Advances in Artificial Intelligence and Machine Learning; Research 5 (1) 3314-3343 Received 26-11-2025; Accepted 5-03-2025; Published 12-03-2025

Evaluation of Large Language Models for Extracting Financial Data from Tables in Annual Reports

Michael Andreas Buholzer michael@buholzer.net
School of Business,
University of Applied Sciences and Arts Northwestern Switzerland
Olten, Switzerland

Frederico Fischer fredae14@hotmail.com
School of Business,
University of Applied Sciences and Arts Northwestern Switzerland
Olten, Switzerland

Geremia Simonella
School of Business,
University of Applied Sciences and Arts Northwestern Switzerland
Olten, Switzerland

Thomas Hanne thomas.hanne@fhnw.ch
Institute for Information Systems,
University of Applied Sciences and Arts Northwestern Switzerland
Olten, Switzerland

Corresponding Author: Thomas Hanne

Abstract

This study evaluates the performance of extracting data from tables using three large language models (LLMs), namely ChatGPT 4, Custom GPT based on ChatGPT 4, and ChatPDF, in extracting and interpreting quantitative data from tables in financial reports. The models were tested on six questions regarding financial data with varying levels of difficulty using three financial reports from different industries and provided in different formats. The results are compared in terms of accuracy, precision, error rates, and qualitative analysis of the output quality. The results indicate that LLMs have a very limited ability to correctly read and interpret data from tables using annual reports. The study also showed that the same reports including the text yielded better results than the tables alone. The results also indicated that a more specific query can lead to slightly better results. However, the study shows that the current LLM technology is still unsuitable for practical applications in similar use cases related to table extraction, in particular where a high reliability of results is required. Thus, the study suggests that future research should focus on improving the capabilities of LLMs in financial data analysis, including the development of more advanced techniques for data extraction and interpretation.

Citation: Michael Andreas Buholzer, et al. Evaluation of Large Language Models for Extracting Financial Data from Tables in Annual Reports. Advances in Artificial Intelligence and Machine Learning. 2025;5(1):190.

Keywords: Large Language Models, LLMs, Table extraction, Table content, Information extraction, Financial analysis, Annual reports, Evaluation.

1. INTRODUCTION

ChatGPT, Bard, and other Natural Language Processing (NLP) chat applications have been the talk of the town since September 2022. The capabilities and results that they provide are currently in the spotlight. In various fields, they achieve excellent qualitative results, but issues like dissemination of false information and misuse are also at the forefront. Our study explores their ability to extract and analyze tabular information, which is assumed to be rather limited in current models. The focus of this research is specifically on the analysis of these functions in the context of financial documents, particularly the annual reports of companies. These documents are examined using various tools and are subject to qualitative and quantitative evaluation. The efficiency and accuracy of large, pre-trained transformer models for interpreting data from tables are to be investigated. No novel extraction solution is intended; rather, emphasis is placed on a thorough analysis and assessment of the current capabilities of these technologies in a specific and practically relevant application area.

1.1 Problem Statement

According to Teubner et al. (2023) [1], development in the chatbot area has been accelerating rapidly since the publication of ChatGPT in November 2022 and has helped to make ChatGPT the fastest-growing consumer technology in history. Pre-trained large language models, such as GPT-3/4 or Bard, have received remarkable results in various application purposes already before 2022 [2]. Although large language models deliver excellent results in areas such as translation or text summarization [3], in other fields, the performance is not well understood or appears questionable. One of these areas is the treatment of quantitative information included in tables such as from annual reports. This exploration is essential to inform and guide both the AI and financial communities about the capabilities, potential improvements, and realistic expectations of applying these sophisticated AI tools to complex, real-world financial tasks. Such investigations enrich academic and professional discourse and serve as a catalyst for future innovations and advances in AI-driven financial analysis.

1.2 Thesis Statement and Research Questions

Large pre-trained transformer models like GPT-3/4 and Bard, despite their proficiency in natural language processing, exhibit distinctive strengths and limitations in extracting and interpreting quantitative data from tables, such as those found in balance sheets or annual company reports. This study analyzes these aspects systematically, offering insights into the performance of models in handling financial data included in pdf files, their accuracy, and their potential application in automated financial analysis.

The following Research Questions (RQs) are considered and correspond to different phases of the research project:

Preparation Phase:

RQ 1: How should data (balance sheets) be prepared to obtain more accurate results when extracting information from tables using an LLM?

RQ 2: What are the best ways to ask natural language queries to the transformer models to elicit accurate and relevant answers from the table data?

Testing Phase:

RQ 3: How does the quality of the results differ in terms of data extraction and possible use/application when examining different LLM tools?

RQ 4: What are the specific strengths and limitations of using large pre-trained transformer models when extracting quantitative table data from tables of annual company reports?

RQ 5: How can different models and tools be compared and evaluated based on their performance in extracting and interpreting table data?

Our paper is structured as follows: In Section 2, we provide a literature review. The research methodology for our study is outlined in Section 3. Results are presented in Section 4. The paper ends with a discussion and conclusions.

2. LITERATURE REVIEW

This literature review explores the evolution and current state of table extraction using Large Language Models (LLMs). It covers the transition from heuristic and machine learning-based methods to advanced neural network approaches, highlighting the recent focus on transformer-based models like Table-GPT, which represent significant advancements in enabling LLMs to interact with and understand tabular data.

For the literature review, different databases, such as Google Scholar, You.com, and JSTOR. In addition, LLMs like ChatGPT and Copilot (Bing) were used to search for specific references. Search engines like Google and Bing were also used to locate specific publications. The following research terms were used for searching: “large transformer models”, “extracting table data and conversion”, “neural network”, “performance extract data tables”, “pdf table extraction”, “analyze balance sheets in pdfs”, “table with LLM”, “table to text generator” and “table to text datasets”. This list is not exhaustive; various combinations of these keywords were also used to expand the search for relevant literature.

2.1 NLP and LLMs

The purpose of a neural language model is to learn the function of words as they appear in sentences, as described by Bengio et al. (2000) [4]. According to Hudson and Cohen (1999) [5], early applications focused on pattern recognition, which was first tested during the 1960s. As described

by Kang et al. (2020) [6], neural language models have become part of Artificial Intelligence (AI) in addition to traditional symbol-based models, such as expert systems. Although AI models were already in use, such as for controls in large manufacturing industries [7], or in financial services [8], it was not until the development of new technologies at the beginning of the 21st century that AI was available for the broader public to use [9, 10]. Natural language processing, also known as NLP, is a part of artificial intelligence and, according to Kang et al. (2020) [6], is designed to enable communication between machine and human language in natural language. Despite significant advancements, this technology was critically viewed in some countries, and in some instances, according to Kasneci et al. (2023) [11], banned from usage. As described by Lauriola et al. (2022) [12], NLP models have experienced a boost in demand due to the technical development of deep learning. As the amount of information available in research has reached a level that is overwhelming for an individual to analyze, Torfi et al. (2020) [13], pointed out how NLP models can help researchers analyze textual information effectively. In practical use, NLP models still require human intervention to extract data from balance sheets and company reports, according to Sage et al. (2021) [14]. Although literature on the extraction of table data is available [15], there is less research on the specific extraction of tables in PDFs.

In the field of linguistics, experts often focus on creating rules to explain how languages work. These rules cover different aspects of language, such as grammar (how sentences are structured), phonology (the study of sounds), morphology (the study of word forms), and the logic behind meanings. This approach, which is particularly noted in the works of linguist Noam Chomsky and many other theoretical linguists, attempts to map the language in a structured way.

However, Large Language Models (LLMs) work in different ways. Instead of using a fixed set of rules, LLMs learn from a vast amount of text data. LLMs look at how words are used in existing texts and learn to predict the likelihood or probability of one word following another. This process involves analyzing a huge number of sentences and texts, which helps the model understand and generate language based on patterns and usage in these texts, not just fixed rules ([16], p. 8).

Since the late 1980s, researchers have attempted to make computers understand and use language by observing patterns in large amounts of text. First, the researchers mostly worked on specific tasks where the model was provided with clear examples to improve, such as translating languages or understanding the structure of sentences. These tasks were easier because they had good datasets, which are collections of examples in which the correct answer is already known, like a sentence and its translation ([16], p. 9).

Around 2015, experts started to focus more on a new approach called self-supervised learning. This method is independent of datasets with correct answers but learns from observing many texts, which is more challenging but also more powerful because there is much more text out there without specific labels or answers ([16], p. 9). In self-supervised learning, tasks are designed such that solving them requires capturing essential features or relationships in the data. The input data are typically augmented or transformed to create pairs of related samples, where one sample serves as the input and the other is used as the supervisory signal [17].

A significant advancement in this field was the introduction of a model called the transformer in 2017 by Vaswani et al. (2017) [18], which focused on the context and position of every word in the text, unlike prior models that processed data sequentially. Initially, the model was intended for translating languages. Soon after, two important transformer models were developed: BERT and

GPT. BERT improved understanding by looking at words that were hidden in a sentence, which helped it to receive context before and after the hidden word. However, BERT had some limitations in terms of generating text. GPT, on the other hand, focused on predicting the next word in a sequence and was found to be simpler and more effective for many tasks. Both models used a two-step approach. First, they were trained on a large number of general texts to understand language (this is called pre-training). Then, they were fine-tuned for specific tasks with more targeted data. This method proved very effective in teaching computers to understand and use human language ([16], p. 10).

2.2 Table Extraction

LLMs are primarily designed for processing and generating text rather than interpreting structured data like tables, especially when embedded in formats like PDF. They are not inherently equipped to interpret the spatial layout of a table in a PDF file. Tables have rows and columns, which create a relational structure. By converting the table into a linear text format, LLMs do not naturally parse and relate data from different cells.

Table recognition aims to extract information from tables, including table detection, table structure recognition, and table content recognition. Table detection is used to locate tables in images. Table structure recognition is used to recognize spatial and ontology structures, and table content recognition is used for text recognition [19].

Kasem et al. (2022) [19], surveyed table detection methods and differentiated them between heuristic-based, machine-learning-based, and deep learning-based methods. Various indicators, such as character alignment and spacing, distance between words and arrangements of lines, local word spacing, line height thresholds, text block arrangement, and line positions, were used in heuristic-based methods. Unsatisfactory results for generic solutions were obtained using these methods. Therefore, machine-learning approaches have gained increasing attention ([19], pp. 26–27).

Kieninger and Dengel (1999, pp. 255–270) [20], used unsupervised learning through word segment clustering, whereas Cesarini et al. (2002) [21], employed a modified XY tree-supervised learning method. Fan et al. (2015) [22], combined both supervised and unsupervised techniques. Wang and Hu (2002, pp. 242–250) [23], used decision trees and SVM classifiers, and Kasar et al. (2013) [24], applied Hidden Markov Models to merge potential table lines. In addition, the Docstrum algorithm proposed by O'Gorman (1993) [25], uses KNN and angle-based techniques to identify text blocks, and F. Shafait and Smith (2010) [26], proposed a versatile table recognition method suitable for various document layouts, which was implemented in an open-source Tesseract OCR engine.

Neural networks are used to help computers understand the layout of documents. These programs were initially used for simple tasks, such as finding tables in a document. However, more advanced programs have begun to solve more difficult tasks, for example, determining the structure of tables and the arrangement of columns ([19], p. 28).

Researchers have tried different methods to improve how well computers recognize tables. Hao et al. (2016) [27], used a Convolutional Neural Network (CNN) to determine if a part of a document was a table. Another researcher, Azka Gilani et al. (2017) [28], developed an improved method that was built on Hao's work to improve the computer understanding of document layouts.

Schreiber et al. (2017) [29], were among the first researchers to use an advanced version of this technique called Faster RCNN. The proposed method allows us to find tables and, in addition, to understand their structure. Following them, other researchers like He et al. (2017) [30], and Arif and Shafait (2018) [31], also applied different methods to help computers better recognize and understand different parts of a document, such as whether a section is a table, or to identify the content of a page more accurately. Reza et al. (2019) [32], used a mix of different technologies, including an approach called Generative Adversarial Network (GAN), to spot tables. Agarwal et al. (2020) [33], used a multistage extension of Mask R-CNN with a dual backbone for detecting tables.

Transformer-based models have attracted increasing attention in recent years. Smock et al. (2022) [34], used the Detection Transformer (DETR) framework, which is a transformer encoder-decoder architecture, on their table dataset. The proposed application was designed to detect tables and understand their structure. Xu et al. (2022) [35], introduced a self-supervised, pre-trained model called Document Image Transformer. The model uses large-scale unlabeled text images for various document analysis tasks, including table detection.

Zha et al. (2023) [36], introduced a framework called Table Generative Pre-trained Transformer (TableGPT). It allows large language models to understand and interact with tabular data using natural language. This means that TableGPT can process and manipulate data organized in rows and columns, such as spreadsheets, databases, and tables. It can also answer questions, perform calculations, or generate summaries based on the table data. It employs global tabular representations to obtain a comprehensive understanding of the entire tables, thereby allowing for complex operations like data manipulation and analysis. Unlike systems that rely on external APIs, TableGPT is self-contained, ensuring data processing and privacy. It supports functionalities such as answering questions, data modification, and report generation, making it user-friendly and accessible for a wide range of data-related tasks. The proposed framework represents a significant leap forward in combining natural language processing with table-based data analysis and manipulation.

Li et al. (2023) [37], introduced a framework called Table-GPT which uses an approach called ”table-tuning” to enhance the ability of large language models like GPT-3 to understand and perform tasks related to relational tables. This approach is important because conventional language models, which are pre-trained predominantly on one-dimensional natural language texts, are suboptimal at handling two-dimensional table data [38]. The concept of table-tuning is inspired by instruction-tuning from the NLP literature, which involves training language models on diverse instruction completion pairs, which describes a prompt or command that is given to the language model and a model’s response or the output that completes the instruction [39]. Table-tuning is a technique that fine-tunes a pre-trained language model on table-related tasks, such as completion, summarization, or question answering. By doing so, the language model can learn to recognize and interpret the structural and semantic features of tables, such as headers, cells, rows, columns, and their relationships. Table-tuning is useful for applications that require working with large amounts of tabular data, such as finance, healthcare, and scientific research [37]. In table-tuning, the focus shifts to using diverse instruction table completion triples, each defining an instance of a table task. This method enhances the ability of language models to understand and work with tables.

The process of synthesizing diverse table tasks involves two key approaches: creating new table tasks for task diversity and synthesizing new table test cases of existing table tasks for data diversity. This methodology exercises the language models’ ability to understand complex two-dimensional table structures. To enhance diversity and prevent overfitting in table-tuning, synthesized table

tasks are further augmented at multiple levels, including instruction-level augmentations. These augmentations involve paraphrasing canonical human-written instructions into different variants using generative models like GPT, and thus adding further diversity to the training data.

2.3 Research Gap

Our literature review shows that there is a gap in the existing literature with respect to a lack of empirical, quantitative evaluations of existing frameworks and models for table extraction, especially in direct comparison with established LLMs such as BART and ChatGPT. While recent frameworks like Table-GPT, have proposed methods to enhance LLMs abilities in handling two-dimensional table data, the practical implementation and performance comparison of these models remain largely unexplored.

To bridge this gap, there is a need for empirical research that conducts quantitative and comparative analyses of various LLMs, including BART, ChatGPT 4, and ChatPDF, in the context of table extraction tasks. This research should aim to do a systematic evaluation of these LLMs across a range of table extraction and interpretation tasks. This involves using standardized datasets to assess each model's ability to extract and interpret table data accurately and efficiently. By addressing this gap, the proposed research contributes to the field by providing a deeper understanding of the capabilities and limitations of current LLMs in the specific context of table extraction. This will guide future advancements in the development of more effective and efficient models for processing tabular data.

3. RESEARCH DESIGN

This section describes the research design that guides the investigation of how well large pre-trained transformer models such as ChatGPT, GPT-3/4, and Bard process and interpret quantitative table data from financial documents. The research design should be suitable for answering the research questions and substantiating the thesis statement.

This research adopts a quantitative strategy, accompanied by a qualitative assessment of the output quality, to evaluate the effectiveness of transformer models in financial data analysis. This mixed-method approach enables an examination of the capabilities of the models, integrating numerical precision with contextual understanding, such as the interpretation of the content in financial reports.

The quantitative phase of the analysis starts with data collection, where a dataset of financial documents is compiled, like annual reports from a diverse range of companies. These documents were selected to represent several industries and complexity levels to test the models' generalizability. The first selection criterion is that the reports are published in English; therefore, only international companies are considered. Second, the reports have the same reporting style. The third selection criterion is the operational field of the selected companies. It should be a logistics company, a company operating in telecommunications or information technology, a pharmaceutical company, or some other production company. To ensure high reporting standards, all companies should be listed on the stock exchange. For the research, three financial annual reports, each representing a distinct type of company and varying in form and detail, were selected for testing using diverse tools

and questions. Although this rather small number limits the generalizability of findings, it appeared as sufficient to gain insight into qualitative strength and weaknesses of the considered LLMs.

Next, the selection and access to pre-trained transformer models that can extract financial data from the collected documents are ensured. These models were benchmarked against each other and against a human baseline to assess relative performance.

Regarding the type of questions to be used for evaluating LLMs as a possible technology to extract financial data from annual reports, we have been in contact with a company in the field of financial consulting which is interesting in developing related technological solutions. Based on their suggestions for suitable test questions addressing data from the balance sheet and the income statement, the following six questions were considered for our tests:

1. What is the revenue?

2. What is the total value of intangible assets?

3. What is the total net income, and is it profit or loss?

4. What are the interest expenditures?

5. What are the personnel costs?

6. What is the return on equity ratio?

All tests were processed with the same selected fixed questions. Three selected LLM models and tools were applied with the same three financial reports in various formats. The number of computational resources used and the time taken to extract data were not explicitly focused on because access to these metrics was not available while performing these tasks.

3.1 Preparation of Reports and Research Strategy

This research employs a strategy that combines elements of both analytical and design-oriented methodologies, with a greater focus on an analytical approach. This strategy is chosen to align with the requirements of our research questions, which focus on the evaluation and understanding of large pre-trained transformer models in the context of financial report analysis.

The core of this research is anchored in an analytical framework. In particular, RQs 1, 3, and 4 focus on understanding the strengths, limitations, and the role of data quality in the performance of transformer models when extracting data from balance sheets. These questions necessitate a detailed analysis of the models' capabilities and limitations in handling quantitative data, reflecting the essence of analytical research. Similarly, RQ 5 adheres to this approach, where a comparative evaluation of different models is undertaken as a task to analytical methodology. This involves a systematic examination of various models to assess their performance metrics, such as accuracy, precision, and error rates, thereby providing a quantifiable and objective analysis of their capabilities.

Although the primary focus is analytical, RQ 2 introduces a design-oriented dimension to the study. This question explores the optimization of query framing to elicit accurate and relevant responses from transformer models, which is a task that inherently involves the design and development of effective interaction methods with the Large Language Models (LLMs). This aspect is crucial because it addresses the practical application of these models and ensures that the interface and interaction mechanisms are conducive to extracting the most precise and relevant information.

This research adopts a predominantly quantitative methodology. This choice is reflective of the nature of the research questions, which are structured to test specific hypotheses about the performance of LLMs in a highly structured and objective manner. The quantitative aspect is evident in the focus on numerical data extraction and the use of statistical tools for performance evaluation. This methodology allows us to derive specific, measurable, and generalizable findings from these empirical investigations.

3.2 Selection of Reports

Public organizations are required to publish their financial statements annually and at a certain standard. This statement led to the decision to select large companies listed on the stock exchange. English financial reports were considered because LLMs are generally better trained in English because of the abundance of English data available for training. As a result, only organizations operating internationally were considered.

The three companies selected for this study are as follows: Roche Group (pharma), Ruag (defense), and Swisscom (telco/it). TABLE 1 provides an overview of the selected reports. The reports with the ID schema “RXA” are the full reports, whereas the reports with the ID schema “RXB” only contain the tables necessary to answer the related questions.

Table 1: Overview of financial reports selected for the testing.

ID	Organization	Pages	Has pictures	Format
R1A	La Roche	191	Only on the first and last pages	A4 Portrait
R1B	La Roche	6	No pictures	A4 Portrait
R2A	Ruag International	57	On 17 pages	A4 Portrait and A4 Landscape
R2B	Ruag International	1	No pictures	A4 Landscape
R3A	Swisscom	186	On 29 pages	A4 Portrait
R3B	Swisscom	20	No pictures	A4 Landscape

The financial reports from those companies are publicly available and typically well documented, ensuring the transparency and accessibility of data for research. They are also heavily regulated, which usually means that their financial reporting is held to high standards. This aspect can provide insights into how well LLMs handle data that must adhere to strict reporting guidelines.

3.3 Selection of LLMs

Currently, various LLMs are publicly available, but the ability to directly upload and analyze PDF documents varies significantly among them. Some do not offer a straightforward method for uploading PDFs, and others are not yet accessible in specific regions, including Switzerland. For instance, due to access limitations at the time of conducting this study, Google Bard could not be used in this research. Another considered option was Bing Chat, which is a Microsoft chatbot integrated into the Edge Browser. However, during preliminary tests, some difficulties were encountered when analyzing PDF documents via Bing Chat. The chatbot's responses often left uncertainty about whether the information was sourced from the open PDF document, the internet, or its own knowledge base. This ambiguity led us to exclude the Bing Chat from the experimental setup. Finally, it was decided to use three LLMs: ChatGPT 4, Custom GPT, and ChatPDF, which are described in the following.

3.3.1 ChatGPT

ChatGPT, an LLM developed by OpenAI, first made its debut in 2022 with the release of version GPT 3.5 which it is was based on. The model has quickly received significant attention due to its advanced natural language processing capabilities. There was no fee for the usage of the LLM; however, even though it can be used for free, there was no possibility to upload PDF files for analysis. Alternatively, the textual content from these files could have been transferred to a text file, which would not be in the scope of the research questions because the goal was to test the quality of the PDF data extraction. Therefore, it was considered not usable for the testing. The subsequent release of ChatGPT version 4 marked a considerable improvement in performance, most notably by introducing the ability to process and analyze PDF documents. This enhancement in functionality makes ChatGPT 4 a suitable tool for this experiment.

3.3.2 Custom GPT

OpenAI has introduced a feature that allows paid customers to create their own versions of GPTs, which are referred to as "Custom GPTs". These models retain the core functionality of ChatGPT 4 but can be tailored to specific applications. This capability was used to develop a custom GPT specialized in interpreting financial reports.

The Custom GPT was configured with a dual focus: maintaining web-browsing capabilities and integrating specialized financial knowledge. To enhance its expertise in financial matters, particularly in the context of Swiss and international accounting standards, the following PDF documents were uploaded as knowledge sources:

- Bewertung der Sachanlagen nach Swiss GAAP FER: eine konzeptionelle Analyse [40]
- Swiss GAAP FER vs. IFRS: A systematic guide to the two main true and fair view accounting standards applied in Switzerland [41]
- IFRS and US GAAP: similarities and differences [42]

- IFRS in your pocket [43]
- FASAB Handbook of Federal Accounting Standards and Other Pronouncements, as Amended [44]

By focusing on the Custom GPT and a specific domain like financial reporting, this research aims to evaluate whether such specialization yields more accurate and relevant results in this field than a general-purpose model. This approach is expected to leverage the model's pre-existing language understanding and reasoning abilities, augmented by the in-depth knowledge from the uploaded documents, to provide insights and analyses tailored to financial reports.

3.3.3 ChatPDF

ChatPDF is an application designed to enhance interactions with PDF documents using AI technology. In particular, ChatPDF leverages a specialized version of OpenAI's GPT language models. This integration allows ChatPDF to understand and interpret the content of PDF documents, transforming them into interactive, conversational interfaces.

Users can ask questions directly to the PDF through ChatPDF's User Interface, and the LLM will respond with answers extracted from the text. This feature is especially useful for large or complex documents, as it allows users to quickly find specific information without manually searching through the entire document. In addition, ChatPDF can summarize lengthy documents, highlight key points, and translate content. The application is designed to be user-friendly, ensuring that even those with minimal technical expertise can benefit from its capabilities.

3.4 Testing Procedure

3.4.1 Structuring of questions and test process

The testing structure was prepared in accordance with the research methodology. To ensure that the capability of the LLMs was tested in a balanced manner, the questions were selected and categorized according to their expected difficulty. Since the selected reports were from 2022, every question relates to 2022. Six questions were defined as shown in TABLE 2. The first question denoted as Q1, was "What is the revenue?". As revenue is the total income generated by a company from its primary operations, revenue is typically clearly stated. Therefore, it was expected that the difficulty for the LLM to search for the figure is easy. Question Q2 was defined for intangible assets. Thus, Q2 is defined as "What is the total of intangible assets?". In general, intangible assets are non-physical assets that offer value to an organization. They can include categories such as goodwill, brands, patents and copyrights, franchises, software and technology, intellectual property, and trademarks. Typically, intangible assets are stated in the consolidated balance sheet. Therefore, the expected difficulty was set to "easy".

The third question "Q3" was about net income. Typically, net income can be stated as a profit or loss (P/L), as it represents the overall profitability of a company. The related question is "What was the total of Net income (P or L)?". As with revenue, net income is generally clearly stated in

Table 2: Overview of test questions with category, difficulty, and further explanation.

ID	Category	Question	Expected difficulty	Clarification
Q1	Revenue	What is the revenue?	Easy – finding the relevant information	Revenue is the total income generated by the company from its primary operations.
Q2	Intangible assets	What is the total of intangible assets?	Easy – finding the relevant information	Intangible assets include categories such as patents, copyrights, and trademarks. In some cases, rights and goodwill are intangible assets.
Q3	Net income	What was the total of Net income (P or L)?	Easy – finding the relevant information	Total net income represents the company's overall profitability.
Q4	Interest expenditures	What are the interest expenditures of the company?	Medium – finding a non-obvious synonym	Interest expenditures are typically costs related to the company debts. However, in the balance sheet of the bank the term used is indeed 'interest expenditures', other synonyms might also be used for corporate balance sheets.
Q5	Personal costs	What are the personal costs?	Medium – finding a non-obvious synonym	Personnel costs are important for estimating a company's efficiency. However, it may not be disclosed directly.
Q6	ROE	What is the ROE?	Difficult – calculating the ratio with correct numbers	Return on Equity (ROE) is a financial ratio that measures a company's profitability in relation to its shareholders' equity. The ROE formula is net income divided by average shareholders' Equity.

financial statements; consequently, the expected difficulty was categorized as easy. The three first questions were all categorized as easy, as it can be expected that the figures are clearly stated, and no calculations were required to search for the correct figure. For question four, Q4, the interest expenditures were selected. The related question is "What are the interest expenditures of the company?". Because interest expenditures can be defined in several ways, the difficulty was set to medium, as it can be expected, that synonyms are included in the financial statements. The second last question, Q5, relates to personal costs. Personal costs are important when estimating an organization's efficiency. Typically, this information is not stated directly in a financial statement. Because the correct category must be found, the difficulty was defined as medium. The final question Q6 addressed the return on equity, commonly known as "ROE". Although ROE is a critical financial ratio, its calculation is not straightforward. To achieve a correct ratio, net income must be divided by average shareholder equity. Because shareholder equity is usually not stated directly in a financial report, the calculation requires several steps. Thus, the question "What is the ROE?" was categorized as difficult.

The testing procedure was conducted in a way to ensure that each query was tested at least once for each report. Each question in TABLE 2 was allocated to a report. As three LLMs were selected for testing, each query was conducted for each tool (defined as T) three times. Therefore, the minimum number of tests for each query and report was nine. Due to the decision to use complete reports and fractions of reports for testing, the number of tests for each organization was 18. Thus, the total number of tests was 108.

To determine that the testing was documented systematically, an Excel file was developed. The file contains six sheets. The sheet "Prompts" contains examples of queries and how they could be provided to the LLM. A second sheet was created for notes and screenshots, if required. The third

sheet is “Reports” which has the reports listed. Following the reports, the sheet “Tools” contains an overview of the three tools that were chosen for testing. For the queries, the sheet “Question” was created that provides an overview of the questions. It also contains the person responsible for testing. The last sheet is “Log”, which contains the testing records. A detailed exposition of the “Log” sheet is provided in the following section.

For the testing, it was considered important to ensure that each test was allocated an identification code. This is defined in column A, where the code is formed from the query, the report ID, the used LLM, and the day and time. For example, an ID can be defined as Q1.R1A.T1.2.14:00

Therefore, the ID is simply the concatenated columns. Column B defines the question or query used for testing. Columns C, D, and E provide information about the day, month, and year when the model was tested, and column G specifies the LLM that was used. In addition, column H specifies the tester, and column K describes the report used. Column M species shows in terms of colors, whether the result was correct or not. The category is defined in column N, where the testing result is qualified. This is a comparison between column O, which contains the expected result, and column P, which shows the result received from the LLM. Column R gives a comment about the result, which each tester had the possibility to define themselves, if just a comment, or the whole answer should be pasted in each cell. The last column S shows the full prompt, which was used for testing.

Guidelines were developed to assess whether a response from an LLM is accurate or not. According to these guidelines, a response is classified as ”correct” if it includes the exact number requested in the query. The format in which this number is presented, such as in millions, billions, or any other numerical representation, is not relevant to the accuracy assessment. A true response is indicated in green in the Excel file. In contrast, a response is considered “semi-correct” if the tester had to help the LLM provide the correct number. This can occur if the LLM does not understand the context or misinterprets the numerical data. In the Excel sheet, such responses are highlighted in orange.

A response is considered “wrong” under two main circumstances. First, if the LLM fails to provide the correct numerical answer that was specifically requested in the prompt, the response is marked as inaccurate. Second, if the LLM is unable to respond to the query, which can occur when the model encounters operational issues like crashing or experiences difficulties in interpreting data from a source, such as a PDF document, the response is also considered inaccurate. A wrong response is indicated in red in the Excel file.

3.4.2 Test execution

The test execution phase was organized and distributed among the team members. Each team member was assigned the responsibility of executing tests for two questions, corresponding to a minimum of 60 individual tests. The tests were conducted over a two-week period using the notebooks of the testers. The LLMs are run on their servers and are served over web applications. Therefore, no special hardware was required. Only a stable internet connection was necessary. It had to be ensured that each team member had access to the respective LLMs. ChatGPT 4 and the Custom GPT were only accessible through a paid subscription, while ChatPDF was free.

Each test was logged in the Log table in Excel. One log entry is shown in TABLE 3: Testing Log as an example. Logfiles were managed in a Teams Group so that each team member had access to.

Table 3: Testing Log

Question	Q1
Day	2
Month	1
Year	2024
Time	14:00
LLM	T1
Tester	Frederico
Prompt	P1
Report	R1A
Result	
Expected Result	66426000000
Received Result	66,426 million
Harmonized Result	66426000000
Comment	Received expected result.
Full Prompt Log	User 2022_La-Roche.pdf PDF What is the total revenue? ChatGPT The total revenue for La-Roche in 2022 was CHF 66,426 million.

4. RESULTS

4.1 Main Findings

Of the 180 tests conducted, 56 were classified as correct, which is only 31.11%. From the 180 tests, only two were classified as "Semi-correct", which counts only to 1.11 %. 122 test results were classified, which is 67.78 % of the 180 tests conducted. As shown in TABLE 4, the first overall overview only gives an estimate of the results based on the tests conducted by the three LLMs.

Table 4: Overview of the 180 test results for all three LLMs

Overall Result		
Category	**Count**	**%**
Correct	56	31.11%
Semi-Correct	2	1.11%
Wrong	122	67.78%
Total	**180**	

TABLE 5 compares the results of the reports; both complete and extracted reports demonstrate high error rates. In total, 58 prompts were classified as incorrect, representing 63.74% of complete reports. In contrast, 70.33% or 64 prompts were classified as incorrect report extractions. Two

prompts were semi-correct, which is 2.20% of the 91 prompts for complete reports, and zero prompts for extractions of reports. The rate of correct results was slightly higher for complete reports, which count to 31 prompts (34.07%). In contrast, is the number of correct prompts for extracted reports is 25, or 27.47%. However, it is not clear whether the results depend on the size of the reports or on the presence of several pictures in the reports.

Table 5: Overview of results drilled down to complete reports and extractions of reports.

Complete report			Extract of report		
Category	**count**	**%**	**Category**	**count**	**%**
Correct	31	34.07%	Correct	25	27.47%
Semi-correct	2	2.20%	Semi-correct	0	0.00%
Wrong	58	63.74%	Wrong	64	70.33%
Total	**91**		**Total**	**89**	

Overall, the results were between 23.33% and 36.67% correct. As shown in TABLE 5, the highest rate of correct results was achieved with the LLM ChatPDF, which was 34.48%. In absolute terms, the custom GPT obtained the highest number of correct results. However, the highest absolute number of incorrect results was achieved using the custom GPT in the experiments with the tables extracted from the reports. The absolute number of incorrect results was 23, which is 76.67%. In summary, the test results were correct by approximately 1/3 and incorrect by 2/3 of the tests.

Table 6: Result distribution compared between LLM and report status.

Complete report			Extract of report		
Category	**Count**	**%**	**Category**	**Count**	**%**
ChatGPT 4			**ChatGPT 4**		
Correct	10	31.25%	Correct	8	27.59%
Semi-correct	1	3.13%	Semi-correct	0	0.00%
Wrong	21	65.63%	Wrong	21	72.41%
ChatPDF			**ChatPDF**		
Correct	10	34.48%	Correct	10	33.33%
Semi-correct	1	3.45%	Semi-correct	0	0.00%
Wrong	18	62.07%	Wrong	20	66.67%
Custom GPT			**Custom GPT**		
Correct	11	36.67%	Correct	7	23.33%
Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	19	63.33%	Wrong	23	76.67%
Total	**91**			**89**	

When the results for complete reports and extractions of reports are compared at the level of each question, the results are more mixed. TABLE 6 shows the distribution of testing results when complete reports were used. The highest proportion of correct results was achieved using the custom GPT, which corresponds to 75 %. In contrast, the highest proportion of incorrect results was found by ChatGPT 4, when the report by Ruag International was tested [45]. The incorrect result was

91.67%, which was the highest among all results for complete reports. In comparison to TABLE 5, the overall results in the class "correct" was less homogeneous.

Table 7: Test results for complete reports and used LLMs.

Complete report								
ChatGPT 4			**ChatPDF**			**Custom GPT**		
La Roche			**La Roche**			**La Roche**		
Correct	4	33.33%	Correct	2	16.67%	Correct	4	36.36%
Semi-correct	1	8.33%	Semi-correct	1	8.33%	Semi-correct	0	0.00%
Wrong	7	58.33%	Wrong	9	75.00%	Wrong	7	63.64%
Ruag International			**Ruag International**			**Ruag International**		
Correct	1	8.33%	Correct	3	33.33%	Correct	1	9.09%
Semi-correct	0	0.00%	Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	11	91.67%	Wrong	6	66.67%	Wrong	10	90.91%
Swisscom			**Swisscom**			**Swisscom**		
Correct	5	62.50%	Correct	5	62.50%	Correct	6	75.00%
Semi-correct	0	0.00%	Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	3	37.50%	Wrong	3	37.50%	Wrong	2	25.00%
Total	**32**		**Total**	**29**		**Total**	**30**	

In comparison to the report extraction, as shown in TABLE 7, the results are sometimes better than those in TABLE 6. For example, the result in the Report "La Roche" was, for ChatGPT 33.33% correct in the whole report, and 44.44% correct for the extracted reports. In this example [46], it is worth noting that the absolute number of correct results was 4 in both cases. The extraction of the "Swisscom" report however [47], performed in all three LLMs worse than in the total report.

Table 8: Average results for each LLM and each report extraction

Extract of report								
ChatGPT 4			**ChatPDF**			**Custom GPT**		
La Roche			**La Roche**			**La Roche**		
Correct	4	44.44%	Correct	4	44.44%	Correct	2	25.00%
Semi-correct	0	0.00%	Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	5	55.56%	Wrong	5	55.56%	Wrong	6	75.00%
Ruag International			**Ruag International**			**Ruag International**		
Correct	4	50.00%	Correct	5	55.56%	Correct	5	55.56%
Semi-correct	0	0.00%	Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	4	50.00%	Wrong	4	44.44%	Wrong	4	44.44%
Swisscom			**Swisscom**			**Swisscom**		
Correct	0	0.00%	Correct	1	8.33%	Correct	0	0.00%
Semi-correct	0	0.00%	Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	12	100.00%	Wrong	11	91.67%	Wrong	13	100.00%
Total	**29**		**Total**	**30**		**Total**	**30**	

A possible reason for this could be that the table in the report extraction was not intended to be analyzed by a machine. Contrary to the "Swisscom" report, the "Ruag International" report seemed to be performed better by all LLMs. The rate of correct result was over 50% in all three instances, whereby it achieved only 33.33% when analyzed by ChatPDF. A possible explanation for the result could be that the Ruag report has a special format and pictures on almost every page. The extraction, however, has no pictures and [45], is only one page long and shorter input documents should make

the information extraction easier . The report extraction of La Roche was overall better, with an overall result of 44.44% in two instances. The custom GPT achieved a slightly worse result, only 25% compared to the result of 36.36% in the analysis of the complete report. Here, a possible solution could be as well, that the extraction was first a consolidated report; hence, the summary was extracted. In addition, there were no pictures in the extraction. The analysis of whether the difficulty of a question influences the performance follows in the next section.

When the three systems are compared over the six questions, as shown in TABLE 8, the results look different. Overall, the custom GPT had a higher correct response rate, and the lowest wrong response rate compared to the other two LLMs. ChatPDF had the highest correct response rate; however, it also had a high incorrect response rate. ChatGPT 4 had the lowest correct response rate, but a similar incorrect answer rate as ChatPDF. As shown in TABLE 2, questions Q1–Q3 were categorized as easy. Even though the answer should be easy to find, the rate of incorrect answers was particularly high in Q1 for ChatPDF, Q3 for the custom GPT, and in Q1 for ChatGPT 4. However, the highest rate of correct answers to Q2, with 83.33% for ChatPDF, and in Q3 for ChatGPT 4 with 75 %. Question 4 was categorized as medium; the overall rate of incorrect results was 100% in all instances. Although Q5 was categorized as medium-difficult as well, the results were slightly better than those of Q4. Q6, which was categorized as difficult, had the highest rate of incorrect answers, which is 100 % in two instances and 50 % in the example of the Custom GPT. As Q1 to Q3 are simple, almost no calculation is required to find the solution. However, Q4 to Q6 require calculations or explanations to find the solution.

In comparison to the analysis of the response rate of complete reports, the report extractions show a different picture. As shown in TABLE 10, the correct answer rate for Q1 in two instances was higher than that in TABLE 9. ChatGPT 4 and the custom GPT both achieved 66.67% correct answers, compared to 25% and 20%, respectively. ChatPDF achieved a similar correct answer rate of 20% in both cases. Interestingly, the performance for Q2 for all three systems was worse than that in TABLE 9. For the extracted reports, ChatGPT 4 had a correct answer rate of 42.86%, which is 23.81% less than that of TABLE 9. ChatPDF also achieved a 16.67% lower correct answer rate than when complete reports were used. The custom GPT had the highest difference of 50.00% less compared to when the complete reports were used. However, for Q3, mixed results were obtained. The performance of ChatGPT 4 was also 25.00% lower than that of complete reports. This is different for ChatPDF, which achieved a 25% higher correct answer rate of 75% compared to 50%, as shown in TABLE 9. The custom GPT achieved a similar result of 50% in both instances. For Q4, all systems performed at 0%; therefore, there were no correct results in any instances. A similar picture is observed for Q5, where the performance was only in one instance compared to when complete reports were used. Q6 also shows a high rate of incorrect answers. The difficulty of the question may explain the high rate of incorrect answers.

In the following, we discuss the result for 3 of the test questions in greater detail. TABLE 11 presents the results of Q1, which asked for revenue. In general, the number can be found easily, as in most cases, it is presented clearly in the financial statements. However, the answer rate was not homogenous but rather mixed. The reason for the high response rate for the La Roche report might be that the overall report was more than 180 pages long. On the other hand, the Ruag International report [46], is in a special format, and it includes pictures on almost every page. The tables are also not placed in a uniform manner throughout the report [45]. In contrast, the correct response rate for

Table 9: Results per question for each LLM for complete reports.

Complete report								
ChatGPT 4			**ChatPDF**			**Custom GPT**		
				Q1				
Correct	1	25.00%	Correct	1	20.00%	Correct	1	20.00%
Semi-correct	1	25.00%	Semi-correct	0	0.00%	Semi-correct	2	40.00%
Wrong	2	50.00%	Wrong	4	80.00%	Wrong	2	40.00%
				Q2				
Correct	4	66.67%	Correct	5	83.33%	Correct	4	66.67%
Semi-correct	0	0.00%	Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	2	33.33%	Wrong	1	16.67%	Wrong	2	33.33%
				Q3				
Correct	3	75.00%	Correct	2	50.00%	Correct	2	50.00%
Semi-correct	0	0.00%	Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	1	25.00%	Wrong	2	50.00%	Wrong	2	50.00%
				Q4				
Correct	0	0.00%	Correct	0	0.00%	Correct	0	0.00%
Semi-correct	0	0.00%	Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	6	100.00%	Wrong	6	100.00%	Wrong	7	100.00%
				Q5				
Correct	2	25.00%	Correct	2	50.00%	Correct	2	50.00%
Semi-correct	0	0.00%	Semi-correct	1	25.00%	Semi-correct	0	0.00%
Wrong	6	75.00%	Wrong	1	25.00%	Wrong	2	50.00%
				Q6				
Correct	0	0.00%	Correct	0	0.00%	Correct	2	50.00%
Semi-correct	0	0.00%	Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	4	100.00%	Wrong	4	100.00%	Wrong	2	50.00%
Total	**32**		**Total**	**29**		**Total**	**30**	

the Swisscom report was 100%. This could imply that the way reports are presented may influence the performance of LLMs [47].

In contrast to the complete report, TABLE 12 presents the response performance of the report extraction. The Swisscom report had a lower correct answer rate than the rate presented in TABLE 11. It appears that the information in the tables of the Swisscom report is now difficult to extract for two of the considered LLMs. The systems performed better with the extraction of the La Roche and Ruag International reports. Again, it may be easier to extract information when the report is short and includes either tables or text, but not both.

(TABLE 13 shows the response rates of the LLMs for Q2 for complete reports.The performance of the LLMs was positive in reports from La Roche and Swisscom, where all three LLMs achieved a 100% correct response rate. In contrast, the LLMs in the example of the Ruag International report, only in the instance of ChatPDF achieved a semi-correct response rate. As discussed previously,

Table 10: Results per question and LLM for extractions of reports.

Extraction of report								
ChatGPT 4			**ChatPDF**			**Custom GPT**		
				Q1				
Correct	2	66.67%	Correct	1	20.00%	Correct	2	66.67%
Semi-correct	0	0.00%	Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	1	33.33%	Wrong	4	80.00%	Wrong	1	33.33%
				Q2				
Correct	3	42.86%	Correct	4	66.67%	Correct	1	16.67%
Semi-correct	0	0.00%	Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	4	57.14%	Wrong	2	33.33%	Wrong	5	83.33%
				Q3				
Correct	2	50.00%	Correct	3	75.00%	Correct	2	50.00%
Semi-correct	0	0.00%	Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	2	50.00%	Wrong	1	25.00%	Wrong	2	50.00%
				Q4				
Correct	0	0.00%	Correct	0	0.00%	Correct	0	0.00%
Semi-correct	0	0.00%	Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	6	100.00%	Wrong	6	100.00%	Wrong	6	100.00%
				Q5				
Correct	1	33.33%	Correct	1	25.00%	Correct	1	25.00%
Semi-correct	0	0.00%	Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	2	66.67%	Wrong	3	75.00%	Wrong	3	75.00%
				Q6				
Correct	0	0.00%	Correct	1	20.00%	Correct	1	14.29%
Semi-correct	0	0.00%	Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	5	100.00%	Wrong	4	80.00%	Wrong	6	85.71%
Total	**28**		**Total**	**30**		**Total**	**30**	

the structure of the report, the use of graphics, and the placement of tables may influence the performance of LLMs in extracting information).

Compared to the complete reports used, the extraction of reports presented in TABLE 14 shows a performance of 100% for ChatPDF for all three reports. In contrast, ChatGPT 4 and the custom GPT system demonstrated lower performance in the reports by Ruag International and Swisscom.

Q6 shows only a correct response in one instance. As mentioned, the question was categorized as difficult because it requires calculations to find the return on equity ratio. The Swisscom report was, compared to the La Roche and Ruag International report, well-structured and easy to interpret. TABLE 15 presents the result of Q6, which contains only one positive response.

Compared to the report extractions (as shown in TABLE 16), the performance of Question Q6 was negative in all but two cases. ChatPDF achieved a positive response rate, and the custom GPT achieved one correct result from two trials. As previously argued, the extraction of the Swisscom

Table 11: Overview of response rates for Q1 when complete reports are used.

Question Q1 - complete reports used								
ChatGPT 4			**ChatPDF**			**Custom GPT**		
La Roche								
Correct	0	0.00%	Correct	0	0.00%	Correct	0	0.00%
Semi-correct	1	100.00%	Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	0	0.00%	Wrong	2	100.00%	Wrong	2	100.00%
Ruag international								
Correct	0	0.00%	Correct	0	0.00%	Correct	0	0.00%
Semi-correct	0	0.00%	Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	2	100.00%	Wrong	2	100.00%	Wrong	2	100.00%
Swisscom								
Correct	1	100.00%	Correct	1	100.00%	Correct	1	100.00%
Semi-correct	0	0.00%	Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	0	0.00%	Wrong	0	0.00%	Wrong	0	0.00%
Total	**4**		**Total**	**5**		**Total**	**5**	

Table 12: Response rates for Q1 when using report extractions.

Question Q1 - report extraction								
ChatGPT 4			**ChatPDF**			**Custom GPT**		
La Roche								
Correct	1	50.00%	Correct	1	100.00%	Correct	1	100.00%
Semi-correct	0	0.00%	Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	1	50.00%	Wrong	0	0.00%	Wrong	0	0.00%
Ruag international								
Correct	1	100.00%	Correct	0	0.00%	Correct	1	100.00%
Semi-correct	0	0.00%	Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	0	0.00%	Wrong	2	100.00%	Wrong	0	0.00%
Swisscom								
Correct	1	100.00%	Correct	0	0.00%	Correct	0	0.00%
Semi-correct	0	0.00%	Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	0	0.00%	Wrong	2	100.00%	Wrong	1	100.00%
Total	**4**		**Total**	**5**		**Total**	**3**	

report is not suitable for interpretation by an LLM. On the other hand, the consolidated report by La Roche is easy to read and can be solved by humans.

As the ROE requires the average shareholder equity, in most cases, the LLMs were unable to identify the shareholder equity. Although this was stated in the report, the systems could not find it. In some cases, a special guide was required to allow the LLMs to generate a response.

Table 13: Response rates for Q2 when using complete reports .

Question Q2 - complete reports used								
ChatGPT 4			**ChatPDF**			**Custom GPT**		
La Roche								
Correct	2	100.00%	Correct	2	100.00%	Correct	2	100.00%
Semi-correct	0	0.00%	Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	0	0.00%	Wrong	0	0.00%	Wrong	0	0.00%
Ruag international								
Correct	0	0.00%	Correct	1	50.00%	Correct	0	0.00%
Semi-correct	0	0.00%	Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	2	100.00%	Wrong	1	50.00%	Wrong	2	100.00%
Swisscom								
Correct	2	100.00%	Correct	2	100.00%	Correct	2	100.00%
Semi-correct	0	0.00%	Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	0	0.00%	Wrong	0	0.00%	Wrong	0	0.00%
Total	**6**		**Total**	**6**		**Total**	**6**	

Table 14: Response rates for Q2 when using extractions of reports.

Question Q2 - report extraction								
ChatGPT 4			**ChatPDF**			**Custom GPT**		
La Roche								
Correct	2	100.00%	Correct	2	100.00%	Correct	0	0.00%
Semi-correct	0	0.00%	Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	0	0.00%	Wrong	0	0.00%	Wrong	2	100.00%
Ruag international								
Correct	1	50.00%	Correct	2	100.00%	Correct	1	50.00%
Semi-correct	0	0.00%	Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	1	50.00%	Wrong	0	0.00%	Wrong	1	50.00%
Swisscom								
Correct	1	25.00%	Correct	2	100.00%	Correct	0	0.00%
Semi-correct	0	0.00%	Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	3	75.00%	Wrong	0	0.00%	Wrong	2	100.00%
Total	**8**		**Total**	**6**		**Total**	**6**	

4.2 Observations and Anomalies

The correct response rate appears to be influenced by the number of images. Overall, the rate of correct answers was not particularly promising. However, as shown in FIGURE 1, the highest numbers of correct answers were obtained by R2B and R3A. As mentioned before, R2B is the extraction of the Ruag International report, which is only one page and has no pictures in it. R3A, the Swisscom report, which has 29 pictures, is generally well-structured and easy to read for a machine. The reports [47], with the lowest overall correct answer rates were R2A and R3B. The

Table 15: Response rates for Q6 when complete reports are used.

Question Q6 - complete reports used								
ChatGPT 4			**ChatPDF**			**Custom GPT**		
La Roche								
Correct	0	0.00%	Correct	0	0.00%	Correct	0	0.00%
Semi-correct	0	0.00%	Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	2	100.00%	Wrong	2	100.00%	Wrong	2	100.00%
Ruag international								
Correct	0	0.00%	Correct	0	0.00%	Correct	1	100.00%
Semi-correct	0	0.00%	Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	1	100.00%	Wrong	1	100.00%	Wrong	0	0.00%
Swisscom								
Correct	0	0.00%	Correct	0	0.00%	Correct	1	100.00%
Semi-correct	0	0.00%	Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	1	100.00%	Wrong	1	100.00%	Wrong	0	0.00%
Total	**4**		**Total**	**4**		**Total**	**4**	

Table 16: Response rates for Q6 when using report extractions.

Question Q6 - report extraction								
ChatGPT 4			**ChatPDF**			**Custom GPT**		
La Roche								
Correct	0	0.00%	Correct	0	0.00%	Correct	0	0.00%
Semi-correct	0	0.00%	Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	2	100.00%	Wrong	2	100.00%	Wrong	1	100.00%
Ruag international								
Correct	0	0.00%	Correct	1	100.00%	Correct	1	50.00%
Semi-correct	0	0.00%	Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	1	100.00%	Wrong	0	0.00%	Wrong	1	50.00%
Swisscom								
Correct	0	0.00%	Correct	0	0.00%	Correct	0	0.00%
Semi-correct	0	0.00%	Semi-correct	0	0.00%	Semi-correct	0	0.00%
Wrong	2	100.00%	Wrong	1	100.00%	Wrong	4	100.00%
Total	**5**		**Total**	**4**		**Total**	**7**	

first is for the Ruag International report, which contains only 57 pages, with overall pictures on 17 pages. Additionally, it is in an unconventional format. R3B is the Swisscom report, which contains only tables. However, the LLMs failed to extract accurate information from these tables. The highest response rate was achieved with the Swisscom report. The second highest response rate was achieved with the extraction of the Ruag international report.

The number of pages appears to have no impact on the correct response rate, as evidenced by RA1 having 191 pages and R3A containing 186 pages. In a direct comparison of the LLMs, the highest

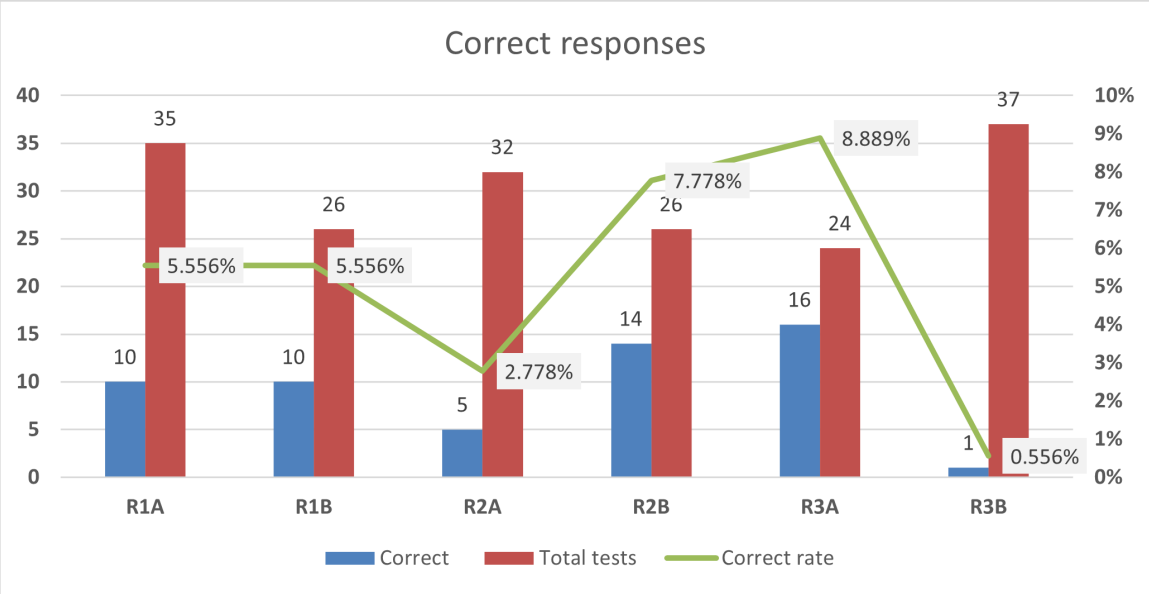

Figure 1: Extraction of correct response rate from 180 conducted tests.

correct response rate was achieved using the custom GPT. However, the lowest correct response rates were also achieved by all three LLMs. FIGURE 2 presents an overview of the absolute numbers of correct responses per report and LLM.

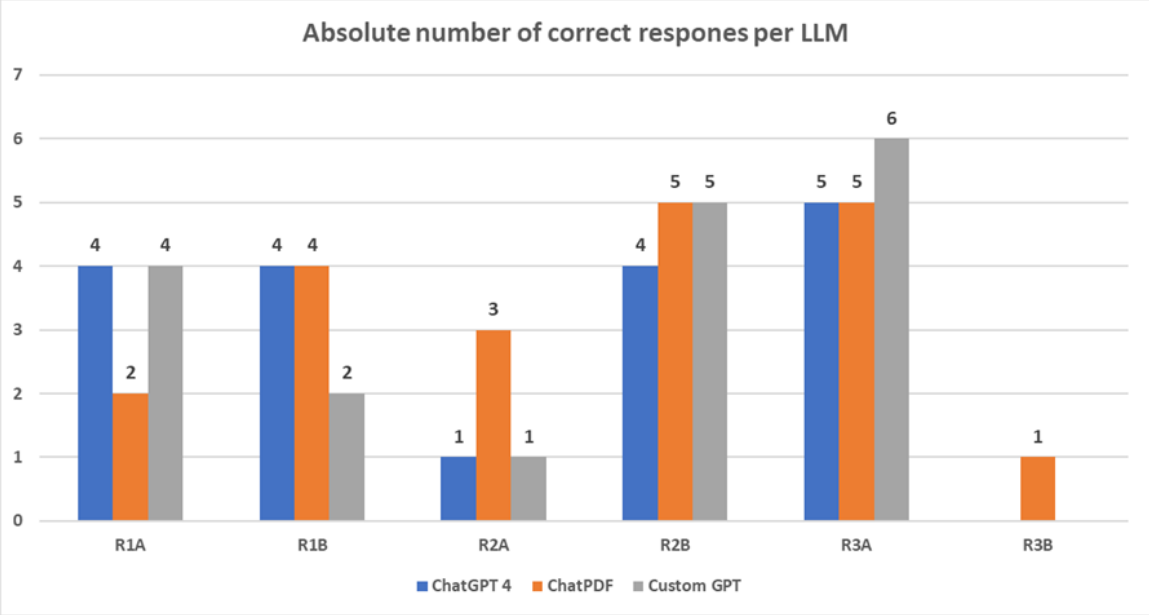

Figure 2: Correct response rates in absolute numbers by LLM.

Overall, there is an indication that the length of the report and the number of pictures may influence the correct response rate. The highest rate of correct responses was achieved by the custom GPT,

with 10% for the Swisscom report, as shown in FIGURE 3. Overall, the results are not very promising and are mixed. There were two instances in which 0% correct results were achieved with ChatGPT 4 and the custom GPT.

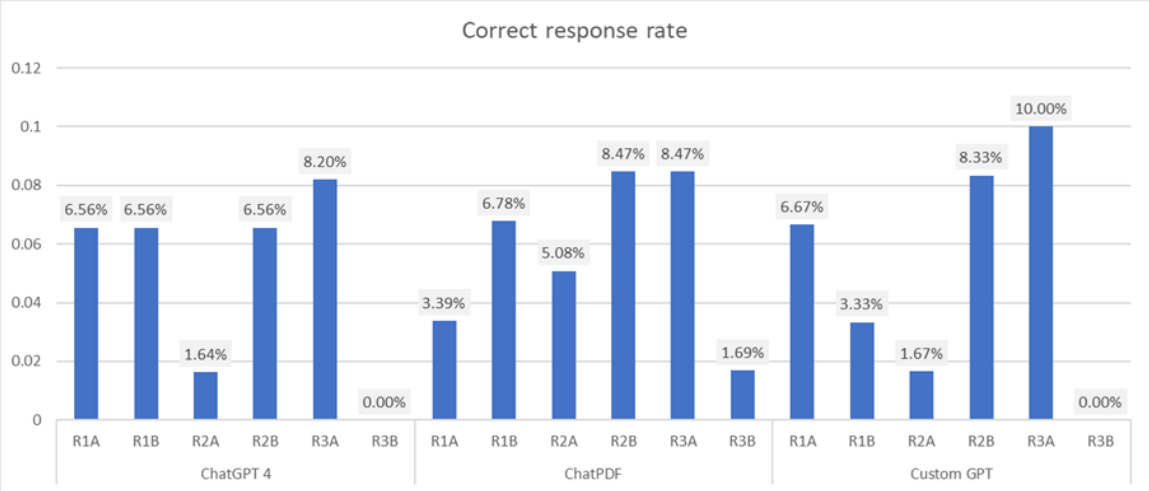

Figure 3: Correct response rates per LLM and report.

However, the overall results are not over 10% and lie on average at 5.2%, while the median is slightly higher at 6.6%. The reliability of the LLM results was inconsistent, with ChatGPT experiencing technical issues in some cases. The length of the report caused the GPT to crash in some cases.

In summary, it can be concluded that LLMs are useful for analyzing textual content and summarizing large documents. The results revealed that the accuracy of the balance sheet evaluation and the overall rate of correct answers were 31.11%. Thus, the quality of LLMs is insufficient for reliable analysis of balance sheets and financial statements.

5. DISCUSSION AND CONCLUSIONS

This research project investigated the quality (RQ 3) and limitations (RQ 4) of current LLMs in extracting data from tables. This study focused on analyzing the quality of data extracted from financial statements, such as annual reports. Tests were conducted using six questions, three financial reports, and three different tools: ChatGPT 4, ChatPDF, and a custom GPT. The results revealed low performance (in quality) in number calculations and significant limitations in the calculation of ratios.

The tests demonstrated differences in the results of the LLMs used (FIGURE 4). ChatPDF delivered the best overall performance, achieving 33.90%, compared to 29.51% for ChatGPT 4 and 30% for the custom GPT. However, these differences are marginal, indicating that there is no significant difference in quality among the three LLM tools used (RQ 5).

In addition to the tests conducted, the research indicates that the preparation of the reports (RQ 1), by redacting the data (the annual reports to the balance sheets and tables only), by removing

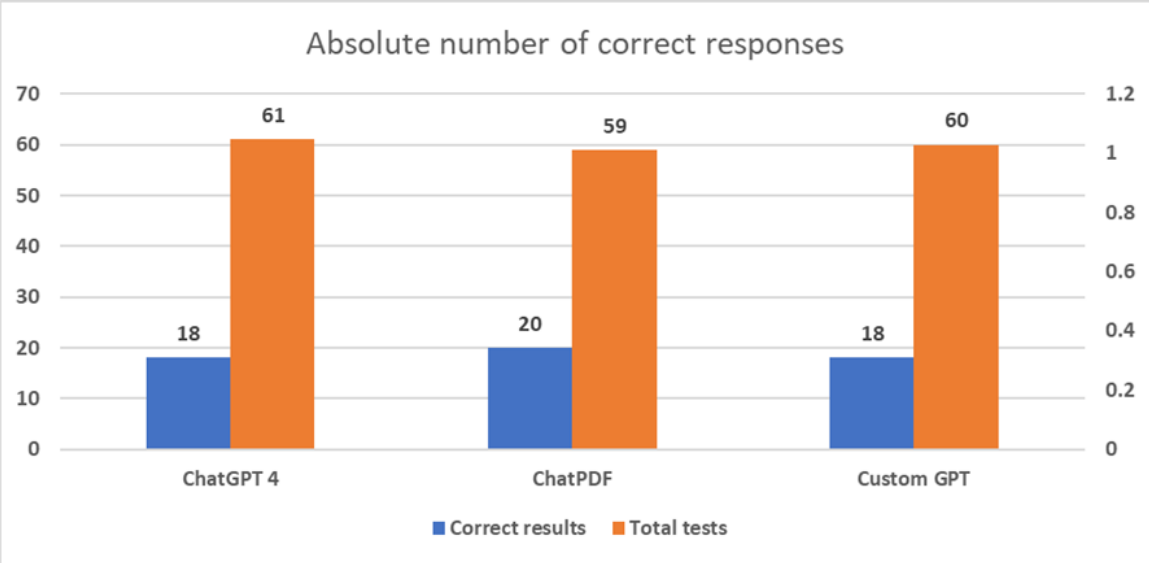

Figure 4: Overall correct responses per LLM in all 180 tests.

text and images, does not improve the results. In contrast, the tests demonstrate that the use of original reports (as in the example of the Swisscom report) yields better results. Furthermore, the test indicates that the LLM tends to extract data from the textual content rather than from the tables.

To identify the best way to obtain more accurate results (RQ 2). The tests conducted were performed using each tool and report with different prompts, and the results could be improved by adding additional information to the prompt. In the test, the correct year of the report and company name were added, which led to a better result or helped the LLM not return the previous year's result. One example is test ID "Q5.R1A.T2.P98.22:55", where the LLM could only find the information with specific instructions. A second example is test ID "Q6.R2B.T3.P1113.14:22" where the result was accurate due to the clear instructions on how the system must solve the question.

5.1 Limitations

To obtain results for the research questions regarding quality (RQ 3) and limitations (RQ 4) , this research project limited the field of tests to the specific field of annual reports with a limited number of questions (Section 3), reports (Section 3.2), and LLM tools (Section 3.3).

One of the main limitations encountered during the testing was the lack of clear and transparent control over the resources used. Despite uploading the PDF files, it was unclear which resources, knowledge, and rules were applied. Only ChatPDF mentioned the page used in the reports. This may have significantly impacted on the results, such as the use of data from web sources or misinterpretation of text in the reports or tables. Consequently, it cannot be definitively determined what caused the correct or incorrect results, whether external sources were used, or whether the LLM was hallucinating. According to Xu et al. (2024) [48], hallucination in LLMs is a natural phenomenon because these models cannot learn all computable functions.

The possibilities of this research were also limited to free versions of LLMs or even paid versions of ChatGPT. During the tests, there was a limitation in the allowed number of requests to use the LLMs, even with a paid version of GPT 4. For the test, Chat GPT Plus was used with a limit of 40 messages within 4 hours [49]. ChatPDF also had a restriction for free users, as seen in the account settings during the tests. The PDF File size can be max 32 MB. It is only possible to make up to 20 questions per PDF, and only two PDFs per day can be uploaded.

5.2 Practical Implications

This study identified certain limitations in using LLMs to extract and analyze data from tables in PDF files. Instances of inaccurate results or the absence of results can lead to misleading information. Reports that heavily feature tabular data often result in incorrect or no extraction, whereas text-rich reports are more likely to yield usable content. However, even text-focused reports require manual verification to ensure accuracy and reliability.

Despite these challenges, LLMs can be notably beneficial in parsing extensive narrative content in long reports. For instance, LLMs can effectively identify specific information, such as shareholders' equity by mentioning the exact page number on which the data are located. They also prove useful in generating concise summaries that encapsulate the content, style, and textual analyses. It is important to note, as discussed in earlier chapters, that the original intent behind the development of LLMs was to process and translate textual information rather than analyze complex tables and financial statements.

In practical applications, it is therefore advisable to prioritize reports that focus on analysis, particularly those that are abundant in textual information. This preference is recommended because current LLMs are proficient in processing and understanding text-based data. Conversely, reports that heavily depend on data presented in tables or in visual formats like images may not be extracted and analyzed effectively by LLMs.

The current generation of LLMs has yet to mature in the capacity to fully handle tasks involving the analysis of financial documents, especially those with tables and graphical data. While their capabilities continue to evolve and show potential, reliance on LLMs for critical analysis should be undertaken with caution. Researchers and practitioners must be aware of these limitations and consider them when integrating LLMs into their workflow.

5.3 Recommendations for Future Research

Future research should focus on creating more detailed and specific prompts to prevent mistakes or misinterpretations when using LLMs to extract and analyze data from financial texts. The development of prompts using contextual aids can enhance the performance of LLMs in this domain. Additionally, it would be beneficial to retest LLMs using the same questions and sources because technology is evolving rapidly. Better results can be achieved when advanced models get released like, for example, GPT-5. These models should be tested to assess improvements in these capabilities.

It may be useful to create language models that specialize in understanding financial papers. Experimenting with financial documents that have been changed a bit, like those with different titles or missing the usual top and bottom parts, could yield valuable insights into the adaptability of LLMs to variations in document structures. This helps them to handle unexpected things that they might see when they are used for real-life tasks. This type of research could also focus on identifying the causes behind incorrect results and developing methods to prevent these issues, for example, by directing LLMs to search within specific sections of the documents.

Since the tools used in the current research were online, it is possible that the LLMs used internet searches or generated hallucinated results. Investigating the differences between cloud-based and on-premise LLM tools can provide a clearer understanding of these dynamics. This comparative analysis is crucial not only for performance evaluation but also to mitigate privacy risks when dealing with nonpublic data. By adopting this approach, the reliability and security of language models used in financial analysis can be enhanced through future research.

References

[1] Teubner T, Flath CM, Weinhardt C, Van Der Aalst W, Hinz O. Welcome to the Era of ChatGPT et al. the prospects of large language models. Business & Information Systems Engineering. 2023;65:95-101.

[2] Han X, Zhang Z, Ding N, Gu Y, Liu X, et al. Pre-trained Models: Past, Present and Future. AI Open. 2021;2:225-250.

[3] Khurana D, Koli A, Khatter K, Singh S. Natural Language Processing: State of the Art, Current Trends and Challenges. Multimedia tools and applications. 2023;82:3713-3744.

[4] Bengio Y, Ducharme R, Vincent P, Jauvin C. A Neural Probabilistic Language Model. Journal of machine learning research. 2003;13:1137-1155.

[5] Hudson DL, Cohen ME. Neural Networks and Artificial Intelligence for Biomedical Engineering. John Wiley & Sons. IEEE Press Marketing. 1999.

[6] Kang Y, Cai Z, Tan CW, Huang Q, Liu H. Natural Language Processing (NLP) in Management Research: A Literature Review. J Manag Anal. 2020;7:139-172.

[7] Parunak HV. Applications of Distributed Artificial Intelligence in Industry. Foundations of distributed artificial intelligence. 1996;2:18.

[8] Ennals R, Ennals R. Approaches to Artificial Intelligence. Artificial Intelligence and Human Institutions. Springer. 1991:1-8.

[9] Liu, B. Sentiment Analysis and Opinion Mining. Springer Nature. 2022.

[10] Och FJ. Minimum Error Rate Training in Statistical Machine Translation. In Proceedings of the 41st annual meeting of the Association for Computational Linguistics. 2003:160-167.

[11] Kasneci E, Seßler K, Küchemann S, Bannert M, Dementieva D, et. al. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. Learning and individual differences. 2023;103:102274.

[12] Lauriola I, Lavelli A, Aiolli F. An Introduction to Deep Learning in Natural Language Processing: Models Techniques and Tools. Neurocomputing. 2022;470:443-456.

[13] Torfi A, Shirvani RA, Keneshloo Y, Tavaf N, Fox EA. Natural Language Processing Advancements by Deep Learning: A Survey. 2020. ArXiv preprint: https://arxiv.org/pdf/2003.01200

[14] Sage C, Douzon T, Aussem A, Eglin V, Elghazel H, et al. Data-Efficient Information Extraction From Documents With Pre-trained Language Models. In E. H. Barney Smith & U. Pal (Eds.), Document Analysis and Recognition – ICDAR Workshops.2021;12917:455–469.

[15] Govindaraju V, Zhang C, Ré C. Understanding Tables in Context Using Standard Nlp Toolkits. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics. 2013:658-664.

[16] Douglas M. R. Large Language Models. 2023. ArXiv Preprint: https://arxiv.org/pdf/2307.05782

[17] Shwartz-Ziv R, Balestriero R, LeCun Y. What Do We Maximize in Self-Supervised Learning? 2022. ArXiv preprint: https://arxiv.org/pdf/2207.10081v1

[18] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, et. al. Attention Is All You Need. Advances in neural information processing systems. 2017;30.

[19] Kasem M, Abdallah A, Berendeyev A, Elkady E, Abdalla M, et. al. Deep Learning for Table Detection and Structure Recognition: A Survey. 2022. ArXiv preprint: https://arxiv.org/pdf/2211.08469

[20] Kieninger T, Dengel A. The T-Recs Table Recognition and Analysis System. In Document Analysis Systems: Theory and Practice: Third IAPR Workshop, DAS'98 Nagano, Japan, November 4–6, 1998. Springer Berlin Heidelberg. 1999:255-270.

[21] Cesarini F, Marinai S, Sarti L, Soda G. Trainable Table Location in Document Images. In 2002 International Conference on Pattern Recognition. IEEE. 2002;3:236-240.

[22] Fan M, Kim DS. Detecting Table Region in PDF Documents Using Distant Supervision. 2015. ArXiv preprint: https://arxiv.org/pdf/1506.08891

[23] Wang Y, Hu J. A Machine Learning Based Approach for Table Detection on the Web. In Proceedings of the 11th international conference on World Wide Web.ACM. 2002:242-250.

[24] Kasar T, Barlas P, Adam S, Chatelain C, Paquet T. Learning to Detect Tables in Scanned Document Images Using Line Information. In 2013 12th International Conference on Document Analysis and Recognition. IEEE. 2013:1185-1189.

[25] O'Gorman L. The Document Spectrum for Page Layout Analysis. IEEE Transactions on pattern analysis and machine intelligence. 1993;15:1162-1173.

[26] Shafait F, Smith R. Table Detection in Heterogeneous Documents. In Proceedings of the 9th IAPR international workshop on document analysis systems. ACM. 2010:65-72.

[27] Hao L, Gao L, Yi X, Tang Z. A Table Detection Method for PDF Documents Based on Convolutional Neural Networks. In 2016 12th IAPR Workshop on Document Analysis Systems (DAS). IEEE. 2016:287-292.

[28] Gilani A, Qasim SR, Malik I, Shafait F. Table Detection Using Deep Learning. In 2017 14th IAPR international conference on document analysis and recognition (ICDAR). IEEE. 2017:771-776.

[29] Schreiber S, Agne S, Wolf I, Dengel A, Ahmed S. Deepdesrt: Deep Learning for Detection and Structure Recognition of Tables in Document Images. In 2017 14th IAPR international conference on document analysis and recognition (ICDAR). IEEE. 2017;1:1162-1167.

[30] He D, Cohen S, Price B, Kifer D, Giles CL. Multi-Scale Multi-Task FCN for Semantic Page Segmentation and Table Detection. In 2017 14th IAPR international conference on document analysis and recognition (ICDAR). IEEE. 2017;1:254-261.

[31] Arif S, Shafait F. Table Detection in Document Images Using Foreground and Background Features. In 2018 Digital Image Computing: Techniques and Applications (DICTA). IEEE. 2018:1-8.

[32] Reza MM, Bukhari SS, Jenckel M, Dengel A. Table Localization and Segmentation Using GAN and CNN. International Conference on Document Analysis and Recognition Workshops (ICDARW). IEEE. 2019;5:152-157.

[33] Agarwal M, Mondal A, Jawahar CV.CDeC-Net: Composite Deformable Cascade Network for Table Detection in Document Images. In 2020 25th international conference on pattern recognition (ICPR).IEEE. 2021:9491-9498.

[34] Smock B, Pesala R, Abraham R. PubTables-1M: Towards comprehensive table extraction from unstructured documents. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:4634-4642.

[35] Li J, Xu Y, Lv T, Cui L, Zhang C, Wei F. DIT: Self-Supervised Pre-training for Document Image Transformer. In Proceedings of the 30th ACM international conference on multimedia. ACM. 2022;3530–3539.

[36] Zha L, Zhou J, Li L, Wang R, Huang Q, et al. Tablegpt: Towards Unifying Tables, Nature Language and Commands Into One GPT. 2023. ArXiv preprint: https://arxiv.org/pdf/2307.08674

[37] Li P, He Y, Yashar D, Cui W, Ge S, et al. Table-Gpt: Table-Tuned GPT for Diverse Table Tasks. 2023. ArXiv preprint: https://arxiv.org/pdf/2310.09263

[38] https://papers.nips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[39] Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, et al. Training Language Models to Follow Instructions With Human Feedback. Advances in neural information processing systems. 2022;35:27730-27744.

[40] Meyer C. Bewertung der Sachanlagen nach Swiss GAAP FER: eine konzeptionelle Analyse. Der Schweizer Treuhänder. 2008;82:317-324.

[41] https://www.deloitte.com/ch/en/services/audit-assurance/services/swiss-gaap-fer.html

[42] https://viewpoint.pwc.com/dt/us/en/pwc/accounting_guides/ifrs_and_us_gaap_sim/assets/pwcifrsusgaap0623.pdf

[43] https://www.iasplus.com/de/publications/global/ifrs-in-your-pocket/2021/file

[44] https://fasab.gov/accounting-standards/

[45] https://annualreport.ruag.com/2022/sites/ar22/files/media_document/2023-03/RUAG_2022_GB_EN.pdf

[46] https://assets.roche.com/f/126832/x/8971737b76/fb22e.pdf

[47] https://reports.swisscom.ch/download/2022/de/swisscom_geschaeftsbericht_gesamt_2022_de.pdf

[48] Xu Z, Jain S, Kankanhalli M. Hallucination Is Inevitable: An Innate Limitation of Large Language Models. 2024. ArXiv preprint: https://arxiv.org/pdf/2401.11817

[49] https://help.openai.com/en/articles/7102672-how-can-i-access-gpt-4