Evaluating Large Language Models for Table Data Extraction from Annual Reports in PDF

Ngoc Bao Nguyen [1], Levi Kammermann [1], and Thomas Hanne [2,*]

[1] School of Business, University of Applied Sciences and Arts Northwestern Switzerland, Olten, Switzerland
[2] Institute for Information Systems, University of Applied Sciences and Arts Northwestern Switzerland, Olten, Switzerland
Email: ngocbao.nguyen@students.fhnw.ch (N.B.N.); levi.kammermann@students.fhnw.ch (L.K.); thomas.hanne@fhnw.ch (T.H.);
*Corresponding author

***Abstract*—This paper assesses which Large Language Model (LLM), ChatGPT 3.5 or Google Bard (Bard), performs better in extracting tabular data from annual reports of major publicly traded companies in Portable Document Format (PDF). Through an examination of 40 experiments with different difficulty levels, this study delves into the capabilities of these LLMs in processing and extracting financial information. The evaluation metrics, including relevance, accuracy, completeness, consistency, and context awareness reveal varying degrees of proficiency when handling complex financial data for both ChatGPT 3.5 and Bard. In our study, ChatGPT 3.5 demonstrated superior performance, particularly on more challenging questions. The findings offer valuable insights into the utility of LLMs in financial data extraction and provide recommendations for future research.**

***Keywords*—Large Language Models (LLMs), table extraction, table content, information extraction, financial analysis, annual reports, evaluation**

I. Introduction

Large pretrained language models, such as ChatGPT 3.5 and Google Bard (Bard), have demonstrated remarkable success in diverse Natural Language Processing (NLP) applications. However, their performance in certain areas remains insufficiently explored, which introduces uncertainties. A prime example of this is their effectiveness in handling quantitative information commonly presented in tabular formats [1] and often involving numerical data. Surprisingly, there is a limited body of research on the use of Large Language Models (LLMs) in the field of accounting. Previous studies have either concentrated on the contextual aspects surrounding numerical data [2] or have completely neglected numerical information [3].

Annual reports serve as the primary medium through which companies convey historical performance, insights into the external environment, strategies, short-term plans, and expectations for future performance. Since these reports are instrumental in communication with stakeholders, a precise analysis of their contents is imperative [4]. The manual extraction of the relevant information from free-text data is a time-consuming, labor-intensive, error-prone, and expensive process. Hence, Information Extraction (IE) has emerged as a solution to automating the extraction of structured information from unstructured text [5]. In particular, artificial intelligence and NLP offers an alternative approach that may automate the information extraction from text [6]. With the emergence of LLMs, exploring the ability of ChatGPT 3.5 and Bard to extract information from financial reports becomes not only intriguing but also vital.

Among the prevailing data formats, Portable Document Format (PDF) documents stand out because they incorporate tables, graphs, numerical data, and textual content. An evident hurdle, evident in both accounting literature and real-world applications, concerns the extraction of financial data from these PDF documents. This challenge becomes notably more complex when dealing with documents characterized by irregular formatting and varying item descriptions [7]. Moreover, limited research has been conducted on the distinctions between Bard and ChatGPT 3.5. Performing a comparison is crucial for determining the most effective large language model for future research.

In the contemporary financial landscape, the use of LLMs like ChatGPT 3.5 and Bard is increasingly used for formerly manual processes, such as extracting and further processing information from financial documents, such as annual reports, which are usually prepared as PDF files. These reports, vital to stakeholders such as investors and analysts, present a unique challenge because of their complex financial language and structured format. Recognizing the significance of accurate information retrieval, there is a growing need to assess the capabilities of such language models. This study compares the reliability of ChatGPT 3.5 and Bard in extracting information, specifically, from tables included in annual reports and addresses the following thesis statement: "ChatGPT 3.5 demonstrates a more reliable ability to extract information from tables included annual reports compared to Bard".

Manuscript received December 3, 2024; revised December 26, 2024; accepted April 8, 2025; published August 18, 2025.

To prove the thesis statement, the following research questions are suggested:

(1) What existing research and literature address the extraction of quantitative data from PDF documents using language models?
(2) How can ChatGPT 3.5 and Bard be used to extract quantitative information from PDF annual reports?
(3) What types of questions could be used in the prompt to assess the capabilities of ChatGPT 3.5 and Bard in extracting information from financial reports?
(4) What criteria can be used to evaluate the performance of ChatGPT 3.5 and Bard in extracting quantitative financial data from PDF annual reports?

The research questions include exploring the existing literature on language models and financial data extraction from PDFs, understanding the operational aspects of ChatGPT 3.5 and Bard for this purpose, investigating the types of questions used in prompts to assess their capabilities, and identifying criteria for evaluating their performance.

Our study is structured as follows: In Section II, related literature is discussed. Section III presents the used research methodology. Results are presented in Section IV. A discussion is provided in Section V and conclusions are presented in Section VI.

II. Literature Review

In recent years, substantial progress has been achieved in NLP with the development of LLMs like the Generative Pre-trained Transformer (GPT-3). Trained on extensive textual data, these models exhibit impressive capabilities, generating text with human-like quality, providing accurate answers to questions, and successfully completing various language-related tasks [8]. As pretrained foundational models, LLMs are self-supervised and can be fine-tuned to handle diverse natural language tasks. This replaces the need for separate network models for each task, and it marks a significant stride toward achieving the remarkable versatility of human language [9].

Prior automated methods often required substantial effort in setup, involving tasks like crafting parsing rules, fine-tuning or re-training models, or a combination of them. These efforts tailored the method to specific tasks. However, the advent of conversational LLMs presents an opportunity for significantly enhanced information extraction methods that require minimal initial effort by providing only a prompt, that includes a task description and task input in our context [10]. Utilizing the strong general language capabilities of conversational LLMs, which encompass their inherent abilities for zero-shot classification, accurate word reference identification, and efficient retention of information in conversation text, has the potential to enhance data extraction. This improvement is further facilitated by prompt engineering, the process of designing questions and instructions to enhance the result quality. without the need for model fine-tuning or other costly procedures [11].

In financial markets, numerous companies release substantial amounts of textual content alongside relevant numerical data to mitigate information asymmetry between companies and investors. This includes periodic disclosure documents like financial results, securities reports, 10-K, and 10-Q, which encompass a significant amount of textual information [12]. With a notable surge in the volume of such content in recent years, there is widespread adoption of NLP methods for the processing of this data [13]. The application of NLP in the field of financial technology is expansive and intricate, encompassing tasks, such as sentiment analysis, named entity recognition, and question answering. While LLMs have proven effective in various tasks, there is currently a lack of literature reporting on an LLM specifically designed for the financial domain [14].

Annual reports serve as the primary reservoir of information that guides external investors in their investment decisions and empowers shareholders to oversee managerial activities [15]. The financial statements encompass the company's financial performance data, providing insights into operational efficiency and financial robustness throughout the reporting period. It presents comprehensive details through the income statement, balance sheet, and cash flow statement. While hardcopy annual reports continue to be prevalent, electronic versions are on the rise and accessible on company websites in PDF or other formats. Interactive online reports are also gaining traction, allowing users to virtually explore and interact with the content [16].

In the accounting domain, unstructured data are widespread, with PDF documents being the dominant format containing tables, graphs, numbers, and texts. A significant challenge encountered in accounting literature and practice is the extraction of financial information from these PDF documents. This challenge becomes particularly pronounced when dealing with documents characterized by inconsistent formatting and varying item descriptions [7].

The extraction of tabular data from financial PDF documents poses unique challenges, primarily due to the integrity and noise issues in the detected table areas. Financial PDFs often contain tables with complex layouts, including borderless and partially bordered structures; thus, their detection and extraction are nontrivial tasks. Advanced object segmentation techniques in image processing have shown potential in detecting the location of figures and tables in noisy PDF financial documents. These methods address the challenge of extracting data from documents with variable quality and layout complexity [17].

Recent algorithms have been developed to extract information from tables with various layouts in scanned financial PDF documents. These algorithms maintain the structure of tables and achieve high accuracy, and they have proven effective in processing complex financial documents. Despite these advancements, limited attention

has been paid to comprehensive solutions for table detection, extraction, and annotation in financial PDF documents. In addition, the lack of documents and related tools using semantic table information further limits information extraction capabilities [18].

Table extraction remains an area that requires further study. There is a clear need for more comparative studies to evaluate the performance of various LLMs, such as ChatGPT 3.5 and Bard, in extracting table data from financial PDFs. So far, LLMs have been used for table extraction but for different types and formats of data. For instance, in Ref. [19] more simple types of tables (from Wikipedia) and provided in formats easier to analyze than PDF files (such as Comma-Separated Values (CSV), JavaScript Object Notation (JSON), Extensible Markup Language (XML), markdown, Hypertext Markup Language (HTML), and Excel (XLSX), are analyzed by using an LLM and other methods. Thus, there is a need for further studies for identifying the most effective models for this specific application [20]. Future research should focus on developing and refining accurate table data extraction techniques for financial documents. This includes exploring advanced LLMs and AI-based methods to improve extraction accuracy and support more effective business decisions and financial analysis [18].

III. Methodology

A. LLMs and Company Selection

This study focuses on comparing the performance of ChatGPT 3.5 and Bard. These two LLMs were chosen for their widespread media attention and usage in research. They provide distinct approaches to realize transformer networks for LLMs. ChatGPT 3.5, which is an advanced LLM, is reshaping the landscape of AI by showcasing human-like capabilities across a spectrum of tasks. These tasks encompass understanding and responding to natural language questions, language translation, code generation, passing professional exams, and even crafting poetry, among its diverse abilities [21].

Bard AI, which was developed by Google, is a notable breakthrough in the field of AI-powered chatbots. Engineered to emulate human-like interactions, Bard addresses a broad spectrum of user inquiries. Trained on an extensive dataset, Bard leverages advanced language models to produce comprehensive and informative responses. At its core, Bard employs the Pathways Language Model 2 (PaLM 2), which is tailored for superior comprehension of facts, logical reasoning, and mathematics. The model's ability to analyze contextual information allows it to deliver precise and comprehensive responses to diverse queries [22].

The comparison between these two LLMs is of both interest and significance. The decision to analyze ChatGPT 3.5 instead of ChatGPT 4 stemmed from being a free version, making it equitable to compare it with another free version, namely Bard. In addition, it is crucial to highlight that despite Bard's capability to upload PDF files, all experiments in this paper involve manually copying and pasting the PDF content into the prompts. This is because ChatGPT 3.5 does not support the direct upload of PDF files.

To use ChatGPT 3.5, individuals should visit the OpenAI website, create an account, and upon signing in, input a prompt into the text box at the bottom that clearly conveys the desired type of response, which is then provided after pressing "Enter". For utilizing Bard, individuals are required to visit the Google Bard website, sign in with a Google account, input a prompt into the text box at the bottom that effectively conveys the desired type of response, and then press "Enter" to receive an answer from Bard.

This paper uses annual financial reports of thirteen companies: Airbnb, Adobe, Amazon, Apple, Berkshire Hathaway, Coca-Cola, Lindt & Sprüngli, Meta, Microsoft, Nestle, PayPal, Starbucks, and Tesla. These companies were chosen based on two specific criteria: reporting annual revenue exceeding $1 billion and being publicly traded. This selection ensures a diverse representation of industries and operational scales, providing a comprehensive understanding of various financial reporting styles and complexities. The chosen companies not only hold significance in their respective sectors but also offer a range of reporting styles and structures that are essential for testing the adaptability and accuracy of language models across diverse contexts.

B. Data Collection and Evaluation Metrics

This study explores seven question types, comprising a total of 10 questions, categorized into three levels of difficulty. Additionally, it employs two types of computational experiments which were conducted by the authors: one-page and seven-page. In the one-page experiment, only the relevant table (either balance sheet or income statement) from the pertinent financial report is included in the prompt. The seven-page experiment includes not only the relevant financial table from the annual report but also three pages preceding and following it in the prompt. This approach aims to compare the performance of Language Models when presented with varying amounts of information. In the one-page experiment, this paper performs three additional experiments for each question using three distinct financial reports (as repeated executions of the same LLM with the same input may lead to different results due to their probabilistic nature). This approach ensures a relatively equitable comparison of results. Conversely, in the seven-page experiment, only one experiment was conducted for each question. In total, 40 experiments were carried out. Table I lists the question types, questions, difficulty level, and number of experiments.

TABLE I. QUESTION TYPES, QUESTIONS, AND NUMBER OF EXPERIMENTS

Number	Question Type	Question	Difficulty Level	Number of one-page experiments	Number of seven-page experiments
1	Finding the value of an obvious term	What is the Operating Income?	easy	3	1
		What is the Income Tax Expense?	easy	3	1
2	Finding the values of obvious terms and judging whether they are negative or positive values	What is the Other Comprehensive Income? Is it a loss? What is the Comprehensive income	easy	3	1
3	Finding the value of non-obvious terms	What are the short-term and long-term liabilities?	medium	3	1
4	Finding the right category for a non-obvious term	What are the Salaries and Wages expenses?	medium	3	1
5	Finding the value of a non-obvious term based on two obvious	What is the Net Worth?	medium	3	1
		What is the Current Ratio	medium	3	1
		What is the Working Capital?	medium	3	1
6	Finding the value of a non-obvious term based on one obvious term and one non-obvious term	What is the Debt-to-Equity Ratio?	difficult	3	1
7	Finding the value of a non- obvious term based on two non-obvious terms	What is the Debt-to-Capital Ratio?	difficult	3	1
Total					40

Our study evaluates the performance of ChatGPT 3.5 and Bard based on five different criteria: relevance, accuracy, completeness, consistency, and context awareness.

- Relevance: Assess whether the answer directly addresses the given question or prompt. A good answer should be relevant to the context.
- Accuracy: Verify the accuracy of the information provided. Ensure that the factual details are correct and that the response aligns with the query context. A value is deemed accurate if it is correct up to two decimal places.
- Completeness: Determine if the answer is comprehensive and covers all relevant aspects of the question. A good response should not be overly brief or excessively verbose. A good answer should also be easy to understand, thereby avoiding ambiguity or confusion.
- Consistency: Check for consistency in the response. Ensure that the information provided does not contradict itself and remains coherent throughout.
- Context Awareness: Evaluate whether the answer demonstrates an understanding of the broader context and provides contextually relevant information.

When a criterion is fulfilled, a grade of 20% is awarded. For a single question and criterion, the score is usually either 0% or 20%. A value between 0% and 20% for a single question may result as an average value if multiple experiments are done (as for the one-page experiments) or when the question comprises several subquestions (as for question 5). The assessment was done by one of the authors (Nguyen). For the overall evaluation of an experimental setup the criterion-specific scores were added, and the following grading scheme of the experimental results was used:

- 100% = Excellent
- 90% = Very Good
- 80% = Good
- 70% = Satisfactory
- 60% = Sufficient
- Below 60% = Insufficient

C. Prompt Engineering

The prompts presented to ChatGPT and Bard follow a specific format:

"What is the value of X in the year Y, based on the following information: Z?"

Here, X represents the financial term, Y denotes the specific year, and Z represents the information copied from the related PDF financial report.

TABLE II. THE DESIGN OF QUESTION TYPES

Question Type	Assessment Competence
1	Identify and retrieve financial terms explicitly stated in the financial report.
2	Identify and retrieve financial terms explicitly stated in the financial report and evaluate them.
3	Identify and retrieve a financial term that is not explicitly stated in the financial report, possibly by identifying synonyms for the financial term.
4	Recognize the financial term and determine its location in the report. This requires the capability to comprehend various term definitions and their contextual usage.
5	Identify a financial term not explicitly stated in the financial report and find its value by performing calculations based on two explicitly mentioned terms.
6	Recognize a financial term that is not explicitly mentioned in the financial report and determine its value through calculations involving one explicitly mentioned term and one not explicitly mentioned term. This may include two calculations.
7	Recognize a financial term that is not explicitly mentioned in the financial report and determine its value through calculations involving two terms that are not explicitly mentioned. This may include multiple calculations.

Upon receiving responses from both ChatGPT 3.5 and Bard, we capture screenshots encompassing both the prompts and their respective answers. Subsequently, we assembled a table outlining the criteria that served as benchmarks for evaluating the effectiveness of ChatGPT and Bard in addressing the posed questions.

Table II shows the design of the seven question types to understand how ChatGPT 3.5 and Bard extract information from financial reports. The difficulty level was increased throughout the questions. "Obvious terms" refer to the financial terms explicitly stated in the financial reports, while "non-obvious terms" refer to those not explicitly mentioned in the financial reports.

IV. Results

A. Results for Different Question Types

This subsection presents the results for each question type. The percentage assigned to each criterion within a particular question type represents the average outcome of that criterion across all experiments conducted for that specific question type. The total percentage for each question type was computed by summing up the results of all five criteria. This indicates the performance of ChatGPT 3.5 and Bard for each question type in both one-page and seven-page experiments.

When it comes to finding the value of an obvious term, in the one-page experiment, ChatGPT 3.5 meets all five criteria and achieves the highest score (100%), whereas Bard's performance is sufficient (66%) (see Table III). Bard fails to provide relevant context regarding the financial term asked in the prompt. In addition, in one out of three experiments, Bard provides an inaccurate value, and in another one out of three experiments, it fails to maintain consistency throughout the answer—initially offering one value and later providing a different one for the same question. In this scenario, ChatGPT 3.5 demonstrated better performance than Bard.

TABLE III. Results for Question Type 1 (Averages for All Annual Reports)

Criteria	One-page Experiment		Seven-page Experiment	
	ChatGPT 3.5 (%)	Bard (%)	ChatGPT 3.5 (%)	Bard (%)
Relevance	20	20	20	20
Accuracy	20	13.3	0	0
Completeness	20	20	0	0
Consistency	20	13.3	0	0
Context Awareness	20	0	0	0
Total	100	66	20	20

Note: A value of 20% indicated a perfect score in one category (all sum up to 100%). A smaller value such as 13.3% indicated that 2/3 of the experiments were successful in the respective category.

In the context of the seven-page experiment, both ChatGPT 3.5 and Bard identify the financial term mentioned in the prompt, thus satisfying the relevance criterion. However, they fall short of meeting the remaining four criteria, and each receives a total score of 20%. This indicates that when the prompt contains not only the relevant financial report but also additional information, both ChatGPT 3.5 and Bard encounter difficulties in analyzing all the contexts to deliver an accurate and comprehensive response. In such cases, the performance of both ChatGPT 3.5 and Bard is considered insufficient.

ChatGPT outperformed Bard in the one-page experiment. Both language models performed equally poorly in the seven-page experiments.

When it comes to finding the value of an obvious term and evaluating whether it is negative or positive, in the one-page experiment, ChatGPT 3.5 and Bard met all the criteria except for context awareness (see Table IV). In two of the three experiments, both models fail to offer pertinent context regarding the financial term mentioned in the prompt. Consequently, each large language model obtained a total score of 86.7%, which is more than good but not enough to be considered very good (90%). In this context, both ChatGPT 3.5 and Bard exhibit equal performance.

TABLE IV. Results for Question Type 2 (Averages for All Annual Reports)

Criteria	One-page Experiment		Seven-page Experiment	
	ChatGPT 3.5 (%)	Bard (%)	ChatGPT 3.5 (%)	Bard (%)
Relevance	20	20	20	20
Accuracy	20	20	0	20
Completeness	20	20	20	20
Consistency	20	20	20	20
Context Awareness	6.7	6.7	20	0
Total	86.7	86.7	80	80

In the context of the seven-page experiment, both ChatGPT 3.5 and Bard successfully identified the financial term mentioned in the prompt, fulfilling the relevance criterion. Furthermore, they deliver comprehensive and easy-to-understand answers, highlighting completeness. Additionally, their responses display coherence and maintain consistency without any contradictions, thus meeting the consistency criterion. Both ChatGPT 3.5 and Bard achieved a score of 80%, which indicates equally good performance for both models.

ChatGPT 3.5 and Bard performed equally well in both the one-page and seven-page experiments.

When tasked with determining the value of non-obvious terms from the financial report in the one-page experiment, Bard performed very well with a score of 93.3%, while ChatGPT 3.5's performance was deemed satisfactory at 73.3% (see Table V). Bard successfully meets all criteria except accuracy, where it falters in providing a correct value in one out of three experiments. ChatGPT 3.5 fails to offer relevant context regarding the financial term mentioned in the prompt (Relevance). Moreover, similar to Bard, ChatGPT 3.5 also falls short in providing a correct value in one out of three experiments. However, ChatGPT 3.5 excels at meeting three criteria: relevance, completeness, and consistency. In this context, Bard performs better than ChatGPT 3.5.

TABLE V. RESULTS FOR QUESTION TYPE 3 (AVERAGES FOR ALL ANNUAL REPORTS)

Criteria	One-page Experiment		Seven-page Experiment	
	ChatGPT 3.5 (%)	Bard (%)	ChatGPT 3.5 (%)	Bard (%)
Relevance	20	20	20	20
Accuracy	13.3	13.3	0	20
Completeness	20	20	0	20
Consistency	20	20	0	20
Context Awareness	0	20	0	20
Total	73.3	93.3	20	100

In the seven-page experiment, while Bard fulfilled all five criteria and achieved the maximum score of 100%, ChatGPT 3.5 failed to meet all criteria except relevance and received an insufficient score of 20%. This indicates that when the prompt contains not only the relevant financial report but also additional information, Bard performed exceptionally well compared to when only the relevant financial report is included in the prompt. In contrast, ChatGPT 3.5 faced difficulties in analyzing all contexts to deliver accurate and comprehensive responses. In such cases, the performance of Bard is clearly better than that of ChatGPT 3.5.

Bard excels over ChatGPT 3.5 in both one-page and seven-page experiments.

When tasked with determining the right category for a non-obvious term in the financial report, in the one-page experiment, ChatGPT 3.5 performed very well with a score of 93.3%, while Bard's performance was deemed satisfactory at 73.3% (see Table VI). Both LLMs met the criteria of relevance, completeness, and consistency and failed to give the correct value in one of the three experiments (Accuracy). Bard fails to offer a relevant context regarding the financial term mentioned in the prompt (Relevance) in all three experiments. In this context, ChatGPT 3.5 outperformed Bard.

TABLE VI. RESULTS FOR QUESTION TYPE 4 (AVERAGES FOR ALL ANNUAL REPORTS)

Criteria	One-page Experiment		Seven-page Experiment	
	ChatGPT 3.5 (%)	Bard (%)	ChatGPT 3.5 (%)	Bard (%)
Relevance	20	20	20	20
Accuracy	13.3	13.3	20	0
Completeness	20	20	20	20
Consistency	20	20	20	0
Context Awareness	20	0	20	0
Total	93.3	73.3	100	40

In the seven-page experiment, while ChatGPT 3.5 fulfilled all five criteria and achieved the highest score of 100%, Bard only met two out of five criteria (Relevance and Completeness) and received an insufficient score of 40%. This indicates that when the prompt contains not only the relevant financial report but also additional information, ChatGPT 3.5 performs exceptionally well when context information is included in the prompt. In contrast, Bard encounters difficulties in analyzing all contexts to deliver an accurate and comprehensive response. In such cases, the performance of ChatGPT 3.5 is clearly better than that of Bard.

ChatGPT 3.5 excels over Bard in both one-page and seven-page experiments.

When tasked with the value of a non-obvious term based on two obvious terms, in the one-page experiment, ChatGPT 3.5 and Bard performed well with scores of 86.7% and 80%, respectively (see Table VII). Both LLMs met the relevance and completeness criteria. While ChatGPT 3.5 always showed consistency in the answers, Bard failed to do so in four out of nine experiments (Consistency). In addition, ChatGPT 3.5 failed to give the correct value in six out of nine experiments (Accuracy). In five out of nine experiments, Bard failed to give the correct value. Both ChatGPT 3.5 and Bard manage to offer a relevant context regarding the financial term mentioned in the prompt (Relevance) in all nine experiments. In this context, ChatGPT 3.5 performs slightly better than Bard.

TABLE VII. RESULTS FOR QUESTION TYPE 5 (AVERAGES).

Criteria	One-page Experiment		Seven-page Experiment	
	ChatGPT 3.5 (%)	Bard (%)	ChatGPT 3.5 (%)	Bard (%)
Relevance	20	20	20	20
Accuracy	6.7	8.9	6.7	20
Completeness	20	20	20	20
Consistency	20	11.1	13.3	20
Context Awareness	20	20	13.3	13.3
Total	86.7	80	73.3	93.3

Note: The scores are the average values for the three subquestions and for all annual reports (with a maximum value of 20) for the two LLMs for one-page and seven-page experiments.

In the seven-page experiment, while Bard fully meets four out of five criteria (relevance, accuracy, completeness and consistency), ChatGPT 3.5 only fully met relevance and completeness. ChatGPT 3.5 fails to provide accurate values in two out of three experiments. Both ChatGPT 3.5 and Bard fail to offer a relevant context regarding the financial term mentioned in the prompt (Relevance) in one of three experiments. Although Bard fully met the consistency criterion, ChatGPT only manages to maintain consistency in two of three experiments.

In the one-page experiment, ChatGPT 3.5 outperformed Bard, whereas in the seven-page experiments, Bard surpassed ChatGPT 3.5 in performance.

When tasked with finding the value of a non-obvious term based on one obvious term and one non-obvious term, in the one-page experiment, ChatGPT 3.5 performed well with a score of 86.6%, while Bard only reached a sufficient score of 60% (see Table VIII). Both LLMs met the following criteria: relevance, completeness, and context awareness. While ChatGPT failed to provide a correct value in one of three experiments, Bard failed to provide an accurate value in all three experiments. Similarly, ChatGPT maintained consistency in two of three answers; however, Bard failed to do so in all three experiments.

TABLE VIII. RESULTS FOR QUESTION TYPE 6 (AVERAGES FOR ALL ANNUAL REPORTS)

Criteria	One-page Experiment		Seven-page Experiment	
	ChatGPT 3.5 (%)	Bard (%)	ChatGPT 3.5 (%)	Bard (%)
Relevance	20	20	20	20
Accuracy	13.3	0	0	0
Completeness	20	20	20	20
Consistency	13.3	0	0	0
Context Awareness	20	20	20	0
Total	86.6	60	60	40

In the seven-page experiments, both ChatGPT and Bard performed worse than in the one-page experiments. While ChatGPT only achieved a sufficient score of 60%, Bard's score was insufficient at 40%. Both LLMs met the criteria of relevance and completeness and failed to provide accurate values and consistency in the answers. While ChatGPT could provide relevant context to the prompt, Bard failed to do so.

ChatGPT 3.5 outperformed Bard in both the one-page and seven-page experiments.

When it comes to finding the value of a non-obvious term based on two non-obvious terms, in the one-page experiment, ChatGPT only reached a sufficient score of 60.1%, and Bard's performance was insufficient at 40% (see Table IX). Both ChatGPT 3.5 and Bard met the relevance and context awareness criteria. While ChatGPT met accuracy, completeness, and consistency in one of three experiments, Bard failed to do so.

TABLE IX. RESULTS FOR QUESTION TYPE 7 (AVERAGES FOR ALL ANNUAL REPORTS)

Criteria	One-page Experiment		Seven-page Experiment	
	ChatGPT 3.5 (%)	Bard (%)	ChatGPT 3.5 (%)	Bard (%)
Relevance	20	20	20	20
Accuracy	6.7	0	0	0
Completeness	6.7	0	0	0
Consistency	6.7	0	0	0
Context Awareness	20	20	20	0
Total	60.1	40	40	20

In the seven-page experiment, both ChatGPT 3.5 and Bard performed insufficiently with 40% and 20%, respectively. ChatGPT met the criteria relevance and context awareness, while Bard only fulfilled the relevance criterion. Both LLMs failed with accuracy, completeness, and consistency.

ChatGPT 3.5 outperformed Bard in both the one-page and seven-page experiments.

B. Ranking of Considered LLMs

Table X lists the rankings of ChatGPT 3.5 and Bard for the one-page and seven-page experiments.

In the one-page experiment, ChatGPT 3.5 had an average grade of 83.8% (good) and was positioned across the spectrum from sufficient to excellent. Bard has an average grade of 71.3% (satisfactory), and its rankings range from insufficient to very good. In five of seven question types, ChatGPT 3.5 outperforms Bard, experiences a loss in one, and has a tie in one.

TABLE X. CHATGPT 3.5 VS. BARD

Question Type	One-page Experiment		Seven-page Experiment	
	ChatGPT 3.5	Bard	ChatGPT 3.5	Bard
1	Excellent (100%) Winner	Sufficient (66%)	Insufficient (20%) Tie	Insufficient (20%) Tie
2	Good (86.7%) Tie	Good (86.7%) Tie	Good (80%) Tie	Good (80%) Tie
3	Satisfactory (73.3%)	Very good (93.3%) Winner	Insufficient (20%)	Excellent (100%) Winner
4	Very good (93.3%) Winner	Satisfactory (73.3%)	Excellent (100%) Winner	Insufficient (40%)
5	Good (86.7%) Winner	Good (80%)	Satisfactory (73.3%)	Very good (93.3%) Winner
6	Good (86.6%) Winner	Sufficient (60%)	Sufficient (60%) Winner	Insufficient (40%)
7	Sufficient (60.1%) Winner	Insufficient (40%)	Insufficient (40%) Winner	Insufficient (20%)
Average	83.8%	71.3%	56.2	56.2

In the seven-page experiment, ChatGPT 3.5 earned an average grade of 56.2% (insufficient) and ranged from insufficient to excellent in rankings. Bard also achieved an average grade of 56.2% (insufficient) and its rankings cover the range from insufficient to excellent. Across seven question types, ChatGPT 3.5 surpasses Bard in three, encounters two losses, and achieves a tie in two.

In summary, ChatGPT 3.5 outperformed Bard in both the one-page and seven-page experiments.

Table XI shows the average performance of both ChatGPT 3.5 and Bard in one-page and seven-page experiments for each question type.

TABLE XI. PERFORMANCE OF CHATGPT 3.5 AND BARD FOR EACH QUESTION TYPE

Question Type	One-page Experiment (%)	Seven-page Experiment (%)
1	83	20
2	86.7	80
3	83.3	60
4	83.3	70
5	83.4	83.3
6	73.3	50
7	50.1	30

As shown in Table XI, for the one-page experiments, both LLMs performed well, ranging from 83% to 86.7%, in question types 1 to 5. They achieved satisfactory results in question type 6 (73.3%) but fell short, registering insufficient performance in question type 7 (50.1%). This indicates that the chosen difficulty levels for each question type were appropriately designed—easy for question types 1 and 2, medium for question types 3, 4, and 5, and difficult for question types 6 and 7.

The same rationale does not hold true for the seven-page experiment. Despite question type 1 being the

easiest, both LLMs performed the worst, with an average of 20%. Question types 2, 3, 4, and 5 yield results ranging from sufficient to good. However, for the challenging question types 6 and 7, the outcomes were indeed insufficient.

V. Discussion

Tables III–IX demonstrate that both ChatGPT and Bard consistently meet the relevance criterion, indicating their ability to recognize the financial terms mentioned in the prompt and attempt to address them in their responses. However, there is no fixed pattern for the remaining four criteria—accuracy, completeness, consistency, and context awareness. This suggests that the responses generated by both language models exhibit a degree of randomness and unpredictability, which impacts the overall quality of the answers. However, ChatGPT 3.5 outperformed Bard on average.

Table X shows that in both the one-page and seven-page experiments, ChatGPT 3.5 outperformed Bard overall. Specifically, in the one-page experiments, on average, ChatGPT 3.5 demonstrated good performance (83.8%), while Bard's performance was considered satisfactory (71.3%). However, both language models face challenges in the seven-page experiments, resulting in insufficient average results (56.2%). This suggests that the quantity of information provided in the prompt significantly influences the performance of ChatGPT and Bard. More information tends to make it more challenging for the models to accurately determine values and provide comprehensive answers.

Table X illustrates that although the overall average performance in the one-page experiments exceeds that of the seven-page experiments, there are exceptions at the level of individual question types. For example, Bard exhibited better performance in question types 3 and 5 during the seven-page experiment than during the one-page experiment. Similarly, for question type 4, ChatGPT 3.5 performs better in the seven-page experiment than in the one-page experiment. However, the remaining question types and experiments align with the overall logic that a seven-page experiment presents greater challenges than a one-page experiment.

Table XI shows that the difficulty level of the question in the prompt influences the outcomes of the one-page experiment. The more difficult the question is, the more challenging it becomes for both ChatGPT 3.5 and Bard to provide a correct and comprehensive answer. However, this pattern is not entirely consistent in the seven-page experiment, where question type 1, the easiest, results in the worst performance for both LLMs. Despite this, for question types 2 to 7, a trend emerges: the more difficult the question, the lower their performance. This suggests a possibility that both models may have the capability to enhance their performance by learning from previous questions.

We consider the results as promising although they indicate the current LLM technology is still far away from providing a reliable solution for table content extraction and questions answering regarding information in annual reports such as balance sheets and income statements. As there is no similar research suitable for comparing our results with, let us mention that even for easier tasks such as cell-type classification the results and utilizing supervised learning, only accuracy rates of 82%–95% were achieved [23].

Regarding the reasons for the better performance of ChatGPT 3.5, let us mention that the results are in line with other comparative evaluations of the two models such as Ref. [24]. It is, however, not possible to provide clear reasons for the different performances as various details of the two LLMs such as model structure, training approaches, and training data are not fully transparent.

VI. Conclusions

Overall, ChatGPT 3.5 demonstrated superior performance compared to Bard in terms of extracting financial information from PDF annual reports. Our study covers a total of forty experiments, based on seven types of questions and three difficulty levels. The evaluation of responses by ChatGPT 3.5 and Bard employed five criteria: relevance, accuracy, completeness, consistency, and context awareness. Despite the observed randomness and unpredictability in the responses generated by both language models, it is evident that the quantity of information provided in the prompt significantly influences their performance. The inclusion of more information poses challenges for both ChatGPT and Bard in accurately determining values and providing comprehensive answers. Additionally, the difficulty level of the question within the prompt directly affects the outcomes of the one-page experiment. The findings suggest that both models have the potential to enhance their performance over time.

The thesis statement posits that ChatGPT 3.5 exhibits a more dependable ability to extract information from PDF annual reports than Bard. After conducting the experiments, the thesis statement was substantiated. We focused on determining which of the two large language models, ChatGPT 3.5 or Bard, displays superior performance in extracting information from PDF financial reports. The evidence indicates that ChatGPT 3.5 outperforms Bard when extracting financial information from annual reports. Following the experiments and subsequent analysis, the approach to addressing the thesis statement remains consistent by addressing the associated research questions.

This paper is among the pioneering efforts to compare the capabilities of ChatGPT 3.5 and Bard in extracting quantitative information from PDF annual reports. Consequently, this study has several inherent limitations. Due to the scarcity of existing scientific papers on the same subject, establishing a standard for research in this area is challenging, making it difficult to benchmark the present study against other studies. The design of the questions in this paper specifically revolves around instructing LLMs to retrieve one or two values, leaving uncertainties about their performance when tasked with simultaneously identifying a higher number of values. The prompts in this study are formulated in clear, precise,

and grammatically correct language, excluding the exploration of results when prompts are composed in complex or grammatically incorrect language. In addition, since all prompts were written in English, the findings may not be representative of other languages. The inclusion of numerical values in the prompts is consistently in the format of thousands or millions, leaving uncertainty about the LLMs' ability to identify results in these units if not explicitly stated. The grading system spans from insufficient to excellent, where all values below 60% are considered sufficient. Consequently, distinctions between, for instance, 20% and 40% remain indistinguishable within this range.

Future research could explore a diverse array of question designs considering prompts that require simultaneous retrieval of a higher number of values or incorporation of complex language structures to assess model adaptability. Additionally, extending the evaluation to include prompts in languages other than English would provide insights into the cross-linguistic performance of the language models. Furthermore, a refined grading system with increased granularity could provide nuanced insights into the proficiency of the models. Finally, as the field evolves, efforts should be made to establish a standardized approach for evaluating language models, promoting benchmarking against established criteria for enhanced comparability across studies. By addressing these areas, future research can contribute to a more thorough understanding of the capabilities and limitations of ChatGPT 3.5 and Bard in the extraction of financial information from PDF annual reports. Another line of future research may be using a hybrid approach of advanced supervised machine learning techniques in combination with LLMs and/or using structured information such as ontologies or knowledge graphs in addition for improving performance.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

NBN and LK conducted the research and wrote the first version of the paper; TH supervised the research, validated the results, and edited the paper; all authors had approved the final version.

References

[1] I. Qureshi, M. Karakaplan, and R. de Padua, "GPT-3 models are few-shot financial reasoners," in *Proc. CS & IT – CSCP 2023*, Toronto, 2023, vol. 13, no. 12, pp. 183–197.

[2] A. Kim and V. V. Nikolaev, "Context-based interpretation of financial information," *Journal of Accounting Research*, 2024. https://doi.org/10.1111/1475-679X.12593

[3] S. Kuester, T. Steindl, and M. Goettsche, "The informational content of key audit matters: Evidence from using artificial intelligence in textual analysis," *SSRN Electronic Journal*, 2023. https://doi.org/10.2139/ssrn.4464713

[4] S. Pasmini, R. Srinivasan, and M. A. Cristina, "Narrative analysis of annual reports: A study of communication efficiency," *SSRN Electronic Journal*, no. 486, 2017. doi: 10.2139/ssrn.2611890

[5] S. Datta, E. Bernstam, and K. Roberts, "A frame semantic overview of NLP-based information extraction for cancer-related EHR notes," *Journal of Biomedical Informatics*, vol. 100, 103301, 2019. doi: https://doi.org/10.1016/j.jbi.2019.103301

[6] T. Gupta, M. Zaki, N. Krishnan *et al*., "MatSciBERT: A materials domain language model for text," *Computational Materials*, vol. 8, no. 1, 102, 2022.

[7] H. Li, H. Gao, C. Wu *et al*., "Extracting financial data from unstructured sources: Leveraging large language models," *SSRN Electronic Journal*, 2023. https://doi.org/10.2139/ssrn.4567607

[8] L. Floridi and M. Chiriatti, "GPT-3: Its nature, scope, limits, and consequences," *Minds and Machines*, vol. 30, pp. 681–694, 2020.

[9] T. J. Sejnowski, "Large language models and the reverse Turing test," *Neural Computation*, vol. 35, no. 3, pp. 309–342, 2023. https://doi.org/10.1162/neco_a_01563

[10] A. Brinkmann, R. Der, R. Shraga *et al*., "Product information extraction using ChatGPT," arXiv preprint, arXiv:2306.14921, 2023.

[11] M. Polak and D. Morgan, "Extracting accurate materials data from research papers with conversational language models and prompt engineering," arXiv preprint, arXiv:2303.05352, 2023.

[12] M. Suzuki, H. Sakaji, M. Hirano *et al*., "Constructing and analyzing domain-specific language model for financial text mining," *Information Processing and Management*, vol. 60, no. 2, 103194, 2023. https://doi.org/10.1016/j.ipm.2022.103194

[13] B. Kumar and V. Ravi, "A survey of the applications of text mining in financial domain," *Knowledge-Based Systems*, vol. 114, pp. 128–147, 2016. https://doi.org/10.1016/j.knosys.2016.10.003

[14] S. Wu, O. Irsoy, S. Lu *et al*., "BloombergGPT: A large language model for finance," arXiv preprint, arXiv:2303.17564, 2023.

[15] J. -H. Luo, X. Li, and H. Chen, "Annual report readability and corporate agency costs," *China Journal of Accounting Research*, vol. 11, no. 3, pp. 187–212, 2018.

[16] P. Swain. (2024). What is an annual report. [Online]. Available: https://de.scribd.com/document/693260645/Annual-Report

[17] C. Zhou, Q. Li, C. Li *et al*., "A comprehensive survey on pretrained foundation models: A history from BERT to ChatGPT," arXiv preprint, arXiv:2302.09419, 2023.

[18] A. Valentina, *Modern Generative AI with Chat GPT and OpenAI Models: Leverage the Capabilities of OpenAI's LLM for Productivity and Innovation with GPT3 and GPT4*, 1st ed. Birmingham, United Kingdom: Packt Publishing Ltd, 2023.

[19] Y. Sui, M. Zhou, M. Zhou *et al*., "Table meets LLM: Can large language models understand structured table data? A benchmark and empirical study," in *Proc. the 17th ACM International Conf. on Web Search and Data Mining,* 2024, pp. 645–654.

[20] C. Tensmeyer, V. Morariu, B. Price *et al*., "Deep splitting and merging for table structure decomposition," in *Proc. 2019 International Conf. on Document Analysis*, Sydney, 2019, pp. 114–121.

[21] I. Alberts, L. Mercolli, T. Pyka *et al*., "Large Language Models (LLM) and ChatGPT: What will the impact," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 50, no. 6, pp. 1549–1552, 2023.

[22] S. Pichai. (February 2023). An important next step on our AI journey. [Online]. Available: https://blog.google/technology/ai/bard-google-ai-search-updates/

[23] K. Kadowaki, Y. Kimura, and H. Ototake, "Towards enhanced information access in finance: A dataset for table structure understanding in annual securities reports," in *Proc. 2024 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics (CIFEr)*, Hoboken, 2024, pp. 1–6.

[24] I. Ahmed, A. Roy, M. A. Kajol *et al*., "ChatGPT vs. Bard: A comparative study," Authorea Preprints, 2023. doi: 10.22541/au.168923529.98827844/v1