computers

MDPI

Article

Assessing Large Language Models Used for Extracting Table Information from Annual Financial Reports

David Balsiger [1], Hans-Rudolf Dimmler [1], Samuel Egger-Horstmann [1] and Thomas Hanne [2,*]

[1] School of Business, University of Applied Science and Arts Northwestern Switzerland, 4600 Olten, Switzerland
[2] Institute for Information Systems, University of Applied Science and Arts Northwestern Switzerland, 4600 Olten, Switzerland
* Correspondence: thomas.hanne@fhnw.ch; Tel.: +41-62-957-22-92

Abstract: The extraction of data from tables in PDF documents has been a longstanding challenge in the field of data processing and analysis. While traditional methods have been explored in depth, the rise of Large Language Models (LLMs) offers new possibilities. This article addresses the knowledge gaps regarding LLMs, specifically ChatGPT-4 and BARD, for extracting and interpreting data from financial tables in PDF format. This research is motivated by the real-world need to efficiently gather and analyze corporate financial information. The hypothesis is that LLMs—in this case, ChatGPT-4 and BARD—can accurately extract key financial data, such as balance sheets and income statements. The methodology involves selecting representative pages from 46 annual reports of large Swiss corporations listed in the SMI Expanded Index from 2022 and copy–pasting text from these into LLMs. Eight analytical questions were posed to the LLMs, and their responses were assessed for accuracy and for identifying potential error sources in data extraction. The findings revealed significant variance in the performance of ChatGPT-4 and another LLM, BARD, with ChatGPT-4 generally exhibiting superior accuracy. This research contributes to understanding the capabilities and limitations of LLMs in processing and interpreting complex financial data from corporate documents.

Keywords: table extraction; larger language models; LLMs; annual reports

check for updates

Citation: Balsiger, D.; Dimmler, H.-R.; Egger-Horstmann, S.; Hanne, T. Assessing Large Language Models Used for Extracting Table Information from Annual Financial Reports. *Computers* **2024**, *13*, 257. https://doi.org/10.3390/computers13100257

Academic Editor: Ming Liu

Received: 4 September 2024
Revised: 25 September 2024
Accepted: 7 October 2024
Published: 9 October 2024

1. Introduction

With the release of Microsoft Copilot, ChatGPT, and other AI tools, the working environment and tools for many jobs are shifting quickly. Although these tools can improve the efficiency and accuracy of users, it is important to be aware of the capabilities and limitations of different applications. In 2020, a discussion was held at Stanford University with scientists from different backgrounds about the capabilities, limitations, and societal impact of Large Language Models (LLMs) [1]. One issue mentioned in this discussion was that identifying and understanding the limitations of these models is critical and should be performed sooner rather than later. Since then, significant research has been conducted to identify the capabilities and limitations of various AI tools. This study adds to this ongoing research with a focus on extracting information included in tables (such as those frequently used in financial documents).

The purpose of this research proposal is to investigate and evaluate the effectiveness of LLMs, such as GPT-4 and BARD, in extracting information from tables in official annual financial reports of companies listed on the SMI Expanded. This research assesses the quality and accuracy of responses generated by LLMs when presented with a predefined set of questions pertaining to the content within these tables. The primary objectives of this study are as follows:

- To determine the capacity of LLMs to accurately extract financial information from tables in official annual reports.

- To quantify and assess the quality of responses generated by LLMs.
- To identify potential limitations and challenges in using LLMs for this specific information extraction task.

Table extraction from PDF files has been a research topic that has been addressed repeatedly during the 2000s, as discussed by Sarawagi [2] and Krapivin et al. [3]. Despite this rather long history of research on table content extraction, existing approaches still have limitations, especially for more complex tables as found, e.g., in financial reports. For this reason, and because LLMs provide further use cases, such as providing information in interactive conversations, LLMs appear to be an interesting alternative to older approaches. However, there is not much research available when it comes to extracting data from tables using LLMs. A possible usage of LLMs is to copy and paste information from PDF files directly into the LLM input and then ask the model for summaries or analytical questions based on the given input. When writing this paper, no research was available that had studied the capabilities and limitations of LLMs based on the accuracy of data extraction when copy–pasting tables from PDF documents into an LLM. When copy–pasting, one directly copies the text of a table out of a PDF file into the temporary storage of the operating system and pastes it into the input of an LLM without aggregating or changing the format or data. This approach is limited to PDF documents that were not scanned and have the table stored as text. During this processing, the location information of the text (included in bounding boxes in the PDF files) is lost.

To obtain a sufficiently large sample of tables in a similar format, annual financial reports from large Swiss companies were chosen as input. This also impacts the importance of this research, as there is a use case for people (e.g., investors or financial analysts) trying to gather information about these and other corporations adopting Large Language Models.

For our study, an experiment-based research design was chosen. The hypothesis considered for this investigation is as follows: We assume that LLMs are able to correctly extract the data from balance sheets and income statements included in PDF files from large Swiss companies when copy–paste is used to input the data.

The following research questions were considered during this study:

- How accurate is the LLM-based interpretation of financial tables from the considered annual reports?
- Where are the possible sources of errors in extraction?
- Are there problems resulting from the copy–pasting of tables from PDF files?

In the next section of our paper, we discuss the related literature. In Section 3, the research methodology is explained. The obtained results are presented and discussed in Section 4. This paper ends with conclusions, presented in Section 5.

2. Literature Review

Starting with a keyword search, many resources can be found when searching for topics like "LLM", "Table Extraction", and "Financial Data AI" or tool-specific searches like "Chat GPT". While searching for comparable work in data extraction [4–6], and [7] proved to be very valuable for backward searching.

In recent years, we have seen a rise in the number of various Large Language Models (LLMs). Although the concept of a chatbot is nothing new, the revolution of a regular chatbot to an LLM is a significant leap in the field of natural language processing and artificial intelligence. A regular chatbot is a software program that can interact with users using predefined rules or scripts, whereas an LLM is a machine learning model that can generate natural language responses based on training with large amounts of data [8].

Popular LLMs include ChatGPT, BARD, and the LLaMa family of models. In addition to large companies like OpenAI and Google, smaller research groups and individuals can also train LLMs [9], 2023). This has led to a huge and recent increase in the number of LLMs. Typically, they are published on a platform called Hugging Face. As of July 2023, 15'821 LLMs were available [10]. This is an example of the widespread use and popularity of these tools.

Although the roots of natural language processing can be traced back to the 1950s [11], we have observed a breakthrough in capability and popularity in recent years.

This can be observed by the quicker release cycles of the most popular LLMs. The first LLM that became widely noticed by the general public was ChatGPT, which was released by OpenAI in November 2022.

With the huge popularity and variety of use cases in which chatbots can be of assistance, it appears that this form of AI has a wide variety of uses. However, there are some limitations to the capabilities of LLMs. Some have already been scientifically tested. For example, Asher et al. [12] examined the limitations of language understanding and found that LLMs have problems capturing important aspects of linguistic meaning. In particular, LLMs cannot learn the notions of semantic entailment or consistency as defined in formal semantics, and they cannot master universal quantification—for example, understanding the difference between linguistic nuances like "every", "some", "many", "most", and so on.

Wolf et al. [13] investigated the alignment problem of LLMs, which is the degree to which a model's behavior matches user intentions and expectations. Their paper proves that there are fundamental limitations to the alignment of LLMs, such as impossibility results, lower bounds, and trade-offs. Impossibility results indicate that there is always a way to make the LLM misaligned by a clever prompt, regardless of how well the LLM is aligned. Lower bounds mean that the more aligned the LLM is, the longer the prompt needs to be to make it misaligned; however, it is still possible. Trade-offs mean that the methods that make the LLM more aligned also make it more likely to be misaligned by a prompt that gives different feedback. It has been proven that the nature and length of a prompt are important to achieve the desired result. They found that all measures to enhance the result quality could be nullified by bad prompting.

In addition, the data extraction capability that we focus on has been previously reviewed. Li et al. [7] and Guo et al. [14] focused on financial data and graphs. They studied how well pretrained language models can analyze financial texts using eight datasets from different tasks and sources. Li et al. [7] also compared the models with fine-tuned and domain-specific methods and provided evidence and insights into their advantages and drawbacks. While obtaining encouraging results for some tasks, the authors also noted that the analyzed models may "fall short on others, particularly when deeper semantics and structural analysis are needed."

The research conducted at this time is still very scarce due to the novelty of LLMs and the ever-changing fast-paced releases of new and improved systems. Therefore, more research on table extraction is still required.

On the other hand, research regarding table identification and extraction from PDF files has a long tradition prior to the advent of LLMs. For instance, Yildiz et al. [15] noted that the extraction of information from tables in a PDF file requires three steps: table detection, table structure recognition, and table functional analysis. One challenge is the correct interpretation of the table because of the tendency toward over-segmentation. Over-segmentation occurs when the extraction method divides the table into too many small regions that do not correspond to meaningful cells or headers. This can make it difficult to identify and classify the table elements or to analyze their relationships. For instance, the algorithm may over-segment the header row into many small segments rather than recognizing the header as a whole. Smock et al. [16] approached this issue by applying a canonicalization procedure that corrects this over-segmentation with an algorithm that handles the data between extraction and analysis. They used this as a part of developing a new model called PubTables-1M, which contains nearly 1 million tables from scientific articles. The PDF documents were processed in sequence of characters using the Needleman–Wunsch algorithm. The proposed method uses dynamic programing to find the optimal global alignment of two sequences based on a scoring system that assigns rewards or penalties for matches, mismatches, and gaps. They showed that with canonicalization the information extraction is improved. They proposed that the application of the suggested methods and canonicalization in additional domains, such as financial ones, is needed.

As mentioned above, there have been several studies conducted to analyze the extraction and aggregation of data from tables and the performance of text recognition; however, there is still a gap in the analysis of the performance of generically trained LLMs in table extraction without aggregating data, especially beforehand. By copying the data and purposely not aggregating them, we test the capabilities exactly in this area to provide new insights.

3. Research Design

For our study, we chose an experiment-based research approach for several reasons. Firstly, as our goal is not to design a novel artifact, design science research is out of focus. Secondly, by gathering quantitative data about the capabilities of LLMs and backing up a generalized statement about LLMs with the analysis of a specific scenario tested with two different LLMs, the experiment–research approach best suits the challenges of our study to address the research questions. The methodology of our study is presented in Figure 1 and is further explained in the following subsections.

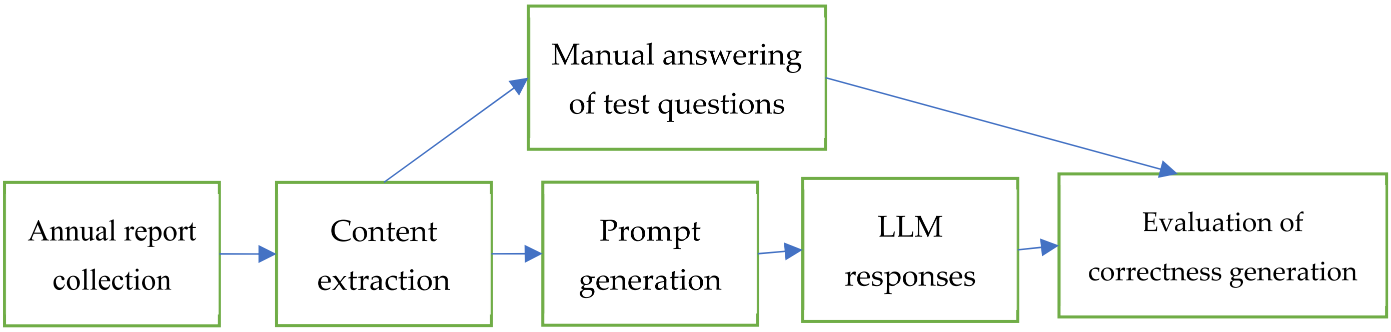

Figure 1. Steps of the conducted research.

3.1. Data Collection

The companies listed on the SIX Swiss Exchange are obliged to publish annual reports due to regulatory requirements. We collected the annual reports of the companies listed in the SMI Expanded index from their websites (see Appendix A (Table A1) for a list of companies with their websites), where the reports are publicly available for download. Let us note that each report is formatted individually, and there are no strict standards regarding content, structure, or the generation procedure of the PDF file. Some companies publish short versions of their annual reports with financial statements. Since we were interested in financial information, we collected the short versions; if available, we focused on actual data, which is why we only collected annual reports from 2022. We assumed that actual annual reports contained tables in text form that could be copied. By the time of collection, 43 reports were available and usable (see Section 4 for further details). Due to limitations in the number of words (500) and the number of characters (4000) in ChatGPT-4, we copied three or four pages from the annual reports, including the tables of the balance sheet and the income statement. The pages chosen were always the page before the table, the page with the table, and the page after the table. If both the income statement and the balance sheet were on the following pages, the pages before and after both tables were included. This provided us with four-to-six pages of input per report based on the distribution of tables in the report. Because the content of these chosen pages was mostly tables with short introductions before and after, in none of the cases was the character limit of either LLM maxed out.

3.2. Test Questions

The following eight test questions were considered for the experiments to be answered by the LLM based on information from the balance sheet or income statement. The questions were motivated by discussions with a company interested in the use case of financial analysis.

- How much are the inventories by the end of 2022?
- How much are the total assets by the end of 2022?
- How much are the current liabilities by the end of 2022?
- How much is total equity by the end of 2022?
- How much was generated in earnings in 2022 before interest and taxes?
- How much gross profit was generated in 2022?
- In what currency is the balance sheet?
- In what currency is the income statement?

3.3. Assessment Criteria

As the landscape of LLMs is increasing, methods to evaluate them have also developed in recent years [17]. As the performance of LLMs increases, the accuracy of evaluations also has to increase to keep up with the development. The most common approaches can be classified as benchmark comparison, evaluation by human assessors, and modeling of human evaluation. The results of these evaluations are often grouped and presented in radar diagrams showing the ability of an LLM in text-specific and dialog-specific abilities, knowledge-specific characteristics, skill-specific abilities, personality and cognitive science features, alignment, reliability and safety-related features, and technical characteristics.

When evaluating LLMs based on regression tasks, common metrics are the mean square error, the root mean square error (RMSE), and the mean absolute error percentage (MAEP). For classification task evaluations, the F1 score, precision, and recall can be used [18]. The F1 score is part of the technical characteristics group, and it evaluates the quality of LLMs by comparing the predictions to the results and calculating the accuracy based on true positives, false positives, true negatives, and false negatives [19]. The F1 score is calculated based on two metrics, the precision and the recall metrics. The precision is calculated by dividing the true positive results by the total number of true positives and false positives. It focuses on calculating how many of the predicted positive events are actually positive and provides a percentage of all correctly predicted positive results.

For the evaluation, the questions' results were summarized by calculating the precision. While there are several possible approaches, such as the F1 score metric, MAEP, or RMSE, we chose the precision metric for further evaluation in our study for the following reasons. Our prediction task is that GPT 4 and BARD should read data out of tables in order to achieve a positive prediction. There is no second class to be predicted. Therefore, the prediction can only be a true positive or false positive since we do not have a negative prediction. This limitation affects the results of the F1 score metric, MAEP, and RMSE, thus decreasing the significance of these metrics, which all require false negative predictions for their calculations. We found that the precision metric is the only one that fits all these limitations.

In the evaluation of answers, there is a possibility that the LLM reads the correct number from the table but presents it incorrectly. These errors can be errors in magnitude, currency, rounding, or usage of commas. Based on approaches chosen by other papers evaluating the correctness of language, the authors decided to further classify the results into correct, semi-correct, and incorrect answers [20]. Semi-correct answers might be cases where the digits are correct but the magnitude (e.g., thousands or millions) is incorrect, where some rounding of numbers occurs, or where the unit of measurement (i.e., the currency) is wrong. This approach was chosen to identify potential error sources, which should be investigated in further research.

3.4. Experiment Execution

In the first step, the 2022 balance sheet of all 50 companies included in the SMI Expanded were downloaded as PDF files from the website of each company. For each company, a new chat was opened within GPT-4 and BARD, and the balance sheet was extracted from the PDF file using the copy–paste function of Windows. Next, the pre-defined questions were asked, and the results were recorded and stored for further evaluation. The correct results for each question were gathered and compared to the results of both LLMs, thereby providing an overview of the number of data points that were correctly extracted from the table.

3.5. Evaluation and Analysis

The collected data points from the experiment were analyzed. We identified patterns of what worked and what did not. We assumed that based on the information we were asking, the result should show that certain questions were more likely answered correctly by one or both of the tested LLMs, and others were more likely answered incorrectly.

4. Results

As described in Section 3.1, we used the data from the annual reports of the SMI Expanded companies. From the 50 companies listed in the index, Lindt, Roche, and Schindler were represented by two types of shares and therefore had the same report. Sandoz did not release their statement when conducting the experiments. Therefore, 46 reports were considered for the analysis. Three of these were inaccessible, as they were protected against copying content. These unusable reports were from Givaudan, Kuehne + Nagel, and Zurich Insurance. Thus, we ended up with 43 usable reports, about which we could ask six questions each. With the two tested models, BARD and Chat GPT-4, the experiments generated 688 valid answers.

4.1. Results for BARD

Following the general trend, BARD struggled, especially with recognizing the EBIT (see Table 1). Regarding inventories, assets, liabilities, and equity, there were more semi-correct answers than completely incorrect answers. There was also only one incorrect answer in naming the currency, which indicates its strength in this area.

Table 1. Results for BARD.

BARD	Full Correct	Semi-Correct	Incorrect	% Full Correct	% Semi-Correct	% Incorrect
Total answers	215	48	81	62.50%	14.00%	23.50%
Total answers excluding currency	131	48	79	50.80%	18.60%	30.60%
Inventories	26	7	10	60.50%	16.30%	23.30%
Assets	27	12	4	62.80%	27.90%	9.30%
Liabilities	24	11	8	55.80%	25.60%	18.60%
Equity	26	12	5	60.50%	27.90%	11.60%
EBIT	15	3	25	34.90%	7.00%	58.10%
Gross profit	13	3	27	30.20%	7.00%	62.80%
Currency Balance sheet	42	0	1	97.70%	0.00%	2.30%
Currency Income statement	42	0	1	97.70%	0.00%	2.30%

4.2. Results for ChatGPT-4

In the results for ChatGPT-4 (see Table 2), almost no semi-correct answers were found, and an almost perfect score in indicating the inventories and a perfect score in specifying the currency was achieved. Like BARD, ChatGPT-4 also struggled to indicate EBIT and gross profit with a precision of only 44.2% and 62.8% of correct answers.

Table 2. Results for ChatGPT-4.

ChatGPT	Full Correct	Semi-Correct	Incorrect	% Full Correct	% Semi-Correct	% Incorrect
Total answers	286	9	49	83.10%	2.60%	14.20%
Total answers excluding currency	200	9	49	77.50%	3.50%	19.00%
Inventories	41	2	0	95.30%	4.70%	0.00%
Assets	39	2	2	90.70%	4.70%	4.70%
Liabilities	36	2	5	83.70%	4.70%	11.60%
Equity	38	1	4	88.40%	2.30%	9.30%
EBIT	19	1	23	44.20%	2.30%	53.50%
Gross profit	27	1	15	62.80%	2.30%	34.90%
Currency Balance sheet	43	0	0	100.00%	0.00%	0.00%
Currency Income statement	43	0	0	100.00%	0.00%	0.00%

4.3. Comparison of Results

When examining the percentages of the combined results (see Table 3), it is obvious that both LLMs had varying success based on the question asked. While the score for guessing currencies was nearly perfect with a precision of 98.8%, they recognized EBIT in only 39.5% of the cases. This question also had the most incorrect answers, with only 4.7% being semi-correct.

Table 3. Combined statistics for BARD and ChatGPT-4.

Combined Statistics	Full Correct	Semi-Correct	Incorrect	% Full Correct	% Semi-Correct	% Incorrect
Total answers	501	57	130	72.80%	8.30%	18.90%
Total answers excluding currency	331	57	128	64.10%	11.00%	24.80%
Inventories	67	9	10	77.90%	10.50%	11.60%
Assets	66	14	6	76.70%	16.30%	7.00%
Liabilities	60	13	13	69.80%	15.10%	15.10%
Equity	64	13	9	74.40%	15.10%	10.50%
EBIT	34	4	48	39.50%	4.70%	55.80%
Gross profit	40	4	42	46.50%	4.70%	48.80%
Currency Balance sheet	85	0	1	98.80%	0.00%	1.20%
Currency Income statement	85	0	1	98.80%	0.00%	1.20%

Comparing all the correct answers across the two LLMs and the overall score, it is clear that ChatGPT-4 outperformed BARD in all areas. Again, the struggle for both LLMs becomes obvious when recognizing EBIT and gross profit. ChatGPT-4 performed between 27.9% and 34.9% better than BARD depending on the question, excluding the question about EBIT, where the difference was only 9.3%. On average, Chat GPT-4 outperformed BARD by 26.7%. This means that it obtained 71 more correct answers to the 344 questions.

4.4. Comparison of BARD and ChatGPT-4 for the Inventory Qestion

If we take a closer look at the inventory question, we see that ChatGPT-4 struggled only two times with the correct magnitude interpretation of the table. BARD demonstrated two types of challenges. In seven cases, the magnitude could not be correctly interpreted. For this topic, a deeper analysis is provided in Section 4.5. In 10 cases, the amount was not listed in the table, and BARD replied with a seemingly random but always incorrect number. Only in four of the fourteen cases did BARD give the answer that it could not find the numbers in the provided data.

4.5. Analysis of Semi-Correct Answers

All 57 semi-correct answers occurred because the LLMs could not answer with the correct magnitude. In 55 of the 57 semi-correct answers provided by the two LLMs, BARD and ChatGPT-4, the returned number was a thousand times too small. Only in two cases was it a hundred thousand times too small. ChatGPT-4 struggled less, with only 3.5% semi-correct answers compared to 18.6% semi-correct answers for BARD.

It is interesting to see that if an LLM had an issue with the magnitude, it mostly interpreted the entire table with the wrong magnitude, resulting in answering all the numerical questions semi-correctly or incorrectly. In total, 13 companies were affected by this phenomenon, but ChatGPT-4 only struggled with 2 of them. Table 4 presents the companies where BARD and ChatGPT-4 could not answer with the correct magnitude. The commonality that 11 of the 13 companies shared was that the table was in the thousands. There were only two companies that also had a magnitude description of thousands, but neither of the LLMs struggled with the answers. All other companies displayed numbers in millions or full numbers.

Table 4. Currency and magnitude descriptions of companies with wrong magnitude answers.

Company	Currency and Magnitude Description	Bard	ChatGPT-4
BALOISE N	in thousands of CHF	x	
BARRY CALLEBAUT N	in thousands of CHF	x	
GALENICA N	in thousand CHF	x	
JULIUS BAER N	CHF m	x	
LINDT N	CHF million	x	
LOGITECH N	(In thousands, except per share amounts)	x	x
MEYER BURGER N	(in CHF 1000)	x	
PSP N	in TCHF	x	
STRAUMANN N	(in CHF 1000)	x	
SWISS PRIME SITE N	in CHF 1000	x	
TECAN GROUP AG N	CHF 1000	x	x
TEMENOS N	USD 1000	x	
VAT GROUP N	In CHF thousand	x	

4.6. Types of Questions and Aggregation

The six questions provided to the language models can be typified into three groups (see Table 5). One group of its own is the two currency questions. Here, the expected answer is no number and can be clearly distinguished from other questions where a number or no answer is expected. The six questions could be further distinguished into two groups. One group of questions is answered by the balance sheet, and the other group is answered by the income statement. Why are the questions different from those in the balance sheet and income statement? One could argue from a technical perspective that it is just another table with other values. However, there are a few key differences in the standardization of terms and numbers used. For example, the balance sheet provides an overview of the companies' assets and liabilities. It is more standardized between industries than income statements. In income statements, rules and terms are not as strongly standardized. The terms vary across industries; for example, gross profit is a term used by industrial or trading companies and does not apply to an income statement of a company in the financial industry.

If we apply this typification to the questions, we can reflect on the answer quality (see Figure 2). In the group in which a text was expected (currency), the two LLMs reached 98.8% of correct answers. For the balance sheet, in which the questions indicate more common terms, the LLMs combined reached 74.7% correct answers. The income statement-related questions received at least 43% correct answers.

Table 5. Question types.

Type	Description	Question
1	Balance Sheet	How much are the Inventories by the end of 2022?
1	Balance Sheet	How much are the total assets by the end of 2022?
1	Balance Sheet	How much are the current liabilities by the end of 2022?
1	Balance Sheet	How much is total equity by the end of 2022?
2	Income Statement	How much earnings before earnings before interest and taxes were generated in 2022?
2	Income Statement	How much gross profit was generated in 2022?
3	Currency	In what Currency is the Balance Sheet?
3	Currency	In what Currency is the Income Statement?

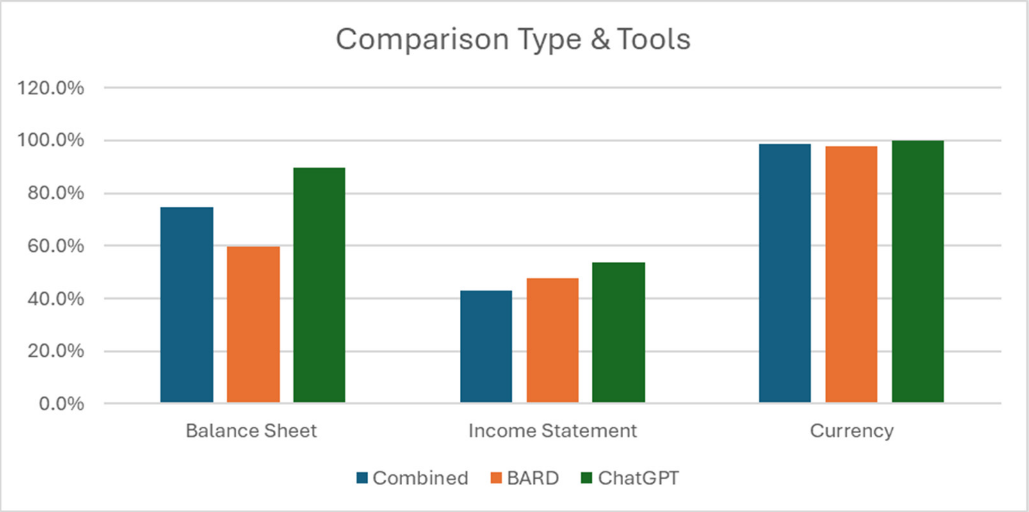

Figure 2. Comparison of question types and tools.

4.7. No Expected Answer Value Leads to Numerical Hallucination

If we take a closer look at the type of expected answers in the balance sheet and income statement groups, we see that the income questions had a much higher share of questions where no value was expected (Table 6).

Table 6. Share of questions where no value was the expected answer.

	# of Questions	# of N/A	Portion of N/A
Balance Sheet	344	40	11.63%
Income Statement	172	68	39.53

The results demonstrate that ChatGPT-4 is more advanced in recognizing when a piece of specific information is not in a table (see Figure 3). In the group of balance sheet questions, ChatGPT-4 provided 90% correct answers, which is better than the 70% accuracy achieved by BARD. In the group of income statement questions, ChatGPT-4 performed only 52.94% correctly; however, it still performed 44.12% better than BARD.

The main reason for the performance gap is that BARD tends to exhibit more numerical hallucinations. For example, if a number has not been labeled correctly in the report or not presented at all, the tools calculate the numbers themselves. This phenomenon was described by Wen Chen and Jung Hsu [21] as numerical hallucinations. They are a byproduct of the design of these language-based frameworks, as there is a gap between linguistic representation and mathematical logic. As an LLM like BARD or ChatGPT becomes trained, it encounters many similar numbers. Because the answer is based on probability, the model produces an answer that might have a fairly high probability because it was often found in the material used for training. However, it is unrelated to the specific value required in this case and is therefore incorrect.

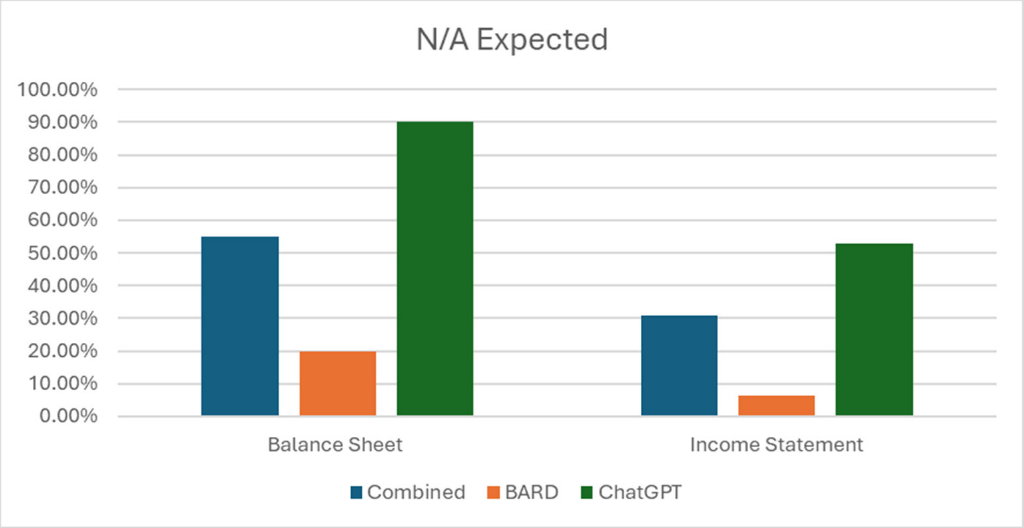

Figure 3. Performance comparison of questions where no value was expected.

To avoid such numerical hallucinations, it is suggested to integrate an external numerical calculator for answering questions that require further calculations. The two tested models do not possess such integrations. Thus, for a use case that relies on such calculations, it would make sense to equip their custom LLM with such.

5. Discussion

Considering the results of our evaluation, it can be seen that ChatGPT performed better in correctly reading numbers out of financial tables than BARD. This difference becomes even more obvious when considering questions that were not addressed in the report. Overall, it can be said that both ChatGPT and BARD could gather data from financial tables when copy–pasted from a PDF file. However, both models had a combined error rate of almost 25% for numerical questions, which is a significant error rate.

When the details of the report varied across industries, the error rate increased substantially, as seen in the income statement questions. In addition, both models had issues when the reports were not from January 1 to December 31 of a calendar year, which also led to incorrect results. It became clear that BARD had more problems with non-existing information, which led to numerical hallucinations and a higher number of errors than ChatGPT.

On the other hand, both LLMs had almost no issues in identifying the currency of a report. ChatGPT achieved an accuracy of 89.5% when it came to balance sheet questions, which is remarkable. When the tables were copy–pasted from a PDF file, both models were able to identify the table and read parts of the table correctly. Therefore, we can confirm our thesis statement that LLMs can extract the data of balance sheets and income statements from large Swiss companies from PDF files with promising accuracy when copy–paste is used to input the data, although some limitations may lead to errors.

The overall accuracy was 81.1% or 75.1% when the answers to the questions about currency were excluded. The accuracy decreased to 72.8%, or 64.1% when magnitude errors were also included as incorrect answers. Possible sources of errors include lack of standardization across industries and financial reports, magnitude errors, and missing values, which lead to incorrect calculations or numerical hallucinations. The number of issues when copying tables from PDF reports appears manageable when LLMs attempt to gather data from them.

6. Conclusions

Our results indicate the potential of LLMs to extract quantitative data, as presented in tables from financial documents. However, the accuracy is mostly not sufficient for use cases that require high reliability of the responses. The experiments yielded rather diverse results for the considered LLMs, for different types of questions, and for different annual reports. This indicates that further advancements in LLMs, improved queries or prompt engineering, and possibly better formatted input documents could lead to better results. We assume that a significant improvement in accuracy may be possible by combining LLMs with other table extraction techniques, which should be developed in future research.

The conducted experiments have some limitations, which are the reduced prompt size, which limited the possible input from four to six pages, the fact that the structure of tables was not considered, the financial numbers asked, which do not exist like this in every industry and report, and the prompting of the EBTI question. Furthermore, the research only focused on large Swiss companies, resulting in the reports being available in English.

Based on the results of this evaluation, we identified that further research is required to evaluate the financial statements of other countries and regions. In addition, the implementation of a mathematical module to increase the accuracy provides a basis for further research. The difference in the accuracy of results between income-statement-related questions and balance-sheet-related questions can also be elaborated. A similar experiment could also be conducted with other specialized models, and a focus could be laid on methods to improve the results, for example, self-changing approaches for the prompt. Lastly, further research is needed on the accuracy of the results if the entire PDF file is used as a prompt and how LLMs perform on other data formats like Excel.

Author Contributions: Methodology, investigation, writing—review and editing: D.B., H.-R.D., and S.E.-H.; writing—original draft preparation, T.H.; supervision, writing—review and editing, T.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset is available on request from the authors.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. Listed companies in the SMI Expanded accessed by 11 November 2023.

Title	Reports on Website
ABB LTD N	https://global.abb/
ADECCO N	https://www.adeccogroup.com/
ALCON N	https://www.alcon.com/
ams-OSRAM	https://www.osram.com/cb/
AVOLTA N	https://www.avoltaworld.com/en
BALOISE N	https://www.baloise.com/
BARRY CALLEBAUT N	https://www.barry-callebaut.com/
BELIMO N	https://www.belimo.com/
BKW N	https://www.bkw.ch/
CLARIANT N	https://www.clariant.com/
EMS-CHEMIE N	https://www.ems-group.com/de/
FLUGHAFEN ZUERICH N	https://www.flughafen-zuerich.ch/
GALENICA N	https://www.galenica.com/
GEBERIT N	https://www.geberit.com/

Table A1. *Cont.*

Title	Reports on Website
GEORG FISCHER N	https://www.georgfischer.com/
GIVAUDAN N	https://www.givaudan.com/
HELVETIA HOLDING N	https://www.helvetia.com/
HOLCIM N	https://www.holcim.com/
JULIUS BAER N	https://www.juliusbaer.com/
KUEHNE+NAGEL INT N	https://home.kuehne-nagel.com/
LINDT N	https://www.lindt-spruengli.com/
LINDT PS	https://www.lindt-spruengli.com/
LOGITECH N	https://www.logitech.com/
LONZA N	https://www.lonza.com/
MEYER BURGER N	https://www.meyerburger.com/
NESTLE N	https://www.nestle.com/
NOVARTIS N	https://www.novartis.com/
PARTNERS GROUP N	https://www.partnersgroup.com/en/
PSP N	https://www.psp.info/
RICHEMONT N	https://www.richemont.com/
ROCHE GS	https://www.roche.com/
ROCHE I	https://www.roche.com/
SANDOZ GROUP N	https://www.sandoz.com/
SCHINDLER N	https://group.schindler.com/en.html
SCHINDLER PS	https://group.schindler.com/en.html
SGS N	https://www.sgs.com/en
SIG Group N	https://www.sig.biz/
SIKA N	https://www.sika.com/
SONOVA N	https://www.sonova.com/
STRAUMANN N	https://www.straumann.com/
SWATCH GROUP I	https://www.swatchgroup.com/
SWISS LIFE HOLDING AG N	https://www.swisslife.com/
SWISS PRIME SITE N	https://sps.swiss/
SWISS RE N	https://www.swissre.com/
SWISSCOM N	https://www.swisscom.ch/
TECAN GROUP AG N	https://www.tecan.com/
TEMENOS N	https://www.temenos.com/
UBS GROUP N	https://www.ubs.com/
VAT GROUP N	https://www.vatvalve.com/
ZURICH INSURANCE N	https://www.zurich.com/

References

1. Tamkin, A.; Brundage, M.; Clark, J.; Ganguli, D. Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models. *arXiv* **2021**, arXiv:2102.02503.
2. Sarawagi, S. Information Extraction. Now Publishers Inc., 2007. Available online: https://www.nowpublishers.com/article/DownloadSummary/DBS-003 (accessed on 30 October 2023).

3. Krapivin, M.; Autaeu, A.; Marchese, M. Large Dataset for Keyphrase Extraction. Technical Report #DISI-09-055, University of Trento. 2008. Available online: http://eprints.biblio.unitn.it/1671/1/disi09055-krapivin-autayeu-marchese.pdf (accessed on 30 October 2023).
4. Naveed, H.; Khan, A.U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Akhtar, N.; Barnes, N.; Mian, A. A Comprehensive Overview of Large Language Models. *arXiv* **2023**, arXiv:2307.06435.
5. Huang, J.; Yang, D.M.; Chi, Z.; Rong, R.; Wang, S.; Nezafati, K.; Xiao, G.; Peterson, E.D.; Zhan, X.; Xie, Y. A Critical Assessment of Using ChatGPT for Extracting Structured Data from Clinical Notes (SSRN Scholarly Paper 4488945). *npj Digit. Med.* **2024**, *7*, 106. [CrossRef] [PubMed]
6. Yin, D.; Dong, L.; Cheng, H.; Liu, X.; Chang, K.-W.; Wei, F.; Gao, J. A Survey of Knowledge-Intensive NLP with Pre-Trained Language Models. *arXiv* **2022**, arXiv:2202.08772.
7. Li, X.; Chan, S.; Zhu, X.; Pei, Y.; Ma, Z.; Liu, X.; Shah, S. Are ChatGPT and GPT-4 General-Purpose Solvers for Financial Text Analytics? A Study on Several Typical Tasks. *arXiv* **2023**, arXiv:2305.05862.
8. Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. A Survey of Large Language Models. *arXiv* **2023**, arXiv:2303.18223.
9. Gao, A.K. Prompt Engineering for Large Language Models. SSRN Report 4504303. 2023. Available online: https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID4504303_code6002566.pdf?abstractid=4504303&mirid=1 (accessed on 30 October 2023).
10. Gao, S.R.; Gao, A.K. On the Origin of LLMs: An Evolutionary Tree and Graph for 15,821 Large Language Models. *arXiv* **2023**, arXiv:2307.09793.
11. Casey, M. Large Language Models: Their History, Capabilities and Limitations. Snorkel AI. 2023. Available online: https://snorkel.ai/large-language-models-llms/ (accessed on 30 October 2023).
12. Asher, N.; Bhar, S.; Chaturvedi, A.; Hunter, J.; Paul, S. Limits for Learning with Language Models. *arXiv* **2023**, arXiv:2306.12213.
13. Wolf, Y.; Wies, N.; Avnery, O.; Levine, Y.; Shashua, A. Fundamental Limitations of Alignment in Large Language Models. *arXiv* **2023**, arXiv:2304.11082.
14. Guo, J.; Du, L.; Liu, H.; Zhou, M.; He, X.; Han, S. GPT4Graph: Can Large Language Models Understand Graph Structured Data? An Empirical Evaluation and Benchmarking. *arXiv* **2023**, arXiv:2305.15066. [CrossRef]
15. Yildiz, B.; Kaiser, K.; Miksch, S. pdf2table: A Method to Extract Table Information from PDF Files. *IICAI* **2005**, *2005*, 1773–1785. Available online: https://www.researchgate.net/publication/220887997_pdf2table_A_Method_to_Extract_Table_Information_from_PDF_Files (accessed on 30 October 2023).
16. Smock, B.; Pesala, R.; Abraham, R. PubTables-1M: Towards comprehensive table extraction from unstructured documents. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 4624–4632. [CrossRef]
17. Tikhonov, A.; Yamshchikov, I.P. Post Turing: Mapping the landscape of LLM Evaluation. *arXiv* **2023**, arXiv:2311.02049.
18. Roberts, A. Precision: Understanding This Foundational Performance Metric. Arize AI. 2022. Available online: https://arize.com/blog-course/precision-ml/ (accessed on 30 October 2023).
19. Sharma, N. Understanding and Applying F1 Score: A Deep Dive with Hands-On Coding. Arize AI. 2023. Available online: https://arize.com/blog-course/f1-score/ (accessed on 30 October 2023).
20. Bawden, R.; Sennrich, R.; Birch, A.; Haddow, B. Evaluating Discourse Phenomena in Neural Machine Translation. *arXiv* **2018**, arXiv:1711.00513.
21. Chen, S.W.; Hsu, H.J. MisCaltral: Reducing Numeric Hallucinations of Mistral with Precision Numeric Calculation. *Res. Sq.* **2023**. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.