Climate Finance Bench

Rafik Mankour*   Yassine Chafai*   Hamada Saleh*

Ghassen Ben Hassine*   Thibaud Barreau*   Peter Tankov†

Abstract

Climate Finance Bench introduces an open benchmark that targets question–answering over corporate climate disclosures using Large Language Models. We curate 33 recent sustainability reports in English drawn from companies across all 11 GICS sectors and annotate 330 expert-validated question–answer pairs that span pure extraction, numerical reasoning, and logical reasoning. Building on this dataset, we propose a comparison of RAG (retrieval-augmented generation) approaches. We show that the retriever's ability to locate passages that actually contain the answer is the chief performance bottleneck. We further argue for transparent carbon reporting in AI-for-climate applications, highlighting advantages of techniques such as Weight Quantization.

Keywords: climate finance; ESG disclosure; retrieval-augmented generation; large language models; question answering; benchmark dataset; sustainability reporting; carbon footprint; quantization; information retrieval

1 Introduction

Climate finance aims to measure and control climate-related financial risks, to redirect financial flows towards the green sectors of the economy, and to provide incentives for the brown sectors to reduce their carbon footprint. All this requires reliable measurement of corporate climate risk exposures and environmental impacts. Yet regulators and investors frequently cite a *climate-data gap*: key indicators such as Scope 1–3 emissions, decarbonization

---

*Institut Louis Bachelier

†CREST, ENSAE, Institut Polytechnique de Paris and Institut Louis Bachelier

targets or capital-expenditure plans are either missing, inconsistently reported or locked in unstructured PDF filings [1].

Climate-related financial disclosure has moved from a voluntary narrative exercise to a quasi-regulatory requirement. Nineteen jurisdictions, including the EU, the United Kingdom and Japan, have either adopted or are piloting mandatory reporting regimes aligned with the *Task Force on Climate-related Financial Disclosures* (TCFD) or the upcoming *IFRS S2* and *European CSRD* standards, covering more than 60% of global GDP and most major capital markets [2].

Credible transition plans, scenario-aligned metrics and financed-emission inventories are now essential inputs to prudential stress tests and capital allocation decisions [3, 4]. As a result, large corporates publish hundreds of pages of sustainability information, often in PDF format, replete with tables, embedded figures and forward-looking statements. Central banks warn that robust, machine-readable climate data are now prerequisite for prudential stress testing and macro-financial stability.

Analysts aiming to extract information from corporate sustainability reports face three persistent bottlenecks:

1. **Heterogeneity of formats.** As there are no common reporting standards, disclosures vary widely in structure, terminology and granularity, hindering automated parsing and cross-company comparison.

2. **Information overload.** Climate or ESG (Environmental, Social and Governance) reports can now exceed hundreds of pages and relevant data are scattered among other information, manual review at portfolio scale is no longer feasible.

3. **Data quality and auditability.** Stakeholders require exact numeric values (e.g. Scope 1 + 2 GHG emissions) and transparent provenance to guard against green-washing.

Large Language Models (LLMs) such as GPT-3 [5] and GPT-4 [6] can read long documents, extract key metrics and draft natural-language explanations. However, vanilla LLMs suffer from two well-known limitations when applied to climate-finance documents: limited context windows that truncate relevant evidence, and a tendency to hallucinate plausible-sounding but unsupported figures [7]. *Retrieval-Augmented Generation* (RAG) addresses both issues by (1) retrieving the most relevant text passages from the source documents and (2) conditioning the LLM on those passages so that answers remain grounded in the original evidence [10, 11].

In this work we systematically evaluate RAG pipelines for climate-finance question answering, making the following contributions:

- We establish a benchmark dataset comprising 33 climate disclosures, 330 analyst-curated questions and expert-validated reference answers that span data extraction, numerical reasoning and logical inference.

- We compare multiple RAG configurations: minimal dense retrieval, hybrid dense + BM25 and reranking, across several LLM back-ends, quantization levels and prompt strategies.

- We provide an open, reproducible test-bed that reports both answer accuracy and the *carbon footprint* of each configuration, so that future research can optimize statistical as well as environmental efficiency.

We assemble a representative sample of reports across all 11 GICS sectors. By emphasizing factual correctness, including exact numbers and units, and by measuring retrieval coverage, we aim to help both researchers and practitioners develop trustworthy, resource-efficient question-answering systems for the climate-finance domain.

2   Literature Review

The rapid commercial rollout of large language models has brought an equally rapid migration toward *retrieval-augmented generation*. In less than two years, RAG has moved from a research curiosity to the backbone of production systems such as Bing Chat, Perplexity.ai and the many domain-specific applications built with LangChain or LlamaIndex. Because it *grounds* outputs in verifiable documents and avoids expensive model fine-tuning, RAG is now the de-facto recipe for deploying LLMs in high-stakes settings ranging from medical question answering to legal contract analysis [8, 9]. The climate-finance domain is no exception: analysts need fact-checked answers, regulators demand provenance, and datasets change too frequently for parametric retraining alone. Against this backdrop, we summarize below the core architectural choices in modern RAG pipelines and how they inform our benchmark design.

Retrieval-Augmented Generation (RAG).   Early RAG systems [10] combine a dense bi-encoder retriever with a seq2seq generator, injecting the top-k retrieved passages into the generation context so that answers remain grounded in source documents. Subsequent work shows that feeding dozens of passages and letting the decoder fuse evidence (the FiD architecture) yields large gains on knowledge-intensive tasks [11]. Best-practice surveys

now recommend a *hybrid pipeline* (dense embeddings + BM25) followed by a cross-encoder reranker for precision, plus prompt instructions that discourage hallucination [15]. In short:

- dense retrieval captures semantic paraphrases,

- sparse lexical search excels at exact figures and units, and

- reranking re-scores candidates jointly with the query to identify the truly relevant ones.

The generator is then tasked to quote or cite its supporting text, a practice that empirically reduces unsupported claims and eases human verification.

Open-source efforts such as FinGPT [12] and ESG-BERT [13] also show promise in adapting general-purpose LLMs to domain-specific tasks like ESG disclosure classification and financial forecasting. We believe that climate-finance QA tasks would benefit from similar fine-tuning initiatives.

The FinanceBench paper [14] introduced a finance-specific benchmark to test LLMs in a question-answering setting. Their dataset covered a selection of 10 000 questions across 40 publicly traded US companies from various sectors, drawn from real financial documents such as 10-K, 10-Q and 8-K filings, and they did a human evaluation over a subset of 150 evaluation questions. Although FinanceBench aimed to provide a broad finance-oriented benchmark, its reported performance evaluations were done primarily via human annotations, with a relatively permissive notion of "correctness" (e.g., minor deviations in units were not treated as fully incorrect) to ensure a good-faith understanding of the capabilities of the models.

Moreover, FinanceBench explored several LLM configurations, including "open book," "closed book," "retrieval," and "long context." Among these, only the "retrieval" setting fully aligns with RAG-based question answering. Inspired by the FinanceBench methodology, we focus here on climate-related disclosures, adopting and extending the retrieval pipeline to better handle complex disclosures such as tables and figures. We also incorporate insights from best-practice studies on retrieval-augmented generation [15] to refine our approach.

In addition to FinanceBench [14], several other benchmarks focus on numerical and reasoning-based question answering from financial documents. FinQA [17], ConvFinQA [18] and TAT-QA [19] emphasize numerical reasoning and hybrid table-text data. These datasets are relevant for evaluating climate-finance models, especially those requiring computation over reported KPIs.

Within the ESG and climate disclosure space, Climate-FEVER [20] and ClimRetrieve [16]

provide valuable baselines for factual verification and information retrieval respectively. While Climate-FEVER focuses on verifying real-world claims, ClimRetrieve offers climate-specific documents and retrieval annotations.

More recently, the Golden Touchstone benchmark [21] provided a bilingual and comprehensive evaluation framework for finance-specific LLMs and BloombergGPT [25] demonstrated how domain pretraining boosts performance in financial QA.

Limitations of current methodologies and proposed improvements

Dataset scope. Current methodologies evaluate sophisticated reasoning but often provide pre-extracted snippets for each question, leaving the retrieval challenge unsolved. Full-document retrieval often covers only U.S. SEC filings or a limited set of sustainability reports.

Retrieval transparency. Few benchmarks disclose the exact retrieval pipeline used in baselines; hybrid search, cross-encoder reranking and document preprocessing are rarely benchmarked side-by-side, hindering reproducibility and progress measurement.

Evaluation methodology. Some benchmarks label an answer as correct only when the exact text string matches the reference answer. The challenge is to avoid letting hallucinations slip through without penalizing semantically correct answers written with different wording, units or abbreviations. A fairer scheme should (i) award partial credit to answers that cover only part of the required information while taking into account their incompleteness and (ii) scale to large datasets. We therefore adopt automated grading with an LLM-as-a-Judge, while keeping a human-in-the-loop design.

Our contribution. **Climate Finance Bench** fills these gaps by:

1. using complete ESG and climate reports (33 documents across all 11 GICS sectors), thereby testing end-to-end retrieval on long, heterogeneous PDFs;

2. releasing a reference *hybrid + reranking* pipeline so that future work has a documented, strong baseline to iterate on;

3. enforcing a non-binary, 3-point grading scheme based on LLM-as-a-Judge;

4. establishing estimates for the carbon emissions associated to the tools we experiment on.

By unifying retrieval, reasoning and environmental accountability, our benchmark aims for trustworthy, resource-efficient question answering in the climate-finance domain.

3   Methods

3.1   Data Collection and Curation

3.1.1   Selection of Reports.

We gathered 33 climate reports from large publicly traded companies spanning multiple regions (e.g., CAC40 and DAX40 in continental Europe, FTSE in the UK, S&P500 in the US) and covering all 11 GICS sectors. We ensured:

- At least one company from Communication Services, Real Estate and Health Care sectors.

- At least two companies from each of the 8 remaining sectors, preferably from different sub-sectors.

- Exactly one recent climate or sustainability report per company, capturing the latest relevant fiscal year.

The full list of selected companies is available in Appendix C.

3.1.2   Question Formulation and Annotation.

We relied on ESG experts to provide 10 questions per report. The questions reflect two modalities (metric-related and domain-related) and three categories:

1. **Pure Extraction**: directly retrieve facts.
   Example: Has the company identified significant decarbonization levers? If yes, detail them.

2. **Numerical Reasoning**: extract figures and/or perform calculations.
   Example: What is the company's carbon intensity (in tCO_2/million USD) for FY 2023? If not reported, compute it by dividing total carbon emissions by that year's revenue.

3. **Logical Reasoning**: combine multiple data points to infer an answer.
   Example: Does the company have a decarbonization trajectory compatible with a 1.5 °C or 2 °C scenario?

Seven analysts carried out the following steps:

- Carefully read the assigned climate report(s).

- Provide written, reference answers (the "Gold Standard") for each of the 10 questions, along with document excerpts, page numbers and an indication of whether the relevant information was found in text, a table, or a figure.

- Follow a unified annotation guide to ensure consistent handling of numerical values, units and references. An adapted version of this guide is available in Appendix D for reference.

Two ESG domain experts resolved ambiguous cases and a final quality-control check was performed to confirm the validity of all annotations. Ultimately, we obtained 330 question–answer pairs.

3.1.3 Resulting Dataset Structure.

We store our dataset in a table of 330 rows and 13 columns. Each row corresponds to a single question about a specific company's report, containing:

- `Company name` and `Fiscal year`.

- `Question ID`, `Question text`, `Type of question`.

- `Answer` (reference/Gold Standard).

- `Documents`, `Pages`, `Document extracts`, `Extract type`.

3.2 Hardware and Environment

We conducted the experiments primarily on a GPU environment available through Kaggle (Notebooks hosted on GCP). This environment provides:

- 60 GB of storage and 30 GB of RAM.

- Access to a GPU P100 (16 GB memory).

For LLMs with publicly available APIs (Claude Sonnet and GPT-4o), as well as Qwen2.5 and DeepSeek R1 via Nebius' API, calls were invoked directly from the Kaggle notebook.

3.3 Data Extraction and Chunking

Our RAG pipeline builds on the foundational architecture introduced by Lewis et al. [10], while incorporating best practices from the FiD architecture [11] and Atlas model [22]. These methods highlight the importance of fusing retrieved passages in the decoder and optimizing passage selection for generation.

In designing the retrieval system, we consulted evaluations from the KILT [23] and BEIR [24] benchmarks, which stress the need for robust lexical-dense hybrid retrieval and reproducible metrics across knowledge-intensive tasks.

PDF Extraction.   We experimented with two main approaches:

- **LangChain loaders** that use the Unstructured library, providing a quick way to parse text.

- **Docling**, which converts PDF files to HTML/Markdown to preserve more structure (e.g., tables, figure captions).

Although Docling can better retain table formatting, it sometimes introduces noise such as HTML tags, which can complicate retrieval.

Chunking Strategy.   To avoid having either overly large text chunks or fragments broken mid-table, we used:

- A base chunk size of 2048 tokens, with an overlap of 204 tokens (i.e., 10%).

- Logic to avoid breaking tables and to merge small paragraphs with relevant headings.

3.4 Vector Indexing and Retrieval

We vectorized all chunks using `sentence-transformers/all-mpnet-base-v2` and stored them in a FAISS index for nearest-neighbor retrieval. We explored two RAG retrieval configurations:

- **Minimal**: return the top $k = 12$ chunks based solely on cosine similarity scores with the question embedding.

- **Best Practices**: a "hybrid" approach combining:

  1. **Semantic and lexical retrieval**: weighted combination of top results from a dense (semantic) retriever (75%) and BM25 (25%).
  2. **Fusion and Reranking**: we first fuse the top 20 chunks using Reciprocal Rank Fusion, select 8 best chunks, then apply a cross-encoder reranker to the next 12 to pick 4 additional relevant chunks, for a total of 12.

3.5 LLMs for Generation

We tested five LLMs under both minimal and hybrid retrieval:

- **Llama3.1 8B Instruct**[1] and **Llama3.1 8B Instruct quantized in 4-bit**.

- **Mistral Nemo Instruct 12B**[2] **quantized in 4-bit**.

- **Claude Sonnet 3.5 2024-06-20**[3].

- **GPT-4o**[4].

Two more LLMs were tested under hybrid retrieval only:

- **Qwen2.5-72B**[5].

- **DeepSeek R1**[6].

We set the temperature to 0.2 (for more deterministic answers) and limited the maximum output to 512 tokens to avoid overly long responses. All queries shared the same two–part prompt shown below, with a system prompt instructing the LLM to answer strictly based on the retrieved chunks, in order to refrain from hallucinating data.

Curly-brace placeholders ({company}, {context}, {question}) are filled at run-time:

---

[1]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[2]https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407
[3]https://www.anthropic.com/index/claude-3-family
[4]https://openai.com/index/hello-gpt-4o
[5]https://huggingface.co/Qwen/Qwen2.5-72B
[6]https://arxiv.org/abs/2501.12948

```
----- System prompt (fixed) -----
You are a documentary assistant.
Answer the question about the mentioned company based on the
provided context that was extracted from climate or sustainability
reports. Do not add any additional notes.
If the answer to the question is missing from the provided context
and you cannot conclude on it on your own, indicate this sincerely.

Here are three examples of the format to follow in your reply:
###
Human: Does the company have a climate change mitigation objective
for FY2023?
AI: Yes, the company aims to become net zero by 2030 on its Scope
1, 2 and 3 emissions.

Human: Does the company have a climate change mitigation objective
for FY2023?
AI: No, the company clarifies its intention not to pursue a
net-zero target.

Human: Does the company disclose a Transition Plan for FY2023?
If yes, highlight its main characteristics.
AI: Not available in the retrieved information.
###

----- User prompt (filled per query) -----
Here are excerpts from documents about the company {company}:
###
{context}
###
Here's the question asked by the user:
Question: <<< {question} >>>
```

To compare full-precision with compressed models under identical conditions and to fit *Mistral Nemo Instruct 12B* on the 16GB GPU available in our local environment, we performed post-training, 4-bit weight quantization with the `bitsandbytes` library (v0.43). Both *Llama3.1 8B Instruct* and *Nemo Instruct 12B* were loaded and no additional fine-tuning was applied. We evaluated the quantized checkpoints in inference-only mode, using exactly the

same decoding hyper-parameters (temperature 0.2, top-p 0.95, `max_new_tokens`=512) as for the baselines. Quantization reduced resident GPU memory from 16GB to 6GB for *Llama3.1 8B Instruct* and from 24GB to 9GB for *Nemo Instruct 12B*, enabling local execution.

4 Results

4.1 Manual Evaluation and LLM-as-a-Judge

We first performed a human evaluation of RAG outputs. Human annotators labeled each answer as either *correct*, *incomplete*, or *incorrect*, with exactness in numeric values and appropriate textual evidence being key factors. However, human evaluation is time-consuming and can be subjective.

To reduce manual overhead, we tested an LLM-based grader (*LLM-as-a-Judge*) that compares each RAG-generated answer to the reference answer, factoring in the context of the original question. We found the highest alignment with human judgments (over 80% agreement) when using Claude with the question prompt included. Thus, for the main experiments, we rely on the LLM-as-a-Judge framework for scalable evaluation.

A full description of the LLM-as-a-Judge evaluation protocol is available in Appendix A.

4.2 Comparing RAG Configurations and LLMs

4.2.1 Minimal RAG.

Under the minimal retrieval approach, smaller or quantized LLMs (Llama3 8B Instruct, Llama3 Instruct 4-bit, Nemo Instruct 4-bit) achieved 35–40% correct responses. Larger models (Claude and GPT-4o) performed significantly better, around 50–55% correct. As shown in Figure 1, GPT-4o and Claude 3.5 outperform smaller 4-bit models by roughly 15 percentage points under the minimal setting.

4.2.2 Hybrid RAG (BM25 + Reranking).

Adding BM25 lexical retrieval and reranking yielded substantial gains for larger LLMs. As illustrated in Figure 2, Claude 3.5 remains in the lead at 62%, but the performance gap has narrowed: DeepSeek R1 attains 60%, while Qwen2.5-72B reaches 44%. In other words, switching from Claude 3.5 to DeepSeek R1 incurs an $\approx$ 2pp drop, even though Claude
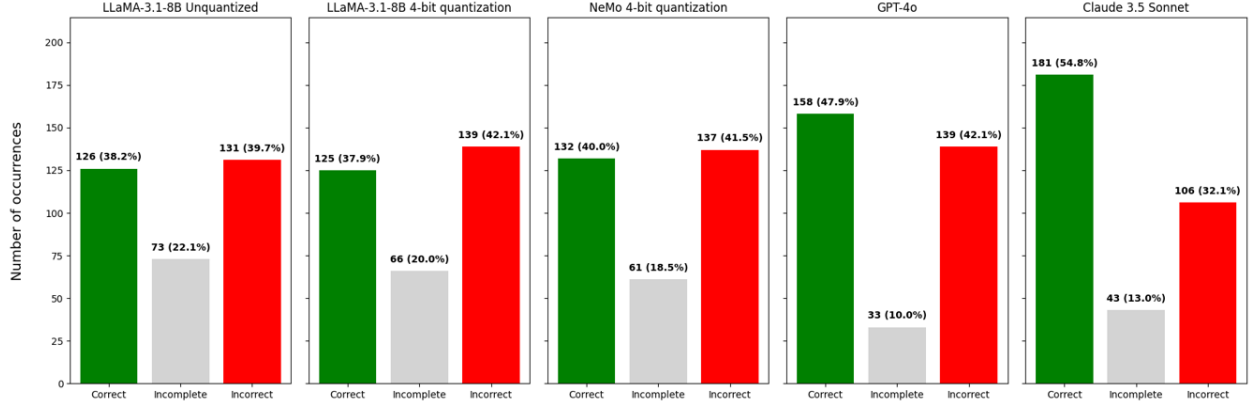
Figure 1: Accuracy breakdown (*correct, incomplete, incorrect*) for the **Minimal RAG** configuration across five LLMs.

consumed about five times less output tokens per answer on average. This suggests that retrieval quality, rather than model capacity, seems to be the current bottleneck in our experiments.
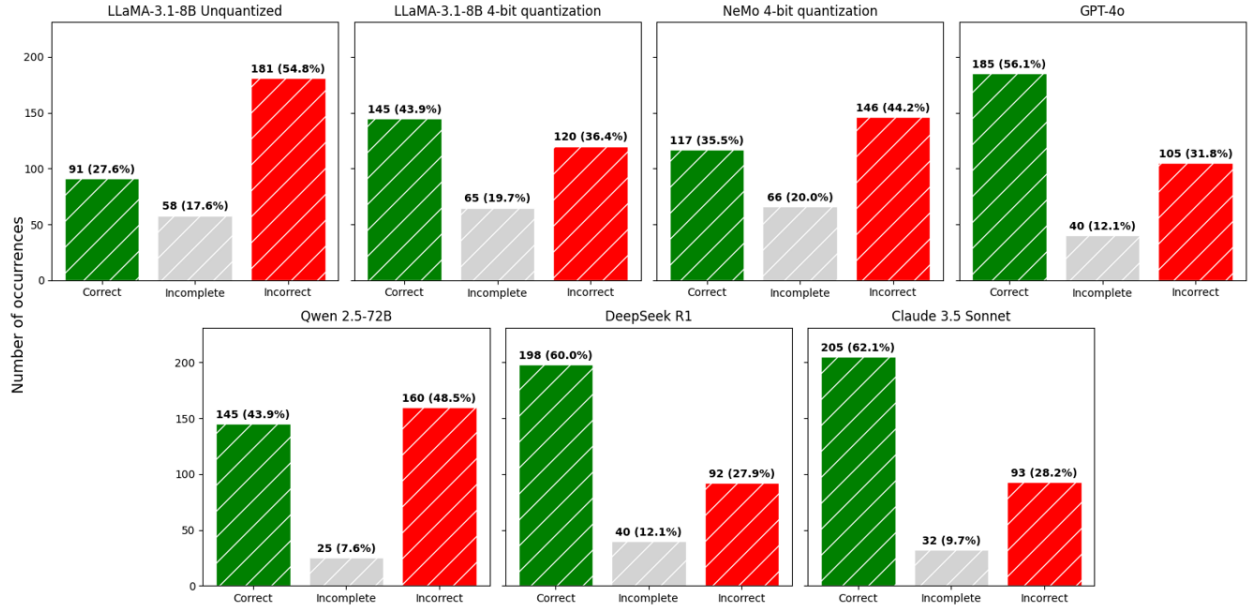

Figure 2: Accuracy breakdown (*correct, incomplete, incorrect*) for the **Hybrid RAG** configuration across the seven LLMs tested.

4.2.3 Minimal vs. Hybrid.

Hybrid retrieval helps large instruction-tuned models but can hurt smaller or highly-quantized ones. Qwen2.5-72B and DeepSeek R1 were not run under minimal retrieval, so direct deltas are unavailable. Incremental improvements on the generation side alone are unlikely to break the 65% barrier without a better retriever. Figure 3 details these jumps: BM25 adds +4.3pp, reranking another +3.0pp, while an un-filtered HTML conversion costs 4.8pp.

4.2.4 Effect of Docling Conversion.

While Docling potentially preserves the layout of tables and figures, we observed a performance decrease of several percentage points, likely due to extra HTML tokens and parsing noise degrading retrieval's performance. This suggests that more advanced post-processing might be required to capitalize on Docling's structural advantages.
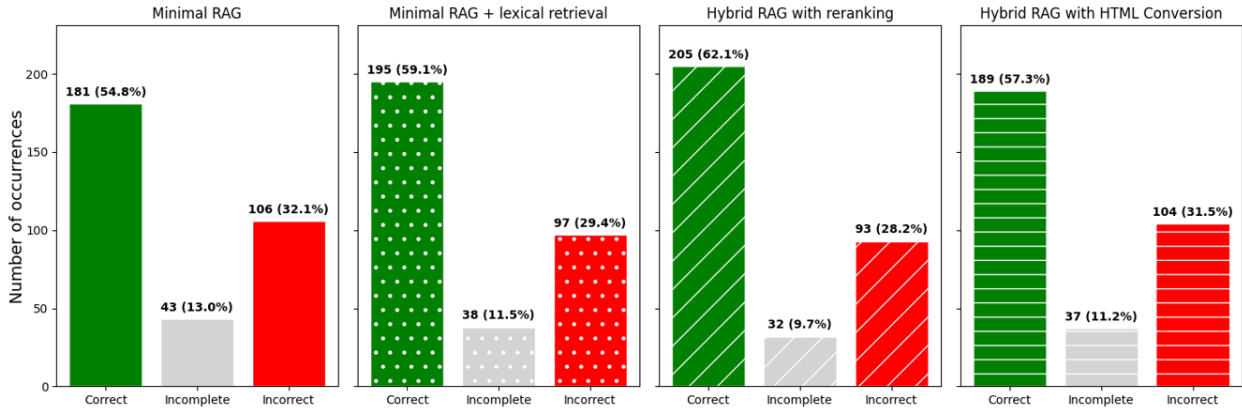

Figure 3: Stepwise impact of successive retrieval upgrades on answer quality (Minimal RAG → + BM25 lexical → + reranking → + HTML conversion). Adding BM25 improves the correct–answer rate from 54.8% to 59.1%, and the hybrid dense–sparse & reranking scheme lifts it further to 62.1%. Introducing Docling's HTML conversion without extra post-processing brings the score down to 57.3%, indicating that raw structural noise can offset earlier gains. Bars show absolute counts (annotated) and the associated share of the 330-question test set.

4.2.5 Quantized vs. Full-Precision.

For smaller-scale Llama models, we found that 4-bit quantization introduced only minor accuracy differences (within 1–2%), while significantly reducing memory usage and energy consumption. Figure 4 illustrates this comparison for LLaMA-3.1 8B, where the

unquantized and quantized versions perform similarly across all three correctness classes. In resource-constrained settings, 4-bit quantization is therefore a compelling strategy to lower computational cost and associated carbon emissions without critically hurting correctness.
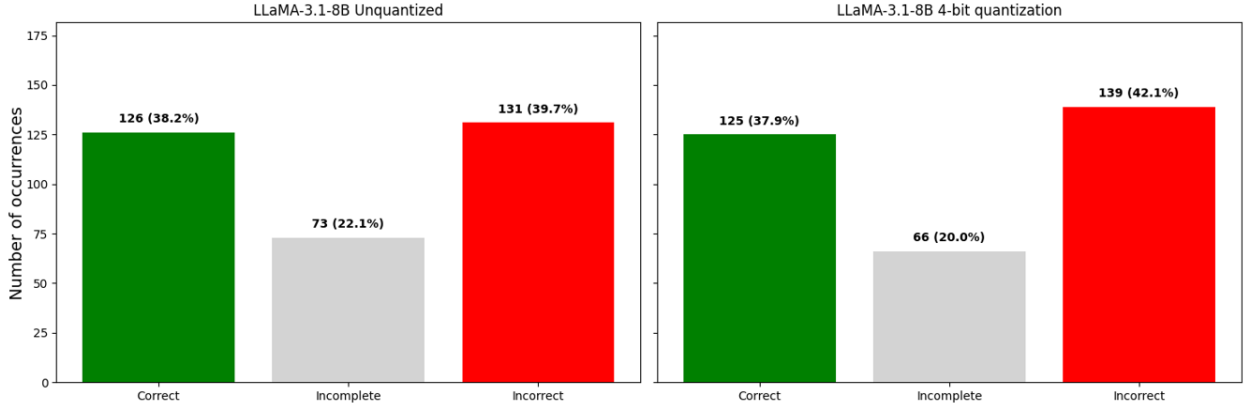

Figure 4: Comparison of **LLaMA 3.1-8B Unquantized** and **4-bit Quantized** under the minimal RAG setting. Quantization leads to negligible accuracy loss while significantly reducing resource usage.

4.3 Performance by Question Type

Our question set includes extraction, numerical reasoning and logical reasoning categories, each varying in difficulty. Figure 5 reveals patterns that are not obvious from aggregate accuracy alone.

In our best configuration (Claude 3.5 + Hybrid retrieval), numerical reasoning questions had the highest correct-answer rate, when one might expect numerical reasoning to be harder than pure extraction (69.7% vs. 65.7% correct). A qualitative error analysis indicates why: in some of the numerical questions, the relevant arithmetic had already been carried out in the source document, so the task collapses into precise retrieval plus unit normalization.

Logical questions expose a retrieval bottleneck as logical-reasoning items require chaining multiple facts scattered across a report. Broad or ambiguous queries (e.g., "Which topics have been assessed to be material?") caused more errors or incomplete answers.

Properly structured queries, with clear numeric or factual targets, were most reliably answered. Future iterations of *Climate Finance Bench* will therefore log intermediate reasoning traces to diagnose whether errors originate from retrieval omissions or reasoning failures and to build custom processes per question in order to improve the success rate of answers.
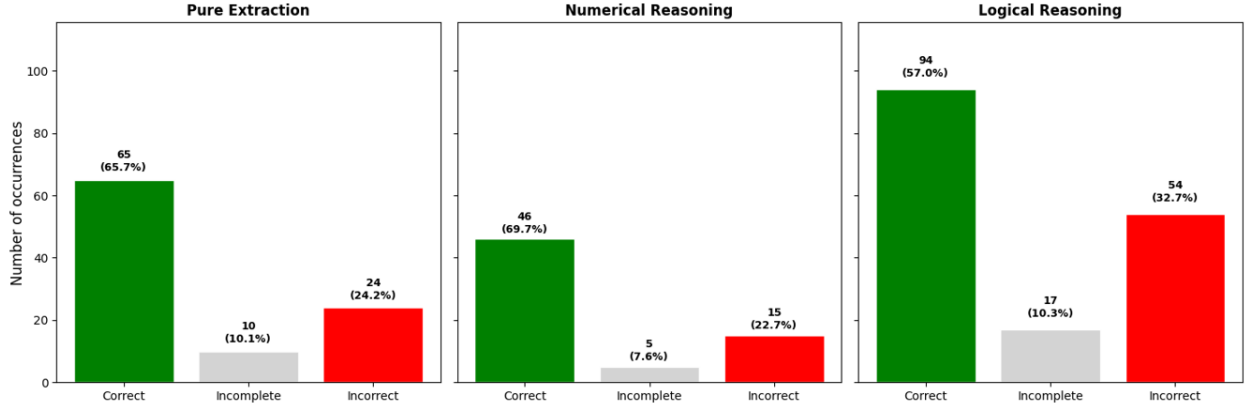
Figure 5: Break-down of answer quality for each question category under the best-performing setup (Claude 3.5 + hybrid retrieval). Numerical reasoning edges out pure extraction, while logical reasoning lags behind because it demands multi-hop synthesis across passages.

4.4 GHG Emissions and Environmental Footprint

Measuring emissions from AI usage aligns with growing interest in sustainable ML practices. Patterson et al. [26] provide methodology to estimate the carbon impact of training large neural networks, while Schwartz et al. [27] advocate for "Green AI," urging the community to prioritize energy efficiency in both training and deployment.

In keeping with the climate focus, we also conducted a rough measurement of the greenhouse gas (GHG) emissions attributed to these experiments:

- For Claude and GPT-4o, we estimated an upper bound per query using external providers like *EcoLogits*, then generalized across all runs.

- For local or Kaggle-based models, we approximated total GPU usage with *CodeCarbon* logs, dividing by the total number of queries to compute a per-query footprint.

- We could not obtain estimates for Qwen and DeepSeek as they were called through Nebius's API.

Aggregating CPU, memory and GPU energy logs, *CodeCarbon* estimates for local runs and per-query upper–bound figures from *EcoLogits* for the proprietary API calls, we estimated the GHG footprint of each model with a confidence interval due to the wide uncertainty bands published for OpenAI's and Anthropic's (Claude) back-end infrastructure.

Table 1 provides emissions per question answered for each model.

Model / configuration	Emissions (g CO_2 eq / query)	Uncertainty (g CO_2 eq / query)
LLaMA 3.1-8B (full precision)	2.79	–
LLaMA 3.1-8B (4-bit)	0.70	–
Mistral NeMo-12B (4-bit)	1.15	–
GPT-4o (API)	7.18	± 4.00
Claude 3.5 Sonnet (API)	8.15	± 4.52
Vector-store build[*]	0.30	n/a

Table 1: Average GHG emissions per query. The asterisk marks a one-off indexing cost if diluted over 330 queries, shown for scale.

Two points emerge: (i) **API calls dominate**: GPT-4o or Claude generate higher emissions than smaller models, disproportionately to the improvement in performance; (ii) **Quantitation pays off**: the 4-bit quantization of Llama3.1 8B reduced emissions by a factor of ≈ 4, yielding an emissions saving of roughly 75% with negligible accuracy loss. Retrieval and indexing remain almost negligible by comparison.

This breakdown illustrates why future work should favor lightweight local models to factor in environmental impacts, whenever accuracy allows, and report vendor-side carbon metrics more transparently if practitioners are to measure carbon impacts with tighter bounds.

A full description of the emissions estimation protocol is available in Appendix B.

5   Limitations

This first release of *Climate Finance Bench* covers only 33 sustainability reports, mostly large-capitalisation firms headquartered in Europe or North America and is therefore not representative of emerging-market issuers, small and medium enterprises, or non-English filings.

Despite a two-step review process, Gold Answers for multi-step reasoning items still contain a degree of subjectivity, which may propagate noise into model-versus-human comparisons.

Because climate disclosures are largely self-reported, any inaccuracy or greenwashing in the underlying documents can bias both retrieval and evaluation.

Sector-wide queries based on information spanning multiple companies have not been considered yet.

Finally, the benchmark depends on PDF extraction and English-language processing.

Heterogeneous web formats and multilingual reports, which are common in practice, remain outside the current scope and should be addressed in future iterations.

6 Conclusion and Practical Implications

Our strongest configuration (Claude 3.5 combined with hybrid dense sparse retrieval and cross-encoder reranking) answered 62% of the 330 question benchmark correctly and a further 10% answers were partially incomplete. Put differently, roughly three-quarters of the outputs were at least directionally useful, while one quarter remained factually wrong or unsupported.

What these numbers mean in practice.

- **Augmented analyst workflows.** At 62 %+10 % accuracy, a RAG pipeline can already serve as first-pass summarization or evidence-surfacing tools. In a typical ESG due-diligence loop, analysts spend the majority of their time locating passages, tables and footnotes. Automating that step can cut reading time even when every answer is manually verified afterwards.

- **Human-in-the-loop is still mandatory.** A 25-30 % error rate remains too high for regulatory disclosure, portfolio weighting or automated sustainability scoring. Every generated answer therefore needs stronger safeguards against hallucination and a review layer. In practice, showing the retrieved snippets alongside the model's answer enables an experienced analyst to validate or override the response in a few seconds.

- **Retrieval dominates further gains.** Numerical and logical questions fail largely because the correct passage never reaches the context window. Hence marginal gains from scale (moving to larger proprietary models) are smaller than gains from smarter retrieval or domain-aware chunking.

- **Environmental trade-offs.** Deploying Claude 3.5 or GPT-4o in production multiplies carbon emissions per query relative to a quantized local models. Organizations that prioritize sustainability can already choose lighter models with human review to keep both error rates and emissions within acceptable bounds.

The benchmark therefore positions current RAG technology not as a replacement for ESG analysts but as a force multiplier that can save time, reduce tediousness and broaden coverage, while leaving final decision-making to human expertise.

7 Moving Forward

We plan to incorporate additional data sources to broaden coverage of companies, sectors and sustainability metrics.

We also anticipate exploring new retrieval strategies (e.g., domain-specific expansions, robust table extraction) and extended evaluation metrics beyond simple correctness, such as faithfulness to source and the ability to handle multi-document summaries.

A first key direction is the development of adaptive answering methods. By tailoring retrieval prompts and selection strategies to the type of question (e.g., numerical, logical, extractive), we can significantly reduce the semantic gap between query and document content. Typed-RAG, for instance, introduces a type-aware decomposition method that improves answer precision for complex question formats [28]. Similarly, task-aware retrieval using instructions improves retrieval relevance and accuracy while reducing the need for model size scaling [29]. These approaches allow RAG systems to better align context construction with task intent, ultimately improving factual accuracy and minimizing hallucination.

Another line of enhancement involves incorporating structured knowledge representations into the pipeline. We plan to prototype a GraphRAG variant, which extracts ESG-specific entities, numeric values and relations to construct a knowledge graph used during retrieval. This hybrid approach supports symbolic reasoning, enables multi-hop inference, and improves retrieval faithfulness [30]. Recent results show that combining document graphs with entity-aware retrieval yields better coverage and more structured answers in focused summarization and QA tasks [31]. This is especially relevant for climate disclosures, where facts are often interdependent and scattered across tables, figures and text.

To improve cost-efficiency and reduce emissions, we also aim to implement dynamic model routing. Inspired by cascade models and energy-aware inference pipelines [32, 33], this method routes queries to lightweight models for simple lookups and reserves larger models for complex or high-risk questions. In a benchmark context, this means using quantized models or symbolic solvers when appropriate, reducing unnecessary compute load while preserving accuracy. Dynamic routing aligns with our sustainability goals by lowering energy use and latency without sacrificing interpretability or correctness.

Finally, we plan to explore the Agentic RAG paradigm, which allows an orchestration agent to select and sequence retrieval tools at runtime. By combining document search, knowledge graph access, and symbolic modules under agentic control, the system can flexibly plan how to answer a question, rather than relying on a fixed RAG pipeline. The ReAct

framework [34] and Toolformer [35] demonstrate that language models equipped with action-selection capabilities can improve factual consistency, reduce hallucinations, and generate more interpretable reasoning traces. In our case, such orchestration can enhance trust in climate-finance QA outputs and ensure that high-stakes queries follow robust, auditable decision paths.

8   Data accessibility

We have made our dataset and code files available in a dedicated repository:

<div align="center">

`github.com/Pladifes/climate_finance_bench`

</div>

All dataset assets (PDF excerpts, questions and gold answers) are distributed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC-BY-NC-SA 4.0) licence.

The repository includes notebooks so that users can:

- generate vector stores from reports and

- run RAG pipelines by choosing from multiple retrieval and LLM options.

Users can experiment with various RAG configurations, test different LLMs and compare results.

9   Acknowledgements

10 Funding

References

[1] Network for Greening the Financial System, "Bridging Data Gaps: Data Availability and Needs for Addressing Climate-Related Financial Risks," NGFS Technical Document, 2022.

[2] Financial Stability Board, "2023 Status Report: Task Force on Climate-related Financial Disclosures," FSB, 2023.

[3] Bank for International Settlements, "Project Gaia - Enabling climate risk analysis using generative AI," March 2024.

[4] Deutsche Bundesbank, "Informing climate risk analysis using textual information – A research agenda ," Technical Report 2024-01, 2024.

[5] T. B. Brown et al., "Language Models are Few-Shot Learners," in *NeurIPS*, vol. 33, pp. 1877–1901, 2020.

[6] OpenAI, "GPT-4 Technical Report," *arXiv preprint arXiv:2303.08774*, 2023.

[7] Z. Ji et al., "Survey of Hallucination in Natural Language Generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1-38, 2023.

[8] G. Mialon et al., "Augmented Language Models: a Survey," *arXiv preprint arXiv:2302.07842*, 2023.

[9] Y. Guo et al., "Retrieval-Augmented Generation for Large Language Models: A Survey," *arXiv preprint arXiv:2305.09675*, 2023.

[10] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *NeurIPS*, vol. 33, pp. 9459–9474, 2020.

[11] G. Izacard and E. Grave, "Leveraging Passage Retrieval with Generative Models for Open-Domain Question Answering," in *ICLR*, 2021.

[12] H. Yang et al., "FinGPT: Open-Source Financial Large Language Model," *arXiv preprint arXiv:2306.06031*, 2023.

[13] S. Mehta et al., "ESG-BERT: A Pre-trained Model for ESG Corporate Disclosures Classification," *arXiv preprint arXiv:2204.11110*, 2022.

[14] P. Islam et al., "FINANCEBENCH: A new benchmark for financial question answering," *arXiv preprint arXiv:2311.11944*, 2023.

[15] X. Wang et al., "Searching for best practices in retrieval-augmented generation," *arXiv preprint arXiv:2407.01219*, 2024.

[16] T. Schimanski et al., "ClimRetrieve: A benchmarking dataset for information retrieval from corporate climate disclosures," *arXiv preprint arXiv:2406.09818*, 2024.

[17] Z. Chen et al., "FinQA: A Dataset of Numerical Reasoning over Financial Data," in *EMNLP*, pp. 3696–3709, 2021.

[18] Z. Chen et al., "ConvFinQA: Exploring the Chain of Numerical Reasoning in Conversational Finance Question Answering," in *EMNLP*, pp. 6869–6884, 2022.

[19] Y. Zhu et al., "TAT-QA: A Question Answering Benchmark on a Hybrid of Tabular and Text Data in Finance," in *EMNLP*, pp. 3277–3288, 2021.

[20] T. Diggelmann et al., "Climate-FEVER: A Dataset for Verification of Real-World Climate Claims," *arXiv preprint arXiv:2012.00614*, 2020.

[21] X. Wu et al., "Golden Touchstone: A Comprehensive Bilingual Benchmark for Evaluating Financial Large Language Models," *arXiv preprint arXiv:2411.06272*, 2024.

[22] G. Izacard et al., "Atlas: Few-shot Learning with Retrieval-Augmented Language Models," in *ICLR*, 2022.

[23] F. Petroni et al., "KILT: A Benchmark for Knowledge Intensive Language Tasks," in *NAACL*, pp. 2523–2544, 2021.

[24] N. Thakur et al., "BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models," in *CIKM*, pp. 2281–2290, 2021.

[25] S. Shen et al., "BloombergGPT: A Large Language Model for Finance," *arXiv preprint arXiv:2303.17564*, 2023.

[26] D. Patterson et al., "Carbon Emissions and Large Neural Network Training," *Communications of the ACM*, vol. 65, no. 7, pp. 86–96, 2022.

[27] R. Schwartz et al., "Green AI," *Communications of the ACM*, vol. 63, no. 12, pp. 54–63, 2020.

[28] D. Lee et al., "Typed-RAG: Type-aware Multi-Aspect Decomposition for Non-Factoid Question Answering," *arXiv preprint arXiv:2503.15879*, 2025.

[29] A. Asai et al., "Task-aware Retrieval with Instructions," in *Findings of the Association for Computational Linguistics*, pp. 3650–3675, 2023.

[30] B. Peng et al., "Graph Retrieval-Augmented Generation: A Survey," *arXiv preprint arXiv:2408.08921*, 2024.

[31] D. Edge et al., "From Local to Global: A GraphRAG Approach to Query-Focused Summarization," *arXiv preprint arXiv:2404.16130*, 2024.

[32] N. Gupta et al., "Language Model Cascades: Token-Level Uncertainty and Beyond," in *International Conference on Learning Representations (ICLR)*, 2024.

[33] P. J. Maliakel et al., "Investigating Energy Efficiency and Performance Trade-offs in LLM Inference Across Tasks and DVFS Settings," *arXiv preprint arXiv:2501.08219*, 2024.

[34] S. Yao et al., "ReAct: Synergizing Reasoning and Acting in Language Models," in *International Conference on Learning Representations (ICLR)*, 2023.

[35] T. Schick et al., "Toolformer: Language Models Can Teach Themselves to Use Tools," *arXiv preprint arXiv:2302.04761*, 2023.

A Appendix A
Automated Grading with a LLM-as-a-Judge

A.1 Human vs. Automatic Evaluation

Human grading. For a sample of 330 RAG answers, each one was labeled by an annotator on a three-level scale:

- **Correct**: matches the gold answer within a narrow tolerance for wording;

- **Incomplete**: on the right track but missing key details;

- **Incorrect**: factually wrong or off-topic.

LLM-as-a-Judge. To scale evaluation we asked a LLM to grade the same answers. Given the *question*, the *RAG answer* and the *gold answer*, the LLM must output only a number corresponding to the labels:

$$2 \text{ (correct)}, \quad 1 \text{ (incomplete)}, \quad 0 \text{ (incorrect)},$$

with no commentary.

A.2 Prompt Design

We tested three variants:

1. **Llama3.1 8B Instruct**, *without* a reminder of the question;

2. **Llama3.1 8B Instruct**, *with* a reminder of the question;

3. **Claude 3.5 Sonnet**, with a reminder of the question.

A.3 Agreement with Human Judgements

A *soft* match counts {correct, incomplete} as agreement, whereas a *hard* match counts only {correct}.

Table 2: Human–LLM agreement on the 330-answer test set

Grader setup	Soft match	Hard match
Llama3.1 8B Instruct (no question)	$272/330 = 82.4\%$	$165/330 = 50.0\%$
Llama3.1 8B Instruct (with question)	$280/330 = 84.4\%$	$179/330 = 54.2\%$
Claude 3.5 Sonnet (with question)	$277/330 = 83.9\%$	$227/330 = 68.7\%$

Claude 3.5 achieved the highest hard-match rate (68.7%) while maintaining soft-match parity with Llama3.1 8B Instruct, and is therefore used for all subsequent automatic evaluations.

A.4 Distribution Shift

Table 3 contrasts the label distributions produced by human graders and by Claude 3.5. The LLM is noticeably stricter: it assigns 9.9 pp fewer *correct* labels and 10.0 pp more *incorrect* labels, while the *incomplete* share stays nearly unchanged.

Table 3: Human vs. Claude 3.5 label distribution on the 330-answer test set

	Correct	Incomplete	Incorrect
Human evaluation	174 (52.7%)	63 (19.1%)	93 (28.2%)
LLM-as-a-Judge	138 (41.8%)	66 (20.0%)	126 (38.2%)

A.5 Type I and Type II Risk

We gauge the reliability of the automatic grader using Table 4.

Table 4: Confusion matrix (Claude 3.5 vs. human labels).

Human label	LLM label		
	Incorrect (0)	Incomplete (1)	Correct (2)
Incorrect (0)	83	6	4
Incomplete (1)	25	24	14
Correct (2)	18	36	120

We consider type I error (false accept) when the LLM judges an answer correct (label 2) while the human grader says it is incorrect or merely incomplete, and type II error (false reject) when the LLM judges an answer incorrect (label 0) or incomplete (label 1) while the human grader deems it correct.

With a set of 330 questions:

$$\text{Type I errors} = 4 + 14 = 18 \quad (5.45\%)$$

$$\text{Type II errors} = 18 + 36 = 54 \quad (16.36\%)$$

The grader is conservative: it underrates correct answers three times more than is over-credits wrong ones. This is desirable in a benchmark where false positives would inflate model scores.

The 16 % Type II rate means some genuinely correct answers are penalized. Reported accuracies for all models are therefore slightly conservative, a trade-off we accept for fully automated scaling.

Take-away. Despite a small distribution shift, Claude 3.5 aligns with human judgements on more than two-thirds of *hard* cases, providing a cost-effective and reproducible grading mechanism for the remainder of our experiments.

B   Appendix B
GHG Emissions Computation Methodology

What is at stake? Large-scale language models already contribute an electricity demand on par with that of some small nations, and their footprint is growing faster than most mitigation pathways allow. If research communities omit credible carbon accounting, three risks arise: *(i) Scientific integrity*: results that overlook energy cost may promote architectures that are infeasible under tightening carbon budgets; *(ii) Legal exposure*: institutions that publish "AI-for-climate" tools without disclosing their own emissions may soon contravene emerging disclosure laws; *(iii) Capital-allocation bias*: investors and policymakers, lacking transparent numbers, could funnel resources toward high-footprint solutions that erode limited global carbon budgets instead of toward lower-impact alternatives. Robust footprint measurement is therefore not an optional add-on but a prerequisite for credible, actionable climate-finance research.

B.1 Why measure the carbon footprint?

Regulatory and fiduciary context. The European *Corporate Sustainability Reporting Directive* (CSRD) and disclosure frameworks such as TCFD[7] and IFRS S2[8] increasingly require firms and not only heavy industry to quantify Scope 2 (electricity) and Scope 3 (up/down-stream) greenhouse-gas (GHG) emissions. Research groups that propose AI tools for climate finance therefore face a dual responsibility: (1) to demonstrate that the use-phase of their models does not undermine the very decarbonisation goals they seek to advance, and (2) to provide transparent numbers that downstream users can incorporate into their own value-chain accounting.

Where do the emissions come from in a RAG pipeline?

- **Model inference.** Each forward pass through a large-parameter LLM drives a GPU at dozens to hundreds of watts. Although training is more energy-intensive per hour, repeated inference requests dominate in a production search-and-answer workload. API calls to proprietary back-ends move this energy use off-premises but do not eliminate it.

- **Retrieval and indexing.** Dense vectorisation of long PDF reports and nearest-neighbour search in FAISS are CPU and memory-heavy and can also leverage GPU if OCR is involved. While it is a one-off cost in our benchmark, a live system that ingests frequent ESG filings must re-index regularly, making this a non-negligible share of total energy.

- **Evaluation.** Using an LLM-as-a-Judge multiplies the number of model invocations, adding its own emissions line item.

- **Data movement and cooling.** Every gigabyte shuttled between object storage, RAM and GPU DRAM and every kilowatt-hour dissipated as heat requires additional electricity that is rarely met with 100 % renewables.

Why are the resulting emissions global? Computation is executed in hyperscale data-centres scattered across the United States, Europe and Asia, each connected to a distinct electricity grid with its own carbon intensity. A single Kaggle session may run in Iowa (coal-heavy mix) today and in Belgium (higher share of wind) tomorrow; similarly, an OpenAI or Anthropic request may land on hardware in Oregon, Virginia or Dublin. Consequently, the

---

[7]Task Force on Climate-related Financial Disclosures.

[8]International Financial Reporting Standards S2.

same Python script yields different real-world emissions depending on where the scheduler places the job and on hourly fluctuations in energy mix. Embodied carbon in GPUs, network switches and cooling infrastructure further spreads the climate impact across manufacturing hubs in Taiwan, South Korea and the wider semiconductor supply chain. Quantifying the footprint therefore requires both local energy logs and global carbon-intensity factors, as reflected in the methodology that follows.

B.2 Study overview

We distinguish two broad sources of energy consumption in our study:

1. **Local inference runs** (vector-store creation, RAG pipelines and LLM generation on Kaggle GPUs/CPUs).

2. **Remote API calls** to proprietary LLM providers (OpenAI's GPT-4o and Anthropic's Claude 3.5), including the LLM-as-a-Judge evaluation phase.

B.3 Remote API calls

For each prompt sent to GPT-4o or Claude we queried the *EcoLogits*[9] service, which returns a minimum and maximum estimate of GHG emissions (in $kg\,CO_2\,eq$) given the token counts and the limited public information on vendor infrastructure. Because the underlying hardware and regional electricity mixes are opaque, we adopt the maximum value as a conservative upper bound, compute the sample mean $\bar{e}_{\mathrm{API}}$ and standard deviation σ_{API} across a calibration batch and then scale by the total number N_{API} of queries:

$$E_{\mathrm{API}} = N_{\mathrm{API}}\,\bar{e}_{\mathrm{API}} \quad \text{with uncertainty } \pm N_{\mathrm{API}}\,\sigma_{\mathrm{API}}.$$

B.4 Local runs (Kaggle)

Instrumentation. We wrapped every Kaggle notebook in a `CodeCarbon`[10] (v 2.8) context manager to log:

- CPU and RAM energy draw on the host machine;

---

[9]https://ecologits.ai/latest/
[10]https://codecarbon.io/

- GPU run-time in seconds, captured via `nvidia-smi`.

Converting energy to emissions. Let $E_{\text{CPU+RAM}}$ and t_{GPU} be the cumulative energy (kWh) and GPU run-time (h) for a given run. For each physical host we resolve its ISO country code via `CodeCarbon`, look up the country-level carbon-intensity CI ($\text{kg}\,\text{CO}_2\,\text{eq/kWh}$) from the Climate Change Performance Index database[11] and compute

$$E_{\text{local}} = CI\,(E_{\text{CPU+RAM}} + t_{\text{GPU}} \times 0.250),$$

because the NVIDIA P100 on Kaggle is rated at $250.00\,\text{W}$ TDP.

Example. For a typical 330-question batch on an unquantized Llama3.1 8B model with a server located in the USA:

$$E_{\text{CPU+RAM}} \approx 0.27\,\text{kWh},$$
$$t_{\text{GPU}} \approx 4.70\,\text{h},$$
$$CI_{\text{USA}} = 0.35\,\text{kg}\,\text{CO}_2\,\text{eq}\,/\,\text{kWh},$$
$$\Rightarrow\ E_{\text{local}} \approx (0.27 + 4.7 \times 0.250)\,(\text{kWh})$$
$$\times\,0.349\,(\text{kg}\,\text{CO}_2\,eq\,/\,kWh)$$
$$\approx 0.50\,\text{kg}\,\text{CO}_2\,\text{eq}.$$

---

[11]`https://ccpi.org`

B.5 Emissions summary.

Table 5: GHG emissions per **330-question** run for each model, including evaluation, and for vector-store generation

Configuration	Emissions $(\mathrm{kg\,CO_2\,eq})$	Uncertainty $(\mathrm{kg\,CO_2\,eq})$
LLaMA 3.1-8B (full-precision)	0.92	0.00
LLaMA 3.1-8B (4-bit)	0.23	0.00
Mistral NeMo-12B (4-bit)[†]	0.38	0.00
GPT-4o (OpenAI API)	2.37	1.32
Claude 3.5 Sonnet (API)	2.69	1.49
Vector-store generation	0.10	n/a

[†] GPU energy for NeMo was approximated using LLaMA 3.1 power draw; the true value may differ.

Table 5 summarises the per-run carbon footprint, including evaluation.

Three patterns stand out:

1. **API calls dominate.** Remote inference on GPT-4o and Claude accounts for $\sim$5 kg CO_2eq—roughly 77.00% of the subtotal for a single RAG configuration. It is expected because these models are substantialy bigger in size compared to the models we ran locally and each prompt triggers an opaque multi-GPU back-end and must include large uncertainty margins.

2. **Quantization pays off.** Compressing LLaMA 3.1 from full precision to 4-bit slashes emissions from 0.92 to 0.23 $\mathrm{kg\,CO_2}$eq ($\approx 75\%$ savings) while retaining answer quality within two percentage points. Similar gains are expected for NeMo, although its figure is an approximation.

3. **Vector stores are not the main emission point.** End-to-end indexing of all sustainability reports emitted just 0.10 $\mathrm{kg\,CO_2}$eq, two orders of magnitude below the API footprint, showing that retrieval costs are negligible compared with repeated LLM inference.

Doubling the experiment to test two RAG pipelines raises the subtotal from 6.60 to 13.19 $\mathrm{kg\,CO_2}$eq, and the overall benchmark footprint (including index generation) to 13.29 ± 5.62 $\mathrm{kg\,CO_2}$eq. The error band is driven almost entirely by the vendor-side uncertainty on GPT-4o and Claude. Further transparency from providers about region-level energy accounting would narrow these bounds and help researchers optimize low-carbon deployments.

B.6 Limitations

- Lack of fine-grained telemetry for GPT-4o/Claude forces us to rely on coarse upper bounds.

- Kaggle does not guarantee the same data-centre region across sessions, so we approximate with the reported country code for each run.

- Carbon-intensity figures rely on national averages; data-centre-specific power-purchase agreements (PPAs) could yield substantially lower actual emissions.

- Provider-side uncertainty accounts only for hardware variation, not for idle overhead, networking, or cooling system coefficient of performance (COP).

- We could not evaluate carbon footprint of Qwen 2.5 and DeepSeek R1 calls because we did not have access to Nebius energy metrics.

C Appendix C

Selected companies and index affiliation

S&P500	CAC40 & DAX40	Other
Apple	Axa	Alibaba Group
AT&T	BNP Paribas	ArcelorMittal SA
Bank of America	BASF	Baoshan Iron & Steel
ExxonMobil	Engie	BP
General Electric	LVMH	Hindustan Unilever
Alphabet (Google)	Orange	Nestlé
Meta Platforms	Sanofi	NTPC
Microsoft	Siemens AG	Roche Holding AG
NVIDIA	Suez	Samsung Electronics
Pfizer	TotalEnergies	Sinopharm
Sysco	Veolia Environnement	SPD Bank

Appendix D

Annotation Guide for Climate Finance Bench

1. Introduction

We build a database that links 33 corporate climate reports to

- ten analyst-style questions per report,

- the reference answers, and

- the evidence passages used to derive those answers,

so that retrieval-augmented generation (RAG) systems can be evaluated on the same tasks.

Rigorous annotation is critical for trustworthy results. This guide sets out the tools, rules, and examples you must follow. **All fields in the database must be written in English.** If in doubt, contact your referents before proceeding.

2. Reports to Annotate

All reports are supplied as PDF files in the shared drive:

<div align="center">

`Company Reports/`

</div>

Verify that every document you use is a climate or sustainability report.

3. Filling in the "Annotation Table" Sheet

Each annotator completes all columns for their assigned rows, except `Validity status of the annotation`, which is for the ESG expert only.

Column-by-column instructions

1. **Annotator's name**: format: `firstname_surname`

2. **Company's name**: exactly the folder name that contains that company's reports.

3. **Fiscal year**

   - Use the fiscal year covered, not the publication year.
   - If the question spans several years, list them: `2020, 2021`.
   - Check that the year matches the report contents.

4. **Question ID**: one ID for each of the ten questions (e.g. `Q1`).

5. **Question**

   - Copy exactly from the master question list.
   - Replace `FYXXXX` with the concrete year, e.g. `FY2020`.

6. **Type of question**: `PE` (Pure Extraction), `NR` (Numerical Reasoning) or `LR` (Logical Reasoning).

7. **Answer**

   - Write a concise English answer once found.
   - If the information is not in the report, write exactly:
     `Not available in the retrieved information.`
   - You may ask internal or external chatbots, but validate the response against the PDF.

8. **Documents**

   - Use the exact PDF filename(s).
   - Multiple documents: list in chronological order, separated by commas (e.g. `doc1.pdf, doc2.pdf`).

9. **Pages**

   - Give the PDF page numbers, not the printed page labels.
   - One document: `doc1{page17, page26}`. Multiple documents: `doc1{page9}, doc2{page1, page27}`.
   - `doc1` = first file named in the **Documents** column, `doc2` = second file named, etc.

10. **Document extracts**

    - Copy-paste the full paragraph / table / figure caption.
    - One extract: `doc1{<extract>}`. Several extracts: `doc1{...}`, `doc1{...}`, `doc2{...}`.
    - For a table or figure, copy textual content rather than screenshots.
    - If the PDF is image-only, screenshot, run OCR (e.g. ChatGPT), and verify the text.

11. **Extract type**: choose from `text`, `table`, `figure`, `text+table`, `text+figure`

12. **Comments**: pick one of these options

    - `Nothing to report` (default)
    - `Uncertain`
    - `Additional comments`

13. **Additional comments**: fill only if the previous field is `Additional comments`. Write a short, clear remark.

14. **Validity status of the annotation**: ESG expert only. Options: `Validated by the expert` / `Modified by the expert`. The default entry is `To be validated`.

4. Annotation Examples

Examples are provided in the workbook:

- *Text answer* → sheet `Examples / Example 1`
- *Table answer* → sheet `Examples / Example 2`
- *Figure answer* → sheet `Examples / Example 3`

5. Quality-Control Procedure

A dedicated ESG expert reviews all annotations, with special attention to rows whose **Comments** field is `Uncertain` or `Additional comments`. The expert then sets the `Validity status of the annotation` to

- `Validated by the expert`, or

- `Modified by the expert` (if corrections were required).

6. Practical Tips

- You may upload the PDF to our internal or external chatbots, ask the question, and then verify the answer in the report.

- If the report is an image (no selectable text), take a screenshot, run OCR, and check the transcription carefully.