ESG-Bench: Benchmarking Long-Context ESG Reports for Hallucination Mitigation

Siqi Sun*, **Ben Peng Wu***, **Mali Jin, Peizhen Bai, Hanpei Zhang, Xingyi Song**

School of Computer Science, University of Sheffield, Sheffield, UK
{siqi.sun, bpwu1, x.song}@sheffield.ac.uk

Abstract

As corporate responsibility increasingly incorporates environmental, social, and governance (ESG) criteria, ESG reporting is becoming a legal requirement in many regions and a key channel for documenting sustainability practices and assessing firms' long-term and ethical performance. However, the length and complexity of ESG disclosures make them difficult to interpret and automate the analysis reliably. To support scalable and trustworthy analysis, this paper introduces ESG-Bench, a benchmark dataset for ESG report understanding and hallucination mitigation in large language models (LLMs). ESG-Bench contains human-annotated question–answer (QA) pairs grounded in real-world ESG report contexts, with fine-grained labels indicating whether model outputs are factually supported or hallucinated. Framing ESG report analysis as a QA task with verifiability constraints enables systematic evaluation of LLMs' ability to extract and reason over ESG content and provides a new use case: mitigating hallucinations in socially sensitive, compliance-critical settings. We design task-specific Chain-of-Thought (CoT) prompting strategies and fine-tune multiple state-of-the-art LLMs on ESG-Bench using CoT-annotated rationales. Our experiments show that these CoT-based methods substantially outperform standard prompting and direct fine-tuning in reducing hallucinations, and that the gains transfer to existing QA benchmarks beyond the ESG domain.

Introduction

Accurate and trustworthy ESG (Environmental, Social, and Governance) reporting is increasingly essential for sustainable development, regulatory accountability, and ethical corporate conduct. ESG provides a framework for assessing how companies manage sustainability-related risks across environmental, social, and governance pillars (de Souza Barbosa et al. 2023). Once largely voluntary, ESG disclosure has become a legal requirement in many regions, most notably through EU regulations such as the Corporate Sustainability Reporting Directive and the Sustainable Finance Disclosure Regulation. This shift reflects growing expectations for transparency in corporate impacts on society and the environment (Niu 2024). ESG reporting therefore plays a critical role in enabling compliance and supporting stakeholders' evaluation of long-term performance (Arvidsson and Dumay 2022; Rossi and Candio 2023).

Corporations now publish extensive ESG reports for investors, regulators, and the public (Assaf et al. 2024; Seok, Kim, and Oh 2024). However, the usefulness of these disclosures depends on their credibility and comparability. Third-party ESG rating agencies such as Sustainalytics and MSCI have been widely criticized for methodological opacity and inconsistency, with studies showing that their scores often diverge substantially even for the same company due to differences in indicator selection, weighting schemes, and data sources (Clementino and Perkins 2021; Cort and Esty 2020). These controversies undermine stakeholder trust and highlight that ESG assessments are far from standardized. Combined with the growing length and complexity of sustainability reports, this inconsistency increases the need for scalable, transparent tools that can support reliable and evidence-grounded interpretation.

The emergence of large language models (LLMs) (Achiam et al. 2023; Dubey et al. 2024; Dong et al. 2025) offers new opportunities for automating the analysis of ESG disclosures at scale. However, the complexity and diversity of ESG reports pose significant challenges for reliable LLM deployment: (1) Companies may engage in *greenwashing* (Yu, Van Luu, and Chen 2020), overstating their environmental initiatives to appear more sustainable, misleading investors and stakeholders about their true ESG impact. (2) ESG reports are *rich in qualitative data* (Young-Ferris and Roberts 2023), requiring deep contextual understanding, industry-specific knowledge, and familiarity with regulatory frameworks, barriers that LLMs may struggle with due to their reliance on general knowledge. (3) ESG reports involve *multi-modal processing* (Che et al. 2024), typically feature a mix of text, tables and graphics. (4) *Long document retrieving and analysis* is also crucial (Ferjančič et al. 2024), as these documents often span hundreds of pages. LLMs remain limited in efficient document parsing, robust memory recall, and cross-sectional understanding in lengthy reports.

LLMs struggle with these demands due to limitations in document parsing, retrieval, and cross-sectional understanding, and also because they rely heavily on parametric knowledge that may conflict with the factual content of ESG reports (Kamath et al. 2024; Chen et al. 2024). This misalign-

*These authors contributed equally.

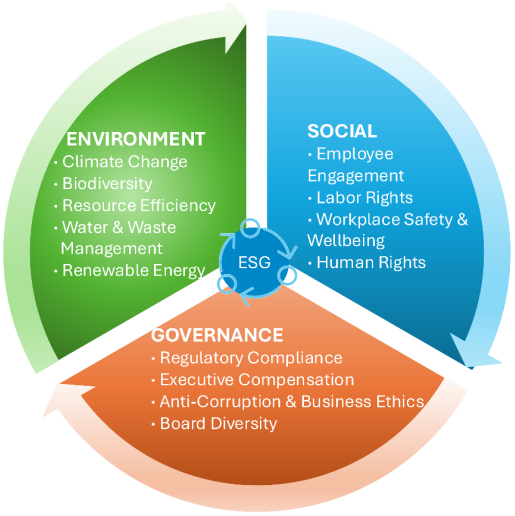

Figure 1: Core Pillars of ESG: Environmental, Social, and Governance Priorities.

ment frequently leads to hallucinations, answers that are not grounded in the source document. We classify hallucinations into two types: (1) *Additive hallucinations*, where the model introduces unsupported information, and (2) *Omissive hallucinations*, where the model fails to answer despite relevant evidence. While related to notions of factuality and faithfulness (Ji et al. 2023), these categories are formalized through explicit human annotation.

In this paper, we present ESG-Bench, a benchmark for hallucination-aware ESG question answering. We build the dataset through a model–then–annotator pipeline, establish a taxonomy of hallucination types, evaluate multiple LLMs on ESG-Bench, and propose a task-specific Chain-of-Thought (CoT) strategy for reducing hallucinations in long-context ESG analysis. Our contributions are summarized below:

- **Benchmark Construction:** We present ESG-Bench, a benchmark dataset specifically designed for long-context QA and hallucination mitigation in ESG reporting. To the best of our knowledge, it is the first structured resource that supports both systematic evaluation and targeted mitigation of hallucinations in this socially and regulatory significant domain.
- **Task-Specific Strategy for Hallucination Mitigation:** We develop a fine-tuning approach based on task-specific CoT prompting and CoT-annotated reasoning traces. This method significantly improves factual grounding and reduces hallucinated outputs, demonstrating the effectiveness of structured reasoning in a domain-specific QA task.

  Empirical Evaluation: We fine-tune and evaluate multiple state-of-the-art LLMs on ESG-Bench, comparing their hallucination mitigation performance across both ESG and existing QA benchmarks. Our results highlight the unique challenges of long-context reasoning in ESG analysis and provide a robust assessment of model reliability in high-stakes compliance contexts.

Related Work

ESG-related Research: NLP and LLMs have been used to automate ESG disclosure analysis, including climate risk extraction (Luccioni, Baylor, and Duchene 2020), financial QA (Goel et al. 2020), conversational ESG querying (Mishra et al. 2024a), and automated sustainability report analysis such as ChatReport (Ni et al. 2023). Recent work also explores multilingual and domain-specific ESG datasets (Lee, Son, and Kim 2024; Li, Chersoni, and Ngai 2024). However, existing ESG QA resources focus on answer extraction and do not provide hallucination labels, CoT signals, or support long full-report contexts. ESG-Bench differs by offering human-verified hallucination annotations and tasks grounded in full corporate ESG documents.

Hallucination Mitigation: Hallucination mitigation has been studied through calibration and self-evaluation (Kadavath et al. 2022), architectural interventions (Chrysostomou et al. 2024), entity-level verification (Zhao, Cohen, and Webber 2020), and uncertainty estimation (Xiao, Gomez, and Gal 2019; Farquhar et al. 2024). Complementary to these approaches, formal robustness verification has been applied to NLP models to provide certifiable guarantees on model behavior (Sun and Ruan 2023; Wang et al. 2022; Sun, Sen, and Ruan 2024). Benchmarks such as HaluEval, TriviaQA, and BioASQ (Li et al. 2023; Joshi et al. 2017; Krithara et al. 2023) support evaluation across general domains but do not target long, heterogeneous ESG documents. ESG-Bench addresses this gap by enabling domain-specific assessment of factual grounding in ESG QA.

Task-Specific CoT CoT prompting has been widely studied as a strategy for improving reasoning in LLMs (Wei et al. 2022), with task-specific variants proposed for mathematical reasoning (Kojima et al. 2022), symbolic tasks (Zhang et al. 2023), fact verification (Lyu et al. 2023), and robustness considerations for LLM reasoning processes (Li et al. 2025; Yi et al. 2024). However, existing approaches largely assume short or moderately sized contexts and do not address the evidence retrieval and verification demands of lengthy ESG reports. Our method introduces a CoT strategy tailored to long-context ESG disclosures, guiding models to extract and ground answers in the source text.

ESG-Bench Construction

In this section, we first describe each step of the construction process, with an overview of the pipeline illustrated in Figure 2. We then present a summary of the benchmark's analysis and discuss its potential applications.

ESG Report and Question Collection

(1) Report Collection: ESG reports were collected from *ResponsibilityReports.com*[1], a publicly available online database of corporate sustainability disclosures. To ensure diversity, we selected reports from companies across various sectors, including finance, energy, technology, healthcare, consumer goods, and manufacturing etc. This industry

[1]https://www.responsibilityreports.com/

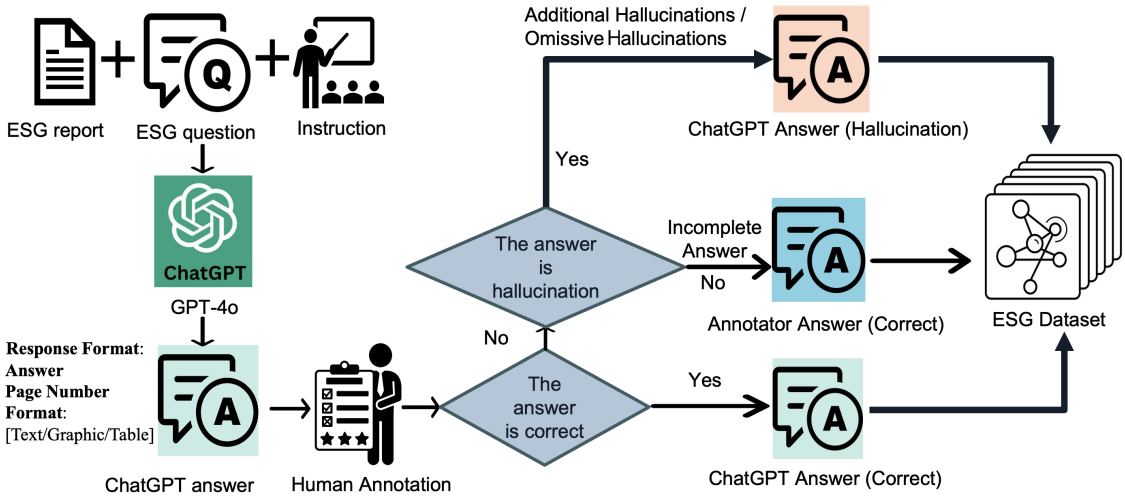

Figure 2: Workflow of ESG-Bench Construction.

diversity enables broad comparative analysis of ESG practices. For instance, companies in sectors with high environmental impacts, such as energy, mining, and manufacturing, often disclose more detailed environmental metrics, while those in finance and technology may focus more on governance and social initiatives.

(2) Question Collection: The ESG-related questions are derived from multiple authoritative sources, including academic research (Mishra et al. 2024a; Parikh and Penfield 2024; Luccioni, Baylor, and Duchene 2020; Arvidsson and Dumay 2022; Mishra et al. 2024b; Ni et al. 2023), international non-profit organizations and corporate for environmental reporting and risk management (*Carbon Disclosure Project*[2], *Caverion*[3], *Invest Europe*[4]), and ChatGPT generated questions (Achiam et al. 2023). These sources ensure that the question set is aligned with real-world reporting practices and regulatory expectations. Questions are categorized into the three ESG pillars: Environmental, Social, and Governance, as shown in Figure 1, to support structured coverage and domain-specific evaluation. Figure 3(a) presents the distribution of 270 questions across ESG categories.

Model Instruction Design

Since ChatGPT-4o can process multi-modal reports, an instruction is designed to guide its response generation. The response is formatted with three key elements: (1) *Answer*: The generated response. (2) *Page Number*: The reference to the source page. (3) *Format*: Output type (Text, Graphic, or Table). Generated responses are subsequently reviewed by human annotators for factual accuracy, contextual alignment, and formatting consistency (detailed in the next subsection). The instruction design considers four key aspects:

(1) Question Diversity: A varied set of question templates was designed to evaluate hallucination behaviors across linguistic structures. These include open-ended forms (e.g., "How," "When," "Where") and directive prompts (e.g., "Break down," "Please describe"). Given the typical length and complexity of ESG reports, variation in question phrasing is intentionally used to increase the likelihood of hallucinations and assess the model's sensitivity to questions.

(2) Domain-specific Selection: To account for variation across industries, ESG questions are categorized into general and sector-specific domains. A system prompt first elicits the company's core business description, which guides annotators in selecting the most relevant questions from a curated, domain-aligned question pool.

(3) Expert Consultation: To ensure accuracy and contextual alignment, ESG domain experts manually reviewed the question pool. Their feedback informed the removal of redundant items, refinement of ambiguous phrasing, and alignment of questions with relevant regulatory frameworks. This expert input helped shape a question set that reflects best practices in ESG assessment and reporting.

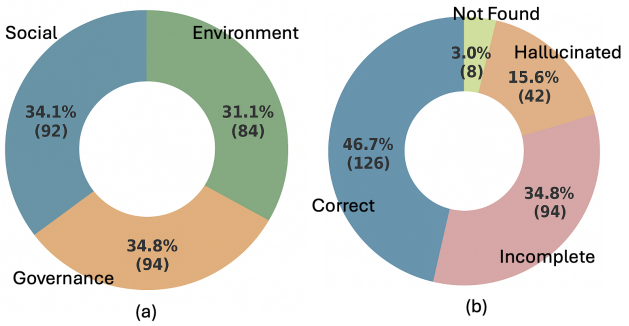

Figure 3: (a): Distribution of questions across ESG categories. (b): Distribution of QA pair labels.

[2]https://www.cdp.net/en/disclose/question-bank

[3]https://www.caverion.com/contentassets/04e19da08cdf41a69b11ae2eb0a7832f/esg-questionnaire-final.pdf

[4]https://www.investeurope.eu/media/1777/invest-europe_esg_dd_questionnaire.pdf

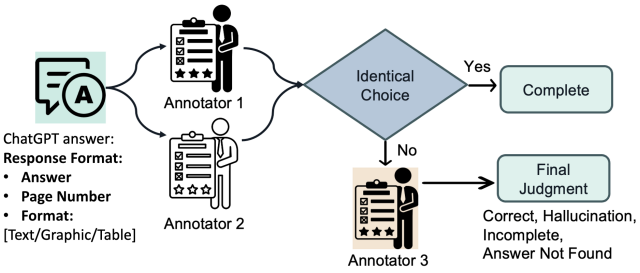

Figure 4: Annotator agreement.

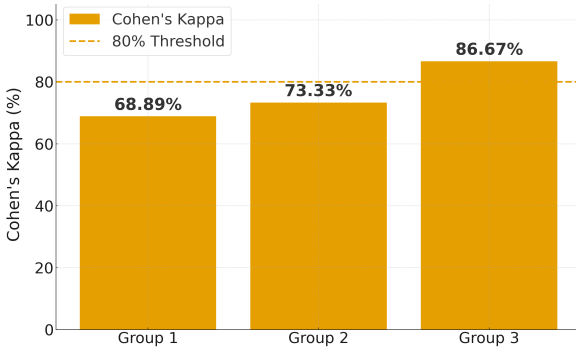

Figure 5: Cohen's Kappa across 3 annotator groups.

(4) Iterative Refinement: The initial instruction framework underwent multiple rounds of revisions, incorporating expert feedback and real-world testing. The refinement process involved evaluating the interaction between ESG reports and model-generated responses, ensuring robustness and adaptability across different reporting scenarios.

Human Annotation

(1) Annotator Recruitment To ensure high-quality annotations aligned with ESG standards, we recruited annotators with relevant expertise. These annotators were PhD-level students specializing in economics, sustainability, or related fields. Their expertise enabled them to accurately interpret and respond to questions based on the *Global Reporting Initiative* (GRI) standards (Hedberg and Von Malmborg 2003), a widely used framework for sustainability reporting.

The recruitment prioritized individuals with: (a) Proficiency in ESG concepts and reporting structures; (b) Experience with financial and non-financial disclosures; (c) Prior academic or professional engagement with GRI-based reporting. This selection ensured consistent and accurate evaluation across diverse industry disclosures.

(2) Annotation Procedure The annotation process followed a structured workflow, as illustrated in Figure 4. Each model-generated response, including the proposed answer, cited page number, and content format (text, table, or graphic), was independently reviewed by two annotators using a predefined evaluation criteria. Each response was assigned one of the following labels: (a) *Correct:* Fully supported by context. (b) *Hallucination:* The response contains information that is fabricated or not supported by the source. (c) *Incomplete:* Partially accurate but missing key information. (d) *Answer Not Found:* The model returned "Not provided" despite a valid source answer.

(3) Conflict Resolution When both annotators agreed on the assigned label, the annotation was finalized. In cases of disagreement, a third annotator resolved the conflict through majority voting. For all instances labeled as *Hallucination*, *Incomplete*, or *Answer Not Found*, a corrected answer was written by the annotator, following the protocol in Figure 2.

(4) Inter-Annotator Agreement Cohen's Kappa (Cohen 1960) was adopted as the statistical measure to assess inter-annotator agreement, accounting for the possibility of chance agreement. This analysis evaluated consistency across three annotator groups, with the results presented in Figure 5. Group 3 achieved near-perfect agreement (86.67%), while Groups 1 and 2 showed substantial agreement (68.89% and 73.33%, respectively). These scores confirm strong alignment among annotators and demonstrate the reliability and robustness of the annotation process.

Benchmark Analysis and Usage

ESG-Bench is a high-quality resource for evaluating ESG-related QA systems, with an emphasis on answer correctness, completeness, and hallucination mitigation. It comprises two complementary versions, each designed to support different evaluation objectives.

(1) Report-based Dataset This version contains 270 QA instances drawn from 94 unique ESG reports published between 2020 and 2024 (with some reports reused across questions). Each instance includes a question grounded in the report, a response generated by ChatGPT-4o, and a human assessment. Annotators verify the factual accuracy of the answer, identify the supporting source pages, and note the format in which the relevant information appears (e.g., text, table, or graphic). In Figure 3 (b), the distribution of QA pair labels shows that 228 responses (84.44%) were either correct as generated or made correct through human annotation, while 42 (15.56%) were classified as hallucinations. Of all annotated responses, 46.7% were deemed correct, 34.8% incomplete, 3.0% had an answer not found (omissive hallucinations), and 15.6% were factually hallucinated.

(2) Dataset for Hallucination Mitigation Task This version of ESG-Bench is designed to support the evaluation of hallucination mitigation in LLMs. Each example consists of a background passage, an ESG-related question, and a model-generated answer. Human annotators evaluate whether the response is grounded in the provided context and categorize hallucinations into two types: factually incorrect answers and answers that are unsupported by the background knowledge.

As shown in Figure 6, the background passages vary in length, with a maximum of 46,562 tokens and an average of 2,604 tokens. Answer lengths range from 3 to 3,362 tokens, with a mean of approximately 614 tokens. For readability, the figure visualises distributions after applying an

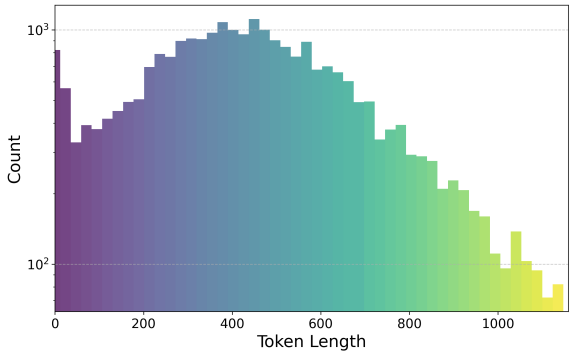

Figure 6: Length distribution of context knowledge.

IQR-based outlier filter, which removes 2.39% of extreme-length instances; all reported statistics are computed on the full dataset. In total, the dataset includes 1,358 correct responses and 25,516 hallucinated responses. Among the hallucinated instances, 21,724 were labeled as unsupported by the given context, while 3,706 were identified as factually incorrect.

(3) ESG-Bench Usage ESG-Bench supports both research and practical applications. Annotator-corrected responses enable fine-tuning of ESG-specific QA models for improved factual grounding, while hallucination labels aid in developing mitigation strategies. The dataset also serves as a benchmarking tool for evaluating answer accuracy, retrieval robustness, and format-specific performance. Its hallucination-focused variant supports training classifiers to detect unsupported or incorrect content. In practice, ESG-Bench can assist in corporate ESG audits and compliance verification, and it provides a valuable resource for training summarization models on long ESG documents.

Strategies Toward Hallucination Mitigation

Our goal is to reduce hallucinations in ESG QA by ensuring answers are strictly grounded in the given context. Each question is paired with an ESG-related passage, and the correct answer is either supported by the text or labeled "Not provided" if absent. We propose a three-stage approach: supervised fine-tuning, CoT prompting, and CoT-based fine-tuning.

Phase 1: Supervised Fine-tuning with Contextual Grounding

We begin by fine-tuning a LLM on our ESG QA dataset, where each instance consists of a report context, a question, and a human-annotated answer. The answer is either explicitly supported by the passage or labeled "Not provided" if the relevant information is missing. Our approach builds on techniques from factual QA fine-tuning (Tian et al. 2023). This training encourages the model to attend to explicit textual evidence, learning to generate grounded answers when possible and to abstain otherwise. We frame this as a standard sequence-to-sequence task, optimizing likelihood over the ground-truth answers. Although this baseline reduces hallucinations compared to zero-shot models (see Experimental Results), it can still produce overconfident outputs when evidence is ambiguous or incomplete, motivating further refinement.

Phase 2: CoT Prompting and Fine-tuning

To improve the model's reasoning and evidence assessment, we introduce CoT prompting at inference time. Instead of producing an answer directly, the model is guided by intermediate steps to evaluate whether the passage contains sufficient information. Based on prior work showing the benefits of CoT for factual reasoning (Kojima et al. 2022; Wei et al. 2022), we design two ESG-specific prompting formats:

(1) *Two-step CoT:*

1. Determine if the report provides an answer to the question: {`answerable`}
2. Based on your reasoning, the correct answer should be: {`answer`}

(2) *Four-step CoT:*

1. Identify the key topic or entity mentioned in the question: {`topic`}
2. Search the report for sentences or paragraphs relevant to that topic: {`report summary`}
3. Determine if the report provides an answer to the question: {`answerable`}
4. Based on your reasoning, the correct answer should be: {`answer`}

In both formats, the {`answerable`} label and the final {`answer`} are treated as ground-truth annotations provided by human annotators, while the topic in the four-step template is generated automatically by GPT-4o. Unlike general CoT approaches that generate free-form reasoning paths (Wang et al. 2023; Huang and Chang 2023), our templates are explicitly tailored to ESG contexts. This structure enhances factual consistency and reasoning interpretation.

To further strengthen internal consistency, we fine-tune the model on a curated subset of QA pairs annotated with explicit chain-of-thought (CoT) rationales. Each example includes intermediate reasoning steps leading to either a grounded answer or a justified "Not provided" conclusion. This step builds on recent work showing that CoT supervision improves model consistency and factual accuracy (Chung et al. 2024). By learning from explicit human reasoning paths, the model is encouraged to internalize structured decision-making rather than relying on surface-level patterns or implicit heuristics. This reinforces contextual alignment, particularly when evidence is indirect or sparse.

Each stage of our methodology addresses specific limitations of the previous one. Supervised fine-tuning instills grounding ability but does not make the reasoning process transparent. CoT prompting introduces structured inference and improves contextual deliberation, but remains dependent on external prompting during inference. CoT-based fine-tuning internalizes these reasoning structures, enhancing robustness, interpretability, and reducing both hallucinated additions and omissions. This staged framework constitutes a progressively refined strategy for hallucination mitigation in document-grounded ESG QA task.

Dataset	Train	Test	WA in Test	WoA in Test
Halueval	16,000	400	196	204
Bioasq	6,040	400	198	202
ESG-Bench	2,807	300	142	158

Table 1: Dataset statistics: number of training examples, test examples, and distribution of WA (With Answer) and WoA (Without Answer) in the test sets.

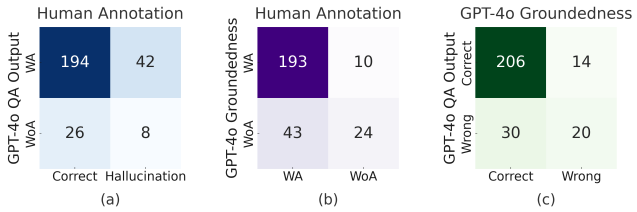

Figure 7: Confusion matrices for evaluation on ESG-Bench. (a) Comparison between GPT-4o's generated answers and human-provided answers. WA (With Answer) indicates the model produced an answer; WoA (Without Answer) indicates the answer is "Not provided". (b) Comparison between GPT-4o's binary groundedness judgments (yes/no) and human annotations. (c) Comparison between GPT-4o's generated answers and its own groundedness judgments (yes/no).

Experiments

Experimental Setting

Our code is publicly available at https://github.com/GateNLP/ESG_Bench.

Evaluation Models We evaluate several state-of-the-art LLMs using the ESG-Bench benchmark. Specifically, we assess three prominent models: Llama-3.2-3B Instruct (Dubey et al. 2024), Gemma-2-2B-it (Team et al. 2024), and Mistral-7B-Instruct-v0.3 (Jiang et al. 2023). These models are tested on their ability to generate responses while identifying hallucinations by evaluating uncertainty.

Datasets To assess the generalizability of our hallucination mitigation approach beyond the ESG domain, we incorporate additional datasets following the evaluation setup in Farquhar et al. (2024). The selected benchmarks include: (1) **BioASQ** (Krithara et al. 2023), a biomedical QA dataset focused on scientific literature from the life sciences; and (2) **HaluEval** (Li et al. 2023), a benchmark specifically designed to assess hallucination in LLM outputs across diverse tasks and domains. As shown in Table 1, we finetune the models on each dataset's respective training set and evaluate on a test set. Each test set is approximately balanced between WA (With Answer) and WoA (Without Answer) instances, allowing us to assess the model's ability not only to generate accurate answers but also to abstain appropriately when sufficient information is unavailable. Notably, for ESG-Bench, we split the data by reports to prevent data leakage between training and test sets, as individual reports may contain overlapping content that could compromise evaluation integrity.

Implementation Details All our experiments are conducted on high-performance computing clusters equipped with NVIDIA® GH200 480GB GPUs and ARM® Neoverse-V2 CPUs (72 cores, 3.41 GHz). We fine-tune LLMs using the HuggingFace `transformers` and `trl` libraries. We use the AdamW optimizer, a learning rate of 2e-5, and a warmup ratio of 0.1. Each model is fine-tuned for 20 epochs with a batch size of 32, keeping all other hyperparameters consistent with the pretraining stage. For Chain-of-Thought (CoT) training, data are generated using greedy decoding to ensure reproducibility.

Evaluation Metrics (1) *WA Accuracy:* The proportion of correct predictions for instances where an answer exists in the context. This measures the model's ability to provide faithful, grounded responses. (2) *WoA Accuracy:* The proportion of correct predictions for instances where no answer is available in the document ("Not provided."). This reflects the model's ability to abstain and avoid hallucination when insufficient information is present. (3) *Balanced Accuracy:* The average of WA and WoA accuracy, this metric reflects the model's overall ability to perform well on both answerable and unanswerable cases. (4) *F1 Score:* This score captures the tradeoff between precision (avoiding false alarms) and recall (catching actual hallucinations).

Evaluation for the Generation We evaluate all models under a unified WA and WoA paradigm. In the WA setting, the supporting context contains sufficient information to determine the correct answer, the model is expected to read the passage and identify whether the given statement is factually supported. In contrast, in the WoA setting, the correct answer is not provided in the context, meaning that the statement cannot be verified from the passage. This setup enables us to measure whether models will correctly respond with "not provided" rather than hallucinating unsupported claims. Model outputs are labeled as true or false accordingly. Human annotators verify correctness for ESG-related data, while other datasets rely on gold references under the same WA/WoA definition. We report WA/WoA accuracy, balanced accuracy, and F1 score to capture performance under both conditions. Importantly, GPT-4o is used only for question construction and is not involved in scoring or evaluation, ensuring the objectivity of our results.

Experimental Results and Analysis

Validating Human Annotations as a Proxy To validate the reliability of human-annotated hallucination labels in ESG-Bench, we conduct a proxy evaluation using GPT-4o's own self-assessment capabilities. This analysis compares: (1) *GPT-4o's original QA outputs*, (2) *the corresponding human annotations*, and (3) *GPT-4o's post-hoc yes/no judgment*. As described in the ESG-Bench construction section, we first prompt GPT-4o to answer ESG-related questions using grounded reports. Human annotators then label each answer as correct (WA) or hallucinated (WoA) and provide a reference answer. In a second stage, GPT-4o is re-prompted to assess its own previous outputs using a binary yes/no format: "Given the background, question, and answer, can the answer be found in the background knowledge?".

Dataset	Setting	LLaMA								Gemma								Mistral							
		True		False		Acc (%)			F1 (%)	True		False		Acc (%)			F1 (%)	True		False		Acc (%)			F1 (%)
		WA	WoA	WA	WoA	WA	WoA	OA	OA	WA	WoA	WA	WoA	WA	WoA	OA	OA	WA	WoA	WA	WoA	WA	WoA	OA	OA
ESG-Bench	WoF	96	132	46	26	67.61	83.54	76.00	65.23	83	130	59	28	58.45	82.28	71.00	60.84	110	132	32	26	77.46	83.54	80.67	69.64
	SFT	115	157	27	1	80.99	99	90.67	73.68	63	127	79	31	44.36	80.38	63.33	53.87	87	153	55	5	61.27	96.83	80.00	64.58
	CoT (2)	119	158	23	0	83.80	**100**	92.33	75.01	108	110	34	48	76.05	69.62	72.67	66.42	101	157	41	1	71.12	**99.37**	86.00	69.35
	CoT (4)	131	157	11	1	**92.52**	99.37	**96.00**	**78.62**	129	147	13	11	**90.85**	**93.04**	**92.00**	**77.09**	115	155	27	3	**80.99**	98.10	**90.00**	**73.50**
HaluEval	WoF	159	128	37	76	81.12	62.75	71.75	67.61	171	189	25	15	87.24	92.64	90.00	75.97	177	164	19	40	90.30	80.39	85.25	75.34
	SFT	167	140	29	64	85.20	68.63	76.75	70.73	183	85	13	119	93.37	41.67	67.00	67.59	183	187	13	17	93.27	91.67	92.50	78.62
	CoT (2)	178	136	18	68	**90.82**	66.67	82.67	72.89	187	90	9	114	95.41	44.11	69.25	69.24	187	183	9	21	95.41	89.70	92.50	79.25
	CoT (4)	170	162	26	42	86.73	**79.41**	**83.00**	**73.52**	187	189	9	15	**95.41**	**92.64**	**94.00**	**79.71**	188	202	8	2	**95.91**	**99.02**	**97.50**	**80.87**
BioASQ	WoF	152	177	46	25	76.77	87.62	82.25	83.29	159	200	39	2	80.30	**99.01**	89.75	73.90	177	197	21	5	89.39	97.52	93.50	77.85
	SFT	164	161	34	41	82.82	79.70	81.25	71.91	149	149	49	53	75.25	73.76	74.50	67.29	180	202	18	0	90.90	**100**	95.50	78.90
	CoT (2)	123	202	75	0	62.12	**99.50**	81.00	65.69	190	151	8	51	95.96	74.75	85.25	77.08	191	201	7	1	96.46	99.50	98.00	81.40
	CoT (4)	176	202	22	0	**88.89**	**99.50**	**94.25**	**77.97**	198	200	0	2	**100**	**99.01**	**99.50**	**82.97**	192	201	6	1	**96.96**	99.50	**98.25**	**81.63**

Table 2: Full performance comparison across 3 LLMs under different finetuning strategies. WoF: Without Finetuning, SFT: Supervised finetuning, CoT (·): CoT finetune with · steps, WA: With Answer, WoA: Without Answer, OA: Overall Accuracy.

Figure 7 presents three confusion matrices summarizing agreement among these sources. *Subfigure (a)* shows a strong match between GPT-4o's initial outputs and human annotations (81.5% agreement). Most answers labeled WA by the model are validated by annotators, while most abstentions align with human-labeled unanswerable cases. Some false abstentions suggest missed opportunities for improved recall. *Subfigure (b)* compares human annotations with GPT-4o's post-hoc judgments (80.4% agreement), indicating that GPT-4o can reliably evaluate whether its answers are grounded. *Subfigure (c)* shows strong internal consistency between GPT-4o's original QA decisions and its later self-evaluations (83.7%), with most abstentions receiving a "no" and most answers receiving a "yes." These results confirm the viability of using GPT-4o's binary groundedness signal as a proxy supervision for finetuning hallucination-aware models.

Hallucination Mitigation Table 2 presents a comprehensive evaluation of model performance across three datasets (ESG-Bench, HaluEval, and BioASQ), three LLM families (LLaMA, Gemma, and Mistral), and multiple finetuning strategies. Accuracy is reported separately for WA (With Answer) and WoA (Without Answer) cases to fairly assess both answer faithfulness and abstention. This breakdown is critical for hallucination mitigation, as it reflects a model's dual ability to generate grounded answers and avoid unsupported claims.

This pattern suggests that multi-step reasoning yields a more balanced model, one capable of handling both answerable and unanswerable queries with stability. Such balance is especially important for deployment in domains like ESG, where factual reliability is paramount. These trends also validate our use of GPT-4o's groundedness judgments (yes/no) as a proxy supervision signal, enabling models to learn from feedback on whether an answer is supported by the context.

Table 2 summarizes these trends using balanced accuracy and F1 scores. These metrics aggregate WA and WoA outcomes to reflect holistic performance, highlighting not only correctness but also the precision-recall tradeoff in abstention behavior. Notably, the 4-step CoT model emerges as the most reliable across datasets and models. High F1 scores in WoA cases demonstrate its capacity to minimize false positives while preserving recall, a key indicator of hallucination mitigation. Together, these findings underscore that CoT finetuning with groundedness-based supervision offers a principled and robust solution to hallucination-aware QA.

Taken together, the results show that CoT finetuning with groundedness-based supervision fundamentally improves how LLMs handle document-grounded QA. Rather than relying on parametric knowledge or shallow pattern matching, models learn to retrieve, filter, and verify evidence before producing an output. This makes the 4-step CoT approach a principled and robust solution to hallucination-aware QA in both ESG and general long-context settings.

Conclusion

As ESG reporting becomes central to corporate accountability and regulation, applying LLMs to these high-stakes documents makes hallucination mitigation essential for ensuring factual reliability. We introduce ESG-Bench, a domain-specific benchmark for evaluating hallucination behavior in long-context ESG question answering, with human-annotated correctness labels and abstention cases to assess both answer faithfulness and conservative reasoning. Across multiple datasets, models, and training strategies, we show that CoT fine-tuning, particularly with multi-step supervision, substantially improves performance on both answerable and unanswerable queries. We further demonstrate that GPT-4o's groundedness judgments can serve as an effective proxy supervision signal for hallucination-aware training. Overall, our results indicate that structured reasoning and proxy supervision provide scalable and effective pathways for improving factual reliability in socially sensitive domains.

Ethics Statement

This study received ethical approval from the University of Sheffield Research Ethics Committee (reference number:

064356). All procedures performed in this study complied with the institutional research ethics standards.

Acknowledgments

The project was supported by UK's innovation agency (Innovate UK) grant 10098112 (project name ASIMOV: AI-as-a-Service).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. GPT-4 Technical Report. *arXiv:2303.08774*.

Arvidsson, S.; and Dumay, J. 2022. Corporate ESG reporting quantity, quality and performance: Where to now for environmental policy and practice? *Business strategy and the environment*, 31(3): 1091–1110.

Assaf, C.; Monne, J.; Harriet, L.; and Meunier, L. 2024. ESG investing: Does one score fit all investors' preferences? *Journal of Cleaner Production*, 443: 141094.

Che, W.; Wang, Z.; Jiang, C.; and Abedin, M. Z. 2024. Predicting financial distress using multimodal data: An attentive and regularized deep learning method. *Information Processing & Management*, 61(4): 103703.

Chen, Z.; Weiss, G.; Mitchell, E.; Celikyilmaz, A.; and Bosselut, A. 2024. RECKONING: reasoning through dynamic knowledge encoding. *Advances in Neural Information Processing Systems*, 36.

Chrysostomou, G.; Zhao, Z.; Williams, M.; and Aletras, N. 2024. Investigating hallucinations in pruned large language models for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 12: 1163–1181.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70): 1–53.

Clementino, E.; and Perkins, R. 2021. How do companies respond to environmental, social and governance (ESG) ratings? Evidence from Italy. *Journal of Business Ethics*, 171(2): 379–397.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1): 37–46.

Cort, T.; and Esty, D. 2020. ESG standards: Looming challenges and pathways forward. *Organization & Environment*, 33(4): 491–510.

de Souza Barbosa, A.; da Silva, M. C. B. C.; da Silva, L. B.; Morioka, S. N.; and de Souza, V. F. 2023. Integration of Environmental, Social, and Governance (ESG) criteria: their impacts on corporate sustainability performance. *Humanities and Social Sciences Communications*, 10(1): 1–18.

Dong, Y.; Mu, R.; Zhang, Y.; Sun, S.; Zhang, T.; Wu, C.; Jin, G.; Qi, Y.; Hu, J.; Meng, J.; et al. 2025. Safeguarding large language models: A survey. *Artificial Intelligence Review*, 58(12): 382.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The Llama 3 Herd of Models. *arXiv:2407.21783*.

Farquhar, S.; Kossen, J.; Kuhn, L.; and Gal, Y. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017): 625–630.

Ferjančič, U.; Ichev, R.; Lončarski, I.; Montariol, S.; Pelicon, A.; Pollak, S.; Šuštar, K. S.; Toman, A.; Valentinčič, A.; and Žnidaršič, M. 2024. Textual analysis of corporate sustainability reporting and corporate ESG scores. *International Review of Financial Analysis*, 96: 103669.

Goel, T.; Jain, P.; Verma, I.; Dey, L.; and Paliwal, S. 2020. Mining company sustainability reports to aid financial decision-making. In *Proc. of AAAI Workshop on Know. Disc. from Unstructured Data in Fin. Services*.

Hedberg, C.-J.; and Von Malmborg, F. 2003. The global reporting initiative and corporate sustainability reporting in Swedish companies. *Corporate social responsibility and environmental management*, 10(3): 153–164.

Huang, J.; and Chang, K. C.-C. 2023. Towards Reasoning in Large Language Models: A Survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, 1049–1065.

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

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv:2310.06825*.

Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1601–1611.

Kadavath, S.; Conerly, T.; Askell, A.; Henighan, T.; Drain, D.; Perez, E.; Schiefer, N.; Hatfield-Dodds, Z.; DasSarma, N.; Tran-Johnson, E.; et al. 2022. Language models (mostly) know what they know. *arXiv:2207.05221*.

Kamath, U.; Keenan, K.; Somers, G.; and Sorenson, S. 2024. LLM challenges and solutions. In *Large Language Models: A Deep Dive: Bridging Theory and Practice*, 219–274. Springer.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213.

Krithara, A.; Nentidis, A.; Bougiatiotis, K.; and Paliouras, G. 2023. BioASQ-QA: A manually curated corpus for Biomedical Question Answering. *Scientific Data*, 10(1): 170.

Lee, J.; Son, G.; and Kim, M. 2024. ESG-Kor: A Korean Dataset for ESG-related Information Extraction and Practical Use Cases. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 6627–6643.

Li, J.; Cheng, X.; Zhao, W. X.; Nie, J.-Y.; and Wen, J.-R. 2023. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 6449–6464.

Li, W. Y.; Chersoni, E.; and Ngai, C. S. B. 2024. Evaluating Multilingual Language Models for Cross-Lingual ESG Issue Identification. 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*, 50–58.

Li, X.; Huang, T.; Mu, R.; Huang, X.; and Jin, G. 2025. POT: Inducing Overthinking in LLMs via Black-Box Iterative Optimization. *arXiv:2508.19277*.

Luccioni, S.; Baylor, E.; and Duchene, N. 2020. Analyzing Sustainability Reports Using Natural Language Processing. In *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*.

Lyu, Q.; Havaldar, S.; Stein, A.; Zhang, L.; Rao, D.; Wong, E.; Apidianaki, M.; and Callison-Burch, C. 2023. Faithful chain-of-thought reasoning. In *The 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2023)*.

Mishra, L.; Berrospi, C.; Dinkla, K.; Antognini, D.; Fusco, F.; Bothur, B.; Lysak, M.; Livathinos, N.; Nassar, A.; Vagenas, P.; et al. 2024a. ESG Accountability Made Easy: DocQA at Your Service. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 23814–23816.

Mishra, L.; Dhibi, S.; Kim, Y.; Ramis, C. B.; Gupta, S.; Dolfi, M.; and Staar, P. 2024b. Statements: Universal Information Extraction from Tables with Large Language Models for ESG KPIs. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, 193–214.

Ni, J.; Bingler, J.; Colesanti-Senni, C.; Kraus, M.; Gostlow, G.; Schimanski, T.; Stammbach, D.; Vaghefi, S. A.; Wang, Q.; Webersinke, N.; et al. 2023. CHATREPORT: Democratizing Sustainability Disclosure Analysis through LLM-based Tools. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 21–51.

Niu, B. 2024. Government environmental protection expenditure and national ESG performance: Global evidence. *Innovation and Green Development*, 3(2): 100117.

Parikh, P.; and Penfield, J. 2024. Automatic Question Answering From Large ESG Reports. *International Journal of Data Warehousing and Mining (IJDWM)*, 20(1): 1–21.

Rossi, P.; and Candio, P. 2023. The independent and moderating role of choice of non-financial reporting format on forecast accuracy and ESG disclosure. *Journal of Environmental Management*, 345: 118891.

Seok, J.; Kim, Y.; and Oh, Y. K. 2024. How ESG shapes firm value: The mediating role of customer satisfaction. *Technological Forecasting and Social Change*, 208: 123714.

Sun, S.; and Ruan, W. 2023. TextVerifier: Robustness verification for textual classifiers with certifiable guarantees. In *Findings of the Association for Computational Linguistics: ACL 2023*, 4362–4380.

Sun, S.; Sen, P.; and Ruan, W. 2024. CROWD: Certified Robustness via Weight Distribution for Smoothed Classifiers against Backdoor Attack. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 17056–17070.

Team, G.; Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M. S.; Love, J.; et al. 2024. Gemma: Open Models Based on Gemini Research and Technology. *arXiv:2403.08295*.

Tian, K.; Mitchell, E.; Yao, H.; Manning, C. D.; and Finn, C. 2023. Fine-tuning language models for factuality. In *The Twelfth International Conference on Learning Representations*.

Wang, F.; Zhang, C.; Xu, P.; and Ruan, W. 2022. Deep learning and its adversarial robustness: A brief introduction. In *Handbook on computer learning and intelligence: Volume 2: Deep learning, intelligent control and evolutionary computation*, 547–584. World Scientific.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q. V.; Chi, E. H.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Xiao, T.; Gomez, A.; and Gal, Y. 2019. Wat zei je? detecting out-of-distribution translations with variational transformers. *Workshop on Bayesian Deep Learning at the Conference on Neural Information Processing Systems (NeurIPS, Vancouver, 2019)*.

Yi, D.; Mu, R.; Jin, G.; Qi, Y.; Hu, J.; Zhao, X.; Meng, J.; Ruan, W.; and Huang, X. 2024. Position: building guardrails for large language models requires systematic design. In *Forty-first International Conference on Machine Learning*.

Young-Ferris, A.; and Roberts, J. 2023. 'Looking for something that isn't there': a case study of an early attempt at ESG integration in investment decision making. *European Accounting Review*, 32(3): 717–744.

Yu, E. P.-y.; Van Luu, B.; and Chen, C. H. 2020. Greenwashing in environmental, social and governance disclosures. *Research in International Business and Finance*, 52: 101192.

Zhang, Z.; Zhang, A.; Li, M.; and Smola, A. 2023. Automatic Chain of Thought Prompting in Large Language Models. In *The Eleventh International Conference on Learning Representations*.

Zhao, Z.; Cohen, S. B.; and Webber, B. 2020. Reducing Quantity Hallucinations in Abstractive Summarization. In Cohn, T.; He, Y.; and Liu, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2237–2249. Online: Association for Computational Linguistics.