Automating Corporate Greenwashing Detection Using Natural Language Processing

Nicole Lin*, Amy Pu*
*Columbia University, New York, NY, USA
Email: {nsl2126, ap4489}@columbia.edu

***Abstract*—Corporate greenwashing, where firms exaggerate or misrepresent their environmental efforts, poses a critical barrier to genuine climate progress. Traditional expert-driven evaluations, such as the Corporate Climate Responsibility Monitor (CCRM), deliver high-quality assessments but demand intensive manual effort, limiting scalability. In this paper, we propose a dataset to assess corporate sustainability action and policies and present an end-to-end automated pipeline that synthesizes advanced PDF extraction, recursive document chunking, retrieval-augmented generation (RAG), and rubric-driven prompt engineering to produce CCRM-style assessments at scale. We evaluate three large language models — ClimateGPT, Qwen, and Mistral[1] — across multiple prompting and retrieval strategies. Our RAG retrieval achieves a question-level accuracy up to 26 % for overall transparency, substantially outperforming naive summarization and truncation baselines. We conclude with a discussion of limitations, open challenges, and future directions for improving automated climate-report analysis. You can find our code and prompts at github.com/amypu99/ml-climate.**

I. Introduction

Over the past decade, heightened regulatory scrutiny, investor activism, and consumer demand have driven corporations to publicize ambitious sustainability commitments. Yet, a growing body of evidence indicates that many organizations engage in *greenwashing*, overstating environmental initiatives without corresponding operational changes. For instance, a company might announce a "net-zero by 2050" target while continuing to expand fossil-fuel operations, or heavily marketing minor recycling programs while omitting major sources of greenhouse gas emissions from its disclosures. This misalignment between corporate rhetoric and actual performance not only misleads stakeholders but also diffuses accountability, slowing genuine progress toward global climate goals.

Despite the urgency of addressing greenwashing, existing evaluation frameworks remain labor-intensive. The Corporate Climate Responsibility Monitor (CCRM), developed by climatology and ESG research specialists, offers rigorous, multi-faceted assessments of corporate climate commitments [1]. CCRM reports rate firms across four domains: (1) emissions tracking and disclosure, (2) setting of climate targets, (3) realized emission reductions, and (4) management of unabated emissions. While these expert-driven analyses are highly detailed, they require months of manual literature review, data extraction, and qualitative judgment—rendering continuous monitoring of thousands of publicly traded companies infeasible.

Recent advancements in Natural Language Processing (NLP) and Large Language Models (LLMs) present an opportunity to automate key aspects of these evaluations. LLMs such as GPT variants can interpret unstructured text, generate summaries, and answer complex queries when provided with appropriate context. However, there are few open datasets available for training models and evaluation, which is the result of several challenges: corporate sustainability reports often exceed tens or hundreds of pages; formatting varies widely across firms; and critical details may appear in footnotes, charts, or tables. Furthermore, naively feeding entire documents into an LLM risks exceeding context windows or diluting relevant information amidst extraneous content.

In this paper, we propose a dataset assessing corporate sustainability action created from a collection of corporate sustainability reports which are used to identify greenwashing. We construct assessment labels from the Corporate Carbon Responsibility Monitor and Transition Pathway Initiative. Our dataset task requires inputs of corporate sustainability reports and outputs a set of scores and answers associated with the corporate sustainability assessment.

Given the dataset, we design a modular pipeline that combines: (1) GPU-accelerated, LLM-based OCR for faithful text extraction; (2) recursive document chunking informed by best-practice research to maintain semantic coherence within token limits; (3) Retrieval-Augmented Generation (RAG) to focus LLM attention on the most pertinent passages; and (4) rubric-driven prompt engineering embedding expert scoring criteria directly into model inputs. By synthesizing these components, our system attempts to replicate essential CCRM and TPI methodology at scale, providing a starting point for automating the production of structured climate disclosures with minimal manual intervention.

Our contributions are threefold: **first**, we introduce a dataset including over 200 documents to assess corporate sustainability action and identify greenwashing; **second**, we construct an end-to-end pipeline with robust data ingestion, chunking strategy, and embedding formal scoring rubrics within prompts to automate the creation of the corporate sustainability monitoring reports; **third**, we provide an empirical evaluation of three domain-adapted LLMs—ClimateGPT, Mistral, and Qwen—across multiple retrieval and prompting strategies, showing that our RAG approach yields up to 26% accuracy

[1] We will refer to models ClimateGPT-7B, Ministral-8B-Instruct-2410, and Qwen2.5-7B-Instruct-1M as ClimateGPT, Mistral, and Qwen throughout the paper.

on transparency scoring tasks.

The remainder of the paper is organized as follows. Section II surveys related work in automated document evaluation, RAG methods, and domain-specific LLMs. Section III details our data collection and preprocessing procedures across successive conceptual sprints. Section IV describes our end-to-end pipeline architecture. Section V presents experimental setup and quantitative results. Section VI discusses insights, limitations, and practical implications. Finally, Section VII concludes and outlines directions for future research.

II. Related Work

Automated assessment of corporate sustainability claims intersects multiple research domains: expert-driven frameworks, retrieval-augmented generation, domain-adapted language models, and prompt engineering strategies.

A. Expert-Driven Climate Frameworks

The Corporate Climate Responsibility Monitor (CCRM) by NewClimate Institute provides in-depth, expert-curated evaluations of corporate climate commitments across four pillars: emissions disclosure, target setting, realized reductions, and management of residual emissions [1]. Each pillar comprises granular sub-criteria, such as the breadth of scope (Scope 1, 2, and 3 disclosures), specificity of reduction targets, third-party verification status, and transparency around emission offset mechanisms. While CCRM sets a high standard, its manual production pipeline—entailing several months of literature review and expert calibration—precludes frequent or broad-scale updates.

Other frameworks include the Science Based Targets initiative (SBTi) [3], which validates corporate emission targets against climate science thresholds, and the Transition Pathway Initiative (TPI), which evaluates firms' readiness for a low-carbon transition via 23 structured questions spanning six maturity levels [4]. Although SBTi and TPI offer more frequent updates, their assessments remain limited in qualitative depth compared to CCRM.

ESG scores are another set of broad ratings that assess a company's performance across three key areas: Environmental, Social, and Governance [5]. The scope of ESG scores is much larger than that of CCRM and TP, as ESG scores also include a wider array of non-climate-related factors, such as labor practices or board diversity [6]. Additionally, ESG scores are often based on a combination of third-party ratings and public disclosures, which may lack consistency or full transparency [6].

B. Corporate Sustainability NLP Research

There has been developing research interest at the intersection of NLP and climate related tasks. Ekimetrics developed Climate Q&A [7], a conversational assistant for question-answering, which uses data from over thousands of pages of scientific reports to generate quick answers. Stammbach et al. [8] proposed an expert-annotated dataset on the task of sentence-level classification for environmental claim detection and provided models trained on the task. More recently, Morio and Manning [9] developed an NLP benchmark for assessing climate policy engagement based on LobbyMap, which aims to categorize a corporation's stance on specific topics. While Morio and Manning's work is a natural predecessor to our paper, their work focuses mostly on capturing companies' stances on certain climate policies. Our work takes this step further and aims to automate the evaluation of the company's climate policies and actions, not just by capturing the company's stance and action but actually assigning transparency and integrity scores. We believe our research most closely resembles a major end goal in the work of climate experts in making complex assessments given the many long environmental claims and documents provided by a company.

C. Domain-Specific Language Models

General-purpose LLMs often lack the specialized vocabulary and nuanced reasoning required for climate and sustainability texts. ClimateBERT, a BERT-based encoder model fine-tuned on climate corpora, achieves improved classification accuracy on tasks like emission category identification and sentiment analysis in environmental reports [10]. It has also been used for analyzing companies' climate-risk disclosures along the Task Force for Climate-related Financial Disclosures (TCFD) categories [11]. In parallel, ClimateGPT extends generative LLMs by fine-tuning on a 4.2B-token climate-specific dataset, improving document-level summarization and question-answering for sustainability texts [12]. These models underline the value of domain adaptation but stop short of automating structured rubric-based scoring.

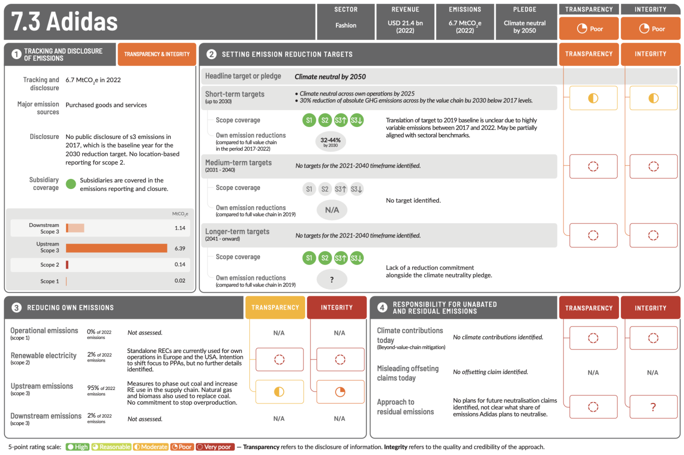
D. Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation combines neural text generation with vector-based retrieval to ground LLM outputs in external knowledge sources [15]. By embedding document passages into a vector store, queries retrieve top-k semantically similar chunks, mitigating hallucinations and improving factual accuracy. Snowflake's engineering research highlights that medium-sized chunks (2K tokens) optimize the trade-off between context completeness and model input constraints, a finding we adopt in our chunking strategy [16].

RAG has been successfully applied in specialized domains, such as legal document summarization, financial report analysis, and medical Q&A, demonstrating its efficacy in knowledge-intensive tasks [17] [18] [19]. However, prior work has typically focused on short-form queries or summary-generation rather than structured rubric-based scoring as needed for climate disclosures.

E. Prompt Engineering for Structured Outputs

Recent studies in prompt engineering emphasize embedding explicit task instructions and evaluation rubrics directly into LLM prompts to achieve reliable, parsable outputs. Brown et al. [20] demonstrate that carefully designed "system messages" and formatting cues significantly reduce LLM hallucinations. Gao et al. [21] further show that embedding



7.3 Adidas

SECTOR	REVENUE	EMISSIONS	PLEDGE	TRANSPARENCY	INTEGRITY
Fashion	USD 21.4 bn (2022)	6.7 $MtCO_2e$ (2022)	Climate neutral by 2050	◔ Poor	◔ Poor

1 TRACKING AND DISCLOSURE OF EMISSIONS — TRANSPARENCY & INTEGRITY

Tracking and disclosure: 6.7 $MtCO_2e$ in 2022

Major emission sources: Purchased goods and services

Disclosure: No public disclosure of s3 emissions in 2017, which is the baseline year for the 2030 reduction target. No location-based reporting for scope 2.

Subsidiary coverage: ● Subsidiaries are covered in the emissions reporting and closure.

	$MtCO_2e$
Downstream Scope 3	1.14
Upstream Scope 3	6.39
Scope 2	0.14
Scope 1	0.02

2 SETTING EMISSION REDUCTION TARGETS

			TRANSPARENCY	INTEGRITY
Headline target or pledge	*Climate neutral by 2050*			
Short-term targets (up to 2030)	• *Climate neutral across own operations by 2025* • *30% reduction of absolute GHG emissions across by the value chain bu 2030 below 2017 levels.*		◑	◑
Scope coverage	S1 S2 S3↑ S3↓	Translation of target to 2019 baseline is unclear due to highly variable emissions between 2017 and 2022. May be partially aligned with sectoral benchmarks.		
Own emission reductions (compared to full value chain in the period 2017-2022)	32-44% by 2030			
Medium-term targets (2031 - 2040)	*No targets for the 2021-2040 timeframe identified.*		◌	◌
Scope coverage	S1 S2 S3↑ S3↓	No target identified.		
Own emission reductions (compared to full value chain in 2019)	N/A			
Longer-term targets (2041 - onward)	*No targets for the 2021-2040 timeframe identified.*		◌	◌
Scope coverage	S1 S2 S3↑ S3↓	Lack of a reduction commitment alongside the climate neutrality pledge.		
Own emission reductions (compared to full value chain in 2019)	?			

3 REDUCING OWN EMISSIONS

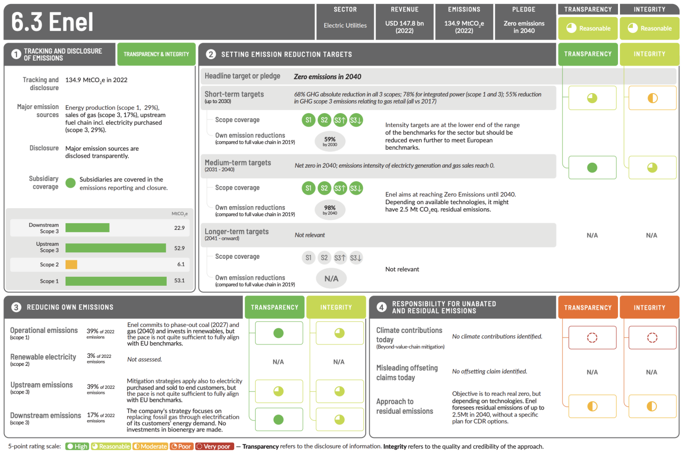
			TRANSPARENCY	INTEGRITY
Operational emissions (scope 1)	0% of 2022 emissions	*Not assessed.*	N/A	N/A
Renewable electricity (scope 2)	2% of 2022 emissions	Standalone RECs are currently used for own operations in Europe and the USA. Intention to shift focus to PPAs, but no further details identified.	◌	◌
Upstream emissions (scope 3)	95% of 2022 emissions	Measures to phase out coal and increase RE use in the supply chain. Natural gas and biomass also used to replace coal. No commitment to stop overproduction.	◑	◔
Downstream emissions (scope 3)	2% of 2022 emissions	*Not assessed.*	N/A	N/A

4 RESPONSIBILITY FOR UNABATED AND RESIDUAL EMISSIONS

		TRANSPARENCY	INTEGRITY
Climate contributions today (Beyond-value-chain mitigation)	*No climate contributions identified.*	◌	◌
Misleading offseting claims today	*No offsetting claim identified.*	N/A	N/A
Approach to residual emissions	No plans for future neutralisation claims identified, not clear what share of emissions Adidas plans to neutralise.	◌	?

5-point rating scale: ● High ◕ Reasonable ◑ Moderate ◔ Poor ◌ Very poor — **Transparency** refers to the disclosure of information. **Integrity** refers to the quality and credibility of the approach.

Fig. 1. Example of a CCRM report with color-coded ratings across different categories.



6.3 Enel

SECTOR	REVENUE	EMISSIONS	PLEDGE	TRANSPARENCY	INTEGRITY
Electric Utilities	USD 147.8 bn (2022)	134.9 $MtCO_2e$ (2022)	Zero emissions in 2040	◕ Reasonable	◕ Reasonable

1 TRACKING AND DISCLOSURE OF EMISSIONS — TRANSPARENCY & INTEGRITY

Tracking and disclosure: 134.9 $MtCO_2e$ in 2022

Major emission sources: Energy production (scope 1, 29%), sales of gas (scope 3, 17%), upstream fuel chain incl. electricity purchased (scope 3, 29%).

Disclosure: Major emission sources are disclosed transparently.

Subsidiary coverage: ● Subsidiaries are covered in the emissions reporting and closure.

	$MtCO_2e$
Downstream Scope 3	22.9
Upstream Scope 3	52.9
Scope 2	6.1
Scope 1	53.1

2 SETTING EMISSION REDUCTION TARGETS

			TRANSPARENCY	INTEGRITY
Headline target or pledge	*Zero emissions in 2040*			
Short-term targets (up to 2030)	*68% GHG absolute reduction in all 3 scopes; 78% for integrated power (scope 1 and 3); 55% reduction in GHG scope 3 emissions relating to gas retail (all vs 2017)*		◕	◑
Scope coverage	S1 S2 S3↑ S3↓	Intensity targets are at the lower end of the range of the benchmarks for the sector but should be reduced even further to meet European benchmarks.		
Own emission reductions (compared to full value chain in 2019)	59% by 2030			
Medium-term targets (2031 - 2040)	*Net zero in 2040; emissions intensity of electricty generation and gas sales reach 0.*		●	◕
Scope coverage	S1 S2 S3↑ S3↓	Enel aims at reaching Zero Emissions until 2040. Depending on available technologies, it might have 2.5 Mt CO_2eq. residual emissions.		
Own emission reductions (compared to full value chain in 2019)	98% by 2040			
Longer-term targets (2041 - onward)	*Not relevant*		N/A	N/A
Scope coverage	S1 S2 S3↑ S3↓	Not relevant		
Own emission reductions (compared to full value chain in 2019)	N/A			

3 REDUCING OWN EMISSIONS

			TRANSPARENCY	INTEGRITY
Operational emissions (scope 1)	39% of 2022 emissions	Enel commits to phase-out coal (2027) and gas (2040) and invests in renewables, but the pace is not quite sufficient to fully align with EU benchmarks.	●	◕
Renewable electricity (scope 2)	3% of 2022 emissions	*Not assessed.*	N/A	N/A
Upstream emissions (scope 3)	39% of 2022 emissions	Mitigation strategies apply also to electricity purchased and sold to end customers, but the pace is not quite sufficient to fully align with EU benchmarks.	◕	◕
Downstream emissions (scope 3)	17% of 2022 emissions	The company's strategy focuses on replacing fossil gas through electrification of its customers' energy demand. No investments in bioenergy are made.	●	◕

4 RESPONSIBILITY FOR UNABATED AND RESIDUAL EMISSIONS

		TRANSPARENCY	INTEGRITY
Climate contributions today (Beyond-value-chain mitigation)	*No climate contributions identified.*	◌	◌
Misleading offseting claims today	*No offsetting claim identified.*	N/A	N/A
Approach to residual emissions	Objective is to reach real zero, but depending on technologies. Enel foresees residual emissions of up to 2.5Mt in 2040, without a specific plan for CDR options.	◑	◑

5-point rating scale: ● High ◕ Reasonable ◑ Moderate ◔ Poor ◌ Very poor — **Transparency** refers to the disclosure of information. **Integrity** refers to the quality and credibility of the approach.

Fig. 2. Example of a CCRM report with color-coded ratings across different categories.

domain rubrics—defining discrete labeling criteria—yields outputs that closely mirror expert annotations. Our work builds on these insights by integrating CCRM scoring rubrics into each prompt, ensuring the model references precise, human-validated standards when generating labels.

While existing literature explores each component—manual frameworks, RAG, domain adaptation, prompt engineering—no prior work unifies them into a scalable, automated pipeline for corporate greenwashing detection. Our study fills this gap by systematically combining OCR, recursive chunking, RAG retrieval, and rubric-driven prompts to mimic expert assessments at scale.

III. Data Collection & Preprocessing

We invested heavily in collecting data and aligning more than 200 corporate sustainability reports into the raw evaluation reports from CCRM and TPI, which we hope will help motivate further research in the intersection of NLP and sustainability. To build a reliable ground-truth dataset and prepare inputs for our automated pipeline, we integrate three data sources—CCRM expert reports, the Transition Pathway Initiative (TPI) CSV, and raw corporate sustainability PDFs—through a multi-stage workflow:

A. 3.1 Corporate Sustainability Reports

For collecting the evidence used to generate the CCRM and TPI evaluation reports, we gathered more than 200 corporate sustainability reports from companies that were evaluated in CCRM, TPI, or both. Although companies evaluated by CCRM and TPI range a diverse set of sectors and countries, we prioritized our dataset collection in the following ways: 1) retrieving reports first from companies evaluated in CCRM due to the high-quality labels, and 2) retrieving reports belonging to companies in the tech industry. We retrieved corporate sustainability reports for companies based on the fiscal year from 2 years before the given evaluation report from CCRM or TPI; in other words, if CCRM evaluated the company Adidas in 2022, we collected the corporate sustainability report for Adidas from 2020.

Sustainability information for a company can come from many different sources including the company website, media reports, CDP responses, company blog posts, financial disclosures, ESG disclosures, and official sustainability reports. Moreover, the form and nomenclature of a company's sustainability report differed across companies, making it challenging to identify which released reports were most suitable for the task at hand. Most companies titled their reports either "Corporate Sustainability Report" or "Corporate Responsibility Report"

Although CCRM discloses the exact sources they used for assessments, TPI does not. We decided as a general rule of thumb to use the official company sustainability report when available as input to our dataset task, and sometimes deferred to ESG disclosures or annual reports when official sustainability reports were not available. We manually retrieved every individual company sustainability report and saved them as PDF files, to be converted into JSONL for input to the language models.

Using ClimateGPT's tokenizer, we provide a summary of the basic statistics regarding the document lengths within our collection of PDF corporate sustainability reports:

- **count**: 227 documents
- **mean**: 81,491 tokens
- **std**: 80,380 tokens
- **min**: 4,545 tokens
- **25% percentile**: 32,811 tokens
- **50% percentile**: 57,049 tokens
- **75% percentile**: 95,328 tokens
- **max**: 467,268 tokens

Note that even the shortest document has a total token length greater than the context window of ClimateGPT. More statistics describing the complete collection of corporate sustainability reports are described in section III-E.

B. 3.2 CCRM Reports (2022–2024)

The Corporate Climate Responsibility Monitor (CCRM) publishes annual PDF reports that award companies multidimensional scores on (i) greenhouse-gas emissions disclosure, (ii) the rigor of reduction targets, (iii) realized emissions reductions, and (iv) management of unabated emissions [1]. Figure 1 shows an example CCRM report, with evaluations for the company along the aforementioned axes. The different axes along which CCRM evaluates are explained as follows [2]:

- **Tracking and disclosure of emissions**: how transparent a company is about their GHG emission footprints and their trajectories
- **Setting specific and substantiated targets**: whether the company's commitments send a clear signal for immediate action to decarbonize the value chain and do not mislead consumers, regulators, and stakeholders
- **Reducing emissions**: whether the company makes encompassing measures for deep emission reductions
- **Responsibility for unabated and residual emissions**: whether the company takes responsibility for unabated emissions and avoiding misleading offsetting claims

Transparency refers to how openly a company discloses the information needed to assess the credibility of its climate responsibility efforts, whereas **integrity** is a measure of the quality, credibility and comprehensiveness of those approaches [2]. Figure 2 in the Appendix section shows an example of a report with relative high transparency and integrity ratings across the 4 sections.

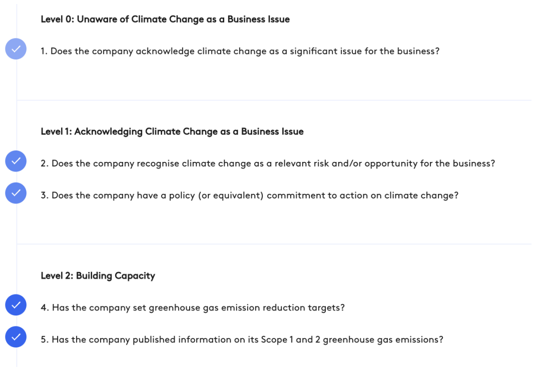
Assessments are available from 2022, 2023, and 2024, which typically draw corporate sustainability disclosures from the previous fiscal year(s). Across these three years, we processed 70 unique firm–year entries spanning energy, finance, consumer goods, and technology sectors. Each company's CCRM report combines narrative analysis, tabular data, and colored legend boxes. Figure 1 & 2 illustrates the graphical layout and colored rubric boxes. To extract structured labels:

- **Image Rendering.** Convert each chosen PDF page into a 300 DPI PNG using Poppler-based utilities [22]. This resolution balances legibility (for small-font tables) with GPU memory limits.
- **Color-Box Detection.** CCRM encodes transparency and integrity rubric categories via color-coded boxes (e.g., "High Transparency"). We applied pixel-region segmentation on RGB thresholds to detect these boxes and mapped each to its corresponding rubric label using nearest-neighbor matching in RGB space (Euclidean distance < 15), following standard color-based segmentation practices [23].
- **Text-Box Extraction.** After manually defining bounding-box templates for key sections (emissions tables, target summaries, governance notes), we batch-extracted these regions and applied OCR to retrieve the underlying text and numerical values [24].

All extracted text snippets and rubric labels were consolidated into a structured CSV file, yielding our ground-truth CCRM subset.

C. 3.3 Transition Pathway Initiative Harominization

The Transition Pathway Initiative (TPI) dataset offers moment-in-time binary responses (Yes/No) to 23 standardized climate-action questions, organized into six maturity levels (0–5) for over 1,200 global firms [4]. Figure 3 shows some example questions from a Transition Pathway Initiative evaluation.



Level 0: Unaware of Climate Change as a Business Issue

✓ 1. Does the company acknowledge climate change as a significant issue for the business?

Level 1: Acknowledging Climate Change as a Business Issue

✓ 2. Does the company recognise climate change as a relevant risk and/or opportunity for the business?

✓ 3. Does the company have a policy (or equivalent) commitment to action on climate change?

Level 2: Building Capacity

✓ 4. Has the company set greenhouse gas emission reduction targets?

✓ 5. Has the company published information on its Scope 1 and 2 greenhouse gas emissions?

Fig. 3. 5 Yes/No questions from Transition Pathway Initiative evaluating Adidas.

We ingested the March 2025 CSV and applied the following to prepare the TPI dataset for merging with the CCRM dataset:

1) **Name Canonicalization.** Standardize both CCRM "Name" and TPI "Company Name" by stripping whitespace, upper-casing, and removing punctuation. Use Levenshtein fuzzy matching (via Python's fuzzywuzzy or rapidfuzz) with a threshold $\geq 85\%$ similarity to align any remaining variants.
2) **Binary Encoding.** Map Yes:1, No:0 for each of 23 question columns. This numeric form will feed interpolation.

D. 3.4 Fuzzy Matching, Interpolation & Merging

1) **Outer Join.** Merge CCRM-extracted CSV and TPI DataFrame on (Standardized Company Name, Year) to preserve all rows from both sides.
2) **Filtering & Year Alignment.** Drop TPI rows whose company doesn't appear in any CCRM entry (we focus on firms with expert labels). For CCRM-only rows (no TPI), set TPI columns to NaN. For TPI-only rows (no CCRM), replicate the Year into "Assessment Date" and leave CCRM fields blank.
3) **Time-Series Interpolation.** For each company and each TPI question, sort records by Year; if there are at least two known numeric values, linearly interpolate missing years and round back to 0/1. Map interpolated 0/1 back to "No"/"Yes". Any gaps at the ends remain blank.

E. 3.5 Final Ground-Truth Dataset Schema

The resulting merged dataset comprises 227 firm–year entries, each containing:

- **CCRM Multi-Class Scores:** Transparency and integrity ratings across four pillars.
- **TPI Binary Responses:** Answers to 23 climate-action questions.
- **Metadata Fields:** Industry sector, headquarters country, fiscal year, and source provenance.

Geographically, 79.3% of firms are U.S.-based, followed by Germany (4.3%) and Japan (3.1%). Sector distribution is led by Consumer Services (27 firms), Technology (25), and Industrials (11). This richly annotated dataset underpins both our empirical evaluation of retrieval and prompt design, and the supervised fine-tuning of ClimateGPT for automated greenwashing detection.

IV. Methodology

We convert raw sustainability reports into structured CCRM-style assessments via six stages: text extraction, document chunking, embedding & retrieval, prompt engineering & inference, modeling variants, and results aggregation.

A. 5.1 Text Extraction with olmOCR

Instead of generic OCR (e.g. Tesseract), we leverage *olmOCR*, an LLM fine-tuned on complex PDF layouts (multi-column, sidebars, footnotes) [24]. We feed the entire corporate sustainability to `olmOCR` on NVIDIA A6000 GPUs (15–20 GB VRAM per document). The output is a JSONL of the PDF text. We then convert the JSONL into Langchain [25] documents where each document consists of:

- `metadata.source`: original filename & page ID
- `page_content`: concatenated text spanning columns, sidebars, and footnotes

B. 5.2 Document Structuring & Recursive Chunking

Due to limitations from LLM context windows and also hard limits given GPU memory, our models could not ingest 200-page reports directly (token lengths can be up to multiple hundreds of thousands). We therefore split each document into ≈2,200-token segments, preserving semantic boundaries:

Algorithm 1 Recursive Document Chunking

```
function SPLIT($D, T$)
    if tokenize$(D).len() \leq T$ then
        return $[D]$
    else
        $s \leftarrow$ find_last_paragraph_break$(D, T)$
        $(D_1, D_2) \leftarrow (D[: s], D[s :])$
        return SPLIT$(D_1, T) \cup$ SPLIT$(D_2, T)$
    end if
end function
$chunks \leftarrow$ SPLIT$(D, 2200)$
return ADDOVERLAP$(chunks, 200)$
```

We implement this via LangChain's `RecursiveTextSplitter (chunk_size=2200, overlap=200)` [25], which first tries to split at a double-newline, then at line breaks, whitespace, and finally at character level.

C. 5.3 Embedding & Retrieval

Each chunk is embedded into a 384-dimensional vector using the `all-MiniLM-L6-v2` [26] sentence transformer so that we can do cosine-similarity lookups. Regardless of the model's context limit or the query length, we retrieve a fixed top-$k = 10$ document chunks (roughly around $22,000$ total tokens) most semantically similar to q, and then remove document chunks depending on the specific model context length and query length. Specifically, for each model with context window length M, we dynamically retrieve $n \leq k$ documents, where $n \leq \frac{M - len(q)}{2200}$.

D. 5.4 Prompt Engineering & Inference

For each set of n document chunks, we build a structured prompt that embeds the full CCRM rubric:

> *Given the following passage, answer Question X:*
> **Rubric: 0=Very Poor; 1=Poor; 2=Moderate; 43=Reasonable; 4=High; -1=Unknown**
>
> `<concatenated top-`k` chunks>`
>
> *Output exactly one label from* {*very poor, poor, moderate, reasonable, high, unknown*}.

We run three LLMs via HuggingFace's `pipeline("text-generation")` in bfloat16 mode:

- **ClimateGPT-7B** [12]: context 4K tokens
- **Ministral-8B-Instruct-2410** [14]: context 128K tokens
- **Qwen2.5-7B-Instruct-1M** [13]: context 1M tokens

Hyperparameters: temperature = 0.7; max output = 256 tokens (CCRM) or 1 token (TPI). We batch one query per GPU but distribute companies across multiple GPUs for throughput.

E. 5.5 Modeling Variants & Experimental Setup

To disentangle effects of retrieval and prompting, we tested the following methods:

1) **Naive truncation**: feed first M tokens without retrieval
2) **Summarize-then-query**: generate document summary, then query
3) **RAG**: query n document chunks, join chunks together, then query

Due to time constraints, we were only able to evaluate all three LLMs on the RAG method. We ran some small sample evaluation using ClimateGPT on naive truncation and summarize-then-query methods, which we found to be less effective than RAG, most likely due to the limitations given the context length. We also experimented with adjusting our prompts with vs. without embedded rubrics.

F. 5.6 Results Aggregation & Evaluation

For each question, we select the first definitive label (non-“unknown”) across chunk responses. Numeric values (e.g. emissions) are extracted via the regex `\d[\d,.]+`, converted to floats, and averaged when multiple candidates occur. We report **Question-Level Accuracy**: fraction of matches vs. ground truth. Since the LLM can output an answer not in one of our category labels, we had to use fuzzy-matching to calculate accuracy when comparing with our ground-truth labels.

This structured methodology enables robust, scalable greenwashing detection across very large, heterogeneous reports while respecting LLM memory and context limits.

V. Experiments and Evaluation

We evaluate our RAG pipeline across three LLMs (ClimateGPT-7B, Ministral-8B-Instruct-2410, Qwen2.5-7B-Instruct-1M). As mentioned, we are unable to report our results on methods using naive truncation and summarize-then-query due to time constraints, but small samples showed that the results were inferior compared to RAG on ClimateGPT. Our primary metric us question-level accuracy. Table I breaks down performance across specific CCRM pillars.

The results show that ClimateGPT consistently outperforms larger-context models, despite its shorter 4K token window. Under RAG, ClimateGPT achieves up to 0.46 accuracy on Climate Contributions integrity, compared to essentially 0 for Qwen and Mistral.

We also evaluate the three LLMs on the questions from TPI, and see the same pattern. Note that the TPI questions are easier to answer because we specify to the model that the answer should be one of ”Yes” or ”No.” Table II breaks down performance across specific TPI questions.

TABLE I
CCRM Per-Pillar Transparency Accuracy (RAG)

Pillar	ClimateGPT	Mistral	Qwen
Overall Transparency	**0.26**	0.07	0.03
Overall Integrity	**0.25**	0.13	0.13
Emissions Disclosure	**0.19**	0.12	0.08
Emissions Reductions	**0.40**	0.13	0.06
Reduction Measures Trans.	**0.15**	0.14	0.14
Reduction Measures Integ.	**0.30**	0.03	0.04
Climate Contributions Trans.	**0.35**	0.03	0.06
Climate Contributions Integ.	**0.46**	0.01	0.00

TABLE II
TPI Per-Question Accuracy (RAG)

Question	ClimateGPT	Mistral	Qwen
Q1L0	0.95	0.27	**0.96**
Q2L1	**0.91**	0.20	0.85
Q3L1	**0.95**	0.19	0.87
Q4L2	**0.88**	0.13	0.71
Q5L2	**0.90**	0.13	0.76
Q6L3	**0.77**	0.03	0.03
Q7L3	**0.94**	0.08	0.70
Q8L3	**0.93**	0.13	0.5
Q9L3	**0.88**	0.03	0.14
Q10L3	0.60	0.1	**0.62**
Q11L3	**0.79**	0.12	0.5
Q12L3	0.13	0.01	**0.14**
Q13L4	**0.88**	0.08	0.55
Q14L4	**0.57**	0.1	0.38
Q15L4	**0.63**	0.09	0.57
Q16L4	**0.72**	0.07	0.35
Q17L4	**0.51**	0.15	0.46
Q18L4	**0.58**	0.06	0.46
Q19L5	0.12	0.13	**0.75**
Q20L5	0.13	0.13	**0.92**
Q21L5	0.05	0.21	**0.97**
Q22L5	0.05	0.17	**0.95**
Q23L5	0.25	0.13	**0.8**

We see the same pattern that ClimateGPT outperforms the other two models in answering most questions.

VI. Discussion

Our comprehensive evaluation across models and retrieval strategies highlights several important findings:

A. Retrieval Focus Outperforms Raw Context Length

Although Qwen and Mistral offer much larger context windows (up to 1M and 128K tokens), simply feeding them more of the raw report did not produce better scores. Instead, selectively retrieving the most semantically relevant 2.2 K-token chunks for each question yielded higher alignment with ground truth. In particular, ClimateGPT—despite its modest 4K token window—achieved the best performance in seven out of eight CCRM pillars. This suggests that previous domain knowledge and instruction tuning may be more important than the long context window for document analysis.

B. Rubric Embedding Dramatically Reduces Hallucinations

We observed that, without explicit rubric definitions embedded in the prompt, model outputs became inconsistent and prone to "creative" but incorrect answers. By placing the full CCRM scoring criteria (e.g. "High.... Very Poor") directly in the system prompt, we anchored the models' reasoning to a fixed, expert-driven standard. In our initial sample analysis, removing rubric text caused ClimateGPT to often produce outputs which were sometimes not an answer to the question at all. This finding underlines the importance of clarity of instructions.

C. Model Size vs. Pipeline Design

While larger models like Qwen and Mistral can handle more text, their performance lagged behind ClimateGPT when all were paired with our RAG pipeline. We attribute this to two factors: (1) diminishing returns from longer contexts once relevant passages are isolated, and (2) potential mismatch between general-purpose pretraining and domain-specific climate reasoning. Fine-tuning on CCRM-style data or integrating chain-of-thought techniques may close this gap, but our results make clear that even smaller, well-prompted models can excel when guided by precise retrieval.

D. Limitations

- **Data Sparsity:** Companies with minimal public disclosures generate fewer relevant chunks, reducing model recall.
- **OCR Errors:** Complex tables or scanned images sometimes yield mis-parsed numbers.
- **Throughput Constraints:** Because of large context sizes, we batch only one query per GPU; scaling to enterprise-level fleets will require asynchronous pipelines and more efficient parallelization.

VII. Conclusion and Future Work

We have presented a robust, scalable NLP pipeline for automated CCRM-style greenwashing detection that successfully handles very large, heterogeneous reports. Through careful integration of high-fidelity OCR, semantic chunking, targeted RAG retrieval, and rubric-driven prompting, our system achieves up to 26% question-level accuracy.

Looking ahead, we envision several enhancements:

1) **Multimodal Fusion:** Incorporate vision-language models to jointly reason over report text and embedded charts or tables, improving performance on qualitative pillars.
2) **Domain Fine-Tuning:** Leverage our ground-truth dataset to fine-tune ClimateGPT (and other LLMs) directly on CCRM-style assessments, potentially boosting both accuracy and consistency.
3) **Incremental Indexing:** Develop an online FAISS pipeline that ingests new reports as they are published, enabling continuous, real-time ESG monitoring.
4) **Advanced Prompting Techniques:** Experiment with chain-of-thought prompting and few-shot exemplars for subjective judgments, with the goal of further narrowing the gap between automated and expert evaluations.

By open-sourcing our code, prompt templates, and vector indexes, we aim to catalyze a community effort toward transparent, scalable, and trustworthy greenwashing detection.

References

[1] NewClimate Institute, "Corporate Climate Responsibility Monitor 2024 Report," 2024. [Online]. Available: https://newclimate.org/CCRM2024

[2] NewClimate Institute, "Corporate Climate Responsibility: Guidance and Assessment Criteria for Good Practice Corporate Emission Reduction and Net-Zero Targets," 2024. [Online]. Available: https://newclimate.org/sites/default/files/2024-04/NewClimate_CCRM2024_Methodology.pdf

[3] Science Based Targets, "Ambitious corporate climate action" [Online]. Available: https://sciencebasedtargets.org/

[4] Transition Pathway Initiative, "Company Framework and Questions," 2024. [Online]. Available: https://transitionpathwayinitiative.org

[5] S&P Global, ESG Scores and Raw Data. [Online]. Available: https://www.spglobal.com/esg/solutions/esg-scores-data

[6] A. Amel-Zadeh and G. Serafeim, "Why and how investors use ESG information: Evidence from a global survey," in *Financial Analysts Journal*, 2018.

[7] T. A. Da Costa *et al.*, "ClimateQ&A, AI-powered conversational assistant for climate change and biodiversity loss," 2024. [Online]. Available: https://huggingface.co/spaces/Ekimetrics/climate-question-answering

[8] D. Stammbach *et al.*, "A dataset for detecting real-world environmental claims," arXiv:2209.00507, 2022.

[9] G. Morio and C. D. Manning, "An NLP Benchmark Dataset for Assessing Corporate Climate Policy Engagement," in *Neural Information Processing Systems*, 2023.

[10] N. Webersinke *et al.*, "ClimateBERT: A Pretrained Language Model for Climate-Related Text", in *AAAI*, 2022.

[11] J. A. Bingler *et al.*, "Cheap talk and cherry-picking: What ClimateBert has to say on corporate climate risk disclosures," in *Finance Research Letters*, 2022.

[12] J. A. Smith *et al.*, "ClimateGPT: A Generative Model for Climate Science Summarization," in *ACL*, 2024.

[13] A. Yang *et al.*, "Qwen2.5-1M Technical Report," arXiv:2501.15383, 2025.

[14] Mistral AI team, "Un Ministral, des Ministraux," 2024. [Online] Available: https://mistral.ai/news/ministraux

[15] P. Lewis *et al.*, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *NeurIPS*, 2021.

[16] Snowflake Engineering Blog, "Impact of Retrieval Chunking on Finance RAG Models," 2023. [Online]. Available: https://www.snowflake.com/en/engineering-blog/impact-retrieval-chunking-finance-rag/

[17] Y. Zhao *et al.*, "Optimizing LLM Based Retrieval Augmented Generation Pipelines in the Financial Domain," in *ACL*, 2024.

[18] H. Li *et al.*, "LexRAG: Benchmarking Retrieval-Augmented Generation in Multi-Turn Legal Consultation Conversation," arXiv:2502.20640, 2025.

[19] Y. Shi *et al.*, "MKRAG: Medical Knowledge Retrieval Augmented Generation for Medical Question Answering," arXiv:2309.16035, 2023.

[20] T. Brown *et al.*, "Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm," arXiv:2302.04023, 2023.

[21] L. Gao *et al.*, "Embedding Domain Rubrics in LLM Prompts for Structured Outputs," in *ICLR*, 2023.

[22] Poppler Contributors, "Poppler PDF Rendering Library," 2024. [Online]. Available: https://poppler.freedesktop.org/

[23] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 4th ed., Pearson, 2018.

[24] J. Poznanski *et al.*, "olmOCR: Unlocking Trillions of Tokens in PDFs with Vision Language Models,“ arXiv:2502.18443, 2025.

[25] Langchain, Langchain. [Online]. Available: https://www.langchain.com/

[26] sentence-transformers/all-MiniLM-L6-v2 [Online] Available: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Appendix

This appendix enumerates the rubric categories derived from the Corporate Climate Responsibility Monitor (CCRM) and the binary question labels from the Transition Pathway Initiative (TPI) dataset, as used in our prompt engineering and evaluation.

A. CCRM Rubric Categories for emissions disclosure

- **Very Poor**: No emissions scope tracked or disclosed, or target-base-year data missing entirely.
- **Poor**: Major emission sources disclosed but at least one material source missing or aggregated.
- **Moderate**: Annual emissions disclosure; breakdown by specific sources (limited granularity); at least one prior year of historical data; explanation for omitted sources; non-GHG climate forcers disclosed; market- and location-based estimates; baseline-year data provided or inferred.
- **Reasonable**: Meets all "High" criteria except emission aggregates use lowest estimate rather than highest.
- **High**: Comprehensive annual emissions disclosure; detailed breakdown by individual Scope 1/2/3 categories; multi-year historical data (2 prior years); explicit justification for omitted sources; disclosure of non-GHG forcers; both market- and location-based estimates using the highest values.
- **Unknown**: Evidence too sparse or conflicting to assign another score.

B. Transition Pathway Initiative (TPI) Question Categories

The TPI framework organizes 23 binary (Yes/No) questions into six maturity levels:

1) **Level 0: Unaware** – Company does not acknowledge climate change as a business issue.
2) **Level 1: Acknowledgment** – Recognizes climate change as a risk and/or opportunity.
3) **Level 2: Capacity Building** – Has policies or commitments to address climate change; sets emission reduction targets.
4) **Level 3: Operational Integration** – Reports Scope 1/2 emissions; has management processes; board oversight; third-party verification; supports mitigation efforts.
5) **Level 4: Strategic Assessment** – Quantitative targets for GHG reduction; integrates risks and opportunities into strategy; links executive remuneration to climate performance; conducts scenario planning; discloses internal carbon pricing and transition actions.
6) **Level 5: Transition Planning** – Commits to phasing out carbon-intensive assets; aligns capital expenditure with decarbonization; ensures consistency between climate positions and trade association memberships.

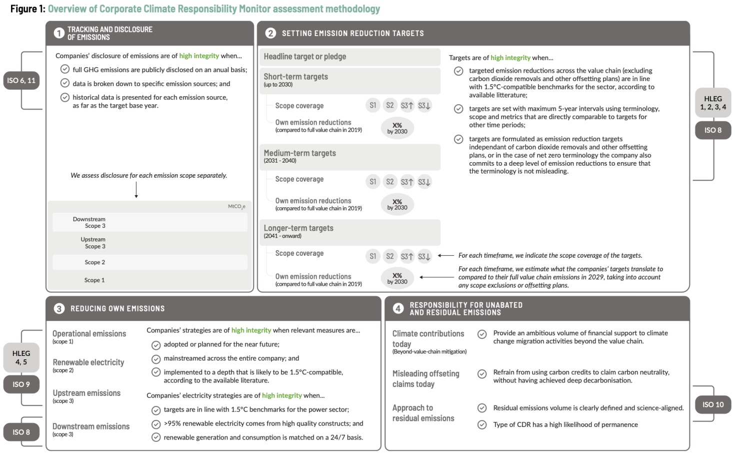


Figure 1: Overview of Corporate Climate Responsibility Monitor assessment methodology
1 TRACKING AND DISCLOSURE OF EMISSIONS
ISO 6, 11
MtCO$_2$e
Downstream Scope 3
Upstream Scope 3
Scope 2
Scope 1
2 SETTING EMISSION REDUCTION TARGETS
Headline target or pledge
Short-term targets (up to 2030)
Medium-term targets (2031 - 2040)
Longer-term targets (2041 - onward)
Scope coverage
S1
S2
S3↑
S3↓
Own emission reductions (compared to full value chain in 2019)
X% by 2030
HLEG 1, 2, 3, 4
ISO 8
3 REDUCING OWN EMISSIONS
Operational emissions (scope 1)
Renewable electricity (scope 2)
Upstream emissions (scope 3)
Downstream emissions (scope 3)
HLEG 4, 5
ISO 9
ISO 8
4 RESPONSIBILITY FOR UNABATED AND RESIDUAL EMISSIONS
Climate contributions today (Beyond-value-chain mitigation)
Misleading offseting claims today
Approach to residual emissions
ISO 10

Companies' disclosure of emissions are of high integrity when...
- ✓ full GHG emissions are publicly disclosed on an anual basis;
- ✓ data is broken down to specific emission sources; and
- ✓ historical data is presented for each emission source, as far as the target base year.

We assess disclosure for each emission scope separately.

Targets are of high integrity when...
- ✓ targeted emission reductions across the value chain (excluding carbon dioxide removals and other offsetting plans) are in line with 1.5°C-compatible benchmarks for the sector, according to available litterature;
- ✓ targets are set with maximum 5-year intervals using terminology, scope and metrics that are directly comparable to targets for other time periods;
- ✓ targets are formulated as emission reduction targets independant of carbon dioxide removals and other offsetting plans, or in the case of net zero terminology the company also commits to a deep level of emission reductions to ensure that the terminology is not misleading.

For each timeframe, we indicate the scope coverage of the targets.

For each timeframe, we estimate what the companies' targets translate to compared to their full value chain emissions in 2029, taking into account any scope exclusions or offsetting plans.

Companies' strategies are of high integrity when relevant measures are...
- ✓ adopted or planned for the near future;
- ✓ mainstreamed across the entire company; and
- ✓ implemented to a depth that is likely to be 1.5°C-compatible, according to the available literature.

Companies' electricity strategies are of high integrity when...
- ✓ targets are in line with 1.5°C benchmarks for the power sector;
- ✓ >95% renewable electricity comes from high quality constructs; and
- ✓ renewable generation and consumption is matched on a 24/7 basis.

- ✓ Provide an ambitious volume of financial support to climate change migration activities beyond the value chain.
- ✓ Refrain from using carbon credits to claim carbon neutrality, without having achieved deep decarbonisation.
- ✓ Residual emissions volume is clearly defined and science-aligned.
- ✓ Type of CDR has a high likelihood of permanence

Fig. 4. Example of corporation with high ratings.

C. CCRM example of good practice corporation

See Figure 4 above.