Can AI Read Like a Financial Analyst? A Financial Touchstone for Frontier Language Models Such as Gemini 2.5 Pro, o3, and Grok 4 on Long-Context Annual Report Comprehension

Jan Spörer
Institute for Computer Science
University of St. Gallen
St. Gallen, St. Gallen, Switzerland
jan.spoerer@unisg.ch

Abstract

Comprehending long, dense annual reports is a critical task for financial analysts that is ripe for AI automation, yet model reliability remains a key concern. To address this, we introduce Financial Touchstone—a new, large-scale benchmark with 2,878 question-context-answer triplets across 480 international annual reports, guaranteed to be unseen by the models we evaluate. We test eleven frontier language models from leading labs, including reasoning-capable models like Google's Gemini 2.5 Pro, Anthropic's Claude Opus, OpenAI's o3, and xAI's Grok 4. Our analysis reveals that while reasoning models achieve high accuracy—with Gemini 2.5 Pro reaching 91.6% and hallucination rates as low as 3.2%—the primary bottleneck is not the models' comprehension but the initial information retrieval step. Model accuracy plummets to 0.2% when the provided context is insufficient. This work demonstrates that future progress in automated financial analysis hinges more on solving the challenge of targeted information retrieval in complex documents than on incremental improvements in model reasoning alone.

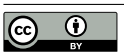
CCS Concepts

• **Computing methodologies** → **Natural language processing**; *Information extraction*; • **Applied computing** → **Economics**.

Keywords

information extraction, question answering, datasets, benchmarks, LLM evaluation, natural language processing, neural networks, language models, financial text, annual reports, financial reporting, corporate disclosure, needle in the haystack, context window, reasoning models, RAG

ACM Reference Format:
Jan Spörer. 2025. Can AI Read Like a Financial Analyst? A Financial Touchstone for Frontier Language Models Such as Gemini 2.5 Pro, o3, and Grok 4 on Long-Context Annual Report Comprehension. In *6th ACM International Conference on AI in Finance (ICAIF '25), November 15–18, 2025, Singapore, Singapore.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3768292.3770417



ICAIF '25, Singapore, Singapore

ACM ISBN 979-8-4007-2220-2/25/11
https://doi.org/10.1145/3768292.3770417

1 Introduction

1.1 Training Set Contamination

Strong and new evaluation datasets are hard to come by in a time when new language models are released frequently, with allegations of test set contamination being a major concern in today's model benchmarks.

With the Financial Touchstone evaluation dataset, we provide a fully unseen dataset of a real-world task. We evaluate most major frontier models on the dataset, contributing to a fair and unbiased evaluation of model performance.

1.2 Financial Datasets

Annual reports are an important part of financial research, as quick market reactions to these reports show. Markets have higher volatility on earnings release days than on other days [3, 14]. This measurable market reaction demonstrates how these reports constitute a primary source for investor decisions. Investors read these reports directly, or read secondary equity research written by financial analysts [2].

Annual reports sometimes exceed 700 pages (see also figures 1 and 2) of dense financial data, regulatory disclosures, and strategic narratives, and contain information that drives trillion-dollar investment decisions. Financial analysts use these reports to set stock price targets [5, 11] and to give investment recommendations [4, 27].

Consider the complexity: a single annual report may contain hundreds of financial metrics scattered across multiple sections, temporal relationships spanning years of performance data, and subtle narrative cues about future strategy buried in executive commentary. Human analysts can connect these disparate pieces into coherent investment theses, but replicating this capability in AI systems is challenging.

1.3 Contributions

Our study addresses this challenge by conducting the most comprehensive open evaluation of AI systems' performance on financial document comprehension. Rather than testing theoretical capabilities, we evaluate performance on actual annual reports using the same questions that professional analysts consider most critical for investment decisions [25].

In summary, the paper makes the following novel contributions that will benefit academic researchers and business professionals alike:

- **Data Scale and International Coverage:** We present a comprehensive benchmark dataset for evaluating language model performance on annual report analysis, containing 2,878 question-answer pairs from 480 corporate annual reports focusing on, but not limited to, the years 2021–2023[1], resulting in a previously unseen scale of 29,530 financial question-context-answer triplets for an open annual report dataset. The dataset exceeds the size and international coverage of prior annual report benchmarks, as prior studies largely focused on reports from the United States [13]. The dataset contains over 83 million tokens, almost entirely high-quality, coherent, manually-written (non-synthetic) financial text. This is more than four times larger than the previously largest study on financial text comprehension with LLMs [25].
- **Coverage of All Flagship Models:** We conduct systematic evaluations across eleven reasoning and non-reasoning models accessible via APIs, enabling reproducible research without local infrastructure requirements.
- **Relevant Model Selection:** We choose the best-performing and most popular models [7], making the study highly relevant to most LLM users and researchers.
- **Efficient Retrieval Architecture:** We implement and evaluate RAG architectures for enhanced financial document analysis, demonstrating measurable improvements in entity relationship modeling and cross-document reasoning.
- **Comprehensive Analysis:** We provide comprehensive comparative analysis revealing how different reasoning architectures perform on real-world financial document comprehension tasks, with detailed error analysis and inter-annotator agreement studies.
- **Open Data and Reproducibility:** We release all data, evaluation frameworks, and RAG implementation to support reproducible research in financial AI applications. All data is readily available in both PDF and OCRed text format.

2 Related Work

Financial text analysis has evolved from early work on sentiment analysis [17, 18] to more sophisticated question answering tasks. The field saw a significant step forward with the introduction of benchmarks designed to test multi-step and numerical reasoning. Datasets like FinQA [6] and TAT-QA [30] pushed models to comprehend semi-structured data combining text and tables. However, these benchmarks typically used contexts of only a few thousand tokens, far shorter than complete annual reports, which can exceed 500,000 tokens.

With the ChatGPT breakthrough of LLMs in 2022, the focus shifted to benchmarking their capabilities on financial tasks. It quickly became apparent that large, general-purpose models like GPT-4 [22] could beat domain-specific models like BloombergGPT [28], even on specialized financial tests. FinanceBench [13] was a key benchmark in this era, evaluating models like GPT-4 and Claude 2 on questions derived from annual reports. Its finding that the top-performing model achieved only 19% accuracy underscored the immense difficulty of the task and highlighted the need for more capable models and better evaluation frameworks. Our work builds directly on this by using a larger, more diverse dataset and evaluating the newest generation of reasoning-capable models.

A central challenge in this domain is processing long documents. While modern transformer architectures now support context windows of up to one million tokens [10], the "needle-in-a-haystack" problem persists [21], where models struggle to locate and utilize specific facts buried in long contexts. This necessitates the use of a Retrieval-Augmented Generation (RAG) system [15], which combines efficient retrieval from a large document base with a powerful generator model for answer formulation. Our work explicitly evaluates the performance and, critically, the failure modes of such a system on complete, unmodified annual reports, filling a key gap in the literature by diagnosing the retriever's role as a primary bottleneck.

3 Data

3.1 Annual Report Collection

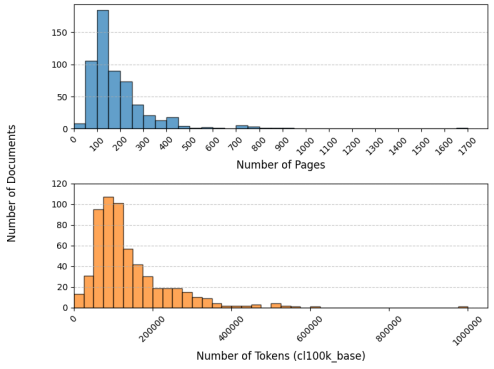
We collected 480 annual reports from publicly traded companies across most global stock markets with relevant market capitalization: Australia, Austria, Belgium, Canada, the Cayman Islands, China, France, Germany, Hong Kong, India, Ireland, Japan, Malaysia, the Netherlands, Portugal, Singapore, Spain, Switzerland, Taiwan, Thailand, the United Kingdom, and the United States. These jurisdictions represent diverse regulatory environments and reporting standards.

Whenever possible, these reports were stratified by company size, country, industry sector, and reporting year (mostly between 2021–2023).

3.2 Question Development

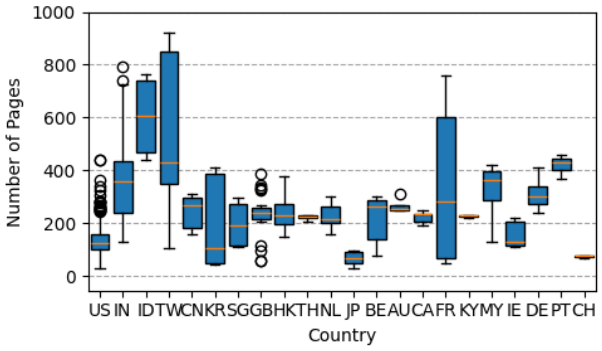
Our questions are based on prior research identifying the most frequently asked questions by professional equity analysts [25]. We selected six question types that cover both quantitative and qualitative information extraction:

Figure 1: Page Count Distribution: Most reports have between 50 and 450 pages. Token Count Distribution: Most reports have less than 300,000 tokens.



[1]We included reports from 2018–2025, with most reports being from 2021–2023. The reports for the year 2025 are by companies with financial years that differ from calendar years. For example, some companies have financial years that last from February 2024 to January 2025, and we included some of these recent reports.

Figure 2: Page Count Distribution by Country: Indian, Indonesian, Taiwanese, and French annual reports were the only countries in the sample that had reports with more than 500 pages. We cut off the chart's y-axis, removing an Indian report with almost 1750 pages from view.



(1) What are the key financial figures and metrics mentioned in this text?
(2) What information is provided about the company's cash flow?
(3) What are the revenue figures mentioned in this text?
(4) What is mentioned about revenue growth or decline?
(5) What business segments or divisions are mentioned and what do they consist of?
(6) What is the legal form of incorporation of the company?

3.3 Current Dataset Statistics

The annual report dataset contains:

- 480 annual reports from publicly traded companies.
- All industry classes from the Global Industry Classification Standard (GICS) are covered.
- 2,878 manually annotated question-context-answer triplets.
- Six questions per report.
- Over 83 million tokens of financial text.
- Temporal span: Focus on 2021–2023 reporting years, with a few reports before and after this time span.
- Covered 20 global stock exchanges, representing most of the world's market capitalization and 22 countries: Shanghai Stock Exchange (CN), Tokyo Stock Exchange (JP), National Stock Exchange (IN), Bombay Stock Exchange (IN), Hong Kong Stock Exchange (HK), Korea Exchange (KR), Taiwan Stock Exchange (TW), London Stock Exchange (GB), Xetra (DE), Toronto Stock Exchange (CA), Australian Securities Exchange (AU), Singapore Exchange (SG, with one company headquartered in the KY), Bursa Malaysia (MY), Thailand Stock Exchange (TH), Indonesia Stock Exchange (ID), NASDAQ (US), NYSE Euronext (US), Euronext (NL, FR, BE, PT, IR), the various BME exchanges (ES), and SIX (CH).

3.4 Inter-Annotator Agreement, Disagreement Triage, and Analysis of Human Error Sources

We performed two types of checks to verify that the manually labeled data is reliable:

(1) For a small subset of the data, we compared two separate annotations to each other and identified discrepancies between them. We triaged the correct solution by reading the report in more depth.
(2) Before we ran the large-scale experiments with all eleven models in full, we ran inference on a subset of the data, and manually reviewed model answers that were marked as wrong.

Our check of final human answers (2) involved reviewing model answers that were marked as wrong, and allowed us to determine if the models or the humans were the true error source. This check confirmed that the approach presented in (1) was effective.

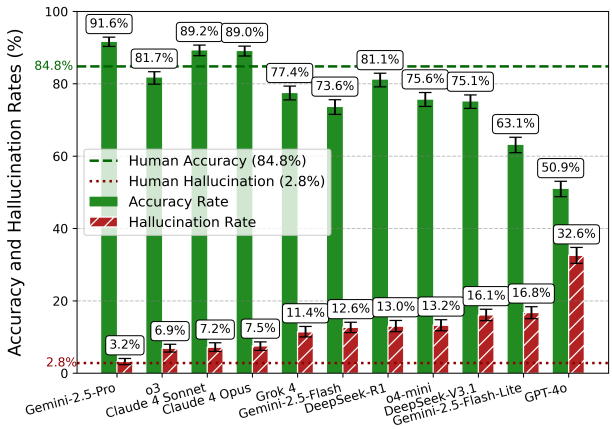
The result of the manual inter-annotator analysis (1) was that 84.8% of human labels are correct, and 2.2% of human labels contain what would be called "hallucinations" in language model terminology, i.e., answers that were not based on the context. We, of course, used the correct answers as the golden source to compare the models' answers to.

We want to point out that real-world human performance may vary from our results, depending on the competence and time commitment of the analyst. The annotators spent multiple weeks on full-time labeling work, and we did not reward them for working fast, giving annotators the time to thoroughly check each of the thousands of answers before marking them as final. In business settings, time constraints are likely tighter, leading to higher error rates among humans.

The main areas of difficulty for human annotators include: exhaustive identification of company segments, which requires considering multiple segmentation dimensions (such as product groups and geographic regions). Also, some annotations contain financial information about subsegments of the company, indicating that the annotator has failed to realize that these financials only relate to a subset of the company. These types of errors can be attributed to a lack of human diligence and attentiveness. Some reports split cash flows and other financials across multiple pages, so that if a human annotator searches for keywords to browse through a PDF, they may miss the context that is written on the prior pages when not paying close enough attention.

Similarly, some annotations failed to mention key financials that were provided on a summary page of key financial performance. While some companies have these pages, others do not have them. The location of these pages is usually early in the annual reports (typically between pages 2–10). Some annotators may have used keyword searches to browse to the correct page, but skipped the summary page while doing so, jumping directly to the three statements (balance sheet, income statement, and cash flow statement). These three statements are only useful as fallbacks or as additions to the information presented on the summary page, as some ratios (such as earnings per share or company-specific metrics such as EBITDA) only appear on the summary page. There is no unified term for these summary pages, with companies using phrases such

Figure 3: Model Performance With 95% Confidence Bars, Excluding Retriever Errors, Ranked by Hallucination Rate: Google Gemini 2.5 Pro has the best recall as well as the lowest hallucination, with OpenAI o3 and the Claude 4 models (Sonnet and Opus) being close followers. Non-reasoning models underperform. The dashed lines show the human baselines.



as "Key Financials," "Our Performance," "Summary of Financial Performance," "xyz at a Glance" as titles to these pages. This heterogeneity sometimes makes it hard for humans to find the summaries in reasonable time, again leading to errors if not enough thought is given to questions about key financials.

4 Methodology

4.1 Task Definition

The benchmark task requires models to extract specific information from annual reports in response to a question. Models must process extensive document contexts and provide accurate, grounded answers without hallucination.

4.2 Overview of Evaluated Closed and Open Models

We employ the following eleven closed-weight and open-weight reasoning models:

- Google Gemini 2.5 Pro (stable) [10]
- Google Gemini 2.5 Flash (stable) [10]
- OpenAI o3 (2025-04-16) [24]
- OpenAI o4-mini (2024-08-06) [24]
- Anthropic Claude Opus 4 [1]
- Anthropic Claude Sonnet 4 [1]
- DeepSeek R1 (0528) [8]
- xAI Grok 4 [29]

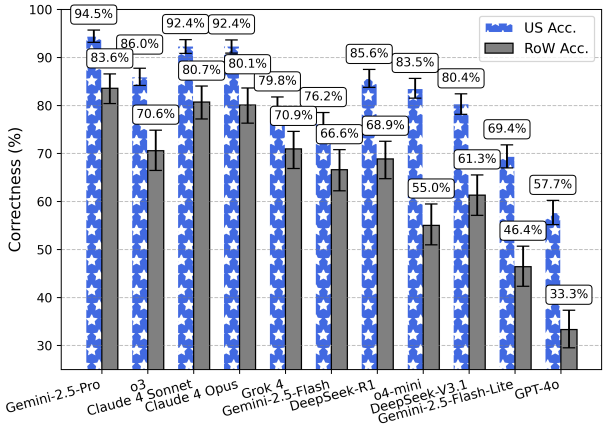
We included the following non-reasoning models:

- Google Gemini 2.5 Flash-Lite Preview (06-17) [10]
- DeepSeek V3.1 [9]
- OpenAI GPT-4o (2024-08-06) [23]

These models were chosen because they include the best-rated language models according to LLM Arena in July 2025 [7]. We did not employ finance-specific models as these are not competitive against large generalized language models anymore.

We opted against including Meta's Llama 3 [16, 26] and Llama 4 [19] models as their performance is not competitive with other labs' open- and closed-weight models.

Figure 4: Model Accuracy With 95% Confidence Bars, USA vs. Rest of World, Ranked by Hallucination Rate (Not Displayed Here): All models work better with American 10-Ks than with other annual reports. y-axis starts at 25%.



4.3 RAG Implementation

We implement a RAG architecture with the following configuration:

The knowledge base contains chunks of 1000 tokens. The chunk overlap is 200 tokens to ensure that no contextual information is lost. We used `cl100k` base encoding for the tokenization. The vector store uses FAISS. OpenAI's `text-embedding-3-small` embeddings construct the embedding space.

We also tried top-10 retrieval, but did not observe a significant increase for the RAG system's performance. We provide all top-5 chunks to the model for the answer generation.

We could further optimize the pipeline by providing the top chunks one-by-one to the model, and by retrieving more than five chunks, but we decided against this path to make the already extensive experiments more computationally permissive.

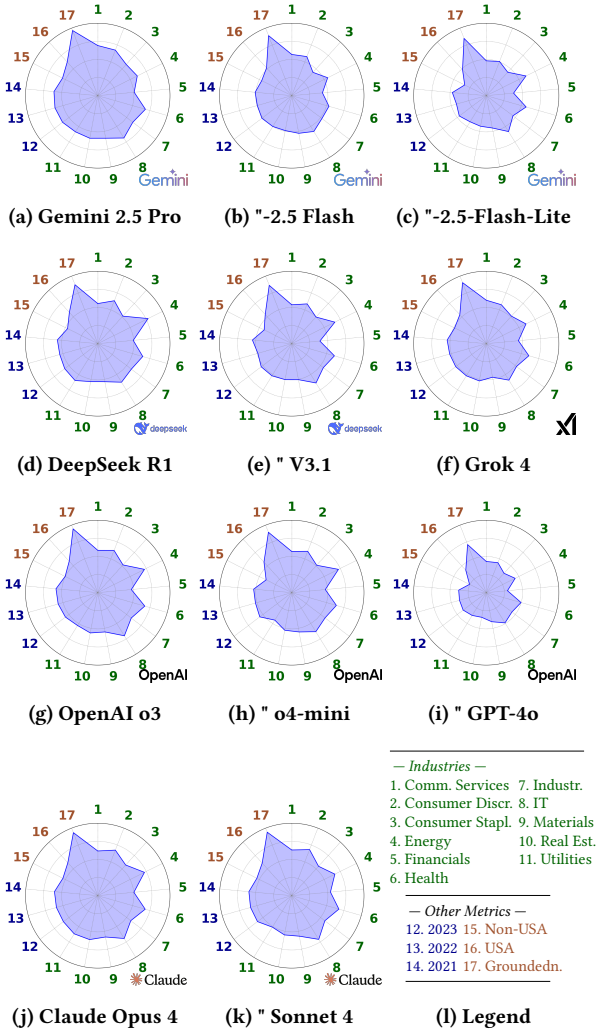
We are aware that changes to the RAG architecture and more powerful embeddings may increase retrieval performance. For that reason, we exclude retrieval errors from most analyses of this study, as in figures 3 and 4. As the retrieval is always the same for all models, the model performance can be perfectly isolated in this way. In other words, the retriever is no confounding variable for model performance.

5 Results

5.1 Model Performance Comparison

We employ multiple evaluation metrics. First, accuracy (recall), which informs about the correctness and completeness of the answers. Second, hallucination rate (opposite term: groundedness or precision) refers to the frequency of claims not supported by source

Figure 5: Multidimensional Performance Comparison of All Models: Each spider chart shows accuracy for industries, years, regions, and overall groundedness (i.e., inverse hallucination rate). Each circle layer is a 25% step.



documents. This metric is equivalent to the precision of the system. Third, we systematically categorize failure points into retrieval failures and generator (LLM) failures. Please note that, as stated in the title of figure 3, the model results exclude retrieval errors. This allows us to compare the pure LLM performance. The human performance, however, does not have such an intermediate step, and thus represents the overall human error rates.

Gemini 2.5 Pro is the best performing available model for financial text today. This holds both for the accuracy (91.6%) and for the hallucination rate (3.2%) of the model. Its accuracy is better than the human baseline, but Gemini 2.5 Pro still has a higher hallucination rate than humans.

o3 has the second-best hallucination rate, but the Claude models (Sonnet and Opus) have a higher accuracy.

Grok 4 is the fifth-best model by hallucination rate, although DeepSeek R1 has a higher accuracy. Gemini 2.5 Flash is between these two models in terms of hallucination prevention, but has a significantly lower accuracy.

o4-mini and DeepSeek V3.1 follow next, with similar accuracy. But DeepSeek has sizably more hallucinations than o4-mini, which is likely a result of o4-mini being a reasoning model.

As the bottom of the list are the non-reasoning models Gemini 2.5 Flash Lite and GPT-4o. GPT-4o has by far the worst accuracy (50.9%) and a 32.6% hallucination rate. Gemini 2.5 Flash Lite, even though it is the second-worst model in the comparison, has only half of GPT-4o's hallucination rate.

Compared to a human baseline, some models have better accuracy, but no model is on par with humans in preventing hallucinations. We found that humans have a hallucination rate of approx. 2.8%, while their accuracy (recall) of 82.8% is already surpassed by Gemini 2.5 Pro, Claude Sonnet 4, and Claude Opus 4.

5.2 Error Analysis

5.2.1 Failure Taxonomy and Analysis. Our analysis of 11,859 failures reveals that retrieval issues dominate (66.5%), followed by hallucination (19.2%) and model comprehension errors (3%). The "Other" category (11.3%) captures complex cases where multiple failure types interact simultaneously, such as format errors combined with temporal confusion or ambiguous questions with partial retrieval, making primary cause attribution impossible.

Table 1: Failure Mode Distribution

Failure Type	Description	Rate
Retrieval Failure	Missing/partial information	66.5%
Hallucination	Fabricated information	19.2%
Model Error	Comprehension failures	3.0%
Other	Complex multi-factor	11.3%

5.2.2 The Cascade Effect. Retrieval failures trigger predictable secondary effects. Fixing retrieval would eliminate 66.5% of total errors, and prevent model hallucinations. Table 1 only lists standalone hallucinations, not hallucinations that are downstream of retrieval errors.

5.2.3 Question Type Impact. Key financials and cash flow questions prove most challenging (figure 6). They often require context that is distributed across multiple pages.

5.2.4 Key Insights.

(1) **Retrieval dominates**: Two-thirds of failures stem from retrieval, not model limitations.
(2) **Cascading failures**: Poor retrieval triggers over half of all hallucinations.
(3) **Model capability is sufficient**: Only 3% true comprehension errors indicate strong model performance when given complete information.
(4) **Intervention hierarchy**: Retrieval enhancement offers error reduction potential that far exceeds possible model improvements.

These findings fundamentally reframe the challenge: financial QA improvement requires better retrieval systems, not more capable models. Complete context retrieval represents the most direct path to accurate financial QA.

5.3 Illustrative Case Studies

To provide a more granular view of model performance, we present four case studies drawn from the benchmark that illustrate the spectrum of difficulty and highlight specific, recurring challenges.

Case Study 1: Universal Difficulty in Data Aggregation. A question asking for a summary of cash flow from BorgWarner's 2023 annual report resulted in a 0% success rate across all eleven models. The task required extracting three distinct figures (operating, investing, and financing cash flow) from a dense, formatted table and aggregating them into a single, coherent answer. The universal failure on this task demonstrates a critical weakness in current models: while proficient at single-point extraction, they struggle with multi-step tasks that require identifying, correctly labeling, and synthesizing several related data points from a single, complex source like a financial statement. This represents a significant hurdle for real-world financial analysis automation.

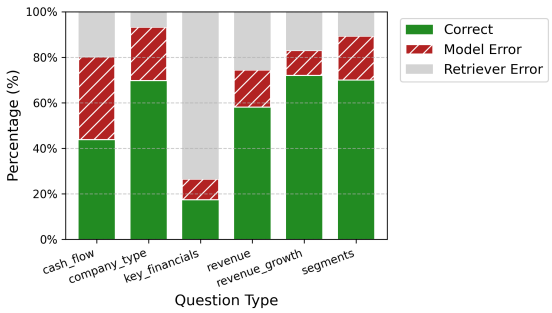
Case Study 2: Straightforward Extraction. In contrast, a question about the cash flow from Imperial Brands' 2022 report was answered correctly by nearly all models. The reason for this high success rate was the simplicity of the source text, where the answer was contained within a single, unambiguous narrative sentence (`...Strong cash performance delivered almost £2.6 billion of free cash flow...`). This "trivial" case shows that when information is presented clearly and requires no aggregation or interpretation of complex formats, models perform with very high reliability.

Case Study 3: A Hallucination Hotspot. A seemingly simple question about the legal form of the Hong Kong Exchanges and Clearing (HKEX) from its 2022 report became a "hallucination hotspot." The correct answer required extracting the single word "Limited" from the company's full name. Instead of performing this simple extraction, many models invented incorrect but plausible-sounding legal forms, such as "Limited Liability Company" or "Corporation." This pattern of failure highlights a tendency for models to overgeneralize or rely on parametric knowledge rather than strictly adhering to the provided context, even when the answer is explicitly present. This is a subtle but critical failure mode, as it can introduce factual inaccuracies into otherwise correct-looking outputs.

Case Study 4: Reasoning vs. Non-Reasoning. One advantage of reasoning over non-reasoning models that we observed when manually reviewing the models' mistakes is that reasoning models seem to systematically scan over the entire context more effectively, reducing the needle-in-the-haystack [21] problem. For example, DeepSeek V3.1 was unable to identify the revenue growth (identified by the synonym "net sales" in the annual report at hand) from the 2021 Steel Dynamics report, even though the retrieved context mentions it in the second chunk. The DeepSeek R1 model, however, was able to correctly answer the question. As more than 1000 tokens preceded the occurrence of the relevant text span containing the revenues, the V3.1 model's attention mechanism likely glossed over this text, not expending enough compute on the given problem. We ran small-scale manual tests to verify this suspicion, and found that indeed the problem is mitigated for non-reasoning models by placing the relevant information toward the end or the beginning of the contexts. We identified some additional error patterns from manual review:

- Reasoning models maintain lower hallucination rates while achieving higher accuracy.
- Reasoning models significantly reduce table-related errors. They can identify the correct rows and columns even in tables that are not legible by human eyes (due to the lack of formatting as a result of OCR).
- Reasoning models reduce temporal confusion errors, while some answers of non-reasoning models display a lack of disambiguation ability in poorly formatted tables.

Figure 6: Average Performance by Question Type: The question about key financials is by far the hardest question for the retriever to answer, while the cash flow question frequently leads to wrong model results.



5.4 Addressing Potential Sources of Experimental Bias in Favor of Selected Models

The lead of Gemini 2.5 Pro over its direct competitors is significant, and begs the question if the experimental conditions unfairly favor this model. We do not think so: The Gemini model family has the largest context window of all models under consideration, but we only utilize a share of less than 10% of this context window here. If we increased the number of tokens retrieved from annual reports, we would expect Gemini to widen its lead as some models' context windows will overflow or get closer to overflowing, only exacerbating Gemini's advantage.

5.5 Checks for Training Data Contamination

Another source of bias may stem from test set leakage via contamination of training data with the annual reports at hand. We do not think this is an issue with our experimental design as the relationship between retriever errors and overall errors shows. When there is a retrieval error, the model produces wrong results in 99.8% of all cases. This means the answers do not originate from the model's inherent knowledge, but from the context.

The comparison of US and non-US in figure 4 suggests an absence of evaluation leakage as well. The US vs. non-US comparison shows that models that are working well on US reports are also working well on non-US reports. If there was significant training data contamination, one would expect to see more inconsistent results here, with some models exhibiting larger gaps between their US vs. their non-US performance, depending on which countries' reports they were trained on.

In addition, from a theoretical point of view, contamination is highly unplausible due to the tiny dataset size that annual reports represent compared to the overall text corpus that these models were (pre-)trained on, giving a low chance that the models were able to internalize year-specific company financials in their model parameters. This holds true especially because we used not only large companies, but also lesser-known companies in our data stratification strategy.

Model performance does not significantly vary by annual report year, with 78.3%, 76.6%, and 79.1% average accuracy in the years 2021, 2022, and 2023, respectively. We conclude that there is no discernible training data leakage from older annual reports into large language models. We would have expected the performance on older reports to be higher than the performance on newer reports if the training data of large language models would be trained on annual reports. This is because usually older data is more readily available in training datasets than newer data.

While this analysis cannot rule out that a leakage effect is present, we wanted to report this finding for full transparency. We think that the empirically validated retrieval dependence shows that the impact of leakage is limited, as the models consistently fail when not given the correct context. This indicates that the relevant answers were not rehearsed from parameterized (trained) world knowledge.

6 Discussion and Conclusion

This paper establishes a critical finding for the field of AI in finance: the primary barrier to automating annual report comprehension is not, as widely assumed, the reasoning capability of language models, but the far more fundamental task of information retrieval. Our comprehensive evaluation provides conclusive evidence. While the latest generation of reasoning-capable LLMs can interpret complex financial data with remarkable accuracy—reaching up to 91.6%—their effectiveness is nullified when the retrieval system fails to provide the correct context, causing accuracy to collapse to a mere 0.2%. This result redefines the problem, shifting the focus from building better reasoners to building better retrievers.

We also definitively show that advanced reasoning architectures are a prerequisite for high-stakes financial analysis. The performance gap between reasoning and non-reasoning models is not incremental but significant; reasoning models deliver a 15+ percentage point uplift in accuracy and are essential for avoiding hallucinations. Their superior ability to parse unstructured tables and maintain context systematically mitigates the needle-in-the-haystack problem.

The implications for the financial industry are immediate and profound. The level of accuracy and reliability demonstrated by the top models in our benchmark confirms that AI is poised to enhance trust and transparency in equity research, directly addressing long-standing issues of analyst bias and conflicts of interest [20].

6.1 Scope and Future Directions

The conclusions of this work are drawn from a deliberately focused scope, which in turn lays a clear foundation for the next wave of research. Our study was confined to **English-language reports**; the path forward involves extending this methodology to assess cross-lingual transfer, a vital step for global financial analysis. The **question coverage**, while representative, was limited to six foundational types. The next frontier is to scale this benchmark to encompass the full breadth of analyst queries, such as the 169 question archetypes in equity research reports identified by [25].

Our use of a standard RAG architecture was a methodological choice designed to isolate and stress-test model comprehension. Having established the retrieval bottleneck, the most urgent future work is to pioneer more advanced retrieval methods. Architectures like GraphRAG [12] represent a promising starting point.

Furthermore, our finding of low inter-model agreement (with the highest Cohen's Kappa being just 0.37 between Grok 4 and o3) points to another key area for future work: **model ensembling**. Future production systems will likely achieve the necessary reliability not through a single model, but through ensembles of top-tier, uncorrelated models. Based on our results, an ensemble combining the top performers from different labs—such as Google's Gemini 2.5 Pro, OpenAI's o3, and Anthropic's Claude Sonnet 4—would be the most promising combination to explore, as it would leverage diverse strengths to maximize accuracy and minimize errors.

In conclusion, this paper provides a robust, unbiased benchmark that answers the question, "Can AI read like a financial analyst?" with a qualified "yes." More importantly, it reveals that the critical question we should be asking is, "Can AI **find** what it needs to read?" Our work demonstrates that solving this retrieval challenge is the central challenge for unlocking the next generation of AI in finance. To accelerate this effort, the complete dataset, evaluation framework, and source code will be made publicly available upon publication.

References

[1] Anthropic. 2025. Introducing Claude 4. *Anthropic Blog* (2025).

[2] Paul Asquith, Michael Mikhail, and Andrea Au. 2005. Information Content of Equity Analyst Reports. *Journal of Financial Economics* 75, 2 (2005), 245–282.

[3] Ray Ball and Sriprakash Kothari. 1991. Security Returns around Earnings Announcements. *The Accounting Review* 66, 4 (1991), 718–738.

[4] Brad Barber, Reuven Lehavy, Maureen McNichols, and Brett Trueman. 2001. Can Investors Profit From the Prophets? Security Analyst Recommendations and Stock Returns. *The Journal of Finance* 56, 2 (2001), 531–563.

[5] Stefano Bonini, Laura Zanetti, Roberto Bianchini, and Antonio Salvi. 2010. Target Price Accuracy in Equity Research. *Journal of Business Finance & Accounting* 37, 9-10 (2010), 1177–1217.

[6] 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. 2021. FinQA: A Dataset of Numerical Reasoning over Financial Data. *ACL Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2021), 3697–3711.

[7] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. In *Forty-first International Conference on Machine Learning (ICML)*.

[8] DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv* (2025).

[9] DeepSeek-AI. 2025. DeepSeek-V3 Technical Report. *DeepSeek Blog* (2025).

[10] Gemini Team, Google. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv* (2025).
[11] Cristi Gleason, Bruce Johnson, and Haidan Li. 2013. Valuation Model Use and the Price Target Performance of Sell-Side Equity Analysts. *Contemporary Accounting Research* 30, 1 (2013), 80–115.
[12] Haoyu Han, Harry Shomer, Yu Wang, Yongjia Lei, Kai Guo, Zhigang Hua, Bo Long, Hui Liu, and Jiliang Tang. 2025. RAG vs. GraphRAG: A Systematic Evaluation and Key Insights. *arXiv* (2025).
[13] Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. FinanceBench: A New Benchmark for Financial Question Answering. *arXiv* (2023).
[14] Wayne Landsman and Edward Maydew. 2002. Has the Information Content of Quarterly Earnings Announcements Declined in the Past Three Decades? *Journal of Accounting Research* 40, 3 (2002), 797–808.
[15] Patrick 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. *Advances in Neural Information Processing Systems (NIPS)* 33 (2020), 9459–9474.
[16] Llama Team, AI @ Meta. 2024. The Llama 3 Herd of Models. *Technical Report* (2024). A detailed contributor list can be found in the appendix of this paper.
[17] Tim Loughran and Bill McDonald. 2011. When is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance* 66, 1 (2011), 35–65.
[18] Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts (Financial Phrase Bank). *Journal of the Association for Information Science and Technology* 65, 4 (2014), 782–796.
[19] Meta AI. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. *Meta AI Blog* (2025).
[20] Roni Michaely and Kent Womack. 1999. Conflict of Interest and the Credibility of Underwriter Analyst Recommendations. *The Review of Financial Studies* 12, 4 (1999), 653–686.
[21] Elliot Nelson, Georgios Kollias, Payel Das, Subhajit Chaudhury, and Soham Dan. 2024. Needle in the Haystack for Memory Based Large Language Models. *ICML 2024 Workshop – Next Generation of Sequence Modeling Architectures* (2024).
[22] OpenAI. 2023. *GPT-4 Technical Report.* Technical Report. The author list is excessively long with more than 200 authors and can thus be found in the technical report only.
[23] OpenAI. 2024. Hello GPT-4o. *OpenAI Blog* (2024).
[24] OpenAI. 2025. Introducing OpenAI o3 and o4-mini. *OpenAI Blog* (2025).
[25] Adria Pop and Jan Spörer. 2025. Identification of the Most Frequently Asked Questions in Financial Analyst Reports to Automate Equity Research Using Llama 3 and GPT-4. *IEEE Swiss Data Science Conference (SDS)* (2025).
[26] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and Efficient Foundation Language Models. *arXiv* (2023).
[27] Kent Womack. 1996. Do Brokerage Analysts' Recommendations Have Investment Value? *The Journal of Finance* 51, 1 (1996), 137–167.
[28] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. BloombergGPT: A Large Language Model for Finance. *arXiv* (2023).
[29] xAI. 2025. Grok 4. *xAI Blog* (2025).
[30] 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. *Annual Meeting of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (2021), 3277–3287.