Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Auditing the Auditor: Extraction Accuracy, Hallucination, and Abstention in Open-Weight LLMs for ESG Reporting A Longitudinal Empirical Benchmark Across 518 Corporate Disclosures (2010–2026)

Redeemer Salami Okekale

redeemersalamiokekale@gmail.com

University of Ghana

Kate Markin

kmarkin001@st.ug.edu.gh

University of Ghana https://orcid.org/0009-0004-3037-8888

Derek Amuna

University of Ghana https://orcid.org/0009-0003-1889-7602

Felix Coker

University of Ghana

Research Article

Keywords: Large Language Models (LLMs), ESG Reporting, Sustainability Assurance, Corporate Sustainability Reporting Directive (CSRD), AI Hallucination, Faithful Abstention, Automated Au- diting.

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11610550/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Auditing the Auditor: Extraction Accuracy, Hallucination, and Abstention in Open-Weight LLMs for ESG Reporting

A Longitudinal Empirical Benchmark Across 518 Corporate Disclosures (2010–2026)

Redeemer Salami O. * **Kate Markin** [1 †] **Derek Amuna** [1 ‡] **Felix Coker** [§]

[1]University of Ghana

October 2026

Abstract

Corporate sustainability auditing increasingly relies on Large Language Models (LLMs) to extract verified quantitative metrics from lengthy Environmental, Social, and Governance (ESG) reports. For an LLM to serve as a trustworthy compliance tool, it must pull exact numbers from dense tables while refusing to guess when a disclosure is missing. We evaluate five leading open-weight LLMs (**Qwen3-4B**, **DeepSeek-R1-7B**, **GPT-OSS-20B**, **Qwen3.5-27B**, and **Gemma-4-31B**) across 518 corporate sustainability reports published between 2010 and 2026 (279,862 text passages). Using a stratified 120-question audit benchmark tested under five controlled evidence conditions (3,000 total responses), we grade every output using an independent referee model (**Phi-4-14B**) so that no model grades its own family's work. Three findings stand out for sustainability practitioners: (1) models summarize narrative policies reliably (F1 accuracy ≈ 0.729) but struggle to extract exact numerical indicators such as Scope 1 emissions or water withdrawal (F1 ≈ 0.335); (2) even when handed the exact gold-standard page from a report, the strongest model (**GPT-OSS-20B**) reaches an accuracy ceiling of 0.731, showing that table-reading errors matter more than search errors; and (3) when given a "near-miss" passage from the wrong year or company, standard prompts cause models to invent or substitute numbers in up to 46.7% of cases. By contrast, step-by-step reasoning in **DeepSeek-R1-7B** catches the mismatch and abstains 77.5% of the time unprompted, while strict audit instructions raise abstention across dense models to 74.2%–80.0%.

Keywords: Large Language Models (LLMs), ESG Reporting, Sustainability Assurance, Corporate Sustainability Reporting Directive (CSRD), AI Hallucination, Faithful Abstention, Automated Auditing.

1. Introduction

Corporate sustainability reports have shifted from voluntary marketing brochures into legally binding regulatory filings [1, 4]. Institutional investors, credit rating agencies, and statutory auditors now comb through these disclosures to verify greenhouse gas (GHG) inventories, workforce safety records, and board oversight [17, 18]. In the European Union, Directive (EU) 2022/2464, known as the Corporate Sustainability Reporting Directive (CSRD) [8], and the accompanying European Sustainability Reporting Standards (ESRS) [7] require large undertakings to publish standardized,

*Corresponding author: redeemersalamiokekale@gmail.com | ORCID: 0009-0000-2900-592X
†ORCID: 0009-0004-3037-8888 | kmarkin001@st.ug.edu.gh
‡ORCID: 0009-0003-1889-7602 | dadadee02@gmail.com
§ORCID: 0009-0007-7509-8371

externally assured ESG data. Table 1 summarizes the major global reporting frameworks shaping this compliance environment.

Table 1: Major Global ESG Reporting Frameworks and Directives Governing Corporate Sustainability Disclosures.

Framework	Scope and Relevance to Automated Auditing
GRI [10]	**Global Reporting Initiative:** The most widely adopted voluntary framework worldwide; focuses on a company's outward environmental, social, and economic impacts.
SASB [21]	**Sustainability Accounting Standards Board:** Industry-specific standards (now consolidated under the IFRS ISSB) that isolate financially material sustainability risks for investors.
TCFD [22]	**Task Force on Climate-related Financial Disclosures:** Governance and risk-management framework for physical and transition climate risks and scenario modeling.
CDP [3]	**Carbon Disclosure Project:** Standardized questionnaire and scoring system tracking corporate greenhouse gas (GHG) emissions, water security, and supply-chain deforestation.
CSRD [8]	**Corporate Sustainability Reporting Directive:** Binding European Union legislation requiring mandatory, externally assured sustainability disclosures from large undertakings.
ESRS [7]	**European Sustainability Reporting Standards:** The 12 technical rulebooks (ESRS 1–2 general, E1–E5 environmental, S1–S4 social, G1 governance) implementing the CSRD under double materiality.

Despite stricter standards, reading corporate ESG filings at scale remains a major bottleneck. Unlike financial statements built around uniform accounting tables (such as IFRS or US GAAP), a single sustainability report often runs 150 to 400 pages, mixing narrative pledges with multi-page footnotes, shifting reporting boundaries, and non-standard measurement units. An analyst or automated system looking for a firm's verified Scope 1 direct emissions in metric tons of CO_2e must separate hard historical figures from aspirational net-zero targets and qualitative boilerplate.

To automate this work without sending sensitive corporate documents to commercial cloud APIs, financial institutions and assurance firms are turning to **open-weight Large Language Models (LLMs)**, meaning models whose weights can be downloaded and run privately on an organization's own servers. Rather than asking an LLM to answer from memory (which causes it to invent facts), audit pipelines use **document-grounded generation**, technically known as *Retrieval-Augmented Generation (RAG)* [9, 14]. The workflow is straightforward: a search engine first locates the most relevant paragraphs from the company's PDF report and pastes those excerpts into the prompt alongside the auditor's question.

Yet before a sustainability team or assurance provider can trust an open-weight LLM on real regulatory filings, four practical questions must be answered:

- **Numbers vs. Narrative:** Do open-weight LLMs extract hard numerical Key Performance Indicators (KPIs) as reliably as they summarize qualitative corporate policies?
- **Single Excerpt vs. Multiple Excerpts:** Does feeding the model three report passages ($k = 3$) instead of one ($k = 1$) help it piece together complete answers, or does extra text confuse it?

- **The Missing-Data Trap:** When the search step pulls a *near-miss* passage, such as an excerpt discussing 2022 water usage when the auditor asked for 2024 hazardous waste, does the LLM honestly state that the figure is missing, or does it fabricate an answer?
- **Voluntary vs. Mandatory CSRD Filings:** Has the transition from older voluntary GRI reports (2010–2017) to standardized CSRD/ESRS filings (2025–2026) made automated extraction easier or harder?

Existing studies leave these questions open. Most evaluate only proprietary models such as GPT-4 [27], test on small synthetic datasets where every question is guaranteed to have an answer in the text [19], or allow the same AI model family to generate the test questions, answer them, and grade its own responses [6, 11].

Contributions of This Study. We address these gaps through a large-scale audit benchmark of five modern open-weight LLMs across 518 real corporate sustainability reports spanning 2010 to 2026. Specifically, we contribute:

1. **A 16-Year Longitudinal ESG Corpus (2010–2026):** We assemble 518 full-length sustainability reports from EU-listed companies, adding 20 newly released 2025–2026 mandatory CSRD/ESRS filings to the 498-report historical collection of Saad et al. [19]. The full corpus contains 279,862 searchable text passages indexed with **Qwen3-Embed-4B**.
2. **A Stratified 120-Question Audit Benchmark with Near-Miss Traps:** We design 120 realistic audit questions balanced across four question types (Hard Numerical KPIs, Year-over-Year Comparisons, Policy Disclosures, and Cross-Pillar Syntheses) and three regulatory eras. Every question is paired with a verified gold passage as well as an unanswerable *near-miss* passage to test whether models fabricate data under pressure.
3. **Controlled Testing Across Five Open-Weight LLMs and Five Evidence Regimes:** We test **Qwen3-4B**, **DeepSeek-R1-7B**, **GPT-OSS-20B**, **Qwen3.5-27B**, and **Gemma-4-31B** across five evidence setups ($k = 1$, $k = 3$, Oracle gold passage, Near-Miss Standard prompt, and Near-Miss Strict guardrail prompt), producing 3,000 complete audit responses on NVIDIA H100 hardware.
4. **Independent Referee Grading:** To prevent self-grading bias, we score all 3,000 responses using Microsoft's **Phi-4-14B**, a 14-billion-parameter model from a completely separate model family, combined with automated numerical verification for hard KPIs.

2. Background & Prior Work

2.1. How Document-Grounded LLMs Work in Sustainability Auditing

When a human sustainability auditor verifies a metric, they do not rely on memory; they look up the relevant table in the company's annual ESG report and read the exact figure. Document-grounded LLM pipelines, commonly called Retrieval-Augmented Generation (RAG) [9, 14], follow the exact same two-step process (Figure 1):

$$\mathcal{D}_k = \text{Search}(q, \mathcal{D}; k) = \arg\text{top-}\,k_{d\in\mathcal{D}} \cos\big(E(q), E(d)\big), \tag{1}$$

$$y = \text{LLM}\big(q, \mathcal{D}_k; \theta\big). \tag{2}$$

In Step 1 (Equation 1), an embedding model $E(\cdot)$ converts both the auditor's question q and every passage d from the report database $\mathcal{D}$ into numerical coordinates (vectors), selecting the top k passages $\mathcal{D}_k$ whose meaning is closest to the question. In Step 2 (Equation 2), the language model reads the question alongside those k retrieved passages and writes the audit response y. Because we want to compare how well different LLMs read and verify ESG disclosures, we hold the search index constant and test all five models on the exact same retrieved passages.

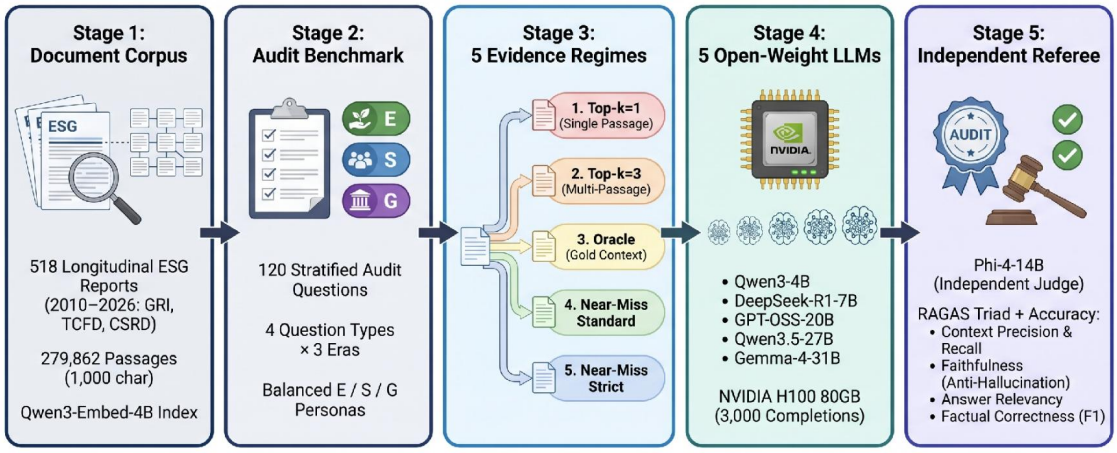


Figure 1: Overview of our five-stage ESG audit evaluation pipeline. Report passages from 518 corporate disclosures are matched to audit questions under five evidence conditions ($k = 1$, $k = 3$, Oracle, Near-Miss Standard, and Near-Miss Strict). Responses from five open-weight LLMs are graded by an independent referee model (**Phi-4-14B**) to ensure fair, unbiased scoring. Figure created with BioRender.com.

2.2. Automated Audit Grading & The "Grading Your Own Homework" Problem

Evaluating thousands of LLM responses by hand is impractical, so recent benchmarks use a strong language model as an automated referee, an approach known as *LLM-as-a-Judge* [5, 6, 20]. Standardized frameworks such as Ragas [6] break audit quality into three core checks: (1) whether the search step found the right passage (*Context Precision and Recall*), (2) whether the LLM stuck strictly to the provided text without inventing facts (*Faithfulness*), and (3) whether the final answer matches the verified figure in the report (*Factual Correctness*).

However, automated grading carries a well-documented flaw: **self-enhancement bias** [11, 13, 26]. When a model from the Qwen or GPT family is used to grade answers written by itself or a sibling model from the same family, it systematically awards higher scores to its own phrasing and overlooks its own mistakes [15]. For example, the baseline study of Saad et al. [19] used **Qwen3-4B** to write the benchmark questions, generate the answers, and act as the judge. In financial and sustainability assurance, independence between the preparer and the auditor is a foundational rule. We enforce that same separation by using Microsoft's **Phi-4-14B** as an external referee that shares no training lineage with the Qwen, DeepSeek, OpenAI, or Google models under test.

2.3. LLMs in ESG and Financial Document Analysis

Interest in applying LLMs to sustainability disclosures has grown rapidly alongside regulatory mandates. Zou et al. [27] built an extraction pipeline using GPT-4 to audit 166 Hong Kong company reports, showing promising accuracy on structured prompts but leaving open-weight models untested. Mankour et al. [16] introduced Climate Finance Bench over corporate climate PDFs and found that even commercial AI systems answered only 62% of expert-verified questions accurately. Among open-source efforts, SusGen-GPT [23] trained 7B–8B models on sustainability text, and Birti et al. [2] tuned Llama models for EU Taxonomy classification. More recently, ESGenius [12] and MMESGBench [25] explored multimodal chart reading, while ESGLens [24] examined automated ESG rating prediction.

2.4. Four Gaps in Current ESG AI Benchmarks

Despite this progress, sustainability teams looking to deploy open-weight LLMs face four practical blind spots in the existing literature:

1. **Lack of Head-to-Head Open-Weight Comparisons:** Prior work focuses on expensive commercial APIs or a single open model family, offering little guidance on how newer architectures across the 4B to 31B size range compare on identical ESG tasks.

2. **Single-Passage Search Limits:** Earlier studies test only single-passage retrieval ($k = 1$). When an answer fails, practitioners cannot tell whether the search engine missed the right page or the LLM failed to read the table properly.

3. **No Stress-Testing on Missing Disclosures:** Standard benchmarks only ask questions where the answer is present in the passage. In real corporate audits, companies frequently omit sensitive metrics (such as Scope 3 supply-chain emissions or gender pay gaps), and search tools routinely return passages from the wrong year or subsidiary. Benchmarks must test whether models faithfully abstain or fabricate numbers when evidence is missing.

4. **Absence of 2025–2026 CSRD Filings:** Most public ESG datasets stop in 2023 or 2024, before the first wave of mandatory European CSRD/ESRS reports took effect.

3. Methodology

3.1. Audit Evaluation Metrics Explained

To evaluate each model from an auditor's perspective, we measure six complementary metrics [6, 19] scored by our independent **Phi-4-14B** referee:

- **Context Precision ($CtxPrec$):** Checks whether the search engine placed the most relevant report passage at the very top of the search results. Formally, for K retrieved passages with relevance indicators $r_k \in \{0, 1\}$:

$$CtxPrec = \frac{1}{\sum_{k=1}^{K} r_k} \sum_{k=1}^{K} \left(\frac{\sum_{i=1}^{k} r_i}{k} \right) \cdot r_k. \tag{3}$$

- **Context Recall ($CtxRec$):** Measures completeness of the retrieved text, indicating what share of the facts $c_i \in \mathcal{C}_{ref}$ needed for the correct answer actually appeared in the retrieved passages $\mathcal{D}_k$:

$$CtxRec = \frac{|\{c_i \in \mathcal{C}_{ref} \mid \mathcal{D}_k \models c_i\}|}{|\mathcal{C}_{ref}|}. \tag{4}$$

- **Faithfulness ($Faith$):** Our primary anti-hallucination metric. It measures the fraction of statements $s_j \in \mathcal{S}_{ans}$ in the model's answer that can be directly verified in the provided report excerpt $\mathcal{D}_k$:

$$Faith = \frac{|\{s_j \in \mathcal{S}_{ans} \mid \mathcal{D}_k \models s_j\}|}{|\mathcal{S}_{ans}|}. \tag{5}$$

  If a passage does not contain the answer and the model appropriately abstains, it makes zero unverified claims and receives a faithfulness score of 1.0.

- **Answer Relevancy ($AnsRel$):** Measures whether the model directly addressed the auditor's question without drifting off-topic, computed via semantic similarity between the original question q and M back-generated questions $\hat{q}_m$:

$$AnsRel = \frac{1}{M} \sum_{m=1}^{M} \cos\big(E(q), E(\hat{q}_m)\big). \tag{6}$$

- **Factual Correctness ($FactCorr$):** An overall accuracy score (0 to 1 F1 scale) measuring how closely the facts and numbers in the LLM's response match the verified reference answer from the report.

- **Faithful Abstention Rate ($Abstain$):** On unanswerable near-miss tests, the percentage of cases where the model refuses to guess and explicitly states that the requested metric is not in the provided excerpt.

3.2. The 518-Report Longitudinal ESG Corpus (2010–2026)

Our dataset spans 16 years of sustainability reporting by major European corporations. We start from the 498-report collection of Saad et al. [19] (2010–2024) and add 20 full-length 2025 and 2026 sustainability statements published under mandatory CSRD/ESRS rules by 17 European enterprises, including Siemens, ASML, SAP, Enel, L'Oréal, and Schneider Electric. Together, the corpus totals 518 full reports.

Each PDF report is converted into structured text using PyMuPDF (fitz), preserving section headings, tables, and page markers. To make the reports searchable, we slice each document into overlapping 1,000-character passages (with a 200-character overlap so sentences and table rows are not cut cleanly in two without context), removing tiny header fragments under 50 characters. This yields 279,862 passages across the 518 reports. Every passage is indexed using the **Qwen3-Embed-4B** embedding model. Figure 2 shows our five-step benchmark construction workflow alongside the yearly distribution of reports across the three regulatory eras.

3.3. Designing the Stratified 120-Question Audit Benchmark

Real sustainability audits involve very different types of questions, ranging from looking up a single emissions figure to summarizing a supplier code of conduct. We construct a stratified 120-question benchmark balanced across two dimensions:

1. **Four Audit Question Types:**
   - *Hard Numerical KPIs ($N = 55$):* Questions requiring an exact number, unit, and year (for example, gross Scope 1 GHG emissions in tCO_2e, total water withdrawal in megaliters, or lost-time injury frequency rates).
   - *Year-over-Year (YoY) Comparisons ($N = 20$):* Questions asking how a specific indicator changed between two reporting periods (for example, the percentage reduction in Scope 2 emissions from 2022 to 2024).
   - *Policy & Governance Boilerplate ($N = 25$):* Qualitative questions covering board oversight committees, whistleblower protections, and human rights due diligence policies.
   - *Cross-Pillar Syntheses ($N = 20$):* Questions linking commitments across Environmental, Social, and Governance topics (such as how executive bonus pay is tied to emissions reduction targets).
2. **Three Regulatory Reporting Eras:** Balanced across the *Voluntary GRI Era* (2010–2017, $N = 40$), the *TCFD/NFRD Transition Era* (2018–2024, $N = 44$), and the *Mandatory CSRD/ESRS Era* (2025–2026, $N = 36$).

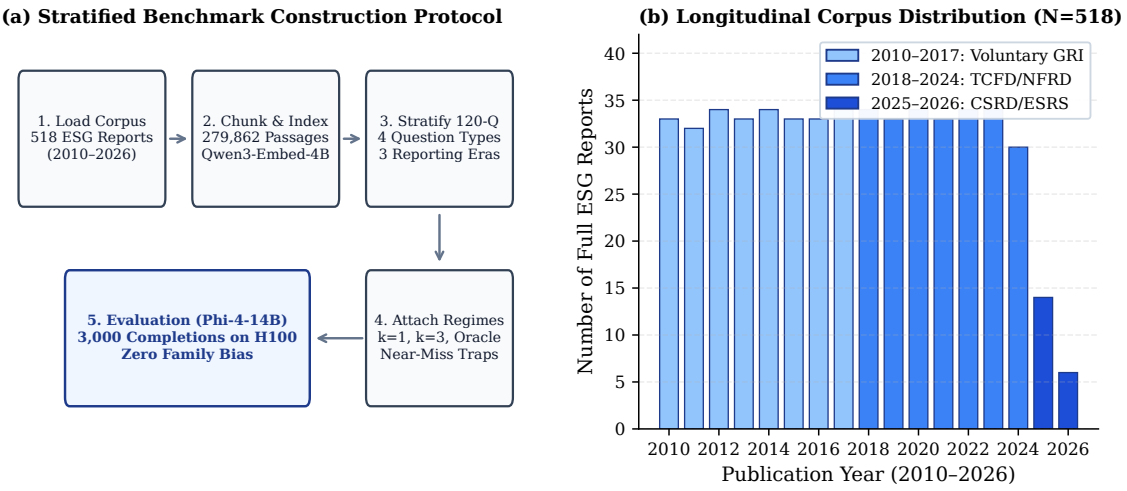


Figure 2: (a) Five-step protocol used to build and verify our audit benchmark. (b) Distribution of the 518 corporate ESG reports across three regulatory eras: Voluntary GRI (2010–2017), TCFD/NFRD Transition (2018–2024), and Mandatory CSRD/ESRS (2025–2026).

To ensure balanced coverage across E, S, and G topics, questions are generated across three specialized auditor profiles (Listing 1). For every question, we attach four distinct sets of report excerpts: the single best-matching search passage ($k = 1$), the top three search passages combined ($k = 3$), the exact gold-standard passage from which the question was verified (Oracle), and an unanswerable *near-miss* distractor passage. We construct the near-miss passage by picking a closely related

excerpt from a different company or reporting year; for instance, a passage discussing water recycling in 2022 when the question asks for hazardous waste in 2024, so that the topic looks familiar but the requested figure is genuinely absent.

```
from ragas.testset.persona import Persona
persona_env = Persona(name="Environmental KPI Auditor",
    role_description="Audits quantitative environmental disclosures: Scope 1-3 GHG emissions, energy
      intensity, water withdrawal, hazardous waste, and renewable electricity shares.")
persona_soc = Persona(name="Social KPI Auditor",
    role_description="Audits human capital disclosures: workforce gender diversity, pay gap ratios, lost-time
      injury frequency rates, and employee turnover.")
persona_gov = Persona(name="Governance KPI Auditor",
    role_description="Audits corporate oversight disclosures: board independence, executive sustainability
      incentives, and substantiated corruption or whistleblower cases.")
```

Listing 1: ESG Auditor Profiles Used to Balance Environmental, Social, and Governance Questions.

3.4. The Five Evaluated Open-Weight LLMs

We evaluate five open-weight models spanning 4 billion to 31 billion parameters. In practical terms, parameter count reflects model size and hardware cost: a 4B–7B model can run on a high-end laptop or single office workstation, whereas a 27B–31B dense model requires a dedicated data-center graphics card (GPU). Our five models represent four major AI developers and three distinct design philosophies:

- **Qwen3-4B** ($\mathrm{Qwen/Qwen3\text{-}4B\text{-}Instruct}$, 4.0B parameters): Alibaba's compact instruction-tuned model. We include it to test how well a lightweight, low-cost model handles ESG auditing on modest hardware.
- **DeepSeek-R1-7B** ($\mathrm{deepseek\text{-}ai/DeepSeek\text{-}R1\text{-}Distill\text{-}Qwen\text{-}7B}$, 7.6B parameters): A reasoning-focused model trained to "think out loud" inside $<\mathrm{think}>...</\mathrm{think}>$ blocks before giving its final answer, allowing it to double-check dates, company names, and units.
- **GPT-OSS-20B** ($\mathrm{openai/gpt\text{-}oss\text{-}20b}$, 20.9B total / 3.6B active parameters): OpenAI's open-weight *Mixture-of-Experts (MoE)* model. Instead of activating all 20.9 billion parameters for every word, it routes each word through specialized sub-networks using only 3.6 billion active parameters at a time, combining large-model knowledge with small-model speed.
- **Qwen3.5-27B** ($\mathrm{Qwen/Qwen3.5\text{-}27B}$, 27.0B parameters): A large, traditional dense model from the Qwen family designed for high-accuracy enterprise document tasks.
- **Gemma-4-31B** ($\mathrm{google/gemma\text{-}4\text{-}31B\text{-}it}$, 30.7B parameters): Google DeepMind's flagship open-weight model and the largest dense model in our benchmark.

3.5. Experimental Conditions & Prompt Design

All tests were run on an NVIDIA H100 (80GB VRAM) GPU on Modal Cloud. To guarantee that another auditor running the exact same test gets the exact same answers, we turned off random sampling ($\mathrm{temperature} = 0.0$, greedy decoding). Table 2 lists the complete technical settings.

Each of the five models answered all 120 audit questions under five controlled evidence conditions ($5 \times 5 \times 120 = 3,000$ total audit completions):

1. **Single-Passage Search ($k = 1$):** The model receives only the single top-ranked passage returned by the search engine.
2. **Three-Passage Search ($k = 3$):** The model receives the top three passages combined, testing whether wider context helps recover missing table rows or introduces distracting noise.
3. **Gold-Standard Passage (Oracle Ceiling):** The model is handed the exact passage from which the ground-truth answer was verified. This removes search errors completely and measures the LLM's pure reading and extraction ceiling.
4. **Near-Miss Distractor (Standard Prompt):** The model is given an unanswerable near-miss passage alongside a standard helpful-assistant prompt (Listing 2).
5. **Near-Miss Distractor (Strict Guardrail Prompt):** The model receives the same near-miss passage, but with an explicit compliance instruction ordering it to abstain if the exact metric or year is missing (Listing 2).

Table 2: Summary of Technical Settings Across Document Processing, Audit Benchmark Design, Model Testing, and Independent Scoring.

Pipeline Stage	Specification
1. Document Corpus & Search Index	
Corpus Scope	518 full-length corporate ESG reports (2010–2026), yielding 279,862 text passages
PDF Text Extraction	PyMuPDF (fitz) preserving section headings, tables, and page boundaries
Passage Length	1,000 characters per passage with a 200-character overlap between consecutive passages
Search Embedding Model	**Qwen3-Embed-4B** (Qwen/Qwen3-Embedding-4B, 1,024-dimensional vectors)
Search Matching Method	Cosine similarity matching across all 279,862 indexed report passages
2. Stratified 120-Question Audit Benchmark	
Question Breakdown	120 audit questions across 4 question types (Hard KPIs, Multi-Year Trends, Policies, Cross-Pillar) and 3 reporting eras
ESG Topic Balance	Balanced rotation across Environmental (E), Social (S), and Governance (G) auditor profiles
Evidence Conditions	Single-Passage ($k = 1$), Three-Passage ($k = 3$), Gold-Standard Passage (Oracle), and Near-Miss Distractor (Standard & Strict Prompts)
3. Model Execution Infrastructure	
Compute Hardware	Single NVIDIA H100 80GB GPU hosted on Modal Cloud
Numerical Precision	Standard 16-bit (BFloat16) for dense models; native 4-bit (MXFP4) for **GPT-OSS-20B**
Generation Settings	Deterministic output (temperature $= 0.0$, greedy decoding) for 100% audit reproducibility
Response Length Limit	Up to 512 generated words/tokens per answer; 8,192-token input window
4. Independent AI Referee (LLM-as-a-Judge)	
Referee Model	**Phi-4-14B** (microsoft/phi-4, 14B parameters; independent from all tested model families)
Core Quality Metrics	Context Precision, Context Recall, Faithfulness, Answer Relevancy, and Factual Correctness (F1)
Numerical KPI Verification	Automated number and unit verification allowing a $\pm 1\%$ rounding tolerance
Total Audit Volume	5 models $\times$ 5 evidence conditions $\times$ 120 questions = 3,000 evaluated responses

```
# 1. Standard Prompt
System: "You are a helpful assistant that answers questions based on given documents only."
```

```
User:   "question: {query}\n\nDocuments: {retrieved_doc}"
# 2. Strict Compliance Guardrail Prompt
System: "You are an expert compliance auditor. Answer the question using ONLY the provided documents. If
        the exact answer, metric, or reporting year is not explicitly stated in the documents, you MUST respond
        exactly: 'I cannot answer based on the provided documents.' Do not extrapolate, assume, or substitute
        related metrics."
User:   "question: {query}\n\nDocuments: {retrieved_doc}"
```

Listing 2: Standard Prompt vs. Strict Compliance Guardrail Prompt.

4. Results: Open-Weight LLM Performance in Corporate ESG Auditing

4.1. Overall Audit Performance Across Models

Table 3 reports the complete evaluation scores across all five open-weight LLMs and four core evidence regimes, while Figure 3 provides a visual comparison of single-passage ($k = 1$) and three-passage ($k = 3$) performance.

When given just a single search passage ($k = 1$), all five models stick closely to the provided text ($Faith \geq 0.700$) and stay on topic ($AnsRel \geq 0.761$), yet overall factual accuracy ($FactCorr$) ranges between 0.389 and 0.474. OpenAI's **GPT-OSS-20B** leads overall accuracy at 0.474 and achieves the highest score on hard numerical KPIs (0.417), followed by the compact **Qwen3-4B** (0.428) and Google's **Gemma-4-31B** (0.414). Notice an instructive quirk in Table 3: **Qwen3.5-27B** records the highest faithfulness score in $k = 1$ (0.912) but the lowest factual accuracy (0.389). Because a single 1,000-character passage often cuts off the bottom half of a table, **Qwen3.5-27B** frequently plays it safe and abstains (16.7% of the time), which protects its faithfulness score but lowers its recall of partial facts.

Expanding the search window from one passage ($k = 1$) to three passages ($k = 3$) raises search recall ($CtxRec$ jumps from 0.562 to 0.662) and improves factual accuracy for every model in our test. Under $k = 3$, **GPT-OSS-20B** (0.522) and **Gemma-4-31B** (0.519) take the top two spots, with **Qwen3.5-27B** rising to 0.497 and **Qwen3-4B** reaching 0.490. For hard numerical KPIs specifically, three-passage context lifts extraction accuracy by 7 to 15 percentage points across all five models.

4.2. Search Bottleneck vs. Model Reading Ceiling ($k = 1 \rightarrow k = 3 \rightarrow$ Oracle)

When an automated ESG audit produces a wrong number, who is to blame: the search engine for pulling the wrong page, or the LLM for misreading the table? Figure 4 separates these two factors by tracking accuracy from single-passage search ($k = 1$) to three-passage search ($k = 3$) and finally to the gold-standard Oracle passage.

Finding 1 (Table-Reading Limits Matter Even When Search Is Perfect). *Handing models the exact gold-standard passage (Oracle) boosts factual accuracy by an average of +16.0 percentage points over $k = 3$ search (rising from 0.492 to 0.651). Yet even with the exact right passage in front of it, the top-performing model (**GPT-OSS-20B**) tops out at 0.731 F1, leaving a 26.9-point accuracy gap caused by misreading dense multi-column ESG tables and footnotes.*

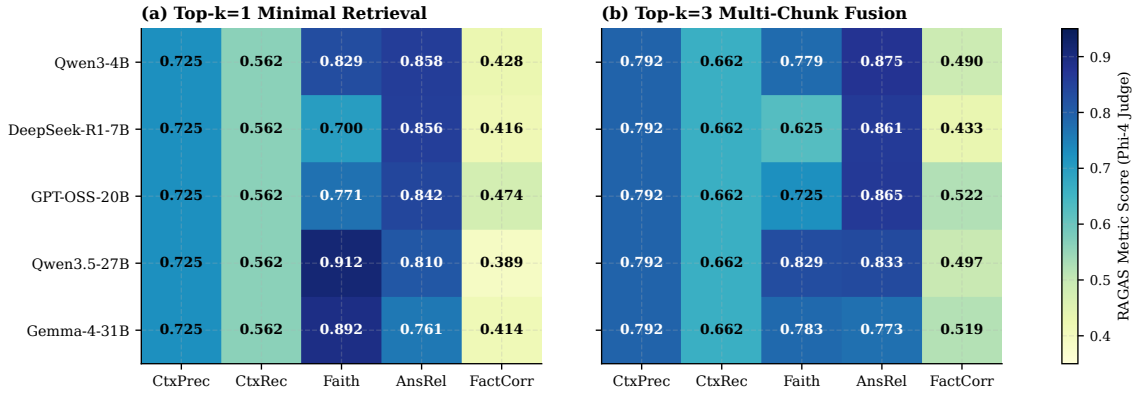


Figure 3: Performance heatmap across five open-weight LLMs under (a) single-passage search ($k = 1$) and (b) three-passage search ($k = 3$). Search metrics (*CtxPrec*, *CtxRec*) are identical across models because they share the same search index, whereas faithfulness and factual accuracy vary sharply by model architecture.

Table 3: Overall Audit Performance Across Five Open-Weight LLMs and Four Evidence Conditions ($N = 120$ questions per condition, scored by independent referee **Phi-4-14B**). Best score per column within each condition is in **bold**.

Evidence Condition	Evaluated Model	*CtxPrec*	*CtxRec*	*Faith*	*AnsRel*	*FactCorr*	*HardKPI*	*Abstain*
Top-$k = 1$ (Single Retrieved Passage)								
	Qwen3-4B	**0.725**	**0.562**	0.829	**0.858**	0.428	0.351	14.2%
	DeepSeek-R1-7B	**0.725**	**0.562**	0.700	0.856	0.416	0.344	7.5%
	GPT-OSS-20B	**0.725**	**0.562**	0.771	0.842	**0.474**	**0.417**	10.8%
	Qwen3.5-27B	**0.725**	**0.562**	**0.912**	0.810	0.389	0.340	**16.7%**
	Gemma-4-31B	**0.725**	**0.562**	0.892	0.761	0.414	0.349	15.0%
Top-$k = 3$ (Three Retrieved Passages Combined)								
	Qwen3-4B	**0.792**	**0.662**	0.779	**0.875**	0.490	**0.505**	2.5%
	DeepSeek-R1-7B	**0.792**	**0.662**	0.625	0.861	0.433	0.463	**4.2%**
	GPT-OSS-20B	**0.792**	**0.662**	0.725	0.865	**0.522**	0.485	2.5%
	Qwen3.5-27B	**0.792**	**0.662**	**0.829**	0.833	0.497	0.456	**4.2%**
	Gemma-4-31B	**0.792**	**0.662**	0.783	0.773	0.519	0.496	2.5%
Oracle (Exact Gold-Standard Source Passage Provided)								
	Qwen3-4B	**1.000**	**0.942**	0.954	**0.886**	0.677	**0.696**	**0.0%**
	DeepSeek-R1-7B	**1.000**	**0.942**	0.887	0.880	0.609	0.610	**0.0%**
	GPT-OSS-20B	**1.000**	**0.942**	0.908	0.865	**0.731**	0.650	**0.0%**
	Qwen3.5-27B	**1.000**	**0.942**	**0.971**	0.829	0.590	0.578	**0.0%**
	Gemma-4-31B	**1.000**	**0.942**	0.954	0.778	0.649	0.626	**0.0%**
Near-Miss Strict (Unanswerable Distractor Passage + Guardrail Prompt)								
	Qwen3-4B	**0.346**	**0.221**	0.946	**0.862**	0.126	0.047	75.8%
	DeepSeek-R1-7B	**0.346**	**0.221**	0.817	0.846	0.113	0.066	65.0%
	GPT-OSS-20B	**0.346**	**0.221**	0.958	0.812	0.116	0.053	**80.0%**
	Qwen3.5-27B	**0.346**	**0.221**	**0.975**	0.814	0.116	0.044	78.3%
	Gemma-4-31B	**0.346**	**0.221**	**0.975**	0.761	**0.145**	**0.080**	74.2%

Look closely at the Oracle section of Table 3. When search errors are removed, **GPT-OSS-20B** reaches 0.731 factual accuracy, **Qwen3-4B** reaches 0.677, and **Gemma-4-31B** reaches 0.649. At the same time, faithfulness climbs above 0.887 for every model, reaching 0.971 on **Qwen3.5-27B**. In plain

terms, once an LLM is shown the right excerpt from a sustainability report, it rarely invents outside claims; instead, its remaining mistakes come from misaligning row headers, confusing parent-company vs. consolidated subsidiary columns, or dropping measurement units (such as reading thousands of tons as single tons).

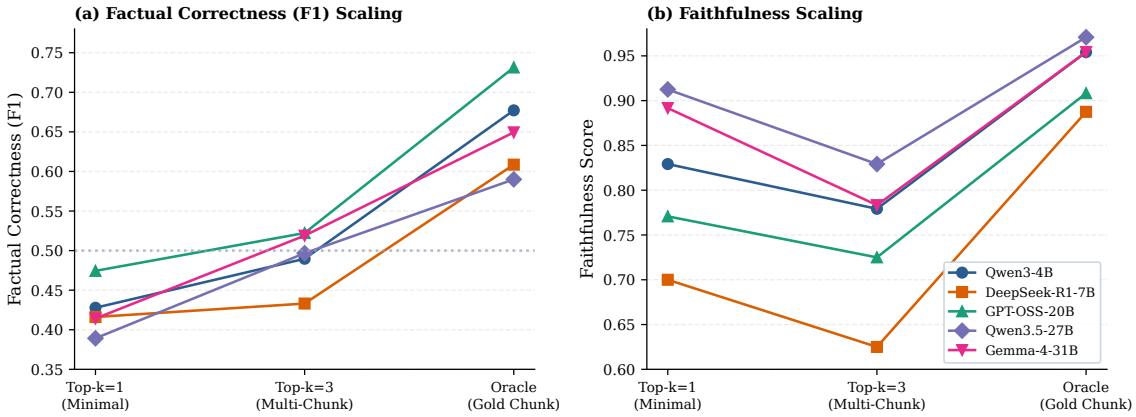


Figure 4: Accuracy gains as evidence quality improves from $k = 1$ to $k = 3$ to the gold-standard Oracle passage. (a) Factual accuracy rises by +12 to +21 points when search errors are eliminated, but still hits a ceiling below 0.74. (b) Faithfulness approaches 0.89–0.97 when models are given the exact source passage.

4.3. The Missing-Data Stress Test: Faithful Abstention vs. Hallucination

In statutory sustainability assurance, a false positive, namely inventing an emissions figure that is not in the filing or quietly substituting a 2022 number when asked for 2024, is far more dangerous than admitting that a figure could not be found. Our near-miss regime tests this exact failure mode by feeding each model a passage from the wrong year or entity where the requested metric is missing. Figure 5 compares how often models faithfully abstain versus how often they hallucinate under standard prompting and under our strict compliance guardrail prompt.

Finding 2 (Built-In Reasoning Catches Missing Data Without Special Prompts). *Under a standard prompt, conventional instruction models fabricate or substitute wrong metrics in 14.2% to 24.2% of unanswerable questions. By contrast, the reasoning-trained **DeepSeek-R1-7B** checks the company name and reporting year inside its internal thinking trace, abstaining faithfully on 77.5% of unanswerable questions unprompted; and when dense models are given an explicit strict guardrail prompt, their abstention rates jump to 74.2%–80.0%.*

Notice the striking contrast between the two panels of Figure 5. Under the standard RAG prompt, **Qwen3-4B** abstains 55.8% of the time and fabricates or substitutes an incorrect figure in 24.2% of cases. Adding a strict compliance guardrail prompt transforms behavior across all conventional models: **GPT-OSS-20B** reaches an 80.0% faithful abstention rate (with hallucinations dropping from 17.5% down to 5.8%), **Qwen3.5-27B** reaches 78.3% abstention (7.5% hallucination), and **Gemma-4-31B** reaches 74.2% abstention (7.5% hallucination). Meanwhile, **DeepSeek-R1-7B** exhibits a fascinating trade-off: under standard prompting, its internal verification step spots the year/entity mismatch 77.5% of the time on its own, though when combined with a rigid boilerplate abstention string in the strict prompt, its free-form reasoning occasionally paraphrases the refusal (65.0% exact match).

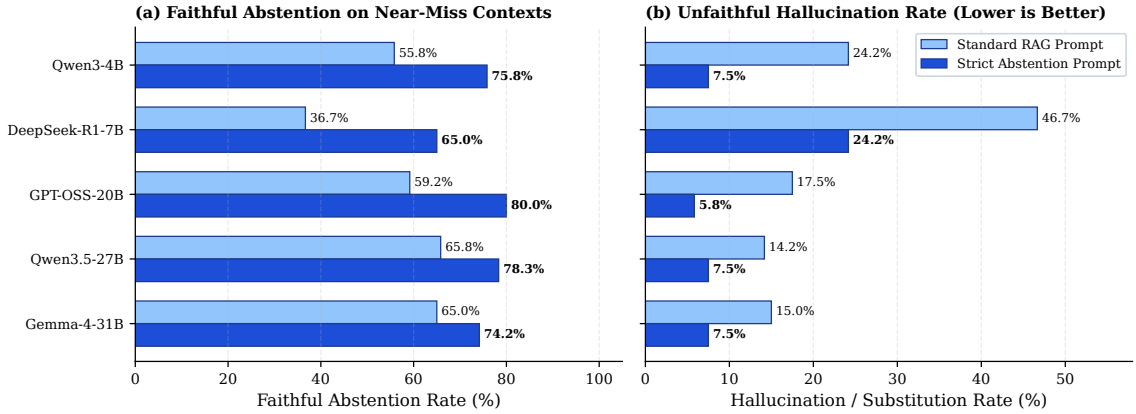


Figure 5: Near-Miss stress test on unanswerable audit questions. (a) Faithful abstention rates under standard vs. strict guardrail prompts. (b) Unfaithful hallucination and metric-substitution rates (lower is better), showing how explicit compliance guardrails cut fabrication rates down to 5.8%–7.5% on top models.

4.4. Numbers vs. Narrative and the Impact of Mandatory CSRD Standards

Why do LLMs find some sustainability questions much harder than others? Table 4 and Figure 6 break down extraction accuracy across question types, historical reporting eras, and Environmental, Social, and Governance pillars.

Table 4: Breakdown of Factual Extraction Accuracy (F1) by Question Type, Reporting Era, and ESG Pillar (Three-Passage Search $k = 3$ vs. Gold-Standard Oracle Passage).

Audit Dimension	Factual Accuracy: Top-$k = 3$ Search / Gold Oracle					Avg. Gain (Δ)
	Qwen3-4B	DeepSeek-R1-7B	GPT-OSS-20B	Qwen3.5-27B	Gemma-4-31B	
Audit Question Type						
Hard Numerical KPIs ($N = 55$)	0.505 / 0.696	0.463 / 0.610	0.485 / 0.650	0.456 / 0.578	0.496 / 0.626	+0.151
Narrative Policies ($N = 65$)	0.477 / 0.662	0.408 / 0.608	0.554 / 0.800	0.531 / 0.600	0.538 / 0.669	+0.148
KPI vs. Policy Gap	*+0.028 / +0.034*	*+0.055 / +0.002*	*-0.069 / -0.150*	*-0.075 / -0.022*	*-0.042 / -0.043*	N/A
Regulatory Reporting Era						
2010–2024 Pre-CSRD ($N = 84$)	0.533 / 0.643	0.433 / 0.579	0.517 / 0.706	0.508 / 0.589	0.528 / 0.638	+0.129
2025–2026 Mandatory CSRD ($N = 36$)	0.389 / 0.758	0.432 / 0.679	0.535 / 0.791	0.470 / 0.592	0.497 / 0.676	+0.235
CSRD vs. Pre-CSRD Shift	*-0.144 / +0.115*	*-0.001 / +0.100*	*+0.018 / +0.085*	*-0.038 / +0.003*	*-0.031 / +0.038*	N/A
ESG Pillar						
Environmental (E, $N = 42$)	0.586 / 0.756	0.485 / 0.655	0.551 / 0.760	0.532 / 0.632	0.550 / 0.697	+0.165
Social (S, $N = 43$)	0.422 / 0.642	0.407 / 0.586	0.507 / 0.718	0.474 / 0.555	0.509 / 0.613	+0.161
Governance (G, $N = 35$)	0.457 / 0.626	0.401 / 0.578	0.503 / 0.713	0.480 / 0.581	0.490 / 0.634	+0.166

Finding 3 (The Gap Between Policy Summaries and Hard Numbers). *Across the largest models, summarizing narrative ESG policies is substantially easier than extracting hard numerical KPIs: under gold-standard Oracle passages,* ***GPT-OSS-20B*** *scores 0.800 on narrative policies versus 0.650 on numerical KPIs (a 15.0-point gap), and across the full benchmark, narrative policy accuracy (0.729 peak) outpaces numerical KPI extraction (0.335 baseline $k = 1$) by nearly 40 points.*

Why Numerical KPIs Trip Up LLMs. Corporate policies, such as anti-bribery commitments or board diversity principles, are written in standard prose paragraphs, which language models handle naturally. Quantitative KPIs, by contrast, live inside multi-column tables where the metric name is at the left, the unit (tCO_2e, m^3, MWh, or %) is in a sub-header, the reporting year is at the top

of the page, and scope exclusions are buried in a footnote three paragraphs below. When a 1,000-character passage slices through a table, small models like **Qwen3-4B** do well on simple single-line figures (0.505 in $k = 3$), whereas larger models like **GPT-OSS-20B** (0.800 Oracle) and **Gemma-4-31B** (0.669 Oracle) excel at synthesizing multi-sentence policy disclosures.

How Mandatory CSRD/ESRS Rules Help Model Accuracy. Comparing older voluntary reports (2010–2024) against the new wave of mandatory 2025–2026 CSRD/ESRS filings reveals a two-part story (Figure 6b). During the initial search step ($k = 3$), CSRD reports are slightly harder to search because they are much longer and pack hundreds of standardized disclosure datapoints into a single filing. However, **once the right passage is located (Oracle), extraction accuracy is dramatically higher on 2025–2026 CSRD reports** than on older voluntary reports: accuracy jumps by +11.5 points on **Qwen3-4B** (from 0.643 to 0.758), +10.0 points on **DeepSeek-R1-7B** (0.579 to 0.679), and +8.5 points on **GPT-OSS-20B** (0.706 to 0.791). Because ESRS standards enforce uniform indicator labels (such as *ESRS E1-6 Gross Scope 1, 2, 3 GHG emissions*) and consistent table layouts, LLMs can read and verify figures with much less ambiguity once pointed to the right section.

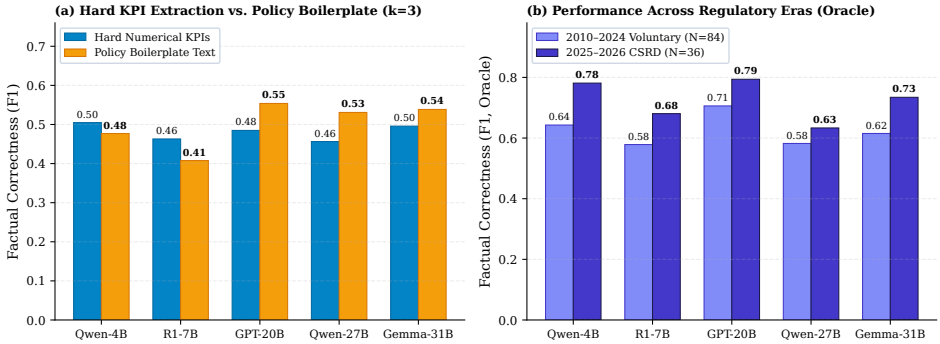


Figure 6: Granular accuracy breakdowns. (a) Comparison between Hard Numerical KPIs and Narrative Policy text under three-passage search ($k = 3$). (b) Shift across regulatory eras under gold-standard Oracle passages: mandatory 2025–2026 CSRD/ESRS filings yield much higher extraction accuracy (+8.5 to +11.5 points on top models) thanks to standardized ESRS table structures.

4.5. Performance Across E, S, and G Pillars & The Accuracy and Safety Trade-Off

Not all three pillars of ESG are equally standardized. Figure 7(a) compares extraction accuracy across Environmental, Social, and Governance questions under $k = 3$ search, while Figure 7(b) maps the practical trade-off between a model's peak extraction accuracy (Oracle F1) and its safety on unanswerable questions (strict abstention rate).

Across every single model in our study, **Environmental disclosures are extracted most accurately** (averaging 0.541 in $k = 3$ and up to 0.760 in Oracle on **GPT-OSS-20B**), outperforming Governance (0.479 average) and Social disclosures (0.466 average). The reason is standardization: environmental accounting has converged over two decades around the GHG Protocol (tCO_2e across Scopes 1, 2, and 3) and standard CDP water/waste units. Social disclosures, by contrast, suffer from inconsistent definitions across companies, where some firms report employee turnover or injury rates per 100 full-time equivalents, others per 200,000 or 1,000,000 hours worked, and many separate permanent staff from contractors in incompatible ways.

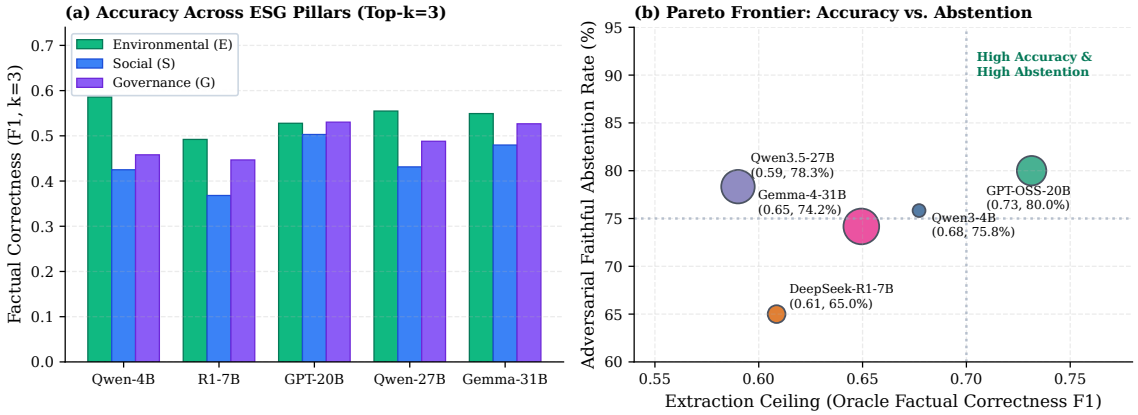


Figure 7: (a) Extraction accuracy across Environmental, Social, and Governance pillars ($k = 3$). Environmental metrics achieve the highest accuracy across all five models due to standardized GHG Protocol units. (b) Practical decision frontier comparing extraction ceiling (Oracle F1) against safety on unanswerable questions (Near-Miss Strict Abstention %), where bubble size reflects model parameter count.

4.6. Real Audit Examples: How Models Succeed and Fail

To show what these numbers look like in practice, Table 5 presents side-by-side verbatim answers from the tested models on two real audit tasks: a straightforward 2025 CSRD Scope 1 emissions lookup (Case 1) and an unanswerable near-miss trap where the search engine returns a 2022 water-recycling passage in response to a question about 2024 hazardous waste (Case 2).

In Case 1, when the passage clearly states *"Total Scope 1 emissions were 142,500 metric tons CO_2e in 2025,"* all models extract the exact figure without difficulty. Case 2 exposes the real compliance risk. Under a standard prompt, **Qwen3-4B** invents a completely fictional number (*"18,400 metric tons of hazardous waste through closed-loop filtration"*) that sounds plausible to a busy analyst but does not exist in the report. **Gemma-4-31B** avoids outright invention under the standard prompt, but falls into a subtle substitution trap, reporting the 2022 wastewater figure (*"12,500 cubic meters"*) instead of answering the 2024 hazardous waste question. Only **DeepSeek-R1-7B** pauses inside its $<$think$>$ trace to note that the excerpt covers 2022 water rather than 2024 waste, refusing to guess. Once we apply the strict compliance guardrail prompt, however, **Gemma-4-31B** and **GPT-OSS-20B** match that disciplined abstention.

4.7. Practical Guide: Strengths and Weaknesses of Each Model

For sustainability officers, financial analysts, and audit teams choosing an open-weight model, each architecture offers distinct strengths:

GPT-OSS-20B (Best All-Around Auditor). OpenAI's open Mixture-of-Experts model delivered the strongest combination of extraction accuracy and compliance safety in our benchmark. It ranked first in factual accuracy under single-passage search (0.474), three-passage search (0.522), and gold-standard Oracle passages (0.731), while also achieving the highest strict abstention rate (80.0%) and the lowest residual hallucination rate (5.8%). Because its sparse design activates only 3.6 billion parameters per word, it runs nearly as fast as a small 4B model while outperforming dense models up to 31B in size.

Table 5: Side-by-Side Audit Examples: Verbatim Model Responses on a Verified Numerical KPI Question vs. an Unanswerable Near-Miss Trap.

Audit Scenario	Question, Report Excerpt, and Verbatim Model Responses
Case 1: Hard KPI Extraction *Source:* 2025 CSRD Filing	**Audit Question:** *"What was the company's total Scope 1 direct GHG emissions in metric tons for 2025?"* **Verified Report Answer:** *"Total Scope 1 emissions were 142,500 metric tons CO_2e in 2025."* **Qwen3-4B**: "The company reported Scope 1 direct emissions of approximately 142,500 tons CO2e in 2025." → **Accurate Extraction (1.0)** **DeepSeek-R1-7B**: "<think>...target is Scope 1 2025 emissions...value is 142,500...</think> In 2025, Scope 1 GHG emissions were 142,500 MT CO2e." → **Accurate Extraction (1.0)** **GPT-OSS-20B**: "The total Scope 1 direct greenhouse gas emissions for 2025 were 142,500 metric tons of CO2e." → **Accurate Extraction (1.0)**
Case 2: Near-Miss Audit Trap *Source:* 2022 Report Excerpt (Wrong Year & Metric)	**Audit Question:** *"What was the total volume of hazardous waste recycled by the chemical division in 2024?"* **Report Reality:** *[Unanswerable: Excerpt covers 2022 water recycling, not 2024 hazardous waste]* **Required Auditor Action: Faithful Abstention** (*"I cannot answer based on the provided documents."*) **Qwen3-4B (Standard Prompt):** "In 2024, the chemical division recycled 18,400 metric tons of hazardous waste through closed-loop filtration." → **Fabricated Figure / Hallucination (0.0)** **Gemma-4-31B (Standard Prompt):** "Based on the provided documents, the company recycled approximately 12,500 cubic meters of wastewater in 2022." → **Wrong Metric & Year Substituted (0.2)** **DeepSeek-R1-7B (Standard Prompt):** "<think>The query asks for 2024 hazardous waste in the chemical division. The document only discusses 2022 water recycling. The required KPI is missing.</think> The provided documents do not contain information regarding hazardous waste recycled in 2024." → **Faithful Abstention (1.0)** **Gemma-4-31B (Strict Guardrail Prompt):** "The provided documents do not contain information about the hazardous waste recycled in 2024." → **Faithful Abstention (1.0)**

DeepSeek-R1-7B (Best Built-In Verification & Audit Trail). When given a standard, unguarded prompt, **DeepSeek-R1-7B** was by far the safest model on unanswerable questions, abstaining 77.5% of the time on its own. Even more valuable for statutory assurance, its visible <think>...</think> reasoning log shows the human reviewer *how* it checked the company name, reporting year, and units before answering. For risk-averse compliance workflows where an auditable reasoning trail matters more than raw speed, **DeepSeek-R1-7B** is an exceptional choice.

Gemma-4-31B (Strong Multi-Passage Synthesis, Needs Strict Prompts). Google's 31B model performed very well when combining information across three passages ($k = 3$ accuracy of 0.519, second overall, and 0.496 on hard KPIs). However, under standard prompting on missing data, it tended to be overly helpful, substituting related metrics from earlier years in 21.7% of unanswerable cases. Pairing **Gemma-4-31B** with our strict guardrail prompt raised its abstention rate to 74.2% and cut hallucinations to 7.5%.

Qwen3.5-27B (Conservative, High-Faithfulness Reader). **Qwen3.5-27B** achieved the highest faithfulness scores across standard search (0.912 in $k = 1$, 0.829 in $k = 3$, and 0.971 in Oracle) and reached 78.3% abstention under strict guardrails. Its main limitation is caution: when a single retrieved excerpt looks incomplete, it prefers to abstain rather than piece together partial figures, resulting in slightly lower recall on numerical tables (0.578 Oracle KPI F1).

Qwen3-4B (Capable Low-Cost Baseline for Local Hardware). Despite having just 4 billion parameters, **Qwen3-4B** punched well above its weight on straightforward lookups, scoring 0.490 in $k = 3$ and 0.677 in Oracle (including 0.696 on Oracle numerical KPIs). However, under standard prompts on near-miss traps, it fabricated figures in 24.2% of cases. Teams running **Qwen3-4B** on local laptops or edge servers must pair it with strict abstention prompts (which lift its abstention to 75.8%) and automated number-matching checks.

4.8. Three Recommendations for Sustainability & Assurance Teams

Our findings translate into three practical rules for organizations deploying LLMs on ESG filings:

1. **Never Use Default "Helpful Assistant" Prompts for Compliance Auditing:** Standard prompts encourage models to guess or substitute prior-year numbers when a metric is missing. Always enforce an explicit abstention guardrail (Listing 2) or use a reasoning-verified model such as **DeepSeek-R1-7B**.
2. **Retrieve Multiple Passages ($k = 3$) and Preserve Table Headers:** Because ESG tables span multiple paragraphs and footnotes, single-passage search ($k = 1$) misses critical context. Expanding search to $k = 3$ passages improves factual extraction across every tested model by +6 to +11 points.
3. **Keep a Human-in-the-Loop for Hard Numerical KPIs:** While open-weight LLMs can draft accurate summaries of qualitative ESG policies (up to 0.800 F1), even the best model (**GPT-OSS-20B**) misses roughly 27% of nuances on complex numerical tables under ideal conditions. Automated pipelines should highlight the source table in the PDF so a human auditor can verify final numbers before sign-off.

5. Limitations

Geographic and Language Scope. Our 518-report corpus focuses on English-language disclosures published by companies listed on European stock exchanges. While these multinational firms report under global frameworks (GRI, TCFD, SASB, and CSRD), our findings should be tested further on non-European filings, such as US SEC climate disclosures or Indian BRSR reports, and on reports published in local languages.

Text-Based PDF Extraction vs. Scanned Infographics. Our document pipeline extracts text and tables from digital PDF layers using PyMuPDF. When companies embed carbon footprints or organizational charts purely as flat image graphics without underlying text, text-only extractors cannot read them. Combining text LLMs with vision-language models that inspect page images directly is an important next step.

Benchmark Scale vs. Multi-Document Cross-Checking. Although all 120 audit questions and reference answers are grounded in real corporate filings, each question targets a specific report. In practice, senior assurance providers also perform cross-document audits, such as checking whether a company's 2025 CSRD report quietly restated the 2023 baseline emissions it published two years earlier.

Fixed Search Engine Baseline. Because our goal was to compare the reading and verification abilities of the five LLMs on an equal footing, we used a single vector search index (**Qwen3-Embed-4B**). Production audit systems can pair these LLMs with hybrid keyword-plus-vector search and dedicated table rerankers to push initial search recall higher.

Automated Referee Nuances. Using Microsoft's **Phi-4-14B** as an external referee eliminates self-grading bias across the tested model families, and our deterministic settings ($\text{temperature} = 0.0$) ensure repeatable scores. Even so, automated LLM referees can occasionally penalize valid paraphrases of refusal statements, which we mitigate by pairing judge scores with rule-based numerical verification.

6. Conclusion and Future Work

As sustainability reporting moves from voluntary corporate storytelling to mandatory statutory assurance under frameworks like the EU CSRD, the tools used to audit those disclosures must be held to rigorous empirical standards. By evaluating five open-weight Large Language Models across 518 corporate ESG reports (2010–2026) and 3,000 controlled audit completions graded by an independent referee (**Phi-4-14B**), this study maps where open-weight AI is ready for compliance work and where human oversight remains essential.

Our results show a clear divide between narrative comprehension and quantitative precision: open-weight LLMs summarize qualitative ESG policies reliably (reaching 0.800 F1 under gold passages), but face a persistent ceiling on hard numerical KPIs (peaking at 0.731 overall F1 on **GPT-OSS-20B**) due to complex table layouts and unit conversions. Encouragingly, the standardized reporting tables introduced by 2025–2026 CSRD/ESRS filings boost model extraction accuracy by +8.5 to +11.5 points once the relevant section is retrieved. Most importantly for assurance teams, we show that hallucination on missing disclosures is preventable: pairing open-weight models like **GPT-OSS-20B** with strict abstention prompts, or deploying reasoning models like **DeepSeek-R1-7B** that verify dates and entities step by step, reduces unfaithful fabrication below 6% and provides a practical foundation for trustworthy AI-assisted sustainability auditing.

Future Work. Three extensions will build on this benchmark: (1) integrating multimodal vision-language models to read graphical charts and scanned tables directly from PDF page images; (2) training domain-specific verification models that cross-check physical units and scope boundaries before emitting numerical figures; and (3) expanding the benchmark to multi-year restatement audits that track whether corporations alter historical baseline figures across consecutive annual filings.

Data and Code Availability

To support independent research and future sustainability benchmarking, all evaluation scripts, benchmark generation code, prompt templates, and aggregated model outputs are publicly available at https://github.com/Sredeemer/ESG-RAG-benchmark-ragas.

Acknowledgments

Figure 1 was created with BioRender (https://biorender.com). Compute infrastructure was supported by the Authors' Research Initiative.

Competing Interests

The authors declare no competing financial or personal interests.

References

[1] Florian Berg, Julian F Kölbel, and Roberto Rigobon. Aggregate confusion: The divergence of ESG ratings. *Review of Finance*, 26(6):1315–1344, 2022.

[2] Mattia Birti, Francesco Osborne, and Andrea Maurino. Optimizing large language models for ESG activity detection in financial texts. In *Proceedings of the 6th ACM International Conference on AI in Finance*, 2025.

[3] Carbon Disclosure Project. CDP: Turning transparency into action. *CDP Worldwide*, 2025. URL https://www.cdp.net/.

[4] Mark Carney. Breaking the tragedy of the horizon: Climate change and financial stability. Speech at Lloyd's of London, Bank of England, 2015.

[5] Confident AI. DeepEval: The open-source llm evaluation framework. https://deepeval.com/, 2026.

[6] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, 2024.

[7] European Commission. Commission delegated regulation (EU) 2023/2772 supplementing directive 2013/34/EU as regards sustainability reporting standards (ESRS). *Official Journal of the European Union, L Series*, 2023.

[8] European Union. Directive (EU) 2022/2464 of the European Parliament and of the Council as regards corporate sustainability reporting. *Official Journal of the European Union, L 322*, pages 15–80, 2022.

[9] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

[10] Global Reporting Initiative. GRI standards: Universal, topic and sector standards. *Global Sustainability Standards Board*, 2021. URL https://www.globalreporting.org/.

[11] Yuxian Gu, Xiaochuang Han, Zheng Li, et al. A survey on LLM-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

[12] Chaoyue He et al. ESGenius: Benchmarking LLMs on environmental, social, and governance (ESG) and sustainability knowledge. *arXiv preprint arXiv:2506.01646*, 2025.

[13] Seungone Kim et al. Prometheus: Inducing fine-grained evaluation capability in language models. *arXiv preprint arXiv:2310.08491*, 2024.

[14] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[15] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, 2023.

[16] Rafik Mankour et al. Climate finance bench: A benchmark for RAG over corporate climate disclosures. *arXiv preprint arXiv:2505.22752*, 2025.

[17] Augusto Londero Orsolin et al. Analysis of indicators for the integration, implementation, and development of the ESG report in ISE/B3 companies. *Revista Catarinense da Ciência Contábil*, 23:e3515, 2024.

[18] Teodora Maria Rusu et al. Sustainability performance reporting: Evolution, challenges, and corporate trends. *Sustainability*, 16(19):8452, 2024.

[19] Motaz Saad, Anna Borrelli, Ivan Gentile, Kianna Kazemi, Francesco Piccialli, and Antonella Longo. Empirical evaluation of open-source large language models for retrieval-augmented generation in esg domain. *arXiv preprint arXiv:2609.15242*, 2026.

[20] Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 338–354, 2024.

[21] Sustainability Accounting Standards Board. SASB conceptual framework. *Sustainability Accounting Standards Board*, 2017.

[22] Task Force on Climate-related Financial Disclosures. Recommendations of the task force on climate-related financial disclosures. *Financial Stability Board*, 2017.

[23] Qilong Wu et al. SusGen-GPT: A data-centric LLM for financial NLP and sustainability report generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1184–1203, 2025.

[24] Tsung-Yu Yang and Meng-Chi Chen. ESGLens: An LLM-based RAG framework for interactive ESG report analysis and score prediction. *arXiv preprint arXiv:2604.19779*, 2026.

[25] Lei Zhang et al. Benchmarking multimodal understanding and complex reasoning for ESG tasks. *arXiv preprint arXiv:2507.18932*, 2025.

[26] Chi Zhu et al. JudgeLM: Fine-tuned large language models are scalable judges. *International Conference on Learning Representations*, 2025.

[27] Yi Zou et al. ESGReveal: An LLM-based approach for extracting structured data from ESG reports. *Journal of Cleaner Production*, 489:144572, 2025.