

ESGRep: Learning Domain-Specific Embeddings for ESG Retrieval-Augmented Question Answering

Vanmanh Nguyen, Vanha Tran(✉), Thiloan Bui, and Truongminh Ngo

FPT University, Hanoi 155514, Vietnam
manhnvhe170552@fpt.edu.vn, {hatv14,loanbt7,minhnt79}@fe.edu.vn

Abstract. Environmental, Social, and Governance (ESG) information plays a critical role in evaluating corporate sustainability and responsible investment practices. However, accurately answering ESG-related questions remains challenging due to the specialized terminology and complex structure of sustainability reports, in particular, when combined with large language models, the answers are prone to hallucinations. In this work, we propose ESGRep, a domain-specific embedding model designed to improve document retrieval in retrieval-augmented generation pipelines for ESG question answering. ESGRep is fine-tuned using contrastive learning with within-report hard negatives, leveraging disclosure content indices from GRI, SASB, and TCFD corporate sustainability report frameworks to construct weakly labeled training triplets without manual annotation. We systematically evaluate ESGRep across six BERT-based architectures and integrate the resulting embeddings into a retrieval-augmented generation (RAG)-based question-answering system, evaluating performance on the ESGenius-QA benchmark. Experimental results show that ESGRep significantly improves downstream QA accuracy across 14 LLMs, increasing accuracy from 71.3% to 77.5% over strong generic baseline embedding models. These results highlight the importance of domain-adapted embeddings for enhancing LLM performance in specialized knowledge domains such as ESG.

Keywords: ESG · Question answering · Large Language Models · RAG

1 Introduction

Environmental, Social, and Governance (ESG) factors have rapidly evolved from niche concerns into central pillars of corporate strategy, regulatory compliance, and investment decision-making. Initiated by the United Nations Global Compact in 2004, the ESG framework provides a structured approach for evaluating a company's sustainability practices, ethical conduct, and long-term value creation beyond traditional financial metrics [6,21]. Environmental factors include a company's carbon footprint, energy consumption, water management, and effects

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16768, pp. 416–431, 2027.
https://doi.org/10.1007/978-3-032-30488-9_26

on biodiversity. Social dimensions address labor practices, employee wellbeing, community engagement, and supply chain accountability. Governance aspects cover board composition, executive accountability, anti-corruption measures, and transparency in disclosure. Together, these three pillars offer stakeholders, including investors, regulators, employees, and civil society, a holistic lens for assessing whether a company operates responsibly and sustainably.

The growing importance of ESG has driven a dramatic expansion in the volume and complexity of sustainability documentation [11]. Authoritative bodies such as the Global Reporting Initiative (GRI)[1], the International Financial Reporting Standards Foundation (IFRS/ISSB) [2], the Task Force on Climate-related Financial Disclosures (TCFD)[3], the Sustainability Accounting Standards Board (SASB)[4], the Carbon Disclosure Project (CDP)[5], and the Intergovernmental Panel on Climate Change (IPCC)[6] collectively produce thousands of pages of technical standards, scientific assessments, and disclosure frameworks each year [36]. Companies across the world are now expected to align their sustainability reporting with multiple overlapping frameworks simultaneously. As a result, the total body of relevant ESG documentation has become vast, multi-layered, and highly specialized, posing a significant challenge for anyone seeking to efficiently locate, interpret, or verify specific ESG claims [17].

In current practice, ESG analysis in settings such as investment research, corporate reporting assurance, and third-party ESG rating typically still relies on expert analysts manually reviewing long sustainability reports, cross-referencing multiple reporting frameworks, and tracing reported claims back to source pages [1,2]. This workflow is time-consuming, expensive, and susceptible to inter-analyst inconsistency, because different evaluators may prioritize different disclosure standards or interpret the same evidence differently [1,2]. The need for scalable, accurate, and verifiable automated tools to assist ESG analysis has never been greater [7].

The promise of large language models (LLMs) to automate ESG analysis has garnered much research attention [5]. LLMs have developed as formidable tools for text understanding and question responding across numerous areas. However, when faced with the technical complexity and multidisciplinary scope of ESG content, assessments on specific benchmarks show that even frontier models display significant knowledge gaps.

Retrieval-Augmented Generation (RAG) addresses limitations of standalone language models by decoupling knowledge storage from generation [14]. Instead of relying solely on information encoded in model weights, RAG retrieves relevant passages from an external corpus at inference time and provides them as context for the LLM, grounding responses in verifiable sources and reducing

[1] https://www.globalreporting.org/.
[2] https://www.ifrs.org/groups/international-sustainability-standards-board/.
[3] https://www.fsb-tcfd.org/.
[4] https://www.cdsb.net/sustainability-accounting-standards-board-sasb.
[5] https://www.cdp.net/en.
[6] https://www.ipcc.ch/.

hallucinations. However, its effectiveness depends heavily on the quality of the retrieval component. Queries and passages are mapped into a shared embedding space and ranked by similarity; if the embedding model fails to capture ESG-specific semantics, the system may retrieve irrelevant or incomplete evidence [27].

These challenges motivate a domain-specific approach to ESG retrieval. In this work, we propose ESGRep, a family of embedding models fine-tuned for dense ESG retrieval. In contrast to prior ESG studies that mainly emphasize classification, information extraction, summarization, or general LLM benchmarking [9,12,26], our work targets the retrieval layer in RAG systems. Unlike prior domain-adapted language models such as ClimateBERT, FinBERT, and E-BERT, which mainly focus on pre-training or downstream classification settings [10,32,38], ESGRep learns retrieval-oriented representations from weak supervision derived directly from disclosure content indices in sustainability reports. We construct a weakly labeled dataset from reports following GRI, SASB, and TCFD standards, treat each disclosure requirement as a query, refine positive passages using GPT-5.2 as a relevance judge, and evaluate six BERT-based backbones on ESGenius-QA. Results show consistent improvements over BM25 and general-purpose embeddings, highlighting the importance of domain-adapted representations for ESG retrieval. Our main contributions are as follows:

(1) ESGRep, a domain-specific embedding model trained with contrastive learning and within-report hard negatives on a weakly labeled corpus of sustainability reports compliant with GRI, SASB, and TCFD standards is proposed, that enables more accurate retrieval for ESG question-answering.
(2) Conducting a systematic comparison of six BERT-variant backbone architectures as ESG embedding models, providing a comprehensive analysis of how model size, architecture, and domain pre-training affect ESG retrieval quality.
(3) We integrate ESGRep into a RAG-based QA pipeline and demonstrate significant improvements in downstream QA accuracy on the benchmark over baselines of 14 models, with analysis of performance across source families and model sizes.

The remainder of this paper is organized as follows. Section 2 reviews related work on ESG NLP, dense retrieval, and domain-specific embedding fine-tuning. Section 3 details the ESGRep training data construction, architecture, and RAG pipeline. Section 4 reports and analyzes experimental results. Section 5 concludes the paper and outlines future directions.

2 Related Work

The field of ESG data extraction and analysis has come a long way, especially with the introduction of advanced LLMs and NLP approaches. This section examines previous research that is relevant to our study, including domain-specific embedding, Retrieval-Augmented Generation, LLM approaches to financial and sustainability contexts, and related ESG resources and benchmarks.

2.1 LLMs in ESG and Sustainability

Due to the growing number and complexity of sustainability disclosures, the use of LLMs in ESG research has becoming increasingly popular [5,19]. Researchers have explored NLP methods for extracting and analyzing ESG information from corporate documents, assessing disclosures, and ensuring compliance with ESG policies [12,26]. LLMs have demonstrated strong performance in many NLP tasks and show promise for automating ESG analysis. However, because of domain-specific vocabulary and intricate reporting systems, implementing general-purpose LLMs to specialist fields like ESG continues to be difficult. Evaluations on dedicated benchmarks reveal that even state-of-the-art models exhibit notable knowledge gaps when handling the technical and interdisciplinary nature of ESG content [9,30].

2.2 RAG and Domain-Specific Embeddings

RAG systems address limitations of LLMs in specialized domains by grounding responses in external knowledge sources [16,31]. RAG lowers hallucination and allows models to use vast texts without memory by recovering pertinent passages at inference time [14]. Its effectiveness for ESG question answering has been demonstrated in prior work [39]. However, RAG performance is fundamentally dependent on the quality of its retrieval component, which selects relevant information from large corpora [4]. This process relies on embedding models that map queries and passages into a shared vector space for similarity-based ranking [35]. Generic embeddings often struggle with ESG-specific terminology and nuanced semantics [29], motivating the development of domain-specific embeddings. Previous research demonstrates that fine-tuned and domain-adapted models significantly improve retrieval accuracy and downstream performance in specialized tasks [8,22,28]. Models such as FinBERT, ClimateBERT, and SEC-BERT further demonstrate the advantages of domain-specific pre-training in ESG-related applications [10,15,32]. Relative to these studies, ESGRep differs in three ways: it optimizes bi-encoder retrieval rather than only document classification or representation learning, it derives weak supervision from report-specific disclosure indices rather than manually labeled QA pairs, and it evaluates retrieval quality through downstream ESG QA performance across multiple LLM families.

2.3 ESG Resources and Benchmarks

In order to support research and assessment in this area, dedicated resources and standards have been developed in response to the growing significance of ESG. The ESGenius benchmark [9] provides a comprehensive framework for evaluating LLMs on ESG knowledge, featuring a collection of 1,136 expert-validated questions (ESGenius-QA) and a corpus of 231 authoritative documents (ESGenius-Corpus). OntoSustain is an ontology developed for mapping quantitative environmental indicators from the GRI and ESRS standards to enable structured ESG data representation [40]. SusGen-3K is a high-quality, category-balanced

dataset comprising approximately 30,000 sentences across seven financial and ESG NLP tasks [34]. ESG-Kor is a dataset designed for information extraction tasks specifically within Korean-language ESG documents [13].

Furthermore, ClimateBERT is a domain-specific language model pre-trained on a specialized corpus of 1.6 million climate-related paragraphs to enhance environmental discourse analysis [32]. E-BERT consists of a set of models optimized for ESG activity detection and reporting in financial texts [38]. ChatReport is an LLM-based tool that provides automated summarization and semantic search capabilities for TCFD-aligned sustainability recommendations [18]. Bronzini et al. utilizes GPT-based pipelines to extract semantic triples and populate knowledge graphs from unstructured ESG disclosures [3]. TCFD-Bench is a benchmark comprising 14 datasets across 8 tasks designed for the rigorous evaluation of sustainability report generation [34]. MMESGBench is a multimodal benchmark that targets complex reasoning and understanding across diverse ESG reporting formats [37]. These resources collectively highlight the ongoing efforts to develop robust tools and datasets for ESG analysis, underscoring the need for domain-adapted approaches to navigate the complexities of sustainability information.

3 Methodology

This section presents the ESGRep framework for learning domain-specific embeddings for ESG retrieval-augmented question answering. We first describe the weak supervision pipeline for constructing training triplets from corporate reports, followed by the encoder architecture and contrastive training procedure, the RAG pipeline that integrates the fine-tuned embeddings with diverse LLMs, and finally the evaluation benchmark.

3.1 Data Construction

Constructing a high-quality training dataset for ESG retrieval is non-trivial. Unlike open-domain QA, there is no large-scale, publicly available collection of ESG-specific (query, relevant passage) pairs. Manual annotation at the passage level is impractical given the length and technical depth of sustainability reports – documents that frequently exceed 100 pages of dense, structured content. We address this by designing a three-stage weak supervision pipeline that produces training signal from the structural properties of sustainability reports following GRI, SASB, and TCFD standards without requiring human annotators.

We collected approximately 2,000 corporate sustainability reports spanning 2013–2024 across diverse industries, produced under GRI, SASB, and TCFD reporting frameworks. We process these reports through a three-stage pipeline: query extraction from disclosure content indices, positive and negative chunk construction, and LLM-based label refinement. Stages 1–2 are illustrated in Fig. 1 and Stage 3 in Fig. 2. Figure 1 shows how report-level structure is converted into retrieval training pairs, while Fig. 2 shows the subsequent filtering step that improves label precision before model training.

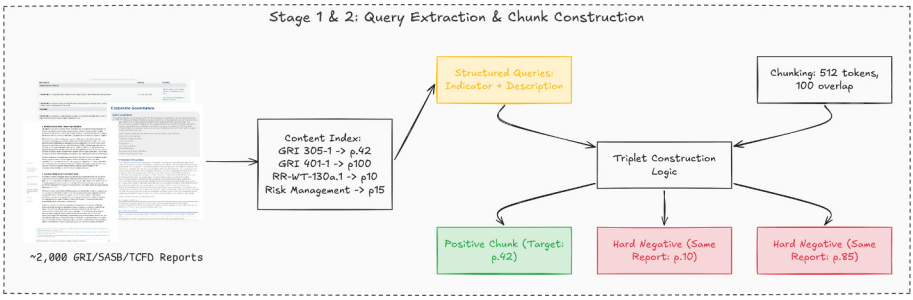

Fig. 1. ESGRep data preprocessing workflow Stages 1 and 2.

Stage 1: Query Extraction. GRI, SASB, and TCFD reports include a disclosure content index – a company-produced table mapping each applicable indicator (e.g., "GRI 305-1: Direct Scope 1 GHG emissions") to the report pages where it is addressed. We treat each indicator description as a natural language retrieval query q, yielding 30–80 queries per report depending on the standard and reporting level.

Stage 2: Positive and Negative Chunk Construction. Each report is segmented into overlapping chunks of 512 tokens with 100-token overlap. Positive chunks are those whose page span overlaps with the pages referenced in the content index for query q. Negative chunks are sampled from the same report but non-overlapping with any referenced page, ensuring negatives share the same company context, writing style, and domain vocabulary as the positive, forcing fine-grained semantic discrimination.

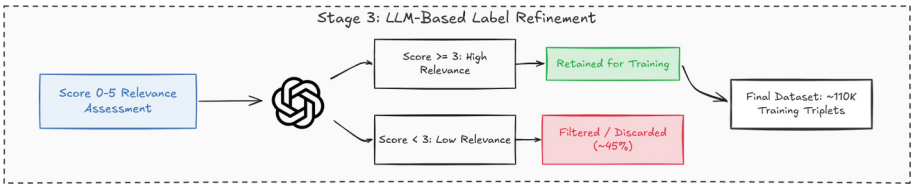

Fig. 2. ESGRep data preprocessing workflow Stage 3.

Stage 3: LLM-Based Label Refinement. Since not every page-referenced chunk directly addresses the disclosure, we apply GPT-5.2 as a relevance judge, scoring each candidate positive (q, c^+) from 0–5 and retaining only those with score ≥ 3. This filters approximately 45% of candidates, yielding 110,000 high-quality triplets from an initial pool of 200,000, split 80/10/10 for train/val/test. The full prompt is provided in Appendix A.1.

3.2 ESGRep Architecture and Training

ESGRep adopts a encoder only architecture in which the query encoder and passage encoder share weights. Each input text is passed through the backbone, and the sentence embedding is computed by mean-pooling the final hidden states across all tokens, followed by L2 normalization.

We fine-tune 6 pre-trained BERT-based encoders as backbone models for ESGRep: RoBERTa-base, DeBERTa-v3-base, BERT-base-multilingual-cased, FinBERT, ModernBERT-base, and DistilBERT-base-multilingual. These models were selected to represent a diverse range of architectural designs (standard transformer, disentangled attention, distilled, and modern efficiency-optimized), parameter counts (109 M to 184 M), and pre-training data profiles (general-purpose English, multilingual, and financial domain).

Training uses MultipleNegativesRankingLoss, which maximizes the cosine similarity between the query and its positive passage while minimizing similarity with all in-batch passages and the sampled hard negatives:

$$\mathcal{L} = -\log \frac{\exp(\text{sim}(q,\, c^{+})/\tau)}{\sum_{j} \exp(\text{sim}(q,\, c_j)/\tau)} \tag{1}$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity and τ is a temperature hyperparameter. We use the AdamW optimizer with a learning rate of 2e-5, cosine learning rate scheduling with 200 warmup steps, weight decay of 0.01, and a batch size of 32. All models are fine-tuned for 3 epochs.

3.3 Evaluation Benchmark: ESGenius-QA

We evaluate on ESGenius-QA, a benchmark of 1,136 expert-validated multiple-choice questions grounded in the ESGenius-Corpus – 231 authoritative ESG documents spanning seven source families (IPCC, CDP, IFRS/ISSB, SASB, GRI, TCFD, UN SDGs). The full benchmark is used as the QA test set and is entirely disjoint from our training data, ensuring that results reflect genuine out-of-distribution generalization from corporate reports to authoritative standards documents.

3.4 Software and Implementation Details

The experimental pipeline was implemented used Transformers [33], and Sentence-Transformers [23,24]. BM25 was used as the sparse retrieval baseline [25]. For Stage 3 label refinement, GPT-5.2 was accessed through the OpenAI API [20]. These software components are reported explicitly to improve reproducibility of the training and evaluation pipeline.

4 Experimental Result

We organize our experimental results as follows. Subsection 4.1 compares the 6 embedding backbones before and after ESGRep fine-tuning. Subsection 4.2

reports per-LLM accuracy to demonstrate that ESGRep's gains generalize across model families. Subsection 4.3 presents ablation studies on negative sampling strategy and the number of retrieved passages. Finally, Subsect. 4.4 analyzes performance across the 7 ESGenius source families to examine in-domain versus out-of-domain transfer.

4.1 Benchmarking Fine-Tuned Embedding Models

Table 1 presents QA accuracy on the ESGenius-QA benchmark for each of the 6 embedding backbones, comparing the pre-trained embedding model without fine-tuning (Baseline) against the ESGRep fine-tuned version. All results use $k = 3$ retrieved passages as context for the LLM, and are averaged across 14 LLMs to isolate the effect of the embedding backbone.

Table 1. QA Accuracy on ESGenius-QA by Embedding Backbone (Avg Across 14 LLMs)

Embedding Backbone	Baseline (no ft)	ESGRep (ft)	Δ^*
RoBERTa-base	0.712	0.774	+0.062
DeBERTa-v3-base	0.716	0.779	+0.063
BERT-base-multilingual-cased	0.708	0.768	+0.060
FinBERT	0.718	0.782	+0.064
ModernBERT-base	0.710	**0.785**	**+0.075**
DistilBERT-base-multilingual	0.703	0.759	+0.056
Average	0.713	0.775	+0.062

*Δ: ESGRep vs. Baseline (no fine-tuning) for same backbone

The results demonstrate a consistent improvement from domain-specific fine-tuning across all 6 backbones: ESGRep provides an average improvement of +0.062 over the corresponding pre-trained baseline, confirming that general-purpose dense embeddings – regardless of architecture – are not well-calibrated for the ESG domain.

Table 1 shows that ESGRep improves every backbone over its untuned baseline, indicating that the gains come from ESG-specific contrastive fine-tuning rather than from one unusually strong pre-trained model. Among the backbone architectures, ModernBERT-base achieves the highest ESGRep accuracy (0.785) with an improvement of 0.075 over its pre-trained counterpart. FinBERT ranks second at 0.782 (+0.064), benefiting from its financial domain pre-training on SEC filings, earnings calls, and analyst reports, which provides a stronger initialization (0.718 baseline) than general-purpose models for ESG-related vocabulary. DeBERTa-v3-base follows at 0.779 (+0.063), while DistilBERT-base-multilingual trails at 0.759 (+0.056), reflecting the capacity limitations of its distilled architecture.

4.2 Per-LLM QA Accuracy Comparison

To demonstrate that ESGRep's improvements generalize across different language models, we report QA accuracy for each of the 14 evaluated LLMs under four retrieval conditions: Zero-shot, BM25, the pre-trained ModernBERT-base baseline (no fine-tuning), and the ModernBERT-base ESGRep model. Table 2 therefore isolates the contribution of retrieval quality while holding the underlying QA benchmark fixed.

Table 2. QA Accuracy by LLM (ModernBERT-base ESGRep vs. Baselines)

LLM	Zero-shot	BM25	Baseline	ESGRep
Qwen3-4B	0.426	0.513	0.544	**0.610**
Qwen3-8B	0.521	0.603	0.631	**0.706**
Qwen3-14B	0.608	0.679	0.704	**0.782**
Qwen3-30B-A3B	0.583	0.659	0.686	**0.762**
Qwen3-235B-A22B	0.693	0.748	0.770	**0.844**
DeepSeek-R1-Distill-Qwen-7B	0.501	0.586	**0.616**	0.613
DeepSeek-R1-Distill-Qwen-14B	0.643	0.708	0.732	**0.808**
DeepSeek-v3.1	0.621	0.690	**0.715**	0.711
DeepSeek-v3.2	0.657	0.718	0.741	**0.816**
gpt-oss-120b	0.700	0.752	0.773	**0.850**
gpt-oss-20b	0.663	0.721	0.743	**0.817**
Gpt-5.2	0.752	0.803	0.822	**0.901**
Claude-4.5	0.724	0.781	0.801	**0.879**
Gemini-3	0.737	0.792	0.811	**0.890**
Average	0.659	0.697	0.721	**0.785**

ESGRep improves QA accuracy for 12 of the 14 evaluated LLMs over the pre-trained dense baseline. The two exceptions are DeepSeek-R1-Distill-Qwen-7B (0.616 to 0.613) and DeepSeek-v3.1 (0.715 to 0.711), both showing slight regressions that may reflect sensitivity to passage ordering or context length in these reasoning-focused architectures.

Among the improving models, the largest absolute gains from Baseline to ESGRep are observed for GPT-5.2 (+0.079), Gemini-3 (+0.079), and Claude-4.5 (+0.078), indicating that even the strongest LLMs benefit substantially from higher-quality retrieved evidence. At the other end of the scale, smaller models benefit most from the full RAG pipeline: Qwen3-4B improves from 0.426 (zero-shot) to 0.610 (ESGRep), an absolute gain of +0.184, demonstrating that domain-specific retrieval can substantially compensate for limited parametric knowledge.

Across all LLMs, the average accuracy increases from 0.659 (zero-shot) to 0.697 (BM25) to 0.721 (Baseline) to 0.785 (ESGRep), a total pipeline gain

of +0.126 over zero-shot. Each stage contributes incrementally, with the fine-tuning step (Baseline→ESGRep, +0.064) providing the largest single improvement. The top-performing LLMs (GPT-5.2, Gemini-3, Claude-4.5) all exceed 0.87 with ESGRep, with GPT-5.2 reaching 0.901.

4.3 Ablation Studies

Figure 3 presents two ablation experiments on the best-performing backbone (ModernBERT-base), averaged across 14 LLMs. Figure 3(a) identifies which negative sampling strategy contributes most to retrieval quality, while Fig. 3(b) shows how retrieval depth affects downstream QA accuracy.

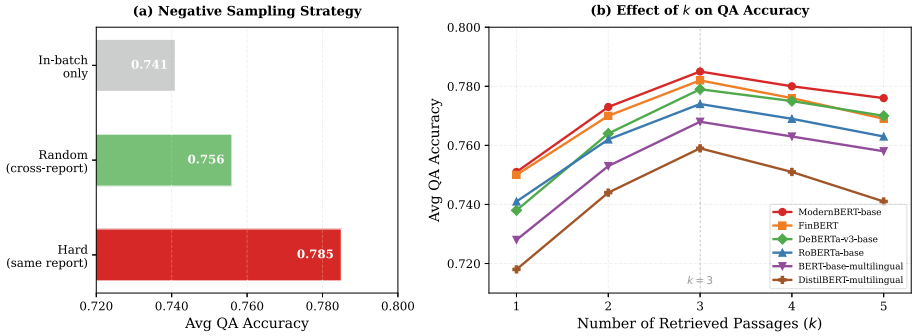

Fig. 3. Ablation studies. (a) Effect of negative sampling strategy on QA accuracy. (b) QA accuracy vs. number of retrieved passages k for each embedding backbone.

Negative Sampling Strategy (Fig. 3a): Within-report hard negatives yield the highest accuracy (0.785), outperforming cross-report random negatives (0.756) and in-batch-only training (0.741). This pattern indicates that retrieval in ESG QA benefits most from difficult negatives drawn from the same report rather than from easy negatives drawn from unrelated documents. Hard negatives force the model to discriminate between passages that share the same company context, writing style, and ESG vocabulary but address different disclosure requirements—precisely the challenge at inference time. Cross-report random negatives are easier to distinguish by company context alone, and in-batch negatives provide the weakest signal.

Number of Retrieved Passages (Fig. 3b): Accuracy peaks at $k = 3$ across all 6 backbones. This confirms that the best downstream QA performance is achieved with a small amount of highly relevant evidence rather than a larger context window. At $k = 1$ and $k = 2$, coverage is limited and the LLM must fall back on parametric knowledge when the ground-truth passage is missed. Beyond $k = 3$, additional passages introduce distracting context that dilutes attention toward the correct evidence—with DistilBERT-multilingual showing

the steepest decline from $k = 3$ to $k = 5$ (-0.018), consistent with the lower retrieval precision of its distilled representations. We adopt $k = 3$ as the default in all main experiments.

4.4 Source Family Analysis

To understand which ESG domains benefit most from domain-specific embedding fine-tuning, we report per-source-family QA accuracy for the best ESGRep model compared to BM25 and the best pre-trained dense baseline. Table 3 distinguishes improvements on source families seen during training from transfer to related but unseen ESG domains.

Table 3. Per-Source-Family QA Accuracy (Best Model, Avg Across 14 LLMs)

Source Family	Zero-shot	BM25	Baseline	ESGRep	Δ^*
IFRS/ISSB	0.651	0.698	0.723	**0.775**	+0.052
SASB	0.673	0.721	0.738	**0.818**	**+0.080**
GRI	0.642	0.678	0.708	**0.796**	+0.088
CDP	0.668	0.704	0.726	**0.778**	+0.052
TCFD	0.680	0.724	0.742	**0.819**	+0.077
IPCC	0.623	0.661	0.694	**0.741**	+0.047
UN SDGs	0.671	0.693	0.718	**0.764**	+0.046

*Δ: ESGRep RAG vs. Baseline RAG

Source family results reveal a clear in-domain advantage. The three source families used to construct training data – GRI (+0.088), SASB (+0.080), and TCFD (+0.077) – show the largest ESGRep gains, confirming that the contrastive fine-tuning procedure effectively learns the retrieval patterns specific to these frameworks' disclosure structures. In contrast, out-of-domain families – IFRS/ISSB (+0.052), CDP (+0.052), IPCC (+0.047), and UN SDGs (+0.046) – show more modest but still positive improvements, indicating that ESGRep transfers partial domain knowledge to related ESG content. Notably, IPCC and UN SDGs show the smallest gains, consistent with their distinct document styles (scientific assessment reports and high-level policy goals) compared to the corporate disclosure format of the training sources. UN SDGs, with the fewest test questions (36 total), should be interpreted with caution due to high variance.

5 Conclusion and Future Work

In this paper, we presented ESGRep, a family of domain-specific bi-encoder embedding models for RAG-based ESG question answering. We introduced a scalable weak supervision pipeline that constructs training triplets from the

disclosure content indices of GRI, SASB, and TCFD corporate sustainability reports without manual annotation, with candidate positives refined by GPT-5.2 relevance scoring. Models are trained with contrastive learning and within-report hard negatives, encouraging fine-grained semantic discrimination on an out-of-distribution test set (ESGenius-QA) drawn from authoritative ESG standards documents.

Experiments across 6 BERT backbones and 14 LLMs show that ESGRep consistently improves QA accuracy over BM25 and generic dense baselines, achieving an absolute gain of +0.064 for the best backbone (ModernBERT-base). Ablations confirm that within-report hard negatives are the primary driver of performance, and per-source-family analysis shows the largest gains on in-domain families – GRI (+0.088), SASB (+0.080), and TCFD (+0.077) – with positive transfer to out-of-domain sources.

The main limitation is a potential bias toward GRI, SASB, and TCFD terminology and the dependence on machine-readable content indices, which limits the training corpus size. Future work will explore multilingual extension to non-English ESG documents, cross-encoder re-ranking to refine retrieved candidates, and expanding the pipeline to additional standards such as ESRS, CDP, and IFRS/ISSB for broader training coverage.

A Appendix

A.1 LLM-as-Judge Prompt for Label Refinement

The following prompt is used in Stage 3 of the training data construction pipeline (Sect. 3.2) to assess whether a candidate positive passage sufficiently addresses a given GRI disclosure requirement. Only pairs receiving a score >= 3 are retained as positive training examples. The score is extracted by parsing the boxed{} tag from the response.

```
You are an expert ESG analyst specializing in sustainability reporting
standards. Your task is to evaluate the relevance of a text passage to
a specific GRI disclosure requirement.

Given the following GRI disclosure requirement and a candidate passage
extracted from a corporate sustainability report, assess how well the
passage addresses the disclosure requirement.

**GRI Disclosure Requirement:**
{query}

**Candidate Passage:**
{passage}

Rate the relevance of the passage to the disclosure requirement on a
scale from 0 to 5:

0 - Completely irrelevant: The passage has no connection to the
    disclosure requirement.
1 - Marginally related: The passage mentions related topics but does
    not address the specific disclosure requirement.
2 - Partially relevant: The passage touches on the disclosure
    requirement but lacks substantive information or only addresses it
     tangentially.
3 - Moderately relevant: The passage provides some meaningful
    information that addresses the disclosure requirement, though it
    may not be comprehensive.
4 - Highly relevant: The passage directly addresses the disclosure
    requirement with substantial and specific information.
5 - Fully addresses: The passage comprehensively and directly addresses
     all aspects of the disclosure requirement with detailed, specific
     data or explanations.

First, provide a brief explanation (2-3 sentences) analyzing how the
passage relates to the disclosure requirement. Then, output your
final score as \boxed{score}.

Example output format:
The passage discusses the company's direct greenhouse gas emissions
and provides specific metrics for Scope 1 emissions, which directly
addresses the GRI 305-1 disclosure requirement. However, it lacks
details on the methodologies used for calculation. \boxed{4}
```

References

1. Berg, F., Kölbel, J.F., Rigobon, R.: Aggregate confusion: the divergence of ESG ratings. Rev. Finan. **26**(6), 1315–1344 (2022)
2. Billio, M., Costola, M., Hristova, I., Latino, C., Pelizzon, L.: Inside the ESG ratings: (dis)agreement and performance. Corp. Soc. Responsib. Environ. Manag. **28**(5), 1426–1445 (2021)
3. Bronzini, M., Nicolini, C., Lepri, B., Passerini, A., Staiano, J.: Glitter or gold? Deriving structured insights from sustainability reports via large language models. EPJ Data Sci. **13**(1) (2024)
4. Cook, T., et al.: Retrieval augmented generation (RAG) for fintech: agentic design and evaluation (2025)
5. Du, K., Mao, R., Xing, F., Mengaldo, G., Cambria, E.: Language models for environmental, social, and governance analysis: a review. Inf. Process. Manag. **63**(4), 104596 (2026)
6. Duan, Y., Yang, F., Xiong, L.: Environmental, social, and governance (ESG) performance and firm value: evidence from Chinese manufacturing firms. Sustainability **15**(17), 12858 (2023)
7. Gharpure, A.: ESG reporting in enterprise financial systems: challenges and innovations. J. Comput. Sci. Technol. Stud. **7**(3), 194–201 (2025)
8. Gupta, A., Rawat, M., Stolcke, A., Pieraccini, R.: Refine on scarce data: retrieval enhancement through fine-tuning via model fusion of embedding models (2024)
9. He, C., et al.: ESGenius: benchmarking LLMs on environmental, social, and governance (ESG) and sustainability knowledge. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pp. 14623–14664. Association for Computational Linguistics (2025)
10. Huang, A.H., Wang, H., Yang, Y.: <scp>finbert</scp>: a large language model for extracting information from financial text. Contemp. Account. Res. **40**(2), 806–841 (2023)
11. Kim, J., Yang, W.: ESG performance in the regulatory transformation era: a systematic thematic review (2020–2024). Front. Sustain. **6** (2026)
12. Lee, H., Kim, J.H., Jung, H.S.: ESG-KIBERT: a new paradigm in ESG evaluation using NLP and industry-specific customization. Decis. Support Syst. **193**, 114440 (2025)
13. Lee, J., Son, G., Kim, M.: ESG-KOR: a Korean dataset for ESG-related information extraction and practical use cases. In: Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 6627–6643. Association for Computational Linguistics (2024)
14. Lewis, P., Perez, E., Piktus, A., Petroni: retrieval-augmented generation for knowledge-intensive NLP tasks (2020). https://doi.org/10.48550/ARXIV.2005.11401
15. Loukas, L., et al.: FiNER: financial numeric entity recognition for XBRL tagging. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics (2022)
16. Luo, H., et al.: Domain-specific retrieval-augmented generation with adaptive embedding and knowledge distillation-based re-ranking. Processes **14**(1), 99 (2025)
17. Mervelskemper, L., Streit, D.: Enhancing market valuation of ESG performance: is integrated reporting keeping its promise? Bus. Strateg. Environ. **26**(4), 536–549 (2016)

18. Ni, J., et al.: CHATREPORT: democratizing sustainability disclosure analysis through LLM-based tools. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 21–51. Association for Computational Linguistics (2023)
19. Ong, K., et al.: Explainable natural language processing for corporate sustainability analysis. Inf. Fusion **115**, 102726 (2025)
20. OpenAI: OpenAI API platform documentation (2025). https://platform.openai.com/docs/overview
21. Ortas, E., Álvarez, I., Garayar, A.: The environmental, social, governance, and financial performance effects on companies that adopt the united nations global compact. Sustainability **7**(2), 1932–1956 (2015)
22. Pacis, F.J., Alyaev, S., Wiktorski, T.: Domain-Adapted Embeddings Model Using Contrastive Learning for Drilling Text Data, pp. 491–503. Springer Nature Singapore (2025)
23. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using Siamese BERT-networks (2019)
24. Reimers, N., Gurevych, I.: Sentence transformers: multilingual sentence, paragraph, and image embeddings using BERT & CO. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 122–131 (2020)
25. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: Bm25 and beyond. Found. Trends Inf. Retr. **3**(4), 333–389 (2009)
26. Schimanski, T., Reding, A., Reding, N., Bingler, J., Kraus, M., Leippold, M.: Bridging the gap in ESG measurement: using NLP to quantify environmental, social, and governance communication. Financ. Res. Lett. **61**, 104979 (2024)
27. Sun, X., Shao, Y., Han, J.: ESG performance drives enterprise high-quality development through financing constraints: based on the background of China's digital transformation. Sustainability **17**(13), 6094 (2025)
28. Szabó, B.N., Gyöngyössy, N.M., Dalos, A.: Applied domain adaptation of LLM-based document embeddings for engineering knowledge retrieval. In: Proceedings of the Intelligent Robotics FAIR 2025, IntRob '25, pp. 29–37. ACM (2025)
29. Tang, Y., Yang, Y.: Do we need domain-specific embedding models? An empirical investigation (2024)
30. Tran, V.K., Tran, K.T., Pham, Q.V., Guhaa, K., Nguyen, H.D.: LLM-powered question answering for environmental, social, and governance (ESG) reports. In: Proceedings of the 2nd ACM Workshop in AI-powered Question & Answering Systems, pp. 49–53. ACM (2025)
31. Wang, X., Chi, J., Tai, Z., Kwok, T.S.T.: FinSage: a multi-aspect rag system for financial filings question answering. In: Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM '25, pp. 6144–6152. ACM (2025)
32. Webersinke, N., Kraus, M., Bingler, J.A., Leippold, M.: ClimateBert: a pretrained language model for climate-related text (2021)
33. Wolf, T., et al.: Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45 (2020)
34. Wu, Q., et al.: SusGen-GPT: a data-centric LLM for financial NLP and sustainability report generation. In: Findings of the Association for Computational Linguistics: NAACL 2025, pp. 1184–1203. Association for Computational Linguistics (2025)

35. Xu, Z., Mo, F., Huang, Z., Zhang, C., Yu, P., Wang, B.: A survey of model architectures in information retrieval (2025)
36. Yadav, S., Samadhiya, A., Kumar, A., Luthra, S., Pandey, K.K.: Environmental, social, and governance (ESG) reporting and missing (M) scores in the industry 5.0 era: broadening firms' and investors' decisions to achieve sustainable development goals. Sustain. Dev. **33**(3), 3455–3477 (2024)
37. Zhang, L., et al.: MMESGBench: pioneering multimodal understanding and complex reasoning benchmark for ESG tasks. In: Proceedings of the 33rd ACM International Conference on Multimedia, MM '25, pp. 12829–12836. ACM (2025)
38. Zhang, M., Shen, Q., Zhao, Z., Wang, S., Huang, G.Q.: Optimizing ESG reporting: innovating with e-BERT models in nature language processing. Expert Syst. Appl. **265**, 125931 (2025)
39. Zhao, Y., et al.: Optimizing LLM based retrieval augmented generation pipelines in the financial domain. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track), pp. 279–294. Association for Computational Linguistics (2024)
40. Zhou, Y., Perzylo, A.: OntoSustain: towards an ontology for corporate sustainability reporting. In: ISWC 2023 Posters and Demos: 22nd International Semantic Web Conference (2023)