Aalto-yliopisto
Sähkötekniikan
korkeakoulu

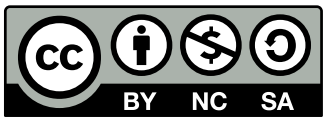
Master's Programme in Data Science

Automating Information Extraction from Financial Reports Using LLMs

A Comparative Study of Text, Image, and Multimodal Approaches

Gabriel Gomes Ziegler

Master's Thesis
2024



Aalto-yliopisto
Sähkötekniikan
korkeakoulu

Author Gabriel Gomes Ziegler

Title Automating Information Extraction from Financial Reports Using LLMs — A Comparative Study of Text, Image, and Multimodal Approaches

Degree programme Data Science

Major ICT Innovation

Supervisor Prof. Bo Zhao

Advisor MS Liliya Shakhpazyan (MSc)

Collaborative partner Datia

Date 26 September 2024 **Number of pages** 51 **Language** English

Abstract

This thesis investigates the application of the latest Large Language Model (LLM)s for the automated extraction of Environmental, Social, and Governance (ESG) indicators from financial reports, a critical task for companies focused on keeping up-to-date sustainability reporting. The study explores the trade-offs in three distinct approaches: text-only, image-only, and multimodal, to evaluate their effectiveness in a real-world diverse dataset of financial reports.

The text-only approach, although effective for documents with structured textual data, struggled with visual-rich content, leading to high residual difference between the extracted and actual values. The image-only approach, while adept at interpreting visual elements, faced challenges with hallucinations and lacked the accuracy of text-based methods that tend to get more perfect matches than the image-based counterparts. The multimodal approach, which combines text and image data, demonstrated superior performance, achieving the highest accuracy with a perfect match rate exceeding 85% for key indicators like Scope 1 and Scope 3 emissions. This method effectively mitigated errors through cross-validation between text and image data, resulting in minimal residuals and reliable data extraction.

The study's findings underscore the possibility of using the three approaches according to the nature of the dataset having a more confident usage when applying the combined multimodal approach in DocumentAI, particularly in format-agnostic scenarios such as the one presented in this study. The thesis also identifies challenges, such as the ambiguity in sub-indicators under Scope 2 emissions, and the importance of examining the tolerance for errors in the context of the application, where residuals can be acceptable or make the system completely unusable. This work contributes to the growing field of DocumentAI, offering insights into the capabilities and limitations of LLM for financial report analysis, and suggests pathways for future advancements in automated information retrieval systems.

Keywords Large Language Models, Document AI, Information Extraction, GPT-4, Deep Learning, Machine Learning, Automated Data Retrieval, Natural Language Processing

Aalto-yliopisto
Sähkötekniikan
korkeakoulu

Tekijä Gabriel Gomes Ziegler

Työn nimi Automating Information Extraction from Financial Reports Using LLMs — A Comparative Study of Text, Image, and Multimodal Approaches

Koulutusohjelma Data Science

Pääaine ICT Innovation

Työn valvoja ja ohjaaja Prof. Bo Zhao

Yhteistyötaho Datia

Päivämäärä 26 September 2024 **Sivumäärä** 51 **Kieli** englanti

Tiivistelmä

Avainsanat Large Language Models, Document AI, Information Extraction, GPT-4, Deep Learning, Machine Learning, Automated Data Retrieval, Natural Language Processing

Preface

This thesis marks the completion of my Double Master's Degree in Europe, awarded by Aalto University and the Technical University of Eindhoven. Over the past two years, I have had the privilege of living in four different countries, forming lifelong friendships, meeting inspiring professors and entrepreneurs, and establishing myself as a Software Engineer specializing in Machine Learning. These experiences are incredibly aligned with the expectations I held when I decided to pursue a Master's Degree overseas. Actually, the outcome absolutely exceeded my expectations.

For all of this, I owe my deepest gratitude to my mother, who has tirelessly worked to provide me and my sister with opportunities that she herself was not afforded. Her unwavering dedication is the reason I stand here today, having fulfilled my dreams. She continues to set an example of the kind of parent I aspire to be in the future. Obrigado, mãe!

I would also like to extend my sincere thanks to EIT Digital for granting me the scholarship that enabled my participation in this program. I am equally grateful to Datia for the opportunity to work on such an engaging and impactful topic, and for allowing me to collaborate with so many remarkable individuals. Lastly, but certainly not least, I want to thank Nano Brasca – my friend and supervisor at Datia – for sharing invaluable ideas and conversations over these past years. It has been a privilege to be surrounded by such inspiring individuals, and I hope to one day inspire others in a similar way.

Otaniemi, 22 September 2024

Gabriel Gomes Ziegler

Contents

1 Introduction

1.1 Overview

The digital age has transformed how information is stored, accessed, and analyzed, with financial reports being a critical source of data for businesses, regulators, and investors. Traditionally, extracting meaningful information from these reports required manual labor or rule-based systems, which were not only time-consuming but also prone to errors. The advent of machine learning, particularly the development of large language models (LLMs), has brought about significant advancements in this field. These models, such as Generative Pre-trained Transformer (GPT) and Bidirectional Encoder Representations from Transformers (BERT), offer unprecedented capabilities in processing unstructured data, making it possible to automate some information extraction tasks with high accuracy and efficiency.

1.2 Context and Motivation

Financial reports are rich sources of information, encompassing a wide range of data, including financial statements, management discussions, and ESG disclosures. In companies that aim to provide up-to-date analysis on financial resources, keeping a database with key indicators from companies all over the world is essential. Since there are many companies and reports are released with high frequency throughout the year, it is not feasible to manually extract this information. Therefore, there is a huge demand for automated systems that can extract this information from documents or at least speed up the process of acquiring and structuring this information. Among these, the extraction of Greenhouse Gas (GHG) emissions data has gained particular importance due to increasing regulatory requirements and the growing emphasis on sustainability. However, the diverse formats and layouts used in these reports pose significant challenges for automated data extraction systems. The need for a robust solution that can handle such variability is more pressing than ever, as businesses and stakeholders demand faster and more accurate insights from financial documents. For these reasons, in this study, we focus on the extraction of GHG emissions data from financial reports, aiming to develop and evaluate methodologies to understand the effectiveness, issues, and trade-offs to consider when implementing automated systems in this process.

1.3 Problem Definition

The core challenge addressed in this thesis is the development and comparison of methodologies for extracting GHG emissions data from non-standardized financial reports. This problem is compounded by the variety of ways in which companies present their data, ranging from structured tables to complex visual elements such as charts and infographics. Traditional approaches, whether manual or rule-based, often fall short in dealing with this complexity, leading to incomplete or inaccurate data

extraction. The rise of LLMs offers a promising avenue to overcome these limitations by leveraging advanced text and image processing capabilities.

1.4 Objectives of the Study

The primary objective of this study is to conduct a comprehensive evaluation of different LLM-based approaches for extracting information from financial reports, particularly focusing on non-standard report formats.

The study aims to:

1. Compare the effectiveness of text-only, image-only, and multimodal approaches in extracting GHG emissions data.
2. Enhance the precision and efficiency of data extraction by optimizing LLM techniques for handling diverse document layouts.
3. Develop a scalable system capable of processing large volumes of financial reports while maintaining high accuracy.
4. Identify the limitations and challenges associated with current LLM technologies in the context of financial data extraction.

By achieving these objectives, this thesis seeks to contribute to the broader field of automated document analysis and provide actionable insights for improving data extraction processes in real-world applications.

1.5 Research Questions

The central research question guiding this thesis is: *"What are the most effective strategies for utilizing LLMs to accurately and efficiently extract financial data from unstructured reports?"* To address this question, several sub-questions are explored:

- What are the trade-offs between text-only, image-only, and multimodal approaches in the context of financial data extraction?
- Do the systems fail similarly across different types of financial reports, or are there specific challenges associated with certain industries or formats?
- Are there more challenging sub-indicators or metrics within the GHG emissions data that require special attention in the extraction process?
- Are there more challenging formats that the indicators are presented for particular approaches, such as text, image, or multimodal?

These research questions are essential for developing a comprehensive understanding of how LLMs can be effectively applied to the task of financial data extraction, particularly in the context of non-standard report formats.

1.6 Scope and Limitations

This study focuses on the extraction of financial data, specifically GHG emissions metrics, from financial reports. The reports used in this study are publicly available and include annual and quarterly statements from various industries. The scope is deliberately confined to GHG emissions data to maintain a clear focus, although the methodologies developed could be extended to other types of financial data in future research. The primary constraints include the variability in report formats, the complexity of financial and environmental terminology, and the current limitations of LLM technologies in fully understanding and processing domain-specific contexts. Additionally, the study is constrained by the computational resources available, which may affect the scalability and speed of the proposed solutions.

Scope

1. Analyzing the effectiveness of text-only, image-only, and multimodal approaches for extracting GHG emissions data from financial reports using GPT-4 models.
2. Evaluating the types of errors and trade-offs posed by different approaches in the context of non-standard report formats.
3. Identifying the challenges and limitations of current LLM technologies in handling financial data extraction tasks.

Limitations

1. Due to limited human resources for dataset labeling, the experiments in this thesis are conducted with fewer samples than ideal for a more robust analysis.
2. Strict time constraints prevented the inclusion of models from competing approaches in the study.
3. The study focuses exclusively on GHG emissions in sustainability and financial reports. Other types of indicators and reports are not considered.

1.7 Structure of the Thesis

This thesis is organized to systematically explore and address the challenges of financial data extraction using LLMs. The structure is as follows:

1. **Introduction**: Provides the context, problem definition, and objectives of the study.
2. **Literature Review**: Discusses the concepts and state of the art in document analysis, LLMs, and information retrieval from financial reports.
3. **Methodology**: Details the dataset used, the experimental setup, and the methodologies implemented for data extraction.

4. **Results**: Presents and analyzes the results of the experiments, comparing the performance of different approaches.

5. **Conclusion**: Summarizes the findings, discusses the implications and limitations, and suggests directions for future research.

6. **References**: Lists the academic and technical sources cited throughout the thesis.

This structure is designed to provide a logical flow of information, guiding the reader through the development, implementation, and evaluation of the research conducted in this thesis.

2 Concepts and State of the Art

Documenting and searching for information is an ancient human practice that can be traced back as far as 3000 BC, when the Sumerians—the first civilization in the world—used clay tablets with cuneiform inscriptions to keep track of legal documents, transaction records, literature, mythological tales, among other information. They also created different categories to differentiate tablets based on their contents in a classification fashion [1]. Similar practices have remained largely relevant throughout history. With the invention of paper and the printing press, the practice of documenting and storing information evolved, allowing for more information to be stored and shared physically. Not so long ago, in the 20th century, the invention of the computer and the internet revolutionized the way information is stored and shared, allowing for information to be stored and shared digitally. This allowed for vast amounts of information to be stored and shared in a way that was never possible before. This period in time is often referred to as the information age, also known as the third industrial revolution, which marks a time when information became increasingly accessible and also a commodity, especially later in the 21st century with the rise of Artificial Intelligence (AI) and Machine Learning (ML) models that feed on large datasets to learn and make predictions. Additionally, the creation of Portable Document Format (PDF)s by Adobe in 1993 [2] established a standard for storing and sharing information in a portable format that could be easily shared and printed, quickly becoming a standard for sharing information, especially in the business world. The digitalized information stored in PDFs soon became a target for information retrieval and data mining techniques, where systems were developed to extract information from these files based on a variety of approaches, such as heuristic-based methods [3, 4, 5], vector space models [6, 7], probabilistic models [8], and more recently, Deep Learning (DL) models, especially those leveraging the transformer architecture [9, 10, 11, 12].

2.1 Information Retrieval

Information Retrieval (IR) is a field of study that focuses on the organization, storage, and retrieval of bibliographic information. IR systems are used to provide a response to a user query with references to documents that contain the information sought by the user. Although the field of IR has been improved and become more popular in recent years with the usage of Natural Language Processing (NLP) techniques dominating the field, such as Retrieval Augmented Generation (RAG)s, the core task of IR has been studied and applied since the 1940s, with pioneers like Vannevar Bush, who in 1945 envisioned the Memex, a theoretical device that would store extensive collections of documents and allow rapid retrieval and cross-referencing of information [13].

2.2 Document AI

The procedure of extracting information from a document—recently referred to as "Document AI" [14] or "Document Understanding" [15]—is a complex problem due

to the diverse nature of data that PDFs allow to store. Such problems often involve cross-modal interactions where information is represented in both natural language text and visual elements such as tables, charts, and images. In the visual domain, document layout analysis has been widely studied and applied using Computer Vision (CV) techniques to detect and extract elements in the document. Document image processing is usually treated as an object detection task where elements such as text, tables, and images are detected and classified [16, 17].

In many scenarios, the information presented in a document requires an approach that can leverage understanding from both the text and visual domains. For instance, consider tables where the values in cells only have meaning when associated with the header row, which is commonly described in the first row or column of the table. Other structures such as charts only convey information when the labels and axes are correctly identified and associated with the data points. These problems are particularly true for financial reports, where information is presented in text, tables, charts, infographics, and other visual elements. The requirement to handle layout invariance is tackled by general-purpose models, such as LLMs, that through pre-training of general-purpose models, learn the position information of the elements as well as the visual information presented with the text, such as font size, color, and style. These visual cues can be learned by visual encoders and combined with the text in the pre-training stage, significantly improving the model's capability to abstract non-standardized information from documents [14].

2.3 Transformers

In recent years, AI systems with DL architectures have been the norm for most of the state-of-the-art models introduced. The first DL models were based on fully-connected neural networks, which are composed of layers of neurons that are connected to each other in a feed-forward fashion [18, 19]. Decades after the first fully-connected neural networks were published, the Convolutional Neural Networks (CNN) architecture was proposed by Yann LeCun in 1998 [20] and later popularized by the AlexNet model in 2012 after winning the ImageNet competition [21]. These models represented a significant improvement in the field of CV and were able to learn hierarchical features from images, allowing for more complex patterns to be learned. Meanwhile, in the field of NLP, the Recurrent Neural Networks (RNN) architecture was proposed in 1986, bringing the concept of sequential processing to neural networks, allowing previous states to be considered in the processing of the current state [22]. The RNN architecture was able to learn sequential patterns from text data, allowing for the development of models that could generate text, translate languages, and more. However, the RNN architecture had limitations in learning long-range dependencies due to the vanishing gradient problem [23]. To address this issue, the Long Short-Term Memory (LSTM) architecture was proposed in 1997 by Sepp Hochreiter and Jürgen Schmidhuber [24]. The LSTM architecture introduced the concept of memory cells that could store information over time, allowing for the learning of long-range dependencies in sequential data. The next breakthrough in the AI field came with the introduction of the transformer architecture in 2017, which revolutionized the

field of NLP by utilizing self-attention mechanisms to learn intrinsic features [25]. The transformer architecture brought significant improvements in the field of NLP, as demonstrated by the BERT model in 2019 [26] and later by the GPT-3 model in 2020 [27], which would become the state-of-the-art models in the field of NLP. Much inspired by the major advancements in the field of NLP with the transformer architecture, researchers started applying the transformer architecture to the field of CV as an alternative to the long-standing CNN architecture that has been the standard for many years [28]. These models, known as visual transformers, have shown to be competitive with CNNs in many tasks and have demonstrated the potential to learn hierarchical features from images in a more efficient way than CNNs [29].

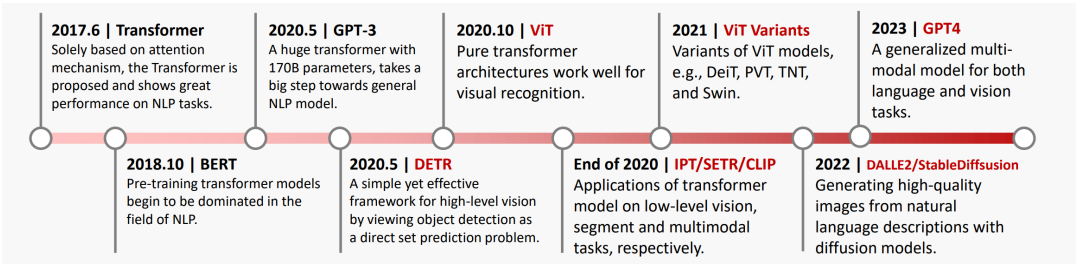


Figure 1: Key milestones in the development of transformer. The vision transformer models are marked in red (Image from 'A Survey on Vision Transformer' [30]).

The original transformer architecture [25] is a sequence-to-sequence model composed of an encoder and a decoder block that are themselves composed of multiple layers of multi-head self-attention mechanisms and feed-forward neural networks.

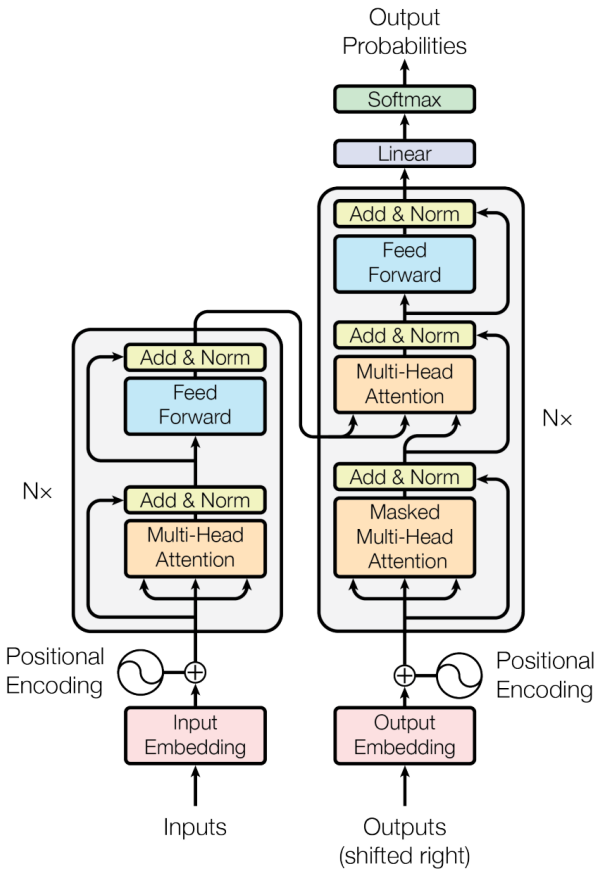


Figure 2: Transformer architecture. The model is composed of an encoder and a decoder block that are themselves composed of multiple layers of multi-head self-attention mechanisms and feed-forward neural networks [25].

The attention function, often described as a map of a query and a set of key-value pairs to an output, outputs a weighted sum of the values based on the similarity of the query to the keys, computing a compatibility function between the query and the corresponding key.

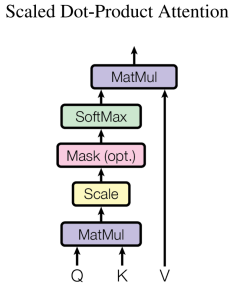


Figure 3: Scaled dot-product attention function [25].

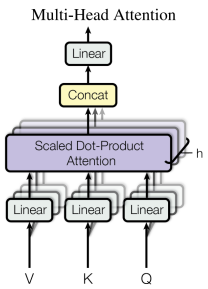


Figure 4: Multi-head attention mechanism [25].

The attention function is computed on a set of queries in a parallel fashion, where

the queries are packed into a matrix Q, the keys into a matrix K, and the values into a matrix V. The attention function is computed as follows [25]:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

2.3.1 Transformers in Vision

Before the introduction of the transformer architecture in the visual domain, traditional supervised techniques suffered from the necessity of large amounts of labeled data to learn hierarchical features from images, limiting the scalability and generalization of such models. The incorporation of models such as BERT, GPT, and DALL-E, which are trained on large amounts of text data and generally with a self-supervised learning approach, leads to a decreased dependence on carefully annotated labels, which might allow room for progress on essential cognitive skills that have been challenging in the traditional supervised learning paradigm [31, 32, 33].

The application of transformer architectures to vision tasks has revolutionized the field. Vision Transformers (ViTs) have demonstrated that by treating image patches as sequences of tokens, similar to words in text, one can leverage the powerful sequence modeling capabilities of transformers to process and analyze visual data. ViTs have shown remarkable performance across various tasks, including image classification, object detection, and segmentation, often surpassing the performance of traditional CNNs [29].

With such a trend of transformer architectures in the visual domain, OpenAI released GPT-4 Vision, which integrates advanced transformer architectures specifically designed for understanding and generating visual content. GPT-4 Vision builds on the principles of its text-based equivalent GPT-4 but adapts the transformer model to handle the unique challenges of image inputs [34]. This includes learning complex spatial relationships and capturing fine-grained details from images, which are crucial for tasks such as image synthesis, enhancement, and interpretation.

The integration of multimodal learning in GPT-4 Vision allows for the processing of textual and visual information, enhancing the model's ability to understand and generate coherent and contextually enriched content. This multimodal capability is particularly valuable in processing documents that contain both text and visual elements, such as financial reports, where information is presented in a variety of formats, including tables, charts, and images. For these reasons and for its availability, GPT-4 Vision is the model of choice for multimodality experiments in this study.

2.4 Large Language Models

LLMs are a class of statistical language models based on neural networks—oftentimes utilizing the transformer architecture—that are large in size and pre-trained on vast amounts of text data. These models represent a significant advancement in the NLP field, allowing complex tasks to be performed with high accuracy and efficiency. This technology has been widely researched and developed by leading big tech companies

such as OpenAI, Google, and Meta [27, 26, 35, 36, 37, 38], which have pushed the boundaries of transformer pre-trained models to the point that they can generate human-like text, understand context, and even perform tasks that require domain-specific knowledge. This milestone signifies not only a significant improvement in the field of NLP but also a major adoption of artificial intelligence agents by the average user, especially with OpenAI's ChatGPT release in 2022, which delivered a fine-tuned GPT-3.5 model utilizing reinforcement learning from human feedback using the same methods as the InstructGPT model [39].

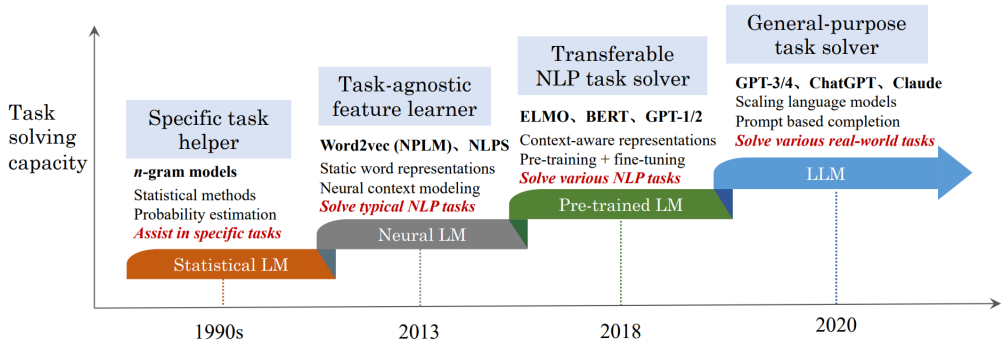


Figure 5: Key milestones of large language models from the task-solving perspective [40].

2.4.1 GPT-4

The fourth GPT release by OpenAI is a large language model with multilingual and multimodal capabilities that allow it to process image and text inputs, producing text outputs. The model was developed with the aim of improving the ability to comprehend and generate natural language text. GPT-4 is often evaluated against human performance on tasks like the bar exam, LSAT, SAT, among others, and has shown to be competitive with human performance by achieving top 10% scores on bar exams, compared to GPT-3.5, which achieved bottom 10% scores [36]. The specifics about the model's architecture and training data are not disclosed by OpenAI, but it is known that the model has been pre-trained to predict the next word in a sentence, using publicly available and licensed data fine-tuned with Reinforcement Learning from Human Feedback (RLHF), which has a great impact on the model's performance [36]. While this new model has shown significant improvements in language understanding, generation, and reduction of hallucinations, it still has limitations in understanding context and generating coherent responses, which is a common issue in large language models [36].

2.4.2 GPT-4V

GPT-4 Vision represents an extension of the capabilities of traditional LLMs into the realm of visual understanding and analysis. By integrating vision-based artificial intelligence technologies with the language processing prowess of GPT-4, this model

can interpret and analyze images, diagrams, and visual data in conjunction with textual information. This multimodal approach enables GPT-4 Vision to perform tasks that require an understanding of both visual and textual content, such as extracting data from charts and graphs in financial reports, identifying key information in documents with complex layouts, and answering questions that depend on visual cues. The development of GPT-4 Vision is a testament to the ongoing advancements in AI, highlighting the move towards more integrated and comprehensive models that can navigate the complexities of human communication and information processing [34].

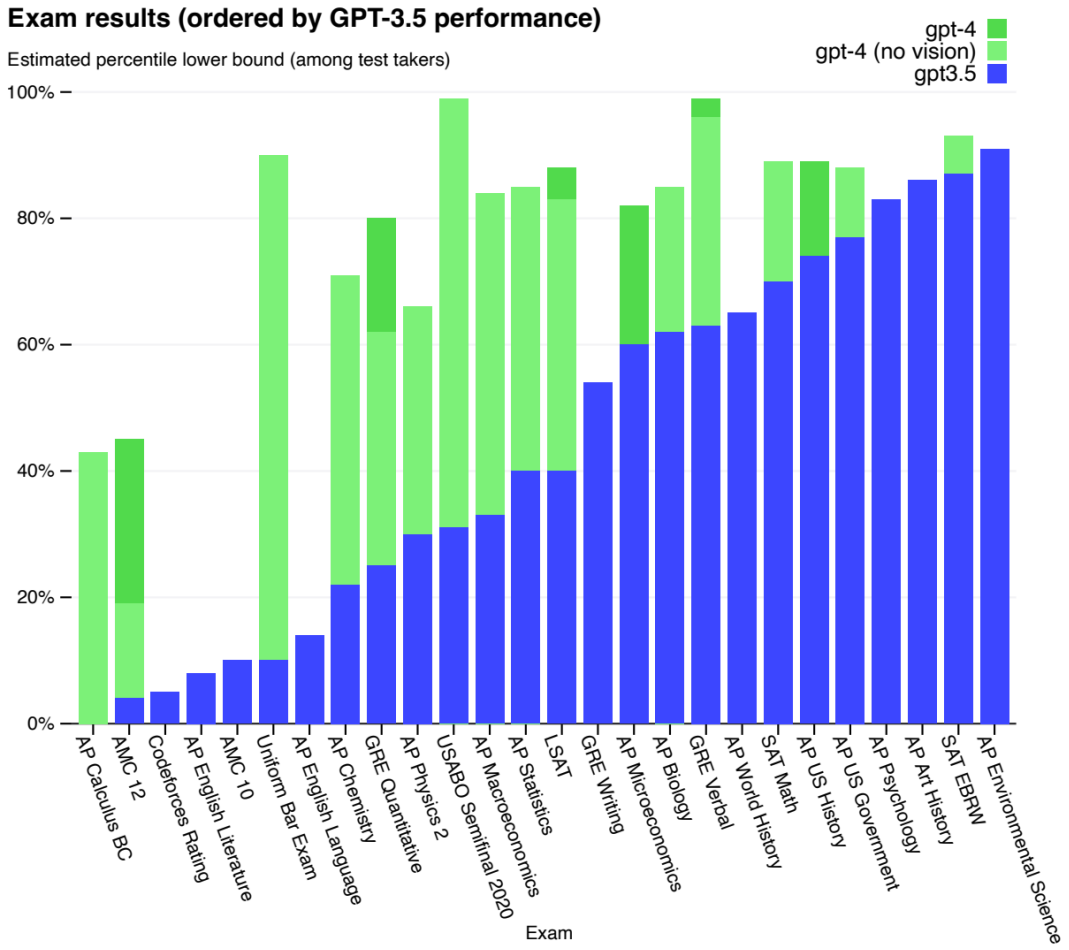


Figure 6

2.5 LLMs for Document AI

LLMs have become a popular strategy in the field of Document AI, transforming how information is extracted, processed, and analyzed from documents. In the context of Document AI, LLMs are utilized to understand the content within documents, ranging from simple text to complex structures like tables and charts, and the relationships between different pieces of information. These models leverage their extensive training on diverse datasets to adapt to the specific challenges posed by document analysis, such

as varying formats, layouts, and the integration of multimodal data. Through techniques such as transfer learning and fine-tuning, LLMs can be specialized to perform tasks including, but not limited to, information extraction, document summarization, and semantic search within documents. Their ability to process and analyze documents at scale significantly reduces the time and effort required for data entry, extraction, and analysis, enabling more efficient and accurate handling of document-based information.

2.6 Question Answering with RAG

RAG represents a novel approach in leveraging LLMs for the task of question answering. RAG combines the generative capabilities of GPT-like models with retrieval-based methods, which search a large corpus of documents to find relevant information that can aid in generating accurate and informative answers. This technique involves two main components: a retriever, which identifies relevant documents or passages given a query, and a generator, which synthesizes the retrieved information into a coherent response. By integrating these two processes, RAG is able to produce answers that are not only contextually relevant but also enriched with details and insights drawn from a wide range of sources. This method has shown significant promise in improving the accuracy and depth of responses provided by AI systems in question answering applications, particularly in domains where detailed and specific knowledge is required, such as academic research and technical support.

2.7 Issues with LLMs for Document AI

2.7.1 Hallucinations

Despite their many advantages, LLMs have known limitations when applied to Document AI tasks. One issue that arises with the generative nature of these models is the potential for generating incorrect or misleading information, especially when the input data is ambiguous or incomplete. These mistakes—oftentimes referred to as hallucinations—occur when the generative model creates plausible and convincing responses that are incorrect. Although it is possible to identify and mitigate these so-called hallucinations, studies have shown via learning theory that these mistakes are inherent to the generative nature of LLMs and cannot be completely extinguished [41]. The fact that these mistakes are realistic increases the difficulty of detecting them and brings uncertainty about the reliability of the information provided by the model in a productive environment. Comprehensive studies have been conducted to understand the causes of hallucinations and have found that these stem from a variety of reasons, including noisy data, poor parametric choices, incorrect attention mechanisms, improper training procedures, among others. There are two distinct categories of hallucinations identified in the literature: intrinsic hallucination and extrinsic hallucination, and they require different strategies to be mitigated [42].

Consider the following source data used as input to an LLM model:

> *The company reported revenues of $1 million in Q1 2022, and $2 million in Q2 2022.*

- **Intrinsic hallucination**: This type of hallucination occurs when the model generates information that contradicts the input data. A case of intrinsic hallucination would be if the model generated the following output:

  > *The company reported revenues of $1 million in Q1 2022, and $3 million in Q2 2022.*

  Here, the model has generated information that is inconsistent with the input data since the revenue reported in Q2 2022 is incorrect according to the source data. This type of hallucination can be particularly problematic in document analysis and is the one that this study aims to mitigate.

- **Extrinsic hallucination**: Extrinsic hallucination occurs when the model generates incorrect information that is not present in the input data but is plausible given the context. For example, if the model generated the following output:

  > *The company is projected to report revenues of $3 million in Q3 2022.*

  This information is not present in the input data and cannot be inferred from the given context, therefore it is classified as an extrinsic hallucination.

Several techniques aim to mitigate these issues, such as using retrieval-based methods [43], fine-tuning a model on a specific domain [27], or retrying the generation process multiple times while validating the output against a defined model, as proposed by tools like *LangChain* [44] and *instructor* [45]. In this study, we show applications of how using defined models with *pydantic* [46] and *instructor* can help mitigate hallucinations in the context of document data extraction.

2.7.2 Interpretability and Explainability

Ever since DL models gained popularity in productive environments, many concerns have been raised regarding the interpretability and explainability of these models, particularly in high-stakes applications such as healthcare, finance, and law, where accountability and transparency are crucial, and these models can have societal impacts, e.g., inequity, discrimination, misuse, economic and environmental impact, ethical concerns, among others. As already demonstrated in several studies, DL models are often treated as black boxes, because of the complex and non-linear nature of their architectures, making it difficult to understand how they arrive at their decisions and generate their outputs [47, 48]. LLMs are no exception to this, as their complex attention mechanisms and deep learning architectures make it challenging to interpret the reasoning behind their predictions and the information they generate.

The lack of interpretability and explainability can be a significant barrier to the adoption of LLMs in critical applications, as it not only raises concerns about the reliability, trustworthiness, and ethical implications of the decisions made by these models but also increases complexity when debugging and improving the models. The uncertainty involving LLMs' outputs poses a challenge for users who need to

understand what kinds of inputs lead to incorrect outputs. This is particularly important in the context of this study, as the information extracted from financial reports is used to make critical business decisions, and the reliability and accuracy of the extracted data are paramount. For these reasons, we investigate nuances in the documents that lead to erroneous predictions so that they can be avoided in the future.

3 Financial Reports Dataset

The dataset used in this study consists of a collection of annual and quarterly reports from various public companies across different industries and countries. We have a total of 1,000 reports—comprising over 32,580 pages across all documents in the dataset—carefully annotated manually with the GHG emissions data reported by the companies for the specified periods. These reports contain a wide range of information, including financial statements, management discussions and analysis, auditor reports, and ESG disclosures, including GHG emissions data, which is the primary focus of this study. The documents are stored in PDF format, with varying layouts, fonts, structures, and number of pages.

We opted to create a diverse dataset with heterogeneous formats to report GHG emission data, including text, tables, charts, and images, to evaluate the effectiveness of different strategies for extracting information from financial reports. This approach reflects the realistic nature of the data that the system would encounter in a real-world scenario in productive environments. Below, we present a few samples to illustrate the diversity of formats and layouts used to report GHG emissions data in financial reports.

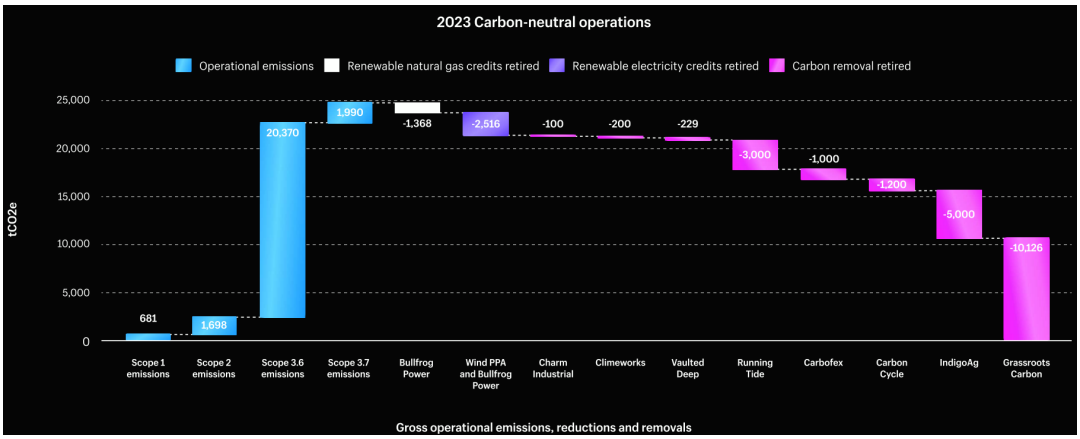


Figure 7: Document reporting GHG emissions data using a bar chart figure.

The figure above shows a case where the report presents the GHG emissions data using a bar chart. Although this is visually rich and easy to comprehend for the human eye, it poses a significant challenge for an Optical Character Recognition (OCR) system to extract the data from the chart. The text may be detected and extracted correctly by the OCR system, but the context of the data will likely be lost, leading to an extraction with high uncertainty. A visual or multimodal approach would be necessary to extract the data from this type of document.

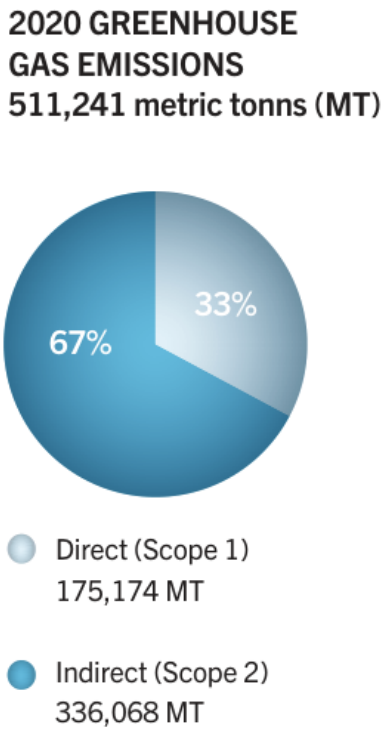


Figure 8: Document reporting GHG emissions data using a pie chart figure.

Similar to figure 7, the figure above shows a case where the report presents the GHG emissions data using a pie chart. In this case, since we would only be interested in the values presented in the text under the image, the OCR-based approach could lead to a successful extraction of the data. However, this is another example that illustrates the diversity of formats companies use to report key information in their financial reports.

GATX Greenhouse Gas (GHG) Emissions
All Global Locations for Rail North America and Rail International

Scope 1 & Scope 2 GHG Emissions

Scope	2019		2020	
	Location-Based Total by Scope (MT CO_2e)	Market-Based Total by Scope (MT CO_2e)	Location-Based Total by Scope (MT CO_2e)	Market-Based Total by Scope (MT CO_2e)
Scope 1 - Direct Emissions	14,258	14,258	15,197	15,197
Scope 2 - Indirect Emissions from Purchased Energy	13,379	12,022	11,321	10,554
Total	27,636	26,279	26,518	25,751

Figure 9: Document reporting GHG emissions data in a table.

One of the most common layouts present in the dataset is the table format. It is an effective way to present structured data over a period in a concise, objective, and easy-to-read manner. This is also one of the best formats for an OCR-based approach to extract the data, as the data is already structured, reducing the likelihood of context loss. However, OCRs are not error-free with tables, especially in cases where the table is structured in a non-conventional way, such as merged cells, rotated text, or cells with no borders. Additionally, Document AI researchers have extensively studied and developed models to detect, extract, and understand tables in documents [49, 3, 17, 10, 50, 51].

3.1 Additional Relevant Data Intricacies

3.1.1 Unit of Measurement

The diversity in such documents is not only evident in the layouts but also in the selection of units reported. Most companies disclose their GHG emissions data in *metric tons of* CO_2 equivalent. However, other companies report their emissions in different units, such as kilograms of CO_2, metric tons of CO_2 per unit of production, and some companies with very large numbers report their emissions in million metric tons of CO_2 equivalent. Since the unit of measurement is crucial for the correct interpretation of the data, it is important to also be able to extract this information from the reports. Therefore, if our system correctly detects the values in the document but fails to output a normalized unit of measurement—in our case, we use metric tonnes—the system output will be considered incorrect and will likely lead to a large residual error since the values will be in different magnitudes.

3.1.2 Scope 2 Emissions Origin

The GHG Scope 2 emissions are divided into three categories: location-based, market-based, and undefined. Location-based emissions are calculated based on the location of the company's operations, market-based emissions are calculated based on the market where the company sells its products, and undefined emissions are those not clearly defined as location-based or market-based. It is important to be able to extract this information from the reports, as it is crucial for the correct interpretation of the data. Therefore, if our system correctly detects the values in the document but fails to output the correct category for the Scope 2 emissions, the system output will be considered incorrect.

4 Strategies for Information Extraction from Financial Reports

Different strategies for extracting information from financial reports have been implemented to compare their effectiveness across diverse report formats and content types. The study brings an approach focused solely on text information, an image-analysis approach that treats the PDF page as an image, and a multimodal approach that combines image and text information to infer the correct indicators disclosed in the document from more than one source of truth. For these experiments, we maintain the same pre-processing steps to ensure that the comparison only takes into account the core feature of extracting data from a given input.

4.1 System Specifications

The experiments proposed in this study were conducted on a machine with the following specifications:

- **CPU**: AMD Ryzen 7 3700X 16 threads at 3.600 GHz
- **Memory**: 32GB at 3200 MHz
- **Operating System**: Linux Manjaro 6.1.55-1
- **Python**: 3.11.5

4.2 Experiments Definition

To set up an information retrieval challenge relevant to our business case and provide an opportunity to compare different strategies for extracting information from financial reports, we establish a simple set of indicators of interest, a desired schema for the extracted data, and a set of metrics to evaluate the performance of the different strategies.

4.2.1 Indicators of Interest

Given Datia's strong presence in the ESG domain, we have chosen to focus on a set of ESG indicators commonly reported in financial statements. Therefore, this challenge focuses on correctly extracting values for the following indicators:

- **GHG Scope 1 emissions**: The total amount of GHG emissions directly produced by a company.
- **GHG Scope 2 emissions**: The total amount of GHG emissions indirectly produced by a company.
  - **Location-based**: Emissions calculated based on the location of the company's operations.

  - **Market-based**: Emissions calculated based on the market where the company sells its products.
  - **Undefined**: Emissions that are not clearly defined as location-based or market-based.

- **GHG Scope 3 emissions**: The total amount of GHG emissions produced in the value chain of a company. Scope 3 emissions can also be broken down into 3.1, 3.3, 3.4, 3.6, 3.7, and 3.11 categories. However, for the purpose of this study, we will only consider the total Scope 3 emissions.

- **Reported Unit**: The unit of measurement used for the emissions.

These indicators are crucial for assessing a company's environmental impact and sustainability practices, and they are often reported in financial statements as part of the company's ESG disclosures. In essence, there are three main indicators analyzed, but since the GHG Scope 2 emissions are divided into three categories, we consider them as separate indicators for this study. It is also important to be able to extract the unit of measurement used for the emissions because different companies may report their emissions in different units, such as metric tons of CO_2 equivalent, kilograms of CO_2, or other units.

4.3 Extracted Data Schema

The extracted data from the financial reports should follow a specific schema to ensure consistency and comparability across different strategies. Since most of the strategies are LLM-based, defining a schema adds complexity to the system, as hallucination could lead to correct data being extracted but in the wrong format. For these reasons, we define a schema that is simple and straightforward, focusing on the key indicators of interest and their values for each year reported. Here is an example of the JavaScript Object Notation (JSON) schema for the extracted data:

```
{
    "metrics": {
        "2022": {
            "scope_1": 88200000.0,
            "scope_2": {
                "location_based": null,
                "market_based": null,
                "undefined": 200000.0
            },
            "scope_3": 38800000.0
        },
        "2023": {
            "scope_1": 75100000.0,
            "scope_2": {
```

```
                "location_based": null,
                "market_based": null,
                "undefined": 200000.0
            },
            "scope_3": 36600000.0
        }
    },
    "extracted_pages": [1, 4, 10],
    "reported_unit": "million_metric_tonnes",
}
```

This schema ensures that the extracted data can be processed by the system and represented in the same unit of measurement, ensuring that the comparison between the different strategies is fair and accurate. The `extracted_pages` field is used to store the page numbers from which the data was extracted, allowing for traceability and validation of the extracted information.

4.4 Evaluation Criteria

To thoroughly assess the performance of the proposed systems for extracting indicators from financial reports, a combination of quantitative and qualitative metrics is employed. These metrics are designed to measure both the accuracy of the extracted data and the robustness of the extraction process against various types of errors. Here, we delineate the key metrics and evaluation criteria used.

4.4.1 Detection Rate

Not only is it important to evaluate the accuracy of the extracted data, but it is also crucial to assess the system's ability to detect the presence of the indicators of interest in the document, as well as avoid detecting false positives. The detection rate is calculated as the proportion of documents in which the system correctly identifies the indicators of interest. In order to assess the detection rate, we utilize Precision, Recall, and Accuracy.

- **Precision** assesses the proportion of data points extracted by the model that are correct and relevant. A high precision rate indicates fewer instances of fabricated metrics and irrelevant data extraction.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{2}$$

- **Recall** measures the system's ability to retrieve all relevant data points from the document. High recall is essential to ensure no significant data is missed.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{3}$$

- **Accuracy** provides an overall measure of the system's performance in detecting the indicators of interest.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}} \tag{4}$$

Ideally, the system should achieve high precision and recall rates to ensure that the extracted data is both accurate and comprehensive. In practice, there is often a trade-off between precision and recall, and the system must be optimized to balance these metrics effectively. In our case, we would prefer to minimize the number of false positives, as those would represent fabricated metrics — perhaps due to LLM hallucinations — thus, we would slightly prioritize precision over recall.

4.4.2 Residual Analysis

While we want a system that has high recall and precision rates to guarantee that the presence of key information is being detected, it is also fundamental that this information is being extracted correctly (or at least as close as possible to the ground truth). For this reason, we also conduct a residual analysis to evaluate how far the extracted values are from the ground truth values. The residual is calculated by taking the absolute difference between the extracted value (`y`) and the ground truth value (`y_hat`).

$$\text{Residual} = |y - \hat{y}| \tag{5}$$

The lower the residual, the closer the extracted value is to the ground truth value, indicating a more accurate extraction process.

4.4.3 Perfect Match Rate

A special case when the metric is correctly detected and also perfectly extracted is what the system ultimately aims to maximize, and in this study, we refer to it as the Perfect Match Rate. This metric is calculated as the proportion of documents in which the system correctly identifies and extracts the indicators of interest without any errors.

$$\text{Perfect Match Rate} = \frac{\text{Perfect Matches}}{\text{Total Documents}} \tag{6}$$

4.4.4 Error Types Analysis

In addition to quantitative metrics, an error analysis is conducted to identify the types of errors made by the system during the extraction process.

- **Missing Data**: Indicates data that should have been extracted but was not. This error impacts the `recall` metric.

- **Fabricated Metrics**: Metrics that are not present in the document but appear in the extracted output — also referred to as extrinsic hallucination. This error impacts `precision`.
- **Incorrect Values**: Reflects errors in the value of the data extracted. These are numerical errors that impact the `residual` metric.

4.5 Shared Processing Steps

For all of the experiments, the system receives a PDF as input, and common pre and post-processing steps are applied to the data to ensure that the system is able to extract the information from the reports.

4.5.1 Pre-processing: Finding Pages of Interest

A shared pre-processing step is to find the pages of interest — those that might contain the information that we are looking for — in the PDF document.

```
def extract_data_from_pdf(pdf_file: str):
  # Load the PDF file
  doc = fitz.open(pdf_file)

  for page in doc:
    # extract text from page
    text = page.get_text("text")

    # function that will look for matches
    # for the indicators of interest
    matches = search_for_keywords(text)

    # this page does not contain
    # any relevant information
    if matches is None:
        return

    # Then different strategies follow...
```

This pre-processing step is used to iterate over the pages in the PDF document and find the pages of interest so that the core engine is only applied to these pages. This is an important step to reduce the processing time and avoid unnecessary processing of pages that do not contain relevant information. However, it is also crucial to ensure that no false negatives are generated, as this could lead to missing pages of interest.

4.5.2 Parsing LLM Outputs

As of the time of writing, the text GPT-4 model has a parameter flag to determine the output format of the model, while the vision model does not have this feature. This

means that for text, it is possible to specify JSON as the output format. However, for the vision model, the output is always a string, which leads to uncertainty about whether the output string can be parsed into a JSON format by the system. Additionally, the parameter flag does not guarantee that the output JSON will contain the correct keys and meet the desired schema. For these reasons, it is important to have a parsing step that validates the output of the LLM models and ensures that the extracted data is in the expected format.

This is where the *pydantic* and *instructor* libraries come into play. We can use *pydantic* to define the model schema and validate fields and types, and *instructor* will handle any imperfections in the model by retrying the generation process n times and validating the output against the defined schema.

4.5.3 Post-processing: Consolidating Information from Different Pages

In an ideal scenario, all the information we are looking for would be contained on a single page of the PDF document, allowing our system to find the page, extract the data, and return it. However, this is not the case in practice, as different companies use varying designs to display their information, sometimes spreading the indicators across multiple pages or even repeating them in different sections of the report. Since our system needs to resolve conflicts in scenarios where multiple pages contain the same information (potentially with different values due to local regulations, units, and so forth), a strategy for consolidating the information is required. After testing various strategies like frequency bags, selecting the output with the most found values, and LLMs, we decided to use the LLM-based approach, leveraging the GPT-4 model to consolidate the information from different pages. Here is a demonstration of how this can be done:

```
def consolidate_data(data: list[str]) -> str:
  """
  Consolidate data from different pages

  Args:
    data: List of JSON strings containing the data
          extracted from different pages

  Returns:
    consolidated_data: Unique JSON string containing
                       the consolidated data
  """
  # Send request with data and instructions to GPT-4
  consolidated_data = openai.chat.completions.create(
    model="gpt-4-turbo-preview",
      response_format={"type": "json"},
      messages=[
        {
```

```
            "role": "system",
            "content": """Consolidate the following data
                          into only one JSON string...""",
        },
        {"role": "user", "content": data},
      ]
    )
  return consolidated_data
```

This is a simple example of how the GPT-4 model can be used to consolidate information from different pages of the PDF document. This process can be further improved by adding more context and validation steps, such as checking for inconsistencies in the data before returning the consolidated information.

4.6 Text-Only Approach with LLMs

In this experiment, we use the GPT-4 architecture, specifically *gpt-4-0125-preview* released in December 2023, to extract information from the text contained in the financial reports without considering any images or visual data. Evidently, this approach is limited to the information present in the text and does not account for any charts or important visual cues that the report might contain. Therefore, it is expected that this system may fail in cases where the indicators are presented in a visual form. This is also the cheapest approach in terms of processing time and cost, as it only requires the GPT-4 model to process the text data. Depending on the use case, if this approach performs sufficiently well, it could be the most cost-effective solution for extracting information from financial reports.

This approach can also be replaced by the usage of RAG systems where the information from the document is stored in a database and the system can perform queries using embeddings to find the most similar documents, or pages, to the query

4.7 Image-Only Approach with LLMs

In this experiment, we use the GPT-4 Vision model to extract information from the images contained in the financial reports without considering any text data. Since OpenAI's visual model is highly accurate in extracting information from images, this approach is expected to have an overall performance superior to the text-only approach because it can generalize the information from images. However, it is also expected to have an increased propensity for hallucinations, as the model cannot validate the information extracted from the images with the text data.

For this experiment, after the preprocessing step, the system applies the following image-processing transforms to each of the pages of interest:

- Dots Per Inch (DPI): The image is set to 300 DPI to ensure that the model can extract the information with high quality. This setting has been shown to be one of the optimal settings for OCR tasks [52].

- Grayscale: The image is converted to grayscale to reduce the amount of information that the model needs to process.
- Downscaling: The image is downscaled to fit OpenAI's GPT-4 Vision criteria:

  > *...images are first scaled to fit within a 2048 x 2048 square, maintaining their aspect ratio. Then, they are scaled such that the shortest side of the image is 768px long [53].*

- Compression: The image is compressed to a 90% quality to significantly reduce the size of the image while maintaining sufficient quality for image analysis.

These transforms aim to reduce the amount of information that the model needs to process while maintaining the quality of the information extracted from the images. This is important not only for model performance but also to keep the number of tokens processed lower, reducing the cost of the operation.

The system sends a request to OpenAI's GPT-4 Vision model with the preprocessed image and instructions for the model to extract the key indicators from the image into a predefined JSON schema. This is necessary and important because not only does the model need to be able to extract the information from the images, but it also needs to produce the information in a format that can be parsed by the system, otherwise, it might raise unexpected errors on the client side.

4.8 Multimodal LLMs to Extract Information from Images and Text

In this experiment, we use a multimodal approach to extract information from the financial reports, leveraging information from both textual and visual domains. This approach combines the strengths of the text-only and image-only methods while mitigating their respective weaknesses by validating the information extracted from the images with the text data and vice versa.

The system follows the same preprocessing steps as the image-only approach before sending the image and instructions to the GPT-4 Vision model as described in 4.7. The difference in this experiment is that the system also extracts the text from the page and creates a set of all the numbers mentioned in the text, which will be used to validate whether the information provided by the model is consistent with the text data or if it is an extrinsic hallucination.

To mitigate intrinsic and extrinsic hallucinations, we implement a simple lookup method that searches for the closest value in the text data to the one extracted from the image, within a defined threshold limit. Values within the threshold are considered valid and are readjusted, while those that fall outside the threshold are considered hallucinations and are discarded.

This system is expected to have the best performance among the three approaches, as it can validate the information extracted from the images with the text data and vice versa. However, it is also expected to have the highest processing time due to the increased complexity of the system.

To address the observed hallucination behavior — particularly errors in the last digits of numeric values — the system could benefit from additional validation checks. For example, implementing a rule that flags unusually large discrepancies in the last digits could help mitigate such errors, ensuring greater accuracy in the final extracted values.

5 Results

In this section, we explore the results of the experiments conducted to extract information from financial reports using different strategies, assessing the metrics defined in Section 4.4. The results are presented in tabular format, showing the performance of each system in terms of detection rate, residual analysis, perfect match rate, and error type analysis. Charts are also used to ease the interpretability of the results.

5.1 Detection Rate

The detection rate is a crucial metric for evaluating the system's ability to identify the indicators of interest in the financial reports while also accounting for the presence of false positives, which could lead to fabricated indicators.

Approach	Indicator	Recall	Precision	Accuracy
Text Only	Scope 1 Emissions	0.97	0.90	0.88
	Scope 2 Emissions	0.05	0.33	**0.74**
	Scope 2 Emissions Location	0.91	0.62	0.63
	Scope 2 Emissions Market	**1.00**	0.48	0.81
	Scope 3 Emissions	0.91	0.78	0.85
Vision Only	Scope 1 Emissions	0.90	0.87	0.81
	Scope 2 Emissions	**0.08**	0.33	0.68
	Scope 2 Emissions Location	0.82	**0.64**	**0.64**
	Scope 2 Emissions Market	0.85	0.38	0.77
	Scope 3 Emissions	0.77	0.83	0.83
Multimodal	Scope 1 Emissions	**0.98**	**0.95**	**0.94**
	Scope 2 Emissions	0.05	**1.00**	0.69
	Scope 2 Emissions Location	**0.97**	0.60	0.63
	Scope 2 Emissions Market	0.92	**0.67**	**0.89**
	Scope 3 Emissions	**0.93**	**0.90**	**0.92**

Table 1: Metrics for different approaches.

In terms of recall, the multimodal approach demonstrates the highest recall rates for most indicators, particularly *Scope 1 emissions* and *Scope 3 emissions*, where it achieves 0.98 and 0.93, respectively. The text-only approach, while generally reliable, struggles with indicators that involve visual elements, such as *Scope 2 emissions*, where its recall drops significantly to 0.05. The image-only approach performs similarly to the text-only approach in some areas, but it is generally less consistent, with recall rates varying widely across different indicators.

Precision is another area where the multimodal approach excels. For most indicators, including *Scope 1 emissions* and *Scope 3 emissions*, the multimodal strategy achieves the highest precision, reaching 0.95 and 0.90, respectively. This indicates that the multimodal approach is not only effective in identifying relevant data but also minimizes the extraction of incorrect or irrelevant information. Notably,

the precision for *Scope 2 emissions* under the multimodal approach reaches 1.00, suggesting that when the system does detect this indicator, it is highly accurate.

Accuracy, which provides an overall measure of performance, is consistently highest for the multimodal approach across most indicators. For *Scope 1 emissions* and *Scope 3 emissions*, the accuracy reaches 0.94 and 0.92, respectively, reinforcing the superiority of the multimodal method in extracting accurate data. The text-only approach, while accurate for certain indicators, suffers in cases where visual data is critical, as seen in its lower accuracy for *Scope 2 emissions*.

The findings suggest that in a real-world scenario, where the goal is to either expedite the extraction of information from financial reports or automate the process entirely, the multimodal approach presents the most promising strategy. Its ability to integrate and cross-validate text and visual data makes it particularly well-suited for handling the diverse and complex nature of financial documents.

5.2 Perfect Match Rate

5.2.1 Text-only

This experiment focused on extracting the information based only on text data from the document pages. Since some documents present the information in a structured way, the text-only approach was able to extract the information with a high precision and recall rate, as shown in Table 2. However, as previously discussed in Section 3, there are documents that present visually rich information, such as charts and tables, which might be challenging for the text-only approach to extract the information correctly. These documents will often produce a mismatched prediction, confusing different metrics (e.g., assigning the value of Scope 1 emissions to Scope 3 emissions or vice versa). Because of the nature of the indicators being analyzed, these metrics can often be in different orders of magnitude, which can lead to high residuals when the system makes a mistake in the extraction process. Therefore, we see that the residual values are high for all the indicators, indicating that the system is not able to extract the information correctly in some cases. However, the perfect match rate is mostly above 70% for all the indicators, indicating that the system was able to correctly predict the indicators in a high proportion of the documents.

Indicator	Metric	Mean	Std	Min	25%	50%	75%	Max
Scope 1 Emissions	residual	621589	2617186	0	0	0	301	16657326
	perfect match	70%						
Scope 2 Emissions	residual	42701	227480	0	0	0	72	1534000
	perfect match	73%						
Scope 2 Emissions Location	residual	1093202	6180676	0	0	0	12355	49315635
	perfect match	53%						
Scope 2 Emissions Market	residual	358920	2260835	0	0	0	0	16997182
	perfect match	79%						
Scope 3 Emissions	residual	785786	5287354	0	0	0	2	46199954
	perfect match	74%						

Table 2: Residual and perfect match metrics for the text-only approach (rounded).

5.2.2 Image-only

The image-only approach focuses on extracting information from the PDF document pages, treating them as images. This approach is expected to perform better than the text-only approach for documents that contain visually rich information, such as charts and tables, as the model is able to interpret the visual nuances of the document. However, it is also expected to have a higher rate of hallucinations, as the model is not able to validate the information extracted from the images with the text data.

Indicator	Metric	Mean	Std	Min	25%	50%	75%	Max
Scope 1 Emissions	residual	1537568	12489496	0	0	0	2435	116455366
	perfect match	51%						
Scope 2 Emissions	residual	28025	167960	0	0	0	601	1534000
	perfect match	67%						
Scope 2 Emissions Location	residual	1608247	12893555	0	0	28	5243	119880861
	perfect match	45%						
Scope 2 Emissions Market	residual	127890	1142411	0	0	0	5	10718989
	perfect match	72%						
Scope 3 Emissions	residual	859218	5387119	0	0	0	215	46199954
	perfect match	66%						

Table 3: Residual and perfect match metrics for the image-only approach (rounded).

From the results above, it is notable that the visual model underperforms compared to the text-only approach all the metrics. This could be due to the fact that rich-visual information is not present in most documents and therefore this model would outperform the text-only approach in a subset of the documents present in the dataset while underperforming in the majority of the documents.

5.2.3 Multimodal

The multimodal approach combines the information extracted from the text and images to predict the indicators of interest from the financial reports. This approach is expected

to have the best performance among the three approaches, as it is able to use the information from both the text and images to validate the information extracted from the images with the text data and vice versa.

Indicator	Metric	Mean	Std	Min	25%	50%	75%	Max
Scope 1 Emissions	residual	752	3601	0	0	0	0	25030
	perfect match	89%						
Scope 2 Emissions	residual	8431	32266	0	0	0	83	239308
	perfect match	69%						
Scope 2 Emissions Location	residual	8681	30929	0	0	0	4834	239308
	perfect match	57%						
Scope 2 Emissions Market	residual	585	2194	0	0	0	0	12560
	perfect match	88%						
Scope 3 Emissions	residual	227871	1169000	0	0	0	0	8254000
	perfect match	85%						

Table 4: Residual and perfect match metrics for the multimodal approach (rounded).

The results confirm that the multimodal approach outperforms the text-only and image-only approaches in almost all the evaluated metrics by a significant margin. Additionally, most indicators have 75% of the documents with a residual of 0, indicating that the system was able to extract the information correctly in most of the cases while also holding a perfect match rate above 85% for metrics *Scope 1 emissions*, *Scope 2 emissions market* and *Scope 3 emissions*.

5.2.4 How the Approaches Compare

When analyzing the perfect match rates for each indicator in each experiment strategy, it is clear that the multimodal approach outperforms the text-only and image-only approaches, which perform similarly, with a slight advantage for the text-only approach. Considering the scopes that do not have subdivisions, it is observable that there is a considerable improvement in the multimodal approach against the text approach, reaching a 19% improvement in the perfect match rate for the *Scope 1 emissions* indicator and an 11% improvement for the *Scope 3 emissions* indicator. In a context where this system is being assessed for use in a real-world scenario — whether to expedite the process of extracting information from financial reports or to automate the process — the multimodal approach appears to be the right strategy to fine-tune and build upon.

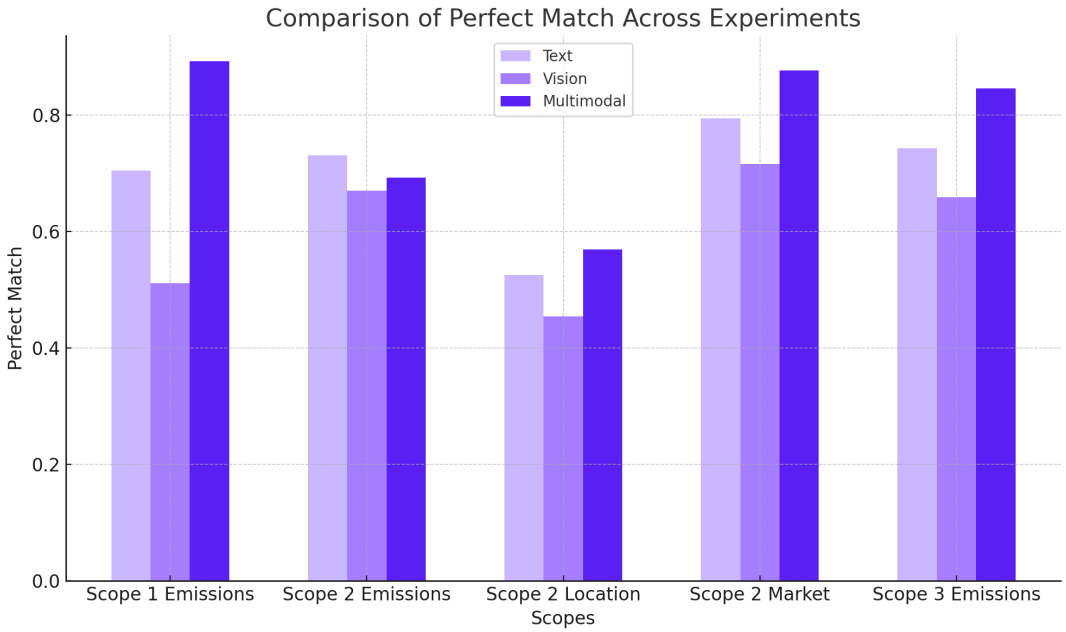


Figure 10: Results of the experiments for the different approaches.

5.3 How Significant Are the System's Errors?

While the perfect match rate provides a meaningful and straightforward way to compare the experiments, it is crucial to understand the severity of the errors when a perfect match is not achieved. For instance, consider a scenario where a company reports its *Scope 1 emissions* as 300,000 MT CO_2, but the system extracts this value as 300,020 MT CO_2. Although this is not a perfect match, the error is so small that it could be considered negligible in some cases. The sensitivity to such errors may vary depending on the specific indicators and the business context in which the extracted data is used.

The charts below represent the residuals for each indicator in each experimental strategy, arranged from the highest to the lowest residual value. Ideally, the best-performing system would have a curve closer to the origin, indicating smaller errors, while the worst-performing system would have a curve closer to the top-right corner of the chart.

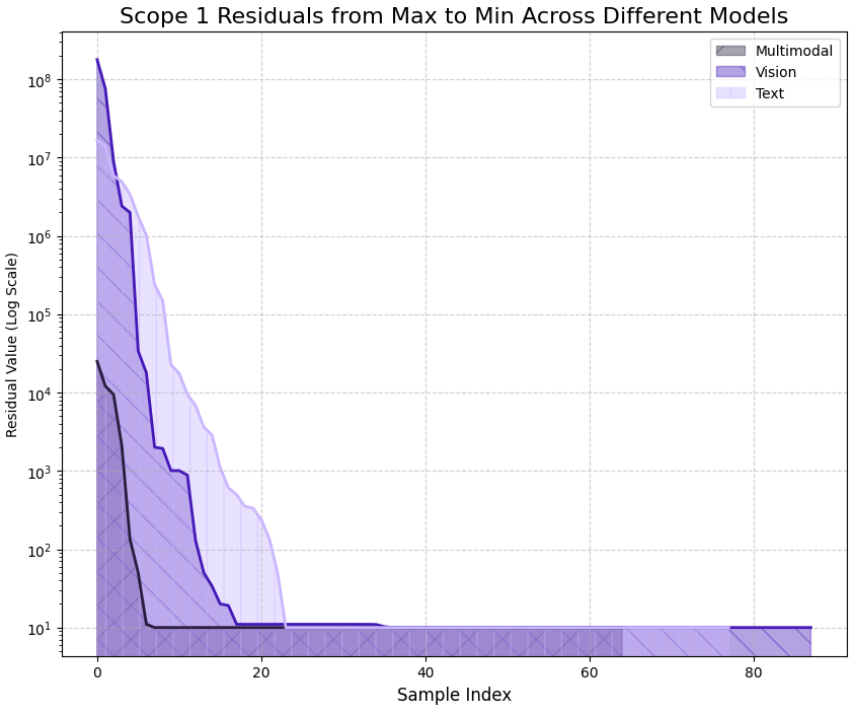


Figure 11: Scope 1 emissions residual

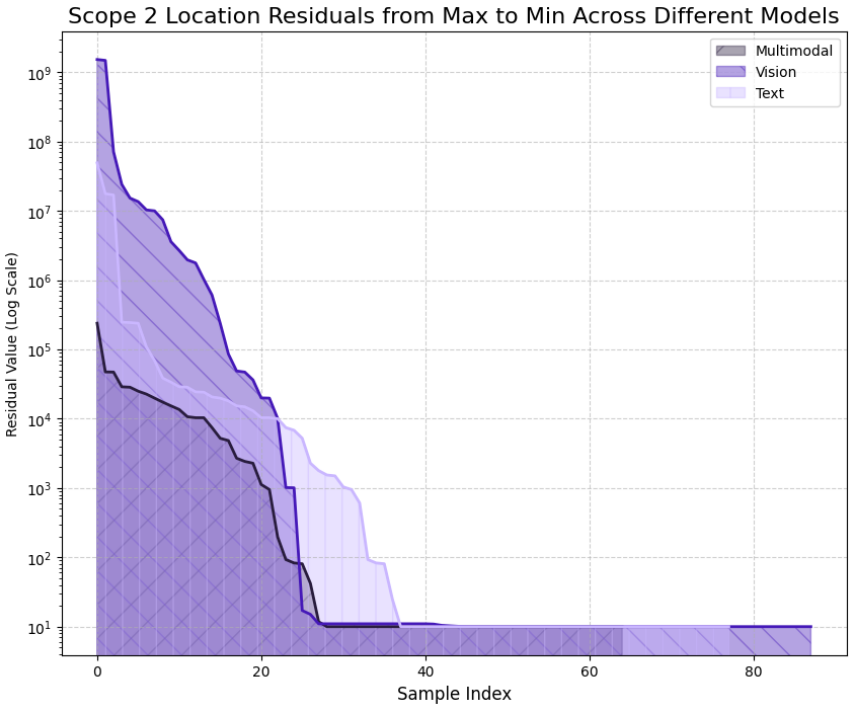


Figure 12: Scope 2 emissions location residual

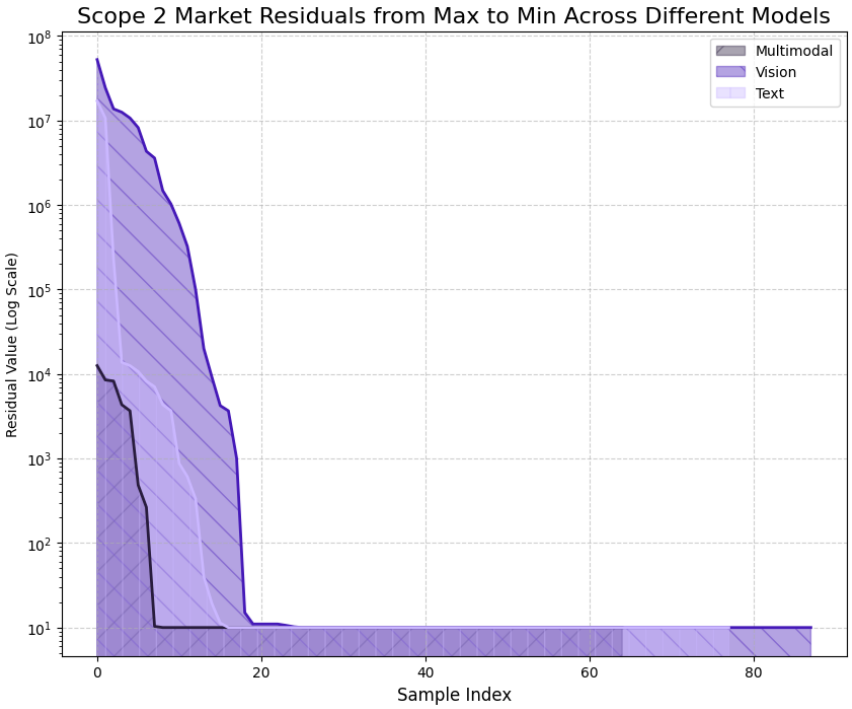


Figure 13: Scope 2 emissions market residual

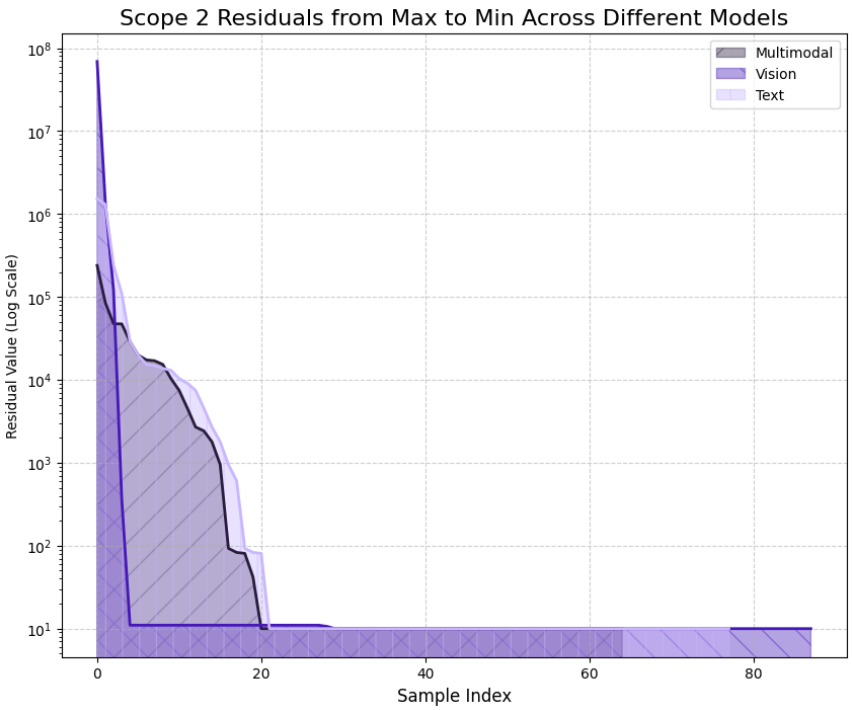


Figure 14: Scope 2 undefined emissions residual

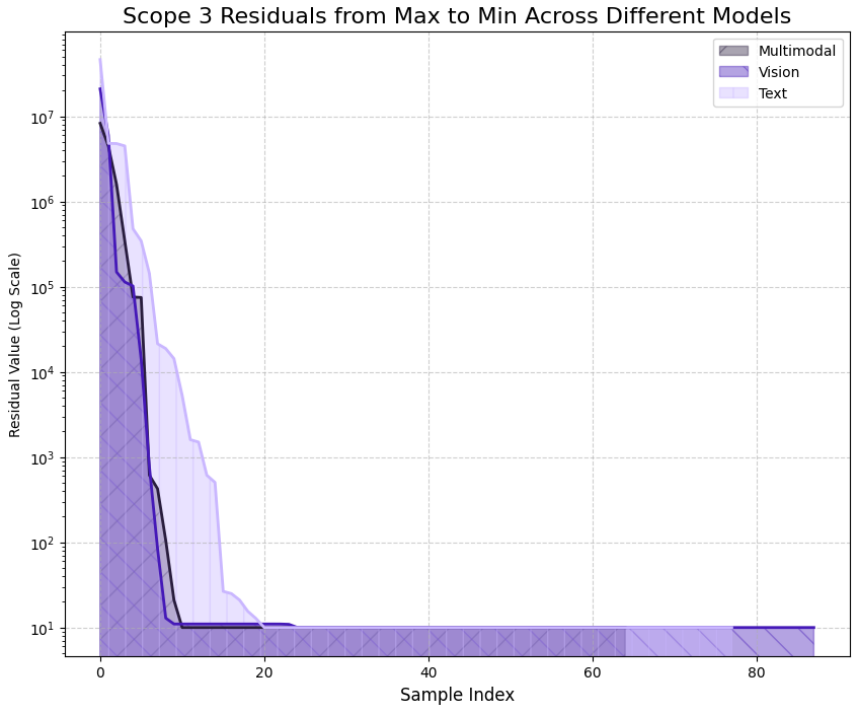


Figure 15: Scope 3 emissions residual

A closer examination of the residuals reveals that the vision model produces samples with higher residuals than the text model, despite the text model generally having a higher perfect match rate. This observation is interesting as it indicates that these models make different types of errors. If the primary goal is to maximize the perfect match rate, the text model outperforms the vision model. However, if minimizing residuals is the main objective, the vision model demonstrates better performance.

Overall, the multimodal approach offers the best performance in terms of residuals. In the case of the *Scope 2 Residuals* indicator chart, the vision model seems to produce fewer samples with high residuals, but this is due to the fact that the vision model predicted fewer samples for this indicator.

For the *Scope 3 Residuals* indicator chart, the vision model performs very closely to the multimodal model, highlighting a weaker performance by the text model for this particular indicator. This is likely due to the loss of information that the text model experiences compared to the vision model. Since the *Scope 3 emissions* indicator often has the largest values — due to encompassing all indirect emissions — the residuals tend to be higher for this indicator if there is confusion with another indicator.

5.4 Considerations

All experiments show poorer performance for the sub-indicators under *Scope 2 emissions* compared to the other indicators. This occurs because *Scope 2 emissions* are divided into three categories: location-based, market-based, and undefined (when the

document does not clearly specify the measurement standard). The system struggles to differentiate between these categories, leading to higher residuals and lower perfect match rates for these indicators.

In many cases, what is considered a mistake may actually be a correct extraction, but the labeling might be unclear. For instance, the model may infer that Company A is reporting 300, 349 MT CO_2 as location-based emissions when the document does not specify the nature of the emissions, and the label refers to it as undefined. This type of error falls into a gray area, where the labeled dataset would need to be revised to ensure that the nature of the emissions is perfectly labeled. However, due to time constraints, this will not be possible in this study.

The metrics and labels reported for Scopes 1 and 3 should not suffer from this uncertainty, making the results more reliable for these indicators and more suitable for comparison between the different approaches.

6 Conclusions

This thesis investigated several approaches for extracting GHG emissions from financial reports using LLMs as the core engine for predictions. The primary goal was to evaluate the effectiveness of text-only, image-only, and multimodal strategies in identifying key environmental metrics from complex financial documents. Each method was assessed based on detection rate, residual analysis, and perfect match rate, providing insights into the strengths and weaknesses of each approach.

The text-only approach performed well, especially when handling structured data like tables and text-based indicators. It achieved high precision and recall for most indicators, but struggled with visually rich elements like charts and complex tables, leading to higher residuals and mismatches. This highlights the challenges of relying solely on text-based methods for visually complex documents.

The image-only approach excelled at interpreting visual data but underperformed overall. While effective at extracting data from images, it suffered from hallucinations and errors due to the lack of textual validation. This method's weaknesses were particularly evident as financial documents often present visual information alongside textual descriptions, with the latter providing more reliable data for extraction.

The multimodal approach, which combines text and image data, proved to be the most effective. By leveraging both sources of information, it achieved the highest accuracy, precision, and recall, with perfect match rates exceeding 85% for key indicators. It also performed well in residual analysis, minimizing errors by cross-validating between text and images, making it the most robust solution for ESG data extraction.

Despite the multimodal approach's success, challenges remain. Differentiating sub-indicators under Scope 2 emissions revealed classification ambiguities, suggesting that refining dataset labeling or validation methods could improve extraction accuracy.

In conclusion, this thesis demonstrates the potential of LLMs in automating ESG data extraction from financial documents. The findings emphasize the importance of multimodal approaches for achieving reliable results in complex document analysis. Future research could focus on better preprocessing techniques, such as identifying key tables and infographics, to reduce irrelevant information. Expanding the scope of indicators and refining evaluation metrics would further advance the understanding of LLMs's capabilities in this field. These insights lay the groundwork for future developments in automated financial report analysis, with significant implications for business and regulatory compliance.

References

[1] I. Finkel and J. Taylor, *Cuneiform*, ser. Ancient scripts. J. Paul Getty Museum, 2015. ISBN 9781606064474. [Online]. Available: https://books.google.com.br/books?id=cf7NrQEACAAJ

[2] Adobe Systems Incorporated, "Pdf timeline," https://www.adobe.com/acrobat/resources/pdf-timeline.html, 2023, accessed: 2024-04-20.

[3] D. Pinto, A. McCallum, X. Wei, and W. B. Croft, "Table extraction using conditional random fields," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, ser. SIGIR '03. New York, NY, USA: Association for Computing Machinery, 2003. doi: 10.1145/860435.860479. ISBN 1581136463 p. 235–242. [Online]. Available: https://doi.org/10.1145/860435.860479

[4] F. Peng and A. McCallum, "Information extraction from research papers using conditional random fields," *Information Processing & Management*, vol. 42, no. 4, pp. 963–979, 2006. doi: https://doi.org/10.1016/j.ipm.2005.09.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306457305001172

[5] H. Fang, T. Tao, and C. Zhai, "A formal study of information retrieval heuristics," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '04. New York, NY, USA: Association for Computing Machinery, 2004. doi: 10.1145/1008992.1009004. ISBN 1581138814 p. 49–56. [Online]. Available: https://doi.org/10.1145/1008992.1009004

[6] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, p. 613–620, nov 1975. doi: 10.1145/361219.361220. [Online]. Available: https://doi.org/10.1145/361219.361220

[7] S. K. Wong, W. Ziarko, V. V. Raghavan, and P. C. Wong, "On modeling of information retrieval concepts in vector spaces," *ACM Trans. Database Syst.*, vol. 12, no. 2, p. 299–321, jun 1987. doi: 10.1145/22952.22957. [Online]. Available: https://doi.org/10.1145/22952.22957

[8] M. E. Maron and J. L. Kuhns, "On relevance, probabilistic indexing and information retrieval," *J. ACM*, vol. 7, no. 3, p. 216–244, jul 1960. doi: 10.1145/321033.321035. [Online]. Available: https://doi.org/10.1145/321033.321035

[9] Z. Wang, Y. Xu, L. Cui, J. Shang, and F. Wei, "Layoutreader: Pre-training of text and layout for reading order detection," 8 2021. [Online]. Available: http://arxiv.org/abs/2108.11591

[10] J. Li, Y. Xu, T. Lv, L. Cui, C. Zhang, and F. Wei, "Dit: Self-supervised pre-training for document image transformer," in *Proceedings of the 30th ACM International Conference on Multimedia*, ser. MM '22. New York, NY, USA: Association for Computing Machinery, 2022. doi: 10.1145/3503161.3547911. ISBN 9781450392037 p. 3530–3539. [Online]. Available: https://doi.org/10.1145/3503161.3547911

[11] M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, and F. Wei, "Trocr: transformer-based optical character recognition with pre-trained models," in *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. doi: 10.1609/aaai.v37i11.26538. ISBN 978-1-57735-880-0. [Online]. Available: https://doi.org/10.1609/aaai.v37i11.26538

[12] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, "Layoutlm: Pre-training of text and layout for document image understanding." Association for Computing Machinery, 8 2020. doi: 10.1145/3394486.3403172. ISBN 9781450379984 pp. 1192–1200.

[13] V. Bush, "As We May Think," *Atlantic Monthly*, vol. 176, no. 1, pp. 641–649, March 1945. doi: 10.1145/227181.227186. [Online]. Available: http://www.theatlantic.com/doc/194507/bush

[14] L. Cui, Y. Xu, T. Lv, and F. Wei, "Document ai: Benchmarks, models and applications," 11 2021. [Online]. Available: http://arxiv.org/abs/2111.08609

[15] N. Subramani, A. Matton, M. Greaves, and A. Lam, "A survey of deep learning approaches for OCR and document understanding," *CoRR*, vol. abs/2011.13534, 2020. [Online]. Available: https://arxiv.org/abs/2011.13534

[16] X. Yang, E. Yumer, P. Asente, M. Kraley, D. Kifer, and C. L. Giles, "Learning to extract semantic structure from documents using multimodal fully convolutional neural networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. doi: 10.1109/CVPR.2017.462 pp. 4342–4351.

[17] S. Schreiber, S. Agne, I. Wolf, A. Dengel, and S. Ahmed, "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)*, vol. 01, 2017. doi: 10.1109/ICDAR.2017.192 pp. 1162–1167.

[18] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," in *Psychological Review*, ser. PR '58. Cambridge, MA, USA: MIT Press, 1958. ISBN 9780262181114 pp. 386–408.

[19] ——, "Principles of neurodynamics: Perceptrons and the theory of brain mechanisms," in *Neural Networks*, ser. NN '61. Cambridge, MA, USA: MIT Press, 1961. ISBN 9780262181114 pp. 386–408.

[20] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, ser. IEEE '98. Washington, DC, USA: IEEE, 1998. ISBN 1558605529 pp. 2278–2324.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, ser. NIPS '12. Red Hook, NY, USA: Curran Associates Inc., 2012. ISBN 9781450319895 pp. 1097–1105.

[22] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[23] S. Hochreiter, "Untersuchungen zu dynamischen neuronalen Netzen," *Diploma Thesis*, July 1991. [Online]. Available: http://www.idsia.ch/~juergen/SeppHochreiter1991ThesisAdvisorSchmidhuber.pdf

[24] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 11 1997. doi: 10.1162/neco.1997.9.8.1735. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017. ISBN 9781510860964 p. 6000–6010.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

[27] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020.

[28] Y.-C. Chen, Y.-S. Liu, Z. Kira, and G. AlRegib, "Pre-trained image processing transformer," *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 3530–3539, 2021.

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 3530–3539, 2021.

[30] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, "A survey on vision transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 87–110, 2023. doi: 10.1109/TPAMI.2022.3152247

[31] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. S. Chatterji, A. S. Chen, K. A. Creel, J. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. E. Gillespie, K. Goel, N. D. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. F. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. S. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. P. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. F. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. H. Roohani, C. Ruiz, J. Ryan, C. R'e, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. P. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. A. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang, "On the opportunities and risks of foundation models," *ArXiv*, 2021. [Online]. Available: https://crfm.stanford.edu/assets/report.pdf

[32] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi, "From recognition to cognition: Visual commonsense reasoning," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. doi: 10.1109/CVPR.2019.00688 pp. 6713–6724.

[33] R. Martín-Martín, M. Patel, H. Rezatofighi, and et al., "JRDB: A Dataset and Benchmark of Egocentric Robot Visual Perception of Humans in Built Environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 6748–6765, 2023. doi: 10.1109/TPAMI.2021.3070543

[34] OpenAI, "Gpt-4v(ision) system card," 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:263218031

[35] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, E. Hesse, H. Wang, J. Dorado, M. Park, J. Foo, E. Steorts, and I. Sutskever, "Learning transferable visual models

from natural language supervision," *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 3530–3539, 2021.

[36] OpenAI, "Gpt-4 technical report," 3 2023. [Online]. Available: http://arxiv.org/abs/2303.08774

[37] G. T. et al., "Gemini: A family of highly capable multimodal models," 2024. [Online]. Available: https://arxiv.org/abs/2312.11805

[38] ——, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," 2024. [Online]. Available: https://arxiv.org/abs/2403.05530

[39] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," 2022. [Online]. Available: https://arxiv.org/abs/2203.02155

[40] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen, "A survey of large language models," 2023. [Online]. Available: https://arxiv.org/abs/2303.18223

[41] Z. Xu, S. Jain, and M. Kankanhalli, "Hallucination is inevitable: An innate limitation of large language models," 2024.

[42] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, no. 12, mar 2023. doi: 10.1145/3571730. [Online]. Available: https://doi.org/10.1145/3571730

[43] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020. ISBN 9781713829546

[44] H. Chase, "LangChain," Oct. 2022. [Online]. Available: https://github.com/langchain-ai/langchain

[45] J. Liu, "Instructor: Structured LLM Outputs," Jun. 2023. [Online]. Available: https://github.com/jxnl/instructor

[46] S. Colvin, E. Jolibois, H. Ramezani, A. G. Badaracco, T. Dorsey, D. Montague, S. Matveenko, M. Trylesinski, S. Runkle, D. Hewitt, and A. Hall, "Pydantic," 3 2024. [Online]. Available: https://docs.pydantic.dev/latest/

[47] D. Castelvecchi, "Can we open the black box of ai?" *Nature News*, vol. 538, no. 7623, p. 20, 2016.

[48] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, aug 2018. doi: 10.1145/3236009. [Online]. Available: https://doi.org/10.1145/3236009

[49] S. Paliwal, V. D, R. Rahul, M. Sharma, and L. Vig, "Tablenet: Deep learning model for end-to-end table detection and tabular data extraction from scanned document images," 2020. [Online]. Available: https://arxiv.org/abs/2001.01469

[50] A. Gilani, S. R. Qasim, I. Malik, and F. Shafait, "Table detection using deep learning," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 01, 2017. doi: 10.1109/ICDAR.2017.131 pp. 771–776.

[51] F. Shafait and R. Smith, "Table detection in heterogeneous documents," in *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, ser. DAS '10. New York, NY, USA: Association for Computing Machinery, 2010. doi: 10.1145/1815330.1815339. ISBN 9781605587738 p. 65–72. [Online]. Available: https://doi.org/10.1145/1815330.1815339

[52] W. Bieniecki, S. Grabowski, and W. Rozenberg, "Image preprocessing for improving ocr accuracy," in *2007 International Conference on Perspective Technologies and Methods in MEMS Design*, 2007. doi: 10.1109/MEMSTECH.2007.4283429 pp. 75–80.

[53] "Openai vision api documentation," https://platform.openai.com/docs/guides/vision, OpenAI, 2024, accessed: 2024-04-11.