

ESG-QA: Building a Dataset for Question Answering on Environmental, Social, and Governance Pillars

**Gabriel Assis[‡], Ayrton Surica[‡], Pedro Kroll[‡], Gabriela Aires[†],
Darian Rabbani[†], Edson Bollis[†], Lucas Pellicer[†], Aline Paes[‡]**

[‡] *Universidade Federal Fluminense, Instituto de Computação, Niterói, RJ, Brazil*

[†] *Instituto de Ciência e Tecnologia Itaú, São Paulo, SP, Brazil*

{assisgabriel@id, ayrtonsurica@id, pedrokroll@id, alinepaes@ic}.uff.br;

{gabriela.aires-mendes, darian.rabbani, edson.bollis, lucas.pellicer}@itau-unibanco.com.br;

Abstract

Environmental, Social, and Governance (ESG) factors are becoming increasingly central to corporate accountability and sustainable development. However, benchmarks for evaluating large language models (LLMs) in this domain remain scarce. To alleviate this gap, we present ESG-QA, a dataset of 87,261 question–answer–context triplets spanning the three ESG pillars. ESG-QA was built using an LLM-based Question Answering (QA) generation pipeline, enhanced with rule-based and semantic filtering, and validated through human inspection of sampled instances, enabling both abstractive QA and retrieval-augmented QA. We benchmark three open-weight LLM families (Llama-3, Gemma-3, and Qwen-3) across multiple dimensions, including correctness, environmental impact, and readability. Results show that Qwen-3 with retrieval achieves the highest absolute QA performance, while Gemma-3 provides the strongest overall balance between correctness, efficiency, and clarity. By releasing ESG-QA and its generation framework, this work establishes a comprehensive benchmark for advancing ESG-oriented QA and promoting more transparent and responsible AI evaluation.

Keywords: ESG, Question Answering, Large Language Models, Benchmarking, Responsible AI

1. Introduction

Environmental, Social, and Governance (ESG) dimensions have become central to contemporary debates on sustainable development and corporate accountability (Zhang et al., 2025; He et al., 2025; Schimanski et al., 2024b). ESG provides a comprehensive framework for assessing how organizations integrate environmental priorities, social demands, and governance practices into their strategic decision-making (Schimanski et al., 2024b).

Once misinterpreted as mere compliance tools, ESG practices are now widely linked to long-term financial performance, serving as mechanisms for risk mitigation and opportunity creation (Friede et al., 2015). Evidence shows that firms with strong ESG commitments benefit from lower capital costs, stronger stakeholder trust, and greater innovation capacity (Friede et al., 2015). This evolution highlights that ESG is not simply a reporting requirement, but a strategic factor shaping competitiveness and resilience. Consequently, both companies and investors increasingly integrate ESG into decision-making, recognizing it as a driver of sustainable growth and financial value creation.

On the other hand, the financial sector’s dynamism and the vast amount of data it generates demand automation and refined analytical techniques (Najem et al., 2025). Advances in Large Language Models (LLMs) have increasingly filled this role, supporting tasks that include investment forecasting (Zhang et al., 2024), fundamentalist anal-

ysis (Nepomuceno et al., 2025; Han et al., 2024), and the automatic generation of documents and reports (Assis et al., 2024). However, despite considerable progress in evaluating these models in terms of corporate aspects such as financial reasoning (Mateega et al., 2025) and mathematical capabilities (Chen et al., 2021), systematic assessments of their handling of ESG-related dimensions remain scarce (Ali et al., 2025). In particular, one contributing factor is the limited availability of comprehensive benchmarks specifically designed for ESG evaluation (Ali et al., 2025; He et al., 2025).

Building on this context, we introduce ESG-QA, a large-scale dataset comprising 87,261 Question & Answer (QA) pairs contextualized across the three ESG pillars. Its construction relied on a Question Answer Generation (QAG) framework powered by LLMs, which transformed an extensive unannotated ESG corpus (Schimanski et al., 2024b) into QA form. ESG-QA enables two complementary modes of evaluation: it can be used to directly assess the knowledge encoded in LLMs through QA pairs alone, and it also supports retrieval-based setups by including their contextual passages. Importantly, the dataset adopts an abstractive QA design (Srivastava and Memon, 2024; Assis et al., 2025), requiring systems to generate answers instead of selecting spans or options, addressing recent critiques of multiple-choice evaluation for advanced LLMs (Balepur et al., 2025).

In addition to releasing the dataset, we developed a reproducible QAG framework that allows

the expansion and replication of similar resources, and we carried out a systematic evaluation of three families of LLMs, namely Llama-3 (Meta-[AI, 2024](#)), Gemma-3 (Gemma-Team, 2025), and Qwen-3 (Yang et al., 2025), using ESG-QA under a comprehensive set of abstractive QA metrics (Srivastava and Memon, 2024). Aligned with ESG principles, we further extended the evaluation to incorporate environmental impact (Budenny et al., 2023) and textual readability¹, aiming to establish a multi-aspect protocol that goes beyond correctness alone.

Our findings indicate that although the smaller Qwen-3 4B tool-enhanced achieved the best absolute QA performance, the Gemma-3 family stands out for its balance of QA correctness, efficiency, and readability. This may stem from Qwen’s training emphasis on tasks such as mathematics, which can affect both generation time and the presentation of answers. Taken together, our resource, its generative evaluation setup, and the accompanying evaluation framework provide a comprehensive foundation for advancing research on ESG-oriented QA and evaluating language models in this domain.

The contributions of this work include (c1) the release of ESG-QA as a large and comprehensive dataset for ESG-focused evaluation, (c2) the design of a reproducible framework for QAG that enables future resource generation, and (c3) a broad evaluation of three major LLM families on ESG-QA, examining their performance not only in terms of correctness but also with respect to sustainability and linguistic clarity, demonstrating the applicability of our proposed benchmark².

2. Related Work

The use of QA has become a standard paradigm for evaluating knowledge and reasoning in automatic systems, including recent LLM-based architectures (He et al., 2025; Zhang et al., 2025). In this regard, several resources have been developed within the financial domain (Mateega et al., 2025; Reddy et al., 2024; Chen et al., 2024, 2021; Maia et al., 2018), though only a limited number explicitly address ESG aspects (He et al., 2025; Zhang et al., 2025; Ali et al., 2025). While LLM-based systems continue to evolve, the demand for updated evaluation resources becomes critical. This section, therefore, discusses recent advances in QA datasets for fin-applications, and, more specifically, aligned with our work, ESG-oriented QA resources.

Corporate Finance QA The financial domain reflects this evolving environment, with resources such as FiQA (Maia et al., 2018), FinQA (Chen

et al., 2021), FinanceQA (Mateega et al., 2025), and SFinD (Strative AI, 2024). These datasets vary in methodology: some are expert-annotated (Chen et al., 2021; Mateega et al., 2025), others are synthetic (Strative AI, 2024), and several combine human-in-the-loop supervision with automated generation (Maia et al., 2018). Despite these differences, all concentrate on corporate financial reasoning, encompassing numerical and tabular reasoning (Chen et al., 2021), retrieval from financial reports (Strative AI, 2024), and finance-specific knowledge and reasoning (Mateega et al., 2025; Maia et al., 2018).

ESG QA More recently, efforts have focused on ESG-related QA. Datasets such as Climate Finance Bench (Mankour et al., 2025) and ClimRetrieve (Schimanski et al., 2024a) primarily emphasize climate- and environmental-related aspects, while others aim to address ESG pillars more comprehensively. Within this last group, MMESG-Bench (Zhang et al., 2025) provides 933 multimodal questions to evaluate multimodal language models using ESG-related corporate PDFs. Its construction combines Qwen-family models for multimodal extraction with human inspection. SustainableQA (Ali et al., 2025) incorporates the European Union taxonomy of sustainable activities and employs GPT-4o in its generation process, resulting in 195,000 QA pairs. While it includes broader ESG aspects across different pillars, its perspective remains largely centered on sustainability. Lastly, ESGenius (He et al., 2025) offers 1,136 multiple-choice QA instances, entirely human-authored, focusing on ESG knowledge.

Our contribution, ESG-QA, complements these ESG-specific resources by providing a large-scale dataset of 87,261 question–answer–context triplets. Unlike prior work focusing on document-based reasoning (Zhang et al., 2025) or multiple-choice setups (He et al., 2025), ESG-QA enables abstractive evaluation, requiring systems to generate answers rather than selecting from candidates. This aligns with recent findings that abstractive QA yields more robust evaluations (Balepur et al., 2025). Moreover, including context supports retrieval-based evaluation scenarios, enabling multiple setups within the same resource.

Finally, while earlier works validate applicability through baseline experiments with LLMs, we extend our evaluation beyond correctness metrics by integrating indicators of environmental impact and linguistic simplicity, thus embedding ESG principles directly into our evaluation methodology. Further details are provided in the following sections.

¹<https://textstat.org/>

² <https://github.com/Melllitau/ESG-QA>

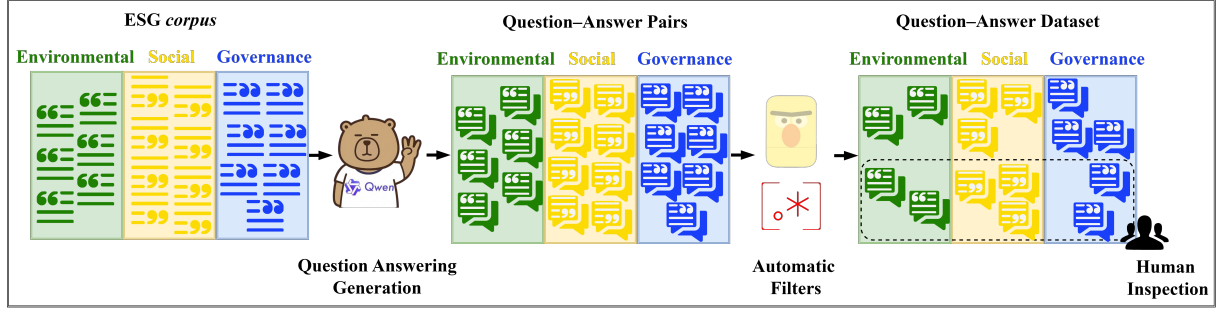


Figure 1: The adopted Question-Answering Generation pipeline.

3. Dataset Construction

This section introduces the construction of the ESG-QA dataset, a generative QA resource grounded in the Environmental, Social, and Governance pillars. Specifically, we present and discuss: (i) the underlying corpus employed as the basis for dataset creation (Section 3.1); (ii) the Question Answer Generation method applied to transform the corpus into a collection of question-answer pairs (Section 3.2); (iii) the filtering procedure designed to select consistent and unbiased questions (Section 3.3); and (iv) the human inspection process conducted to validate the generated resource (Section 3.4). The overall workflow is illustrated in Figure 1.

3.1. Data

We selected the corpus proposed by Schimanski et al. (2024b) as the basis for building our QA resource. This corpus was originally used to pre-train the ESGBERT family of models (Schimanski et al., 2024b), which were designed to classify texts into domains associated with the ESG pillars and to support ESG rating tasks from English-language textual communications. The corpus is public, large-scale, and diverse, with data drawn from annual reports, corporate social responsibility documents, and news articles. Its combination of real-world coverage, accessibility, and variety led us to adopt it as the basis of our dataset. Table 1 reports ESGBERT corpus statistics, with each instance represented as a sentence following the original setup³.

subset	#sentences	#words	#tokens
environment	2,100,586	56,775,185	71,172,980
social	1,787,198	51,230,511	63,800,537
governance	3,011,546	91,914,234	113,864,357

Table 1: ESGBERT corpus statistics by subset.

Table 1 also highlights another characteristic of the corpus that motivated its selection: it is pre-divided into data aligned with the three ESG pillars. This structure enables the creation of QA

pairs specific to each pillar, which, in turn, supports evaluations of QA methods for each component. Schimanski et al. (2024b) applied a word-based filtering procedure, resulting in frequent terms such as *environment*, *energy*, *emission*, *water*, and *diversity* for the environmental domain; *safety*, *community*, *social*, *human*, and *education* for the social domain; and *management*, *corporate*, *compliance*, and *standard* for the governance domain. Collectively, this aspect, along with the previous ones, establishes the corpus as an appropriate foundation for our QAG process.

3.2. Question Answering Generation

The ESGBERT base corpus is a sentence-level resource originally developed for pre-training encoder-based language models. Adapting it for QA required a procedure to generate grounded QA pairs from the corpus sentences. The corpus scale required an automated procedure. We therefore formulated the task of transforming the corpus into QA pairs as follows.

Problem Setup. Let $\mathcal{P} = \{E, S, G\}$ denote the ESG pillars (Environmental, Social, and Governance), $\mathcal{C} = \bigcup_{P \in \mathcal{P}} \mathcal{C}_P$ the set of contexts (sentences from the original corpus, separated according to their pillar $P \in \mathcal{P}$), $\mathcal{Q}$ the space of questions, and $\mathcal{A}$ the space of answers. Given a pillar P , a context $C \in \mathcal{C}$, and a number $N \in \mathbb{N}$ of question-answer pairs to be generated, we define a QAG function

$$\text{QAG} : \mathcal{P} \times \mathcal{C} \times \mathbb{N} \longrightarrow (\mathcal{Q} \times \mathcal{A})^N$$

such that

$$\text{QAG}(P, C, N) = \{(q_i, a_i)\}_{i=1}^N,$$

where each $q_i \in \mathcal{Q}$ is a question and each $a_i \in \mathcal{A}$ is its corresponding answer, both grounded in C and aligned with pillar P .

The literature has proposed several approaches for QAG. Encoder-based pre-trained language

³<https://huggingface.co/ESGBERT>

models (PLMs) are commonly used (Ushio et al., 2023), where contexts are provided as input, and the models extract answers and generate aligned questions. However, this approach is often limited to extracting exact answers from the context. More recently, decoder-based LLMs have achieved strong results through prompt-based generation, often without fine-tuning, and allow for more flexible question–answer creation (Paula et al., 2024). Therefore, to instantiate the proposed problem setup, we adopted an LLM-based strategy and employed the Qwen-3 32B model (Yang et al., 2025), balancing state-of-the-art open-weights performance (Artificial-Analysis, 2025) with our available computational resources, while also drawing on prior applications of models from the same series to similar tasks (He et al., 2025).

Prompt

You are an ESG-focused Q&A generator. You will receive a *context* tied to the {pillar} pillar of ESG. Based strictly on that context, generate {N} question–answer pairs whose answers can be fully derived from the context only. Reply in the same language of the context. Do **not** use external knowledge. Every question must be clearly linked to ESG and grounded in the provided context.

Avoid ALL questions that refer to the text itself, such as “What is this text about?” or “What does the context mention?”. Focus on the **substantive content and facts** presented in the context, not on the act of writing, reporting, or structure.

Do NOT ask vague or underspecified questions, such as “What role did support play?”—always specify what kind of support, from whom, and to what. Prefer **self-contained questions** that make sense without needing to re-read the context.

Avoid using generic terms like “organization”, “company”, or “institution” if the context does not mention a specific entity. Instead, focus on concrete actors, groups, locations, or initiatives that are explicitly referenced in the context.

When possible, **prefer open-ended questions** that promote reflection or generalization, rather than simply repeating facts from the context. For example, ask “**What practices could be adopted to improve transparency?**” instead of “**What practices were adopted by the company?**”, unless the latter is explicitly required by the context. Such questions should still be grounded in the context, but written in a way that makes them meaningful.

If fewer than {N} pairs are possible, fill remaining slots exactly with “<n>. Not possible” for Question / Answer.

Return **only** the following {N} blocks (no extra text):

1. Question: ...1. Answer: ...
2. Question: ...2. Answer: ...
- ...
- {N}. Question: ...{N}. Answer: ...

Context {context}

Figure 2: LLM prompt used to generate ESG-grounded QA pairs.

Figure 2 illustrates the prompt adopted for our QAG task. The prompt explicitly receives as input the number of pairs N , the ESG pillar, and the target context. It also exemplifies the flexibility of an LLM,

as the model can be guided by detailed instructions during generation. For instance, explicit constraints were introduced to prevent the model from focusing on textual form rather than substantive content, to encourage open-ended questions, and to avoid vague references to generic entities such as “the institution” or “the organization”. Additional guidance was included to shape the writing style of the generated questions. These refinements were incorporated iteratively based on analyses conducted during human inspection of successive generations.

We instantiated the model using LlamaIndex⁴ in combination with Ollama⁵, configured with a `context_window` of 10,000 tokens and a maximum `num_predict` of 4,000 tokens. This setup allows the prompt and input context to be processed while remaining flexible for long outputs. As input, we randomly sampled 13,000 instances from each of the three subsets of the ESGBERT corpus. This subset was selected due to hardware constraints that prevented processing the entire corpus. The decoding parameters were set to `temperature` = 0.8 and `top_p` = 0.9, allowing creativity in generation while keeping it below the maximum level to prevent absurd outputs (Li et al., 2025). Lastly, the LLM was instructed to try to generate $N = 3$ QA pairs for each given context.

After model execution, a total of 109,106 pairs were generated. This number is smaller than the direct product of $3 \times 3 \times 13,000$, since cases in which the generator returned “Not possible” as an answer were discarded.

3.3. Filtering

Following the automatic generation of question–answer pairs, two filtering mechanisms were employed: (3.3.1) a rule-based approach leveraging string matching, and (3.3.2) a neural approach using an encoder PLM. The goal was to suppress low-quality and undesirable generations while preserving informative QA pairs. Further details are provided in the following sections.

3.3.1. Rule-based Filtering

The rule-based filter was implemented using regular expression (regex) matching and applied with three main objectives: to remove defective generations produced by the models, to discard questions with low general relevance, and to mitigate bias.

Regarding defective generations, we removed cases in which the model signaled non-generation through outputs other than the “Not possible” instruction given in the prompt. We also excluded

⁴<https://www.llamaindex.ai/>

⁵<https://ollama.com/>

answers and questions that consisted solely of “...” or empty strings.

We also chose to remove pairs containing the token `context`. This decision followed an inspection of the generations, which revealed that such a strategy offered a low-cost solution for filtering shallow questions, such as “*What does the context mention?*” or “*What entities are mentioned in the context?*”. Although the prompt was designed to discourage such outputs (see Figure 2), our inspection indicated that the model still tended to produce them when the input context was exhausted in terms of deeper knowledge-based questions, rather than signaling with “Not possible”. We therefore opted to eliminate these types of questions, retaining those more closely aligned with ESG purposes. We acknowledge that this approach could potentially remove certain false negatives (e.g., cases where `context` was legitimately used). However, our inspection confirmed that the vast majority of such cases occurred in low-quality generations.

Finally, we removed questions and answers that explicitly referred to partisan political contexts, using a list of keywords such as {MAGA, woke, Biden, Trump, Conservative, Liberals, Progressive}, among others, available in our source code. This filter was introduced to avoid value judgments regarding policies promoted by different political spectra, as such references were identified in the original corpus. We emphasize that this choice does not reflect any normative judgment about right or wrong; rather, it seeks to keep the dataset as politically neutral as possible, focusing strictly on ESG-related aspects. An example of a pair removed by the low-quality filter, which would also be captured by the political-partisan filter, is presented in Appendix A.

3.3.2. Semantic Alignment Filtering

Although regex-based filtering can effectively eliminate evidently low-quality QA pairs, it is insufficient for capturing finer nuances of generation and alignment. For instance, an answer may be fluent and coherent but still fail to address the generated question. In previous work, adequacy is typically achieved through PLM-based filters (Paula et al., 2024). One example is the roundtrip approach (Alberti et al., 2019), where different models extract answers from the same context-question pair, and alignment is assessed by comparing these extractions. Alternatively, models can be applied directly to predict adequacy as a value (Wang et al., 2022).

Inspired by these approaches, we propose the use of PLMs in a Natural Language Inference (NLI)-based configuration (Yu et al., 2024). Whereas standard NLI tasks typically rely on the labels `Entailment` and `Contradiction`, we reformulate

them as `Aligned` and `Non-aligned`. Formally, given a question $q \in \mathcal{Q}$ and an answer $a \in \mathcal{A}$, the classifier f_θ should predict:

$$f_\theta(q, a) = \begin{cases} \text{Aligned}, & \text{if } a \text{ answers } q, \\ \text{Non-aligned}, & \text{otherwise.} \end{cases}$$

We adapted existing datasets for training PLMs on this task to operationalize this evaluation. Specifically, we selected four datasets: three from the financial domain (SFINd (Strative AI, 2024), FIQA (Maia et al., 2018), and FINANCEQA (Mateega et al., 2025)) and one general-domain dataset (the traditional SQuAD (Rajpurkar et al., 2016)). Each dataset provides naturally aligned pairs, which we treat as positive instances of the `Aligned` class. Negative instances were randomly created by re-pairing questions and answers at imbalance rates of 0.35 and 0.45.

For evaluation, we adopted a cross-fold-inspired strategy in which a classifier was trained on one dataset and evaluated on the others. In addition, based on the literature (Chen et al., 2025; Martinho et al., 2025), we selected three PLMs for the task, namely DeBERTa (He et al., 2021), FinBERT (Araci, 2019), and ModernBERT (Warner et al., 2024), and we also applied a soft-voting ensemble. The models were trained using the `Transformers` (Wolf et al., 2020) framework with default settings.

Due to space constraints, Table 2 reports only the best observed configuration, which used a 0.35 imbalance rate and training on the SFINd dataset. Results for all other configurations are available in our source code and data.

Model	fiQA	FinanceQA	SQuAD	Avg.
FinBERT	0.697	0.659	0.655	0.670
DeBERTa	0.875	0.666	0.688	0.743
ModernBERT	0.753	0.659	0.609	0.674
ModB+DeB	0.864	0.659	0.691	0.738
ModB+FinB	0.737	0.597	0.680	0.671
FinB+DeB	0.834	0.655	0.705	0.731
FinB+DeB+ModB	0.810	0.644	0.703	0.719

Table 2: Macro F1-score for models trained on SFINd (higher is better).

While the ensembles achieved the best performance on the general-purpose SQuAD dataset, DeBERTa outperformed the other models on the financial datasets and also on average. For this reason, DeBERTa was selected as the filtering model.

Table 3 reports the global filtering results, combining the PLM-based filter discussed in this section with the regex-based filter described previously, resulting in a final set of 87,261 QA pairs.

Removal strategy	Env.	Soc.	Gov.
regex:context	4,230	3,520	3,323
regex:bias	7	13	18
DeBERTa filter	3,867	3,421	3,446
Total removed	8,104	6,954	6,787
Total preserved	28,658	29,792	28,811

Table 3: Removed items by filtering strategy and ESG subset (counts).

3.4. Human Inspection

We conducted a human inspection with four annotators, one of whom was an ESG expert, and the others had prior experience in natural language processing (NLP) evaluation. Each QA pair was inspected by at least one annotator for **self-sufficiency** (whether the question is understandable without the context), **answer adequacy** (whether the answer properly addresses the question using the context), **political bias** (whether it conveys political partisanship), and **ESG adherence** (whether it is aligned with an ESG topic)⁶.

Criterion	Ratio (%)
Self-sufficiency	62.30
Answer adequacy	99.80
Non-Political bias	100.00
ESG adherence	100.00

Table 4: Human inspection: criterion ratios (% of QA pairs meeting each criterion).

In total, 976 QA pairs, stratified by the ESG pillars, were inspected by human annotators. Table 4 summarizes the results. For adequacy, political neutrality, and ESG alignment, the outcomes were nearly perfect, highlighting the robustness of the adopted automated QAG framework. However, approximately 37.7% of the pairs were considered non-self-sufficient. Importantly, annotations indicated that these pairs still satisfied the other criteria but required the context to be properly answered. Unlike the pairs removed by the regex-based filter, they nonetheless retain relevance for ESG. This proportion highlights the challenge an automatic system would face in answering such questions in isolation, while also demonstrating the potential of evaluating the dataset in enhanced settings, such as RAG (Fan et al., 2024), where supporting contexts are available for retrieval. We therefore do not regard this finding as negative.

⁶The exact evaluation protocol can be found at [to-be-deanonymized](#).

4. The ESG-QA Dataset

This section discusses the statistics of the 87,261 QA pairs generated in the proposed ESG-QA dataset. The dataset was split into 61,082/8,727/17,452 instances for training, validation, and testing, respectively, following an approximate 70/10/20 ratio while maintaining stratification across the ESG pillars. Moreover, the 976 pairs inspected by human annotators were selected as a curated subset of the test set, providing a manually validated evaluation set.

Table 5 reports the dataset statistics. The values are balanced across both splits and pillars, indicating a well-structured distribution. In particular, the statistics regarding the answers show that they are not limited to short spans (such as one to five words), which supports the use of the resource for evaluating the generative capabilities of models. In addition, we computed the average exact-match overlap between answers and contexts, which was low. Taken with the inspection confirming that the pairs are still grounded in the contexts, this finding shows that the answers go beyond direct copying and reflect abstraction, making the dataset suitable for assessing abstractive QA rather than only extractive approaches.

Additionally, Table 6 provides examples of QA pairs together with their corresponding contexts as included in the dataset. The examples further indicate that the answers are abstractive yet remain grounded, and they illustrate both broad, general questions — such as the example in the social pillar — and more specific, entity-oriented ones, exemplified in the environmental and governance pillars. This behavior underscores the resource’s potential for evaluating diverse ESG knowledge.

5. Large Language Models Evaluation

This section presents an evaluation of LLMs on the proposed resource, exploring different configurations to demonstrate their potential applications.

5.1. Experimental Setup

We evaluated the ability of three families of open-weight LLMs to answer questions on our constructed dataset, namely Llama-3 (Meta-AI, 2024), Gemma-3 (Gemma-Team, 2025), and Qwen-3 (Yang et al., 2025). These families were selected based on their widespread adoption, demonstrated performance in ESG-related contexts (He et al., 2025), and the computational resources available during the study. This selection further allows us to evaluate smaller open models, consistent with our goal of promoting responsible and sustainable AI. The evaluation was conducted under two experimental setups, as follows.

statistic	Train			Validation			Test		
	Env.	Soc.	Gov.	Env.	Soc.	Gov.	Env.	Soc.	Gov.
Avg. words/tokens (question)	12.94 / 14.49	13.09 / 14.65	13.25 / 14.88	12.91 / 14.43	13.04 / 14.62	13.19 / 14.76	12.92 / 14.44	13.10 / 14.68	13.12 / 14.75
Avg. words/tokens (answer)	17.61 / 19.92	18.25 / 20.73	18.81 / 21.31	17.68 / 19.97	18.31 / 20.78	18.80 / 21.28	17.54 / 19.79	18.31 / 20.75	18.35 / 20.79
Avg. words/tokens (context)	28.36 / 32.84	30.04 / 34.79	32.04 / 36.94	28.28 / 32.77	30.04 / 34.76	32.43 / 37.48	28.28 / 32.55	28.96 / 33.42	30.52 / 35.34
Avg. words/tokens (overall)	58.92 / 67.24	61.38 / 70.18	64.11 / 73.13	58.86 / 67.17	61.39 / 70.16	64.42 / 73.53	58.74 / 66.78	60.37 / 68.85	61.99 / 70.87
Avg. Answer-Context overlap	0.030	0.023	0.027	0.023	0.023	0.026	0.033	0.022	0.027
#instances	20,060	20,854	20,168	2,866	2,980	2,881	5,732	5,958	5,762

Table 5: Dataset statistics by split and pillar.

Pillar	Question	Answer	Context
Env.	What specific emissions from waste management are included in the corporate carbon footprint according to the GHG Protocol Guidance?	Emissions included are those from CO ₂ and biogenic carbon in waste, as well as CH ₄ from the decomposition of biogenic materials in landfills or waste-to-energy (WTE) technologies.	Following GHG Protocol Guidance, CO ₂ and biogenic carbon contained in waste and CH ₄ emissions from the decomposition of biogenic materials in landfills or waste-to-energy (WTE) technologies are captured in the overall corporate carbon footprint.
Soc.	What primary risk factors make homeless youth more vulnerable to sex and labor trafficking?	Homeless youth are more vulnerable to sex and labor trafficking due to poverty, unemployment, a history of sexual abuse, and a history of mental health issues.	Homeless youth are vulnerable to both sex and labor trafficking because they tend to experience a higher rate of the primary risk factors to trafficking: poverty, unemployment, a history of sexual abuse, and a history of mental health issues.
Gov.	What are the key principles of sustainability that Trinity recognizes as critical to its long-term value?	The key principles of sustainability recognized by Trinity include environmental stewardship, safety and quality assurance, corporate social responsibility, governance, and diversity and inclusion.	We recognize that further integrating the key principles of sustainability — including actions and transparency in environmental stewardship, safety and quality assurance, corporate social responsibility, governance, and diversity and inclusion — is critical to enhancing Trinity's long-term value.

Table 6: Example QA pairs by ESG pillar.

Zero-shot Setting. Under this setup, the models were provided with one question q_i at a time from the ESG-QA test partition and instructed to produce a direct answer without additional support or any examples, thereby testing the knowledge encoded in their parameters.

Knowledge Base (KB) Setting. In this setup, the models were given access to an information retrieval tool based on embedding vector indices, with default retrieval settings of $top_k = 3$ and the use of the `bge-small-en-v1.5` model (Xiao et al., 2023), which balances efficiency with performance comparable to state-of-the-art solutions. The knowledge base consisted of all context units c_i from the entire ESG-QA dataset, providing a usage scenario close to real-world conditions. The models were then asked to instructed the test-set questions q_i , following the previous setup.

Evaluation To evaluate the model responses, we adopted an open-generation QA (OQA) setting, in line with recent work showing that discriminative evaluations may fail to capture the true knowledge of LLMs (Balepur et al., 2025). Accordingly, we employed OQA metrics, combining reference-based evaluation methods that measure lexical overlap (e.g., F1, BLEU, METEOR, and ROUGE) with semantic-oriented metrics (e.g., BERTScore) (Srivastava and Memon, 2024). In addition, consistent with responsible AI practices, we used the `eco2AI` (Budenny et al., 2023) framework to measure environmental impact, including equivalent CO₂ emissions and energy consumption. Beyond

response correctness, we also assessed the readability of the outputs using the `textstat` package⁷. Taken together, these approaches enabled a comprehensive evaluation of model performance.

Environment Settings All experiments were conducted on an NVIDIA L40S GPU, with the models instantiated via Ollama and integrated into Llamaindex⁸ under both configurations. The experiments were executed in Rio de Janeiro, which is relevant for computing environmental impact indicators. The `context_window` was set to 10,000 tokens, with other hyperparameters left at default.

5.2. Results

Table 7 reports the results of the models evaluated on ESG-QA. Best zero-shot results are underlined, while best KB results are **bold**⁹. Overall, models enhanced with a KB achieved the highest scores, with Qwen-3 4B ranking first, followed by Gemma-3 12B and Llama-3 8B. This outcome was somewhat expected, as access to external knowledge generally improves performance. Notably, however, the best result came from a smaller model, which is relevant for GPU-constrained scenarios,

⁷<https://textstat.org/>

⁸<https://www.llamaindex.ai/>

⁹We also computed evaluation on the human-inspected question pairs, which align with the reported results and indicate overall resource quality. Results are omitted due to space constraints. Llama 1B+KB is not reported, as the model failed to use the tool effectively.

Mode/Model	Rank (OQA)	Inference Consumption (↓)					Generative Metrics (↑)								
		Time (h)	Energy (kWh)	CO ₂ eq (kg)	vRAM (GB)	F1	METEOR	BLEU	ROUGE				BERTScore		
									R1	R2	RL	RLs	prec.	recall	f1
gemma-3 4B	8	5.723	1.003	0.113	9.846	0.191	0.256	0.030	0.247	0.096	0.208	0.216	0.849	0.887	0.867
gemma-3 4B + KB	11	108.500	15.643	1.769	10.742	0.085	0.098	0.015	0.102	0.025	0.087	0.087	0.819	0.847	0.832
gemma-3 12B	5	<u>4.399</u>	1.022	0.116	11.658	<u>0.223</u>	0.280	0.041	<u>0.280</u>	<u>0.115</u>	<u>0.237</u>	<u>0.242</u>	<u>0.866</u>	0.892	<u>0.878</u>
gemma-3 12B + KB	2	36.480	8.819	0.997	12.476	0.332	0.475	0.144	0.371	0.242	0.325	0.325	0.876	0.923	0.899
llama-3.2 1B	10	4.913	<u>0.918</u>	<u>0.104</u>	15.630	0.120	0.237	0.020	0.142	0.060	0.116	0.121	0.819	0.884	0.850
llama-3.1 8B	6	5.479	1.393	0.157	31.843	0.211	<u>0.333</u>	<u>0.043</u>	0.248	0.113	0.206	0.214	0.847	0.903	0.874
llama-3.1 8B + KB	3	36.981	8.016	0.906	32.833	0.326	0.463	0.130	0.368	0.228	0.316	0.316	0.879	0.920	0.899
qwen-3 4B	9	35.209	6.891	0.779	<u>5.862</u>	0.154	0.255	0.018	0.187	0.070	0.150	0.160	0.812	0.888	0.848
qwen-3 4B + KB	1	64.860	15.813	1.788	6.820	0.340	0.446	0.138	0.385	0.238	0.340	0.340	0.886	0.916	0.900
qwen-3 8B	7	24.271	6.928	0.783	14.869	0.194	0.315	0.034	0.228	0.103	0.186	0.193	0.832	<u>0.903</u>	0.866
qwen-3 8B + KB	4	41.472	10.326	1.168	15.550	0.283	0.444	0.099	0.323	0.195	0.274	0.278	0.861	0.923	0.891

Table 7: Results for open-generation QA performance and efficiency.

Mode/Model	Rank (Read.)	Direct. (↑)	Inverse (↓)									
		FRE	FKG	Fog	SMOG	ARI	CLI	Linsear	DaleChall	DiffWords	Syllables	Sentences
Reference Answer	—	24.372	14.649	18.125	15.609	16.213	16.041	14.644	12.750	8.768	37.758	1.020
gemma-3 4B	2	14.722	14.376	18.083	14.204	15.752	17.547	<u>10.121</u>	13.625	17.261	72.975	2.738
gemma-3 4B + KB	1	40.813	11.419	13.768	11.950	14.751	13.581	8.241	11.604	6.824	47.025	1.856
gemma-3 12B	3	11.087	15.314	19.156	15.169	17.170	18.286	11.864	13.847	13.894	57.710	2.026
gemma-3 12B + KB	6	11.111	17.214	21.386	18.093	18.942	17.929	18.174	13.548	22.554	103.978	2.400
llama-3.2 1B	4	<u>20.417</u>	15.620	19.301	16.943	17.098	<u>15.907</u>	14.958	<u>12.583</u>	55.222	320.200	7.498
llama-3.1 8B	7	10.917	17.136	21.252	17.766	18.889	17.643	17.446	13.631	31.276	149.127	3.994
llama-3.1 8B + KB	8	6.427	19.382	23.471	19.662	21.897	17.718	23.081	13.420	19.585	91.276	1.679
qwen-3 4B	5	13.860	14.808	18.562	15.158	17.236	17.974	11.175	14.129	37.562	181.045	7.071
qwen-3 4B + KB	10	-0.444	19.207	23.380	18.670	21.553	18.670	20.179	14.364	18.206	83.147	1.741
qwen-3 8B	9	5.589	16.790	20.963	17.041	19.112	19.348	14.524	14.371	38.100	167.348	5.278
qwen-3 8B + KB	11	4.440	17.723	21.918	18.132	19.816	19.623	17.264	14.104	30.860	136.227	3.385

Table 8: Model performance on text readability and complexity metrics.

making it a viable option. Nevertheless, Qwen’s strong performance was accompanied by higher latency and increased energy and CO₂eq impact compared, for instance, to Gemma-3, which, despite slightly higher VRAM usage, produced faster responses. This may be explained by Qwen’s reasoning-oriented design, leading to more tokens.

In the zero-shot setup, Gemma-3 12B achieved the best performance, followed by Llama-3 8B. Interestingly, Qwen-3 4B dropped significantly, suggesting that the knowledge required to answer may not be encoded in its weights but instead leveraged through retrieval. This also contrasts with the findings of He et al. (2025), who reported Llama-3 8B, Gemma-3 12B, and Qwen-3 4B, in that order, as the top-performing variants of these families on ES-Genius. These results indicate that OQA evaluation can reveal new nuances in model performance.

Table 8 provides a complementary perspective by focusing on the readability of the answers. While the Gemma-3 family consistently shows stronger readability indicators, Qwen models rank lowest. This may relate to their post-training emphasis on reasoning-oriented tasks such as mathematics, potentially reducing readability. Thus, the Gemma-3 family emerges as a promising candidate for QA generation, balancing efficiency and readability.

However, the findings derived from our dataset

also reveal that the evaluated models still present room for improvement. While semantic similarity is captured, as shown by BERTScore, other measures expose gaps in assertiveness, which our dataset makes evident, particularly in zero-shot scenarios. This is especially relevant considering that the evaluated models are among the most widely accessible, given their manageable size and open weights distribution. Altogether, the results indicate the utility of our dataset for uncovering important needs in the ESG domain.

6. Conclusions

This work presented ESG-QA, a new resource for ESG evaluation through QA. We also introduced our QAG pipeline, which combines LLMs as generators, PLMs applied as filters, regex-based rules, and human inspection. Using this resource, we evaluated three families of LLMs under two setups, zero-shot and retrieval-enhanced. In addition, we conducted an evaluation aligned with responsible AI principles, incorporating environmental impact and readability. Our findings highlight the Gemma-3 family as achieving a balance of correctness, efficiency, and readability. By making ESG-QA and its evaluation framework available, we aim to foster future research on ESG-oriented QA and to support

the broader community in building more transparent, efficient, and responsible language models.

7. Limitations

Although we manually inspected a sample of the generated QA pairs, we did not conduct a full evaluation of the entire corpus. Human validation was performed on sampled subsets, and residual artifacts or biases from the LLM-based generation pipeline may remain. Our analysis focused only on QA pairs retained after filtering, potentially excluding relevant examples, highlighting the need for improved filtering and validation strategies.

Because the dataset is automatically generated, it may reflect stylistic or structural patterns inherited from the underlying models rather than naturally occurring human-authored QA data. Moreover, our evaluation relied primarily on quantitative metrics, with limited systematic qualitative analysis of naturalness, usefulness, and contextual adequacy. Future work should incorporate deeper qualitative assessment and examine the impact of retrieval effectiveness and context relevance on QA performance.

ESG-QA is intended primarily as a benchmarking resource for domain-specific LLM evaluation rather than as a direct end-user educational tool. The current study is limited to English and to three open-weight model families (up to 12B parameters). Expanding to additional languages, larger models, and broader architectural families remains important future work.

Despite these limitations, our goal was to provide a scalable resource and reproducible framework to support further research on ESG-oriented QA and responsible model evaluation.

8. Ethical Considerations

The ESG-QA dataset raises some ethical considerations. As it was derived from corporate reports, social responsibility documents, and news articles, it may inherit biases from these sources, including cultural or regional imbalances and the overrepresentation of certain ESG perspectives. Although partisan political content was filtered to maintain neutrality, ideological framings and implicit assumptions about sustainability, governance, or social responsibility may persist. The dataset is intended as a benchmark to foster research on ESG-oriented language technologies, and it should not be used to draw conclusions about individual companies or to inform investment recommendations, as such applications go beyond its scope and could be misleading without human expert oversight. Furthermore, although the generation process combined automated filtering with partial human inspection,

residual inaccuracies and biases may persist in the generated QA pairs.

Acknowledgements

This research was supported by CNPq (307088/2023-5); FAPERJ (SEI-260003/002930/2024, SEI-260003/000614/2023); and CAPES (Finance Code 001). The authors would also like to thank Instituto de Ciência e Tecnologia Itaú (ICTi). Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Itaú Unibanco and Instituto de Ciência e Tecnologia Itaú. This document is not and does not constitute or intend to constitute investment advice or any investment service. It is not and should not be deemed to be an offer to purchase or sell, or a solicitation of an offer to purchase or sell, or a recommendation to purchase or sell any securities or other financial instruments. All data used in this study comply with the Brazilian General Data Protection Law. ChatGPT (OpenAI) assisted with grammar and LaTeX, with all suggestions reviewed by the authors, who take full responsibility for the final content.

9. Bibliographical References

- Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. [Synthetic QA Corpora Generation with Roundtrip Consistency](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173, Florence, Italy. Association for Computational Linguistics.
- Mohammed Ali, Abdelrahman Abdallah, and Adam Jatowt. 2025. [Sustainableqa: A comprehensive question answering dataset for corporate sustainability and eu taxonomy reporting](#).
- Dogu Araci. 2019. [FinBert: Financial Sentiment Analysis with Pre-trained Language Models](#).
- Artificial-Analysis. 2025. Qwen3-32b instruct eval. <https://artificialanalysis.ai/models/qwen3-32b-instruct-reasoning?intelligence-tab=intelligence>. Accessed: 2025-10-13.
- Gabriel Assis, Cláudia Freitas, and Aline Paes. 2025. [Exploring Brazil's LLM Fauna: Investigating the Generative Performance of Large Language Models in Portuguese](#). *Journal of the Brazilian Computer Society*, 31(1):940–972.

- Gabriel Assis, Daniela Vianna, Gisele L. Pappa, Alexandre Plastino, Wagner Meira Jr, Altigran Soares da Silva, and Aline Paes. 2024. [Analysis of material facts on financial assets: A generative AI approach](#). In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing*, pages 103–118, Torino, Italia. Association for Computational Linguistics.
- Nishant Balepur, Rachel Rudinger, and Jordan Lee Boyd-Graber. 2025. [Which of these best describes multiple choice evaluation with LLMs? a\) forced B\) flawed C\) fixable D\) all of the above](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3394–3418, Vienna, Austria. Association for Computational Linguistics.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20*, Red Hook, NY, USA. Curran Associates Inc.
- Semen Budenny, Vladimir Lazarev, Nikita Zakharenko, Alexey Korovin, Olga Plosskaya, Denis Dimitrov, Vladimir Arkhipkin, Ivan Oseledets, Ivan Barsola, Ilya Egorov, Aleksandra Kosterina, and Leonid Zhukov. 2023. [eco2AI: Carbon Emissions Tracking of Machine Learning Models as the First Step Towards Sustainable AI](#). *Doklady Mathematics*.
- Chung-Chi Chen, Chin-Yi Lin, Cheng-Chih Chiu, Hen-Hsen Huang, Alaa Alhamzeh, Yu-Lieh Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2025. Overview of the NTCIR-18 Finarg-2 Task: Temporal Inference of Financial Arguments. In *18th NTCIR Conference on Evaluation of Information Access Technologies*. NII Institutional Repository.
- Jian Chen, Peilin Zhou, Yining Hua, Loh Xin, Kehui Chen, Ziyuan Li, Bing Zhu, and Junwei Liang. 2024. [FinTextQA: A dataset for long-form financial question answering](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6025–6047, Bangkok, Thailand. ACL.
- Zhiyu Chen, Wenhui Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. [FinQA: A dataset of numerical reasoning over financial data](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic. ACL.
- Wenqi Fan, Yajuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. [A survey on RAG meeting LLMs: Towards retrieval-augmented large language models](#). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '24*, page 6491–6501, New York, NY, USA. Association for Computing Machinery.
- Gunnar Friede, Timo Busch, and Alexander Bassen. 2015. [ESG and financial performance: aggregated evidence from more than 2000 empirical studies](#). *Journal of Sustainable Finance & Investment*, 5(4):210–233.
- Gemma-Team. 2025. [Gemma 3 Technical Report](#).
- Xuwen Han, Neng Wang, Shangkun Che, Hongyang Yang, Kunpeng Zhang, and Sean Xin Xu. 2024. [Enhancing investment analysis: Optimizing ai-agent collaboration in financial research](#). In *Proceedings of the 5th ACM International Conference on AI in Finance, ICAIF '24*, page 538–546, New York, NY, USA. Association for Computing Machinery.
- Chaoyue He, Xin Zhou, Yi Wu, Xijia Yu, Yan Zhang, Lei Zhang, Di Wang, Shengfei Lyu, Hong Xu, Xiaoqiao Wang, Wei Liu, and Chunyan Miao. 2025. [Esgenius: Benchmarking llms on environmental, social, and governance \(esg\) and sustainability knowledge](#).
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. [DeBERTa: Decoding-enhanced BERT with disentangled attention](#). In *Proceedings of the International Conference on Learning Representations (ICLR)*. OpenReview.net.
- Lujun Li, Lama Sleem, Niccolo' Gentile, Geoffrey Nichil, and Radu State. 2025. [Exploring the impact of temperature on large language models: Hot or cold?](#) *Procedia Computer Science*, 264:242–251. International Neural Network Society Workshop on Deep Learning Innovations and Applications 2025.

- Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. [Www'18 open challenge: Financial opinion mining and question answering](#). In *Companion Proceedings of the The Web Conference 2018, WWW '18*, page 1941–1942, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.
- Rafik Mankour, Yassine Chafai, Hamada Saleh, Ghassen Ben Hassine, Thibaud Barreau, and Peter Tankov. 2025. [Climate finance bench](#).
- Leonardo Martinho, Hugo Dutra, Gabriel Assis, Jonnathan Carvalho, and Aline Paes. 2025. [When modernity enhances tradition and specialization: A bert ensemble for temporal financial argument detection](#). In *Anais do XIII Symposium on Knowledge Discovery, Mining and Learning*, pages 17–24, Porto Alegre, RS, Brasil. SBC.
- Spencer Mateega, Carlos Georgescu, and Danny Tang. 2025. [FinanceQA: A Benchmark for Evaluating Financial Analysis Capabilities of Large Language Models](#).
- Meta-AI. 2024. [The Llama 3 Herd of Models](#).
- Mistral-AI. 2023. [Mistral 7B](#). *CoRR*, abs/2310.06825.
- Rihab Najem, Ayoub Bahnasse, Meryem Fakhouri Amr, and Mohamed Talea. 2025. [Advanced AI and big data techniques in e-finance: a comprehensive survey](#). *Discover Artificial Intelligence*, 5(1):102.
- João Nepomuceno, Eduardo Rottschaefer, Gabriel Assis, Thiago Castro Ferreira, and Aline Paes. 2025. Agent ai meets fundamentalist financial analysis: Limits and possibilities. In *Proceedings of the 35th Brazilian Conference on Intelligent Systems (BRACIS 2025)*, Fortaleza, CE, Brazil.
- Felipe Paula, Cassiana Roberta Lizzoni Michelin, and Viviane Moreira. 2024. [Evaluation of Question Answer Generation for Portuguese: Insights and Datasets](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5315–5327, Miami, Florida, USA. Association for Computational Linguistics.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [SQuAD: 100,000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.
- Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, Michael Krumbick, Charles Lovering, and Chris Tanner. 2024. [DocFinQA: A long-context financial reasoning dataset](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 445–458, Bangkok, Thailand. ACL.
- Tobias Schimanski, Jingwei Ni, Roberto Spacey Martín, Nicola Ranger, and Markus Leippold. 2024a. [ClimRetrieve: A benchmarking dataset for information retrieval from corporate climate disclosures](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17509–17524, Miami, Florida, USA. Association for Computational Linguistics.
- Tobias Schimanski, Andrin Reding, Nico Reding, Julia Bingler, Mathias Kraus, and Markus Leippold. 2024b. [Bridging the gap in ESG measurement: Using NLP to quantify environmental, social, and governance communication](#). *Finance Research Letters*, 61:104979.
- Akchay Srivastava and Atif Memon. 2024. [Toward Robust Evaluation: A Comprehensive Taxonomy of Datasets and Metrics for Open Domain Question Answering in the Era of Large Language Models](#). *IEEE Access*, 12:117483–117503.
- Asahi Ushio, Fernando Alva-Manchego, and Jose Camacho-Collados. 2023. [An Empirical Comparison of LM-based Question and Answer Generation Methods](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14262–14272, Toronto, Canada. Association for Computational Linguistics.
- Xiaoqiang Wang, Bang Liu, Siliang Tang, and Lingfei Wu. 2022. [QReLScore: Better Evaluating Generated Questions with Deeper Understanding of Context-aware Relevance](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 562–581, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. [Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference](#).
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu,

Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. [C-pack: Packaged resources to advance general chinese embedding](#).

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chu-jie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Kekun Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. [Qwen3 technical report](#).

Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. 2024. [Natural Language Reasoning, A Survey](#). *ACM Comput. Surv.*, 56(12).

Chong Zhang, Xinyi Liu, Zhongmou Zhang, Mingyu Jin, Lingyao Li, Zhenting Wang, Wenyue Hua, Dong Shu, Suiyuan Zhu, Xiaobo Jin, Sujian Li, Mengnan Du, and Yongfeng Zhang. 2024. [When ai meets finance \(stockagent\): Large language model-based stock trading in simulated real-world environments](#).

Lei Zhang, Xin Zhou, Chaoyue He, Di Wang, Yi Wu, Hong Xu, Wei Liu, and Chunyan Miao. 2025. [Mmesgbench: Pioneering multimodal understanding and complex reasoning benchmark for esg tasks](#). To appear in Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), Dublin, Ireland.

10. Language Resource References

Strative AI. 2024. [SFinD-S Dataset](#). Strative. Hugging Face Datasets.

11. Appendices

A. Example of a QA pair removed by our filters

Question

What was one of the key goals of the work described in the context?

Answer

One of the key goals was undoing the damage of the Trump years.

Context from ESGBERT corpus

Working with community and Indigenous-led coalitions, we focused on undoing the damage of the Trump years and pursuing public lands solutions to climate change, species extinction, and expanding equitable access to nature.