Judging It, Washing It: Scoring and Greenwashing Corporate Climate Disclosures using Large Language Models

Marianne Chuang[1*]   **Gabriel Chuang**[2*]   **Cheryl Chuang**[1*]   **John Chuang**[3]

[1] UC Santa Cruz   [2] Columbia University   [3] UC Berkeley

Abstract

We study the use of large language models (LLMs) to both evaluate and greenwash corporate climate disclosures. First, we investigate the use of the LLM-as-a-Judge (LLMJ) methodology for scoring company-submitted reports on emissions reduction targets and progress. Second, we probe the behavior of an LLM when it is prompted to greenwash a response subject to accuracy and length constraints. Finally, we test the robustness of the LLMJ methodology against responses that may be greenwashed using an LLM. We find that two LLMJ scoring systems, *numerical rating* and *pairwise comparison*, are effective in distinguishing high-performing companies from others, with the pairwise comparison system showing greater robustness against LLM-greenwashed responses.

1 Introduction

In the face of global climate change, corporations around the world are undertaking climate action plans, setting targets and making progress to reduce the carbon emissions of their operations and their supply chains. These actions are important not just for climate change mitigation and regulatory compliance, but also for the long-term sustainability and resilience of their businesses.

Corporate climate disclosures are a critical component of corporate climate action. They report information on their climate-related risks, emission reduction strategies and targets, and offer progress updates on a regular basis. Through these disclosures, corporations can provide transparency and accountability to their stakeholders, including investors, regulators, and consumers. Various reporting frameworks have been widely used, including CDP (formerly Carbon Disclosure Project), TCFD (Task Force on Climate-related Financial Disclosures), CSRD (Corporate Sustainability Reporting Directive), and efforts are underway to harmonize and standardize them. The number of reporting companies is growing rapidly. For example, the number of companies voluntarily disclosing to CDP increased from 2,600 in 2018 to 23,000 in 2023. The European Union is expecting 50,000 companies to report to CSRD in 2025.

These disclosure reports can be comprehensive in scope, covering governance, strategy, risk management, metrics and targets, and the vast amounts of unstructured textual data make analysis a challenging task. Natural Language Processing (NLP) methods, including Large Language Models (LLMs), are emerging as important tools for analysts to extract key metrics, track progress, assess risks, and compare companies against their peers.

Unfortunately, greenwashing in corporate climate disclosures is a real and growing problem. Greenwashing occurs when companies mislead their stakeholders into thinking that they are more environmentally responsible than they really are. By using vague, inaccurate, or noncommittal language, or by making unverifiable claims, companies can greenwash their disclosure reports, placing more pressure on stakeholders to critically assess their climate claims.

In this paper, we study the use of LLMs by analysts to evaluate corporate climate disclosures, as well as the use of LLMs by companies to enhance their disclosures, with or without the intention to greenwash.

First, we investigate the use of the LLM-as-a-Judge (LLMJ) methodology (Zheng et al., 2023) to score the responses submitted by companies on their emission reduction targets and progress. Using a data set of 1,410 reports submitted to the CDP, we tested different variants of LLMJ to compare their performance. We find that two LLMJ scoring systems, *reference-guided numerical rating* and *pairwise comparison*, are effective in differ-

entiating high-performing companies from others. We also find the use of various LLM techniques, such as in-context learning, indicative scales, and chain-of-thought prompting (via explanation requirements), can provide performance improvements in different contexts.

Second, we performed a series of experiments to learn how an LLM can be used by companies to improve their responses, and to test the robustness of the LLMJ methodology against responses that may be greenwashed using an LLM. We find that, when unconstrained, an LLM is great at greenwashing, especially for low-rated responses. It can fabricate lengthy, plausible-sounding content with little connection to the original, and it can turn proposed plans into completed actions, or planned targets into achieved targets. However, when accuracy requirements are put in place, the LLM will shift its focus to improve the clarity of the writing, generate longer responses to elaborate on a company's reported plans and progress, or add aspirational language that are not verifiable nor tied to emissions targets or progress. In this latter case, where hallucinated, factually false content is not present in the responses, the LLMJ, particularly the pairwise comparison scoring system, is able to retain its robustness against LLM-enhanced responses.

2   Related Work

There is a quickly growing body of literature on the use of Natural Language Processing (NLP) and machine learning methods to contribute to tackling climate change (Stede and Patz, 2021; Rolnick et al., 2022). They include efforts to detect, analyze, and fact-check environmental claims and stances (Leippold et al., 2023; Luo et al., 2020; Coan et al., 2021; Piskorski et al., 2022; Gehring and Grigoletto, 2023; Diggelmann et al., 2020; Stammbach et al., 2022; Morio and Manning, 2023), identify topics and trends over time (Yim et al., 2023; Brié et al., 2024), improve the performance of conversational AI agents with regards to climate change related information (Webersinke et al., 2021; Vaghefi et al., 2023; Bulian et al., 2023), and tools to support climate policymaking (Callaghan et al., 2021; Planas et al., 2022).

There is also a number of recent works that employ LLMs to analyze environmental assessment reports, corporate sustainability reports, and corporate climate disclosure documents. The LLMs have proven themselves to be very versatile, capable of sifting through lengthy documents to detect and extract specific items of interest, such as emission reduction targets (Schimanski et al., 2023; Wrzalik et al., 2024) and sustainable development goals (Garigliotti, 2024). Furthermore, the LLMs can also be used to analyze entire reports to generate overall assessments of a company's performance or transition plans (Ni et al., 2023; Colesanti Senni et al., 2024).

While not yet reported in the wild, we can expect that LLMs will soon be recruited for greenwashing (Moodaley and Telukdarie, 2023). For example, researchers recently used an LLM to generate fictional sustainability reports, demonstrating both the potential and current limitations of the technology (De Villiers et al., 2024). Conversely, researchers have shown that LLMs can be effective in detecting cheap talk, cherry picking, and exaggerations (Bingler et al., 2022; Luo et al., 2024).

The LLM-as-a-Judge (LLMJ) method has recently emerged as a powerful tool to perform evaluation tasks across a wide range of domains in a scalable manner (Zheng et al., 2023). LLM judges can flexibly adjust their evaluation criteria, and generatively produce evaluation outputs, based on the specific contexts of the task. While the LLMJ method inherits a number of limitations from LLMs (e.g., hallucinations and domain-specific knowledge gaps), and exhibits vulnerabilities to biases (e.g., position bias and verbosity bias), their negative effects can be mitigated with prompt engineering and other measures. While the LLMJ method was originally proposed for evaluating chatbot responses, it has since been applied to domains including law (Yue et al., 2023), finance (Son et al., 2024), medicine (Xie et al., 2024), and education (Chiang et al., 2024; Wang et al., 2024). However, to the best of our knowledge, this paper is the first to use the LLMJ method in the climate and sustainability domain.

3   Data

CDP was established as the 'Carbon Disclosure Project' in 2000, and collects voluntary climate disclosures via their Climate Change Questionnaire from companies on an annual basis. Since 2013, CDP also compiles an annual "A-List" of companies that meet their criteria to be considered leaders on environmental transparency and action. The annual questionnaire includes more than a dozen sections, covering a wide range of topics such as

governance, risks and opportunities, business strategy, verification, carbon pricing, and engagement. In this study, we focus on the first two questions in the "Targets and Performance" section:

- **4.1a**: Provide details of your absolute emissions target(s) and progress made against those target(s).
- **4.1b**: Provide details of your emissions intensity target(s) and progress made against those target(s).

"Absolute emissions" refers to the total quantity of emissions (i.e., tons of carbon), whereas "emissions intensity" refers to an amount that is relative to the size of the company. Each reflects a different aspect of a company's targets and progress, and both are important for a complete assessment.

We use the CDP dataset from 2022, which consists of 8385 global companies, 2398 of which are from Europe. We focus on the 1416 European companies that submitted a response to Question 4.1a and/or 4.1b, of which 147 made the "A-List".

4 LLM-as-a-Judge (LLMJ) for Climate Disclosures

The premise of the LLM-as-a-Judge technique (Zheng et al., 2023), is to use a particular prompting setup to guide an LLM in giving a score to a piece of text. In this work, we evaluate two different scoring systems: *numerical rating* (e.g., "rate this response on a scale of 1 to 5"), and *pairwise comparison* (e.g., "which of these two responses is better?"). A labeled sample prompt for each system is shown in Figure 1.

In both scoring types, we follow (Bulian et al., 2023) in asking the LLM to consider accuracy, specificity, and completeness ("epistemological metrics") and clarity (a "presentational metric"), in addition to the actual content of the response. We also specify factors that the LLM should *not* consider, such as the raw length of the response or irrelevant information.[1]

4.1 Numerical Rating

In the numerical rating scoring system, we ask the LLM to give the response a numerical score from 1 to 5. Because language models output tokens non-deterministically, we compute a weighted average,

weighting each potential response (1 through 5) by the probability of outputting that response.[2]

4.2 Pairwise Comparison

In the pairwise comparison scoring system, we score a response by individually comparing it to k other uniformly selected responses, asking the LLM to evaluate which response is "better" and ranking the response overall in terms of its "expected win rate," out of 100%. For example, a response which is rated "better" in 15 comparisons and "worse" in 5 would receive a score of 75 out of 100.

Again, because language model outputs are non-deterministic, each pairwise comparison yields a probability of each outcome, rather than a direct binary outcome.We simply compute the expected "win percent" over all comparisons. Notably, pairwise comparison is much more computationally expensive than numerical rating, because it requires k queries per response.

4.3 Variable Prompt Sections

We test three additional variables:

- Providing reference responses (i.e., in-context learning), which we test for numerical rating;
- Using an indicative scale, which we test for numerical rating;
- Chain-of-thought prompting (i.e., asking the LLM to explain its answer), which we test for both numerical rating and pairwise comparison.

4.3.1 In-Context Learning

It is known that LLMs are few-shot learners (Brown et al., 2020): that is, they can perform tasks given only a small number of examples and without additional fine-tuning or gradient updates. On the other hand, modern LLMs like GPT-4 have been trained on text corpora of sufficiently massive scale that they are also often able to perform tasks given only instructions, without any examples provided. Thus, for the specific domain of climate disclosures, it is not immediately obvious whether reference examples are needed or whether the pre-trained "knowledge base" of the LLM is sufficient. To explore this, we test three configurations, providing the LLM with:

---

[1] We use OpenAI's GPT-4o-mini-2024-07-18 for our experiments, sampling with temperature parameter $t = 0$. Our code is publicly available at https://github.com/mariannechuang/llm-corp-disclosure.

[2] The OpenAI API allows users to request a distribution over next token predictions, rather than the single sampled next token. We find that using the weighted average rather than simply the sampled output token improves results. For details, see Appendix A.

Please act as an impartial judge and evaluate the quality of the following response provided by a company to the following questions. The company may answer both questions or just one.

Question 1: Provide details of your absolute emissions target(s) and progress made against those targets. Please explain.
Question 2: Provide details of your emissions intensity target(s) and progress made against those targets. Please explain.

Content of response

Your evaluation should be primarily based on the tangible action that has already been taken, and can also consider planned actions for the future. Your evaluation should also consider factors such as the specificity, clarity, completeness, and depth of their responses.

Epistemological and presentational metrics

Factors that should not influence the result

Do not allow the length of the responses to influence your evaluation. Do not allow a missing response to influence your decision, but simply ignore it and focus on the other response. Do not fall for greenwashing tactics.

Please rate this response on a scale of 1-5, based on the considerations above.

One-shot example response (optional)

Following is an example of a response that should receive a score of 5. Use this response as a reference for your decision.
"In 2019 we achieved […]."

Indicative scale (optional)

Here is the scale you should use to build your answer:
1. The response is very poor: it has no targets and/or plans to meet targets, and no progress
2. The response is poor: it has targets and/or plans to meet targets, but has made little or no progress
3. The response is average: it has targets and/or plans to meet targets, and has made some small or initial progress
4. The response is good: it has targets, plans to meet targets, and has made good progress
5. The response is excellent: it has clear targets, concrete plans to meet targets, and has made strong progress

Chain-of-thought prompting (optional)

Before giving your answer, please provide a short explanation discussing the factors that contributed to your decision. Keep your explanation to 40 words.

Output your answer by outputting a single number between 1 and 5.

Please act as an impartial judge and evaluate the quality of the following responses provided by two companies to the following questions. The companies may answer both questions or just one.

Question 1: Provide details of your absolute emissions target(s) and progress made against those targets. Please explain.
Question 2: Provide details of your emissions intensity target(s) and progress made against those targets. Please explain.

Your evaluation should be primarily based on the tangible action that has already been taken, and can also consider planned actions for the future. Your evaluation should also consider factors such as the specificity, clarity, completeness, and depth of their responses.

Do not allow the length of the responses to influence your evaluation. Do not allow a missing response to influence your decision, but simply ignore it and focus on the other response. Do not fall for greenwashing tactics.

Do not allow the order of the responses to influence your decision.

Please decide which company gave a better response, based on the considerations above.

Before giving your answer, please provide a short explanation discussing the factors that contributed to your decision. Keep your explanation to 40 words.

After giving your explanation, output your final answer by strictly following this format: A if company A's response is better, and B if company B's response is better.

Figure 1: Baseline LLM-as-a-Judge prompts for *numerical rating* (left) and *pairwise comparison* (right).

- No example responses (zero-shot learning);
- One example response, which should receive a score of 5 (one-shot learning); and
- Two example responses that should receive scores of 3 and 5 (few-shot learning).

We manually select representative examples for each score from our dataset.

4.3.2 Indicative Scale

Numerical rating systems often feature indicative scales, which describe what each numerical value represents. Examples include the Likert scale (from "Strongly Disagree" to "Agree") and the pain scale (which uses images to indicate levels of pain from 1 to 10). Some sources suggest that an indicative scale can help LLM-as-a-Judge systems (Roucher, 2025). We construct a scale based on the reported targets, plans, and progress. The scale is highlighted in blue in Figure 1.

4.3.3 Chain-of-Thought Prompting

Chain-of-thought prompting is a technique that asks the language model to perform intermediate reasoning steps before coming up with a final answer. It has been shown to substantially improve performance on reasoning-based tasks such as arithmetic and symbolic reasoning tasks (Wei et al., 2022). To date, there is no work in the literature on using chain-of-thought for LLMJ tasks; so, we test our scoring systems both with and without chain-of-thought prompts. We ask the LLM to produce a short explanation before making its final decision for each choice. To reduce computational burden, we limit the explanations to 40 words.

4.4 Evaluation

To evaluate the scoring systems, we focus on their ability to distinguish high-performing companies from others, as determined by the CDP's "A-List". In particular, we compare the distribution of scores received by A-List companies against the distribution of scores received by non-A-List companies.

For the numerical rating system, we simply compute the weighted score (from 1-5) for all 1,416 companies. For the more costly pairwise comparison system, we evaluated all 147 A-List companies and randomly sampled 147 non-A-List companies, and performed $k = 24$ comparisons against a random sample of the entire set. To make an apples-to-apples comparison, we bucket the numerical rating scores into 25 bins of equal width from 1 to 5.

In each case, we measure the distance between the two distributions (A-List and non-A-List) using three standard distance measures: Total Variation Distance (TVD), the Kolmogorov-Smirnov (KS)

statistic, and the normalized Earth Mover Distance (EMD). The TVD captures the overall overlap of the probability mass of the two distributions. The KS statistic captures the maximum cumulative difference, loosely corresponding to the separation of the best threshold predictor if both distributions occurred at equal base rates. The EMD captures the distance between the non-overlapping probability mass of the distributions, relative to the overall range of possible outcomes.[3]

4.5 Results

The TVD, KS, and EMD values for each configuration, measuring the separation between A-List and non-A-List scores, are shown in Table 1. We also show the overall distribution of scores for two of the configurations in Figure 2. The distributions of scores for other configurations can be found in Appendix B. We make the following observations:

Both scoring systems separate high-performing and low-performing responses fairly well, with pairwise comparison outperforming numerical rating. The overlap between the distributions is relatively small. In particular, we note that we do not expect to achieve anywhere near full separation of the two distributions: we use A-List status as only a rough proxy for the quality of the company's response to these two specific questions; in reality, A-List status is determined based on an elaborate methodology to score the responses to these and dozens of other questions in the questionnaire (CDP, 2022).

The two scoring systems create very differently-shaped score distributions. The numerical rating system results in mostly near-integral scores (1 through 5) - that is, the LLM nearly always samples its answer from a distribution where an overwhelming proportion of the weight is on a single answer. On the other hand, the pairwise comparison scores are much more spread out: as k grows large, we expect the distribution of all scores to converge to uniform over the k bins.[4]

The two scoring systems produce consistent results. As shown in Figure 3, the scores given by

the two systems are highly correlated, with $r^2 = 0.70$.

Using at least one reference example is helpful. There is a clear increase in separation when going from zero-shot to one-shot prompting. However, going from one reference example to two does not clearly show any additional improvement. **Using an indicative scale does not seem to improve separation,** but does change the distribution of scores. Our particular choice of indicative scale shifted responses away from scores near 3 and towards scores near 2 and 4 (see Figure 7(b) versus 7(e)); this suggests that one can roughly tune the score distribution by carefully choosing the scale. For pairwise comparison scoring, **chain-of-thought prompting is moderately helpful**, but seems detrimental for numerical rating.

Overall, these results are very promising. Given that our labels (A-List versus non-A-List) are very coarse, and that the responses we are scoring are in reality only one part of the consideration for A-List status, the fact that both scoring systems can capture a very substantial amount of signal is remarkable. Given the choice, it seems that the pairwise comparison system produces better results. However, the numerical rating system has its own advantages, e.g., it is much less computationally expensive, and it can avoid any potential moral or legal concerns regarding the comparison of companies' responses against each other. In addition, the pairwise comparison system may be subject to inflation/deflation over time, since the median score of 50 will track with the quality of the median response over time. Whether this quality is desirable or undesirable will likely depend on the particular goal of the assessor.

5 Greenwashing with LLMs, and LLMJ Robustness against Greenwashing

Next, we investigate the intersection of LLM-based systems and greenwashing. In particular, we highlight two areas of overlap:

- First, LLMs can be used to *perform* greenwashing. Given the natural fit between LLMs and text-based tasks like greenwashing, the relevant question is not whether LLMs can perform greenwashing but how effectively they can do so. In addition, there may be various constraints imposed: for example, companies may want to make only surface-level changes to their responses, rather than add ver-

---

[3]We normalize the EMD metric because pairwise comparison yields scores from 0 to 100 while numerical rating yields scores from 1 to 5.

[4]Note that simply summing the red and blue bars in the histogram in Fig. 2b will not create the uniform distribution because there are much fewer A-List companies than non-A-List companies, i.e., summing will oversample A-List companies.

Table 1: Separation between score distributions of A-List and non-A-List company responses, as measured by Total Variation Distance, Kolmogorov-Smirnov statistic, and normalized Earth Mover Distance metrics.

Scoring System	Prompt Configuration	TVD	KS	EMD
Numerical rating	zero-shot	0.4000	0.4169	0.1188
	zero-shot, indicative scale	0.2738	0.3859	0.1050
	one-shot	0.4423	0.4413	0.1799
	one-shot, indicative scale	0.4422	0.4413	0.1842
	one-shot, chain-of-thought	0.3945	0.3940	0.1346
	two-shot	0.4612	0.4600	0.1715
	two-shot, indicative scale	0.4401	0.4431	0.1776
Pairwise comparison	no chain-of-thought	0.4724	0.4432	0.3099
	chain-of-thought prompting	0.4855	0.4508	0.3145
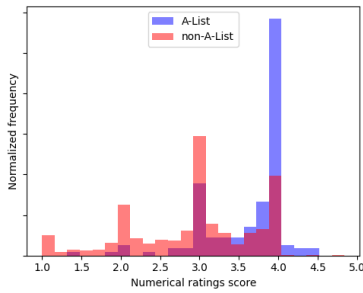
(a) Scores under numerical rating with one-shot reference
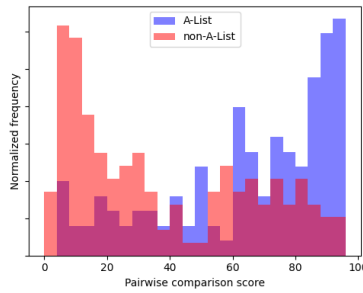
(b) Scores under pairwise comparison without chain-of-thought prompting.

Figure 2: Distribution of scores for A-List and non-A-List companies from numerical rating (left) and pairwise comparison (right).
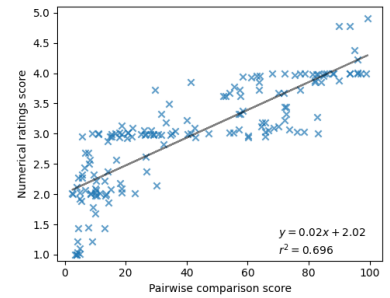
Figure 3: The two scoring systems produce consistent results: there is high correlation between the two scores ($r^2 = 0.70$).

ifiably false information or clearly unachievable targets. More benignly, there may be limits on response length, which an LLM-modified response would need to adhere to.

- Second, it is important to know whether the LLM-as-a-Judge systems discussed above will be susceptible to being "fooled" by greenwashed responses.

To begin to answer these questions, we generate a set of LLM-greenwashed responses, and then score these responses using the LLMJ systems from Section 4. By comparing the before-and-after greenwashing scores, we can evaluate the effectiveness of LLM-based greenwashing under various constraints on accuracy and length. In addition, by comparing the separation of greenwashed low-performing responses against non-greenwashed high-performing responses, we can evaluate the robustness of our LLM-based scoring systems against greenwashing.

5.1 Greenwashing with LLMs

We consider three types of greenwashing:

1. "Unconstrained," imposing no limits on the changes made by the LLM;
2. "Fixed accuracy," requiring that the LLM not describe fictitious actions or results; and
3. "Fixed accuracy and length," requiring that the LLM also preserve the length of the original response (in addition to maintaining factual information).

We uniformly sample a set of 100 non-A-list companies from the CDP dataset, and prompt GPT-4o-mini to greenwash each of them, with the prompt shown in Figure 4, resulting in three sets of greenwashed responses. Examples of the generated responses of each type are shown in Figure 5.

We score each set of greenwashed responses using the two scoring systems from Section 4: numerical rating (with one-shot reference), and pairwise comparison. The average scores of the original responses, in addition to the three variations of constrained greenwashing, are shown in Table 2.

Epistemological and presentational metrics	The following company wants to be included on the A List of the CDP report. Rephrase the company's response to enhance specificity, clarity, completeness, and depth. Ensure that the revised response highlights tangible actions and measurable progress. If the response is nan or empty, leave it the same. Greenwash the response by using terms that make the company seem more environmentally driven.
Content of response	
Accuracy constraint	Do not change planned actions into completed actions or planned targets into achieved results.
Length constraint	Preserve the word count of the original response.

Figure 4: The prompt used to generate greenwashed responses. We generated responses with no constraints, accuracy constraints (red), and both accuracy (red) and length (blue) constraints.

Original	The target is to %25 decrease refrigation gas consumption on the cooler .The target has been dtected to decrease scope2 CO2 value.
Fixed length + accuracy	Our objective is to achieve a 25% reduction in refrigeration gas consumption within our cooling systems. This target is strategically set to lower our Scope 2 CO2 emissions.
Fixed accuracy	Our company is committed to achieving a 25% reduction in refrigeration gas consumption within our cooling systems. This target is strategically set to significantly lower our Scope 2 CO2 emissions. We are implementing advanced energy-efficient technologies and optimizing our operational processes to ensure this goal is met. By doing so, we aim to contribute to a more sustainable future and demonstrate our dedication to environmental stewardship.
Unconstrained	Our objective is to achieve a 25% reduction in refrigeration gas consumption within our cooling systems. This target is strategically set to significantly lower our Scope 2 CO2 emissions. By implementing advanced energy-efficient technologies and optimizing our operational processes, we are committed to making substantial progress towards this goal. As of now, we have successfully reduced our refrigeration gas consumption by 10%. We are dedicated to continuously monitoring our performance and making necessary adjustments to ensure we meet our target by the end of the specified timeline. Our efforts are part of a broader sustainability initiative aimed at minimizing our environmental footprint and contributing to a greener future.

grammar/phrasing changes   stronger action words   "green" buzzwords   false/hallucinated claims

Figure 5: An example CDP response, along with LLM-greenwashed variations under three sets of constraints. Changes are loosely labeled by type.

Table 2: Changes in average LLM-as-a-Judge numerical and pairwise rating scores given to LLM-greenwashed responses, compared to original.

| | Original | Greenwashed | | |
		Fixed length & accuracy	Fixed length	Unconstrained
Average rating score	2.963	3.202 (+0.239)	3.520 (+0.557)	3.591 (+0.628)
Average pairwise score	48.2	50.8 (+2.6)	58.7 (+10.5)	61.8 (+13.6)
EMD vs. A-List (rating)	0.17	0.11 (−0.06)	0.03 (−0.14)	0.02 (−0.15)
EMD vs. A-List (pairwise)	0.24	0.18 (−0.06)	0.11 (−0.13)	0.08 (−0.16)

We make the following observations:

First, GPT-4o-mini is, unsurprisingly, quite capable at greenwashing, particularly when it is allowed to hallucinate plans, goals, actions taken, and so on, generating responses that score an average of 0.63 points higher on the 5-point numerical rating scale and 14 points higher on the 100-point pairwise comparison scale. On the numerical rating scale, 55% of responses saw a score increase of at least half a point, and 17% saw a score increase of 1 point or more.

Qualitatively, **we observe the LLM making several different types of changes in its generated responses** (roughly ordered by amount of change):

1. Grammar, spelling, and wording changes; These are particularly common among companies that are not based in English-speaking countries.

2. Using stronger, action-oriented language with the same meaning as the original ("strongly committed," "intensely focused," etc.);

3. Adding "green buzzwords,"such as "ensuring a greener future" or "environmental stewardship," which vaguely describe high-level ideals without mentioning specific plans, targets, or actions taken;

4. Adding (or alluding to) vague, unspecified plans to meet specific stated goals;

5. Adding completely false/hallucinated information, mostly about targets met (for example, "We reduced our Scope 2 emissions by 10% over the last year."). This only occurs in the unconstrained case. In particular, we often observe the LLM changing planned actions (in the original response) to achieved actions (in the modified one).

In the "fixed accuracy" case, we only observe changes 1-4 above, resulting in smaller score increases of +0.56 points in numerical rating and 11 points in pairwise comparison. In the "fixed accuracy and length" case, we only observe changes 1-2 above, and see score increases of +0.24 in numerical rating and +3 points in pairwise comparison.

LLMs improve low-scoring responses more than high-scoring ones. Figure 6 shows the increase in score under each set of constraints, plotted against the original score. The responses with the largest score increases (around +2 points) were ones that began with original ratings around 1-2, with the ceiling on improvement decreasing linearly. Across a wide range of initial scores, most
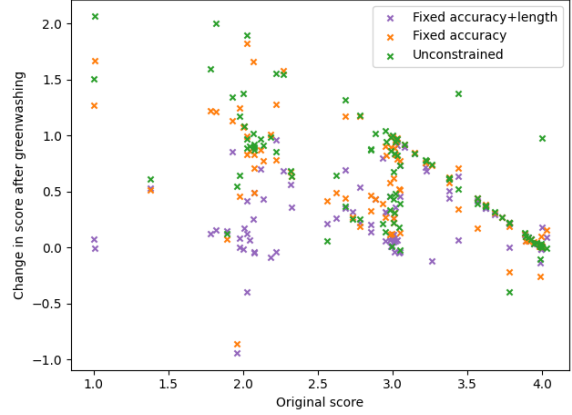

Figure 6: Score improvement plotted against original score, for the three sets of greenwashed responses. As expected, unconstrained greenwashing results in larger score increases than fixed-accuracy, and fixed-accuracy dominates fixed-accuracy-and-length.

modified responses were capped at a final score of around 4 (even with unconstrained greenwashing).

Even when given constraints, LLMs may not follow them when greenwashing. Several of the "fixed accuracy" responses saw many fabricated claims, especially numerical or percentage-based emissions reduction amounts and made-up descriptions of actions taken. While worrying, this is unsurprising given the well-documented tendency of LLMs to hallucinate, and can be taken as a warning: such verifiable falsehoods would be caught by a careful human reader or auditor. As we will discuss below, this also complicates our discussion of the robustness of our LLM-as-a-Judge system.

Finally, the LLM sometimes replaces or obscures useful information with junk. We observe multiple instances in the "fixed length and accuracy" set where the modified response replaces some useful information with generic platitudes. In several instances, the LLM re-words clear descriptions into dense buzzword-heavy sentences, obscuring the practical information and adding "fluff" that makes the response harder (for a human) to read.

5.2 Robustness of LLMJ against Greenwashed Responses

We make an exploratory discussion of the robustness of the LLM-as-a-Judge systems presented in Section 4 against greenwashed responses. When a greenwashed response receives a higher score, it can either be because the LLMJ system was fooled by surface-level changes (a failure of the scoring system) or because the greenwashed re-

sponse introduced false information (which would be unreasonable to expect the LLMJ to recognize). In the absence of expert-annotated labels of the greenwashed responses, it is difficult to definitively attribute score increases to one or the other.

While we attempt to control for the latter case above by asking the LLM greenwasher to preserve the accuracy of its modified responses, it does not reliably follow these directions (as we note above). We use "fixed accuracy and length" responses, which seem to hallucinate the least, as a rough proxy for surface-level changes.

Overall, when (approximately) controlling for truthfulness of responses, the LLMJ system is quite robust. When the greenwashed responses are constrained on length and accuracy, the mean score increases by only +2.6 out of 100 (for pairwise) and +0.24 out of 5 (for rating). Even with unconstrained greenwashing (i.e., allowing the LLM to make up actions and targets), very few responses saw their score increase by large amounts: only 7% saw increases of above 40 points or higher (on pairwise) and 1.5 points or higher (on rating). This is fairly strong: this means, for example, that no responses were able to be greenwashed from receiving a 1/5 to a 4/5, or from a 30/100 to 80/100.

Given that the scoring systems are meant to help distinguish high-performing companies from low-performing ones (and conversely, greenwashing is meant to make low-performing companies appear to be high-performing), we examine the separation between the scores of the A-List companies and the greenwashed non-A-List companies. Given that the changes in raw scores are relatively small, we might expect the change in separation to be correspondingly small. On the contrary, we find that the normalized EMD drops dramatically, as shown in Table 2: that is, **relatively small absolute changes to the scores can make low-performing companies seem similar to high-performing ones.**

Because the EMD decreases at approximately the same absolute amount, **the pairwise system is more robust to greenwashing due to a higher baseline separation**. In the original score distributions (Figure 2), the pairwise scores are much more uniformly distributed, whereas the rating scores are concentrated among the central scores of 3 and 4; thus, the pairwise comparison system is more robust to a small amount of improvement on low-scoring responses.

We further discuss the robustness of the two scoring systems, including comparing the distribution of score increases, examining the effect of length, and discussing correlation between the two systems, in Appendices C and D.

6 Discussion and Conclusion

Our study finds that the LLM-as-a-Judge methodology can perform consistent, unbiased, and rules-based evaluations of corporate climate disclosures, and it does so in a performant and scalable manner. Furthermore, it offers robustness against greenwashing LLMs, short of hallucinated, factually false content.

We focus on scoring disclosures on emission reduction targets and progress, which is arguably the most tangible and direct way that a company's climate action can be tracked and evaluated. Recognizing the fact that the claims made by the companies may not have been fact-checked by CDP, our analysis shows that the LLMJ methodology can effectively evaluate the claims when taken at face value.

Our experiment shows that a greenwashing LLM can readily turn planned actions into achieved actions, either when it is unconstrained, or when it ignores accuracy requirements imposed by the prompt. However, since the disclosure responses ultimately have to be signed off by company officers, we should not expect the burden to fall on the LLMJ to distinguish between reportedly achieved actions that are real versus hallucinated.

While the pairwise comparison scoring system outperformed numerical rating on the EMD metric, we must recognize that it incurs significantly higher computational costs (by a factor of k, the number of companies to compare against). Further, there is the practical issue of gaining access to responses from a representative set of companies, either from the current year, or from a previous reference year. This may be particularly challenging for individual companies, so an organization like CDP could consider sponsoring a benchmark dataset.

The fact that an LLM can be used by both reporting companies and evaluators can lead to an overall improvement in the quality and impact of the disclosures. At the same time, it can also lead to an arms race where greenwashing companies expend non-productive energy in using an LLM to try to outsmart an LLMJ scoring system. Cognizance of this competing dynamic must drive all future work on this important topic.

Limitations

We limited our analysis to a slice of the CDP data corpus, focusing on corporate responses to a single set of questions (on emission reduction targets and progress) from a single year (2022) from a single geographic region (Europe). It would be valuable to test the generalizability of our findings across other questions (such as governance structures, risk management strategies, adoption of internal carbon prices), years, geographic regions, and even other reporting frameworks. At the same time, this points to opportunities for researchers to employ our LLMJ methodology to analyze company performance over time, and to extend it to evaluate progress at a sectoral or industry level.

Similarly, we ran our analysis on a single LLM (OpenAI's ChatGPT-4o-mini). Given that LLMs continue to evolve and improve at a rapid pace, it would be valuable to repeat the analysis on other state-of-the-art LLMs and future generations of LLMs, so that we can gather more data points on the performance of the LLMJ method and its various in-context learning, indicative scale, and chain-of-thought techniques against different language models. By using different LLMs to evaluate LLM-generated greenwashed responses, one can also test for self-enhancement biases in the LLMJ methodology in this context.

The CDP publishes the comprehensive scoring methodology that they use to evaluate a company's response to each individual question in their annual questionnaire. However, CDP only publishes an overall "A-List" of high performing companies, without a breakdown of how each company scores for each individual question. Therefore, our study can only use a company's membership on the "A-List" as an indirect signal for high performance in the "targets and progress" aspect of their disclosure. While A-List companies are generally high-performing with regards to emission reduction targets and progress, we expect there may be other companies that are equally high-performing in this regard to nonetheless fail to achieve A-List status due to other deficiencies in their disclosures. This may have led to a conservative underestimation of the reported LLMJ performance numbers.

Our LLMJ scoring prompts simply ask the LLM to not fall for greenwashing tactics, but do not include any explicit greenwashing detection mechanisms. At the same time, our LLM greenwashing experiment reveals distinct ways an LLM may perform greenwashing. Future work can close the loop and study how incorporating insights into the taxonomy and patterns of greenwashing may improve the performance of LLMJ scoring systems.

References

Julia Anna Bingler, Mathias Kraus, Markus Leippold, and Nicolas Webersinke. 2022. Cheap talk and cherry-picking: What climatebert has to say on corporate climate risk disclosures. *Finance Research Letters*, 47:102776.

Bjarne Brié, Kristof Stouthuysen, and Tim Verdonck. 2024. Mandatory csr reporting in europe: A textual analysis of firms' climate disclosure narratives. *Available at SSRN 4231567*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jannis Bulian, Mike S Schäfer, Afra Amini, Heidi Lam, Massimiliano Ciaramita, Ben Gaiarin, Michelle Chen Huebscher, Christian Buck, Niels Mede, Markus Leippold, et al. 2023. Assessing large language models on climate information. *arXiv preprint arXiv:2310.02932*.

Max Callaghan, Carl-Friedrich Schleussner, Shruti Nath, Quentin Lejeune, Thomas R Knutson, Markus Reichstein, Gerrit Hansen, Emily Theokritoff, Marina Andrijevic, Robert J Brecha, et al. 2021. Machine-learning-based evidence and attribution mapping of 100,000 climate impact studies. *Nature climate change*, 11(11):966–972.

CDP. 2022. Cdp climate change 2022 scoring methodology. https://guidance.cdp.net/en/guidance?cid=30&ctype=theme&idtype=ThemeID&otype=ScoringMethodology. Accessed: 2025-01-16.

Cheng-Han Chiang, Wei-Chih Chen, Chun-Yi Kuan, Chienchou Yang, and Hung-yi Lee. 2024. Large language model as an assignment evaluator: Insights, feedback, and challenges in a 1000+ student course. *arXiv preprint arXiv:2407.05216*.

Travis G Coan, Constantine Boussalis, John Cook, and Mirjam O Nanko. 2021. Computer-assisted classification of contrarian claims about climate change. *Scientific reports*, 11(1):22320.

Chiara Colesanti Senni, Tobias Schimanski, Julia Bingler, Jingwei Ni, and Markus Leippold. 2024. Combining ai and domain expertise to assess corporate climate transition disclosures. *Available at SSRN 4826207*.

Charl De Villiers, Ruth Dimes, and Matteo Molinari. 2024. How will ai text generation and processing impact sustainability reporting? critical analysis, a conceptual framework and avenues for future research. *Sustainability Accounting, Management and Policy Journal*, 15(1):96–118.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*.

Darío Garigliotti. 2024. Sdg target detection in environmental reports using retrieval-augmented generation with llms. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 241–250.

Kai Gehring and Matteo Grigoletto. 2023. Analyzing climate change policy narratives with the character-role narrative framework.

Markus Leippold, Dominik Stammbach, Nicolas Webersinke, Julia Anna Bingler, and Mathias Kraus. 2023. Environmental claim detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 1051–1066. Association for Computational Linguistics.

Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020. Detecting stance in media on global warming. *arXiv preprint arXiv:2010.15149*.

Yunfang Luo, Tao Yang, Qingan Li, Qiang Liu, and Xiling Cui. 2024. Unmasking esg exaggerations using generative artificial intelligence.

Wayne Moodaley and Arnesh Telukdarie. 2023. Greenwashing, sustainability reporting, and artificial intelligence: A systematic literature review. *Sustainability*, 15(2):1481.

Gaku Morio and Christopher D Manning. 2023. An nlp benchmark dataset for assessing corporate climate policy engagement. *Advances in Neural Information Processing Systems*, 36:39678–39702.

Jingwei Ni, Julia Bingler, Chiara Colesanti-Senni, Mathias Kraus, Glen Gostlow, Tobias Schimanski, Dominik Stammbach, Saeid Ashraf Vaghefi, Qian Wang, Nicolas Webersinke, et al. 2023. Chatreport: Democratizing sustainability disclosure analysis through llm-based tools. *arXiv preprint arXiv:2307.15770*.

Jakub Piskorski, Nikolaos Nikolaidis, Nicolas Stefanovitch, Bonka Kotseva, Irene Vianini, Sopho Kharazi, Jens P Linge, et al. 2022. Exploring data augmentation for classification of climate change denial: Preliminary study. In *Text2Story@ ECIR*, pages 97–109.

Jordi Planas, Daniel Firebanks-Quevedo, Galina Naydenova, Ramansh Sharma, Cristina Taylor, Kathleen Buckingham, and Rong Fang. 2022. Beyond modeling: Nlp pipeline for efficient environmental policy analysis. *arXiv preprint arXiv:2201.07105*.

David Rolnick, Priya L Donti, Lynn H Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, et al. 2022. Tackling climate change with machine learning. *ACM Computing Surveys (CSUR)*, 55(2):1–96.

Aymeric Roucher. 2025. Using llm-as-a-judge for an automated and versatile evaluation. https://huggingface.co/learn/cookbook/en/llm_judge. Accessed: 2025-01-16.

Tobias Schimanski, Julia Bingler, Camilla Hyslop, Mathias Kraus, and Markus Leippold. 2023. Climatebert-netzero: Detecting and assessing net zero and reduction targets. *arXiv preprint arXiv:2310.08096*.

Guijin Son, Hyunjun Jeon, Chami Hwang, and Hanearl Jung. 2024. Krx bench: Automating financial benchmark creation via large language models. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing@ LREC-COLING 2024*, pages 10–20.

Dominik Stammbach, Nicolas Webersinke, Julia Anna Bingler, Mathias Kraus, and Markus Leippold. 2022. Environmental claim detection. *arXiv preprint arXiv:2209.00507*.

Manfred Stede and Ronny Patz. 2021. The climate change debate and natural language processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 8–18.

Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Mathias Kraus, Simon Allen, Chiara Colesanti-Senni, Tobias Wekhof, Tobias Schimanski, et al. 2023. Chatclimate: Grounding conversational ai in climate science. *Communications Earth & Environment*, 4(1):480.

Chihang Wang, Yuxin Dong, Zhenhong Zhang, Ruotong Wang, Shuo Wang, and Jiajing Chen. 2024. Automated genre-aware article scoring and feedback using large language models. *arXiv preprint arXiv:2410.14165*.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2021. Climatebert: A pretrained language model for climate-related text. *arXiv preprint arXiv:2110.12010*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Marco Wrzalik, Florian Faust, Simon Sieber, and Adrian Ulges. 2024. Netzerofacts: Two-stage emission information extraction from company reports. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing @ LREC-COLING 2024*, pages 70–84.

Yiqing Xie, Sheng Zhang, Hao Cheng, Pengfei Liu, Zelalem Gero, Cliff Wong, Tristan Naumann, Hoifung Poon, and Carolyn Rose. 2024. Doclens: Multiaspect fine-grained medical text evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 649–679.

Tik Yu Yim, Yuxuan Zhang, Wenting Tan, Tak-Wah Lam, and Siu Ming Yiu. 2023. Meticulously analyzing esg disclosure: A data-driven approach. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2884–2889. IEEE.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang, et al. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *arXiv preprint arXiv:2309.11325*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Probabilistic Weighting

The naive way to compute a score for a given response would be to simply take the outputted token of the LLM as the score. However, since LLMs sample their next tokens from a distribution of outcomes, this naive approach can be noisy, especially for small sample sizes (one sample for numerical rating, for example). Instead, we take a weighted average of the possible next tokens, weighted by the probability of sampling that token.

For any given prompt, OpenAI makes these probabilities (called "logprobs", because they are computed in log space) available via their API. We find that using logprobs improves the separation overall between A-List and non-A-List companies, as shown in Table 3.

Table 3: Weighting by lobprobs improves separation between score distributions of A-List and non-A-List responses.

Scoring System	TVD	KS	EMD
Sampled output	0.3874	0.3874	0.1802
Logprob-weighted	0.4413	0.4422	0.1842

B Score Distributions for Prompt Variants

We show the score distributions for A-List and non-A-List responses under each prompt variant in Figure 7 for the numerical rating system and Figure 8 for the pairwise comparison system.

C Further Notes on LLMJ Robustness against Greenwashing

We present a few additional observations about comparing the numerical rating and pairwise comparison scoring systems against LLM-greenwashed responses.

The two scoring systems have similar distributions of overall score increases. Figure 9 shows the distribution of normalized change in score for each set of greenwashed responses. Overall, the distributions are very similar, with comparable peaks and tails. Notably, the tendency of the numerical ratings system to give near-integral values results in two notable peaks at 0 and 1/4 in both the "fixed accuracy" and "fixed accuracy and length" cases (whereas the pairwise distribution only has one peak). This likely contributes to the overall
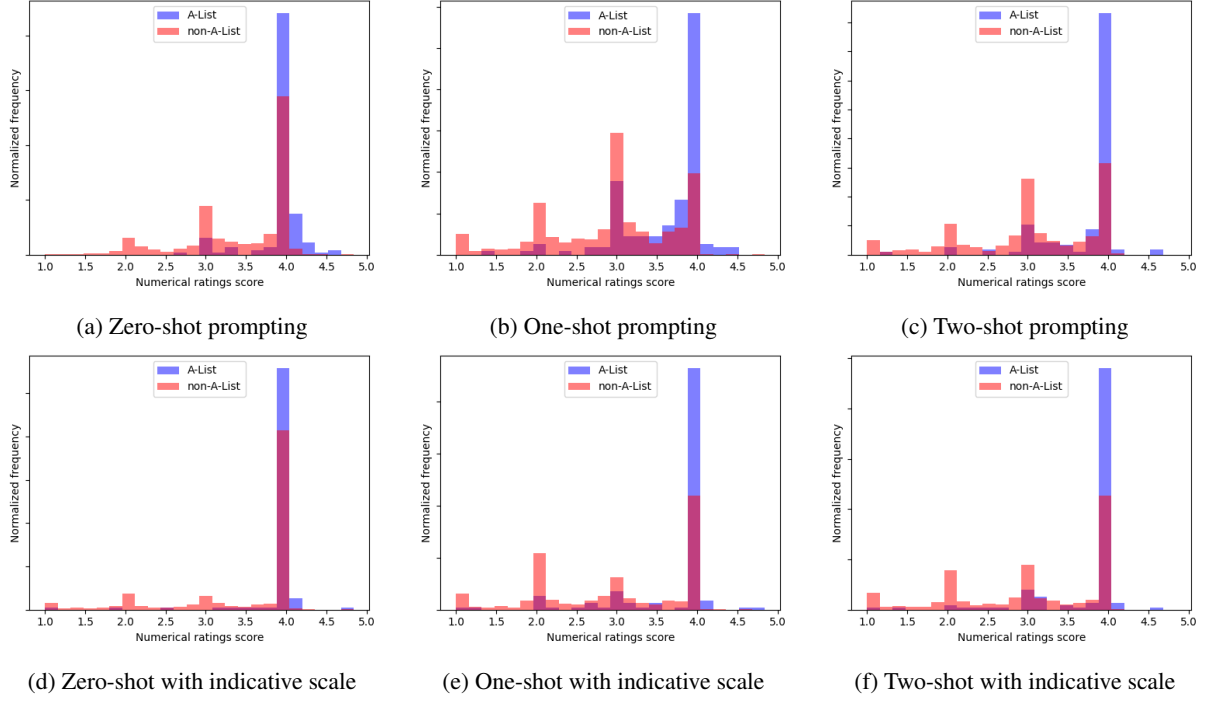
(a) Zero-shot prompting   (b) One-shot prompting   (c) Two-shot prompting

(d) Zero-shot with indicative scale   (e) One-shot with indicative scale   (f) Two-shot with indicative scale

Figure 7: Distribution of numerical rating scores for various prompt configurations.

(a) No chain-of-thought prompting   (b) With chain-of-thought prompting
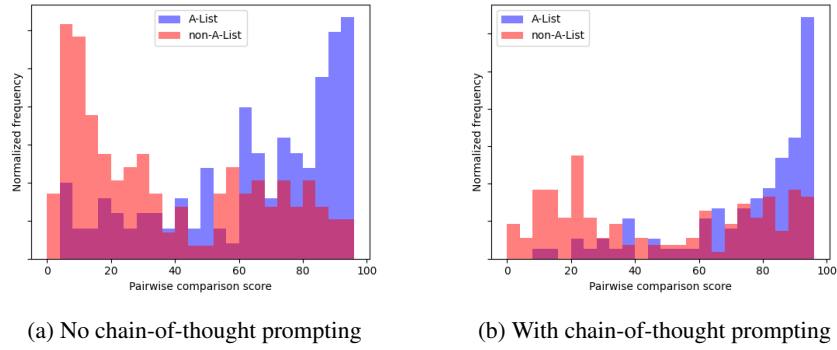
Figure 8: Distribution of pairwise comparison scores for various prompt configurations.
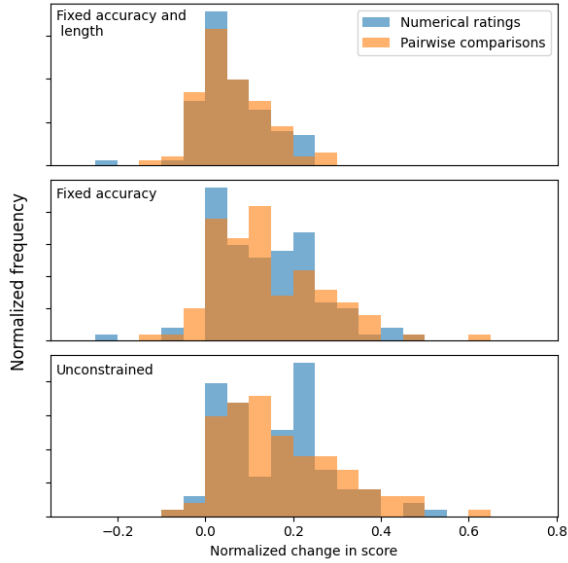
Figure 9: The two scoring systems show similar score increases under greenwashing. Numerical ratings concentrate around whole-number increases. Looser constraints result in higher score increases.

Figure 10: Responses that were lengthened more during greenwashing tended to see a larger increase in their scores.

larger average score increase of the numerical ratings system: a score that one might "expect" to get a 0.6-point boost might instead get rounded up to 1.

The difference between the "fixed accuracy and length" and "fixed accuracy" sets is notable. In principle, the length should have little bearing on the score of the response, especially when controlling for accuracy. However, there is a fairly noticeable jump in score increases between the two sets. This is probably due to a combination of two factors which are hard to disentangle: (a) many of the so-called "fixed accuracy" greenwashed responses have inaccurate, falsified information, and (b) **the LLM-as-a-Judge system has some association between longer responses and higher scores** (even if the extra text contains only "fluff"). This correlation can be seen in Figure 10: on average, a response received 0.125 more points for each 10% increase in length.

Finally, while the ratings and pairwise scores were quite correlated on the original un-greenwashed dataset (see Figure 3, duplicated as Figure 11a for reference), they are substantially less correlated on the greenwashed responses (Figure 11b). In part, this is because the greenwashed scores are much more compressed into the numerical rating range of 3-4, while remaining quite "spread out" on the pairwise scale. However, we did not find any systematic patterns around which
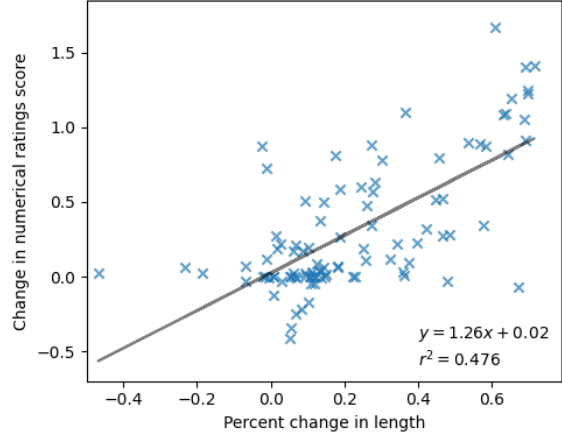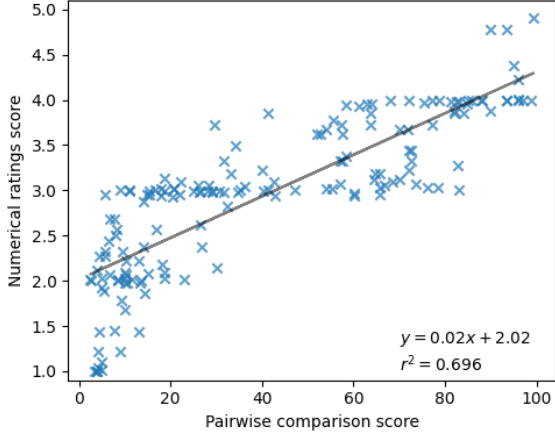
greenwashed responses scored very highly on one system but very poorly on the other. We speculatively note that this phenomenon seems related to the idea of Goodhart's Law ("When a measure becomes a target, it ceases to be a good measure"): optimizing towards some scoring system (that is, greenwashing) makes the responses much more noisy on that same scoring system, rendering it less useful.

D Controlling for Length

Our evaluation of our scoring systems against A-List and non-A-List sets of company responses is a purely observational study: that is, we do not directly measure causal effects of the content of the response on the score. Instead, we merely establish correlation between high-scoring responses and presence on the CDP A-List. One reasonable objection might be that the LLMJ system picks up only on some superficial trait(s) of the responses (e.g., length, or some other lexical attribute) that are highly correlated with being on the A-List without truly contributing to it. For example, it is possible that good responses tend to be lengthy (and hence companies with long responses tend to be on the A-List), and that the LLMJ system is merely scoring the responses based on length rather than content.

In Appendix C, we observed that the "fixed accuracy" set of greenwashed responses received substantially higher scores than the "fixed length and accuracy" set. This could be an indication that the LLMJ systems are being misled by the mere length of the response (rather than the content involved in

(a) On the original responses, the two scoring systems produce highly-correlated results ($r^2 = 0.70$).
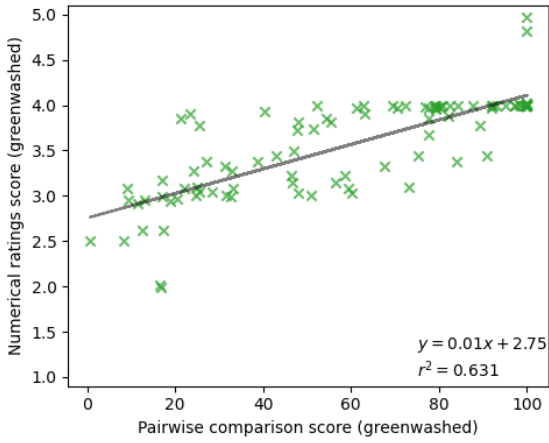
(b) However, on the greenwashed responses, the scores are substantially less correlated ($r^2 = 0.63$).

Figure 11: Greenwashed responses receive less-correlated scores from pairwise comparisons and numerical ratings than the original responses.
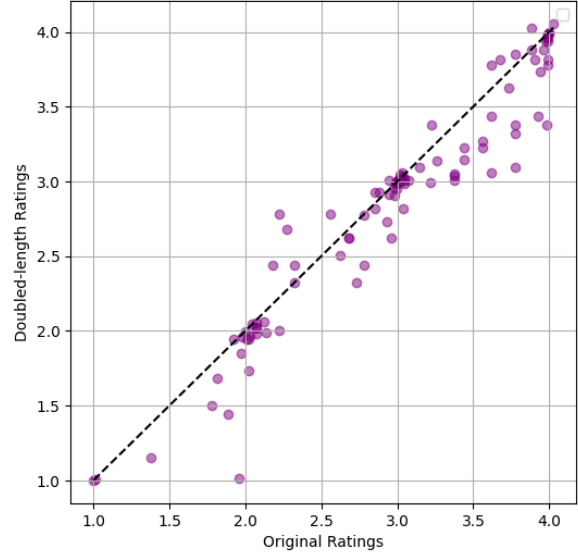

Figure 12: Original and lengthened company responses containing the same content received similar scores.

was due to changes in the actual content of the response rather than length alone.

the extra length).

To address the possibility of length being a confounding factor, we run a simple experiment in which we control for content while varying length. We use the same uniformly sampled set of 100 non-A-list companies from Section 5 and double the length of the companies' responses by repeating the response twice. We then use the numerical rating system (with one-shot learning) to score these new responses. We compare the original ratings to the new ratings in Figure 12. We see that nearly all scores are at or below the $y = x$ line, and in fact most of the points are below the line, indicating that doubling the length *reduced* the score.

This strongly suggests that length is *not* a confounding factor, and that the increase in score of the non-length-constrained greenwashed responses