Erasmus University Rotterdam
Erasmus School of Economics
Master Thesis Data Science and Marketing Analytics

Exploring the effectiveness of Large Language Models in greenwashing detection

Mattia Fornasiero (573735)

Supervisor:	M. van de Velden
Second assessor:	MG. de Jong
Date final version:	August 16, 2024

1 Abstract

This study investigates the potential of Large Language Models (LLMs) to detect greenwashing in marketing claims, a deliberate corporate action aimed at deceiving stakeholders regarding a company's sustainability efforts. To evaluate the effectiveness of LLMs in identifying greenwashing, a framework named Green Lantern was developed, combining Chain of Thought reasoning with Retrieval-Augmented Generation (RAG). The framework's performance was tested against simpler baselines. Various LLMs, including OpenAI's GPT-4 and Google's Gemini-1.5, were tested across different frameworks. The results indicate that while LLMs yield promising results, by performing better than the random chance baseline, their performance varies significantly based on the framework used. The simpler Single Agent with Retrieval framework outperformed others in terms of accuracy and F1 scores, highlighting the importance of framework design. The Green Lantern framework, when paired with GPT-4o, achieved an accuracy of 0.58. However, limitations such as biases and over-reliance on pre-training data were observed in the simpler frameworks, suggesting that LLMs are not yet fully reliable for greenwashing detection. Although the Green Lantern framework proposed in this study performed worse compared to simpler baselines, it still outperformed a random chance baseline and demonstrated greater robustness than its simpler counterparts. This study contributes to LLM-based fact-checking and provides the first structured dataset for greenwashing detection, emphasizing the need for further research to enhance LLM reliability in this domain.

2 Introduction

The rise in public concern about the environmental impact of products and services, combined with the lower cost of capital in more sustainable companies, has driven many corporations to adopt greener marketing strategies. However, this shift is not without drawbacks. A recent study reported that 53% of the claims in the EU market have been found to provide vague, misleading, or unfounded information (Commission, 2022). Moreover, a survey for executives published by Google Cloud revealed that 58% of executives agree that green hypocrisy exists and that their organizations have overstated their sustainability efforts (Google, 2022). In the same report, only 36% of respondents indicated that their organizations have measurement tools in place to quantify their sustainability efforts. Practices such as overstating sustainability efforts and making claims without quantifying actual performance fall under the umbrella term "greenwashing." Although this term lacks a universally agreed-upon definition, it generally refers to deliberate corporate actions that include misleading elements, with the intent to deceive stakeholders (de Freitas Netto et al., 2020).

With regards to green marketing, regulators are lagging behind large corporations. Only in recent years have the EU and US legislated on the matter. However, given the complex nature of greenwashing and the volume of marketing claims made by companies, it can be challenging for regulators to promptly identify greenwashing and take action.

Since 2023, the introduction of Large Language Models (LLMs) has led to significant advancements in various Natural Language Processing (NLP) areas. This technology, characterized by its relatively low cost, is being adopted by numerous businesses worldwide for automation purposes (Media, 2023). LLMs are a type of foundational model trained on a broad set of unlabeled data and can be applied to various tasks with minimal fine-tuning (IBM, 2023a). LLMs are designed to process and generate text like a human, in addition to other forms of content (IBM, 2023b).

The goal of this research is to investigate whether large language models (LLMs) can effectively identify greenwashing by fact-checking green marketing claims. The framework proposed in this study, named Green Lantern, utilizes a chain-of-thought technique, which involves a series of intermediate reasoning steps (Wei et al., 2022). Additionally, Green Lantern allows the model to access external information from sustainability reports. This research aims to serve as an initial exploration of LLM capabilities in the field of greenwashing detection.

2.1 Research Question

The issue of greenwashing is relevant for various stakeholders, including regulators, policy-makers, corporations, and society as a whole. Greenwashing is complex to identify by nature, as it involves active misleading practices. The recent advancements in the LLMs and their proven ability to perform well in fact-checking tasks (Wei et al., 2024) make them a potentially suitable solution for detecting greenwashing.

Developing a framework that can assess marketing claims or assist humans in the evaluation process could increase the volume of processed claims and help protect stakeholders from the negative effects of greenwashing. Therefore, the aim of this study is to investigate the potential role of LLMs in detecting greenwashing in marketing claims.

To better frame the problem, the primary research question for this study is:

"How effective are pre-trained LLMs at detecting greenwashing in marketing claims?"

This primary question can be further divided into two sub-questions, which will help in thoroughly addressing the main question:

- What is the accuracy of LLMs in predicting whether a claim is greenwashed or not?
- How do different frameworks impact the effectiveness of LLMs in detecting greenwashing?

3 Related Work and Background

The following literature review aims to explore the issue of greenwashing, the reasons behind its implementation, and its costs and benefits. Furthermore, background information on Large Language Models will be provided, focusing on the technical structure of these models and their components, and explaining the most common techniques to enhance their performance and limit their shortcomings. The final part investigates the main technologies and methodologies can be employed to identify greenwashing in marketing claims, such as LLM based fact-checking and Sustainability Report Analysis.

3.1 Greenwashing

The term *greenwashing* is an umbrella term that refers to the deceptive practices of companies conveying a misleading impression or providing false information about their environmental efforts or the sustainability of their products or services. It is important to note that there is no officially agreed-upon definition. Therefore, in recent years, numerous studies have attempted to establish a more rigorous definition of this term.

A systematic review by de Freitas Netto et al. (2020) finds that most studies on greenwashing use the definitions by TerraChoice and the Oxford Dictionary, where the phenomenon is described as a *deliberate corporate action with misleading elements, focused on deceiving stakeholders*. Moreover, the systematic review identifies an alternative definition of the term as proposed by Seele and Gatti (2017), which highlights the need for an accusatory element for greenwashing to occur. The formal definition proposed is the *presence of an external accusation toward an organization with regard to presenting a misleading green message*. The definition from Seele and Gatti (2017) will be the one adopted for this study.

In summary, greenwashing can be considered as a type of corporate hypocrisy. Shifting the focus from the actual sustainability performance to how the claim is communicated and perceived (Balluchi et al., 2020).

3.1.1 Forms of greenwashing

As reported in de Freitas Netto et al. (2020), various frameworks exist to define greenwashing, each proposing different forms in which it can occur. The marketing firm TerraChoice, now part of UL Solutions, developed a framework known as the *7 Sins of Greenwashing* to help identify companies that engage in greenwashing (Solutions, 2007). This framework includes

issues such as *environmental claims lacking proof*, *vague or misleading terms* (e.g. "all-natural"), *false endorsements*, *irrelevant claims* (e.g. promoting CFC-free products when CFCs are already banned), *the lesser of two evils* claims that distract from broader impacts, and the *sin of fibbing*, which consists of outright false claims.

de Freitas Netto et al. (2020) integrated these insights with those from other studies to define two main classifications of greenwashing: *Claim Greenwashing*, which involves misleading textual arguments about environmental benefits, and *Executional Greenwashing*, where nature-evoking elements create false perceptions of a brand's environmental friendliness.

According to Carlson et al. (1993), there are different ways in which *Claim Greenwashing* can be carried out. In their study, the researchers identified two attributes of green claims: claim type and claim deceptiveness. See Tables 1 & 2.

Claim Type	**Description**
Product Orientation	Claims centering on the ecological attribute of a product
Process Orientation	Claims centering on the ecological high performance of a production process technique, and/or an ecological disposal method
Image Orientation	Claims centering on enhancing the eco-friendly image of an organization, like claims that associates an organization with an environmental cause or activity with elevated public support
Environmental Fact	Claims that involve an independent statement that is factual in nature from an organization about the environment at large, or its condition
Combination	Claims having two or more of the categories above

Table 1: Claim Types

Claim Deceptiveness	**Description**
Vague/Ambiguous	Claims that are overly vague, ambiguous, too broad, and/or lacking a clear definition
Omission	Claims missing the necessary information to evaluate its validity
False/Outright Lie	Claims that are inaccurate or a fabrication
Combination	Claims having two or more of the categories above
Acceptable	Claims that do not contain a deceptive feature

Table 2: Claim Deceptiveness

3.1.2 Reasons behind greenwashing

In recent years there has been a substantial amount of research that investigates the impact of green marketing and its advantages for organizations. As investigated by Papadas et al. (2019), green marketing positively affects competitiveness and consequently rewards companies with an improved financial performance. Nyilasy et al. (2014) found that the current shift in customer behaviour seems to be favouring brands and companies that position themselves as more sustainable.

The motivation behind greenwashing in marketing claims can be identified in a corporation's desire to access the so-called *green premium* thatcustomers are willing to pay, without having to endure the high costs of actually reducing the environmental impact of their products or services (Zhang et al., 2021) (Lee and Raschke, 2023).

With regards to financial benefits that could drive corporate greenwashing, although it is true that there is a positive correlation between greenwashing companies and lower debt cost, Attig et al. (2021) found that creditors, and private lenders in particular, tend to employ complex pricing structures to mitigate the effect of greenwashing, discouraging the practice. This suggests that the primary motivation behind greenwashing may be the desire to position the brand in a favorable manner, rather than the access to direct financial benefits.

3.1.3 Costs of greenwashing

The topic of green marketing has been studied thoroughly since the beginning of the new millennium (Bhardwaj et al., 2023)(Kumar, 2016). However, its negative effects and the impact of greenwashing have received much less attention. Greenwashing poses several risks for businesses, the most significant being reputational damage and a loss of consumer trust. Zhang et al. (2018) investigated the effect of green word-of-mouth and found that it had a strong negative effect on customer purchase intentions. Generally, as reported by Yang et al. (2020) in a systematic review, *when greenwashing occurs, it will harm the interests of not only consumers but also society as a whole, despite offering significant benefits to existing stakeholders.*

Moreover, greenwashing practices contribute to the issue of green skepticism (Leonidou and Skarmeas, 2017), which can diminish the positive effects of green marketing, not only for companies that actively engage in this practice, but for the market as a whole.

Another potential risk is incurring legal costs resulting from customer class actions or breaches of regulations, particularly within the EU with the introduction of the "Green Claims" directive (Tank, 2024). Recent US rulings (Ferris et al., 2023) have demonstrated

the complexity surrounding greenwashing, highlighting a lack of regulations that makes it challenging to bring cases to court. Following these rulings, the Federal Trade Commission has begun working on a new set of rules to better regulate the US market (FTC, 2022).

3.2 Large Language Models (LLMs)

Large Language Models are designed to process and generate text like a human, in addition to other forms of content (e.g. images, videos), based on the vast amount of data used to train them (IBM, 2023b). LLMs can be referred to as *foundational* models, which are models trained on a broad set of unlabeled data that can be used for different tasks, with minimal fine-tuning (IBM, 2023a).

Large Language Models work by predicting the next token based on a given input. In the case of text generation a token is a unit of text that the model processes, which can be a word, subword, or character. When working with other forms of content, such as images, a token can represent a pixel or a group of pixels. To predict the next token, the model generates a list of probabilities for each possible token and then samples from these probabilities to determine the next token. By repeating this process, an entire sentence or paragraph can be generated.

Given their design, LLMs are able to perform a variety of complex natural language processing tasks, such as machine translation, text summarization and question answering (Radford et al., 2019).

Pre-trained LLMs quickly became a suitable solution for automating simpler business processes worldwide (Media, 2023) thanks to their ability to generate human-like text. LLMs can be used to address a wide range of applications, such as as assisting with programming tasks, generation of marketing content and copy writing. The most recent pre-trained LLMs include GPT-3.5 and GPT-4 developed by OpenAI (Achiam et al., 2023) and Gemini 1.0 and Gemini 1.5 developed by Google (Team et al., 2023).

3.2.1 Technical overview

The aim of the following explanation is to provide a high-level overview of how the Generative Pre-Trained Transformer (GPT), a type of large language model, works. GPTs are predecessors and simpler versions of more advanced models, such as GPT-3 and GPT-4. Models like Google's Gemini 1.0 and 1.5 operate in a similar fashion. GPTs are based on the transformer, a neural network architecture introduced by Vaswani et al. (2017). However, to fully grasp the processes of the transformer, it is necessary to present the concept of attention.

Attention Mechanisms: This concept was introduced in the context of machine translation by Bahdanau et al. (2014), significantly improving the performance of neural machine translation systems. It enables a neural network to make use of contextual information, text that appears before or around the selected token, for predicting the current token.

At a high level, attention mechanisms work by assigning an attention score to each token in the context. The score represents the importance of each token in relation to the current token being predicted. The attention scores are used to create a weighted sum of the token's embeddings. With an embedding being a representation of a text element in a high-dimensional vector space, used to capture semantic meaning. The attention mechanism applies these scores to the embeddings through a weighted sum, producing a new context vector that better captures the relevant information needed to predict the current token (Anthropic, 2021).

The attention mechanism is typically used in sequence-to-sequence settings, such as translation tasks. In the latest LLMs, more advanced mechanisms based on attention, such as self-attention or multi-headed attention, are used.

The Transformer: Introduced by Vaswani et al. (2017), the transformer is a neural network architecture composed of layers that utilize self-attention mechanisms and feedforward neural networks. The key innovation brought by the transformer is the use of self-attention mechanisms, which, unlike traditional attention mechanisms, allow the model to consider other words in the same sentence or sequence to understand the relationships between them (Vaswani et al., 2017). The self-attention mechanism is typically used to capture relationships within a single sequence or sentence. These mechanisms enable the model to access all elements of the input sequence simultaneously. This addition allows the model to capture long-range dependencies and relationships within the sequence, leading to a better understanding of context.

Moreover, self-attention mechanisms enable the parallel processing of input data rather than sequential processing. This parallelization allows transformers to handle much larger datasets and learn more complex patterns than previous models, such as recurrent neural networks (RNNs) and long short-term memory (LSTM) networks (Anthropic, 2021). In practice, a transformer takes the embeddings of the input text as input, processing them through multiple transformer layers, which are composed of:

- *Multi-head self-attention*: Multiple instances of the self-attention mechanism are implemented in parallel, allowing the model to focus on different aspects of the sequence simultaneously.

- *Feedforward Neural Networks*: The output for each token is then passed through a feedforward neural network. This non-linear transformation helps further process the information.

- *Layer Normalization*: After each layer of the transformer, normalization is applied to stabilize the training process.

After processing the input data, the transformer outputs a vector that is passed through a softmax function, an operation that converts the output into a vector of probabilities for each possible token being next. The model can then sample among the possible outputs and finally decode the results, outputting a token. Temperature, a parameter influencing the balance between predictability and creativity in generated text, is used in the softmax function: when the value is set to zero, the token with the highest probability will be selected, and when temperature has higher values, the probabilities will become more balanced.

A transformer can therefore be seen as a neural network with many layers, each containing tunable parameters, such as the weights of the feedforward neural networks and the attention mechanisms. These tunable parameters are what allow the model to learn during the training phase.

3.2.2 Retrieval Augmented Generation (RAG)

LLMs have shown limitations in tasks requiring specific domain knowledge (Kandpal et al., 2023), such as lower quality of generated answers. Moreover, the limited size of the context window, which is the amount of tokens that can be processed in a single instance, is not large enough to process whole documents.

Retrieval-Augmented Generation (RAG), introduced by Lewis et al. (2020), is a technique that improves LLM performance on knowledge-intensive tasks. This technique consists of embedding text coming from external sources, such as PDF files, into vector form and storing it in a vector database, a database that indexes and stores vector embeddings for fast retrieval.

When a query is made, the documents that are most similar to the query, are retrieved from the database. To find the most similar documents, the distance between the embedded query and embedded documents is utilized, therefore the documents that have the minimum distance amongst each other in the high dimensional embedding can be said to be most similar. The goal of this step is to provide only relevant parts of the report as input for the LLM.

This approach was found to improve the performance of LLMs in knowledge intensive tasks by reducing the size of the input (Li et al., 2024), by only selecting the necessary

documents.

3.2.3 Chain of Thought (CoT)

Wei et al. (2022) explored the idea of *Chain of Thought*, dividing a task in a series of intermediate reasoning steps. The research found that this approach significantly improves the ability of LLMs in arithmetic, commonsense, and symbolic reasoning tasks.

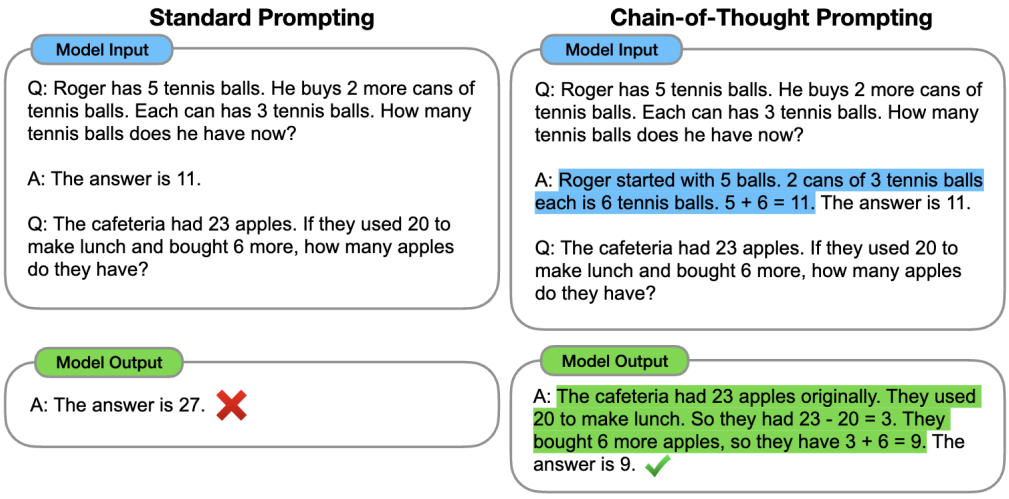

Figure 1: Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. Chain-of-thought reasoning processes are highlighted (Wei et al., 2022)

An example of Chain of Thought approach is shown in Figure 1, where the Chain-of-Thought step is added within the input, mimicking a chat between a human and an agent. However, an alternative way of implementing Chain-of-Thought prompting is that of adding multiple LLMs in a sequence with preset instructions, which will be the technique used in this study.

3.2.4 Reasoning and LLMs

Reasoning in humans can be defined as the process of thinking about something in a logical and systematic way, using evidence and past experiences to get to a conclusion or make a decision (Wang et al., 2023) (McHugh and Way, 2018).

The term *reasoning* in large language models has often been misused as there is no clear definition of what it refers to. In the literature, the term *reasoning* is often used to refer to the simulation of informal reasoning (Huang and Chang, 2022). Informal reasoning is

defined as a less structured approach to reasoning used where strict formal rules may not apply. Informal reasoning mostly relies on intuition, experience, and common sense.

Given their probabilistic nature, as they sample from a distribution of likely outputs, LLMs cannot perform traditional human reasoning. However, these models demonstrated to be able to replicate arithmetic, symbolic and commonsense reasoning and achieve a satisfying performance in primitive reasoning tasks (Wei et al., 2022).

Although these models are not able to *think*, they have shown the ability to perform well in tasks that would usually require human reasoning and strategy tasks. As shown by Wei et al. (2022) and Zhang et al. (2024) breaking down a complex problem in smaller parts and using multiple LLMs to obtain a result, improves the performance on tasks where human reasoning is traditionally needed. In this study, the term *reasoning* will be used to refer to the ability of LLMs to replicate informal reasoning.

3.2.5 Hallucinations

Hallucinations in the context of LLMs generally refer to responses that are not consistent with human cognition and facts (DeHaven and Scott, 2023a). Another conventional classification of LLM hallucinations distinguishes between intrinsic hallucinations, which occur when an LLM's output contradicts the input provided by the user, and extrinsic hallucinations, which occur when LLM outputs cannot be verified by the information in the input (Zhang et al., 2023) (Huang et al., 2021). Further studies explore different types of hallucinations and their underlying causes; however, for the scope of this study, the definition provided by DeHaven and Scott (2023a) is sufficient, and intrinsic and extrinsic hallucinations will be considered equivalent.

Hallucinations occur due to the nature of large language models which, as highlighted by Azamfirei et al. (2023), follow a probabilistic approach in text generation and therefore solely rely on patterns learned during training.

Techniques to avoid hallucinations or reduce their frequency are currently being developed. The most notable examples consist in inserting specific instructions in the prompt as shown by Xu et al. (2023), and in the implementation of *self-reflection* to allow the model to detect and correct its own possible hallucinations (Ji et al., 2023a) (Ji et al., 2023b).

3.3 LLM-based Greenwashing Detection

Recently, there has been increasing interest in applying Large Language Models for detecting greenwashing (EY, 2024). However, so far, LLM effectiveness has not been tested in greenwashing detection tasks in an academic setting. However, LLMs effectiveness has been

tested in tasks that are closely related to that of greenwashing detection: Sustainability Report Analysis and Fact-Checking.

3.3.1 LLM-based Sustainability Report Analysis

The development and introduction of advanced LLMs, such as OpenAI GPT3.5 and GPT4, brought researchers to investigate the possibility of applying these technologies to automate the analysis of sustainability related documentation. Two recent studies have focused on this issue:

ChatREPORT: is a LLM-based system to automate the analysis of corporate sustainability reports (Ni et al., 2023). Domain experts were tasked with evaluating the sustainability reports, providing high-quality data for the model training. The ChatReport framework allows for Report Embedding, Report Summarization, Customized Question Answering, and Task Force on Climate-related Financial Disclosures (TFCD) Conformity assessment. TFCD is a set of guidelines for non-financial disclosures, such as sustainability reports. In the study the risk of hallucinations addressed by attaching source numbers to retrieved chunks, allowing human experts to efficiently check whether the model produces misinformation.

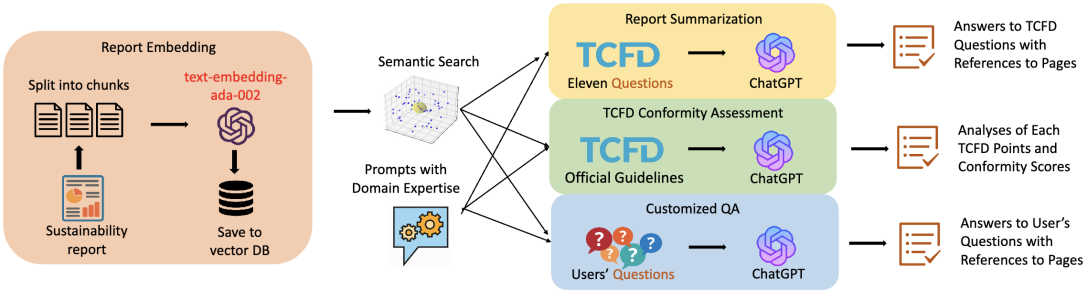

Figure 2: ChatReport Pipeline (Ni et al., 2023)

ESGReveal: This research stems from the development of ChatREPORT and it mainly focuses on sustainability on structured extraction of relevant data from the sustainability report. ESGReveal includes an ESG metadata module for criteria queries, a report preprocessing module for building databases, and an LLM agent module for structured data extraction. The structure followed by the framework is that of Global Reporting Initiative and TFCD, which are widely used guidelines for sustainability reporting. The model demonstrated an accuracy of 76.9% in data extraction and 83.7% in disclosure analysis (Zou et al., 2023) respectively, an increase of over 20% compared to baseline.

3.3.2 LLM-based Fact-Checking

LLMs proved to be useful in the field of fact-checking, especially thanks to their ability to retrieve relevant text and to *reason*, using a Chain of Thought approach (Wei et al., 2022). A study by Hoes et al. (2023), found that ChatGPT accurately categorized statements in 72% of cases, with significantly greater accuracy at identifying true claims (80%) than false claims (67%), compared with a random guess baseline. Starting from these results, more studies have been conducted in the field of fact checking, mainly focusing on political issues:

FactGPT: Choi and Ferrara (2024) introduces a framework tailored to automate the claim matching phase of fact-checking. The study is based on a dataset of social media posts related to public health. To evaluate various models performance in claim matching, the researchers employed a *textual entailment task*. This type of tasks consists of instructing the model to categorize relationships between pairs of statements into three classes: Entailment, Neutral, and Contradiction. FACT-GPT showcases remarkable performance, with the best performing model correctly classifying 73% of the observations. The framework was not tested on other datasets, therefore the only baseline available is random chance, which was outperformed by almost 40%. These findings highlight the possible complementary role of LLMs alongside human expertise for fact checking applications.

Self-Checker: Li et al. (2023) introduces a framework for fact-checking by guiding LLMs in a context where the model is not provided with any example and with few examples. This approach is particularly beneficial in low-resource settings, Self-Checker presents efficient mechanism for constructing fact-checking systems.
This framework performance was compared with that of models fine-tuned specifically for fact-checking tasks, on the BingCheck, FEVER and WiCe datasets. The performance of Self-Checker was better than simpler baselines, such as single LLM agent, and chain of thought, but resulted worse than more complex baseline models, such as BEVERS (DeHaven and Scott, 2023b) . However, the framework showed room for improvement in different areas as the framework tested had minimal fine-tuning, showcasing 56% on FEVER and 71% on WiCe, which are two commonly used datasets for fact-checking tasks.

Although the aforementioned fact-checking frameworks reach a satisfying performance in specific settings, specifically that of political fact checking and LLM answer verification, these techniques cannot address the complexities related to greenwashing detection, as they are unable to access the necessary information to form a fact-based decision.

4 Study Overview

As highlighted in the literature review, a research gap was identified in studies researching the performance of LLMs in detecting greenwashing. This study aims to address this gap by integrating recent findings from two research areas that can possibly be complementary: the use of LLMs in fact-checking and sustainability information retrieval techniques.

Regarding the first research subquestion, a specialized dataset will be developed to evaluate the accuracy of LLMs in this domain, and the performance of four different models will be explored. Additional robustness checks will be conducted to provide more comprehensive results and better address the original research question. A lack of robustness in the results would suggest that the model is not suitable for greenwashing detection.

For the second research subquestion, three distinct frameworks will be tested to determine the impact of various techniques. Two baseline frameworks will be implemented: one using a single LLM instance to generate responses, both with and without Retrieval-Augmented Generation (RAG). The primary framework, however, will incorporate both RAG and Chain of Thought (CoT) reasoning. This approach is designed to simulate intermediate reasoning steps while enabling the model to retrieve external information from sustainability reports.

5 Data collection

Due to the absence of a publicly available dataset on greenwashing in marketing claims, the data was collected manually, visiting company websites and retrieving past marketing campaigns, and published on GitHub to facilitate future research. As greenwashing is a complex issue, the data collection process required a robust structure: a clear definition of greenwashing, a balanced dataset with a sufficient amount of observations.

The dataset uses a binary scoring system, categorizing claims as either greenwashed or not, given the challenge of accurately assessing the degree of greenwashing without expert consultation. A claim is classified as greenwashed only in the presence of an accusation from a reputable NGO, such as ClimateEarth, or a governmental authority, such as the ASA. To be classified as not greenwashed, a claim needs to be supported by two elements, a third party certification related to the claim and the lack of formal greenwashing accusations.

Moreover, the sustainability report relevant to the company and year of each claim, needs to be extracted. This will be the source of information used by the model to certify the veracity of the claim.

The number of retrieved observations is 90, with 48 identified as greenwashed and 42 as

not greenwashed. The observations were manually collected online and the selected marketing campaigns included both large and smaller corporations and were conducted prior to December 2023 to ensure the availability of related sustainability reports.

For each observation, 7 key elements were retrieved:

- **Claim**: the specific environmental or sustainability claim made by the company.
- **Company name**: name of the company that made the claim.
- **Year**: year in which the claim was made.
- **Url**: the link to the advertisement, news article or ruling where the claim can be found.
- **Accusation** in case the claim is greenwashed, a brief description of why the claim is greenwashed.
- **Company Description** a brief description of the company
- **Third-party certifications** a list third-party certifications related to the claim, such as B-Corp or FSC.

5.0.1 Pre-processing of external information

A sustainability report discloses non-financial performance policies, methodologies, and metrics to stakeholders, including investors, employees, customers, and the public (IBM, 2024). In the context of this study, the information contained in the sustainability reports is crucial for accurately evaluating whether a claim is greenwashed or not. To achieve this result Retrieval Augemented Generation will be applied, as shown in Section 6.3.1.

As explained in Section 3.2.2, RAG is a technique that allows to retrieve similar documents in a database to the one queried, based on semantic similarity. For this reason it is essential for the information to be stored in a structured way. To ensure the correct functioning of the technique, the reports need to be split into smaller parts, embedding them in a vector format, and storing them in a vector database. These smaller parts of the report will be referred to as *documents* for the rest of this study.

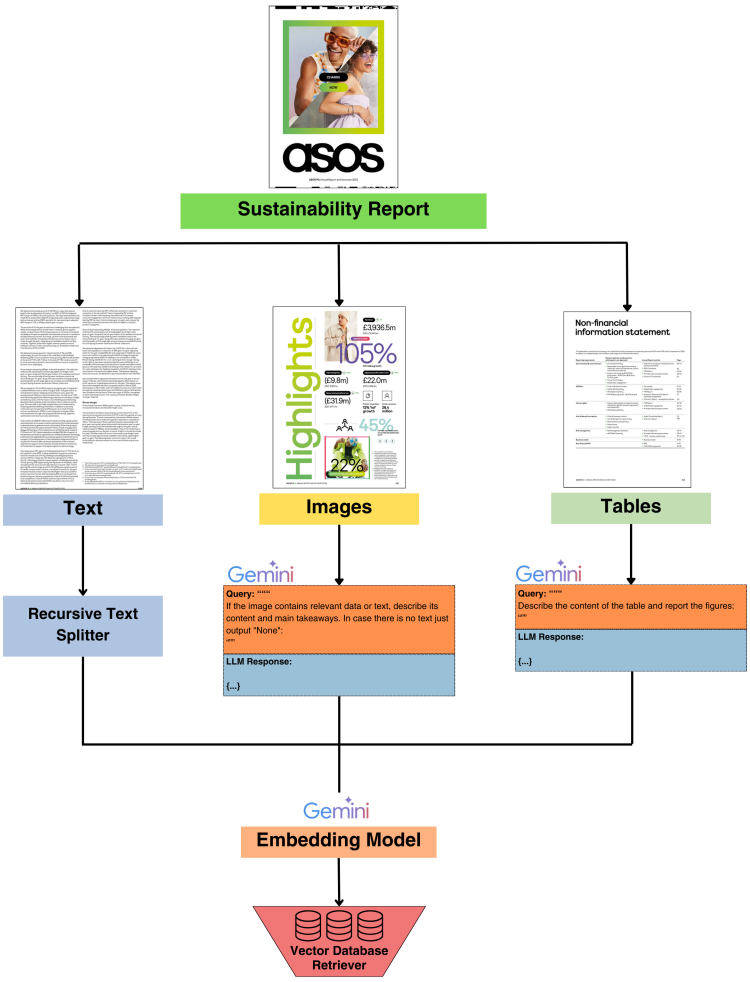

Figure 3: Sustainability Report Pre-processing

Starting from the previously collected data, a sustainability report from the relevant year and company was retrieved for each claim. All the reports are in PDF format and contain text, images, and tables. Images and tables contain a significant amount of information, especially since many companies now make use of infographics and tables.

Text object were extracted using a text-splitter, while non-text objects, such as images and tables, were processed using Gemini 1.5 Flash. This model was tasked with extracting text from tables and describing the images provided. This specific LLM model was chosen for its speed and cost efficiency.

The last step of the pre-processing consists of the embedding of all extracted elements, now in text format, using the Google Embeddings model-004 the text is embedded into a vector and successively stored in a FAISS database, which is a type of vector database developed by Meta.

6 Methodology

6.1 Key Concepts

Before exploring the methodology, a few key concepts need to be clarified:

Framework: In the context of LLMs a framework refers to the structure that organizes different Large Language Models and external components. The same framework can be used with different backbone models.

Agent: an agent refers to a software entity that performs tasks autonomously. Agents can process inputs, make decisions, and execute actions without human intervention (Cheng et al., 2024). In the context of this study, the term Agent refers to an LLM that is prompted with specific instructions.

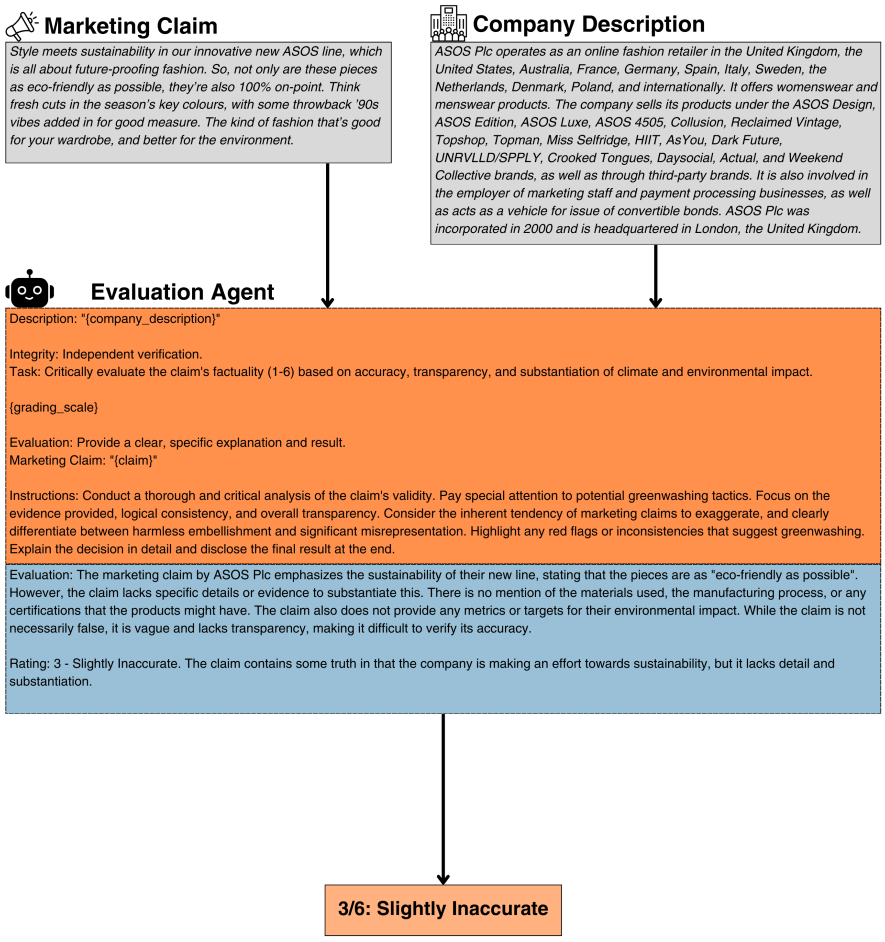

Figure 4: Example of Single Agent using ASOS claim. LLM input is colored in orange and the output is colored in blue.

6.2 Baseline Frameworks

To compare and evaluate the results of the main framework proposed in this paper, two baseline models will be implemented:

Single agent - No external information: shown in Fiugre 4the first baseline framework will be a single agent, that will be tasked to evaluate the claim, given the company description and guidelines to evaluate a claim, all in the same prompt. The scoring system used and the specific format of the output is explained in Section 6.3.3.

Single agent - Retrieval: The second baseline framework, shown in figure 5 also involves a single agent. The LLM will be tasked to evaluate the claim, given the company description and guidelines to evaluate the claim. However, in this case some background information will be provided to the model; more specifically the claim itself will be used to retrieve documents in the vector database. All the information will be served in a single prompt, the scoring system used and the format of the output is reported in Section 6.3.3.

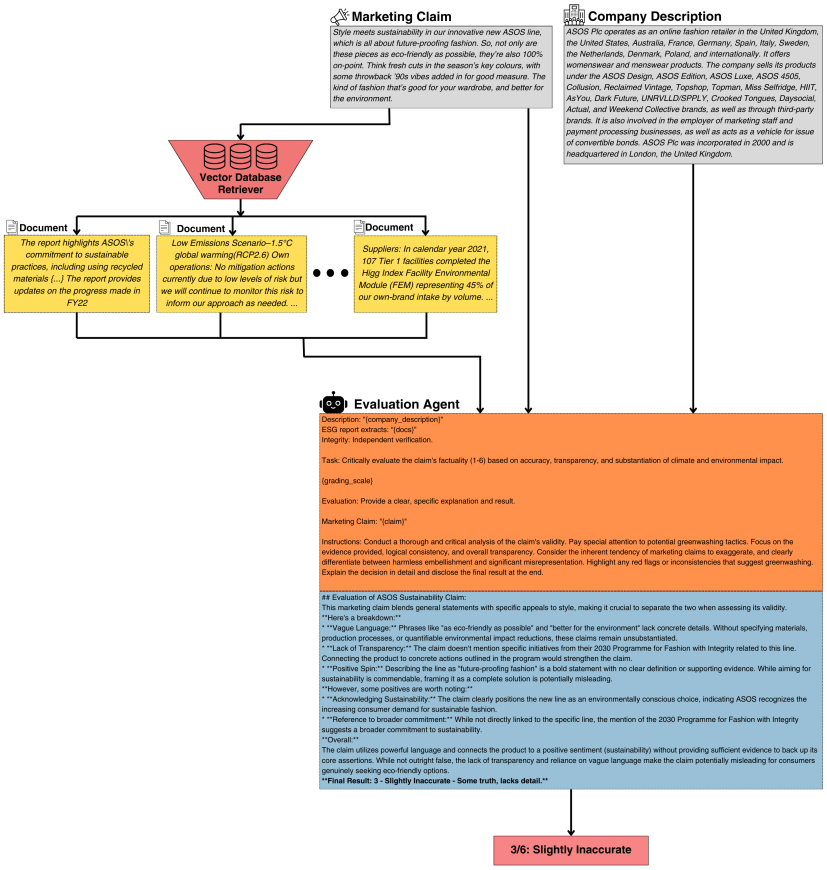

Figure 5: Example of Single Agent - Retrieval using ASOS claim. LLM input is colored in orange and the output is colored in blue.

Both the baseline models, are provided with the data collected in Section 5, and in the case of the Single Agent with Retrieval, the vector database is the one created in the pre-processing phase in Section 5.0.1. In this case, the marketing claim is directly used as a query to the vector database, with the documents retrieved being the semantically closest documents to the marketing claim itself.

6.3 Proposed Framework: Green Lantern

6.3.1 Pipeline

The framework proposed in this paper is a multi-agent system named **Green Lantern**, which combines two commonly used techniques to improve performance in classification tasks when using LLMs: Chain of Thought (CoT) and Retrieval Augmented Generation (RAG). The goal of this approach is to break down claims into smaller, more manageable parts that are easier to verify. The implementation of this technique also allows for the retrieval of relevant documents, providing the evaluation agent with all the necessary and pertinent information to form a correct decision.

The pipeline of Green Lantern contains six steps:

First Evaluation Agent: The first agent is tasked with making an initial evaluation by predicting the claim type and claim focus. Claim type refers to the structure of the claim and can be one of the options listed in Table 1, while claim focus indicates whether the claim is about a specific product or the organization.

MiniCheck Agent: This technique, proposed by Tang et al. (2024), is specific to fact-checking settings. It involves using LLMs to subdivide a claim into individual facts that need to be checked to establish the veracity of the claim.. The agent will generate a maximum of three individual facts derived from the initial claim.

Question Generating Agent: This technique involves using an LLMs to generate multiple similar questions that can be used as input for the retriever. This agent generates two questions for each individual fact generated in the previous step, to help retrieve useful information from the vector database.

Retrieval: For each question generated in the previous step, three individual passages will be retrieved from the vector database. These text passages, each approximately 1,000 tokens in length and unique, will be passed on to the *Document Evaluation Agent*. This approach ensures the context is not overly long, as longer contexts are associated with worse performance in reasoning tasks (Li et al., 2024).

Document Evaluation Agent: In this penultimate step, the agent evaluates the re-

trieved information to determine whether the questions produced by the Question Generating Agent are answered in the provided documents.

Final Evaluation Agent: Lastly, an agent is prompted with instructions on how to evaluate greenwashing. This agent considers the claim, company description, questions, and evaluations made in previous steps. The specific format of the output is explained in Section 6.3.3.

The pipeline of the proposed framework is illustrated using an example in Figure 6, and the complete prompts are reported in the appendix.

6.3.2 Implementation Details

The baseline frameworks and proposed framework were ran using four different models, OpenAI's **"gpt-4o"** and **"gpt-4-turbo"** and Google's **"Gemini-1.5-flash"** and **"Gemini-1.5-pro"**. Each model was used to run all the steps.

The Chain-of-Thought, in the form of intermediate reasoning steps was implemented using multiple LLM instances. The framework was implemented using LangChain, an online tool and python library, while the retrieval was performed using a FAISS database, a vector database developed by Meta . The temperature, the LLM parameter that controls the level of predictability in the output, was set to 0 for all the models, to ensure consistency.

The prompts for all the frameworks can be found starting from Table 10 in the Appendix.

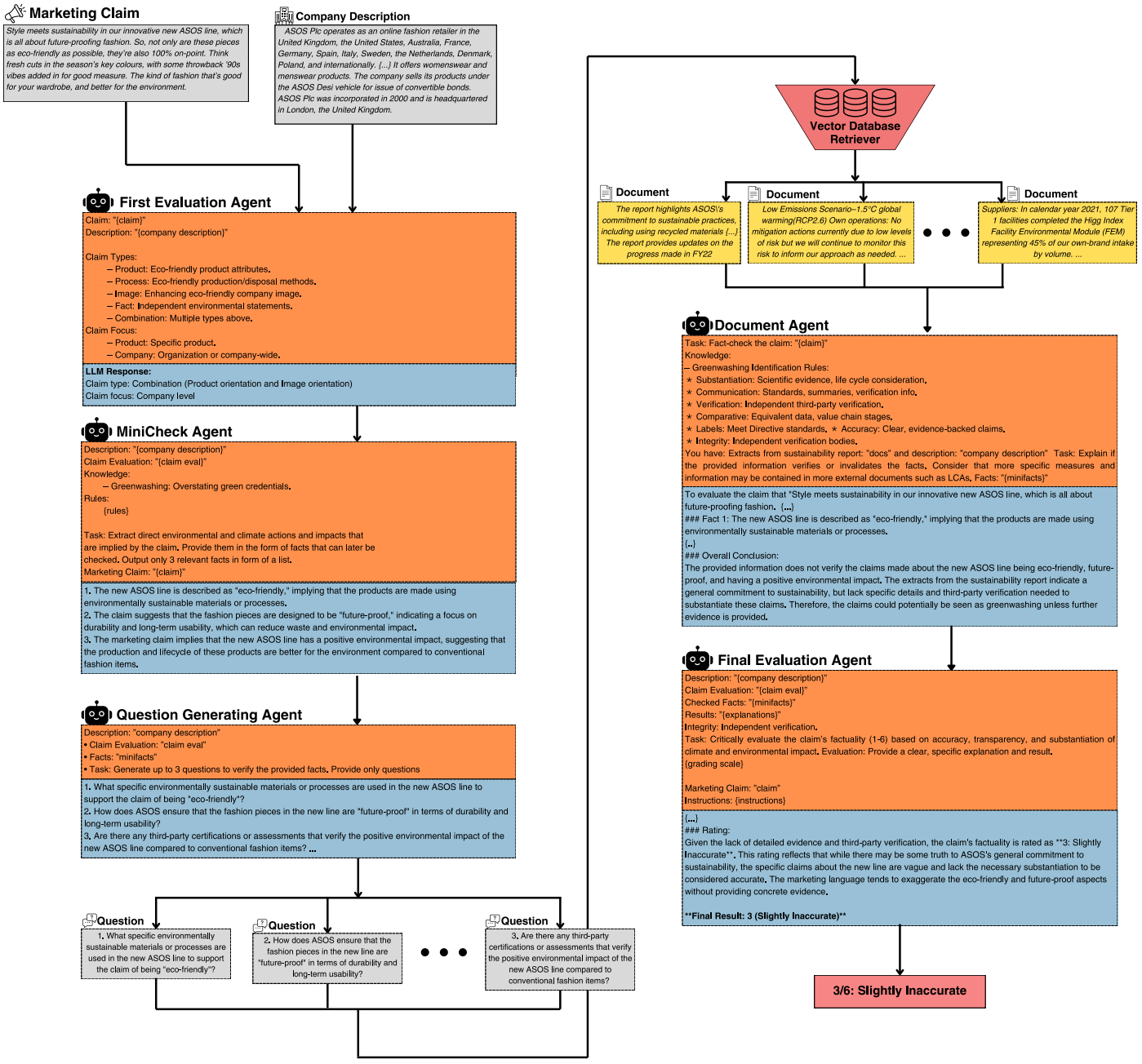

Figure 6: Example of Green Lantern framework using ASOS claim. LLM input is colored in orange and the output is colored in blue.

6.3.3 Scoring System

The scoring system, provided as the prompt of the the *Final Evaluation Agent*, is a six levels scale that matches the scale used by Quelle and Bovet (2024).

The scoring scale chosen for this study consists in a score from 1 to 6:

Claim Accuracy Scale

1: Highly Inaccurate - Misleading, false, no benefits (e.g., claiming zero emissions without evidence).
2: Moderately Inaccurate - Misleading, minimal benefits (e.g., overemphasizing minor green initiatives).
3: Slightly Inaccurate - Some truth, lacks detail (e.g., vague claims without specific metrics).
4: Slightly Accurate - Mostly accurate, minor omissions (e.g., generally truthful but with some exaggeration).
5: Moderately Accurate - Accurate, transparent, minor omissions (e.g., detailed claims with minor missing information).
6: Highly Accurate - Accurate, transparent, substantial benefits (e.g., comprehensive, evidence-backed claims).

Figure 7: Scoring Scale

This grading system was selected because the issue of greenwashing is complex by nature and classifying a claim in a binary fashion could prove challenging even for a human evaluator. Moreover, it provides a more interpretable outcome for researchers and allows to easily spot possible hallucinations. This is because the output is accessible at each step of the chain, and it is possible for the human evaluator to assess the correctness of the responses.

The claims collected as described in Section 5, are in a binary format. Therefore, to compare the results it is necessary to convert the output of the model to a binary scale: the claims graded 3 or lower will be classified as greenwashed and those with 4 or more points will be classified as not greenwashed.

6.4 Evaluation Metrics

To assess the performance of the proposed model and compare it with baseline frameworks, two key evaluation metrics were utilized: accuracy and F1 scores.

Accuracy measures the proportion of correct predictions made by a model out of all predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

TP = True Positives, FP = False Positives, TN = True Negatives, FN = False Negative

The F1 score is another accuracy measure that also provides insights on the reliability of the model. This measure is useful in situations where the cost of false positives and false negatives are different. In practice, the F1 score measures the trade-off between precision and recall. Precision measures the ratio of true positives to all the observations classified as positives. While the recall measures the ratio of true positive to the actual positive observations. In this case it is important to note that a wrong accusation of greenwashing has a higher cost, therefore a metric such as F1 score is necessary to measure the trade-off between precision and recall of the frameworks tested.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN}$$

6.5 Handling Hallucinations

Even though the final output is a rating, the intermediate output of every step of the chain-of-thought was saved. This allows to easily spot hallucinations and pinpoint where they happen. Following the approach of similar studies, a sample of 10 responses for each model, will be used to assess and evaluate the hallucination rate of the Green Lantern framework, given that 4 models will be tested, 40 chains of answers will be checked. As hallucinations refer to the factuality of the output, a response will be reported as hallucinated if the chain of answer is incoherent or if the data mentioned are not contained in the report. Only the proposed framework will be tested for hallucinations, as the simpler baseline models are expected to base their answer on the pre-training, with an implicit risk of hallucinations.

7 Results

The Green Lantern framework achieved an acceptable performance only when used with the GPT-4o model. This result aligns with the findings from fact-checking studies discussed in

the literature review. However, the proposed framework was outperformed by both simpler baseline frameworks.

As shown in Table 3, the combination of a single prompt and retrieval using the gemini-1.5-flash model achieved the best performance, with an accuracy of 0.7. In contrast, the proposed model, when used in combination with GPT-4, performed the worst, with an accuracy of 0.48. This indicates that, in terms of accuracy, the simpler approach was more effective for this task.

To further contextualize these results, random guessing would yield an accuracy of 0.5 in a binary classification scenario. The performances of *gemini-1.5-pro* (0.52) and *gpt-4* (0.48) are close to the random guessing baseline, highlighting the limited effectiveness of the Green Lantern framework in terms of accuracy.

Model	**Green Lantern**	**Single**	**Single Retrieval**
gemini-1.5-flash	0.56	0.65	0.70
gemini-1.5-pro	0.54	0.60	0.66
gpt-4o	0.58	0.58	0.59
gpt-4	0.48	0.58	0.60

Table 3: Balanced Accuracy of different models and frameworks

The F1 scores of the different models indicate their effectiveness in balancing precision and recall across various frameworks. The *gemini-1.5-pro* model consistently achieves high F1 scores, making it the best overall choice for scenarios where precision and recall need to be well-balanced. The *gemini-1.5-flash* model performs particularly well in the "Single" and "Single Retrieval" frameworks. The *gpt-4* model shows the lowest performance, struggling to balance precision and recall effectively. Given that incorrect greenwashing accusations can be more costly, models with higher F1 scores, like *gemini-1.5-pro*, are preferable.

Model	**Green Lantern**	**Single**	**Single Retrieval**
gemini-1.5-flash	0.62	0.67	0.75
gemini-1.5-pro	0.68	0.70	0.71
gpt-4o	0.56	0.39	0.37
gpt-4	0.43	0.63	0.53

Table 4: F1 Scores of different models and frameworks

In contrast with the initial hypothesis, the proposed framework yielded the overall worst per-

formance, in terms of balanced accuracy, out of the tested frameworks. Given the complexity of large language models, it is difficult to expain the results and interpret the reasons behind them. Nonetheless, the presented metrics only provide a partial understanding of the results. The following subsection aims at exploring the robustness of the results, to thoroughly assess the suitability of the framework for greenwashing detection tasks.

Type	**Model**	**TP**	**FP**	**FN**	**TN**
Green Lantern	gemini-1.5-flash	18	25	15	33
Green Lantern	gemini-1.5-pro	4	39	3	45
Green Lantern	gpt-4o	29	14	24	24
Green Lantern	gpt-4	25	18	29	19
Single	gemini-1.5-flash	26	17	15	33
Single	gemini-1.5-pro	12	31	5	43
Single	gpt-4o	41	2	36	12
Single	gpt-4	20	23	15	33
Single Retrieval	gemini-1.5-flash	24	19	8	40
Single Retrieval	gemini-1.5-pro	22	21	10	38
Single Retrieval	gpt-4o	43	0	37	11
Single Retrieval	gpt-4	35	8	28	20

Table 5: Combined Confusion Matrices for Different Frameworks and Models. TP: True Positives, FP: False Positives, FN: False Negatives, TN: True Negatives

7.1 Robustness

It is important to note that Large Language Models (LLMs) are pre-trained on a vast amount of documents and websites available online. All the greenwashing accusations and marketing claims reported in the dataset are accessible online. Although the prompts were designed to base the responses on the structure of the claims, there is a clear risk of spillover effects, when the target variable is contained in the training data, which would assess the robustness of the results.

Given the results of the experiment, we can hypothesize that the Single and Single Retrieval frameworks perform better due to a higher reliance on pre-training rather than on processing the claims. To test this hypothesis, claims belonging to four pairs of companies were selected and switched. As those claims will now lack substantiation because of the switch, the correct answer should be *greenwashing*. A check of the content of the sustainability reports was conducted to ensure that the claims are unsubstantiated. The framework

will be tested on these new combinations of claims and company descriptions.

Two claims were selected from companies well-known for their commitment to environmental sustainability and certified by third parties, such as Fairphone and ChopValue, while the other two claims were from companies with a negative brand image regarding sustainability and climate efforts, such as Procter Gamble (PG) and Tesco.

Tables 6 through 9 report the results of this robustness check. The tables contain the raw score, before it is converted to the binary scale, given by the different combinations of frameworks and models. The colors green and red are used to highlight whether the results is correctly indentified. As hypothesized, the Green Lantern framework seems to be more robust and is able to assess the claim correctly more consistently.

	Green Lantern		Single		Single Retrieval	
	New	Original	New	Original	New	Original
Gemini-1.5-flash	1	4	5	4	5	5
Gemini-1.5-pro	3	3	3	5	4	4
GPT-4o	2	4	4	5	4	5
GPT-4	3	4	3	4	3	4

Table 6: Results obtained by switching Fairphone claim with the one from HSBC. The claim is now greenwashed as it is not substantiated. Cells with values less than or equal to 3 are highlighted in green, and cells with values 4 or higher are highlighted in red.

	Green Lantern		Single		Single Retrieval	
	New	Original	New	Original	New	Original
Gemini-1.5-flash	4	4	5	2	5	6
Gemini-1.5-pro	2	3	3	4	3	3
GPT-4o	3	5	5	3	6	3
GPT-4	1	4	1	6	3	3

Table 7: Results obtained by switching ChopValue claim with the one from Amazon. The claim is now greenwashed as it is not substantiated. Cells with values less than or equal to 3 are highlighted in green, and cells with values 4 or higher are highlighted in red.

	Green Lantern		Single		Single Retrieval	
	New	Original	New	Original	New	Original
Gemini-1.5-flash	3	1	5	3	3	3
Gemini-1.5-pro	3	3	4	3	4	3
GPT-4o	3	4	5	5	5	5
GPT-4	2	4	4	3	4	5

Table 8: Results obtained by switching P&G's claim with the one from Rituals. The claim is now greenwashed as it is not substantiated. Cells with values less than or equal to 3 are highlighted in green, and cells with values 4 or higher are highlighted in red.

	Green Lantern		Single		Single Retrieval	
	New	Original	New	Original	New	Original
Gemini-1.5-flash	3	3	5	5	5	3
Gemini-1.5-pro	2	3	NA*	3	3	4
GPT-4o	1	3	3	5	3	5
GPT-4	1	4	4	4	5	5

Table 9: Results obtained by switching Tesco's claim with the one from Deutsche Post. The claim is now greenwashed as it is not substantiated. Cells with values less than or equal to 3 are highlighted in green, and cells with values 4 or higher are highlighted in red. *The model outputed: "This task asks to evaluate a climate claim that does not belong to Tesco ..."

8 Discussion

This section highlights on interpreting the results and on examining the factors contributing to the performance not meeting expectations. Moreover, the main challenges encountered in the study will be explored.

8.1 Robustness

From the results obtained in the previous section, it can be observed how the Green Lantern framework seems to be more robust and better resist to the adversarial attacks attempted. A surprising outcome was observed in the Single Agent framework, where the model identified that the claim did not originate from Tesco but from Deutsche Post, and did not provide a final answer. This single example clearly indicates the presence of a spillover effect and an over reliance on the pre-training for the *Single* framework. A potential reason for the performance drop in the simpler frameworks could be their dependence on pre-training for decision-making. This reliance makes the simpler frameworks unsuitable for greenwashing detection, as over-dependence on previously seen data can lead to poor performance in settings where there is no access to prior information.

While a more comprehensive assessment of the Green Lantern framework's performance could be conducted, these examples highlight the potential of the proposed framework as a more robust evaluator.

Another factor affecting the robustness of the results is the limited sample size. The dataset comprised only 90 observations, and retrieving accurate instances of greenwashed or non-greenwashed claims proved more challenging than anticipated, for both types of claims. Similar research has relied on evaluations by a team of experts, a solution that could not be implemented in this study, mainly for budgeting reasons.

8.2 Challenges of RAG

As explained in the Methodology section, RAG consists of two main parts: document embedding and document retrieval. With regard to text extraction and document embedding, only Google Gemini models were used, as they offer a combination of performance and cost-effectiveness. Experimenting with and comparing the performance of different embedding models could result in improved performance.

The extraction of tables and images from the reports proved to be challenging, especially with images saved in non-standard formats and tables that spanned across two different pages. These issues affected approximately 12% of the extracted images and tables, calculated on a

sub-sample of 10 reports.

Another approach to information embedding could also be explored to improve performance. The framework ESG Reveal, proposed by Zou et al. (2023), consists of extracting relevant information such as metrics, commitments, and actions from sustainability reports and storing this data in a structured way. Such an approach could ensure improved retrieval of information; however, this improvement would significantly increase the costs.

8.3 Information Availability and Online Access

Another area of concern involves the data and information made available to the models. While sustainability reports explain the environmental efforts of the company, especially in larger companies, the specific impact and environmental information about a product may not be reported.

To address this issue, an attempt was made to conduct an online search. The goal was to enable the model to use a search engine (e.g. Google Search) to confirm information not found in the sustainability report, particularly by locating the product page on the company website. However, this approach introduced other issues: alongside relevant documents and product pages, news articles containing greenwashing accusations were also retrieved, invalidating the experiment's results.

The decision not to further explore this route was primarily due to budget constraints, as the number of tokens processed for each observation quickly scaled up.

8.4 Hallucinations

The sample of 10 observations per model led to the analysis of 40 chains of answers. Out of the 40 chains of answers 6 were flagged as hallucinations. However, it is important to notice that these hallucinations are all in relation to the lack of relevant information in the sustainability report. This problem only occurred with larger companies and conglomerates, such as Romerquelle, owned by Coca-cola. A possible reason for this could be that large corporations do not disclose specific information in the sustainability report.

This finding is in line with the expectations, as the chain of thought approach, combined with a temperature of 0 should minimize these type of events.

8.5 Scoring system

The chosen scoring system design was initially based on the Good On You scoring system (You, 2024). However, after an initial run, some problem were found in the way the model

was assigning the score. Moving to a custom scale from 1 to 6, made it more consistent with similar fact checking studies and seemed to allow the LLMs to make more coherent decisions.

Using the sample taken to check for possible hallucination, something unexpected was discovered. The framework is extracting relevant information from the reports and classifying it in a somewhat correct manner, up until the scoring part, where a generous score was given. This occurred in 9 of the 40 observations considered, with 6 of them being generated either by the GPT-4o or GPT-4 model. An example of these type of mistakes is shown in Figure 8, where the model correctly recognizes that the 73% figure used by Oatly cannot be used as a universal claim, which is a similar conclusion to the ASA ruling. But fails to score it correctly.

This shows there is room for improvement, as this area needs refinement and further investigation. A possible future development in this area could be the development of a more structured framework, or a more rigid set of instructions for the LLM to follow.

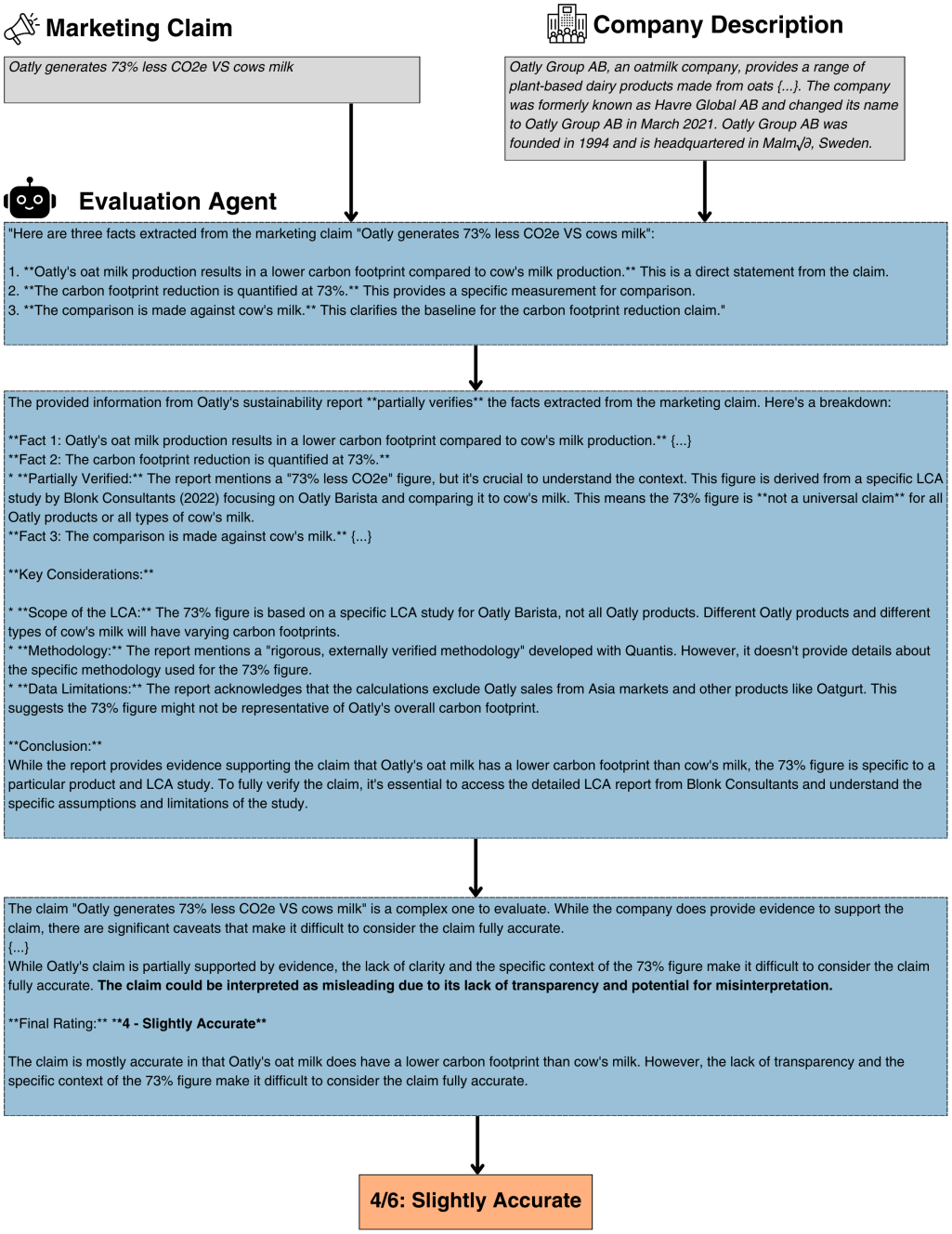

Figure 8: Green Lantern framework with Gemini-1.5-flash response on Oatly's claim. This figure only contains the outputs of the intermediate steps of the Green Lantern Framework

9 Conclusion

The aim of this study was to evaluate whether pre-trained LLMs can accurately identify greenwashing. Different models, including OpenAI's GPT-4o and GPT-4-turbo and Google's Gemini-1.5-flash and Gemini-1.5-pro, were tested with three different frameworks: *Green Lantern* using a combination of Chain of Thought and RAG, S*ingle Agent Retrieval* using only RAG, and *Single Agent* using a single LLM instance.

In response to the primary research question *"How effective are pre-trained LLMs at detecting greenwashing in marketing claims?"* the results indicate that while LLMs show high level of accuracy against random chance baseline, their effectiveness varies significantly depending on the framework and model used. The simpler frameworks, specifically the Single Agent with Retrieval, demonstrated superior performance in terms of accuracy and F1 scores. This indicates that LLMs can be effective at predicting whether a claim is greenwashed or not, when compared with a random chance baseline.

Regarding the impact of different frameworks, the Green Lantern framework, although it did not perform as well as anticipated in absolute terms, demonstrated the robustness of LLMs in greenwashing detection, achieving an accuracy of 0.58 when paired with GPT-4o. This suggests that framework design plays a crucial role in improving the effectiveness of LLMs for this purpose. There remains significant room for improvement in areas such as the scoring system, dataset size, data quality, and the type of information provided to the model.

An additional analysis of the results uncovered important limitations: vulnerability to pre-existing biases and potential over-reliance on pre-training data can lead to inaccurate predictions and reduce the effectiveness of LLMs in detecting greenwashing.

The results and limitations analyzed in the study indicate that while LLMs have potential, they are not yet fully reliable for greenwashing detection tasks without further refinement. Overall, this research contributes to the field of LLM-based fact-checking and serves as a first exploration of LLM capabilities in greenwashing detection. It also produced a first public and structured dataset for greenwashing detection. While LLMs have shown potential, this study highlights the need for further research and development to overcome their current limitations.

10 Appendix

	Prompt	Variables
1	• **Claim:** "claim" • **Description:** "company_description" • **Claim Types:** – Product: Eco-friendly product attributes. – Process: Eco-friendly production/disposal methods. – Image: Enhancing eco-friendly company image. – Fact: Independent environmental statements. – Combination: Multiple types above. • **Claim Focus:** – Product: Specific product. – Company: Organization or company-wide. • **Task:** As a sustainability expert, identify the claim type and focus. No explanation needed.	claim, company_description
2	• **Context:** • **Description:** "company_description" • **Claim Evaluation:** "claim_eval" • **Knowledge:** – Greenwashing: Overstating green credentials. • **Rules:** – Substantiation: Claims need scientific evidence and life cycle consideration. – Communication: Include standards, summaries, and verification info. – Verification: Third-party verification required. – Comparative: Ensure comparable data and value chain coverage. – Labels: Meet Directive requirements. – Accuracy: Claims must be clear, detailed, and evidence-backed. – Integrity: Independent verification bodies. • **Task:** Extract direct environmental and climate actions and impacts that are implied by the claim. Provide them in the form of facts that can later be checked. Output only 3 relevant facts in form of a list. • **Marketing Claim:** "claim"	claim, company_description, claim_eval
3	• **Description:** "company_description" • **Claim Evaluation:** "claim_eval" • **Facts:** "minifacts" • **Task:** Generate up to 3 questions to verify the provided facts. Provide only questions.	claim, company_description, claim_eval, minifacts

4	• **Task:** Fact-check the claim: "claim" • **Knowledge:** – Greenwashing Identification Rules: ∗ Substantiation: Scientific evidence, life cycle consideration. ∗ Communication: Standards, summaries, verification info. ∗ Verification: Independent third-party verification. ∗ Comparative: Equivalent data, value chain stages. ∗ Labels: Meet Directive standards. ∗ Accuracy: Clear, evidence-backed claims. ∗ Integrity: Independent verification bodies. • **You have:** Extracts from sustainability report: "docs" and description: "company_description" • **Task:** Explain if the provided information verifies or invalidates the facts. Consider that more specific measures and information may be contained in more external documents such as LCAs. • **Facts:** "minifacts"	claim, company_description, docs, minifacts
5	• **Description:** "company_description" • **Claim Evaluation:** "claim_eval" • **Checked Facts:** "minifacts" • **Results:** "explanations" • **Integrity:** Independent verification. • **Task:** Critically evaluate the claim's factuality (1-6) based on accuracy, transparency, and substantiation of climate and environmental impact. 1: Highly Inaccurate - Misleading, false, no benefits (e.g., claiming zero emissions without evidence). 2: Moderately Inaccurate - Misleading, minimal benefits (e.g., overemphasizing minor green initiatives). 3: Slightly Inaccurate - Some truth, lacks detail (e.g., vague claims without specific metrics). 4: Slightly Accurate - Mostly accurate, minor omissions (e.g., generally truthful but with some exaggeration). 5: Moderately Accurate - Accurate, transparent, minor omissions (e.g., detailed claims with minor missing information). 6: Highly Accurate - Accurate, transparent, substantial benefits (e.g., comprehensive, evidence-backed claims). • **Evaluation:** Provide a clear, specific explanation and result. • **Marketing Claim:** "claim" • **Instructions:** Conduct a thorough and critical analysis of the claim's validity. Pay special attention to potential greenwashing tactics. Focus on the evidence provided, logical consistency, and overall transparency. Consider the inherent tendency of marketing claims to exaggerate, and clearly differentiate between harmless embellishment and significant misrepresentation. Highlight any red flags or inconsistencies that suggest greenwashing. Explain the decision in detail and disclose the final result at the end.	claim, company_description, claim_eval, minifacts, explanations
6	• **Final evaluation:** final_eval • **Task:** Extract and report the final score as a number from 1 to 6. Only report the score.	final_eval

Table 10: Prompts from Green Lantern Framework

#	Prompt	Variables
1	• **Description:** "company_description" • **Integrity:** Independent verification. • **Task:** Critically evaluate the claim's factuality (1-6) based on accuracy, transparency, and substantiation of climate and environmental impact. – 1: Highly Inaccurate - Misleading, false, no benefits (e.g., claiming zero emissions without evidence). – 2: Moderately Inaccurate - Misleading, minimal benefits (e.g., overemphasizing minor green initiatives). – 3: Slightly Inaccurate - Some truth, lacks detail (e.g., vague claims without specific metrics). – 4: Slightly Accurate - Mostly accurate, minor omissions (e.g., generally truthful but with some exaggeration). – 5: Moderately Accurate - Accurate, transparent, minor omissions (e.g., detailed claims with minor missing information). – 6: Highly Accurate - Accurate, transparent, substantial benefits (e.g., comprehensive, evidence-backed claims). • **Evaluation:** Provide a clear, specific explanation and result. • **Marketing Claim:** "claim" • **Instructions:** Conduct a thorough and critical analysis of the claim's validity. Pay special attention to potential greenwashing tactics. Focus on the evidence provided, logical consistency, and overall transparency. Consider the inherent tendency of marketing claims to exaggerate, and clearly differentiate between harmless embellishment and significant misrepresentation. Highlight any red flags or inconsistencies that suggest greenwashing. Explain the decision in detail and disclose the final result at the end.	claim, company_description
2	• **Final evaluation:** final_eval • **Task:** Extract and report the final score as a number from 1 to 6. Only report the score.	final_eval

Table 11: Prompts from Single Agent Framework

#	Prompt	Variables
1	• **Description:** "company_description" • **Sustainability Report Extracts:** "docs" • **Integrity:** Independent verification. • **Task:** Critically evaluate the claim's factuality (1-6) based on accuracy, transparency, and substantiation of climate and environmental impact. – 1: Highly Inaccurate - Misleading, false, no benefits (e.g., claiming zero emissions without evidence). – 2: Moderately Inaccurate - Misleading, minimal benefits (e.g., overemphasizing minor green initiatives). – 3: Slightly Inaccurate - Some truth, lacks detail (e.g., vague claims without specific metrics). – 4: Slightly Accurate - Mostly accurate, minor omissions (e.g., generally truthful but with some exaggeration). – 5: Moderately Accurate - Accurate, transparent, minor omissions (e.g., detailed claims with minor missing information). – 6: Highly Accurate - Accurate, transparent, substantial benefits (e.g., comprehensive, evidence-backed claims). • **Evaluation:** Provide a clear, specific explanation and result. • **Marketing Claim:** "claim" • **Instructions:** Conduct a thorough and critical analysis of the claim's validity. Pay special attention to potential greenwashing tactics. Focus on the evidence provided, logical consistency, and overall transparency. Consider the inherent tendency of marketing claims to exaggerate, and clearly differentiate between harmless embellishment and significant misrepresentation. Highlight any red flags or inconsistencies that suggest greenwashing. Explain the decision in detail and disclose the final result at the end.	claim, company_description, docs
2	• **Final evaluation:** final_eval • **Task:** Extract and report the final score as a number from 1 to 6. Only report the score.	final_eval

Table 12: Prompts from Single Agent with Retrieval Framework

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., and Anadkat, S. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anthropic (2021). A mathematical framework for transformer circuits.

Attig, N., Rahaman, M. M., and Trabelsi, S. (2021). Greenwashing and bank loan contracting: does environmental disclosure quality matter to creditors? *Available at SSRN 3880113*.

Azamfirei, R., Kudchadkar, S. R., and Fackler, J. (2023). Large language models and the perils of their hallucinations. *Critical Care*, 27(1):120. ID: Azamfirei2023.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Balluchi, F., Lazzini, A., and Torelli, R. (2020). Csr and greenwashing: A matter of perception in the search of legitimacy. *Accounting, accountability and society: Trends and perspectives in reporting, management and governance for sustainability*, pages 151–166.

Bhardwaj, S., Nair, K., Tariq, M. U., Ahmad, A., and Chitnis, A. (2023). The state of research in green marketing: A bibliometric review from 2005 to 2022. *Sustainability*, 15(4).

Carlson, L., Grove, S. J., and Kangun, N. (1993). A content analysis of environmental advertising claims: A matrix method approach. *Journal of advertising*, 22(3):27–39.

Cheng, Y., Zhang, C., Zhang, Z., Meng, X., Hong, S., Li, W., Wang, Z., Wang, Z., Yin, F., and Zhao, J. (2024). Exploring large language model based intelligent agents: Definitions, methods, and prospects. *arXiv preprint arXiv:2401.03428*.

Choi, E. C. and Ferrara, E. (2024). Fact-gpt: Fact-checking augmentation via claim matching with llms. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 883–886.

Commission, E. (2022). Green claims.

de Freitas Netto, S. V., Sobral, M. F. F., Ribeiro, A. R. B., and Soares, G. R. d. L. (2020). Concepts and forms of greenwashing: a systematic review. *Environmental Sciences Europe*, 32(1):19. ID: de Freitas Netto2020.

DeHaven, M. and Scott, S. (2023a). Bevers: A general, simple, and performant framework for automatic fact verification. *arXiv preprint arXiv:2303.16974*.

DeHaven, M. and Scott, S. (2023b). Bevers: A general, simple, and performant framework for automatic fact verification. *arXiv preprint arXiv:2303.16974*.

EY (2024). Can artificial intelligence uncover greenwashing?

Ferris, T., Lawlor, J., and Ketterer, E. (2023). Guidance for 'sustainable' claims after dismissal of hm 'greenwashing' class action.

FTC (2022). Ftc seeks public comment on potential updates to its 'green guides' for the use of environmental marketing claims.

Google (2022). Report: What it will take for ceos to fund a sustainable transformation.

Hoes, E., Altay, S., and Bermeo, J. (2023). Leveraging chatgpt for efficient fact-checking. *PsyArXiv.April*, 3.

Huang, J. and Chang, K. C. (2022). Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

Huang, Y., Feng, X., Feng, X., and Qin, B. (2021). The factual inconsistency problem in abstractive text summarization: A survey. *arXiv preprint arXiv:2104.14839*.

IBM (2023a). What are foundation models?

IBM (2023b). What are large language models (llms)?

IBM (2024). What is a sustainability report?

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023a). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Ji, Z., Yu, T., Xu, Y., Lee, N., Ishii, E., and Fung, P. (2023b). Towards mitigating llm hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843.

Kandpal, N., Deng, H., Roberts, A., Wallace, E., and Raffel, C. (2023). Large language models struggle to learn long-tail knowledge. In *Proceedings of the International Conference on Machine Learning*, pages 15696–15707. PMLR.

Kumar, P. (2016). State of green marketing research over 25 years (1990-2014): Literature survey and classification. *Marketing Intelligence amp Planning*, 34:137–158.

Lee, M. T. and Raschke, R. L. (2023). Stakeholder legitimacy in firm greening and financial performance: What about greenwashing temptations?. *Journal of Business Research*, 155:113393. ID: 271680.

Leonidou, C. N. and Skarmeas, D. (2017). Gray shades of green: Causes and consequences of green skepticism. *Journal of Business Ethics*, 144(2):401–415. ID: Leonidou2017.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., and Rocktäschel, T. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Li, M., Peng, B., Galley, M., Gao, J., and Zhang, Z. (2023). Self-checker: Plug-and-play modules for fact-checking with large language models. *arXiv preprint arXiv:2305.14623*.

Li, T., Zhang, G., Do, Q. D., Yue, X., and Chen, W. (2024). Long-context llms struggle with long in-context learning. *arXiv preprint arXiv:2404.02060*.

McHugh, C. and Way, J. (2018). What is good reasoning? *Philosophy and Phenomenological Research*, 96(1):153–174.

Media, O. (2023). O'reilly releases 2023 generative ai in the enterprise report revealing unprecedented 67% rate of adoption, growing demand for ai programming and data analysis skills. Accessed: 2024-07-25.

Ni, J., Bingler, J., Colesanti-Senni, C., Kraus, M., Gostlow, G., Schimanski, T., Stammbach, D., Vaghefi, S. A., Wang, Q., and Webersinke, N. (2023). Chatreport: Democratizing sustainability disclosure analysis through llm-based tools. *arXiv preprint arXiv:2307.15770*.

Nyilasy, G., Gangadharbatla, H., and Paladino, A. (2014). Perceived greenwashing: The interactive effects of green advertising and corporate environmental performance on consumer reactions. *Journal of Business Ethics*, 125(4):693–707. ID: Nyilasy2014.

Papadas, K.-K., Avlonitis, G. J., Carrigan, M., and Piha, L. (2019). The interplay of strategic and internal green marketing orientation on competitive advantage. *Journal of Business Research*, 104:632–643.

Quelle, D. and Bovet, A. (2024). The perils and promises of fact-checking with large language models. *Frontiers in Artificial Intelligence*, 7:1341697.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Seele, P. and Gatti, L. (2017). Greenwashing revisited: In search of a typology and accusation-based definition incorporating legitimacy strategies. *Business Strategy and the Environment*, 26(2):239–252.

Solutions, U. (2007). Sins of greenwashing.

Tang, L., Laban, P., and Durrett, G. (2024). Minicheck: Efficient fact-checking of llms on grounding documents. *arXiv preprint arXiv:2404.10774*.

Tank, E. T. (2024). Green claims directive.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., and Hauth, A. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, , and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, Y., Reddy, R. G., Mujahid, Z. M., Arora, A., Rubashevskii, A., Geng, J., Afzal, O. M., Pan, L., Borenstein, N., and Pillai, A. (2023). Factcheck-gpt: End-to-end fine-grained document-level fact-checking and correction of llm output. *arXiv preprint arXiv:2311.09000*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Wei, J., Yang, C., Song, X., Lu, Y., Hu, N., Huang, J., Tran, D., Peng, D., Liu, R., Huang, D., Du, C., and Le, Q. V. (2024). Long-form factuality in large language models.

Xu, C., Sun, Q., Zheng, K., Geng, X., Zhao, P., Feng, J., Tao, C., and Jiang, D. (2023). Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Yang, Z., Nguyen, T. T. H., Nguyen, H. N., Nguyen, T. T. N., and Cao, T. T. (2020). Greenwashing behaviours: Causes, taxonomy and consequences based on a systematic literature review. *Journal of business economics and management*, 21(5):1486–1507.

You, G. O. (2024). How we rate fashion brands.

Zhang, J., Ouyang, Y., Ballesteros-Pérez, P., Li, H., Philbin, S. P., Li, Z., and Skitmore, M. (2021). Understanding the impact of environmental regulations on green technology innovation efficiency in the construction industry. *Sustainable Cities and Society*, 65:102647. ID: 280276.

Zhang, L., Li, D., Cao, C., and Huang, S. (2018). The influence of greenwashing perception on green purchasing intentions: The mediating role of green word-of-mouth and moderating role of green concern. *Journal of Cleaner Production*, 187:740–750. ID: 271750.

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., and Chen, Y. (2023). Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Zhang, Y., Mao, S., Ge, T., Wang, X., de Wynter, A., Xia, Y., Wu, W., Song, T., Lan, M., and Wei, F. (2024). Llm as a mastermind: A survey of strategic reasoning with large language models. *arXiv preprint arXiv:2404.01230*.

Zou, Y., Shi, M., Chen, Z., Deng, Z., Lei, Z., Zeng, Z., Yang, S., Tong, H., Xiao, L., and Zhou, W. (2023). Esgreveal: An llm-based approach for extracting structured data from esg reports. *arXiv preprint arXiv:2312.17264*.