

Project proposal

● Graded

Group

Sivaramakrishnan H

Sai Prathik Ravipalli

James Bardowski

...and 2 more

 [View or edit group](#)

Total Points

100 / 100 pts

Question 1


Proposal

 100 / 100 pts

✓ - 0 pts Correct

- 10 pts The proposal should be at least 3 full pages of content, excluding references and AI disclosure.

- 5 pts Did not complete AI disclosure "Free Response" section

 The proposed project aims to investigate how companies engage in "greenwashing" through their sustainability reports using textual analysis techniques. The idea is well-motivated and really interesting!

The methods are really comprehensive, considering both encoder- and decoder-based model. It would be insightful to compare the performance of these models on your specific tasks. To fine-tune the Mistral model using LoRA, you might find Unclot (https://github.com/unslothai/unsloth) useful.

Once you obtain all the inference results, it could be helpful to analyze the correlation between the company scores generated by your method and an external ranking, such as Newsweek's "America's Most Responsible Companies" for 2024 (https://www.newsweek.com/rankings/americas-most-responsible-companies-2024). A strong correlation can be a strong external validation for your method.

Using pdfminer to extract text data from pdfs is a good idea, but you should also manually verify some of the documents to make sure that there are no gibberish text.

Overall, the project is really promising, and I look forward to reading your final report!

— Chau

Question assigned to the following page: [1](#)

Evaluating Scope 3 Emissions Claims through Language Models

Sai Prathik Ravipalli

sravipalli@umass.edu

Neal Chokshi

nchokshi@umass.edu

Sivaramakrishnan H

shariharasub@umass.edu

James Bardowski

jbardowski@umass.edu

Themis Splagounias

tsplagounias@umass.edu

1 Introduction

In the present era of climate crisis awareness, the importance of environmental responsibility within the corporate world cannot be overstated. While many companies tout their commitment to sustainability, a phenomenon known as "greenwashing" has emerged, where organizations exaggerate or misrepresent their environmental efforts to appear more eco-friendly than they truly are. They resort to greenwashing in their sustainability reports, primarily to create a misleadingly positive environmental image in order to benefit from growing investor and consumer interest. Our proposal outlines a method to analyze corporate emissions reports using language models, detecting greenwashing instances and providing metrics for the transparency of environmental claims.

While significant research has been conducted to explore greenwashing, especially in the context of ESG (Environmental, Social, Governance) reports, there is little work done on investigating the claims made related to Scope 3 emissions. Scope 3 emissions are emissions which are not directly produced by the company or its owned assets, but rather from upstream and downstream activities the company is indirectly responsible for. (Greenhouse Gas Protocol, 2023).

Scope 3 emissions, though indirect, constitute a staggering 90% of an average company's carbon footprint according to CDP (2020). Their accurate estimation and reporting are thus paramount, yet pose a considerable challenge due to their indirect nature. Thus, we introduce a method to extract categorical Scope 3 emissions data from publicly available sustainability reports and assign ratings to claims made about Scope 3 based on their specificity, to ultimately construct a transparency and Scope 3 effort score for each company. We will employ existing models on relevant tasks as base-

lines, determine their abilities for Scope 3 statement detection and vagueness analysis, and then experiment with fine tuning models to evaluate what achieves the best results.

2 Related work

LMs For Greenwashing Detection. A method to detect greenwashing was proposed by Vinella et al. (2023) using ClimateBERT (Webersinke et al., 2022) along with an expert human-annotated dataset to identify instances of greenwashing or nonspecific claims. Their model was trained to detect a small set of attributes that contribute to greenwashed text, which is then used as input to an equation which assigns each company a greenwashing risk score as "ground-truth" labeling. This research highlights the potential for using encoder-only models with fine-tuning to classify text specificity. We plan to experiment with various fine-tuning methods and compare performance in detecting vague text.

Sustainability Report Evaluation. There have been multiple efforts across the research community to employ LLMs on corporate sustainability reports in order to extract data and gauge the language used to describe their efforts. Ni et al. (2023) presents CHATREPORT, an LLM-based tool which aims to democratize long sustainability reports and includes a novel scoring system to assess how closely companies conform to TCFD (Task Force on Climate related Financial Disclosures) standards. It additionally provides a method for users to interact with an LLM to ask specific questions about the report.

Bronzini et al. (2024) proposes a method for structured data extraction from sustainability reports using LLMs. A complex data preparation pipeline including a PDF parser and sentence segmenting is used to extract structured data from un-

Question assigned to the following page: [1](#)

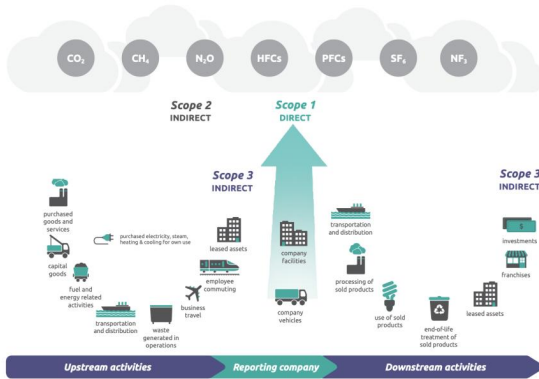


Figure 1: Categories of Scope 3 Emissions (Greenhouse Gas Protocol, 2023)

structured PDFs. They also present a method to sort claims within ESG categories using a semantic text search, useful for constructing a claim categorization pipeline.

Zou et al. (2023) presents another method for ESG data extraction from PDF reports using LLMs. Using a pre-processing stage, they were able to effectively structure report data for use as a knowledge base which an LLM can use to support ESG data retrieval. These methods further the possibilities of enabling LLMs access to unstructured data, specifically through PDF processing tools, and we will be referencing them as we conduct our data collection tasks.

ESG Standards and Scope 3. Emphasis on ESG reporting, particularly in greenhouse gas (GHG) emissions, has led to widespread adoption of the GHG Protocol Corporate Standard (Greenhouse Gas Protocol Initiative, 2004). This standard provides a framework for companies to calculate and categorize their emissions. This protocol also delineates Scope 3 emissions into 15 distinct categories (Greenhouse Gas Protocol, 2023), emphasizing the complexities within indirect emissions accounting. [1] Our Scope 3 analysis will be done with regard to these categories.

Scope 3 emission estimation was proposed by Jain et al. (2023), where a novel LLM based text mining method was described in order to extract financial transaction information from enterprise ledger transaction records as pertinent to Scope 3. This data was then used to estimate a carbon footprint. A study by Serafeim and Velez Caicedo (2022) with a similar estimation objective was conducted using publicly available financial statements to estimate Scope 3 within 15 emission cat-

egories. Both studies provide valuable insights into Scope 3 accountability, but we believe reports created by corporations should also be analyzed as it is a more accessible resource for interested parties.

ESGBert. We plan to utilize ESGBert, a domain-specific language model shown to outperform baseline models in ESG classification tasks (Mehra et al., 2022).

3 Approach

Our main task is to utilize a deep learning model to determine if companies are reporting on Scope 3 emissions and evaluate the specificity of claims they are making. Although off-the-shelf models such as GPT-4 or Gemini may be suitable for our purposes, these models do not present a cost-effective solution to the problem. Given the vast number of publicly listed companies, with an estimated 4,000 in the U.S. and over 50,000 globally (World Federation of Exchanges, 2022), the cost of using a publicly available model to analyze corporate reports presents significant challenges. Assuming an average report spans 20,000 words, employing GPT-4 for both Scope 3 detection and vagueness identification would incur substantial costs, amounting to \$1,600 for US companies and exceeding \$16,000 for the analysis of reports from all companies worldwide within a one-year time frame (based on GPT-4 pricing data at the time of this proposal). We seek to create a model of smaller scale that can operate at significantly reduced costs, yet rival the results of a general purpose public LLM.

For achieving this goal we split the task into steps detailed in the following sub-sections:

3.1 Data Collection and Extraction

Corporations are increasingly required to release ESG-related reporting annually to communicate their ESG performance to investors. These reports cover a multitude of topics, and only a small subset of an overall report discusses Scope 3 emissions. Due to a lack of requirements around the reporting format, companies do not present information in a standardized manner, making data extraction more involved. We plan to use PDF parsing tools like pdfminer to extract the text data before we feed it into our model. Tabular data, images, and text order will be ignored to ease the preprocessing effort.

Question assigned to the following page: [1](#)

3.2 Data Annotation

We are planning to annotate text from 10 Corporate ESG reports which would give us over 10k sentences. Parsed text from 10 ESG reports will be annotated using GPT4(gpt-4-0125-preview) at sentence/paragraph level on two dimensions:

- Whether the text is about Scope 3 emissions and if so which subcategory it belongs to. There are 15 subcategories for Scope 3 emissions.
- The specificity level of each claim. We will rate claims based on one of three classes: specific, ambiguous, generic. In "specific" statements, there might concrete goals, timelines or strategies mentioned which are clear and can be measure while vague statements would lack some or most of these details making it unclear how the company would achieve its environment goals. For example a concrete statement is "By 2030, we plan to cut 30% of our emissions by using renewable sources of energy" while a vague statement would be like "We plan to reduce our emissions in the coming years".

Additionally, a pre-existing dataset to identify vague claims in ESG text (Kumar, 2023) will be adapted using our criteria for further finetuning and testing to support our custom dataset.

If the amount of data proves to be insufficient for the task, attempts to augment it with techniques such as back translation, synonym replacement, text paraphrasing will be made.

3.3 Baseline

We plan to use one LLM each with different model sizes: Phi2, Gemma 7B, Qwen1.5 14B, Mistral to create baselines for our classification tasks. These models, while smaller than GPT-4, can still be effective with in-context learning techniques. This approach will enable us to assess performance and cost trade-offs.

3.4 Model

3.4.1 Encoder Based Model

Because we are targeting two related tasks, the approach to modeling can bifurcate into two distinct methodologies:

- Joint Training with Multi-Task Objective: This method entails the simultaneous training of a single model to address both tasks

concurrently. By optimizing a shared objective function that encompasses the goals of both tasks, the model learns to capture the dependencies and correlations between them, potentially leading to enhanced performance and efficiency.

- Individual Model Training: Training separate models dedicated to each task independently. This approach allows for tailored optimization and fine-tuning specific to the requirements and nuances of each task. Although it may require more computational resources and maintenance overhead, it offers flexibility in model architecture and training parameters, potentially yielding optimized performance for each individual task.

We plan to experiment with various fine-tuning methods and explore the capabilities of models tuned to do multitask classification against two single task models. Since they are related tasks multitask training can be beneficial (Caruana, 1997), but detecting vagueness may have a different distribution of classes than Scope 3 emissions, which we predict can negatively affect multitask learning.

Network Architecture: BERT with the classification layer. We will be fine-tuning "bert-base-en", a specific BERT model, which has been pre-trained on ESG data available on the HuggingFace Model Hub (Mehra et al., 2022). The domain-trained model can provide improved performance (Beltagy et al., 2019).

Loss: There are two tasks to complete, both of which are classification based; and both can be trained with a classification loss function.

3.4.2 Decoder Based Model

We will also train a decoder based model using a PEFT fine-tuning method like LoRA.

Network Architecture: Mistral 7B.

Loss: Same loss as language modelling, i.e. cross entropy loss.

3.5 Inference and Scoring

Once we have selected the model based on the evaluation on the test set, we will run inference on the remaining corporate ESG reports data. We

Question assigned to the following page: [1](#)

can then score companies using a simple heuristic based on whether they are reporting about Scope 3 emissions, how many categories are addressed, and how specific are the claims.

3.6 Schedule

We have decided to work collectively on all the sub tasks below.

1. Acquire data, parse data, annotate data and set baseline (3-4 weeks)
2. Multitask Encoder Model and Decoder Model Training (2 weeks)
3. Analyze the results to score companies (3-4 weeks)
4. Work on final reports (1 weeks)

4 Data

- Sustainability Reports - Sustainability reports will be gathered from ResponsibilityReports.com. We plan on collating and parsing reports for the top 50 companies by market cap, then running inference on them.(ResponsibilityReports, 2024)
- Scope 3 Emission category - There are no annotated datasets available for this.
- Vagueness in Scope 3 statements - There is annotated data on whether a ESG statement is vague or not (Kumar, 2023). This dataset contains 1.1k statements and their vagueness classification. There are also other vague sentence datasets to complement this data.

5 Tools

We will employ the pdfminer Python library to parse PDF documents, enabling us to extract text and structural information from them efficiently. For training and finetuning our models, we will use PyTorch.

To enhance the quality of our dataset, we will utilize the OpenAI APIs, leveraging their advanced natural language processing capabilities to annotate our data.

Additionally, we will integrate the Together API into our workflow to establish baselines utilizing LLMs, providing valuable benchmarks for our

model performance evaluation. To execute our experiments and develop our codebase we will utilize Google Colab, a cloud-based platform that offers free access to GPUs and TPUs, facilitating efficient model training and experimentation. We also have personal GPUs which can be used for finetuning if required.

6 AI Disclosure

- Did you use any AI assistance to complete this proposal? If so, please also specify what AI you used.
 - Yes, ChatGPT 3.5

If you answered yes to the above question, please complete the following as well:

- If you used a large language model to assist you, please paste *all* of the prompts that you used below. Add a separate bullet for each prompt, and specify which part of the proposal is associated with which prompt.
 - What is the seminal paper for Multi task learning?(used for citing Caruana, 1997 paper)
- **Free response:** For each section or paragraph for which you used assistance, describe your overall experience with the AI. How helpful was it? Did it just directly give you a good output, or did you have to edit it? Was its output ever obviously wrong or irrelevant? Did you use it to generate new text, check your own ideas, or rewrite text?
 - It was difficult to identify which paper to cite for some of the technical tasks. We tried asking GPT and only once the answer was correct.

References

- Beltagy, I., Lo, K., and Cohan, A. (2019). SciBERT: A pre-trained language model for scientific text. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.
- Bronzini, M., Nicolini, C., Lepri, B., Passerini, A., and Staliano, J. (2024). Glitter or gold? deriving structured insights from sustainability reports via large language models.

Question assigned to the following page: [1](#)

- Caruana, R. (1997). Multitask learning. *Machine Learning*, 28(1):41–75.
- CDP (2020). Cdp scope 3 emissions report 2020.
- Greenhouse Gas Protocol (2023). Scope 3 Calculation Guidance. Technical report.
- Greenhouse Gas Protocol Initiative (2004). The Greenhouse Gas Protocol: A Corporate Accounting and Reporting Standard.
- Jain, A., Padmanaban, M., Hazra, J., Godbole, S., and Weldemariam, K. (2023). Supply chain emission estimation using large language models. *arXiv preprint arXiv:2308.01741*.
- Kumar, A. (2023). Esg prospectus clarity category.
- Mehra, S., Louka, R., and Zhang, Y. (2022). Esgbert: Language model to help with classification tasks related to companies environmental, social, and governance practices. *arXiv preprint arXiv:2203.16788*.
- Ni, J., Bingler, J., Colesanti-Senni, C., Kraus, M., Gostlow, G., Schimanski, T., Stambach, D., Vaghefi, S. A., Wang, Q., Webersinke, N., et al. (2023). Chatreport: Democratizing sustainability disclosure analysis through llm-based tools. *arXiv preprint arXiv:2307.15770*.
- ResponsibilityReports (2024). The most complete and up-to-date listings of responsibility reports on the internet.
- Serafeim, G. and Velez Caicedo, G. (2022). Machine learning models for prediction of scope 3 carbon emissions. *Available at SSRN*.
- Vinella, A., Capetz, M., Pattichis, R., Chance, C., and Ghosh, R. (2023). Leveraging language models to detect greenwashing.
- Webersinke, N., Kraus, M., Bingler, J., and Leippold, M. (2022). ClimateBERT: A Pretrained Language Model for Climate-Related Text. In *Proceedings of AAAI 2022 Fall Symposium: The Role of AI in Responding to Climate Challenges*.
- World Federation of Exchanges (2022). Number of listed companies.
- 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.

