Automating Corporate Greenwashing Detection Using Natural Language Processing

Nicole Lin and Amy Pu

May 12, 2025

1 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 for multi-label classification 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.

2 Introduction

Over the last ten years, increasing pressure from regulators, investors, and consumers has pushed many companies to make bold sustainability pledges—often without matching action. This "greenwashing" can take the form of lofty net-zero targets alongside growing fossil-fuel operations, or spotlighting small recycling programs while ignoring major emission sources. Such gaps between words and deeds mislead stakeholders and stall real climate progress.

Producing expert reviews like the Corporate Climate Responsibility Monitor (CCRM) provides deep, four-part ratings (disclosure, targets, reductions, unabated emissions) but takes months of manual effort, making it impossible to watch thousands of firms continuously. At the same time, modern NLP tools and large language models (LLMs) offer a path to automate parts of that workflow—but large, variably formatted reports and context-window limits pose major challenges.

We address these issues by building:

1. **A labeled dataset** of 200+ sustainability reports, with assessment scores drawn from CCRM and the Transition Pathway Initiative.
2. **A scalable pipeline** that:
   - Uses GPU-accelerated OCR for reliable text extraction
   - Chunks documents recursively to stay within token limits
   - Applies Retrieval-Augmented Generation (RAG) to zero in on key passages
   - Embeds formal scoring rubrics directly into prompts
3. **An evaluation** of three adapted LLMs (ClimateGPT, Mistral, Qwen), achieving up to 26% accuracy on transparency scoring.

Our work demonstrates a first step toward lightweight, automated monitoring of corporate climate claims, cutting manual review time while retaining expert-level rigor.

3 Related Work

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

[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.

and prompt engineering strategies.

3.1 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.

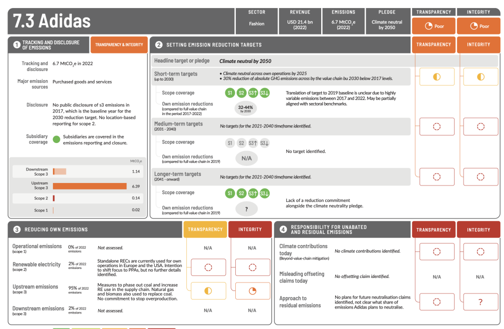

Figure 1: Example of a CCRM report with color-coded ratings across different categories.

3.2 Corporate Sustainability NLP Research

There has been developing research interest at the intersection of NLP and climate related tasks. Ekimetrics developed Climate Q&A (5), a conversational assistant for question-answering, which uses data from over thousands of pages of scientific reports to generate quick answers. Stammbach et al. (6) 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 (7) 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.

3.3 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 (8). It has also been used for analyzing companies' climate-risk disclosures along the Task Force for Climate-related Financial Disclosures (TCFD) categories (9). 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 (10). These models underline the value of domain adaptation but stop short of automating structured rubric-based scoring.

3.4 Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation combines neural text generation with vector-based retrieval to ground LLM outputs in external knowledge sources (13). 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 (14).

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 (15) (16) (17). However, prior work has typically focused on short-form queries or summary-generation rather than structured rubric-based scoring as needed for climate disclosures.

3.5 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. (18) demonstrate that carefully designed "system messages" and formatting cues significantly reduce LLM hallucinations. Gao et al. (19) further show that embedding 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.

4 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:

4.1 Corporate Sustainability Reports

We collected 227 sustainability reports from companies covered by CCRM, TPI, or both. To build our core dataset, we first targeted those firms with CCRM evaluations—since their expert-curated labels are the most granular—and then supplemented with tech-sector companies to match our ML-and-climate focus. For each firm, we pulled the report published two fiscal years before its CCRM/TPI score (e.g., Adidas's 2020 report for its 2022 evaluation).

These documents came from a mix of official "Corporate Sustainability Report" or "Corporate Responsibility Report" PDFs, ESG disclosures, annual financial filings, CDP submissions, and even company blogs or press releases when no formal sustainability report existed. Because naming conventions and formatting varied so widely, we manually inspected each candidate, downloaded the best-fit PDF, and standardized filenames before converting them into JSONL via our OCR pipeline.

Once all 227 PDFs were in JSONL form, we ran ClimateGPT's tokenizer over the full text to get token counts. The resulting distribution ranged from 4,500 tokens (shorter slide-style reports) up to nearly 470,000 tokens (dense, graphics-heavy PDFs), with a median around 57,000.

Table 1: Token Statistics Across Documents

Count	Mean	Std	Min	25%	50%	75%	Max
227	81,491	80,380	4,545	32,811	57,049	95,328	467,268

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 4.5.

4.2 CCRM Reports (2022–2024)

The Corporate Climate Responsibility Monitor (CCRM) publishes annual PDF reports that award companies multi-dimensional 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).

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 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 (20). 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 (21).
- **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 (22).

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

4.3 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).

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.

4.4 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.

4.5 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.

5 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.

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)(22). 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 (23) documents where each document consists of metadata and page content.

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:

We implement this via LangChain's `RecursiveTextSplitter (chunk_size=2200, overlap=200)` (23), which first tries to split at a double-newline, then at line breaks, whitespace, and finally at character level.

5.3 Embedding & Retrieval

Each chunk is embedded into a 384-dimensional vector using the `all-MiniLM-L6-v2`(24) 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 $n \leq \frac{M - len(q)}{2200}$.

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** (10): context 4K tokens
- **Ministral-8B-Instruct-2410** (12): context 128K tokens
- **Qwen2.5-7B-Instruct-1M** (11): 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.

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.

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** which is the 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.

6 Experiments and Evaluation

We evaluate our RAG pipeline across three LLMs (ClimateGPT-7B, Ministral-8B-Instruct-2410, Qwen2.5-7B-Instruct-1M). Due to time constraints, we are unable to report our results on methods using naive truncation and summarize-then-query, but small samples showed that the results were inferior compared to RAG on ClimateGPT. Our primary metric is question-level accuracy. Table 2 breaks down performance across specific CCRM pillars.

Table 2: 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

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 3 and Table 4 breaks down performance across specific TPI questions.

Table 3: TPI Per-Question Accuracy (RAG) — (Q1–Q12)

Model	Q1L0	Q2L1	Q3L1	Q4L2	Q5L2	Q6L3	Q7L3	Q8L3	Q9L3	Q10L3	Q11L3	Q12L3
ClimateGPT	0.95	**0.91**	**0.95**	**0.88**	**0.90**	**0.77**	**0.94**	**0.93**	**0.88**	0.60	**0.79**	0.13
Mistral	0.27	0.20	0.19	0.13	0.13	0.03	0.08	0.13	0.03	0.10	0.12	0.01
Qwen	**0.96**	0.85	0.87	0.71	0.76	0.03	0.70	0.50	0.14	**0.62**	0.50	**0.14**

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

Table 4: TPI Per-Question Accuracy (RAG) — (Q13–Q23)

Model	Q13L4	Q14L4	Q15L4	Q16L4	Q17L4	Q18L4	Q19L5	Q20L5	Q21L5	Q22L5	Q23L5
ClimateGPT	**0.88**	**0.57**	**0.63**	**0.72**	**0.51**	**0.58**	0.12	0.13	0.05	0.05	0.25
Mistral	0.08	0.10	0.09	0.07	0.15	0.06	0.13	0.13	0.21	0.17	0.13
Qwen	0.55	0.38	0.57	0.35	0.46	0.46	**0.75**	**0.92**	**0.97**	**0.95**	**0.80**

7 Discussion

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

7.1 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.

7.2 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.

7.3 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.

7.4 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.

8 Conclusion and Future Work

We have presented a 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] 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

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

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

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

[9] 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.

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

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

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

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

[14] 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/

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

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

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

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

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

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

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

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

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

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