SustainableQA: A Comprehensive Question Answering Dataset for Corporate Sustainability and EU Taxonomy Reporting

Anonymous ACL submission

Abstract

The growing demand for corporate sustainability transparency, particularly under new regulations like the EU Taxonomy, necessitates precise data extraction from large, unstructured corporate reports, a task for which Large Language Models and Retrieval-RAG systems require high-quality, domain-specific question-answering datasets. To address this, we introduce SustainableQA, a novel dataset and a scalable pipeline that generates comprehensive QA pairs from corporate sustainability and annual reports by integrating semantic chunk classification, a hybrid span extraction pipeline, and a specialized table-to-paragraph transformation. To ensure high quality, the generation is followed by a novel automated assessment and refinement pipeline that systematically validates each QA pair for faithfulness and relevance, repairing or discarding low-quality entries. This results in a final, robust dataset of over 195,000 diverse factoid and non-factoid QA pairs, whose effectiveness is demonstrated by initial fine-tuning experiments where a compact 8B parameter model significantly outperforms much larger state-of-the-art models. SustainableQA proves to be a highly effective resource for developing and benchmarking advanced knowledge assistants capable of navigating complex sustainability compliance data[1].

1 Introduction

The global financial landscape is undergoing a paradigm shift, driven by growing demands for corporate transparency in sustainability practices (Venturelli, 2024; Edwards, 2005). Regulatory frameworks such as the European Union's Corporate Sustainability Reporting Directive (CSRD) and the EU Taxonomy for Sustainable Activities (Odobaša and Marošević, 2023; Hummel and Jobst, 2024) require companies to provide detailed disclosures on their environmental and social impact. The CSRD mandates broad ESG reporting, while the EU Taxonomy offers a classification system for environmentally sustainable activities. These requirements have increased the length and complexity of corporate sustainability reports, which constitute a key source of information for financial and regulatory decision-making (Berg et al., 2022; Bronzini et al., 2024; Greenomy, 2025). Yet such reports are often published as unstructured PDFs, making the extraction of verifiable information resource-intensive and error-prone.

The research community has then increasingly turned to Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) systems (Zou et al., 2025; Wu et al., 2024; Bronzini et al., 2024; Setty et al., 2024; Wang et al., 2024). However, the efficiency of any such AI system is fundamentally dependent on the availability of high-quality, domain-specific training and evaluation data. This dependency highlights a significant research gap: the absence of publicly available, large-scale question-answering (QA) datasets specifically tailored for sustainability reporting (Ferjančič et al., 2024; He et al., 2025). This scarcity forces researchers and developers to either create small, private datasets for their experiments, a process that limits the generalizability and comparability of results or rely on general-domain models that lack the necessary expertise to navigate the nuances of regulatory terminology (Schimanski et al., 2024b).

The creation of such a dataset is challenging due to how information is distributed within reports and the complex structure of the EU Taxonomy itself. Our analysis revealed that sustainability information is highly fragmented in corporate reports. While our initial focus was on the EU Taxonomy, we observed that crucial data points and justifications are frequently located within broader ESG and general sustainability chapters. A narrow focus on "EU Taxonomy" sections alone would therefore yield an incomplete and decontextualized set of

[1] Code and data will be made publicly available.

QA pairs. This observation prompted an expansion of our scope: to build a useful resource, we must construct a comprehensive knowledge base that interconnects the three critical pillars—EU Taxonomy, ESG, and Sustainability.

The challenge is further compounded by EU Taxonomy information being scattered across poorly formatted, multi-page tables that require extensive cross-referencing with narrative sections throughout reports. To address these difficulties, we developed a novel pipeline to generate high-quality QA dataset from complex corporate sustainability reports. We employ a multi-stage process that includes: (1) semantic passage classification to identify relevant content across the entire document; (2) a hybrid span extraction pipeline that combines a fine-tuned NER model, rule-based methods, and LLM-driven refinement to ensure the precise identification of factoid answers; and (3) a specialized table-to-paragraph transformation strategy that leverages a large-context model to make complex tabular data accessible for QA generation.

2 Related Work

Question-answering (QA) research has evolved from general-purpose benchmarks like SQuAD (Rajpurkar et al., 2016), Natural Questions (Kwiatkowski et al., 2019), and HotpotQA (Yang et al., 2018) toward domain-specific datasets addressing specialized reasoning requirements.

Corporate Finance QA. The corporate reporting domain was pioneered by finance-focused datasets. FinQA (Chen et al., 2021), TAT-QA (Zhu et al., 2021), and MultiHiertt (Zhao et al., 2022) established paradigms for hybrid text-table reasoning and numerical computation over semi-structured financial documents, demonstrating the importance of multi-modal evidence integration.

Sustainability and ESG QA. As sustainability disclosure requirements proliferated, research adapted to ESG contexts. Systems like ChatReport (Ni et al., 2023) emphasized evidence provenance, while ClimRetrieve (Schimanski et al., 2024a) and GRI-QA (Contalbo et al., 2025) provided retrieval benchmarks and table-based QA aligned with climate disclosures and reporting standards.

The EU Taxonomy Gap. However, existing work addresses voluntary ESG frameworks. The EU Taxonomy differs fundamentally by establishing legally binding disclosure requirements under the CSRD (European Parliament and Council, 2022), mandating structured reporting of activity eligibility, alignment, and compliance with technical screening criteria. While conceptual frameworks for EU Taxonomy automation have been proposed (Hetfleisch et al., 2024) and limited classification datasets exist for specific sectors (Birti et al., 2025), no public resource supports comprehensive question answering over real-world EU Taxonomy disclosures.

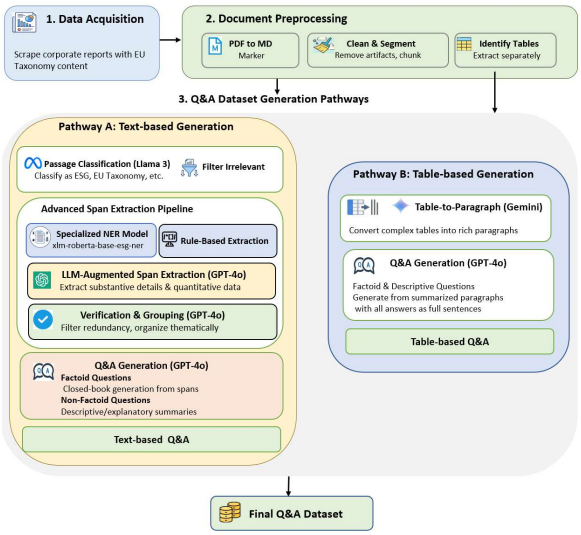

Figure 1: SustainableQA dataset generation pipeline.

SustainableQA addresses this gap by offering more than 195,000 QA pairs derived from 61 corporate sustainability and EU Taxonomy reports, supporting both fact-based questions (e.g., "What percentage of CapEx is aligned?") and explanatory ones (e.g., "Why does this activity fail DNSH?"), enabling research in hybrid reasoning, table-to-text grounding, and regulation-aware QA.

Table 1 summarizes prior datasets; additional discussion is given in Appendix A.

3 Dataset Construction and Analysis

As shown in Figure 1, the dataset pipeline is structured as a multi-stage process encompassing data acquisition, preprocessing, content classification, and question-answering generation.

3.1 Data Acquisition

The initial phase involves the acquisition of relevant corporate documents. This was achieved through web scraping of stock exchange websites to gather publicly available annual reports and standalone sustainability reports. We focused on companies listed on European stock exchanges with dedicated EU Taxonomy sections, prioritizing re-

Table 1: Comparison of datasets, systems, and tools for sustainability reporting NLP. Scale shows dataset size with source document count in parentheses where applicable.

Resource	Type	Task	Scale	Focus Area	Key Contribution
ChatReport	System	QA with citations	1,015 reports	ESG (TCFD)	Traceable evidence provenance
ReportParse	Toolkit	Parsing/segmentation	—	ESG	Unified parsing infrastructure
ClimRetrieve	Dataset	Retrieval benchmark	8,500+ triads (30 reports)	Climate	Relevance-labeled retrieval
Climate Finance Bench	Dataset	QA evaluation	330 pairs (33 reports)	Climate/ESG	Analyst-style questions in RAG
GRI-QA	Dataset	Table QA	4,089 pairs	GRI standards	Calculated answers from GRI tables
ESG-Kor	Dataset	Sentence classification	118,946 sentences	ESG (Korean)	Cross-lingual ESG extraction
CarbonPDF-QA	Dataset	PDF QA	1,735 documents	Carbon footprint	Robust QA from noisy PDFs
ESG-Activities	Dataset	Activity classification	1,325 segments	EU Taxonomy (transport)	Activity-level labeling
SustainableQA	**Dataset**	**Multi-type QA**	**195,287 pairs (61 reports)**	**EU Taxonomy+ESG+Sustainability**	**Factoid, non-factoid, & table QA**

ports that also contain ESG or broader sustainability content, yielding 61 corporate reports.

3.2 Document Preprocessing

Raw PDF reports undergo streamlined preprocessing to convert them into a structured, clean, and manageable format. Each PDF is first transformed into Markdown text using the Marker library (Paruchuri, 2024), which preserves structural elements. The Markdown is then cleaned to remove non-substantive elements such as footnotes, images, and page markers, while consolidating blank lines and removing empty heading sections. Finally, the cleaned text is segmented into semantically coherent passages based on markdown headings, with a word-count constraint (e.g., `max_words=350`) applied to ensure each passage remains within an optimal context window for subsequent LLM-based processing. In parallel, tables are identified within the Markdown output to be processed via a specialized pipeline.

3.3 Question-Answering Dataset Generation

The core Q&A generation process integrates content classification, advanced span extraction, and the creation of diverse question types to maximize information coverage (see Appendix C for all prompts used in the generation pipeline).

Passage Classification. To ensure relevance and aggregate dispersed information, each passage undergoes classification using Llama 3.3 (70B) (Grattafiori et al., 2024) into four categories: "EU Taxonomy," "ESG," "Sustainability," or "Unknown." Unknown passages are filtered out to retain only domain-relevant content scattered throughout reports.

3.3.1 Advanced Span Extraction Pipeline for Factoid Q&A

For passages classified as relevant, a multi-stage, hybrid pipeline is used for extracting key text spans that will serve as answers for factoid questions.

Specialized NER Model Application and Fine-tuning for span Extraction: The initial pass leverages a pre-trained Named Entity Recognition (NER) model, `xlm-roberta-base-esg-ner` (Vutukuri, 2023), which is designed for ESG-related entity recognition. To boost its performance, we fine-tune this model on the "ESG-only" subset of the `ExponentialScience/ESG-DLT-NER` dataset (Exponential Science, 2023), focusing on `B-ESG` and `I-ESG` tags. This optimized the model's ability to accurately identify domain-specific ESG and sustainability concepts.

Rule-Based and Dictionary-Based span Extraction: To capture entities and patterns potentially missed by the NER model, we leverage regular expressions for highly structured data (e.g., regulations, standards, quantitative data). This is complemented by a spaCy *PhraseMatcher* operating with comprehensive ESG and sustainability dictionaries to identify key phrases, thereby ensuring robust candidate span detection.

Two-Stage LLM-Driven Refinement: The initial collected candidate spans are then subjected to a two-stage extension and refinement process using GPT 4o (Achiam et al., 2023). **Stage 1: LLM-Augmented Span Extraction.** In the first stage we extend the initial set of candidate spans using LLM. The LLM is prompted to additionally extract substantive details, quantitative data, and specific regulatory terms from the passages, explicitly excluding generic or trivial phrases. These LLM-generated spans are then aggregated with the previously collected spans (the outputs from the NER, rule-based, and *PhraseMatcher* components), forming an extended set of initial candidates of answers based on which we will generate questions.

Stage 2: Contextual Verification, Filtering, and Thematic Organization. The candidate spans undergo then the second LLM-based processing stage that performs three functions: (1) contextual

verification to ensure that the spans are actually present in the source text, (2) filtering to eliminate redundant or suboptimal entries, and (3) thematic grouping of spans into semantically coherent clusters with descriptive labels. This grouping strategy is introduced so that questions can be generated not only based on individual spans used as their answers but also on the groups of multiple spans that are semantically related. The latter helps to create complex questions which require answers composed of multiple related spans.

3.3.2 Text-based Question-Answering

With relevant passages identified and key spans extracted, we can now generate diverse QA pairs (see Appendix E for examples) for each passage using advanced LLMs.

Factoid Q&A Generation: For every passage, we generate comprehensive factoid QA pairs using GPT-4o through a structured approach. First, we create questions based on individual spans. Next, we create group-level questions that require multiple spans as complete answers from each thematic cluster. All questions maintain exact correspondence to their extracted passages, ensuring direct answerability from the provided context while following "closed-book" constraints, thus guaranteeing accurate and verifiable responses across different structural types.

Non-Factoid Q&A Generation: In addition to factoid questions, we create non-factoid (descriptive/explanatory) QA pairs for each passage using GPT-4o. These questions require comprehensive textual analysis rather than isolated fact retrieval, eliciting detailed answers that explain relationships, describe processes, define concepts, or discuss implications within the passage. The generated responses typically span 1-4 sentences.

3.3.3 Table-based Question-Answering

Finally, a specialized approach is integrated to generate questions from tabular data.

Table-to-Passage Transformation: Given the large and complex tables in corporate reports, especially those related to EU Taxonomy that often span multiple pages, we convert each table into clear summarized passages using Gemini 2.5 Flash Chat (Google DeepMind, 2025), leveraging the model's large context window capability. This process extracts essential tabular data alongside contextual information from surrounding textual content.

QA Generation from Transformed Passages: Following the table-to-passage transformation, we generate QA pairs with `GPT-4o`, encompassing direct numerical queries, explanatory questions about regulatory relationships, and comprehensive questions requiring integration of multiple data points. We standardize tabular answers as complete sentences to enhance evaluation robustness and preserve EU Taxonomy-specific information given the limited tabular passages relative to text passages.

3.4 Dataset Composition and Analysis

Our data generation pipeline produced a comprehensive dataset of 195,287 QA pairs sourced from 61 corporate reports. As detailed in Table 2, the dataset is composed of 88,792 factoid (F) and 102,539 non-factoid (NF) questions derived from 8,067 text passages, complemented by 3,956 QA pairs extracted from 218 tables. The content is distributed across three key categories: ESG, EU Taxonomy, and Sustainability. A key characteristic of the dataset is the distinction in answer length: factoid answers are concise (avg. 4.2 words), while non-factoid answers are descriptive (avg. 32.5 words), targeting contextual understanding.

To assess the complexity of the generated factoid questions, we analyzed the distribution of answer spans across all 88,792 factoid QA pairs. The analysis reveals that while the vast majority (83.3%) of questions are answered by a single, contiguous text span, the dataset also includes complex questions requiring multiple spans, with a heavy-tailed distribution extending up to 10 spans. This diversity challenges models to perform both simple entity extraction and more complex information aggregation, with over 95% of all factoid questions answerable with four or fewer spans.

When disaggregated by category (Figure 2), questions related to the EU Taxonomy are demonstrably more complex, exhibiting the highest mean number of spans (1.45) and the largest share of multi-span answers (21.2%), as detailed in Table 3. ESG questions show similar complexity patterns with 16.9% requiring multiple spans, while Sustainability questions are relatively simpler with only 15.4% multi-span answers.

4 Automated Quality Assessment and Refinement

To ensure the reliability and utility of the dataset for training and evaluating specialized language models, we introduce a comprehensive quality as-

Table 2: Overall dataset statistics.

Category	passages	QA Pairs			Avg. Length (words)		
		F	NF	Total	Q-F	Q-NF	Ans-F/NF
ESG	4,320	48,260	55,139	103,399	12.2	13.6	4.2/32.5
EU Tax.	747	8,260	8,906	17,166	12.7	14.5	4.7/33.5
Sustain.	3,000	32,272	38,494	70,746	12.1	13.4	4.0/32.0
Text Subt.	**8,067**	**88,792**	**102,539**	**191,331**	**12.2**	**13.6**	**4.2/32.5**
Tables	218	3,956		3,956	15.8		23.6
Total	**8,285**	—		**195,287**	—		

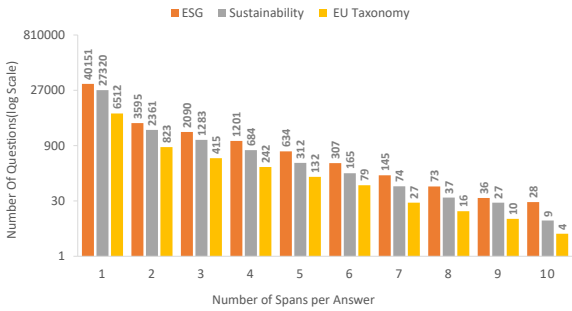

Figure 2: Span distribution by category.

Table 3: Summary of answer spans for factoid questions.

Category	Mean	Med.	Std	Single	Multi
Overall	1.36	1	1.02	83.3%	16.7%
ESG	1.37	1	1.06	83.1%	16.9%
EU Tax.	1.45	1	1.12	78.8%	21.2%
Sustain.	1.32	1	0.94	84.6%	15.4%

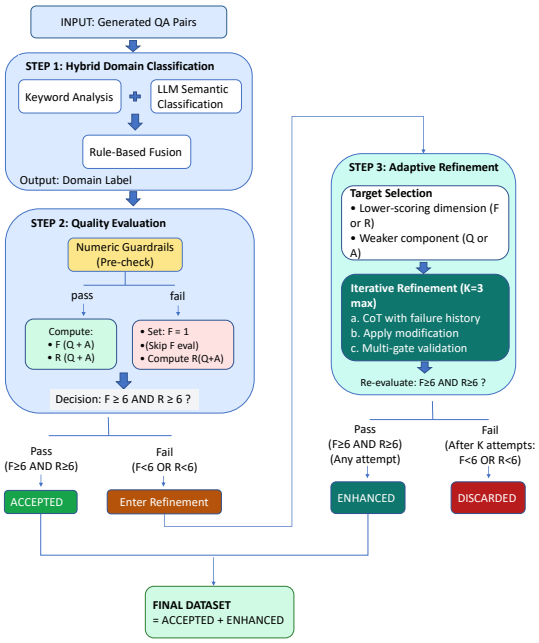

Figure 3: Three-stage quality assessment and refinement pipeline with variables: F (faithfulness score), R (relevance score), Q (question), A (answer), K (maximum refinement attempts).

sessment and refinement framework that addresses critical challenges in automatically generated QA datasets, including hallucinations, domain misalignment, and factual inconsistencies.

Our methodology begins with a hybrid domain classification approach that combines curated keyword analysis with LLM-based semantic understanding through rule-based fusion logic, ensuring accurate domain labeling before quality assessment, as incorrect labels would invalidate all subsequent relevance assessments (Figure 3).

We then employ a dual-metric evaluation architecture inspired by RAGAS (Es et al., 2023) that independently assesses *faithfulness* (factual grounding and freedom from hallucination) and *relevance* (domain-specific utility for sustainability professionals) for both questions and answers. We use a suite of specialized LLM prompts for evaluation, modification, and correction tasks, with representative examples in Appendix D and the complete set is submitted as supplementary material. The assessment adapts to question types, employing span-level analysis for factoid questions, conceptual grounding evaluation for non-factoid questions, and numerical fidelity checks for table-based questions. Both metrics use 1–10 scoring scales, with overall scores computed by averaging their respective question and answer components.

Critically, for faithfulness assessment, we implement deterministic numeric guardrails that validate quantitative claims before LLM evaluation, enforcing strict tolerances to account for rounding differences and measurement precision (± 0.5 percentage points for percentages, $\leq 2\%$ relative error for other numerical values) to prevent hallucinations in sustainability reporting data. When numeric guardrails fail, faithfulness is set to 1.0 (faithfulness failure), and only relevance is evaluated to determine whether the pair requires faithfulness-only or dual-dimension refinement. QA pairs are accepted into the final dataset only if both overall faithfulness and relevance scores meet a configurable quality threshold (default 6.0), determined empirically to balance data quality with retention rates.

QA pairs that score below the acceptance threshold on either dimension are not discarded but instead enter an adaptive refinement pipeline designed to repair quality issues through targeted modification strategies. The system employs a hierarchical approach to identify the primary source of quality degradation, first targeting the lower-scoring dimension (faithfulness or relevance), then focusing on the weaker component (question or an-

swer) within that dimension. For factoid answers, we implement targeted span refinement that modifies only problematic spans rather than regenerating entire answers, extracting and correcting non-verbatim or low-relevance spans individually to significantly improve efficiency. The refinement process incorporates chain-of-thought (CoT) prompting with explicit failure awareness, inspired by self-refinement approaches (Madaan et al., 2023), where the system learns from unsuccessful attempts to construct increasingly specific prompts for subsequent retries (up to K=3 attempts).

Every modification must pass a multi-gate validation protocol (including length appropriateness, readability, coherence, and numeric consistency checks) before re-evaluation, ensuring quality improvement without introducing new errors. When all refinement attempts fail to elevate a QA pair above the acceptance threshold, the pair is deleted from the dataset.

Applying the above-described framework to 205,600 initial QA pairs achieved an overall retention rate of 95.0%, improving dataset quality from 74.0% initial acceptance to 89.8% final quality across 195,287 retained pairs, with successfully refined pairs demonstrating average improvements of +1.42 for faithfulness and +1.19 for relevance. Complete methodology details, including evaluation rubrics, validation protocols, and scoring formulas, are provided in Appendix B.

5 Experiments and Results

We conducted evaluations across multiple dimensions. First, we benchmarked retrieval methods to assess SustainableQA's utility for RAG systems. Second, we evaluated seven state-of-the-art language models under three prompting strategies: zero-shot, zero-shot with context, and few-shot learning(one example per question type). The dataset was partitioned into training (80%), validation (10%), and test (10%) splits, maintaining balanced category distribution.

For factoid questions, we employed span-based evaluation metrics including Exact Match (EM), Precision, Recall, and F1-score. For non-factoid and tabular questions, we utilized BERTScore, BLEU, ROUGE-L, and METEOR to evaluate semantic similarity, lexical overlap, and text quality.

Table 4 presents evaluation results across all models, prompting strategies, and question types, revealing several key findings:

Fine-tuning Superiority: The fine-tuned Llama (Llama 3.1 8B-FT, trained with context) achieved best performance across most metrics: 56.84% EM, 59.70% F1 for factoid questions, and superior performance in non-factoid (91.70% BERTScore, 55.98% ROUGE-L, 60.99% METEOR) and tabular questions (86.83% BERTScore, 75.22% ROUGE-L, 78.76% METEOR). This demonstrates the dataset's effectiveness for domain-specific adaptation.

Prompting Strategy Effectiveness: Zero-shot with context consistently outperformed both pure zero-shot and few-shot approaches across all models. For factoid questions, Zero+Context achieved substantially higher exact match rates (10.12-43.15%) compared to zero-shot (0.28-2.89%) and few-shot (6.73-17.36%) strategies. For non-factoid questions, Zero+Context showed dramatic improvements in BERTScore (84.61-89.97%) versus zero-shot (74.84-81.96%) and few-shot (81.52-86.42%), suggesting that domain-specific context is more valuable than general examples for sustainability QA.

Model Performance Hierarchy: Among base models, GPT-4o and Llama 3.3 70B demonstrated superior performance across most configurations. Llama 3.3 70B achieved the highest recall (60.76%) in factoid questions under contextual prompting, while GPT-4o excelled in tabular question answering with 87.24% BERTScore, 68.28% ROUGE-L and 71.70% METEOR scores.

Task-Specific Challenges: Factoid questions proved challenging, with exact match scores ranging from 0.28% to 56.84%, highlighting the complexity of precise span extraction in sustainability reporting. However, substantially higher F1 scores indicate that models capture relevant information despite imperfect span alignment, suggesting understanding of sustainability concepts.

5.1 Retrieval Performance Benchmark

To validate SustainableQA's utility as a benchmark for Retrieval-Augmented Generation (RAG), we evaluated the performance of four retrieval models on the dataset's test split. We benchmarked the lexical model BM25 against three popular dense retrievers: DPR-single, BGE, and Contriever. Performance was assessed using standard metrics, including Recall@k, NDCG@k, and Mean Reciprocal Rank (MRR).

As detailed in Table 5, the results reveal a clear performance hierarchy. The lexical retriever,

Table 4: Model Performance across Question Types and Prompting Strategies. Bold = best within strategy; Underline = overall best.

Strategy	Model	Factoid				Non-Factoid				Tabular			
		EM	F1	P	R	BERT	ROU	MET	BLU	BERT	ROU	MET	BLU
Zero-shot	GPT-4o (Achiam et al., 2023)	**2.89**	**4.12**	**4.05**	4.38	76.15	14.68	15.21	4.12	**74.82**	**61.34**	**64.57**	**34.43**
	Llama 3.3 70B (Grattafiori et al., 2024)	1.78	3.64	3.19	**4.71**	78.03	18.93	21.28	7.89	73.15	28.59	37.50	14.40
	Qwen2.5 7B (Yang et al., 2025)	0.47	0.62	0.61	0.65	79.87	21.98	24.76	9.49	71.84	36.11	45.77	16.59
	Gemma3 12B (Team et al., 2025)	2.05	3.48	3.24	4.10	81.12	**26.19**	27.19	10.48	72.47	29.49	31.11	13.95
	Llama 3.1 8B (Grattafiori et al., 2024)	0.74	1.53	1.27	2.40	74.84	12.25	13.95	4.79	71.26	33.61	37.92	18.35
	Mistral 7B (Jiang et al., 2023)	0.28	0.54	0.49	0.68	**81.96**	25.28	**29.91**	**12.12**	71.53	39.95	47.23	19.14
Zero+Context	GPT-4o (Achiam et al., 2023)	41.52	52.19	50.38	57.94	89.71	47.90	**56.96**	**29.29**	**87.24**	**68.28**	**71.70**	**52.26**
	Llama 3.3 70B (Grattafiori et al., 2024)	42.89	**54.68**	**52.75**	**60.76**	89.16	44.22	55.62	23.21	86.58	60.77	63.30	52.19
	Qwen2.5 7B (Yang et al., 2025)	31.24	38.43	38.04	40.53	84.61	35.38	43.35	23.52	83.47	57.45	66.06	35.85
	Gemma3 12B (Team et al., 2025)	**43.15**	53.19	51.93	57.54	**89.93**	**47.07**	54.58	28.48	84.92	47.35	41.70	31.14
	Llama 3.1 8B (Grattafiori et al., 2024)	32.18	46.80	43.94	54.85	89.97	46.10	54.60	26.22	84.18	55.25	55.46	41.83
	Mistral 7B (Jiang et al., 2023)	10.12	15.72	15.15	17.72	85.98	34.59	46.44	16.47	83.71	66.69	74.09	36.67
Few-shot	GPT-4o (Achiam et al., 2023)	**17.36**	**23.51**	**22.74**	**24.87**	**86.42**	**34.87**	**38.94**	**18.76**	**81.29**	**64.84**	**66.71**	**37.75**
	Llama 3.3 70B (Grattafiori et al., 2024)	13.28	19.17	18.46	20.59	85.07	31.52	34.68	16.34	79.51	40.66	41.10	21.55
	Qwen2.5 7B (Yang et al., 2025)	7.82	11.94	11.47	12.73	82.18	27.83	33.24	13.47	76.92	37.00	51.00	14.70
	Gemma3 12B (Team et al., 2025)	10.47	16.23	15.61	17.38	83.69	29.94	31.17	14.82	78.16	35.22	35.51	14.58
	Llama 3.1 8B (Grattafiori et al., 2024)	8.94	14.76	14.18	15.83	81.52	21.74	23.29	10.18	76.48	30.52	29.83	13.87
	Mistral 7B (Jiang et al., 2023)	6.73	10.81	10.37	11.59	81.87	26.34	33.92	10.26	76.27	37.63	36.95	12.93
Finetune	Llama 3.1 8B-FT	**56.84**	**59.70**	**59.95**	**60.14**	**91.70**	**55.98**	**60.99**	**37.99**	**86.83**	**75.22**	**78.76**	**57.56**

BM25, achieved the strongest results across both factoid and non-factoid questions. Among the dense models, BGE was the most competitive, while DPR and Contriever showed substantially lower performance.

This performance variance highlights the distinct challenges of the sustainability domain. The success of BM25 is attributable to its strength in exact lexical matching, which is highly effective for the specialized terminology, regulatory identifiers (e.g., "CSRD," "DNSH"), and numerical KPIs prevalent in the reports. Conversely, the performance gap for dense models like DPR and Contriever stems from a dual mismatch: First, a **domain mismatch**, as their general pre-training lacks exposure to specialized sustainability language. Second, a **structural mismatch**, as these models are typically trained on shorter texts (∼100 words). When forced to encode our longer passages (up to 350 words) into a single fixed-size vector, the critical details required for an exact match can become less prominent, hindering precise retrieval. BGE's stronger performance suggests that more recent dense models are better at adapting to this domain, but these results confirm that handling long-form, terminology-dense documents remains a key challenge.

Table 5: Retrieval performance on SustainableQA test set. Best results in **bold**.

Type	Method	R@1	R@5	R@10	N@5	N@10	MRR
Factoid	BM25	**55.4**	**75.3**	**80.6**	**66.3**	**68.0**	**64.4**
	BGE	38.4	60.0	67.5	49.9	52.4	48.4
	DPR	12.5	25.1	31.6	19.0	21.1	18.9
	Contriever	12.8	26.3	33.6	19.8	22.2	19.7
Non-Fact.	BM25	**50.6**	**72.8**	**78.9**	**62.6**	**64.6**	**60.6**
	BGE	47.4	70.6	77.2	59.9	62.0	57.9
	DPR	15.7	30.0	37.8	23.1	25.6	22.9
	Contriever	14.3	29.5	37.1	22.3	24.7	21.9

5.2 Impact of Answer Span Complexity

To assess complexity when multi-component answers are required, we analyzed the best-performing models on factoid questions grouped by the number of answer spans (1–5 spans, covering 95% of questions). As shown in Table 6, performance degrades gradually with increasing span count. Llama 3.1 8B-FT achieves 59.4% EM on single-span questions, declining to 48.7% on 5-span questions (18.0% relative degradation). GPT-4o and Llama 3.3 70B exhibit similar patterns with 18.3% and 17.5% degradation respectively. F1 scores show greater stability with only 6–7% degradation across all models, indicating that models maintain strong partial matching even when exact matches become harder.

These results demonstrate that while multi-span extraction increases difficulty, the degradation remains modest and consistent across different model architectures, validating the dataset's suitability for training and evaluating QA systems on complex sustainability reporting tasks.

Table 6: Performance degradation across span complexity.

Model	EM (%)					F1 (%)					Deg 1 vs 5 (%)	
	1	2	3	4	5	1	2	3	4	5	EM	F1
Llama 3.1 8B-FT	59.4	57.1	54.6	51.8	48.7	61.0	60.2	59.3	58.2	56.9	18.0	6.7
GPT-4o	43.6	41.8	39.9	37.8	35.6	53.5	52.8	52.0	51.1	50.1	18.3	6.4
Llama 3.3 70B	45.1	43.4	41.5	39.4	37.2	55.9	55.2	54.4	53.5	52.5	17.5	6.1

5.3 Human Evaluation

To assess the dataset quality, we conducted human evaluation with three annotators who possess sustainability related knowledge on 3,000 stratified QA pairs across ESG (1,200), EU Taxonomy (600), and Sustainability (1,200) categories, including factoid (1,800) and non-factoid (1,200) question types.

Evaluators rated four dimensions on 5-point Likert scales, with evaluation criteria adapted to question type: **Question Quality** assessed clarity and answerability; **Answer Accuracy** evaluated exact span correctness for factoid questions and factual grounding for non-factoid questions; **Context Appropriateness** verified that answers could be derived from the provided context; and **Practical Utility** measured relevance for sustainability reporting tasks. The evaluation achieved substantial inter-annotator agreement across all dimensions: Question Quality (Krippendorff $\alpha = 0.76$), Answer Accuracy ($\alpha = 0.79$), Context Appropriateness ($\alpha = 0.74$), and Practical Utility ($\alpha = 0.71$).

Table 7 presents the detailed evaluation results. SustainableQA demonstrated high quality ratings with overall scores of 4.3/5.0 for Question Quality, 4.2/5.0 for Answer Accuracy, 4.1/5.0 for Context Appropriateness, and 3.9/5.0 for Practical Utility. Factoid questions consistently outperformed non-factoid variants across all dimensions (4.3 vs. 3.8 average), reflecting their greater verifiability and objective nature. Among domains, EU Taxonomy pairs achieved the highest scores across all dimensions (average 4.4/5.0), followed by ESG (4.1/5.0) and Sustainability (4.0/5.0), indicating strong alignment with regulatory reporting requirements. Answer span complexity analysis showed quality perception decline from single-span (4.5/5.0) to five-span questions (4.0/5.0), confirming that extraction difficulty correlates with human-assessed quality scores.

5.4 Applicability

SustainableQA is designed to support both academic and industry use cases at the intersection of AI and sustainable finance. For academic research, it provides a benchmark for training and evaluating domain-adapted language models capable of handling complex, multi-span questions grounded in EU sustainability regulations. For external stakeholders such as regulators, analysts, and investors, it enables the development of QA and RAG systems that extract and verify critical ESG and Taxonomy-aligned data with traceable evidence. Internally, companies can leverage the dataset to train AI assistants that streamline sustainability reporting and automate compliance checks under CSRD and EU Taxonomy disclosure requirements.

Table 7: Human evaluation results across dimensions, question types, and domains (5-point Likert scale).

Category	Quality	Accuracy	Context	Utility
Overall	4.3	4.2	4.1	3.9
By Question Type				
Factoid	4.5	4.4	4.3	4.1
Non-factoid	4.0	3.9	3.8	3.6
By Domain				
ESG	4.3	4.2	4.1	3.9
EU Taxonomy	4.6	4.5	4.4	4.2
Sustainability	4.2	4.1	4.0	3.8

6 Conclusion

We introduced in this paper, SustainableQA, a large-scale, comprehensive QA dataset designed to address the critical need for high-quality training and evaluation data in corporate sustainability and EU Taxonomy reporting. Our comprehensive evaluations demonstrate that fine-tuning on SustainableQA significantly enhances model performance, enabling a compact 8B parameter model to outperform significantly larger state-of-the-art models across different prompting strategies. The detailed analysis of answer complexity points to the challenges of multi-span extraction and establishes the dataset as a robust benchmark for evaluating model capabilities in this domain.

Limitations

While SustainableQA represents a significant contribution to sustainability reporting question answering, some limitations should be acknowledged. First, the dataset's scope is primarily focused on European corporate sustainability reports adhering to EU Taxonomy regulations, which may limit generalizability to other regulatory frameworks and international contexts. Second, the current dataset focuses exclusively on textual content and tabular data, excluding images, charts, and infographics that frequently appear in sustainability reports and often convey critical information such as trend visualizations and performance metrics. Future work could incorporate multimodal question answering capabilities to address visual information extraction.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Florian Berg and 1 others. 2022. Aggregate confusion: The divergence of esg ratings. Review of Finance, 26(5):1315–1344.

Mattia Birti, Francesco Osborne, and Andrea Maurino. 2025. Optimizing large language models for esg activity detection in financial texts. arXiv preprint arXiv:2502.21112.

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.

Marco Bronzini, Carlo Nicolini, Bruno Lepri, Andrea Passerini, and Jacopo Staiano. 2024. Glitter or gold? deriving structured insights from sustainability reports via large language models. EPJ Data Science, 13(1):41.

Jian Chen, Peilin Zhou, Yining Hua, Yingxin Loh, Kehui Chen, Ziyuan Li, Bing Zhu, and Junwei Liang. 2024. Fintextqa: A dataset for long-form financial question answering. arXiv preprint arXiv:2405.09980.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and 1 others. 2021. Finqa: A dataset of numerical reasoning over financial data. arXiv preprint arXiv:2109.00122.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. arXiv preprint arXiv:2210.03849.

Michele Luca Contalbo, Sara Pederzoli, Francesco Del Buono, Venturelli Valeria, Francesco Guerra, and Matteo Paganelli. 2025. Gri-qa: a comprehensive benchmark for table question answering over environmental data. In Findings of the Association for Computational Linguistics: ACL 2025, pages 15764–15779.

Andres R Edwards. 2005. The sustainability revolution: Portrait of a paradigm shift. new society publishers.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. Ragas: Automated evaluation of retrieval augmented generation. In arXiv preprint arXiv:2309.15217.

European Parliament and Council. 2022. Directive (eu) 2022/2464 of the European Parliament and of the Council of 14 December 2022 amending Regulation (eu) no 537/2014, Directive 2004/109/ec, Directive 2006/43/ec and Directive 2013/34/eu, as regards corporate sustainability reporting. Text with EEA relevance.

Exponential Science. 2023. Esg-dlt-ner dataset. Named Entity Recognition dataset for ESG and Distributed Ledger Technology.

Urša Ferjančič, Riste Ichev, Igor Lončarski, Syrielle Montariol, Andraž Pelicon, Senja Pollak, Katarina Sitar Šuštar, Aleš Toman, Aljoša Valentinčič, and Martin Žnidaršič. 2024. Textual analysis of corporate sustainability reporting and corporate esg scores. International Review of Financial Analysis, 96:103669.

Google DeepMind. 2025. Gemini 2.5 flash. https://deepmind.google/technologies/gemini/flash/. Accessed: [Your access date].

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.

Greenomy. 2025. The essentials of the EU taxonomy: A guide to accelerate your green transition.

Chaoyue He, Xin Zhou, Yi Wu, Xinjia Yu, Yan Zhang, Lei Zhang, Di Wang, Shengfei Lyu, Hong Xu, Xiaoqiao Wang, and 1 others. 2025. Esgenius: Benchmarking llms on environmental, social, and governance (esg) and sustainability knowledge. arXiv preprint arXiv:2506.01646.

Ruben Hetfleisch, Maximilian Nowak, and Fazel Ansari. 2024. Automating eu taxonomy reporting: Can generative ai facilitate corporate sustainability reporting? In Global Conference on Sustainable Manufacturing, pages 54–62. Springer.

Katrin Hummel and Dominik Jobst. 2024. An overview of corporate sustainability reporting legislation in the european union. Accounting in Europe, 21(3):320–355.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. Preprint, arXiv:2310.06825.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. Transactions of the Association for Computational Linguistics, 7:453–466.

Jaeyoung Lee, Geonyeong Son, and Misuk Kim. 2024. Esg-kor: A korean dataset for esg-related information extraction and practical use cases. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 6627–6643.

Aman Madaan, Niket Tandon, Prakhar Gupta, Kathleen Hall, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. In Advances in Neural Information Processing Systems (NeurIPS).

Rafik Mankour, Yassine Chafai, Hamada Saleh, Ghassen Ben Hassine, Thibaud Barreau, and Peter Tankov. 2025. Climate finance bench. arXiv preprint arXiv:2505.22752.

Gaku Morio, Soh Young In, Jungah Yoon, Harri Rowlands, and Christopher Manning. 2024. Reportparse: A unified nlp tool for extracting document structure and semantics of corporate sustainability reporting. In Proceedings Of The Thirty-Third International Joint Conference On Artificial Intelligence (IJCAI) 2024, pages 8749–8753. International Joint Conference on Artificial Intelligence (IJCAI) Organization.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human-generated machine reading comprehension dataset.

Jingwei Ni, Julia Bingler, Chiara Colesanti-Senni, Mathias Kraus, Glen Gostlow, Tobias Schimanski, Dominik Stammbach, 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.

Rajko Odobaša and Katarina Marošević. 2023. Expected contributions of the european corporate sustainability reporting directive (csrd) to the sustainable development of the european union. EU and comparative law issues and challenges series (ECLIC), 7:593–612.

Vik Paruchuri. 2024. Marker.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250.

Tobias Schimanski, Jingwei Ni, Roberto Spacey, Nicola Ranger, and Markus Leippold. 2024a. Climretrieve: A benchmarking dataset for information retrieval from corporate climate disclosures. arXiv preprint arXiv:2406.09818.

Tobias Schimanski, Andrin Reding, Nico Reding, Julia Bingler, Mathias Kraus, and Markus Leippold. 2024b. Bridging the gap in esg measurement: Using nlp to quantify environmental, social, and governance communication. Finance Research Letters, 61:104979.

Spurthi Setty, Harsh Thakkar, Alyssa Lee, Eden Chung, and Natan Vidra. 2024. Improving retrieval for rag based question answering models on financial documents. arXiv preprint arXiv:2404.07221.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. arXiv preprint arXiv:2503.19786.

Andrea Venturelli. 2024. Towards a sustainable future: Competencies, regulations, and paradigm shifts in education. In The Routledge Handbook of Accounting for the Sustainable Development Goals, pages 517–526. Routledge.

Santosh Vutukuri. 2023. xlm-roberta-base-esg-ner.

Shuting Wang, Jiejun Tan, Zhicheng Dou, and Ji-Rong Wen. 2024. Omnieval: An omnidirectional and automatic rag evaluation benchmark in financial domain. arXiv preprint arXiv:2412.13018.

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.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. arXiv preprint arXiv:1809.09600.

Kaiwen Zhao, Bharathan Balaji, and Stephen Lee. 2025. Cf-rag: A dataset and method for carbon footprint qa using retrieval-augmented generation. arXiv preprint arXiv:2508.03489.

Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. 2022. Multihiertt: Numerical reasoning over multi hierarchical tabular and textual data. arXiv preprint arXiv:2206.01347.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. arXiv preprint arXiv:2105.07624.

Yi Zou, Mengying Shi, Zhongjie Chen, Zhu Deng, ZongXiong Lei, Zihan Zeng, Shiming Yang, Hongxiang Tong, Lei Xiao, and Wenwen Zhou. 2025. Esgreveal: An llm-based approach for extracting structured data from esg reports. Journal of Cleaner Production, 489:144572.

A Detailed Related Work

The evolution of question-answering (QA) systems has been fundamentally shaped by large-scale, general-purpose benchmarks that establish both task definitions and evaluation paradigms. Foundational datasets such as SQuAD (Rajpurkar et al., 2016) and Natural Questions (Kwiatkowski et al., 2019) pioneered extractive machine reading comprehension from Wikipedia articles and real search queries, respectively, while HotpotQA (Yang et al., 2018) introduced multi-hop reasoning capabilities with explicit supporting fact supervision. For retrieval-oriented QA, MS MARCO (Nguyen et al., 2016) established large-scale passage ranking and answer synthesis paradigms for open-domain settings. While these resources advanced core methodological approaches, they fail to capture the specialized linguistic patterns of regulatory discourse or the table-intensive evidence structures characteristic of corporate reporting domains.

Finance-Specific Question Answering: Within corporate reporting, the finance domain emerged as the first to emphasize hybrid evidence integration (text+tables) and numerical reasoning capabilities. FinQA (Chen et al., 2021) introduced program-supervised numerical reasoning over financial disclosures, establishing a paradigm for computational approaches to quantitative analysis. TAT-QA (Zhu et al., 2021) extended this framework to mixed tabular-textual contexts derived from authentic financial reports, while MultiHiertt (Zhao et al., 2022) pioneered hierarchical multi-table reasoning combined with narrative text analysis. ConvFinQA (Chen et al., 2022) introduced conversational, multi-turn chains of financial reasoning, and FinTextQA (Chen et al., 2024) shifted toward long-form answers grounded in financial textbooks and regulatory sources. Collectively, these datasets established the necessity for systems capable of retrieving, computing over, and explaining information from lengthy, semi-structured documents—capabilities equally essential for sustainability reporting analysis.

Sustainability and ESG Question Answering

As non-financial disclosure requirements proliferated globally, research efforts adapted QA pipelines and tooling to sustainability and Environmental, Social, and Governance (ESG) reporting contexts. CHATREPORT (Ni et al., 2023) demonstrated LLM-based answering systems with traceable citation mechanisms over sustainability reports, emphasizing the critical importance of evidence provenance in regulatory contexts. ReportParse (Morio et al., 2024) provided unified parsing and segmentation infrastructure to support downstream extraction tasks, while "Glitter or Gold?" (Bronzini et al., 2023) explored ESG information extraction via retrieval-augmented generation (RAG) and graph-based analytical approaches over report corpora. These contributions improved report ingestion capabilities and evidence tracing methodologies but did not provide large-scale, open QA datasets suitable for comprehensive end-to-end evaluation.

Targeted ESG and Climate Benchmarks: Subsequently, more targeted benchmarks emerged to evaluate ESG-specific knowledge extraction and format handling capabilities. ClimRetrieve (Schimanski et al., 2024a) established retrieval benchmarks over climate-related questions grounded in corporate climate disclosures, while Climate Finance Bench (Mankour et al., 2025) targeted QA evaluation over 33 sustainability reports using analyst-style questions within RAG frameworks. GRI-QA (Contalbo et al., 2025) focused specifically on table-based QA using tables aligned with Global Reporting Initiative (GRI) standards. Complementary research explored robust QA from noisy PDF documents in carbon-footprint reporting contexts (CF-RAG/CarbonPDF-QA) (Zhao et al., 2025). Beyond QA-focused benchmarks, ESG-Kor (Lee et al., 2024) provided 118,946 annotated sentences from Korean corporate reports for ESG sentence extraction and classification, demonstrating cross-lingual applicability of sustainability NLP methods. While these resources address retrieval evaluation, table-centric QA, and PDF robustness respectively, they fall short of providing large-scale, report-native QA datasets centered on EU-Taxonomy-relevant content.

EU Taxonomy and Regulatory Automation: Unlike voluntary ESG frameworks, the EU Taxonomy establishes legally binding disclosure requirements under the CSRD. Companies are obligated to report both the eligibility and alignment of their economic activities, supported by narrative justifications, structured KPIs (e.g., turnover, CapEx, OpEx), and references to the applicable technical screening criteria (TSC) and Do-No-Significant-Harm (DNSH) requirements. While previous studies have proposed conceptual frameworks for automating EU Taxonomy compliance—often leveraging knowledge graphs or large language models

(LLMs) (Hetfleisch et al., 2024)—they lack empirically grounded datasets for training or evaluation. ESG-Activities (Birti et al., 2025) provides 1,325 annotated text segments from non-financial disclosures labeled against EU Taxonomy activity descriptions in the transport sector, though this focuses on activity classification rather than comprehensive question answering. To date, no public resource supports question answering (QA) over real-world EU Taxonomy disclosures.

SustainableQA addresses this gap by offering more than 195,000 QA pairs derived from 61 corporate sustainability and EU Taxonomy reports. The dataset includes both fact-based questions (e.g., What percentage of CapEx is aligned?") and explanatory ones (e.g., Why does this activity fail DNSH?"), supporting research in hybrid reasoning, table-to-text grounding, and regulation-aware QA. This enables scalable solutions for compliance automation, verifiable reporting, and transparent monitoring as EU sustainability regulations continue to evolve.

B Automated Quality Assessment and Refinement

To ensure the reliability and utility of the dataset for training and evaluating specialized language models, we introduce a comprehensive quality assessment and refinement framework. This methodology addresses critical challenges inherent in automatically generated QA datasets, including hallucinations, domain misalignment, and factual inconsistencies, which are particularly problematic in regulatory contexts like sustainability reporting that demand high precision. The framework employs 35 specialized LLM prompts for evaluation, modification, and correction (see Appendix D for representative examples; complete set is submitted as supplementary material).

B.1 Hybrid Domain Classification

The initial step in our pipeline is to verify and, if necessary, correct the domain classification of each source text passage. An incorrect domain label can invalidate all subsequent relevance assessments for the QA pairs generated from that text. To address limitations in purely lexical or semantic approaches, we employ a hybrid method that combines the precision of keyword analysis with the contextual understanding of a Large Language Model (LLM).

Our method first assesses the passage against curated keyword sets for each domain. It then uses an LLM to perform a semantic classification. The final classification is determined by a rule-based fusion logic that analyzes the confidence scores from both methods. This logic is designed to be robust: **(1)** If both methods agree with high confidence, the classification is confirmed. **(2)** In cases of disagreement, the method with a significantly higher confidence score prevails. For instance, a high-confidence LLM assessment can override a low-confidence keyword match, which is common in passages with ambiguous terminology. **(3)** If both methods exhibit low confidence, the original classification is preserved, and the item is flagged for manual expert review. This conservative approach prevents the system from making low-confidence automated judgments in the most ambiguous cases, prioritizing accuracy over automation. This hybrid approach produces a single, reliable effective classification for each passage, ensuring that all QA pairs derived from it are evaluated against a consistent and accurate domain context.

B.2 Dual-Metric Evaluation Architecture

Our quality assessment is built upon two orthogonal evaluation dimensions: *faithfulness*, which measures the degree to which an answer is grounded in the source context, and *relevance*, which assesses its alignment with the established sustainability domain. We evaluate both the question and the answer independently for each metric, and compute final scores by averaging their respective components. Both metrics are scored on a 1–10 scale. A QA pair is accepted into the final dataset only if it meets a configurable acceptance threshold (`quality_threshold`, default 6.0), which was determined empirically to balance the need for high-quality data with reasonable dataset retention rates.

B.2.1 Faithfulness Assessment

Faithfulness is a measure of factuality and freedom from hallucination. To accommodate the varied structures of generated answers, we employ type-specific evaluation strategies.

Factoid Questions. In extractive QA, faithfulness is synonymous with verbatim accuracy. We implement a span-level analysis where comma-separated answers are decomposed into individual spans $\{s_1, s_2, ..., s_n\}$. An LLM judge evaluates each span against the source text using a 3-tier scoring system: *Score 10 (Verbatim)* for an exact match;

Score 5 (Partial Match) for spans with minor paraphrasing; and *Score 1 (Non-Verbatim)* for largely paraphrased or hallucinated spans. The overall answer faithfulness is calculated using a proportional weighting formula:

$$F_{\text{answer}} = \text{mean}(f_i | i \in R) \times \frac{|R|}{|T|} \quad (1)$$

where T is the set of all spans, R is the subset of "faithful" spans (scoring ≥ 6), and f_i is the faithfulness score of a span i. This ensures that an answer composed of many high-quality spans scores better than one with fewer, even if their average scores are similar.

Non-factoid Questions. For answers that require synthesis or explanation, we assess conceptual grounding. The evaluation determines whether all claims and inferences in the answer can be reasonably and logically derived from the information present in the source context, without requiring external knowledge.

Table Questions. For answers generated from tabular data, the evaluation prioritizes numerical and categorical fidelity. It verifies that all data points, labels, and relationships mentioned in the answer are an accurate representation of the information in the source table.

The question's faithfulness is also evaluated independently to assess whether it is answerable from the given context. The overall faithfulness score is the average of the question and answer scores:

$$F_{\text{overall}} = \frac{F_{\text{question}} + F_{\text{answer}}}{2} \quad (2)$$

Deterministic Numeric Guardrails. Given the critical importance of quantitative data in sustainability reporting, we implement deterministic pre-LLM checks, or "guardrails," to prevent numerical hallucinations. For example, for an answer stating "Taxonomy-aligned CapEx decreased by 15% in 2023," our guardrail extracts the numeric claim and verify its consistency with the source context. To do so, we apply two validation rules. First, for percentage-based claims, we allow a deviation of up to 0.5 percentage points to account for rounding differences. Second, for general numerical values such as emissions, revenue, or energy use, we require unit compatibility and enforce a maximum relative error of 2%, calculated as the absolute difference between answer and context divided by the context value (with a small constant to avoid division by zero):

$$\text{Relative Error} = \frac{|v_{\text{answer}} - v_{\text{context}}|}{\max(|v_{\text{context}}|, \epsilon)} \leq 0.02 \quad (3)$$

If a numeric claim fails this validation, its faithfulness score is automatically set to 1.0. This *Guardrail Penalty* ensures that any QA pair with factually incorrect numbers is immediately flagged.

B.2.2 Relevance Assessment

Relevance evaluation measures the domain-specific utility and appropriateness of a QA pair for a sustainability professional. This assessment is guided by the effective classification determined in Section B.1. The core principle is to determine if a QA pair provides meaningful, domain-specific insight. A key distinction from faithfulness is that relevance is query-dependent.

Factoid Questions. We employ a context-aware span analysis where each span is scored based on its contribution to answering the specific question within the target domain. For example, in the ESG domain, a span like "Scope 1 emissions" would receive a high relevance score, whereas a generic span like "our corporate headquarters" would score low. The overall answer relevance is calculated using the same proportional weighting formula as faithfulness.

Non-factoid Questions. Relevance is assessed based on their conceptual depth and alignment with the domain's strategic priorities. A high-scoring QA pair might explore the rationale behind a company's climate strategy (Sustainability), the implementation of its governance policies (ESG), or its process for ensuring DNSH compliance (EU Taxonomy).

Table Questions. Relevance is determined by whether the question targets key performance indicators (KPIs) central to sustainability reporting. A question about the percentage of taxonomy-aligned revenue would be highly relevant, while a question about the number of rows in the table would not.

The final overall relevance score is the average of the question and answer relevance scores:

$$R_{\text{overall}} = \frac{R_{\text{question}} + R_{\text{answer}}}{2} \quad (4)$$

B.2.3 Final Decision Rule

The retention decision follows a strict conjunctive rule, requiring that a QA pair demonstrate high quality across both dimensions to be included in the final dataset:

$$\text{RETAIN} \iff F_{\text{overall}} \geq \texttt{quality_threshold} \wedge R_{\text{overall}} \geq \texttt{quality_threshold} \quad (5)$$

B.3 Adaptive Refinement Pipeline

QA pairs that score below the quality threshold on either faithfulness or relevance are not immediately discarded. Instead, they enter a sophisticated refinement pipeline that attempts to repair their quality issues using chain-of-thought (CoT) learning and targeted modification strategies.

B.3.1 Modification and Validation

Modification Target Selection. To maximize efficiency, we identify the primary source of the quality issue using a hierarchical approach. The system first targets the lower-scoring dimension (faithfulness or relevance) and then, within that dimension, identifies the weaker component (question vs. answer) as the primary target for modification.

Chain-of-Thought Refinement. We observe that modification failures often follow repeatable patterns. Therefore, we implement a failure-aware CoT prompting strategy that explicitly learns from previous, unsuccessful refinement attempts. The system maintains a history of failures and constructs increasingly specific prompts for subsequent retries, with a bounded attempt budget (up to K attempts; default K=3).

Targeted Span Refinement. For factoid answers, regenerating the entire answer is inefficient. Instead, we use a targeted approach that fixes only problematic spans. The system identifies spans with low scores (non-verbatim text or weak domain relevance), then repairs each individually by correcting the text to match the source exactly or replacing it with better alternatives from context.

After each repair attempt, the modified content must first pass validation checks (detailed below) before evaluation. Once validated, we evaluate the corrected span—if the span score exceeds the threshold, we recompute the overall answer scores (faithfulness and relevance as defined in §4.2). If both overall scores now meet acceptance criteria, we adopt the modified answer (KEEP_ENHANCED). If not, we continue repair attempts up to K iterations.

When all K attempts fail to improve a span above threshold, we apply selective retention: remove the problematic span and re-evaluate the remaining answer. Due to the proportional weighting formula (Equation 1), even high-quality spans can be penalized by low-scoring ones—for example, an answer with two spans scoring 8.0 and 9.0 plus one scoring 3.0 yields an overall score of 5.7, below the 6.0 threshold; removing the problematic span raises the score to 8.5. If the reduced answer achieves acceptable overall scores, we keep the QA pair with the problematic span removed. If overall scores remain below threshold even after removing bad spans, we discard the entire QA pair (DELETE_LOW_QUALITY).

Validation Protocol. To guarantee quality improvement without introducing new errors, every modified QA pair must pass a series of automated validation checks before re-scoring. A failure at any check results in immediate rejection of the modification, counting as one failed attempt within the K-attempt budget. *Gate 1 (Length Appropriateness)*—A rule-based check ensuring the modification is not unnaturally long or short. *Gate 2 (Readability)*—An LLM-based check assessing linguistic quality, clarity, and grammatical correctness. *Gate 3 (Coherence)*—An LLM-based check validating the logical consistency between the modified question and answer. *Gate 4 (Numeric Consistency)*—A final re-application of the numeric guardrails to ensure no new quantitative errors were introduced.

B.4 Framework Application Results

Implementation Setup. We implemented our framework using Llama-3.3-70B-Instruct as the primary LLM judge with a quality threshold of 6.0 on the 1-10 scale. The initial dataset comprised 205,600 QA pairs before quality assessment.

Initial Quality Assessment. Before refinement, 74.0% (152,144 pairs) met the acceptance threshold, with notable variation across categories. EU Taxonomy pairs demonstrated highest initial quality (81.2%), followed by ESG (75.1%) and Sustainability (70.4%). Table-based questions achieved strongest performance (83.7%), largely reflecting EU Taxonomy tabular data accuracy, while text-based factoid (72.5%) and non-factoid (73.1%) pairs showed moderate quality. Initial faithfulness scores (mean: 7.3) exceeded relevance scores (mean: 6.9) across all categories, with EU Taxonomy achieving highest scores in both dimensions (faithfulness: 7.8, relevance: 7.4).

Refinement Outcomes. The pipeline processed 53,456 QA pairs failing initial thresholds. Of these, 43,143 pairs (80.7%) were successfully enhanced to meet acceptance criteria through targeted refinement, while 10,313 pairs (19.3%) were deleted due to irreparable quality issues. The final dataset con-

tains 195,287 QA pairs with 89.8% overall quality, achieving 93.4% for EU Taxonomy, 90.5% for ESG, and 87.9% for Sustainability. Successfully refined pairs showed average improvements of +1.42 for faithfulness and +1.19 for relevance when targeted, demonstrating the refinement pipeline's effectiveness with an overall retention rate of 95.0%.

C Generation Pipeline Prompts

This section provides the complete prompts used in the dataset generation pipeline.

C.1 Classification Prompt

System: You are an expert assistant specializing in accurately classifying text chunks related to corporate reporting into predefined categories: EU Taxonomy, ESG, Sustainability, or Unknown. Your response must be only the single chosen category label.

INPUT
Title: {title or 'N/A'}
Subtitle: {subtitle or 'N/A'}
Context: {content}

TASK
Based primarily on the Context provided, classify this text chunk into one of the following categories. Use the Title and Subtitle for additional context if needed. Return *only* the single, most appropriate class label (no explanation, reasoning, or punctuation). If specific terms (e.g., EU Taxonomy regulations) are present, prioritize them over broader themes.

CATEGORIES
- *EU Taxonomy:* Specifically mentions EU Taxonomy regulations, principles, alignment criteria, eligibility, screening criteria, DNSH (Do No Significant Harm), Minimum Safeguards, or related KPIs (CapEx, OpEx, Turnover alignment).
- *ESG:* Focuses on Environmental, Social, or Governance factors, metrics, risks, opportunities, reporting frameworks (like GRI, SASB, TCFD, CSRD), materiality assessments, stakeholder engagement, policies, or specific ESG initiatives.
- *Sustainability:* Covers broader sustainability topics like circular economy, climate action goals, biodiversity efforts, resource efficiency, sustainable products/practices, or general corporate responsibility themes.
- *Unknown:* Does not clearly fit into any of the above categories or lacks sufficient information for classification.

C.2 Span Extraction Pass A

System: You are a span extractor that prioritizes recall and outputs results in JSON format. Your goal is to find as many valid spans as possible, adhering strictly to the formatting and content guidelines.

TASK
Extract concise, meaningful spans relevant to a given classification.

INPUT
Context: "{content}"
Classification: "{classification}"

EXTRACTION RULES

Include These:
- Entities, criteria, activities, and technical concepts
- Quantitative data (e.g., '50% reduction', '10 MWh', '€5 million')
- Specific dates/timeframes (e.g., 'by 2030', 'FY2023')
- Named standards/initiatives (e.g., 'GRI Standards', 'TCFD recommendations')
- Specific regulations/policies (e.g., 'CSRD Article 8', 'EU Taxonomy')
- Clearly defined risks or targets (e.g., 'net-zero target')

Exclude These:
- Generic terms like 'company', 'organization', 'report', 'year', 'data', 'information', 'impacts', 'approach', 'process', 'management', 'performance', 'strategy', 'framework'
- Generic adjectives like 'important', 'significant' unless part of a named entity
- Common functional phrases like 'in order to', 'as well as', 'responsible for'

FORMAT REQUIREMENTS
- Extract spans verbatim. Most spans should be 1-5 words
- Critical named entities or specific metrics may be 6-7 words if necessary
- Do not modify, rephrase, or summarize
- Prioritize the most complete and specific verbatim phrase

OUTPUT
{{ "spans": ["<span 1>", "<span 2>", ...] }}

C.3 Span Grouping Pass B

System: You are a highly precise data structuring assistant. Your SOLE task is to critically evaluate candidate spans, filter them based on context and rules, group the valid ones, and output ONLY a valid JSON object in the specified format.

TASK
Organize spans into meaningful, specific thematic groups, minimizing use of the "Individual" category.

INPUT
Context: "{content}"
Classification: "{classification}"
Candidate Spans: "{candidate_lines}"

PROCESS (SEQUENTIAL)
Step 1: Filter & Consolidate
- Keep only spans found verbatim in Context (case-insensitive, original casing)
- Remove trivial/generic standalone spans
- If spans are redundant (substrings, acronym/expansion, exact synonyms), keep only the single most complete version

Step 2: Thematic Grouping (Primary Focus)
- Actively group spans sharing a clear, specific, meaningful common theme
- Form groups even with 2-3 highly related spans
- Create a concise (2-5 words), highly specific 'label' for each group
- Prioritize creating thematic groups

Step 3: "Individual" Category (Use Sparingly)
- Only truly disparate spans that cannot be thematically linked go here
- Double-check before assigning to "Individual"

OUTPUT FORMAT
Return ONLY one valid JSON object:
{{
"groups": [

{{
"label": "<Group Label>",
"spans": ["<Span 1>", "<Span 2>"]
}},
{{
"label": "Individual",
"spans": ["<Span>"]
}}
]
}}

C.4 Factoid Q&A Generation

System: You are an expert Q&A generator. You create factoid questions strictly based on provided text and spans, ensuring the spans are the exact verbatim answers. You output only valid JSON.

TASK
Generate high-quality, factoid Question-Answer pairs based strictly and solely on the provided Context.

INPUT DATA
- Context: "{content}" (The source text)
- Spans (JSON): {json.dumps(spans)} (ONLY valid answers)
- Classification: {classification} (Required tag)

CORE RULES (NON-NEGOTIABLE)
1. *Answer Validity:* The 'answer' field MUST be EXACT, VERBATIM SPAN(S) ONLY from the input Spans. NO additions, omissions, paraphrasing, or changes.
2. *Contextual Accuracy:* The 'answer' MUST be the PRECISE, COMPLETE, CORRECT response to the 'question', based STRICTLY on the provided Context.
3. *Conditional Generation:* Generate ONLY IF you can form a natural, clear question for which the span(s) are the perfect, complete, verbatim answer. Prioritize accuracy over quantity.
4. *Metadata:* Each output object MUST include "type": "Factoid" and "tag": "{classification}".

GENERATION STRATEGY
Span Handling:
- For each group (excluding "Individual"): create one question answered by ALL spans combined (e.g., "SpanA, SpanB"); create separate questions for EACH span individually
- For each "Individual" span: create one question for that single span

Question Formulation: Questions must be grammatical, natural, standalone, and vary in structure (What, When, Which, Percentage, Value, etc.).

EXPLICITLY AVOID
- Trivial questions (answer obvious from span alone)
- Questions embedding the answer text
- "What is [span]?" questions unless Context gives explicit definition
- "Why" or "How" questions

VALIDATION CHECKLIST
Before finalizing, review EVERY pair:
1. Does 'answer' meet Rule 1 (EXACT VERBATIM)? (Y/N)
2. Does 'answer' meet Rule 2 (PRECISE, COMPLETE, CORRECT)? (Y/N)
3. Does 'question' meet Rule 3 and avoid all "Explicitly Avoid" types? (Y/N)

DISCARD any pair failing any check.

OUTPUT
{{"qa_pairs": [...]}}
Each object: "question", "answer", "type": "Factoid", "tag": "{classification}".

C.5 Table Q&A Generation

System: You are an expert AI assistant specialized in analyzing EU Taxonomy financial reports. Your task is to generate comprehensive Question-Answering pairs based only on the provided paragraph text.

INTERNAL THOUGHT PROCESS
(Do Not Show in Output)
1. *Understand Context:* Identify all explicit numerical data (percentages, monetary values), key terms, classifications (Taxonomy-aligned, eligible, non-eligible, transitional, enabling), specific economic activities, and environmental objectives (CCM, CCA, WTR, CE, PPC, BIO).
2. *Identify Implicit Information:* Identify relationships, definitions, reasons, or processes. For example, how different figures relate or what a classification implies.
3. *Categorize Information:* Systematically categorize by EU Taxonomy objective and alignment status (e.g., all turnover data, then CAPEX, then OPEX, then specific activities).
4. *Formulate Diverse Questions:* Ensure a mix of: Factoid questions (specific numbers, percentages, names); Explanatory questions (relationships, purposes, implications, definitions); Holistic questions (synthesizing information from multiple parts).
5. *Focus on Activities & Statuses:* Ask specific questions about which activities are classified under each Taxonomy status (aligned, eligible, non-eligible, transitional, enabling) with their associated figures/percentages.
6. *Self-Correction/Validation:* Cross-reference each Q&A pair against the original paragraph. Verify: The question is non-trivial and specific; The answer is comprehensive, accurate, and directly supported; The Q&A pair is distinct; Avoid questions about table title or caption.

KEY GUIDELINES
1. *Source Constraint:* ALL answers MUST be directly extracted or logically inferred SOLELY from the provided paragraph. Do NOT use outside knowledge.
2. *Comprehensiveness:* Generate questions covering EVERY piece of information (explicit and implicit).
3. *Holistic Questions:* Include questions requiring synthesis of information from multiple parts. This should be the last question, summarizing the core essence.
4. *Specific Activity Focus:* Ask specific questions about Taxonomy-aligned, Taxonomy-eligible, or Taxonomy-non-eligible activities, including 'Transitional' or 'Enabling' designations with their figures/percentages.

OUTPUT
Provide a single JSON object: {"qa_pairs": [...]}
Each object: "question" (string), "answer" (string). No preceding or trailing text.

C.6 Non-Factoid Q&A Generation

System: You are an expert Q&A generator. You create descriptive, non-factoid questions and concise answers based strictly on the provided text. You output only valid JSON.

TASK

Generate 7-15 high-quality, distinct, non-factoid (descriptive/explanatory) Question-Answer pairs STRICTLY from the provided Context.

INPUT
- Context: "{content}"
- Classification: "{classification}"

CORE CONSTRAINTS
1. *Strict Context Adherence:* ALL content MUST be derived SOLELY from information EXPLICITLY STATED or DIRECTLY INFERABLE within the Context. NO external knowledge.
2. *Non-Factoid Focus:* Questions MUST require descriptive/explanatory answers (relationships, purposes, definitions, processes, implications). AVOID simple factoid questions answerable by a single data point.
3. *Distinctness:* Each Q&A pair must offer unique insight. Avoid redundancy.
4. *Answerability:* Every question MUST be fully answerable using ONLY the Context.

QUESTION FORMULATION
Frame questions to elicit explanations/descriptions of:
- Relationships (e.g., "How does X relate to Y?")
- Processes/Mechanisms (e.g., "Explain the mechanism for X")
- Reasons/Justifications (e.g., "What reasons are provided for Z?")
- Implications/Consequences (e.g., "What are the implications of X?")
- Definitions/Clarifications of concepts within the text

For Why/How questions: Ask for explanations of purpose, reasoning, methods, or processes (e.g., "Explain the rationale behind X" or "Describe the method for X"). Use company names from Context for specificity.

ANSWER FORMULATION
Answers must be comprehensive yet focused (1-4 sentences, approximately 25-70 words). Ensure the answer fully addresses the question based ONLY on the text. Avoid trivial/short-phrase answers.

VALIDATION CHECKLIST
Review EACH pair before output:
1. Is the question genuinely non-factoid (descriptive/explanatory)? (Y/N)
2. Is the Q&A pair distinct from others? (Y/N)
3. Is the question fully answerable ONLY from Context? (Y/N)
4. Is the answer comprehensive, accurate, and derived ONLY from Context? (Y/N)

DISCARD any pair failing ANY check.

OUTPUT
{{"qa_pairs": [...]}}
Each object: "question", "answer", "type": "Descriptive, non-factoid", "tag": "{classification}".

C.7 Non-Factoid Q&A Generation

System: You are an expert Q&A generator. You create descriptive, non-factoid questions and concise answers based strictly on the provided text. You output only valid JSON.

TASK
Generate 7-15 high-quality, distinct, non-factoid (descriptive/explanatory) Question-Answer pairs STRICTLY from the provided Context.

INPUTS
- Context: "{content}"
- Classification: "{classification}"

CORE CONSTRAINTS
1. *Strict Context Adherence:* ALL content MUST be derived SOLELY from information EXPLICITLY STATED or DIRECTLY INFERABLE within the Context. NO external knowledge.
2. *Non-Factoid Focus:* Questions MUST require descriptive/explanatory answers (relationships, purposes, definitions, processes, implications). AVOID simple factoid questions answerable by a single data point.
3. *Distinctness:* Each Q&A pair must offer unique insight. Avoid redundancy.
4. *Answerability:* Every question MUST be fully answerable using ONLY the Context.

QUESTION FORMULATION
Frame questions to elicit explanations/descriptions of:
- Relationships (e.g., "How does X relate to Y?")
- Processes/Mechanisms (e.g., "Explain the mechanism for X")
- Reasons/Justifications (e.g., "What reasons are provided for Z?")
- Implications/Consequences (e.g., "What are the implications of X?")
- Definitions/Clarifications of concepts within the text

For Why/How questions: Ask for explanations of purpose, reasoning, methods, or processes (e.g., "Explain the rationale behind X" or "Describe the method for X"). Use company names from Context for specificity.

ANSWER FORMULATION
Answers must be comprehensive yet focused (1-4 sentences, approximately 25-70 words). Ensure the answer fully addresses the question based ONLY on the text. Avoid trivial/short-phrase answers.

VALIDATION CHECKLIST
Review EACH pair before output:
1. Is the question genuinely non-factoid (descriptive/explanatory)? (Y/N)
2. Is the Q&A pair distinct from others? (Y/N)
3. Is the question fully answerable ONLY from Context? (Y/N)
4. Is the answer comprehensive, accurate, and derived ONLY from Context? (Y/N)

DISCARD any pair failing ANY check.

OUTPUT FORMAT
Return ONLY: {{"qa_pairs": [...]}}
Each object must have: "question" (string), "answer" (string), "type": "Descriptive, non-factoid", "tag": "{classification}".

C.8 Table Q&A Generation

Prompt:
You are an expert AI assistant specialized in analyzing financial reports, particularly those adhering to the EU Taxonomy framework. Your task is to generate a comprehensive set of Question-Answering (QA) pairs based only on the provided paragraph text.
Internal Thought Process (Do Not Show in Output):
1. Understand Context: Thoroughly read and understand the entire paragraph, identifying all explicit numerical data (percentages, monetary values), key terms, classifications (Taxonomy-aligned, eligible, non-eligible, transitional, enabling), specific economic activities, and environmental objectives (CCM, CCA, WTR, CE, PPC,

BIO).

2. Identify Implicit Information: Identify any implicit relationships, definitions, reasons, or processes described within the text. For example, how different figures relate or what a certain classification implies in the given context.
3. Categorize Information: Systematically categorize the identified information by relevant EU Taxonomy objective and alignment status (e.g., all turnover data, then all CAPEX, then all OPEX, then specific activities under each).
4. Formulate Diverse Questions: Formulate questions to cover every distinct piece of information identified. Ensure a mix of: Factoid questions (Direct questions about specific numbers, percentages, names); Explanatory questions (Questions requiring a descriptive answer about relationships, purposes, implications, or definitions as presented in the text); Holistic questions (Questions that require synthesizing information from multiple parts of the paragraph to provide a complete answer or draw broader conclusions).
5. Focus on Activities & Statuses: Pay special attention to asking specific questions about which activities are classified under each Taxonomy status (aligned, eligible, non-eligible, transitional, enabling), along with their associated financial figures or percentages. Clearly differentiate between aligned and eligible where relevant.
6. Self-Correction/Validation: Before finalizing each Q&A pair, meticulously cross-reference it against the original paragraph to ensure accuracy and strict adherence to the 'source constraint'. Verify that: The question is non-trivial and specific; The answer is fully comprehensive, accurate, and directly supported by the text; The Q&A pair is distinct and does not overlap significantly with other generated pairs; Avoid generating questions about the table title or caption.

Key Guidelines for QA Generation:

1. Source Constraint: ALL answers MUST be directly extracted or logically inferred SOLELY from the provided paragraph. Do NOT use any outside knowledge or make assumptions.
2. Comprehensiveness: Generate questions that cover EVERY piece of information, both explicit (figures, percentages, categories, names, classifications) and implicit (relationships, purposes, implications, definitions as presented in the text).
3. Holistic Questions: Include questions that require synthesizing information from multiple parts of the paragraph to form a complete answer or draw broader conclusions/summaries from the text. This should be the last question for each paragraph, summarizing the core essence.
4. Specific Activity Focus: Ask specific questions about which activities are classified as Taxonomy-aligned, Taxonomy-eligible, or Taxonomy-non-eligible, including any designated as 'Transitional' or 'Enabling'. Clearly state their associated figures or percentages.

Output Format: Provide the output as a single JSON object with a key "qa_pairs" whose value is a JSON array of objects, where each object has a "question" key and an "answer" key. Ensure the output is only this JSON object, with no preceding or trailing text.

D LLM Evaluation and Refinement Prompts

This section provides representative examples of the LLM prompts used in the quality assessment and refinement framework (Section 4 and Appendix B). The complete set of 35 prompts is submitted as supplementary material.

D.1 Domain Classification

Task: Classify text into EU Taxonomy/ESG/Sustainability/Unknown domains using keyword-guided semantic understanding.

System: You are a sustainability reporting expert. Classify the following text chunk into one of the categories below.

Input:
- Title: {title or 'N/A'}
- Subtitle: {subtitle or 'N/A'}
- Context: {passage_text}

Categories and Key Indicators:
EU Taxonomy:
Key terms: {taxonomy_keywords}
Focus: EU Taxonomy regulations, alignment criteria, DNSH, Minimum Safeguards, CapEx/OpEx/Turnover alignment.
ESG:
Key terms: {esg_keywords}
Focus: Environmental/Social/Governance factors, reporting frameworks (GRI, SASB, TCFD, CSRD), materiality, stakeholder engagement.
Sustainability:
Key terms: {sustainability_keywords}
Focus: Circular economy, climate action, biodiversity, resource efficiency, sustainable practices.
Unknown: Does not fit categories or lacks sufficient information.

Instructions:
1. Analyze context primarily, using title/subtitle for insight
2. Prioritize EU Taxonomy if specific terms present
3. Consider semantic meaning, not just keyword presence
4. Provide confidence score (1-10) and brief reasoning

Output Format:
Classification: [EU Taxonomy|ESG|Sustainability|Unknown]
Confidence: [1-10]
Reasoning: [Brief explanation]

D.2 Faithfulness Evaluation

D.2.1 Factoid Question Faithfulness

Task: Determine if factoid question can be answered with verbatim spans from context.

System: You are an expert evaluator for factoid span extraction.

Span Extraction Rules:
1. Find exact text pieces in context that answer the question
2. Spans must be word-for-word identical (no paraphrasing)
3. Multiple spans can be combined if needed, separated

by commas
4. No interpretation or external knowledge allowed

Evaluation Criteria (1-10):
- 10: Perfect - Clear verbatim spans directly answer the question
- 9: Excellent - Clear verbatim spans with minimal combination needed
- 8: Very Good - Good verbatim spans with minor combination required
- 7: Good - Adequate verbatim spans but require careful extraction
- 6: Acceptable - Some verbatim spans but incomplete coverage
- 5: Weak - Limited verbatim spans with notable gaps
- 4: Below Average - Few relevant verbatim spans
- 3: Poor - Very limited verbatim span availability
- 2: Very Poor - Almost no usable verbatim spans
- 1: Failing - No verbatim spans can answer the question

Context: "{context}"
Question: "{question}"

First identify any verbatim spans that could answer the question, then evaluate answerability.

Format:
REASONING: [Identify specific spans found and evaluate coverage]
SCORE: [Number from 1-10]

D.2.2 Non-Factoid Answer Faithfulness

Task: Assess if non-factoid answer derives ALL information from context with no external knowledge.

System: You are a meticulous evaluator of nonfactoid answer faithfulness. Verify every claim is supported by context.

Requirements:
- All claims must be supported by the context
- No external knowledge or assumptions
- Accurate representation of context information
- Can include reasonable inferences from context

Evaluation Criteria (1-10):
- 10: Perfect - All information directly from context
- 9: Excellent - Context-based with minimal interpretive language
- 8: Very Good - Context-based with minor interpretive language
- 7: Good - Mostly context-based with minimal reasonable inferences
- 6: Acceptable - Mostly context-based with some reasonable inferences
- 5: Fair - Context-based but with notable external additions
- 4: Below Average - Some context-based content with external additions
- 3: Poor - Limited context grounding with significant external content
- 2: Very Poor - Minimal context grounding, mostly external content
- 1: Failing - Significant external content or contradicts context

Context: "{context}"
Question: "{question}"
Answer: "{answer}"

Check each claim in the answer against the context. Identify any external knowledge or unsupported statements.

Format:
REASONING: [Your analysis of context grounding]
SCORE: [Number from 1-10]

D.2.3 Factoid Answer Faithfulness (Span-Level JSON)

Task: Evaluate verbatim accuracy of each factoid answer span using 3-scale scoring.

System: You are evaluating verbatim accuracy of factoid answer spans{domain_text}. Check if each span appears EXACTLY word-for-word in the context.

Context: {context}
Question: {question}

Spans to Evaluate:
- {span_1}
- {span_2}
- ...

Instructions:
1. For each span, check word-by-word if it appears exactly in context
2. Score using 3-scale scoring:
   - 10: PERFECT VERBATIM - all words appear exactly in context
   - 5: PARTIAL VERBATIM - some words verbatim, some not
   - 1: NOT VERBATIM - most/all words not found exactly in context
3. For spans scoring < 10, identify specific verbatim problems

Return strict JSON:
{
"spans": [
{
"span": "<exact span text>",
"verbatim_score": <10, 5, or 1>,
"verbatim_problem": "<issue if score < 10>",
"non_verbatim_words": ["<word1>", "<word2>"],
"found_in_context": true|false
}
],
"reasoning": "<brief verbatim assessment summary>"
}

Focus on exact word matching - number formats, terminology, punctuation must match exactly.

D.3 Relevance Evaluation

D.3.1 Factoid Answer Relevance (Span-Level JSON)

Task: Assess domain relevance of each factoid answer span with evidence.

System: You are a careful, honest expert. Return STRICT JSON only (no Markdown).

User Prompt: You are an expert evaluator for factoid answer relevance in the {classification} domain. Your task is to assess how relevant each span is for {classification} analysis, using the full CONTEXT. Do NOT calculate any overall/final score. Only score each span and provide supporting evidence.

Context: {context}
Question: {question}

Spans to Evaluate:
- {span_1}
- {span_2}
- ...

Domain Anchors ({classification}): {relevant_keywords}

Instructions:
1. For each span, evaluate its relevance to {classification} analysis based on domain value
2. Score each span 1-10:
    - 8-10: Essential {classification} KPI/metric with clear domain connection
    - 6-7: Important {classification} data with domain relevance
    - 4-5: General business data with weak {classification} connection
    - 1-3: Administrative/operational data with no {classification} relevance
3. Identify domain anchors (keywords) that support the relevance assessment

Return strict JSON:
{
"spans": [
{
"span": "<span text>",
"evidence": "<context evidence supporting {classification} relevance>",
"anchors": ["relevant_keyword1", "relevant_keyword2"],
"score": <1-10>
}
],
"reasoning": "<2-3 sentence justification based on {classification} domain value>"
}

Do NOT calculate or return any overall/final score. Score and explain each span individually.

D.3.2 Non-Factoid Answer Relevance

Task: Evaluate answer's domain relevance prioritizing contextual accuracy and practical utility.

System: You are an expert in {classification} strategy evaluation.

Classification Context:
- ESG: Environmental, Social, and Governance factors in business operations and reporting
- EU Taxonomy: European Union's classification system for sustainable economic activities
- Sustainability: Broader sustainable development, environmental stewardship, and corporate responsibility

Important: Non-factoid answers often explain business processes, strategies, or organizational aspects that support {classification} goals. These can be highly relevant even without deep theoretical analysis.

Relevance Criteria (Prioritizing Practical Utility):
- 10: Directly answers question with clear {classification} context relevance and accuracy
- 9: Well-grounded answer with strong {classification} connection and practical utility
- 8: Good contextual answer with clear {classification} relevance for domain professionals
- 7: Adequate contextual answer with noticeable {classification} connection
- 6: Some contextual relevance with {classification} domain applicability
- 5: Limited {classification} connection but contextually accurate
- 4: Weak {classification} relevance, mostly generic but accurate content
- 3: Minimal {classification} connection
- 2: Very limited {classification} relevance
- 1: Not relevant to {classification} domain or contextually inaccurate

Evaluation Principle: Contextually accurate answers that explain business processes, strategies, or organizational aspects supporting {classification} goals should score well (7-9) even without deep theoretical analysis, as practical implementation is crucial for {classification} success.

Answer: "{answer}"

Format:
REASONING: [Explain practical {classification} domain utility and contextual accuracy]
SCORE: [Number from 1-10]

D.4 Question Modification

D.4.1 Factoid Question Faithfulness Modification

Task: Improve factoid question to better target verbatim spans from context while preserving answer consistency.

System: You are an expert in {classification} reporting and factoid question improvement.

Question Analysis: The question "{question}" may have issues with verbatim answerability. Analyze what makes it difficult to answer with exact spans from context.

Modification Approach:
1. IDENTIFY PROBLEMS: Determine why current question is hard to answer with exact spans
2. TARGET SPECIFIC DATA: Focus on data points (numbers, names, dates) actually present in context
3. ALIGN LANGUAGE: Use terminology that matches context exactly
4. ENSURE EXTRACTABILITY: Make sure answers can be word-for-word from context

Answer Consistency Constraint: The modified question MUST target the same data as the answer: "{answer}". Your modified question should elicit this exact same factoid answer from the context. Ensure the question targets the specific spans/data points that produce this answer.

Preserve {classification} Relevance:
- Maintain {classification} domain focus where possible
- Keep domain-specific terminology when supported by context
- Ensure question remains valuable for {classification} analysis

Requirements:
- Question must be DIFFERENT from "{question}"
- Must target verbatim extractable information
- Should allow multiple potential spans for comprehensive answers
- MUST elicit this specific answer: "{answer}"

Context: "{context}"
Current Question: "{question}"
Target Answer: "{answer}"

Format:
REASONING: [Analyze current question's problems, explain specific changes and why they improve verbatim

answerability and preserve answer consistency]
MODIFIED_QUESTION: [Substantially different question targeting verbatim extractable data that produces the target answer]

D.4.2 Non-Factoid Question Relevance Modification

Task: Enhance non-factoid question's domain relevance and analytical depth while maintaining context grounding.

System: You are an expert in non-factoid question modification for {classification} domain.

{classification} Key Terms: {relevant_keywords}

Question Analysis: The question "{question}" may lack sufficient {classification} domain focus or analytical depth. Analyze what domain-specific improvements are needed.

Modification Approach:
1. IDENTIFY DOMAIN GAPS: Determine what {classification}-specific improvements are needed
2. INCORPORATE FRAMEWORKS: Use {classification} analytical frameworks and concepts
3. TARGET PROFESSIONAL INSIGHTS: Focus on analysis valuable for {classification} professionals
4. MAINTAIN ANALYTICAL NATURE: Keep explanatory/analytical character

Preserve Faithfulness Quality:
- Ensure modified question can be answered from available context
- Do not create questions requiring extensive external knowledge
- Keep analytical nature while ensuring context provides foundation

Requirements:
- Question must be DIFFERENT from "{question}"
- Must enhance {classification} professional relevance
- Should target domain-specific analytical insights

Context: "{context}"
Current Question: "{question}"

Format:
REASONING: [Analyze current question's domain relevance, explain specific changes and why they enhance {classification} analytical focus]
MODIFIED_QUESTION: [Substantially different question with enhanced {classification} domain relevance and analytical depth]

D.5 Answer Modification

D.5.1 Non-Factoid Answer Relevance Enhancement (CoT)

Task: Enhance non-factoid answer's domain relevance while maintaining strict context faithfulness.

System: You are an expert in corporate sustainability reporting and analytical answer improvement.

CoT Instructions: {cot_instructions}

Context: "{context}"
Question: "{question}"
Original Answer: "{answer}"

{classification} Domain Focus:
- ESG: Focus on environmental impact, social responsibility, governance practices, stakeholder value, risk management
- EU Taxonomy: Focus on sustainable economic activities, taxonomy alignment, technical screening criteria, environmental objectives
- Sustainability: Focus on sustainable development, resource efficiency, long-term value creation, impact measurement

Enhancement Strategy:
1. ANALYTICAL DEPTH: Provide deeper analysis relevant to {classification} professionals and decision-makers
2. DOMAIN PERSPECTIVE: Frame insights from {classification} operational and strategic viewpoints
3. PROFESSIONAL VALUE: Target content that {classification} practitioners would find actionable
4. CONTEXTUAL GROUNDING: Keep all enhancements strictly based on provided context - do not add external information
5. FLEXIBLE APPROACH: Enhance analytical, numerical, or explanatory content as appropriate
6. CONTEXT LIMITATION: All {classification} domain enhancements must be derivable from the given context

Modification Requirements:
- Enhance {classification} domain analytical perspective using only context-supported information
- Strengthen professional value for {classification} stakeholders and practitioners
- Improve strategic relevance to {classification} business objectives and frameworks
- Maintain answer type (analytical/numerical/explanatory) while enhancing domain focus
- All modifications must stay within the boundaries of the provided context
- Answer must be DIFFERENT from original with clear {classification} relevance improvement

Format:
REASONING: [Explain how {classification} domain relevance was enhanced using only context-supported information while preserving faithfulness]
MODIFIED_ANSWER: [Enhanced answer with stronger {classification} domain focus derived strictly from context information]

D.5.2 Table Answer Faithfulness Enhancement (CoT)

Task: Modify table answer to be perfectly faithful to table context while handling both specific and summary questions.

System: You are an expert in corporate sustainability reporting and tabular data analysis.

CoT Instructions: {cot_instructions}

Table Context: "{context}"
{table_caption_info}
Question: "{question}"
Original Answer: "{answer}"

Important: The TABLE CONTEXT has been converted to paragraph format. Original table structure (rows, columns, cell positions) is not preserved in this format.

Table Faithfulness Requirements:
1. DATA ACCURACY: All numerical values, percentages, names, and categories must match the TABLE CONTEXT exactly

2. CALCULATION PRECISION: Any mathematical operations, totals, or derived values must be mathematically correct
3. NO INVENTED DATA: Do not create, assume, or extrapolate data points not present in the TABLE CONTEXT
4. SUMMARY ACCURACY: For summary questions, ensure aggregations and generalizations are based on actual TABLE CONTEXT
5. REFERENCE PRECISION: Accurately describe data relationships and patterns present in the TABLE CONTEXT

Table Question Types Handling:
- SPECIFIC DATA EXTRACTION: Direct numerical values, names, categories from TABLE CONTEXT
- SUMMARY/ANALYTICAL: Trends, patterns, comparisons, totals derived from TABLE CONTEXT
- COMPARATIVE: Relationships between different table elements
- CATEGORICAL: Groupings and classifications present in TABLE CONTEXT

Modification Process:
1. Cross-verify all numerical values and data points against TABLE CONTEXT
2. Correct any calculation errors or mathematical inaccuracies
3. Remove claims not supported by TABLE CONTEXT
4. For summary questions, ensure generalizations are data-based
5. Ensure answer must be DIFFERENT from original while improving data accuracy

Validation Checklist:
- Do all numbers exactly match the TABLE CONTEXT?
- Are calculations mathematically correct based on available data?
- Are summary statements supported by actual table content?
- Is the data interpretation accurate and complete?

Format:
REASONING: [Explain what data accuracy issues were corrected and how the answer now faithfully represents table information]
MODIFIED_ANSWER: [Accurate answer based precisely on TABLE CONTEXT, appropriate for question type]

D.6 Span-Level Correction

Task: Fix non-verbatim words in factoid span using exact context text.

System: Fix verbatim issues in {classification} spans without redundancy.

Context: {context}
Current Span: "{span_text}"
Verbatim Problem: {verbatim_problem}
Other Existing Spans: {existing_content}
Domain: {classification}

Requirements:
1. Fix non-verbatim words using EXACT word-for-word replacements from context
2. Keep verbatim words unchanged
3. CRITICAL: Modified span must NOT overlap with existing spans: {existing_content}
4. Avoid any redundancy (maximum 40% word overlap with existing spans)
5. Return "REMOVE" if cannot fix without creating redundancy

Examples:
- Fix "2.8M" → "2,800,000" (if found exactly in context)
- Fix "admin costs" → "administrative expenses" (if found exactly)
- Keep verbatim words unchanged

Return only the corrected span or "REMOVE":

E Question-Answer Examples

Table 8: Sample factoid (F) vs. non-factoid (NF) questions (answers are truncated with "..." for space).

Type	Question	Answer
F	What SDGs are mentioned in the context?	SDG 13: Climate action, SDG 16:...
NF	Why does activity 3.10 fail to meet the substantial contribution criterion for the manufacture of hydrogen?	Because the quantified life-cycle GHG emission savings... are not verified, which is necessary to fulfill the criterion.