion]*

Turn 2 — User:

{question_2}

Turn 2 — Assistant:

Answer: *[model completion]*

... continued for all turns ...

Figure 12: Prompt template for conversational tasks. The image is provided once at the first turn; subsequent turns contain only the question. In the *Conv. Gold* setting, ground-truth answers replace model completions in the history.

User:

You will receive a question, a reference answer, and an answer to evaluate. Your task is to determine whether the prediction is correct or incorrect.

Evaluation rules:
1. A prediction is correct if it accurately answers the question based on the reference answer.
2. If the answer involves a numerical or financial value, consider as equivalent any expressions representing the same value (e.g., 20% = 0.2; 1,000,000 = 1 million; 2,200,000 = 2.2M; 12.3 = 12.3).
3. If the question is multiple-choice, an answer is correct if it matches exactly one of the correct options (by letter, number, or text).
4. If the prediction is an exact paraphrase of the reference, it is correct.
5. Do not take into account phrasing or style, only factual or numerical accuracy.
6. Respond only with "Correct" or "Incorrect".

Examples:
Question: {example question}
Reference: {example reference}
Answer to evaluate: {example prediction} Expected evaluation: {example expected evaluation}
[5 examples total of numerical and MCQ cases with correct and incorrect answer]

Answer to evaluate:
Question: {question}
Reference answer: {reference}
Answer to evaluate: {prediction}

Assistant:
Evaluation: *[model completion]*

Figure 13: LLM-as-judge evaluation template. Five examples covering numerical equivalence and multiple-choice formats are included in the full prompt.