

ESGenius: Benchmarking LLMs on Environmental, Social, and Governance (ESG) and Sustainability Knowledge

Chaoyue He¹ Xin Zhou¹ Yi Wu¹ Xijia Yu¹ Yan Zhang¹ Lei Zhang¹
Di Wang¹ Shengfei Lyu¹ Hong Xu¹ Xiaoqiao Wang² Wei Liu² Chunyan Miao¹

¹Alibaba-NTU Global e-Sustainability CorpLab (ANGEL), Singapore, ²Alibaba Group, China

Abstract

We introduce **ESGenius**, a comprehensive benchmark for evaluating and enhancing the proficiency of Large Language Models (LLMs) in Environmental, Social and Governance (ESG) and sustainability-focused question answering. **ESGenius** comprises two key components: (i) **ESGenius-QA**, a collection of **1,136** MCQs generated by LLMs and rigorously validated by domain experts, covering a broad range of ESG pillars and sustainability topics. Each question is systematically linked to its corresponding source text, enabling transparent evaluation and supporting Retrieval-Augmented Generation (RAG) methods; and (ii) **ESGenius-Corpus**, a meticulously curated repository of **231** foundational frameworks, standards, reports, and recommendation documents from **7** authoritative sources. Moreover, to fully assess the capabilities and adaptation potential of the model, we implement a rigorous two-stage evaluation protocol—*Zero-Shot* and *RAG*. Extensive experiments across **50** LLMs (ranging from 0.5B to 671B parameters) demonstrate that state-of-the-art models achieve only moderate performance in zero-shot settings, with accuracies typically around 55-70%, highlighting ESGenius’s challenging nature for LLMs in interdisciplinary contexts. However, models employing RAG demonstrate significant performance improvements, particularly for smaller models. For example, “DeepSeek-R1-Distill-Qwen 14B” improves from 63.82% (zero-shot) to 80.46% with RAG. These results demonstrate the necessity of grounding responses in authoritative sources for enhanced ESG understanding. To the best of our knowledge, ESGenius is the first benchmark curated for LLMs and the relevant enhancement technologies, focusing on ESG and sustainability topics. We publicly release the full ESGenius benchmark suite at <https://github.com/ANGEL-NTU/ESGenius>, including documentation, evaluation code, model implementations, and the dataset. An interac-

tive web portal with a real-time leaderboard and heatmap visualizations is available at <https://angel-ntu.github.io/ESGenius/>.

1 Introduction

Environmental, Social, and Governance (ESG) knowledge encompasses a vast domain of sustainability and corporate responsibility information that LLMs must effectively process to serve emerging business needs. This field spans across critical agendas from climate change and emissions tracking (IPCC, 2023), to workplace safety and human rights (Global Reporting Initiative, 2023). Such knowledge is codified in numerous technical frameworks and standards—major ones including GRI (Initiative, 2023), SASB (Board, 2023), TCFD (Board, 2017), ISSB (International Sustainability Standards Board, 2023), and CDP (CDP Worldwide, 2023)—which are constantly evolving.

While LLMs show promise in processing complex ESG documents and providing relevant answers to user queries, their capabilities in this interdisciplinary domain remain *largely unevaluated*. Considering the high stakes involved, where incorrect responses about ESG requirements or metrics could lead to serious compliance violations or misguided sustainability initiatives, this lack of assessment is particularly problematic. However, there exists **no comprehensive question-answering (QA) benchmark specifically designed to evaluate how well LLMs understand and reason about ESG concepts**. Most existing QA datasets either completely omit ESG topics or address them superficially. This gap leaves researchers and practitioners without a reliable way to measure and improve LLMs’ ESG knowledge comprehension and question-answering abilities.

To close this gap, we present **ESGenius**, a curated benchmark that targets MCQ-QA as the core evaluation task for ESG and sustainability knowledge. Our contributions are fourfold: (1) **ESGe-**

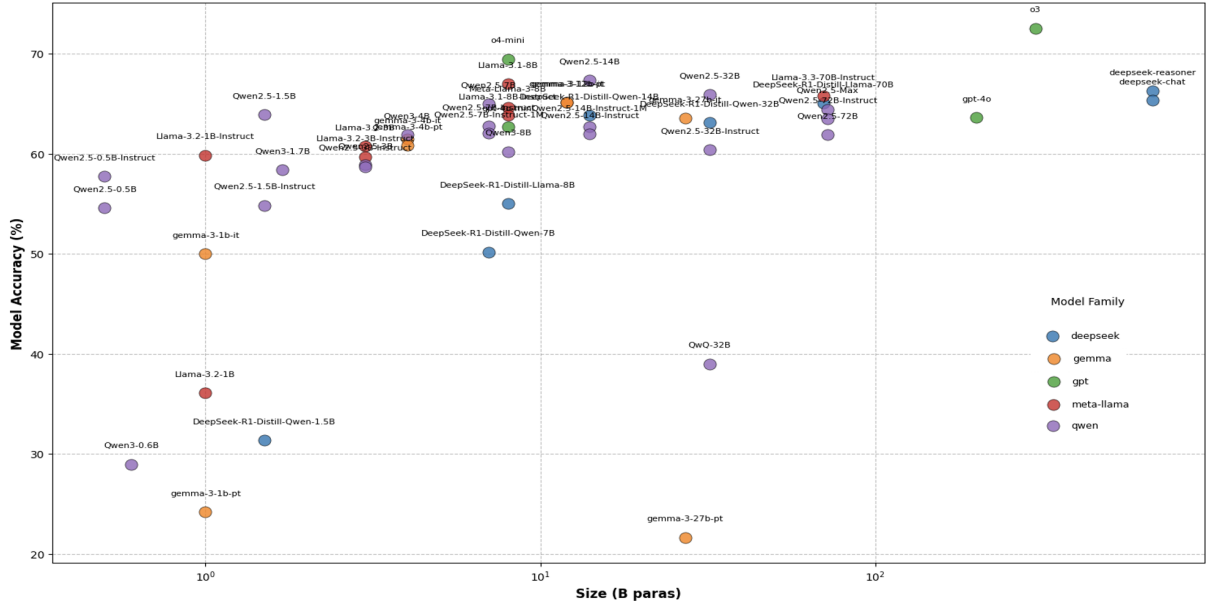


Figure 1: Relationship between model size and zero-shot accuracy across 50 LLMs evaluated on the ESGenius benchmark. Model sizes are plotted on a \log_{10} scale (in billions of parameters), with accuracy shown as percentages. There is a moderate positive correlation between model size and performance, suggesting larger models generally perform better. Dot colours denote 5 model families (DeepSeek, Gemma, GPT, Meta-Llama, Qwen). For proprietary API models, parameter counts are based on industry estimates (see Table 2 for details).

nius Benchmark: A comprehensive evaluation framework comprising two integrated components: (i) **ESGenius-QA**, a collection of 1,136 MCQs across various ESG pillars and sustainability topics, generated through SOTA LLM approaches and validated by domain experts. Each question is explicitly mapped to supporting evidence from authoritative source texts, enabling transparent evaluation and facilitating RAG applications; and (ii) **ESGenius-Corpus**, a collection of 231 ESG documents and frameworks enabling efficient knowledge retrieval from 7 major authoritative sources. (2) **Evaluation Protocol:** We implement a comprehensive two-stage evaluation protocol consisting of zero-shot testing and RAG to systematically assess LLM capabilities. This assessment framework provides valuable insights into the current limitations and future potential of LLMs in understanding ESG. (3) **Evaluation Analysis:** Testing across a diverse set of 50 LLMs (ranging from 0.5B to 671B parameters) reveals significant performance gaps in zero-shot settings, with most models achieving 55-70% accuracy and the best model (o3) achieves a top score of 72.54%, as shown in Figure 1. However, models demonstrate substantial potential for improvement through RAG approaches, with “DeepSeek-R1-Distill-Qwen 14B” improving from 63.82% (zero-shot) to 80.46%. (4) **Open Source**

Initiative: To foster community engagement and collaborative advancement, we have made our complete benchmark suite publicly available at <https://github.com/ANGEL-NTU/ESGenius>. This includes comprehensive documentation, evaluation code, model implementations, and the full ESGenius dataset. We also maintain an interactive web portal at <https://angel-ntu.github.io/ESGenius/> featuring a real-time leaderboard and detailed performance visualizations through heatmaps (see Appendix Figures 7 and 8), enabling researchers to track progress and identify areas for improvement in ESG-focused language models.

The remainder of this paper is organized as follows: Section 2 surveys prior efforts on ESG knowledge resources, question-answering benchmarks, and retrieval-augmented generation; Section 3 details the construction of the **ESGenius** benchmark, describing both the **ESGenius-QA** dataset and the **ESGenius-Corpus**; Section 4 outlines our experimental protocols—including zero-shot and RAG settings—and presents a comprehensive evaluation of 50 LLMs; Section 5 discusses the benchmark’s current limitations and avenues for future extension; and Section 6 concludes the paper and highlights directions for subsequent research.

2 Related Work

This section reviews relevant literature across three key areas in using LLMs for ESG benchmarking, including knowledge sources, QA benchmarks and RAG.

2.1 ESG & Sustainability Knowledge Resources

ESG data have traditionally been guided by voluntary reporting frameworks such as the Global Reporting Initiative (GRI) and the Sustainability Accounting Standards Board (SASB). These frameworks define taxonomies of ESG topics (e.g., emissions, labor practices, board governance) and indicators, but the unstructured text in corporate sustainability reports poses significant challenges for computational use. To impose structure, researchers have proposed ontologies and knowledge bases for ESG. For instance, (Zhou and Perzylo, 2023) introduced *OntoSustain*, aligning GRI and EU standards to capture key sustainability indicators, though broad coverage and automated population remain difficult.

Several ESG-focused text corpora have emerged to support NLP research in this domain. One recent example is *SusGen-30K*, a dataset of 30k instances spanning multiple financial NLP tasks (e.g., sentiment, Q&A) and an ESG report generation task (Wu et al., 2024). Meanwhile, (Chang et al., 2024) built a benchmark labeling sentences from corporate sustainability reports covering 36 ESG topics. These datasets remain relatively restricted, especially compared to general-domain corpora, and focus primarily on classification or generation tasks rather than MCQ question answering, thus motivating the need for broader coverage like **ES-Genius**.

Prior NLP work on sustainability texts has largely centered around information extraction and text classification. Early methods relied on topic models or lexicon-based approaches (Raghupathi and Raghupathi, 2020), but these often fell short in capturing nuanced ESG terminology. More recent efforts leverage LLMs. For instance, (Bronzini et al., 2023) employed a GPT-based pipeline to extract triples (*company*, *action*, *ESG issue*) from sustainability disclosures, populating a knowledge graph covering 500+ ESG topics. Similarly, (Marodon et al., 2022) unveils *SDG Prospector*, an AI-driven DistilRoBERTa NLP tool that mines 237 public development banks’ 2016-2020 reports to

expose which Sustainable Development Goals dominate their strategic narrative.

Researchers have also developed specialized language models for ESG. (Webersinke et al., 2021) introduced *ClimateBERT*, continuing BERT’s pre-training on 1.6M climate-related paragraphs, enabling more effective classification of risk disclosures. In industry, (Ni et al., 2023) proposed *ChatReport*, which uses GPT-based summarization and semantic search to assess a company’s report against TCFD recommendations. These domain-specific approaches illustrate the promise of LLMs for ESG tasks but also highlight the need for standardized benchmarks for a thorough evaluation.

2.2 QA & Evaluation Benchmarks

Question answering (QA) benchmarks have driven many NLP advancements over the past decade, starting with the TREC QA tracks (Voorhees and Tice, 2001) and continuing through datasets like *SQuAD* (Rajpurkar et al., 2016). While these resources predominantly test general factual or reading comprehension, specialized QA datasets have emerged to measure higher-level reasoning and domain expertise. For example, *OpenBookQA* (Mihaylov et al., 2018) challenged models to combine a small “open book” of scientific facts with common sense, while *BoolQ* (Clark et al., 2019) featured yes/no queries requiring nuanced text understanding. Datasets such as *DROP* (Dua et al., 2019) emphasize discrete reasoning skills, involving counting or arithmetic over paragraphs.

Further expansions include large-scale evaluation suites like *MMLU* (Massive Multitask Language Understanding) (Hendrycks et al., 2021), covering dozens of subjects at college exam level, and domain-specific tasks such as *PubMedQA* (Jin et al., 2019) for biomedical research. Despite these efforts, there remains a dearth of QA benchmarks devoted to ESG and/or sustainability. Thus, **ES-Genius** fills in this gap with a large set of expert-validated MCQs spanning across environmental, social, and governance topics. By drawing from comprehensive frameworks (e.g., IPCC, GRI, SASB), **ESGenius** offers a high-stakes, real-world testbed that is specific to ESG.

2.3 RAG and Knowledge Grounding

When tackling specialized or dynamically updated domains, RAG has emerged as a powerful method for accurate and interpretable QA. Lewis et al. (2020) introduced *RAG*, an approach that com-

bines a parametric language model with a non-parametric memory of documents. This approach improved performance on knowledge-intensive QA tasks (e.g., NaturalQuestions) relative to purely parametric models, as it allows the LLM to ground responses in retrieved evidence.

Other retrieval-generation frameworks such as *Fusion-in-Decoder* (Izacard and Grave, 2021) and *REALM* (Guu et al., 2020) illustrated diverse strategies for fetching relevant text chunks to guide the model’s reasoning. Recently, these methods have been applied to high-stakes sectors including finance and climate science. For instance, Vaghefi et al. (2023) integrated GPT-4 with a 3,000-page IPCC report to answer climate change questions more accurately. Similarly, Ni et al. (2023) used semantic search to retrieve TCFD-related segments from a company’s sustainability report, prompting an LLM to summarize potential climate risks. This grounding is critical in ESG contexts, where traceability and authoritative references are essential.

3 The ESGenius Benchmark

Developing reliable AI tools for the complex ESG and sustainability domain requires high-quality, domain-specific data. Our **ESGenius benchmark** provides a comprehensive foundation to evaluate and enhance LLMs in this context. It is designed to encompass the key dimensions of ESG, drawing from globally recognized standards and authoritative sources. The benchmark comprises two interconnected components: **ESGenius-QA**, a challenging question-answering dataset covering environmental, social, and governance topics with explicit linkages to textual evidence; and **ESGenius-Corpus**, encompassing a diverse collection of authoritative sources in ESG and sustainability. The detailed coverage of sources and statistics are presented in Appendix D.

3.1 ESGenius-QA

QA Generation and Validation Principles. The **ESGenius-QA** dataset comprises 1,136 carefully curated multiple-choice questions (MCQs) with ground truth answers, validated by ESG domain experts. The questions systematically cover essential ESG topics across three core pillars: (i) **Environmental:** Climate change mitigation and adaptation, comprehensive carbon accounting (Scope 1, 2, 3 emissions), energy efficiency, water resource management, biodiversity conservation, waste re-

duction, and pollution control; (ii) **Social:** Labor practices and standards, human rights protection, diversity and inclusion initiatives, workplace health and safety, community engagement and impact assessment, data privacy and protection; (iii) **Governance:** Board structure and independence, executive compensation frameworks, business ethics and compliance, shareholder rights and engagement, enterprise risk management, and regulatory adherence. The dataset is generated through our specialized pipeline (detailed in Appendix A.1) and undergoes rigorous validation by domain experts to ensure accuracy and relevance, as described in 3.1.

Descriptions. The **ESGenius-QA** dataset is structured to reflect the complexity and diversity of real-world ESG assessments through a standardized multiple-choice format. Each question contains four options (A-D) with a single correct answer, plus a dedicated “Not sure” option (Z) to assess model uncertainty. This design enables systematic evaluation of both performance and confidence. The dataset carefully balances conciseness with domain-specific precision: questions and options employ precise ESG terminology while maintaining clarity and focus, with contextual prompts provided where necessary for essential background. The specialized lexicon throughout authentically represents real-world ESG assessment scenarios. Table 1 presents a detailed token-level analysis of questions, options, and source text. Key terminology distributions are visualized through word clouds for the source text (Figure 4), questions (Figure 2), and answer options (Figure 3).

Corpus-Question Mapping. Each question in **ESGenius-QA** is precisely mapped to relevant page(s) and text passages within authoritative source documents (GRI, SASB, IPCC, etc.) containing the information required for correct answers. ESG domain experts meticulously annotate these mappings, ensuring both accuracy and relevance. This explicit linking between questions and their supporting evidence in source texts enables effective RAG approaches.

QA Quality Control. To ensure the highest quality of questions and answers, we implement a rigorous multi-stage validation process. After initial generation through our LLM pipeline mentioned in Appendix A.1, each question is assigned to domain experts. These experts carefully validate that:

(1) the question-answer pair is correct according to authoritative sources, (2) the question is clearly worded and tests meaningful domain knowledge, (3) distractors are plausible but unambiguously incorrect, and (4) sufficient context is provided where needed. Questions failing any validation criteria are either refined or removed from the dataset.

3.2 ESGenius-Corpus

The **ESGenius-Corpus** underpins every evaluation in this work, unifying authoritative frameworks, corporate questionnaires, and peer-reviewed scientific assessments that span the full breadth of environmental, social, and governance (ESG) concerns. A quantitative snapshot of the collection appears in Table 3 for IPCC climate-science assessments, Table 4 for the GRI Standards, Tables 5–6 for SASB’s industry-specific disclosures, Tables 7–8 for IFRS/ISSB guidance, Table 9 for TCFD materials, Table 10 for CDP questionnaires, and Table 11 for the UN Sustainable Development Goals. These statistics confirm (i) broad topical coverage across all three ESG pillars, (ii) deep sector-level granularity via SASB industry standards, and (iii) an expanding climate-finance focus through IFRS-ISSB, TCFD, and CDP additions.

Sourcing and Collection. The corpus integrates sources along three tiers: (1) *Core reporting standards* include the Global Reporting Initiative (GRI) Standards (Initiative, 2023), Sustainability Accounting Standards Board (SASB) Standards (Board, 2023), and International Sustainability Standards Board (IFRS-ISSB) Standards (International Sustainability Standards Board, 2023); (2) *Specialized reports and frameworks* extend coverage with IPCC Assessment Reports (IPCC, 2023), TCFD Guidelines (Board, 2017), and CDP questionnaires (CDP Worldwide, 2023); and (3) *Overarching global targets* are provided by the UN Sustainable Development Goals (United Nations, 2015).

Extensibility. The architecture is deliberately modular, allowing rapid incorporation of emerging regulations (e.g., EU CSRD, ISSB updates), sector-specific guidance (e.g., sustainable finance taxonomies), and certification frameworks. This forward-looking design safeguards the benchmark’s relevance as the ESG landscape evolves, ensuring that ESGenius remains a reliable yardstick for measuring LLM progress on sustainability understanding.

Copyrights and Privacy Handling. The **ESGenius-Corpus** comprises publicly available documents and open-access materials. For proprietary standards and frameworks (e.g., GRI, SASB, CDP), we provide references to their official source locations, allowing users to access them directly, with only brief excerpts referenced for academic purposes. The corpus documentation explicitly lists included sources and provides detailed guidance on accessing external materials, ensuring reproducible research while adhering to intellectual property rights. To protect privacy, the corpus excludes all personal and sensitive information through a rigorous filtering process - sections containing individual employees, survey respondents, or contact information are systematically removed. The dataset’s anonymity is verified through both automated filters and comprehensive manual auditing.

4 Experiments and Results

Through **ESGenius**, we benchmark **50** LLMs spanning open-source checkpoints—DeepSeek (Guo et al., 2025), Meta-Llama 3 (Grattafiori et al., 2024), Google Gemma 3 (Team et al., 2025), and Alibaba Qwen 2.5/3 (Yang et al., 2024, 2025) from Hugging Face (Hugging Face, 2024) and Transformers library (Wolf et al., 2020)—as well as proprietary APIs (GPT, DeepSeek, Qwen-Max). All experiments run on a DGX A100 node (4 × 80 GB A100 GPUs) with fixed random seeds (SEED=42) for reproducibility. Our evaluation suite comprises two progressively stronger settings: (1) **Zero-shot prompting** (§4.1.1), which probes the intrinsic ESG knowledge encoded during generic pre-training, and (2) **Long Context RAG** (§4.1.2), a RAG baseline that prepends relevant evidence to each prompt as long context. The RAG experiments are conducted on 43 open-source models.

For each model–setting combination, we compute exact-match accuracy on single-answer questions. Since LLMs may return raw responses that do not directly conform to our required format, we perform response validation to ensure valid multiple-choice answers (see details in Appendix A.2). This produces comprehensive leaderboards in Table 2 for both zero-shot and RAG approaches (full results in Appendix C Table 2).

4.1 Evaluation Protocols

4.1.1 Zero-Shot Prompting

Inference Protocol. Each model is provided the canonical question template (see Appendix B) and an example zero-shot question is shown in Appendix A.3. Response generation employs greedy decoding (temperature=0, top_p=1, top_k=0), with a maximum sequence length of 1024 tokens.

Implementation. Models are evaluated in half precision (FP16) using standard batched inference on GPU. Results are logged incrementally with error handling to ensure robustness.

4.1.2 Long Context RAG

Retrieval Protocol. For each question, the pre-linked source text from **ESGenius-QA** is retrieved from the ESG knowledge corpus. This text serves as context, prepended to the prompt template (see Appendix B.2) and an example zero-shot question is shown in Appendix A.3. This simple but effective approach ensures that models have access to relevant domain knowledge during inference.

Implementation. The evaluation pipeline closely mirrors the zero-shot setup, with the key difference being the augmented prompts. We implement context injection by prepending the retrieved source text to each question prompt before feeding it to the model. The same decoding parameters (temperature=0, top_p=1) and evaluation metrics are used.

4.2 Main Results

Drawing from the experimental results presented in Table 2, we highlight three key findings and provide detailed analysis in the following sections.

4.2.1 Key Findings

(1) **ESG Concept Understanding Remains Challenging for Current LLMs.** Our evaluation reveals that the top-performing model, “o3”, attained an accuracy of 0.7254, whereas the majority of other models scored approximately 0.65. This performance level is notably comparable to, or lower than, that observed in benchmarks from other domains (Guha et al., 2023), underscoring the challenges current LLMs face in accurately comprehending and reasoning about complex ESG concepts. Further details are provided in Section 4.2.2.

(2) **Smaller Models Augmented with Domain-Specific RAG Can Surpass Larger Zero-Shot Models.** Performance evaluation demonstrates that

applying RAG significantly enhances the capabilities of smaller models on ESG question answering. For example, the accuracy of “Qwen3 4B” increased from 0.6188 (zero-shot) to 0.7905 using RAG. Similarly, “Qwen2.5 1.5B” rose from 0.5484 (zero-shot) to 0.6972. These improvements demonstrate that targeted retrieval is more effective than raw model scale for complex domain-specific knowledge.

(3) **Reasoning Capabilities Enhance LLM Performance on ESG Understanding.** Models augmented with explicit reasoning capabilities demonstrate enhanced performance on ESG tasks compared to non-reasoning models of comparable scale. For instance, the reasoning variant of “o3” attained an accuracy of 0.7254, exceeding the scores of similarly sized models without such enhancements (see Figure 1). This underscores the significant benefit derived from incorporating reasoning-focused mechanisms into LLMs for specialized domains. More details in Section 4.2.4.

4.2.2 Challenging Nature of ESGenius

Our comprehensive evaluation, detailed in Table 2 (see Appendix C), reveals that the ESGenius-QA dataset is highly challenging for a broad range of contemporary LLMs. This includes proprietary APIs such as OpenAI’s GPT-4o (Zero-shot: 0.6364), DeepSeek models (“DeepSeek-R1” Zero-shot: 0.6629; “DeepSeek-V3” Zero-shot: 0.6532), and Alibaba’s Qwen2.5-Max (Zero-shot: 0.6444), as well as leading open-source families like Meta Llama (e.g., “Meta-Llama-3 8B” Zero-shot: 0.6461), Google Gemma (e.g., “Gemma-3 12B with instruction tuning” Zero-shot: 0.6514), and Alibaba Qwen (e.g., “Qwen2.5 14B” Zero-shot: 0.6734). The best zero-shot performance in the table is 0.7254 from the proprietary “o3” model, while the highest RAG-enhanced score is 0.8336 with the open-source “Gemma-3 27B with instruction tuning” model. This gap underscores both the discriminative nature of ESGenius and the benchmark’s heightened ESG reasoning demands.

4.2.3 RAG Performance

Integrating ESG-specific knowledge through long context RAG frequently yields substantial performance gains over zero-shot capabilities. For instance, the “DeepSeek-R1-Distill-Qwen” series demonstrates significant improvements across model sizes: the 1.5B model improves by 37.36% (0.3134 to 0.4305), the 7B model by 29.63%

(0.5018 to 0.6505), and the 14B model by 26.07% (0.6382 to 0.8046). Other notable examples include “Gemma-3 27B” (non-instruction-tuned), which achieves a remarkable 141.57% improvement (0.2165 to 0.5229), and “QwQ 32B” with a substantial 95.23% increase (0.3900 to 0.7614). Similarly, “Qwen3 4B” improves by 27.75% (0.6188 to 0.7905), and the instruction-tuned “Qwen2.5 1.5B” model increases by 27.13% (0.5484 to 0.6972). However, RAG does not universally enhance performance. Some models actually perform worse under the RAG configuration in this study—notably, “Qwen2.5 0.5B” (standard) shows a slight decline of 1.14% (0.5458 to 0.5396), while its instruction-tuned counterpart experiences a more significant 7.48% decrease (0.5775 to 0.5343). These results indicate that RAG’s effectiveness is closely tied to model architecture and intrinsic capabilities, though it generally enables smaller models to surpass the zero-shot performance of their larger counterparts.

4.2.4 Reasoning Models Analysis

Table 2 flags “Reasoning Focus” (Rea: Yes) models. Those explicitly oriented or fine-tuned for reasoning, such as “DeepSeek-R1-Distill-Qwen” and “DeepSeek-R1-Distill-Llama” (Open Source) and proprietary offerings like “o3” (300B, Rea: Yes, ZS: 0.7254) or “o4-mini” (3B, Rea: Yes, ZS: 0.6945), frequently excel in zero-shot mode. For example, the reasoning-oriented “DeepSeek-R1-Distill-Qwen” series shows strong RAG performance and gains (14B model: 0.8046, 1.5B model improves by 37.36%), while “QwQ 32B” (Rea: Yes) also achieves a 95.23% improvement under RAG. Comparisons with non-reasoning peers reinforce this pattern. For instance, “o4 mini” (3B, Rea: Yes) reaches 0.6945 zero-shot, well above non-reasoning 3B models such as “Llama-3.2 3B” (ZS: 0.6074) or “Qwen2.5 3B” (ZS: 0.5889). These observations suggest that training or architectural choices emphasizing reasoning bolster ESG question-answering, especially for multi-step logical tasks.

4.2.5 Instruct-Tuned Models Analysis

Our analysis of instruction-tuned models (marked “I-T: Yes” in Table 2) reveals varied performance in zero-shot settings. Several models demonstrate clear benefits from instruction tuning: Google’s “Gemma-3 1B” improves substantially from 0.2421 to 0.5000, while “Gemma-3 27B” shows an

even more dramatic increase from 0.2165 to 0.6356. Similar positive trends appear in Alibaba’s “Qwen2.5 0.5B” (0.5458 to 0.5775) and Meta’s “Llama-3.2 1B” (0.3609 to 0.5986). However, instruction tuning can also lead to performance degradation in some cases: “Meta Llama-3.1 8B” declines from 0.6699 to 0.6382, “Qwen2.5 1.5B” drops from 0.6391 to 0.5484, and “Qwen2.5 14B” decreases from 0.6734 to 0.6197. The impact of instruction tuning becomes particularly interesting when combined with RAG. In this context, instruction-tuned models frequently demonstrate superior improvement margins compared to their standard counterparts. For instance, the instruction-tuned version of “Gemma-3 12B” achieves 0.8380 (28.64% improvement) while its non-instruction-tuned variant reaches only 0.6857 (5.26% improvement). Similarly, the instruction-tuned “Llama-3.1 8B” attains 0.7993 (25.24% improvement) compared to the standard model’s 0.7650 (14.20% improvement). While instruction tuning may not consistently enhance zero-shot performance, its impact becomes particularly interesting when combined with RAG. In this context, instruction-tuned models frequently demonstrate superior improvement margins compared to their standard counterparts.

4.2.6 Error Analysis

Sorting the heat-map (see Section 4.2.7) columns by failure rate exposes a long tail of adversarial-grade questions that no model answered correctly. Appendix C.1 reproduces three such zero-accuracy questions. Question 432 from the *SDG 2024 Report* probes subtle regional disparities in per-capita renewable electricity capacity. Question 635, drawn from *IPCC AR6 WG III*, asks for the single most direct barrier to ESCO adoption; models consistently mis-attributed the bottleneck to regulation or awareness, missing the report’s emphasis on asymmetric information and split incentives within firms. Question 1006, based on the *SASB Chemicals* standard, tests the definition of a recordable workplace incident. Two broad error patterns emerge across these and similar questions: first, a tendency to rely on surface keyword overlap between prompt and distractor, and second, over-generalisation to well-known policy frameworks when a question hinges on a narrower ESG standard. These observations align with the substantial performance gains reported under RAG (§4.2), underscoring the necessity of grounding model outputs in the correct source material.

4.2.7 Interactive Visualization

Figure 7 presents an interactive heat-map visualization of model performance across 50 LLMs (ranging from 0.5 to 671 billion parameters) and all 1,136 MCQs in **ESGenius**. The visualization arranges models (rows) by their overall zero-shot accuracy and questions (columns) by difficulty, creating a clear left-to-right progression from predominantly red cells (incorrect/unanswered questions) to green cells (correct answers). Interactive features enhance analysis - hovering over any cell reveals detailed information including the model name and rank, question ID, difficulty score, complete prompt, answer options, ground truth, and the model's prediction. An example of this interactive tooltip is shown in Figure 8. This dynamic visualization transforms raw accuracy data into an intuitive diagnostic tool for analyzing model performance patterns and systematic weaknesses. The interactive visualization is publicly accessible at <https://angel-ntu.github.io/ESGenius/>.

5 Limitations

While **ESGenius** aspires to provide a holistic evaluation of ESG understanding, several caveats must be acknowledged:

- *Coverage trade-offs*—the benchmark covers only seven sources, which may not comprehensively cover all ESG topics and sustainability issues. While these sources represent key frameworks and standards, they exclude many specialized industry guidelines, regional regulations, and emerging sustainability frameworks.
- *Expert dependency*—crafting high-fidelity questions and validating answers hinges on scarce domain specialists, which introduces both capacity limits and potential disciplinary bias. The specialized nature of ESG expertise makes scaling the question creation process especially difficult.
- *Format constraints*—standardized multiple-choice questions facilitate large-scale scoring but compress the nuanced reasoning and synthesis often required in real-world ESG analysis. This format may not fully capture the complexity of real-world sustainability decision-making.
- *Standard volatility*—frameworks such as ISSB guidance and IPCC assessments evolve quickly, obliging continual corpus refreshes and version control to stay current. This rapid evolution makes maintaining benchmark relevance an ongoing challenge.
- *Limited language*—the present focus on English, globally dominant standards risks underrepresenting regional regulations and non-English corpora, potentially missing important local sustainability contexts and practices.
- *Modality gaps*—the text-only design excludes visual disclosures (e.g., charts, geospatial maps) and time-series data that are central to many ESG reports, limiting the benchmark's ability to test comprehension of multi-modalities.
- *Simplified scoring*—binary accuracy overlooks partial correctness and fails to credit sophisticated but incomplete reasoning, calling for finer-grained metrics that can better capture nuanced understanding.
- *Copyright limits*—licensing restrictions preclude the inclusion of certain proprietary ESG standards, thereby limiting the potential for truly exhaustive coverage and potentially missing key industry-specific frameworks.

6 Conclusion

We presented **ESGenius**, the first benchmark dedicated to probing LLMs on the breadth and depth of ESG and sustainability knowledge, unifying authoritative standards, corporate questionnaires, and scientific assessments into a multi-dimensional QA dataset. A two-stage evaluation protocol—*Zero-Shot* followed by *RAG*—reveals that grounding responses in curated evidence boosts accuracy by roughly 15–30 percentage points, and that 4 to 72 billion parameter models equipped with targeted knowledge can surpass much larger zero-shot models, underscoring the primacy of domain integration over sheer scale. Looking ahead, we aim to incorporate emerging regulations, automate corpus refreshes, extend coverage to non-English sources, and design finer-grained metrics that capture partial credit and nuanced ESG reasoning. Overall, **ESGenius** is designed to serve as a dedicated benchmark to facilitate the development of sustainability knowledge-aware AI systems in the complex ESG domain.

Acknowledgments

This research is supported by the RIE2025 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) (Award I2301E0026), administered by A*STAR, as well as supported by Alibaba Group and NTU Singapore through Alibaba-NTU Global e-Sustainability CorpLab (ANGEL).

References

2015. The 2030 agenda for sustainable development's 17 sustainable development goals (sdgs).
2015. Transforming our world: The 2030 agenda for sustainable development.
- 2016a. GRI 201: Economic performance 2016.
- 2016b. GRI 202: Market presence 2016.
- 2016c. GRI 203: Indirect economic impacts 2016.
- 2016d. GRI 204: Procurement practices 2016.
- 2016e. GRI 205: Anti-corruption 2016.
- 2016f. GRI 206: Anti-competitive behavior 2016.
- 2016g. GRI 301: Materials 2016.
- 2016h. GRI 302: Energy 2016.
- 2016i. GRI 304: Biodiversity 2016.
- 2016j. GRI 305: Emissions 2016.
- 2016k. GRI 306: Effluents and waste 2016.
- 2016l. GRI 308: Supplier environmental assessment 2016.
- 2016m. GRI 401: Employment 2016.
- 2016n. GRI 402: Labor/management relations 2016.
- 2016o. GRI 404: Training and education 2016.
- 2016p. GRI 405: Diversity and equal opportunity 2016.
- 2016q. GRI 406: Non-discrimination 2016.
- 2016r. GRI 407: Freedom of association and collective bargaining 2016.
- 2016s. GRI 408: Child labor 2016.
- 2016t. GRI 409: Forced or compulsory labor 2016.
- 2016u. GRI 410: Security practices 2016.
- 2016v. GRI 411: Rights of indigenous peoples 2016.
- 2016w. GRI 413: Local communities 2016.
- 2016x. GRI 414: Supplier social assessment 2016.
- 2016y. GRI 415: Public policy 2016.
- 2016z. GRI 416: Customer health and safety 2016.
- 2016aa. GRI 417: Marketing and labeling 2016.
- 2016ab. GRI 418: Customer privacy 2016.
2018. GRI 303: Water and effluents 2018.
- 2018ab. GRI 403: Occupational health and safety 2018.
2019. GRI 207: Tax 2019.
2020. GRI 306: Waste 2020.
2020. Task force on climate-related financial disclosures guidance on risk management integration and disclosure.
- 2020eb. Task force on climate-related financial disclosures guidance on scenario analysis for non-financial companies.
2021. Consolidated Set of the GRI Standards 2021.
- 2021ab. GRI 1: Foundation 2021.
- 2021ac. GRI 11: Oil and gas sector 2021.
- 2021ad. GRI 2: General disclosures 2021.
- 2021ae. GRI 3: Material topics 2021.
- 2021af. The GRI standards — a guide for policy makers.
- 2021ag. A practical guide to sustainability reporting using GRI and SASB standards.
2021. Task force on climate-related financial disclosures guidance on metrics, targets, and transition plans.
- 2021eb. Task force on climate-related financial disclosures implementing the recommendations of the task force on climate-related financial disclosures.
2022. GRI 12: Coal sector 2022.
- 2022ab. GRI 13: Agriculture, aquaculture and fishing sectors 2022.
- 2022ac. GRI Standards Glossary 2022.
2022. Task force on climate-related financial disclosures overview.
- 2022eb. Tcfd workshop – session 1: Fundamentals and overview of tcfd.
- 2022ec. Tcfd workshop – session 2: Governance.
- 2022ed. Tcfd workshop – session 3: Strategy.
- 2022ee. Tcfd workshop – session 4: Risk management.
- 2022ef. Tcfd workshop – session 5: Metrics and targets.

2023. Comparison — IFRS S2 climate-related disclosures with the tcfD recommendations.
2023. GRI special report 2023 — mapping resilience for the sustainable development goals.
- 2023ab. IFRS S1 accompanying guidance on general requirements for disclosure of sustainability-related financial information.
- 2023ac. IFRS S1 basis for conclusions on general requirements for disclosure of sustainability-related financial information.
- 2023ad. IFRS S1: General requirements for disclosure of sustainability-related financial information.
- 2023ae. IFRS S2 basis for conclusions on climate-related disclosures.
- 2023af. IFRS S2: Climate-related disclosures.
- 2023ag. IFRS S2 industry-based guidance on implementing climate-related disclosures.
- 2023ah. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 1 — apparel, accessories & footwear.
- 2023ai. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 10 — metals & mining.
- 2023aj. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 11 — oil & gas — exploration & production.
- 2023ak. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 12 — oil & gas — midstream.
- 2023al. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 13 — oil & gas — refining & marketing.
- 2023am. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 14 — oil & gas — services.
- 2023an. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 15 — asset management & custody activities.
- 2023ao. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 16 — commercial banks.
- 2023ap. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 17 — insurance.
- 2023aq. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 18 — investment banking & brokerage.
- 2023ar. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 19 — mortgage finance.
- 2023as. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 2 — appliance manufacturing.
- 2023at. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 20 — agricultural products.
- 2023au. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 21 — alcoholic beverages.
- 2023av. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 22 — food retailers & distributors.
- 2023aw. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 23 — meat, poultry & dairy.
- 2023ax. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 24 — non-alcoholic beverages.
- 2023ay. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 25 — processed foods.
- 2023az. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 26 — restaurants.
- 2023ba. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 27 — drug retailers.
- 2023bb. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 28 — health care delivery.
- 2023bc. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 29 — health care distributors.
- 2023bd. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 3 — building products & furnishings.
- 2023be. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 30 — managed care.
- 2023bf. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 31 — medical equipment & supplies.
- 2023bg. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 32 — electric utilities & power generators.
- 2023bh. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 33 — engineering & construction services.
- 2023bi. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 34 — gas utilities & distributors.

- 2023bj. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 35 — home builders.
- 2023bk. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 36 — real estate.
- 2023bl. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 37 — real estate services.
- 2023bm. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 38 — waste management.
- 2023bn. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 39 — water utilities & services.
- 2023bo. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 4 — e-commerce.
- 2023bp. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 40 — biofuels.
- 2023bq. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 41 — forestry management.
- 2023br. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 42 — fuel cells & industrial batteries.
- 2023bs. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 43 — pulp & paper products.
- 2023bt. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 44 — solar technology & project developers.
- 2023bu. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 45 — wind technology & project developers.
- 2023bv. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 46 — aerospace & defense.
- 2023bw. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 47 — chemicals.
- 2023bx. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 48 — containers & packaging.
- 2023by. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 49 — electrical & electronic equipment.
- 2023bz. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 5 — household & personal products.
- 2023ca. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 50 — industrial machinery & goods.
- 2023cb. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 51 — casinos & gaming.
- 2023cc. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 52 — hotels & lodging.
- 2023cd. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 53 — leisure facilities.
- 2023ce. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 54 — electronic manufacturing services & original design manufacturing.
- 2023cf. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 55 — hardware.
- 2023cg. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 56 — internet media & services.
- 2023ch. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 57 — semiconductors.
- 2023ci. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 58 — software & it services.
- 2023cj. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 59 — telecommunication services.
- 2023ck. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 6 — multi-line and specialty retailers & distributors.
- 2023cl. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 60 — air freight & logistics.
- 2023cm. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 61 — airlines.
- 2023cn. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 62 — auto parts.
- 2023co. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 63 — automobiles.
- 2023cp. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 64 — car rental & leasing.
- 2023cq. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 65 — cruise lines.

2023cr. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 66 — marine transportation.

2023cs. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 67 — rail transportation.

2023ct. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 68 — road transportation.

2023cu. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 7 — coal operations.

2023cv. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 8 — construction materials.

2023cw. IFRS S2 industry-based guidance on implementing climate-related disclosures vol 9 — iron & steel producers.

2023. SASB standard — advertising & marketing 2023.

2023cb. SASB standard — aerospace & defense 2023.

2023cc. SASB standard — agricultural products 2023.

2023cd. SASB standard — air freight & logistics 2023.

2023ce. SASB standard — airlines 2023.

2023cf. SASB standard — alcoholic beverages 2023.

2023cg. SASB standard — apparel, accessories & footwear 2023.

2023ch. SASB standard — appliance manufacturing 2023.

2023ci. SASB standard — asset management & custody activities 2023.

2023cj. SASB standard — auto parts 2023.

2023ck. SASB standard — automobiles 2023.

2023cl. SASB standard — biofuels 2023.

2023cm. SASB standard — biotechnology & pharmaceuticals 2023.

2023cn. SASB standard — building products & furnishings 2023.

2023co. SASB standard — car rental & leasing 2023.

2023cp. SASB standard — casinos & gaming 2023.

2023cq. SASB standard — chemicals 2023.

2023cr. SASB standard — coal operations 2023.

2023cs. SASB standard — commercial banks 2023.

2023ct. SASB standard — construction materials 2023.

2023cu. SASB standard — consumer finance 2023.

2023cv. SASB standard — containers & packaging 2023.

2023cw. SASB standard — cruise lines 2023.

2023cx. SASB standard — drug retailers 2023.

2023cy. SASB standard — e-commerce 2023.

2023cz. SASB standard — education 2023.

2023da. SASB standard — electric utilities & power generators 2023.

2023db. SASB standard — electrical & electronic equipment 2023.

2023dc. SASB standard — electronic manufacturing services & original design manufacturing 2023.

2023dd. SASB standard — engineering & construction services 2023.

2023de. SASB standard — food retailers & distributors 2023.

2023df. SASB standard — forestry management 2023.

2023dg. SASB standard — fuel cells & industrial batteries 2023.

2023dh. SASB standard — gas utilities & distributors 2023.

2023di. SASB standard — hardware 2023.

2023dj. SASB standard — health care delivery 2023.

2023dk. SASB standard — health care distributors 2023.

2023dl. SASB standard — home builders 2023.

2023dm. SASB standard — hotels & lodging 2023.

2023dn. SASB standard — household & personal products 2023.

2023do. SASB standard — industrial machinery & goods 2023.

2023dp. SASB standard — insurance 2023.

2023dq. SASB standard — internet media & services 2023.

2023dr. SASB standard — investment banking & brokerage 2023.

2023ds. SASB standard — iron & steel producers 2023.

2023dt. SASB standard — leisure facilities 2023.

2023du. SASB standard — managed care 2023.

2023dv. SASB standard — marine transportation 2023.

2023dw. SASB standard — meat, poultry & dairy 2023.

- 2023dx. SASB standard — media & entertainment 2023.
- 2023dy. SASB standard — medical equipment & supplies 2023.
- 2023dz. SASB standard — metals & mining 2023.
- 2023ea. SASB standard — mortgage finance 2023.
- 2023eb. SASB standard — multiline and specialty retailers & distributors 2023.
- 2023ec. SASB standard — non-alcoholic beverages 2023.
- 2023ed. SASB standard — oil & gas – exploration & production 2023.
- 2023ee. SASB standard — oil & gas – midstream 2023.
- 2023ef. SASB standard — oil & gas – refining & marketing 2023.
- 2023eg. SASB standard — oil & gas – services 2023.
- 2023eh. SASB standard — processed foods 2023.
- 2023ei. SASB standard — professional & commercial services 2023.
- 2023ej. SASB standard — pulp & paper products 2023.
- 2023ek. SASB standard — rail transportation 2023.
- 2023el. SASB standard — real estate 2023.
- 2023em. SASB standard — real estate services 2023.
- 2023en. SASB standard — restaurants 2023.
- 2023eo. SASB standard — road transportation 2023.
- 2023ep. SASB standard — security & commodity exchanges 2023.
- 2023eq. SASB standard — semiconductors 2023.
- 2023er. SASB standard — software & it services 2023.
- 2023es. SASB standard — solar technology & project developers 2023.
- 2023et. SASB standard — telecommunication services 2023.
- 2023eu. SASB standard — tobacco 2023.
- 2023ev. SASB standard — toys & sporting goods 2023.
- 2023ew. SASB standard — waste management 2023.
- 2023ex. SASB standard — water utilities & services 2023.
- 2023ey. SASB standard — wind technology & project developers 2023.
2023. Sustainable development goals — briefing book 2023.
2023. Task force on climate-related financial disclosures 2023 status report.
2024. Cdp full corporate scoring introduction 2024.
2024. GRI 101: Biodiversity 2024.
- 2024ab. GRI 14: Mining sector 2024.
2024. IFRS sustainability disclosure taxonomy 2024 — IFRS S1 general requirements for disclosure of sustainability-related financial information and IFRS S2 climate-related disclosures.
- 2024cb. Progress on corporate climate-related disclosures — 2024 report.
2024. Progress towards the sustainable development goals — report of the secretary-general 2024.
- 2024eb. The sustainable development goals report 2024.
2024. Sustainable development report 2024 — the sdgs and the un summit of the future.
- 2025a. 2025 cdp–iclei track and states & regions questionnaire and guidance.
- 2025b. Cdp full corporate questionnaire april 2025 — module 7.
- 2025c. Cdp full corporate questionnaire april 2025 — modules 1–6.
- 2025d. Cdp full corporate questionnaire april 2025 — modules 8–13.
- 2025e. Cdp sme questionnaire april 2025 — modules 14–21.
2025. Exposure draft — amendments to greenhouse gas emissions disclosures (proposed amendments to IFRS S2; comments due 27 june 2025).
- 2025cb. Exposure draft — basis for conclusions on amendments to greenhouse gas emissions disclosures (proposed amendments to IFRS S2; comments due 27 june 2025).
- Financial Stability Board. 2017. Task force on climate-related financial disclosures recommendations. <https://www.fsb-tcfd.org/recommendations/>. Accessed May 2025.
- Sustainability Accounting Standards Board. 2023. Sasb standards. <https://sasb.org/standards/>. Accessed May 2025.
- Marco Bronzini, Carlo Nicolini, Bruno Lepri, Andrea Passerini, and Jacopo Staiano. 2023. Glitter or gold? deriving structured insights from sustainability reports via large language models. *arXiv preprint arXiv:2310.05628*.
- CDP Worldwide. 2023. Cdp climate change questionnaire guidance. <https://www.cdp.net/en/guidance/guidance-for-companies>. Accessed May 2025.

- Qi Chang, Xuan Yang, Zihan Ding, Bin Liu, and Wei Lan. 2024. An nlp benchmark dataset for predicting the completeness of esg reports.
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of NAACL-HLT*, pages 2924–2936.
- Dheeru Dua, Ananth Gottumukkala, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of NAACL-HLT*, pages 2368–2378.
- Mathieu Fenniak, Matthew Stamy, pubpub zz, Martin Thoma, Matthew Peveler, exiledkingcc, and PyPDF2 Contributors. 2022. [The PyPDF2 library](https://pypdf2.readthedocs.io/en/latest/meta/CONTRIBUTORS.html). See <https://pypdf2.readthedocs.io/en/latest/meta/CONTRIBUTORS.html> for the full contributor list.
- Global Reporting Initiative. 2023. Gri 400: Social standards series. Technical report, GRI.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, and 1 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36:44123–44279.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shitong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-augmented language model pre-training. In *Proceedings of ICML*, pages 3929–3938.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, and Akul Arora. 2021. Measuring Massive Multitask Language Understanding. In *Proceedings of ICLR*.
- Inc. Hugging Face. 2024. Hugging face hub. <https://huggingface.co>. Software library version 0.23.0.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Global Reporting Initiative. 2023. Gri standards. <http://www.globalreporting.org/how-to-use-the-gri-standards/gri-standards-english-language/>. Accessed May 2025.
- International Sustainability Standards Board. 2023. Ifrs sustainability disclosure standards: Ifrs s1 *General Requirements for Disclosure of Sustainability-related Financial Information* and ifrs s2 *Climate-related Disclosures*. <https://www.ifrs.org/news-and-events/news/2023/06/issb-issues-ifrs-s1-and-ifrs-s2/>. Accessed May 2025.
- IPCC. 2018. Global warming of 1.5 °c. an ipcc special report on the impacts of global warming of 1.5 °c above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty. Technical report, Intergovernmental Panel on Climate Change, Cambridge, UK and New York, NY, USA.
- IPCC. 2019. Climate change and land: an ipcc special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems. Technical report, Intergovernmental Panel on Climate Change, Geneva, Switzerland. In press.
- IPCC. 2019cb. *Ipcc special report on the ocean and cryosphere in a changing climate*. Technical report, Intergovernmental Panel on Climate Change, Cambridge, UK and New York, NY, USA. In press.
- IPCC. 2021. *Climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change*. Technical report, Intergovernmental Panel on Climate Change, Cambridge, United Kingdom and New York, NY, USA. In press.
- IPCC. 2022. *Climate change 2022: Impacts, adaptation, and vulnerability. contribution of working group ii to the sixth assessment report of the intergovernmental panel on climate change*. Technical report, Intergovernmental Panel on Climate Change, Cambridge, UK and New York, NY, USA.
- IPCC. 2022cb. *Climate change 2022: Mitigation of climate change. contribution of working group iii to the sixth assessment report of the intergovernmental panel on climate change*. Technical report, Intergovernmental Panel on Climate Change, Cambridge, UK and New York, NY, USA.
- IPCC. 2023. *Climate change 2023: Synthesis report. contribution of working groups i, ii and iii to the sixth assessment report of the intergovernmental panel on climate change*. <https://www.ipcc.ch/assessment-report/ar6/>. Accessed May 2025.
- IPCC. 2023cb. *Climate change 2023: Synthesis report. contribution of working groups i, ii and iii to the sixth assessment report of the intergovernmental panel on climate change*. Technical report, Intergovernmental Panel on Climate Change, Geneva, Switzerland.

- Gautier Izacard and Edouard Grave. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of EACL*, pages 874–880.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Dmitry Kislyuk, Myle Ott, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of NeurIPS*.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- Régis Marodon, Jean-Baptiste Jacouton, and Adeline Laulanie. 2022. The proof is in the pudding. revealing the sdgs with artificial intelligence. Technical report.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *Proceedings of EMNLP*, pages 2381–2391.
- Jingwei Ni, Julia Bingler, Chiara Colesanti-Senni, Mathias Kraus, Glen Gostlow, Tobias Schimanski, Dominik Stambach, Saeid Ashraf Vaghefi, Qian Wang, Nicolas Webersinke, and 1 others. 2023. Chatreport: Democratizing sustainability disclosure analysis through llm-based tools. *arXiv preprint arXiv:2307.15770*.
- Wullianallur Raghupathi and Viju Raghupathi. 2020. Analyzing sustainability disclosures using topic modeling. *Sustainability*, 12(10):4001.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP*, pages 2383–2392.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivi re, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.
- Qwen Team. 2025. *Qwq-32b: Embracing the power of reinforcement learning*.
- United Nations. 2015. Transforming our world: The 2030 agenda for sustainable development. <https://sdgs.un.org/2030agenda>. Accessed May 2025.
- Saeid Ashraf Vaghefi, Dominik Stambach, Veruska Muccione, Julia Bingler, Jingwei Ni, Mathias Kraus, Simon Allen, Chiara Colesanti-Senni, Tobias Wekhof, Tobias Schimanski, and 1 others. 2023. Chatclimate: Grounding conversational ai in climate science. *Communications Earth & Environment*, 4(1):480.
- Ellen M. Voorhees and Dawn M. Tice. 2001. The TREC-8 Question Answering Track Report. In *The Eighth Text REtrieval Conference (TREC-8)*.
- Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2021. Climatebert: A pretrained language model for climate-related text. *arXiv preprint arXiv:2110.12010*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Cl ment Delangue, Anthony Moi, Pi ric Cistac, Tim Rault, R mi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. *Trans-formers: 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.
- Qilong Wu, Xiaoneng Xiang, Hejia Huang, Xuan Wang, Yeo Wei Jie, Ranjan Satapathy, Bharadwaj Veeravalli, and 1 others. 2024. Susgen-gpt: A data-centric llm for financial nlp and sustainability report generation. *arXiv preprint arXiv:2412.10906*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, and 1 others. 2025eb. Qwen2.5-1m technical report. *arXiv preprint arXiv:2501.15383*.
- Alan Zhou and Martin Perzylo. 2023. Ontosustain: An ontology for esg and eu standards. In *International Semantic Web Conference (ISWC)*.

Appendix

A ESGenius-QA

A.1 Automated Preparation of Candidate MCQs

This appendix details the fully-automated preparation stage that converts raw knowledge sources into a pool of *candidate* multiple-choice questions

(MCQs). (The complete algorithm is detailed in Algorithm 1.) These candidate questions subsequently undergo expert review, editing, and validation before potential inclusion in **ESGenius-QA**. The goals of the preparation stage are twofold: (i) *breadth*—to cover as many distinct passages as possible from the source corpus, and (ii) *difficulty*—to generate questions that force LLMs to reason beyond surface-level facts.

Input corpus. Documents are first collected from authoritative sources in PDF format. Each PDF is then processed independently via the pipeline to enable precise tracking of corpus coverage. For encrypted PDFs, the pipeline attempts in-memory decryption where possible; if decryption fails, the file is skipped with a warning. Throughout processing, the system maintains detailed logs of extracted page counts and flags any instances of unsupported encryption formats.

Text extraction and chunking process. Using PyPDF2 (Fenniak et al., 2022), every page’s text is cleaned (whitespace collapse and control-character stripping) and stored in a page map. A *chunk selector* then samples passages that satisfy length constraints

$$150 \leq |\text{chunk}| \leq 4500 \text{ characters,}$$

optionally concatenating the next page when the combined length remains under the upper bound. Each chunk is tagged with its page span (e.g. “12–13”) so that provenance is preserved.

Prompted multiple-choice question generation. For every selected chunk the pipeline sends a carefully engineered prompt (see A.1) to the **Qwen Max’s DashScope** endpoint¹ that requests *exactly one* “extremely challenging” MCQ. Mandatory instructions enforce (i) deep reasoning across multiple sentences, (ii) near-miss distractors, (iii) answerability from the excerpt alone, and (iv) strict JSON output. A fixed seed (42) controls chunk sampling so that runs are reproducible.

Structural validation and metadata augmentation. The returned JSON is sanitised, parsed, and validated. Missing or malformed keys trigger rejection. The correct answer text is cross-checked with the declared key, and a fallback correction is applied if necessary. A universal **Z** op-

tion (“Not sure”) is appended to every item to enable abstention analysis. Each validated MCQ is wrapped in a metadata envelope containing {id, source_pdf, page_range, difficulty, generation_timestamp}.

Iterative difficulty refinement process. Because the pipeline aims for a target zero-shot accuracy of $\leq 50\%$ on an external evaluator (we use Qwen Max as the evaluator), it performs up to 15 iterative refinement rounds to achieve this threshold:

1. Test the current pool; label each MCQ with the model’s chosen answer and correctness.
2. Remove questions that the evaluator answers correctly.
3. Replace the removed questions by sampling new chunks and repeating the generation step, maintaining the original pool size.
4. Abort early if (a) accuracy falls below the target, (b) authentication errors reach the threshold (default 5), or (c) no new valid chunks remain.

Consistency checking and question deduplication. Beginning with round 2 the pipeline checks that an MCQ’s evaluator result is stable across rounds; inconsistent questions are discarded. After refinement, a pairwise LLM-based similarity check removes near-duplicate questions that probe the same underlying concept. Only the first occurrence is retained.

Output artifacts generation and storage. Two JSON files are written atomically after every processed PDF: `all_candidates_untested.json` stores *every* syntactically valid question, whereas `final_round_tested.json` retains only the post-refinement pool together with evaluator metadata. Incremental saves ensure that partial progress survives crashes, and IDs are globally unique across sessions.

Transition to expert review process. The resulting candidate pool is then passed to the expert review process detailed in Section 3.1. This marks the transition from automated preparation to rigorous human validation and refinement of each question.

¹qwen-max-2025-01-25, temperature 0, deterministic for reproducibility.

Exact Prompt Template

Context: Based *only* on the following text excerpt from page(s) {page_range} of the document '{doc_name}':
'''
{text_excerpt}
'''

Task: Generate ONE **extremely challenging** multiple-choice question (MCQ) with a target difficulty level of '{difficulty}' specifically designed to be difficult for advanced LLMs like Qwen-Max. The question must rigorously test deep comprehension, critical reasoning, inference, and synthesis *strictly* based on the provided text excerpt. Avoid simple factual recall.

Mandatory Instructions for Difficulty:

- Question Focus:** The question must target subtle details, implicit relationships, logical consequences, nuanced comparisons, or the underlying purpose/function of specific information within the text. It *must* require synthesizing information from *multiple* distinct sentences or points within the excerpt.
- Distractor Design (CRITICAL):** Create highly plausible but definitively incorrect distractors. These distractors should:
 - * Be semantically very close to the correct answer or other concepts in the text.
 - * Represent common misconceptions or logical fallacies related to the topic, *but* be directly contradicted by the provided text.
 - * Exploit potential ambiguities or require careful reading to identify subtle inaccuracies.
 - * Avoid being trivially wrong or easily eliminated. All options (A, B, C, D) should seem plausible on a superficial reading.
- Avoid External Knowledge:** The question *MUST* be unambiguously answerable *solely* from the information contained within the provided text excerpt. The difficulty must come from the complexity of reasoning *within* the text, not from needing outside information.
- Direct Questioning:** The text of the generated `question` itself must *not* contain phrases like "Based on the text excerpt", "According to the document...", "In the provided context...", etc. Frame the question directly about the content.
- Advanced Strategies:** Employ techniques like identifying cause-effect relationships only implied, analyzing the necessity or sufficiency of conditions mentioned, or evaluating the scope/limitations of claims made in the text. Carefully consider using negation/exception questions ("Which of the following is NOT...") only if they significantly increase cognitive load beyond simple recall.

Self-Correction Check (Internal Thought Process before outputting): Does this question *genuinely* require multi-step reasoning or deep inference based *only* on the text? Are the distractors *truly* deceptive and require careful analysis of the excerpt to disprove? Is it significantly harder than a simple keyword search or basic comprehension question?

Output Format: Respond *ONLY* with a single, valid JSON object matching this exact structure:

```
{
  "question": "String: The extremely challenging question text (must not refer to the prompt or text source).",
  "options": { "A": "String: Highly plausible but incorrect Option A.", "B": "String: Highly plausible but incorrect Option B.", "C": "String: The single, unambiguously correct answer based only on the text.", "D": "String: Highly plausible but incorrect Option D." },
  "correct_answer_text": "String: The exact text of the correct answer option (must match one of the options exactly).",
  "correct_option_key": "String: The key ('A', 'B', 'C', or 'D') corresponding to the correct answer."
}
Important: Ensure the `correct_option_key` points to the option that IS the correct answer. Randomize the position of the correct answer (A, B, C, or D). Do not include any introductory text, explanations, or markdown formatting like ```json ... ``` before or after the JSON object. Output only the JSON.
```

Algorithm 1: Automated pipeline for generating candidate MCQs

```
1 Hyper-parameters:
2 Chunk length  $L \in [150, 4500]$  chars;   LLM = qwen-max-2025-01-25,  $T=0$ ;   seed = 42;
   max rounds  $R_{\max} = 15$ ;   target acc.  $\leq 50\%$ ;   auth-error cap  $E_{\max} = 5$ .
Input : PDF folder knowledge_source
Output : all_candidates_untested.json, final_round_tested.json
3 foreach PDF  $d$  in folder do                                     // independent handling
4   decryptIfPossible( $d$ )
5   if unsupported encryption then continue
6   pages  $\leftarrow$  EXTRACTPAGES( $d$ )                                // cleaned text
7   chunks  $\leftarrow$  SELECTCHUNKS(pages,  $L$ )
8   foreach chunk  $c$  in chunks do
9      $mcq \leftarrow$  PROMPTLLM( $c$ )
10    if VALID( $mcq$ ) then
11      append option Z ("Not sure")
12      add metadata {id,  $d$ , page_range, difficulty, timestamp}
13      save to all_candidates_untested.json
14    end
15  end
16  round  $\leftarrow 1$ ;
17  acc  $\leftarrow 1$ ;
18  errs  $\leftarrow 0$ 
19  while round  $\leq R_{\max}$  and acc  $> 0.5$  do
20    acc  $\leftarrow$  EVALUATEPOOL()
21    remove correctly answered MCQs
22    if errs  $\geq E_{\max}$  or no new chunks then break
23    replenish via SELECTCHUNKS  $\rightarrow$  PROMPTLLM
24    round  $\leftarrow$  round + 1
25  end
26  drop unstable MCQs across rounds
27  deduplicate with LLM similarity check (keep first)
28  atomically write surviving pool  $\rightarrow$  final_round_tested.json
29 end
30 Expert review: domain specialists verify facts, polish wording, rebalance difficulty, and
   accept/reject questions for release.
```

Metric	Value
<i>Questions</i>	
Entries/Tokens/Vocab	1 136 / 40 983 / 2 896
Mean (Median)	36.08 (35)
Q ₁ – Q ₃	30 – 40
Range	18 – 94
<i>Options</i>	
Entries/Tokens/Vocab	5 680 / 88 304 / 6 810
Mean (Median)	15.55 (17)
Q ₁ – Q ₃	12 – 21
Range	1 – 54
<i>Source Text</i>	
Entries/Tokens/Vocab	1 136 / 550 200 / 18 826
Mean (Median)	484.33 (467)
Q ₁ – Q ₃	390 – 586
Range	33 – 984

Table 1: Token-level profile of the **ESGenius-QA**. Source texts provide dense evidence (median 467 tokens) supporting higher-order reasoning.

A.2 LLM’s Raw Response Validation

Given an LLM’s raw response, the routine first guards against NONE or empty inputs, then accepts “Z” directly as the special *unsure* option. For other strings it proceeds in two stages:

Direct acceptance: If every character in the trimmed string is drawn from {A,B,C,D}, return the deduplicated, *sorted* set (e.g., “DCB” → “BCD”).

Robust cleanup: Otherwise the string is sanitized: (i) replace all non-alphabetic symbols by spaces, (ii) drop any word containing lowercase letters, (iii) retain only words composed of A–D, Z. The result is deduplicated and sorted as above. If the cleaned string is empty, contains a lone “Z”, or mixes “Z” with other letters, the function outputs INVALID_ANSWER_MARKER.

This design accepts the minimal valid alphabet while aggressively filtering free-form text, punctuation, and lower-case distractors that often appear in LLM generations, to guarantee that the response is a valid multiple-choice answer.

A.3 ESGenius-QA Example Question Structures

Zero-shot Evaluation Example

Question ID: 1

Question. According to the *IPCC AR6 Synthesis Report*, which statement would most likely *increase the vulnerability of biodiversity and ecosystem services* while simultaneously *reducing carbon uptake*, given the relationships described?

- A. Implementing targeted management strategies for ocean ecosystems without addressing global warming.
- B. Failing to rebuild overexploited fisheries while achieving significant reductions in fossil-fuel dependency.
- C. Limiting global warming but neglecting land-restoration efforts and cooperative decision-making with Indigenous Peoples.
- D. Prioritizing disaster-risk management and early-warning systems without integrating ecosystem considerations into climate policies.
- Z. Not sure

Correct Answer: A

RAG Evaluation Example

Question ID: 580

Reference. Page 213, SUS_Report.pdf

Question. Which of the following conclusions about Fiji's progress toward sustainable development can be most reasonably inferred from the data trends and gaps presented in the *Sustainable Development Report 2024*?

- A. Fiji has achieved near-universal access to clean water but faces significant challenges in reducing urban slum populations.
- B. Fiji's environmental sustainability efforts are hindered primarily by high deforestation rates and low biodiversity protection.
- C. Fiji excels in reducing income inequality, as evidenced by a low Gini coefficient and minimal reliance on imports for nitrogen emissions.
- D. Fiji demonstrates strong performance in internet accessibility but shows untracked research and development expenditure.
- Z. Not sure

Correct Answer: D

Retrieved Source Text. Performance by Indicator5. Country Profiles *Sustainable Development Report 2024 – The SDGs and the UN Summit of the Future* 201

* Imputed data point, ** Not applicable NA = Data not available

FIJI

SDG 9 – Industry, Innovation and Infrastructure

Rural population with access to all-season roads (%) 96.7 (2024) ••

Population using the internet (%) 85.2 (2022) •A

Mobile broadband subscriptions (per 100 population) 76.1 (2021) •A

Logistics Performance Index: Infrastructure score (worst 1–5 best) 2.2 (2023) •G

The Times Higher Education Universities Ranking: Average score of top 3 universities (worst 0–100 best) 30.5 (2024) ••

Articles published in academic journals (per 1 000 population) 0.6 (2022) •A

Expenditure on research and development (% of GDP) NA NA ••

SDG 10 – Reduced Inequalities

Gini coefficient 30.7 (2019) ••

Palma ratio 1.1 (2019) ••

SDG 11 – Sustainable Cities and Communities

Proportion of urban population living in slums (%) 9.4 (2020) •D

Annual mean concentration of PM_{2.5} (µg/m³) 7.4 (2022) •D

Access to improved water source, piped (% of urban population) 98.4 (2022) •A

Population with convenient access to public transport in cities (%) 19.2 (2020) ••

SDG 12 – Responsible Consumption and Production

Municipal solid waste (kg/capita/day) 0.6 (2011) ••

Electronic waste (kg/capita) 6.1 (2019) ••

Production-based air pollution (DALYs per 1 000 population) NA NA ••

Air pollution associated with imports (DALYs per 1 000 population) NA NA ••

Production-based nitrogen emissions (kg/capita) NA NA ••

Nitrogen emissions associated with imports (kg/capita) NA NA ••

Exports of plastic waste (kg/capita) 0.6 (2022) •A

SDG 13 – Climate Action

CO₂ emissions from fossil-fuel combustion and cement production (t CO₂/capita) 1.2 (2022) •A

GHG emissions embodied in imports (t CO₂/capita) NA NA ••

CO₂ emissions embodied in fossil-fuel exports (kg/capita) 0.0 (2022) ••

SDG 14 – Life Below Water

Mean area that is protected in marine sites important to biodiversity (%) 16.5 (2023) •D

Ocean Health Index: Clean Waters score (worst 0–100 best) 74.1 (2023) •D

Fish caught from overexploited or collapsed stocks (% of total

catch) 9.0 (2018) •A
Fish caught by trawling or dredging (%) 0.0 (2019) ••
Fish caught that are then discarded (%) 7.3 (2019) •D
Marine biodiversity threats embodied in imports (per million population) 0.3 (2018) ••

SDG 15 – Life on Land

Mean area that is protected in terrestrial sites important to biodiversity (%) 11.2 (2023) •D
Mean area that is protected in freshwater sites important to biodiversity (%) 0.1 (2023) •D
Red List Index of species survival (worst 0–1 best) 0.69 (2024) •G
Permanent deforestation (% of forest area, 3-year average) 0.1 (2022) •A
Imported deforestation (m²/capita) NA NA ••

SDG 16 – Peace, Justice and Strong Institutions

Homicides (per 100 000 population) 2.2 (2020) •D
Crime is effectively controlled (worst 0–1 best) NA NA ••
Unsented detainees (% of prison population) 19.9 (2021) •A
Birth registrations with civil authority (% of children under 5) 86.6 (2021) ••
Corruption Perceptions Index (worst 0–1 best) 52.0 (2023) ••
Children involved in child labor (%) 16.7 (2021) ••
Exports of major conventional weapons (TIV constant million USD per 100 000 population) * 0.0 (2023) ••
Press Freedom Index (worst 0–1 best) 71.2 (2024) •A
Access to and affordability of justice (worst 0–1 best) NA NA ••
Timeliness of administrative proceedings (worst 0–1 best) NA NA ••
Expropriations are lawful and adequately compensated (worst 0–1 best) NA NA ••

SDG 17 – Partnerships for the Goals

Government spending on health and education (% of GDP) 9.3 (2021) •A
Government revenue excluding grants (% of GDP) 19.0 (2021) •G
Corporate Tax Haven score (best 0–100 worst) * 0 (2021) ••
Statistical Performance Index (worst 0–100 best) 63.2 (2022) •S
Index of countries' support to UN-based multilateralism (worst 0–100 best) 88.3 (2023) ••

SDG 1 – No Poverty

Poverty headcount ratio at \$2.15/day (2017 PPP, %) 1.6 (2024) •D
Poverty headcount ratio at \$3.65/day (2017 PPP, %) 7.3 (2024) •D

SDG 2 – Zero Hunger

Prevalence of undernourishment (%) 6.6 (2021) •A

Prevalence of stunting in children under 5 years of age (%) 7.2 (2021) ••

Prevalence of wasting in children under 5 years of age (%) 4.6 (2021) ••

Prevalence of obesity, BMI \geq 30 (% of adult population) 33.8 (2022) •G

Human Trophic Level (best 2–3 worst) 2.2 (2021) •D

Cereal yield (tonnes per hectare of harvested land) 4.1 (202. . .)

B Prompt Template

For multiple-choice questions (4 options + 1 Not sure with 1 answer), we use the following prompt templates:

B.1 Zero-Shot Prompt

You are an expert in ESG (Environmental, Social, Governance) and Sustainability related topics. Answer the question with a single letter based on authoritative knowledge. Each option content is case-sensitive.

Question: [Question text]

Options:

A: [Option A text]

B: [Option B text]

C: [Option C text]

D: [Option D text]

Z: Not sure

Answer: <Model's response goes here>

B.2 RAG Prompt

Context: [source text]

You are an expert in ESG (Environmental, Social, Governance) and Sustainability related topics. Answer the question with a single letter based on authoritative knowledge. Each option content is case-sensitive.

Question: [Question text]

Options:

A: [Option A text]

B: [Option B text]

C: [Option C text]

D: [Option D text]

Z: Not sure

Answer: <Model's response goes here>

C Main Experimental Results Table

Type	Family	Model	Size	S.G.	I-T	Rea	Zero-shot	RAG	Improvement
Open Source	DeepSeek	DeepSeek-R1-Distill-Qwen	1.5B	M	No	Yes	0.3134	0.4305	37.36%
		DeepSeek-R1-Distill-Qwen	7B	L	No	Yes	0.5018	0.6505	29.63%
		DeepSeek-R1-Distill-Qwen	14B	L	No	Yes	0.6382	0.8046	26.07%
		DeepSeek-R1-Distill-Qwen	32B	XL	No	Yes	0.6312	0.8143	29.01%
		DeepSeek-R1-Distill-Llama	8B	L	No	Yes	0.5502	0.6250	13.60%
		DeepSeek-R1-Distill-Llama	70B	XL	No	Yes	0.6505	0.8170	25.60%
	Google Gemma	Gemma-3(Team et al., 2025)	1B	M	No	No	0.2421	0.2526	4.33%
		Gemma-3	1B	M	Yes	No	0.5000	0.5977	19.54%
		Gemma-3	4B	M	No	No	0.6083	0.6860	12.77%
		Gemma-3	4B	M	Yes	No	0.6144	0.7518	22.36%
		Gemma-3	12B	L	No	No	0.6514	0.6857	5.26%
		Gemma-3	12B	L	Yes	No	0.6514	0.8380	28.64%
		Gemma-3	27B	L	No	No	0.2165	0.5229	141.57%
		Gemma-3	27B	L	Yes	No	0.6356	0.8336	31.15%
	Meta Llama	Meta-Llama-3(Grattafiori et al., 2024)	8B	L	No	No	0.6461	0.7324	13.36%
		Llama-3.1	8B	L	No	No	0.6699	0.7650	14.20%
		Llama-3.1	8B	L	Yes	No	0.6382	0.7993	25.24%
		Llama-3.2	1B	M	No	No	0.3609	0.3680	2.00%
		Llama-3.2	1B	M	Yes	No	0.5986	0.6452	7.79%
		Llama-3.2	3B	M	No	No	0.6074	0.6831	12.48%
		Llama-3.2	3B	M	Yes	No	0.5968	0.7218	20.95%
		Llama-3.3	70B	XL	Yes	No	0.6576	0.7887	20.00%
	Alibaba Qwen	Qwen2.5(Yang et al., 2024)	0.5B	S	No	No	0.5458	0.5396	-1.14%
		Qwen2.5	0.5B	S	Yes	No	0.5775	0.5343	-7.48%
		Qwen2.5	1.5B	M	No	No	0.6391	0.6928	8.40%
		Qwen2.5	1.5B	M	Yes	No	0.5484	0.6972	27.13%
		Qwen2.5	3B	M	No	No	0.5889	0.7632	29.60%
		Qwen2.5	3B	M	Yes	No	0.5871	0.5211	-11.24%
		Qwen2.5	7B	L	No	No	0.6496	0.8055	23.99%
		Qwen2.5	7B	L	Yes	No	0.6276	0.7967	27.27%
		Qwen2.5	14B	L	No	No	0.6734	0.8231	22.22%
		Qwen2.5	14B	L	Yes	No	0.6197	0.8231	32.83%
		Qwen2.5	32B	XL	No	No	0.6593	0.8081	22.55%
		Qwen2.5	32B	XL	Yes	No	0.6039	0.8247	36.57%
		Qwen2.5	72B	XL	No	No	0.6188	0.7201	16.39%
		Qwen2.5	72B	XL	Yes	No	0.6347	0.8257	29.78%
		Qwen2.5-1M(Yang et al., 2025eb)	7B	L	Yes	No	0.6206	0.8063	29.92%
		Qwen2.5-1M	14B	L	Yes	No	0.6268	0.8222	28.01%
		QwQ(Team, 2025)	32B	XL	No	Yes	0.3900	0.7614	95.23%
		Qwen3(Yang et al., 2025)	0.6B	S	No	No	0.2896	0.0942	-67.47%
		Qwen3	1.7B	M	No	No	0.5836	0.6937	18.87%
		Qwen3	4B	M	No	No	0.6188	0.7905	27.75%
		Qwen3	8B	L	No	No	0.6021	0.6708	11.41%
Proprietary API	DeepSeek	DeepSeek-R1(Guo et al., 2025)	671B	XXL	No	Yes	0.6629	-	-
		DeepSeek-V3(Liu et al., 2024)	671B	XXL	No	No	0.6532	-	-
	Alibaba Qwen	Qwen2.5-Max	72B†(MoE)	XL	No	Yes	0.6444	-	-
	OpenAI GPT	GPT-4o-mini	8B†	L	No	No	0.6268	-	-
		GPT-4o(Hurst et al., 2024)	200B†	XXL	No	No	0.6364	-	-
		o4-mini	3B†	M	No	Yes	0.6945	-	-
		o3	300B†	M	No	Yes	0.7254	-	-

Table 2: Comprehensive ESGenius results showing LLM performance under Zero-shot and RAG settings. S.G. denotes Size Group (S: Small ($\leq 1B$), M: Medium (1–7B), L: Large (7–30B), XL: Extra Large (30–100B), XXL: Extra Extra Large ($>100B$)); I-T: Instruction Tuning; Rea: Reasoning Focus. Improvement (%): $\frac{\text{RAG} - \text{Zero-shot}}{\text{Zero-shot}} \times 100$. “–” indicates scores currently unavailable due to technical constraints. † indicates industry size estimates.

C.1 3 Examples with Very Low Accuracy Across All Models

Question ID: 432

Question. Which statement accurately reflects the relationship between renewable energy adoption and regional disparities as indicated in *The Sustainable Development Goals Report 2024*?

Options:

A: Developing countries are projected to surpass developed countries in per capita renewable electricity capacity within the next decade based on current growth rates.

B: The disparity in renewable energy capacity between least developed countries (LDCs) and developing countries is expected to close within 15 years if LDCs maintain their current growth trajectory.

C: The installed renewable electricity capacity in least developed countries (LDCs) grew at a faster compound annual growth rate than in developed countries over the past seven years.

D: Landlocked developing States have achieved a higher per capita renewable electricity capacity than small island developing countries but still lag significantly behind developing countries overall.

Z: Not sure

Correct Answer: D

Accuracy: 0%

Question ID: 635

Question. Which factor is most directly responsible for the limited adoption of ESCO business models despite their potential to mitigate financial risks and provide expertise in energy efficiency projects, according to the *Climate Change 2022: Mitigation of Climate Change. Working Group III Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*?

Options:

A: The absence of stringent regulatory frameworks governing energy codes.

B: The insufficient involvement of the public sector in transportation infrastructure projects.

C: The lack of awareness among financial institutions about energy efficiency benefits.

D: The prevalence of asymmetric information and split incentives within firms.

Z: Not sure

Correct Answer: C

Accuracy: 0%

Question ID: 1006

Question. Which of the following best describes a necessary condition for an injury or illness to be classified as a recordable incident under the entity's disclosure requirements in the *Chemicals – Sustainability Accounting Standard*?

Options:

A: The injury or illness must result in at least one day away from work or require medical treatment beyond first aid.

B: The injury or illness must be diagnosed by a licensed healthcare professional, regardless of its impact on the employee's work status.

C: The injury or illness must occur within the establishment but does not need to involve exposure to harmful substances or heavy machinery.

D: The injury or illness must lead to restricted work, job transfer, or loss of consciousness, even if it is not diagnosed by a physician.

Z: Not sure

Correct Answer: B

Accuracy: 0%

D ESGenius-Corpus

The corpus integrates authoritative frameworks, corporate questionnaires, and scientific assessments that comprehensively cover environmental, social, and governance (ESG) considerations. A detailed quantitative overview of the collection is presented across multiple tables: Table 3 summarizes IPCC climate-science assessments, Table 4 covers GRI Standards, Tables 5–6 detail SASB's industry-specific disclosures, Tables 7–8 outline IFRS/ISSB standards and guidance, Table 9 catalogs TCFD materials, Table 10 documents CDP questionnaires, and Table 11 captures UN Sustainable Development Goals content. The distribution of pages and questions across these sources is visualized in Figures 5 and 6, respectively. From this comprehensive dataset, several key insights emerge:

1. **Breadth versus depth.** Altogether, these seven sources comprise **231 distinct documents** spanning approximately **19,600 pages**. While bodies such as SASB and IFRS/ISSB each contribute dozens of relatively concise files, the Intergovernmental Panel on Climate Change (IPCC) anchors the corpus with just seven reports that total over 10,000 pages.
2. **Standards-driven core.** Internationally recognized ESG standards and disclosure frameworks—GRI, SASB, IFRS/ISSB, TCFD, and

CDP—form the backbone of **ESGenius**. They reflect the most widely adopted practices for sustainability reporting, management, and climate-risk disclosure, offering a robust foundation for both quantitative and qualitative benchmarking.

3. **Disclosure and reporting frameworks.** Market-facing reporting initiatives (SASB, GRI, TCFD, CDP) collectively emphasize implementation guides and sector- or topic-specific questionnaires, resulting in many individual questions but fewer pages per document. This modular structure facilitates domain-specific data collection and comparability across diverse industries.
4. **Science-heavy climate assessments.** IPCC assessment reports present the opposite pattern: fewer individual documents but extremely high page counts. This science-heavy text ensures deep coverage of climate-change fundamentals, impacts, and mitigation pathways—an essential knowledge base informing the standards and regulations in the broader ESG ecosystem.
5. **Sustainable development anchor.** The UN’s Sustainable Development Goals (SDGs) underpin cross-sector and cross-country sustainability objectives. Although comparatively compact in page count, these seminal UN publications situate corporate ESG strategies within the global 2030 Agenda, ensuring broader alignment with international development priorities.
6. **Imbalanced density highlights practical challenges.** Marked disparities between the distribution of documents and the distribution of pages underscore the varied scope of ESG sources: some (e.g. IPCC) are exhaustive scientific compendiums, while others (e.g. GRI, SASB, IFRS) comprise slimmer but more numerous reference standards. For researchers and practitioners alike, tasks ranging from large-scale summarization to specialized technical queries must navigate this imbalance of item count versus depth.

depth (tens of thousands of pages) required for evaluating advanced language models on ESG-focused reasoning, retrieval, and generation tasks. Future expansions will incorporate emerging disclosure rules and further national or sectoral guidelines, maintaining the corpus’s comprehensive coverage over time.

Taken together, these characteristics demonstrate that **ESGenius** provides both the *breadth* (multiple standards, guidance, and scientific anchors) and the

Original Document Title	Year	Size	Pages	No. Qs
<i>Climate Change 2023 — Synthesis Report</i> (IPCC, 2023cb)	2023	4.9 MB	186	18
<i>Climate Change 2022: Mitigation of Climate Change. Working Group III Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change</i> (IPCC, 2022cb)	2022	74.2 MB	2 042	19
<i>Climate Change 2022: Impacts, Adaptation and Vulnerability. Working Group II Contribution to the IPCC Sixth Assessment Report</i> (IPCC, 2022)	2022	378 MB	3 675	14
<i>Climate Change 2021: The Physical Science Basis. Working Group I Contribution to the IPCC Sixth Assessment Report</i> (IPCC, 2021)	2021	275 MB	2 409	14
<i>Climate Change and Land: An IPCC Special Report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems</i> (IPCC, 2019)	2019	28 MB	874	16
<i>The Ocean and Cryosphere in a Changing Climate: A Special Report of the Intergovernmental Panel on Climate Change</i> (IPCC, 2019cb)	2019	59.4 MB	765	17
<i>Global Warming of 1.5 °C: An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty</i> (IPCC, 2018)	2018	65 MB	631	19
Total	-	884.5MB	10 582	117

Table 3: Comprehensive metadata for the **seven IPCC reports** curated in the **ESGenius**. This collection represents the complete Sixth Assessment Report (AR6) cycle and key Special Reports from 2018-2023. The AR6 materials include the 2023 Synthesis Report and three Working Group contributions covering physical science (WG1), impacts & adaptation (WG2), and mitigation (WG3). Three thematic Special Reports address land use (SRCCL), oceans & ice (SROCC), and 1.5°C warming pathways (SR15). Totalling **10,582 pages** and **117 evaluation questions**, these authoritative climate science assessments form a crucial knowledge foundation for ESG analysis. Each report is with document sizes ranging from 4.9MB to 378MB. The collection provides comprehensive coverage of climate science, impacts, and policy responses that inform modern ESG frameworks like TCFD, CSRD, and ISSB standards. All documents are sourced directly from IPCC (<https://www.ipcc.ch>) and represent peer-reviewed, UN-mandated scientific assessments.

Original Document Title	Year	Size	Pages	No. Qs
<i>Consolidated Set of the GRI Standards 2021</i> (gri, 2021)	2021	19 MB	677	46
<i>GRI 1: Foundation 2021</i> (gri, 2021ab)	2021	1.2 MB	39	6
<i>GRI 2: General Disclosures 2021</i> (gri, 2021ad)	2021	1.2 MB	58	11
<i>GRI 3: Material Topics 2021</i> (gri, 2021ae)	2021	1.1 MB	30	5
<i>GRI 11: Oil and Gas Sector 2021</i> (gri, 2021ac)	2021	2.2 MB	93	14
<i>GRI 12: Coal Sector 2022</i> (gri, 2022)	2022	2.1 MB	86	16
<i>GRI 13: Agriculture, Aquaculture and Fishing Sectors 2022</i> (gri, 2022ab)	2022	2.5 MB	95	17
<i>GRI 14: Mining Sector 2024</i> (gri, 2024ab)	2024	2.6 MB	100	15
<i>GRI 101: Biodiversity 2024</i> (gri, 2024)	2024	1.3 MB	50	10
<i>GRI 201: Economic Performance 2016</i> (gri, 2016a)	2016	862 KB	16	2
<i>GRI 202: Market Presence 2016</i> (gri, 2016b)	2016	834 KB	15	1
<i>GRI 203: Indirect Economic Impacts 2016</i> (gri, 2016c)	2016	817 KB	11	3
<i>GRI 204: Procurement Practices 2016</i> (gri, 2016d)	2016	821 KB	11	3
<i>GRI 205: Anti-corruption 2016</i> (gri, 2016e)	2016	855 KB	16	1
<i>GRI 206: Anti-competitive Behavior 2016</i> (gri, 2016f)	2016	829 KB	13	1
<i>GRI 207: Tax 2019</i> (gri, 2019)	2019	947 KB	21	3
<i>GRI 301: Materials 2016</i> (gri, 2016g)	2016	831 KB	13	1
<i>GRI 302: Energy 2016</i> (gri, 2016h)	2016	859 KB	19	1
<i>GRI 303: Water and Effluents 2018</i> (gri, 2018)	2018	1 MB	28	4
<i>GRI 304: Biodiversity 2016</i> (gri, 2016i)	2016	845 KB	15	1
<i>GRI 305: Emissions 2016</i> (gri, 2016j)	2016	936 KB	26	4
<i>GRI 306: Effluents and Waste 2016</i> (gri, 2016k)	2016	640 KB	15	1
<i>GRI 306: Waste 2020</i> (gri, 2020)	2020	1.7 MB	30	1
<i>GRI 308: Supplier Environmental Assessment 2016</i> (gri, 2016l)	2016	851 KB	14	1
<i>GRI 401: Employment 2016</i> (gri, 2016m)	2016	860 KB	16	2
<i>GRI 402: Labor/Management Relations 2016</i> (gri, 2016n)	2016	855 KB	13	1
<i>GRI 403: Occupational Health and Safety 2018</i> (gri, 2018ab)	2018	1.1 MB	35	6
<i>GRI 404: Training and Education 2016</i> (gri, 2016o)	2016	837 KB	15	0
<i>GRI 405: Diversity and Equal Opportunity 2016</i> (gri, 2016p)	2016	856 KB	15	1
<i>GRI 406: Non-discrimination 2016</i> (gri, 2016q)	2016	853 KB	12	1
<i>GRI 407: Freedom of Association and Collective Bargaining 2016</i> (gri, 2016r)	2016	870 KB	13	1
<i>GRI 408: Child Labor 2016</i> (gri, 2016s)	2016	868 KB	14	0
<i>GRI 409: Forced or Compulsory Labor 2016</i> (gri, 2016t)	2016	912 KB	13	1
<i>GRI 410: Security Practices 2016</i> (gri, 2016u)	2016	841 KB	11	4
<i>GRI 411: Rights of Indigenous Peoples 2016</i> (gri, 2016v)	2016	863 KB	14	1
<i>GRI 413: Local Communities 2016</i> (gri, 2016w)	2016	885 KB	16	2
<i>GRI 414: Supplier Social Assessment 2016</i> (gri, 2016x)	2016	853 KB	14	1
<i>GRI 415: Public Policy 2016</i> (gri, 2016y)	2016	816 KB	12	1
<i>GRI 416: Customer Health and Safety 2016</i> (gri, 2016z)	2016	825 KB	12	1
<i>GRI 417: Marketing and Labeling 2016</i> (gri, 2016aa)	2016	837 KB	14	1
<i>GRI 418: Customer Privacy 2016</i> (gri, 2016ab)	2016	825 KB	12	1
<i>GRI Standards Glossary 2022</i> (gri, 2022ac)	2022	680 KB	23	2
<i>A Practical Guide to Sustainability Reporting Using GRI and SASB Standards</i> (gri, 2021ag)	2021	1.5 MB	42	4
<i>The GRI Standards — A Guide for Policy Makers</i> (gri, 2021af)	2021	7.1 MB	19	2
Total	-	70.863MB	1,826	201

Table 4: Metadata for the **GRI Standards** collection in **ESGenius**. This comprehensive collection spans 2016-2024 and comprises: (1) Universal Standards (GRI 1-3) establishing core reporting principles, (2) Sector Standards (GRI 11-14) for high-impact industries, and (3) Topic Standards covering economic (200 series), environmental (300 series), and social (400 series) aspects. Key features include: **Coverage:** 1,826 pages across 40+ standards documents; **Evaluation:** 201 domain-specific questions; **Scope:** Comprehensive ESG disclosure requirements spanning corporate governance, environmental impact, and social responsibility. All standards sourced from GRI (<https://www.globalreporting.org/>), the leading authority in sustainability reporting frameworks.

Original Document Title	Year	Size	Pages	No. Qs
<i>Apparel, Accessories & Footwear - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023cg)</i>	2023	410 KB	21	4
<i>Appliance Manufacturing - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023ch)</i>	2023	340 KB	13	1
<i>Building Products & Furnishings - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023cn)</i>	2023	369 KB	18	1
<i>E-commerce - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023cy)</i>	2023	402 KB	24	3
<i>Household & Personal Products - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023dn)</i>	2023	371 KB	19	1
<i>Multiline & Specialty Retailers & Distributors - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023eb)</i>	2023	399 KB	24	3
<i>Toys & Sporting Goods - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023ev)</i>	2023	346 KB	13	0
<i>Coal Operations - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023cr)</i>	2023	490 KB	41	8
<i>Construction Materials - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023ct)</i>	2023	414 KB	27	3
<i>Iron & Steel Producers - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023ds)</i>	2023	392 KB	24	4
<i>Metals & Mining - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023dz)</i>	2023	521 KB	47	8
<i>Oil & Gas – Exploration & Production - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023ed)</i>	2023	511 KB	47	6
<i>Oil & Gas – Midstream - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023ee)</i>	2023	407 KB	26	4
<i>Oil & Gas – Refining & Marketing - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023ef)</i>	2023	433 KB	31	5
<i>Oil & Gas – Services - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023eg)</i>	2023	413 KB	28	4
<i>Asset Management & Custody Activities - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023ci)</i>	2023	404 KB	24	3
<i>Commercial Banks - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023cs)</i>	2023	394 KB	23	3
<i>Consumer Finance - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023cu)</i>	2023	372 KB	20	3
<i>Insurance - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023dp)</i>	2023	408 KB	26	3
<i>Investment Banking & Brokerage - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023dr)</i>	2023	420 KB	28	4
<i>Mortgage Finance - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023ea)</i>	2023	366 KB	17	2
<i>Security & Commodity Exchanges - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023ep)</i>	2023	362 KB	17	1
<i>Agricultural Products - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023cc)</i>	2023	426 KB	30	3
<i>Alcoholic Beverages - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023cf)</i>	2023	400 KB	26	4
<i>Food Retailers & Distributors - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023de)</i>	2023	452 KB	37	5
<i>Meat, Poultry & Dairy - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023dw)</i>	2023	442 KB	34	6
<i>Non-Alcoholic Beverages - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023ec)</i>	2023	419 KB	30	3
<i>Processed Foods - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023eh)</i>	2023	434 KB	32	5
<i>Restaurants - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023en)</i>	2023	418 KB	30	5
<i>Tobacco - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023eu)</i>	2023	337 KB	12	1
<i>Biotechnology & Pharmaceuticals - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023cm)</i>	2023	415 KB	29	4
<i>Drug Retailers - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023cx)</i>	2023	378 KB	21	3
<i>Health Care Delivery - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023dj)</i>	2023	432 KB	33	6
<i>Health Care Distributors - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023dk)</i>	2023	358 KB	17	1
<i>Managed Care - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023du)</i>	2023	371 KB	20	3
<i>Medical Equipment & Supplies - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023dy)</i>	2023	380 KB	22	2
<i>Electric Utilities & Power Generators - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023da)</i>	2023	458 KB	35	4
<i>Engineering & Construction Service - SUSTAINABILITY ACCOUNTING STANDARD (sas, 2023dd)</i>	2023	402 KB	26	4

Table 5: Metadata for SASB industry-specific disclosure standards — **Part I.**

Original Document Title	Year	Size	Pages	No. Qs
<i>Gas Utilities & Distributors - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023dh)	2023	372 KB	19	0
<i>Home Builders - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023dl)	2023	383 KB	21	1
<i>Real Estate - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023el)	2023	285 KB	38	5
<i>Real Estate Services - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023em)	2023	349 KB	15	0
<i>Waste Management - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023ew)	2023	430 KB	31	5
<i>Water Utilities & Services - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023ex)	2023	436 KB	32	4
<i>Biofuels - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023cl)	2023	380 KB	20	1
<i>Forestry Management - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023df)	2023	363 KB	18	2
<i>Fuel Cells & Industrial Batteries - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023dg)	2023	381 KB	21	3
<i>Pulp & Paper Products - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023ej)	2023	390 KB	24	2
<i>Solar Technology & Project Developers - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023es)	2023	410 KB	28	5
<i>Wind Technology & Project Developers - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023ey)	2023	358 KB	17	2
<i>Aerospace & Defence - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023cb)	2023	439 KB	27	4
<i>Chemicals - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023cq)	2023	447 KB	36	7
<i>Containers & Packaging - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023cv)	2023	422 KB	30	5
<i>Electrical & Electronic Equipment - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023db)	2023	385 KB	23	3
<i>Industrial Machinery & Goods - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023do)	2023	361 KB	17	2
<i>Advertising & Marketing - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023)	2023	368 KB	18	1
<i>Casinos & Gaming - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023cp)	2023	353 KB	15	1
<i>Education - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023cz)	2023	366 KB	18	1
<i>Hotels & Lodging - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023dm)	2023	371 KB	19	2
<i>Leisure Facilities - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023dt)	2023	343 KB	13	1
<i>Media & Entertainment - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023dx)	2023	357 KB	15	1
<i>Professional & Commercial Services - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023ei)	2023	371 KB	18	2
<i>Electronic Manufacturing Services & Original Design Manufacturing - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023dc)	2023	379 KB	22	3
<i>Hardware - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023di)	2023	392 KB	24	3
<i>Internet Media & Services - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023dq)	2023	417 KB	27	4
<i>Semiconductors - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023eq)	2023	409 KB	27	4
<i>Software & IT Services - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023er)	2023	426 KB	29	4
<i>Telecommunication Services - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023et)	2023	409 KB	26	2
<i>Air Freight & Logistics - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023cd)	2023	394 KB	24	4
<i>Airlines - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023ce)	2023	371 KB	19	2
<i>Auto Parts - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023cj)	2023	373 KB	20	2
<i>Automobiles - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023ck)	2023	378 KB	22	3
<i>Car Rental & Leasing - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023co)	2023	334 KB	11	3
<i>Cruise Lines - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023cw)	2023	414 KB	26	3
<i>Marine Transportation - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023dv)	2023	400 KB	24	2
<i>Rail Transportation - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023ek)	2023	381 KB	21	3
<i>Road Transportation - SUSTAINABILITY ACCOUNTING STANDARD</i> (sas, 2023eo)	2023	368 KB	17	1
Total		30.431MB	1,864	236

Table 6: Comprehensive metadata for all **77 industry-specific SASB Sustainability Accounting Standards** issued in 2023 and now stewarded by the IFRS Foundation’s ISSB. This corpus, distilled for our **ESGenius** system, comprises 77 documents totalling **30.431 MB**, **1,864 pages**, and **236 multiple-choice questions**. The standards span the full economy—from *Apparel* to *Transportation*—and cover financially material sustainability topics including greenhouse gas emissions, water management, data security, and workforce diversity. These standards provide investors with decision-relevant ESG information for valuation, risk assessment, and stewardship. The complete collection is freely accessible at <https://www.sasb.org/>.

Original Document Title	Year	Size	Pages	No. Qs
<i>IFRS S1</i> (ifr, 2023ad)	2023	307 KB	48	8
<i>IFRS S2</i> (ifr, 2023af)	2023	297 KB	46	8
<i>Progress on Corporate Climate-related Disclosures — 2024 Report</i> (ifr, 2024cb)	2024	1.3 MB	164	27
<i>IFRS S1 Basis for Conclusions on General Requirements for Disclosure of Sustainability-related Financial Information</i> (ifr, 2023ac)	2023	341 KB	57	10
<i>IFRS S2 Basis for Conclusions on Climate-related Disclosures</i> (ifr, 2023ae)	2023	337 KB	55	11
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures</i> (ifr, 2023ag)	2023	2.3 MB	538	33
<i>IFRS S1 Accompanying Guidance on General Requirements for Disclosure of Sustainability-related Financial Information</i> (ifr, 2023ab)	2023	201 KB	17	2
<i>Exposure Draft Amendments to Greenhouse Gas Emissions Disclosures Proposed amendments to IFRS S2 Comments to be received by 27 June 2025</i> (ifr, 2025)	2025	239 KB	24	4
<i>Exposure Draft Basis for Conclusions on Amendments to Greenhouse Gas Emissions Disclosures Proposed amendments to IFRS S2 Comments to be received by 27 June 2025</i> (ifr, 2025cb)	2025	238 KB	25	4
<i>Comparison IFRS S2 Climate-related Disclosures with the TCFD recommendations</i> (ifr, 2023)	2024	105 KB	12	1
<i>IFRS Taxonomy IFRS Sustainability Disclosure Taxonomy 2024 IFRS S1 General Requirements for Disclosure of Sustainability-related Financial Information and IFRS S2 Climate-related Disclosures</i> (ifr, 2024)	2024	1.1 MB	76	10
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 1 — Apparel, Accessories & Footwear</i> (ifr, 2023ah)	2023	194 KB	10	4
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 2 — Appliance Manufacturing</i> (ifr, 2023as)	2023	179 KB	7	3
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 3 — Building Products & Furnishings</i> (ifr, 2023bd)	2023	203 KB	13	1
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 4 — E-Commerce</i> (ifr, 2023bo)	2023	199 KB	11	2
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 5 — Household & Personal Products</i> (ifr, 2023bz)	2023	191 KB	9	2
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 6 — Multiline and Specialty Retailers & Distributors</i> (ifr, 2023ck)	2023	180 KB	7	0
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 7 — Coal Operations</i> (ifr, 2023cu)	2023	216 KB	14	1
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 8 — Construction Materials</i> (ifr, 2023cv)	2023	232 KB	17	1
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 9 — Iron & Steel Producers</i> (ifr, 2023cw)	2023	210 KB	14	1
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 10 — Metals & Mining</i> (ifr, 2023ai)	2023	204 KB	12	1
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 11 — Oil & Gas – Exploration & Production</i> (ifr, 2023aj)	2023	244 KB	20	2
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 12 — Oil & Gas – Midstream</i> (ifr, 2023ak)	2023	196 KB	10	3
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 13 — Oil & Gas – Refining & Marketing</i> (ifr, 2023al)	2023	212 KB	13	1
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 14 — Oil & Gas – Services</i> (ifr, 2023am)	2023	202 KB	11	1
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 15 — Asset Management & Custody Activities</i> (ifr, 2023an)	2023	202 KB	11	4
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 16 — Commercial Banks</i> (ifr, 2023ao)	2023	189 KB	8	1
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 17 — Insurance</i> (ifr, 2023ap)	2023	223 KB	15	1
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 18 — Investment Banking & Brokerage</i> (ifr, 2023aq)	2023	201 KB	10	2
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 19 — Mortgage Finance</i> (ifr, 2023ar)	2023	188 KB	7	1
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 20 — Agricultural Products</i> (ifr, 2023at)	2023	223 KB	17	1
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 21 — Alcoholic Beverages</i> (ifr, 2023au)	2023	209 KB	13	1
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 22 — Food Retailers & Distributors</i> (ifr, 2023av)	2023	219 KB	15	1
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 23 — Meat, Poultry & Dairy</i> (ifr, 2023aw)	2023	235 KB	19	1
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 24 — Non-Alcoholic Beverages</i> (ifr, 2023ax)	2023	217 KB	15	1
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 25 — Processed Foods</i> (ifr, 2023ay)	2023	219 KB	15	0
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 26 - Restaurants</i> (ifr, 2023az)	2023	202 KB	11	3
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 27 — Drug Retailers</i> (ifr, 2023ba)	2023	186 KB	7	2
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 28 — Health Care Delivery</i> (ifr, 2023bb)	2023	201 KB	11	2
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 29 — Health Care Distributors</i> (ifr, 2023bc)	2023	180 KB	6	2
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 30 — Managed Care</i> (ifr, 2023be)	2023	174 KB	6	1

Table 7: IFRS / ISSB sustainability disclosure materials — **Part I.**

Original Document Title	Year	Size	Pages	No. Qs
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 31 — Medical Equipment & Supplies</i> (ifr, 2023bf)	2023	183 KB	7	2
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 32 — Electric Utilities & Power Generators</i> (ifr, 2023bg)	2023	257 KB	24	3
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 33 — Engineering & Construction Services</i> (ifr, 2023bh)	2023	222 KB	16	1
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 34 — Gas Utilities & Distributors</i> (ifr, 2023bi)	2023	210 KB	13	1
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 35 — Home Builders</i> (ifr, 2023bj)	2023	214 KB	13	1
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 36 — Real Estate</i> (ifr, 2023bk)	2023	285 KB	34	3
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 37 — Real Estate Services</i> (ifr, 2023bl)	2023	197 KB	9	2
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 38 — Waste Management</i> (ifr, 2023bm)	2023	215 KB	12	1
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 39 — Water Utilities & Services</i> (ifr, 2023bn)	2023	241 KB	19	1
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 40 — Biofuels</i> (ifr, 2023bp)	2023	208 KB	14	0
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 41 — Forestry Management</i> (ifr, 2023bq)	2023	209 KB	12	1
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 42 — Fuel Cells & Industrial Batteries</i> (ifr, 2023br)	2023	202 KB	10	3
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 43 — Pulp & Paper Products</i> (ifr, 2023bs)	2023	202 KB	18	1
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 44 — Solar Technology & Project Developers</i> (ifr, 2023bt)	2023	216 KB	14	1
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 45 — Wind Technology & Project Developers</i> (ifr, 2023bu)	2023	195 KB	8	2
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 46 — Aerospace & Defence</i> (ifr, 2023bv)	2023	204 KB	9	2
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 47 — Chemicals</i> (ifr, 2023bw)	2023	224 KB	16	2
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 48 — Containers & Packaging</i> (ifr, 2023bx)	2023	232 KB	17	2
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 49 — Electrical & Electronic Equipment</i> (ifr, 2023by)	2023	202 KB	10	2
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 50 — Industrial Machinery & Goods</i> (ifr, 2023ca)	2023	196 KB	9	3
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 51 — Casinos & Gaming</i> (ifr, 2023cb)	2023	189 KB	7	3
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 52 — Hotels & Lodging</i> (ifr, 2023cc)	2023	201 KB	9	4
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 53 — Leisure Facilities</i> (ifr, 2023cd)	2023	188 KB	7	1
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 54 — Electronic Mfg Services & ODM</i> (ifr, 2023ce)	2023	196 KB	8	1
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 55 — Hardware</i> (ifr, 2023cf)	2023	200 KB	10	2
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 56 — Internet Media & Services</i> (ifr, 2023cg)	2023	200 KB	9	2
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 57 — Semiconductors</i> (ifr, 2023ch)	2023	222 KB	14	0
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 58 — Software & IT Services</i> (ifr, 2023ci)	2023	208 KB	11	3
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 59 — Telecommunication Services</i> (ifr, 2023cj)	2023	205 KB	10	1
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 60 — Air Freight & Logistics</i> (ifr, 2023cl)	2023	202 KB	11	2
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 61 — Airlines</i> (ifr, 2023cm)	2023	201 KB	9	3
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 62 — Auto Parts</i> (ifr, 2023cn)	2023	192 KB	8	3
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 63 — Automobiles</i> (ifr, 2023co)	2023	193 KB	8	3
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 64 — Car Rental & Leasing</i> (ifr, 2023cp)	2023	189 KB	7	2
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 65 — Cruise Lines</i> (ifr, 2023cq)	2023	206 KB	10	3
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 66 — Marine Transportation</i> (ifr, 2023cr)	2023	208 KB	10	2
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 67 — Rail Transportation</i> (ifr, 2023cs)	2023	201 KB	9	1
<i>IFRS S2 Industry-based Guidance on implementing Climate-related Disclosures Vol 68 — Road Transportation</i> (ifr, 2023ct)	2023	200 KB	9	3
Total	-	20.81MB	1,866	238

Table 8: Comprehensive metadata for the entire **IFRS / ISSB sustainability-disclosure corpus** (2023–2025): it comprises the universal core standards *IFRS S1* (general sustainability-related financial disclosure) and *IFRS S2* (climate-related disclosure), their bases for conclusions, accompanying guidance, the 2024 climate-progress report, the 2024 Sustainability Disclosure Taxonomy, the 2025 greenhouse-gas exposure draft (with basis), a comparison of *IFRS S2* with the TCFD recommendations, and the 68-volume *IFRS S2 Industry-based Guidance* that maps SASB’s sector-specific materiality into the new framework. Together the 77 PDFs, total **20.81 MB**, span **1 866 pages**—well over two thousand when including ancillary matter—and yield **238** benchmark MCQs for **ESGenius**. *IFRS S1* sets the universal disclosure baseline, while *IFRS S2* details climate-specific metrics, mirroring the TCFD architecture and enriched with SASB’s sectoral depth; the accompanying materials establish a globally consistent, investor-focused baseline that links decision-relevant sustainability information—such as greenhouse-gas emissions, transition plans, climate resilience, data security, and workforce diversity—directly to financial statements. All documents are freely available at <https://www.ifrs.org/>.

Original Document Title	Year	Size	Pages	No. Qs
<i>Task Force on Climate-related Financial Disclosures Overview</i> (tcf, 2022)	2022	11.1 MB	25	2
<i>Task Force on Climate-related Financial Disclosures Guidance on Metrics, Targets, and Transition Plans</i> (tcf, 2021)	2021	12 MB	79	11
<i>Task Force on Climate-related Financial Disclosures Implementing the Recommendations of the Task Force on Climate-related Financial Disclosures</i> (tcf, 2021eb)	2021	1.1 MB	88	18
<i>TCFD Workshop – Session 1: Fundamentals and Overview of TCFD</i> (tcf, 2022eb)	2022	3.2 MB	40	7
<i>TCFD Workshop – Session 2: Governance</i> (tcf, 2022ec)	2022	1.4 MB	20	1
<i>TCFD Workshop – Session 3: Strategy</i> (tcf, 2022ed)	2022	2.7 MB	43	8
<i>TCFD Workshop – Session 4: Risk Management</i> (tcf, 2022ee)	2022	1.4 MB	34	4
<i>TCFD Workshop – Session 5: Metrics and Targets</i> (tcf, 2022ef)	2022	3 MB	53	9
<i>Task Force on Climate-related Financial Disclosures Guidance on Scenario Analysis for Non-Financial Companies</i> (tcf, 2020eb)	2020	3.7 MB	133	22
<i>Task Force on Climate-related Financial Disclosures Guidance on Risk Management Integration and Disclosure</i> (tcf, 2020)	2020	5.2 MB	52	9
<i>Task Force on Climate-related Financial Disclosures 2023 Status Report</i> (tcf, 2023)	2023	19.3 MB	161	32
Total	-	60 MB	728	123

Table 9: Comprehensive metadata for the **Task Force on Climate-related Financial Disclosures (TCFD)** corpus included in **ESGenius**. The collection spans 2020–2023 and comprises 11 documents totaling **60 MB** across **728 pages**, yielding **123** benchmark questions. The materials include core guidance documents on implementation, metrics, targets, scenario analysis and risk management, a complete five-part workshop series covering fundamentals through metrics, and the 2023 status report. These documents establish the foundation for global climate-risk disclosure practices and form the conceptual framework adopted by ISSB’s IFRS S2. All materials are accessible at <https://www.fsb-tcfd.org/>.

Original Document Title	Year	Size	Pages	No. Qs
<i>CDP Full Corporate Questionnaire April 2025 – Modules 1–6</i> (cdp, 2025c)	2025	5.1 MB	447	29
<i>CDP Full Corporate Questionnaire April 2025 – Module 7</i> (cdp, 2025b)	2025	4.9 MB	482	41
<i>CDP Full Corporate Questionnaire April 2025 – Modules 8–13</i> (cdp, 2025d)	2025	4.7 MB	435	44
<i>CDP SME Questionnaire April 2025 – Modules 14–21</i> (cdp, 2025e)	2025	2.1 MB	174	38
<i>2025 CDP–ICLEI Track and States & Regions Questionnaire and Guidance</i> (cdp, 2025a)	2025	2.5 MB	430	31
<i>CDP Full Corporate Scoring Introduction 2024</i> (cdp, 2024)	2024	293 KB	19	2
Total	-	19.6 MB	1,987	185

Table 10: Comprehensive metadata for the **Carbon Disclosure Project (CDP)** knowledge base curated in **ESGenius**: the collection comprises six hyper-linked PDFs including the *2025 Full Corporate Questionnaire* split across three modules (114 questions, 1,364 pages), the *2025 SME Questionnaire* (38 questions, 174 pages), the *2025 CDP–ICLEI Track and States & Regions Questionnaire* (31 questions, 430 pages), and the *2024 Full Corporate Scoring Introduction* (2 scoring criteria, 19 pages). Together the corpus totals **19.6 MB** and spans **1,987 pages**, delivering **185 unique, standardised questions** that power the world’s largest voluntary platform for climate, water-security, and deforestation disclosure. All files are directly accessible via the linked titles, with consolidated resources available at <https://www.cdp.net/>.

Original Document Title	Year	Size	Pages	No. Qs
<i>The 2030 Agenda for Sustainable Development's 17 Sustainable Development Goals (SDGs)</i> (sdg, 2015)	2015	424 KB	19	1
<i>Transforming Our World: The 2030 Agenda for Sustainable Development</i> (tra, 2015)	2015	378 KB	41	4
<i>Progress towards the Sustainable Development Goals — Report of the Secretary-General</i> (sdg, 2024)	2024	518 KB	26	2
<i>The Sustainable Development Goals Report 2024</i> (sdg, 2024eb)	2024	8.6 MB	51	6
<i>Sustainable Development Goals – Briefing Book 2023 (UN Office for Partnerships)</i> (sdg, 2023)	2023	4.6 MB	35	7
<i>GAR Special Report 2023 — Mapping Resilience for the Sustainable Development Goals</i> (gar, 2023)	2023	19.5 MB	51	5
<i>Sustainable Development Report 2024 — The SDGs and the UN Summit of the Future (includes the SDG Index and Dashboards)</i> (sdr, 2024)	2024	39.3 MB	512	12
Total	-	73.33 MB	735	37

Table 11: Comprehensive metadata for the **United Nations Sustainable Development Goals (SDGs)** corpus incorporated in **ESGenius**: the set begins with the 2015 adoption texts—the 19-page plain-language overview of the 17 Goals and the 41-page General Assembly resolution “Transforming Our World” that enshrines the 2030 Agenda—then follows implementation through the Secretary-General’s *Progress towards the SDGs 2024* and the flagship *SDGs Report 2024*, is complemented by the UN Office for Partnerships’ *SDG Briefing Book 2023* and deepened by thematic analyses such as UNDRR’s 2023 *Global Assessment Report on Resilience* (GAR) and the independent *Sustainable Development Report 2024* with its widely cited SDG Index and global dashboards. Across these seven key PDFs—totalling **73.33 MB**, **735 pages**, and distilled into **37** benchmark MCQs that anchors national, corporate, and investor sustainability strategies in the universally agreed 17-goal framework. Full SDG resources are freely available at <https://sdgs.un.org/goals>.

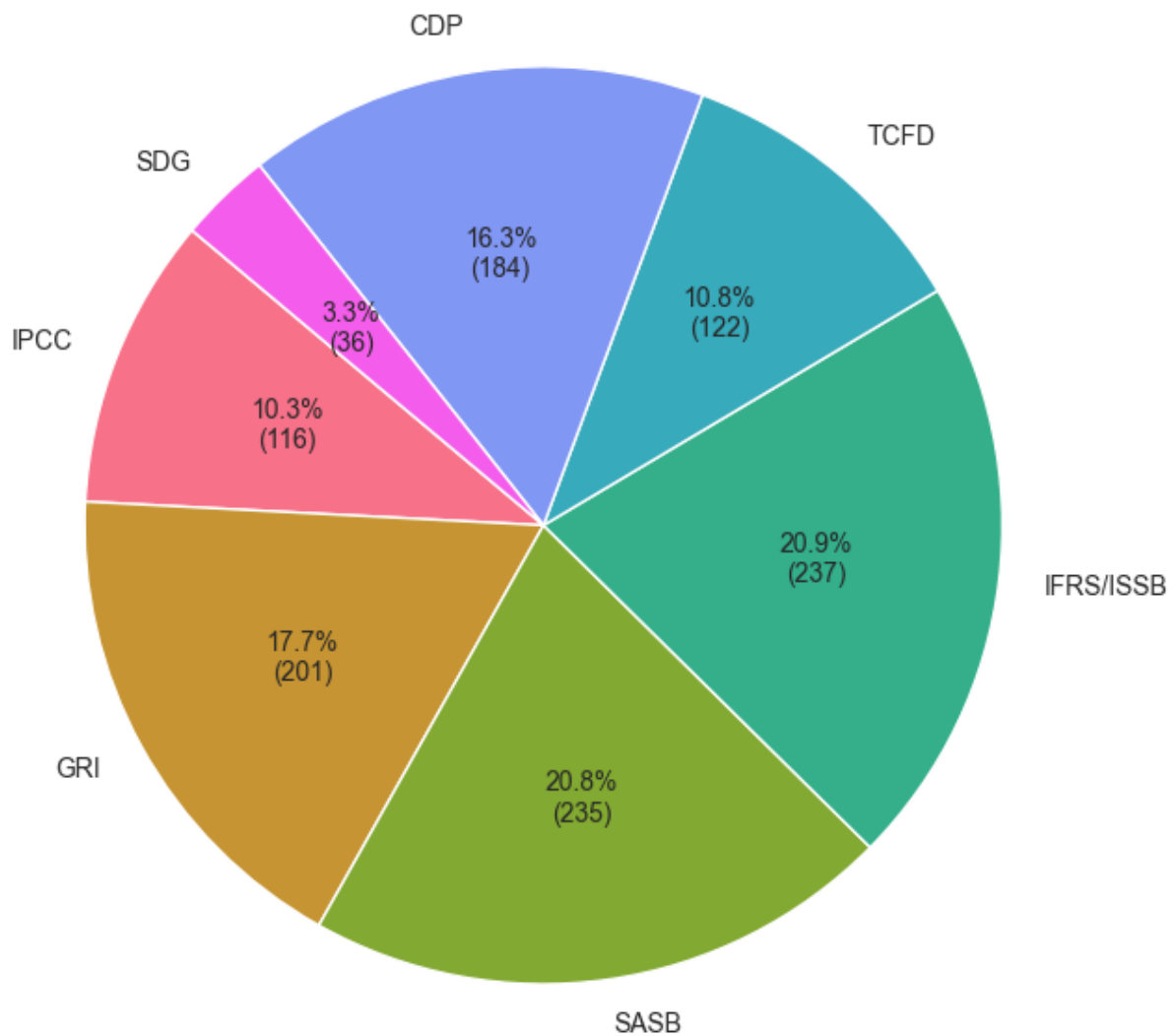


Figure 5: Relative **question distribution** (benchmark MCQs) derived from each source family. In contrast to the page distribution, questions are more evenly spread: the largest contributors are IFRS/ISSB (**20.9 %**, 237 Qs) and SASB (**20.8 %**, 235 Qs), followed by GRI (**17.7 %**, 201 Qs) and CDP (**16.3 %**, 184 Qs). TCFD (**10.8 %**, 122 Qs) and IPCC (**10.3 %**, 116 Qs) provide focused climate-risk and climate-science coverage, while the SDG set supplies a compact but essential sustainability anchor (**3.3 %**, 36 Qs).

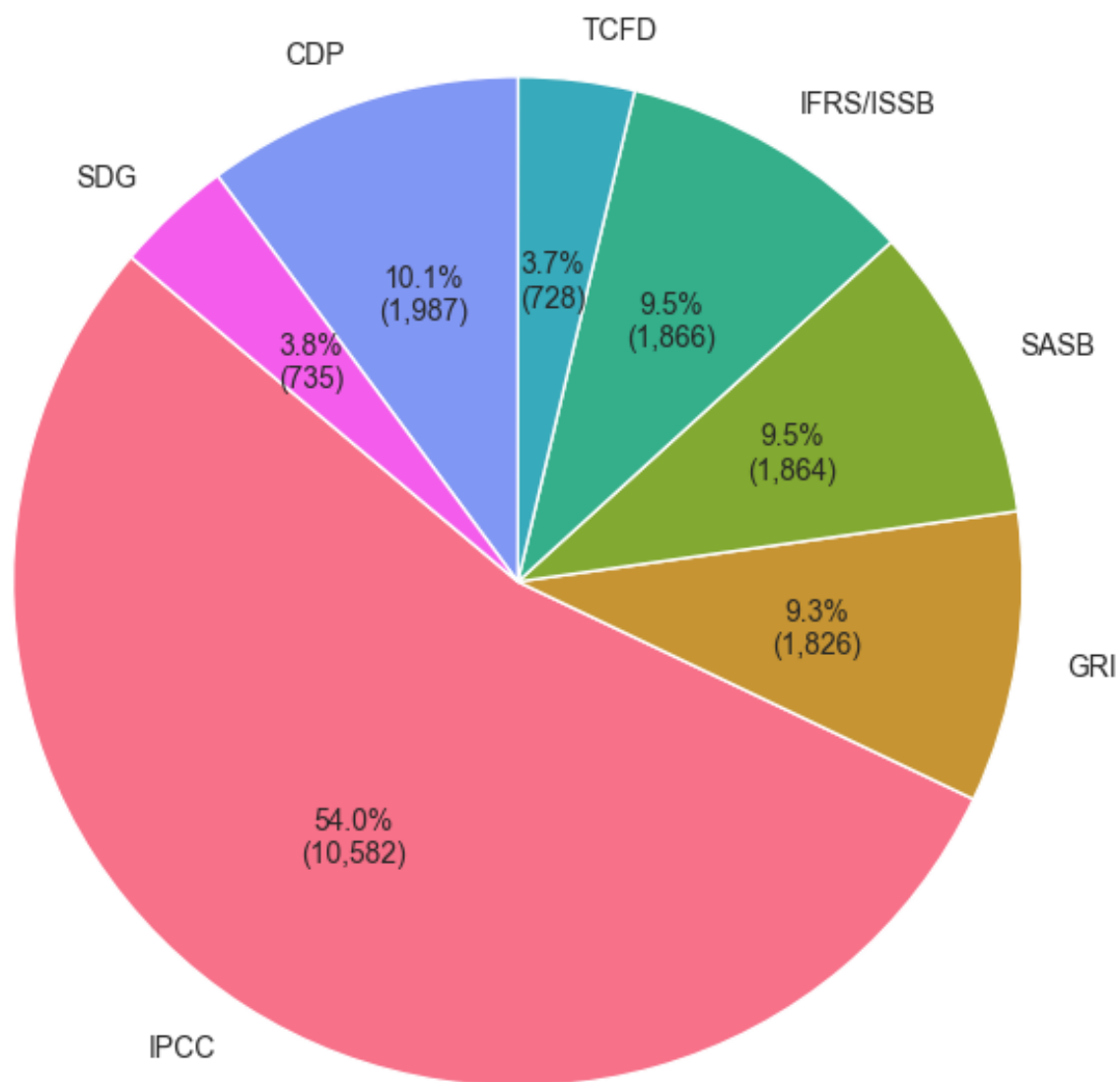


Figure 6: Relative **page-count distribution** of the **ESGenius-Corpus** across its 7 source families. The Inter-governmental Panel on Climate Change (IPCC) alone accounts for a majority of the material—**54 %**, or **10 582 pages**—reflecting the encyclopaedic scope of its assessment reports. Standards and disclosure frameworks contribute smaller but still substantive shares: CDP (**10.1 %**, 1 987 pp.), IFRS/ISSB (**9.5 %**, 1 866 pp.), SASB (**9.5 %**, 1 864 pp.), and GRI (**9.3 %**, 1 826 pp.). Guidance-oriented sources such as TCFD (**3.7 %**, 728 pp.) and the UN Sustainable Development Goals corpus (**3.8 %**, 735 pp.) round out the collection.

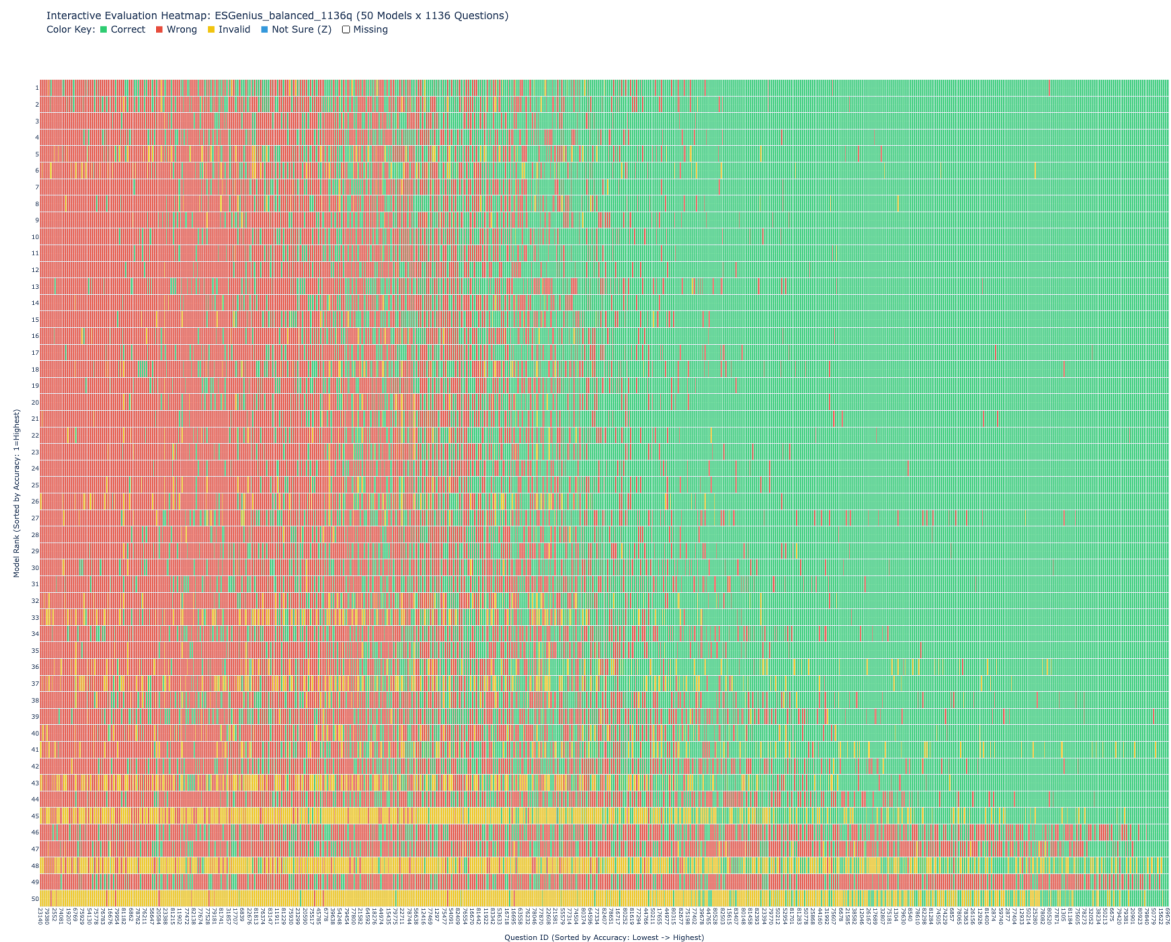


Figure 7: Example zero-shot performance heatmap showing model accuracy patterns across different ESG question types and topics within ESGenius (based on initial data). Darker colors indicate higher accuracy. Interactive version url will be provided in the camera-ready version.

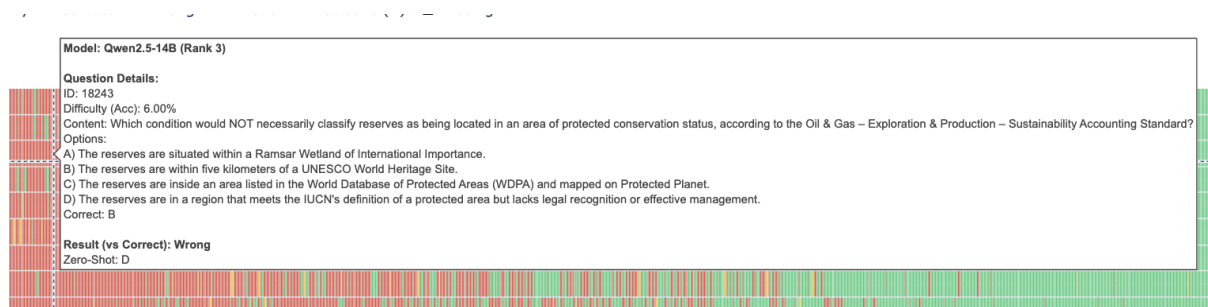


Figure 8: Conceptual information display for interactive heatmap cells (Figure 7), showing model and accuracy score.