

Empirical Evaluation of Open-Source Large Language Models for Retrieval-Augmented Generation in ESG Domain

Motaz Saad

University of Salento

motazk.saad@unisalento.it

ORCID: 0000-0002-1080-7276

Anna Borrelli

University of Naples Federico II

anna.borrelli3@unina.it

Ivan Gentile

IFAB Foundation

ivan.gentile@ifabfoundation.org

Kianna Kazemi

IFAB Foundation

kianna.kazemi@ifabfoundation.org

Francesco Piccialli

University of Naples Federico II

francesco.piccialli@unina.it

Antonella Longo

University of Salento

antonella.longo@unisalento.it

Abstract—Environmental, Social, and Governance (ESG) reporting has emerged as a critical component of corporate accountability, with Large Language Models (LLMs) through Retrieval-Augmented Generation (RAG) showing significant potential for automating KPI extraction from reports. However, the performance characteristics of open-source LLMs in ESG domain-specific tasks remain inadequately understood, creating challenges for informed model selection.

This paper addresses the performance characteristics of open-source LLM deployment in ESG contexts through a structured evaluation framework and an ESG-RAG evaluation resource grounded in 498 real-world ESG reports. We conduct rigorous performance evaluation of seven open-source models ranging from 2B to 30B parameters using a corpus of 498 ESG reports from EU-listed companies (2010–2024), from which 100 synthetic QA pairs are drawn for evaluation from a larger 284-pair persona-based synthetic dataset covering Environmental, Social, and Governance information needs. Our evaluation employs persona-driven query generation and established RAGAS metrics including contextual recall, precision, relevance, faithfulness, answer relevancy, and factual correctness.

Our evaluation of glm-4.7-flash (30B MoE), nemotron-3-nano:4b (4B), qwen3:4b-instruct (4B), gemma3:4b (4B), gemma4:e4b (8B), gemma4:e2b (2B), and ministral-3:8b (8B) reveals notable differences in performance characteristics across different model architectures and sizes. Retrieval is strong but not perfect under LLM-based judging (context recall ≈ 0.58 – 0.61 , context precision ≈ 0.78 – 0.81 , context relevance 0.965 – 0.985 , LLM-judged context precision with reference ≈ 0.83 – 0.88). Generation diverges most on faithfulness (0.607 – 0.822) and least on answer relevancy (0.760 – 0.881): glm-4.7-flash leads faithfulness (0.822), qwen3 leads factual correctness (0.449), and ministral-3 leads answer relevancy (0.881). Factual correctness across all models (0.387 – 0.449) highlights ongoing opportunities for domain-specific fine-tuning. This evaluation provides organizations with data-driven insights for selecting open-source models tailored to their specific ESG reporting requirements.

Index Terms—LLM, open-source models, performance evaluation, RAG, ESG

I. INTRODUCTION

ESG reporting refers to the disclosure of data related to a company’s environmental, social, and governance practices. Originally rooted in ethical investing, it has evolved into a key framework for assessing non-financial risks and performance that can influence a company’s long-term sustainability and investor attractiveness [31]. ESG reporting demonstrates a company’s commitment to sustainability, ethical conduct, and regulatory compliance. Investors, policymakers, and other stakeholders increasingly depend on ESG disclosures to assess corporate performance beyond traditional financial indicators, making the accurate and efficient extraction of ESG Key Performance Indicators (KPIs) a pressing priority. However, ESG data is often complex, heterogeneous, and embedded in diverse, unstructured documents, creating significant challenges in achieving consistency and reliability.

The environmental component of the ESG evaluates how a company affects and is affected by the natural world, focusing on metrics such as carbon emissions (including Scope 1, 2, and 3), energy use, water consumption, waste management, and climate risk strategies. The social dimension centers on the company’s relationships with key stakeholders, including employees, suppliers, and communities. It addresses labor practices, workplace safety, human rights, diversity and inclusion, community involvement, and consumer protection. The governance aspect pertains to the systems and structures guiding corporate behavior, including board composition and diversity, executive compensation, transparency, shareholder rights, and anti-corruption policies. Together, these three pillars form a comprehensive lens through which organizations are evaluated for long-term sustainability and ethical stewardship [29].

Another dimension of complexity is that there are different ESG Reporting Standards and Frameworks. Table I summarizes some of the most common frameworks.

Artificial Intelligence (AI), especially Large Language Mod-

TABLE I
SUMMARY OF MAJOR ESG REPORTING FRAMEWORKS

Framework	Description
GRI [19]	Global Reporting Initiative: Most widely used ESG framework, focused on a company's overall sustainability impact across environmental, social, and economic dimensions.
SASB [33]	Sustainability Accounting Standards Board: Provides industry-specific standards that highlight financially material ESG issues relevant to investors.
TCFD [35]	Task Force on Climate-related Financial Disclosures: Offers guidance on disclosing climate-related financial risks and opportunities, including governance and scenario analysis.
CDP [10]	Carbon Disclosure Project: Specializes in environmental disclosures, particularly greenhouse gas emissions, water security, and forest risks.
CSRD [14]	Corporate Sustainability Reporting Directive: European Union regulation requiring standardized ESG disclosures from large companies with third-party assurance and digital tagging.
ESRS [12]	European Sustainability Reporting Standards: 12 EFRAG-developed standards (ESRS 1–2 cross-cutting, ESRS E1–E5 environmental, ESRS S1–S4 social, ESRS G1 governance) that operationalise CSRD disclosure requirements under a double-materiality lens.

els (LLMs), has emerged as a powerful tool for automating ESG KPI extraction, streamlining data processing, and improving reporting accuracy. Retrieval-Augmented Generation (RAG) techniques further enhance this capability by integrating relevant external data sources into LLM outputs, thereby reducing hallucination and increasing the precision of ESG insights.

Open-source models offer compelling advantages for ESG applications, including deployment flexibility and customization potential. However, the performance characteristics of open-source models in ESG domain-specific tasks remain inadequately understood, creating challenges for informed model selection.

This paper addresses this gap by conducting a comprehensive empirical evaluation of open-source LLMs within RAG frameworks applied to ESG KPI extraction. It provides an assessment framework that evaluates model performance across multiple dimensions including contextual understanding, faithfulness, and factual correctness, enabling organizations to select appropriate open-source solutions tailored to their specific ESG reporting needs.

The paper aims to (1) highlight the growing importance of AI-driven ESG KPI extraction in enhancing sustainability reporting, (2) evaluate the performance characteristics of open-source LLMs in ESG domain-specific RAG applications through systematic benchmarking, (3) analyze the practical implications and deployment considerations of employing open-source models for this purpose, and (4) offer comparative insights to guide organizations toward technically effective open-source solutions for their ESG reporting requirements.

By focusing on both the performance and practical dimensions of open-source AI adoption in ESG analytics, this paper

provides valuable guidance for decision-makers seeking to balance technical capabilities with deployment flexibility and optimize resource allocation in sustainability initiatives. The evaluation framework enables organizations to make informed decisions that account for both the quality of ESG information extraction and the practical advantages of open-source solutions, ensuring sustainable and scalable AI deployment strategies.

This work makes two contributions. First, it provides an empirical evaluation of five open-source LLMs for ESG-focused RAG under a controlled retrieval setup. Second, it introduces an ESG-RAG evaluation resource (available at <https://github.com/motazaad/ESG-RAG-benchmark-ragas>) consisting of a processed corpus of 498 ESG reports, 284 synthetic QA pairs grounded in source documents with persona-based coverage across ESG pillars, and accompanying evaluation outputs and scripts to support reproducible comparison.

II. STATE OF THE ART

Retrieval-Augmented Generation (RAG) is a cutting-edge architecture in NLP that fuses document retrieval with language generation, addressing a key limitation of large language models (LLMs): their inability to access external, up-to-date knowledge at inference time. Originally proposed by Facebook AI Research, RAG systems leverage real-time access to external corpora, enabling models to generate more factually accurate, context-aware answers [23]. Figure 1 shows the key components of a typical RAG system:

- **Retriever:** Selects relevant documents. The retrieval quality directly impacts generation accuracy.
- **Generator:** A pretrained LLM that synthesizes an output based on the retrieved documents plus the query.

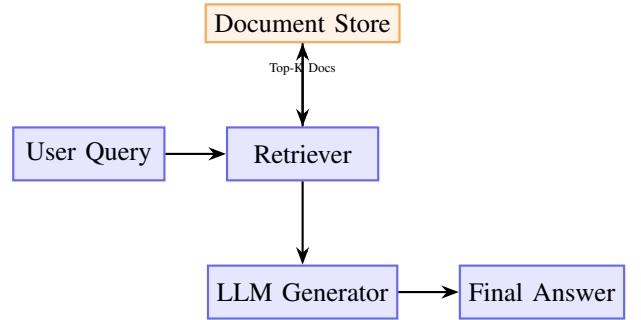


Fig. 1. RAG pipeline

The Mechanism of Retrieval-Augmented Generation (RAG) is as follows:

- 1) **User Query:** Let the input be a natural language query q .
- 2) **Retriever:** A dense retriever encodes the query into a vector and retrieves the top k documents from an external corpus D . This can be formalized as:

$$\{d_i\}_{i=1}^k = \text{Retriever}(q, D)$$

where each d_i is a document retrieved based on semantic similarity.

- 3) **Generator:** For each retrieved document d_i , the generator model produces a response conditioned on the pair (q, d_i) :

$$y_i = \text{Generator}(q, d_i)$$

This architecture enables dynamic knowledge integration without retraining the generator, as updating the external data is sufficient to refresh the model’s knowledge.

RAG has evolved beyond the original retrieve-then-generate paradigm. Gao et al. [17] categorize its development into three generations—Naive, Advanced, and Modular RAG—showing how pre-retrieval query rewriting, post-retrieval re-ranking, and adaptive generation progressively improve factual grounding. These architectural advances make RAG the preferred approach for knowledge-intensive tasks requiring access to specialized corpora such as ESG reports.

A. RAG Evaluation Frameworks

Systematic RAG evaluation has been formalized through dedicated frameworks. Es et al. propose RAGAS [15], a reference-free evaluation library measuring faithfulness, answer relevancy, context precision, and context recall using LLM-as-judge scoring—the framework adopted in this study. Saad-Falcon et al. introduce ARES [32], which fine-tunes lightweight LM judges on synthetic training data for automated pipeline assessment. In addition to research-oriented frameworks such as RAGAS and ARES, practitioner-oriented tooling such as DeepEval supports end-to-end and component-level RAG evaluation with metrics for answer relevancy, faithfulness, contextual relevancy, precision, and recall [11]. More broadly, the use of LLMs as evaluators has been studied in general NLG and LLM benchmarking settings through frameworks such as G-Eval [25], JudgeLM [42], and Prometheus [22], while recent survey work highlights both the practical value of LLM-as-a-judge systems and their reliability challenges [20]. Lima et al. [37] provide a taxonomic framework for synthetic benchmark generation in domain-specific RAG, demonstrating that dataset diversity and generation strategy substantially affect evaluation validity—a finding that directly motivates the persona-driven, quality-filtered protocol used in our evaluation dataset construction.

B. LLMs and RAG for ESG Applications

The application of LLMs to ESG analysis has attracted growing research attention across extraction, generation, and benchmarking tasks. ESGReveal [43] applies RAG with GPT-4 to extract structured ESG data from 166 Hong Kong company reports, demonstrating strong retrieval-augmented extraction performance but evaluating only a single proprietary model. Climate Finance Bench [26] benchmarks RAG pipelines over corporate climate disclosure PDFs with expert-annotated QA pairs, finding that even the strongest proprietary system achieves only 62% accuracy—highlighting task difficulty and leaving open-source models entirely unevaluated.

On the open-source side, SusGen-GPT [39] fine-tunes 7–8B open-source LLMs on a multi-task ESG and financial corpus, showing that small open-source models can approach proprietary performance on sustainability report generation, but without evaluating them within a RAG pipeline. Birti et al. [7] optimize Llama and Gemma models for ESG activity detection and EU taxonomy classification, confirming competitive open-source performance on classification tasks, again without a RAG evaluation setting.

Benchmarks like ESGenius [21] and MMESGBench [41] demonstrate that RAG significantly boosts LLM performance on ESG tasks, especially for smaller models, while multimodal approaches outperform text-only systems on visually grounded disclosures. Fine-tuning datasets such as ESG-CID [1] enable domain-specific retrieval optimization. Practical tools like ESG-Consultant [28] show promise in compliance support, while ESGLens [40] extends ESG-focused RAG toward interactive report analysis with source traceability and score prediction. Studies in low-resource languages—such as Turkish ESG report analysis—reveal that traditional IR methods like BM25 can outperform transformer-based retrievers when paired with proprietary generators [4].

C. Research Gap

Despite these advances, there remains limited work comparing multiple open-source LLMs as generators within a shared ESG RAG pipeline under controlled retrieval, prompting, and evaluation conditions. Existing work either benchmarks ESG knowledge more broadly with RAG as one condition rather than as a controlled cross-model pipeline study [21, 41], introduces ESG QA benchmarks or domain resources without a shared open-source generator comparison over one pipeline configuration [1], or presents ESG-focused RAG systems without systematic cross-family generator evaluation [43, 26, 40]. This gap makes it difficult for organizations to make evidence-based decisions about open-source model selection for ESG applications under practical local-deployment constraints.

D. ESG Reports: Structure and Data Extraction Challenges

ESG reports follow established frameworks such as GRI, SASB, TCFD, and ESRS [10, 19, 14] to ensure comparability and transparency while focusing on material topics most relevant to stakeholders.

Key Structural Elements: ESG reports typically include executive summaries with CEO messages; company profiles and ESG strategy alignment; governance structures and risk management frameworks; materiality assessments identifying priority topics; environmental performance data (emissions, resource use, climate targets); social performance metrics (workforce diversity, human rights, community impact); governance disclosures (board composition, executive compensation tied to ESG); quantitative KPIs with benchmarking; framework alignment tables; and third-party assurance statements.

Data Presentation Formats: ESG reports employ diverse layouts including structured tables for KPI comparisons and framework mappings, infographics and visual narratives

for stakeholder engagement, interactive dashboards in digital formats, materiality matrices plotting stakeholder concerns against business impact, charts and graphs for trend visualization, and geographic maps showing operational impacts.

Data Extraction Challenges: Extracting ESG KPIs from these varied formats presents significant automated processing challenges. Key difficulties include unstructured narrative text with non-standardized phrasing, multi-modal content combining text and visuals that resist traditional parsing methods, inconsistent table formatting across companies and sectors, visual data encoding through charts and color-coding, and variable language expression where identical metrics may be described differently (e.g., "CO2 footprint reduced by 15%" versus "Scope 1 emissions fell by 15%").

A typical extraction challenge occurs when emission values appear within pie charts without clear legends, while accompanying narratives provide only vague statements like "value-chain emissions saw a significant drop," with actual figures relegated to footnotes. This complexity necessitates advanced techniques combining visual parsing, natural language understanding, and contextual inference for effective ESG KPI retrieval.

III. METHODOLOGY

A. Evaluation framework (RAGAS)

We evaluate with RAGAS [15], an open-source library designed specifically for RAG assessment. RAGAS provides metric implementations for: (i) generation quality—faithfulness (does the answer stay grounded in retrieved context?) and answer relevancy (does it answer the user’s question?); and (ii) retrieval quality—context precision (signal-to-noise in retrieved chunks) and context recall (were all relevant chunks retrieved). We use the library’s evaluate() API to compute scores over our evaluation set (details below). RAGAS uses an LLM-judge to score most metrics (reference-free), while context recall relies on ground-truth annotations or synthesized gold answers to estimate true positives and false negatives. The evaluation metrics that we use in our evaluation include:

- **Faithfulness:** penalizes hallucinations by checking whether answer claims are supported by the retrieved passages.
- **Answer relevancy:** measures how directly and completely the answer addresses the query.
- **Context precision:** fraction of retrieved content that is actually useful (high precision leads to less distractor text).
- **Context recall:** fraction of all relevant content that was retrieved (high recall leads to fewer misses).

Together, these metrics expose common failure modes: low precision (over-stuffed context), low recall (missed evidence), low faithfulness (hallucination), and low answer relevancy (off-topic or incomplete answers). The RAG triad as shown in Figure 2 is composed of three RAG evaluation metrics: answer relevancy, faithfulness, and contextual relevancy. Each metric

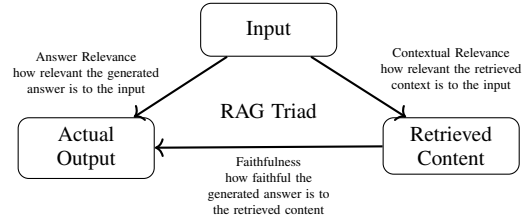


Fig. 2. RAG Triad

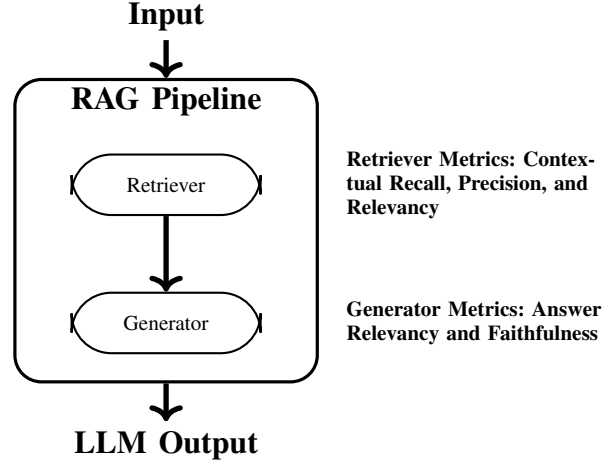


Fig. 3. RAG Evaluation

corresponds to a specific component of the RAG pipeline, making the triad a useful diagnostic tool for identifying which component limits overall performance. Beyond these four core metrics, our evaluation further incorporates LLM Context Precision With Reference, Context Relevance, and Factual Correctness; the complete set of seven metrics is detailed in Section III-E6.

Figure 3 illustrates that RAG evaluation requires measuring both components separately. Metrics must ensure the retriever finds the right information and that the generator produces accurate, relevant answers based on that retrieved context. This aligns with the synthetic data generation process described in Section III-C1, where question-answer pairs are created along with the relevant source chunks to evaluate both retrieval and generation performance.

B. Dataset Selection and Preparation

Our evaluation utilizes a dataset of 498 ESG reports collected from publicly available corporate websites, spanning 15 years (2010-2024) of sustainability reporting from EU-listed companies. This corpus provides coverage across different years and document sizes, enabling robust evaluation of model performance across diverse ESG reporting contexts.

1) *Selection Strategy:* The document selection process followed a systematic approach intended to provide broad coverage across years and document sizes:

- **Temporal Coverage:** Documents were sampled to ensure balanced representation across all 15 years, with 6-7 documents selected per year depending on corpus availability.
- **Size Diversity:** Documents were selected from different size quartiles to ensure variety in document complexity, with selected documents ranging from 11,137 to 1,482,863 bytes.
- **Random Selection:** Within each stratum (year and size category), random selection was used to reduce manual selection bias.

The resulting dataset provides broad coverage of ESG reporting evolution while maintaining manageable computational requirements for multiple model evaluations.

2) *Dataset Characteristics:* The selected 498-document dataset exhibits the following characteristics:

- **Temporal Distribution:** Balanced coverage across 2010-2024, enabling evaluation of model performance on ESG reporting evolution.
- **Size Distribution:** Mean size of 282,119 bytes with median of 230,348 bytes, providing variety in document complexity.
- **Content Diversity:** Covers environmental, social, and governance topics across different industries and reporting frameworks.

3) *PDF-to-Markdown Conversion:* The source corpus consists of PDF files. Prior to indexing, each document was converted to plain-text Markdown using a custom pipeline built on PyMuPDF (`fitz`). The pipeline proceeds as follows:

- 1) **Block extraction:** Each page is processed via PyMuPDF’s `get_text("blocks")` API, which returns text blocks with bounding-box coordinates, preserving the spatial layout of the original PDF.
- 2) **Heading detection:** Blocks are classified as headings using two heuristics: (a) text length under 50 characters with at most 8 words, or (b) all-uppercase text under 100 characters. Heading level is assigned by length and casing—short all-caps blocks become `# H1`, blocks under 50 characters become `## H2`, and remaining heading-classified blocks become `### H3`.
- 3) **Text cleaning:** Body text undergoes whitespace normalization (collapsing multiple spaces), insertion of spaces at lowercase-to-uppercase and digit-to-letter transitions to repair common PDF extraction artefacts, and removal of isolated page-number lines.
- 4) **Page markers:** A separator is appended after each page to preserve page-boundary information for downstream chunking.

This process preserves the logical document structure required for accurate chunk-level retrieval while producing clean plain text suitable for embedding. Tables and figures embedded purely as images are not captured by this text-extraction approach and therefore fall outside the scope of this evaluation.

4) *Dataset Availability:* The complete ESG-RAG evaluation resource is publicly available at <https://github.com/motazsaad/ESG-RAG-benchmark-ragas>. The repository in-

cludes all 498 ESG reports in processed Markdown format, the generated query-answer pairs in CSV and JSON formats, complete RAGAS evaluation scores for all seven models, document selection metadata, and evaluation scripts for reproducing the experimental setup.

C. Synthetic Data Generation for RAG Evaluation

Evaluating a RAG pipeline requires a labeled dataset containing (i) queries that reflect real user information needs and (ii) reference answers grounded in the source documents. Since no such labeled benchmark existed for our ESG corpus described in Section III-B, we constructed one using a custom synthetic generation protocol inspired by prior RAG evaluation workflows [15, 2], illustrated in Figure 4. We used `qwen3:4b-instruct` as the generator; this model also appears in the evaluated set, which introduces a self-enhancement risk that we quantify in Section V.

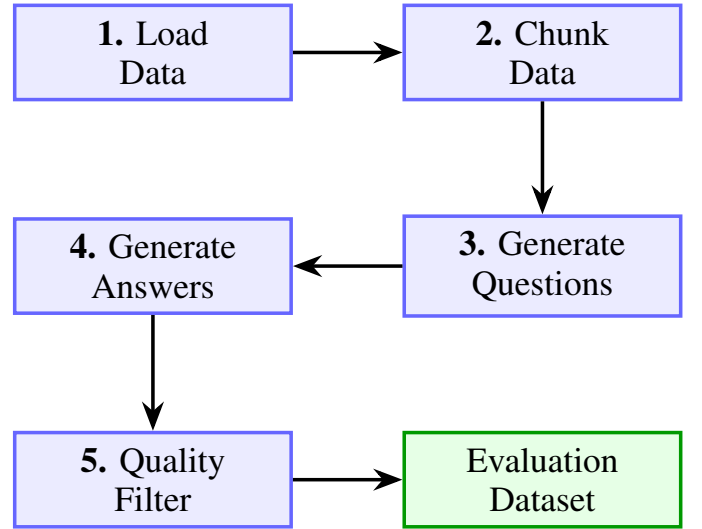


Fig. 4. Five-step synthetic dataset generation protocol for RAG evaluation [15, 2]. Data flows in a snake pattern: documents are loaded and chunked (Steps 1–2); questions and reference answers are derived *directly* from each sampled chunk (Steps 3–4); a length-based quality filter removes degenerate pairs (Step 5).

1) *Validity of the Synthetic Evaluation Protocol:* A legitimate concern with synthetic evaluation datasets is whether machine-generated QA pairs can accurately reflect real-world RAG performance. In our protocol, every question and reference answer is derived solely from the real 498-report ESG corpus, and each retained QA pair records its originating document source. This keeps the benchmark grounded in actual disclosures and preserves document-level provenance for later audit and retrieval analysis.

The protocol also applies simple but explicit controls to reduce degenerate samples. Source chunks shorter than 50 characters are discarded before generation, and generated QA pairs are retained only when both the question and the reference answer exceed 10 characters. Combined with persona rotation across the Environmental, Social, and Governance pillars, these design choices are intended to produce a benchmark

that captures persona-guided ESG information needs while maintaining broad topical coverage.

At the same time, the benchmark should be understood as a custom RAGAS-inspired synthetic resource rather than a direct implementation of the full RAGAS testset-generation pipeline. The dataset was generated with `gwen3:4b-instruct`; one of the evaluated models shares this generator family, and the same model family also serves as the RAGAS judge. We therefore treat the benchmark as a practical comparative instrument within the scope of this study and explicitly acknowledge the resulting self-enhancement risk in Section V.

2) *Generation Protocol*: The pipeline proceeds in five steps:

- 1) **Load Data**: Each ESG document in the 498-report corpus is loaded from the document store.
- 2) **Chunk Data**: Documents are split into fixed-size chunks of 1000 characters with a 200-character overlap (step size 800 characters). Chunks shorter than 50 characters after stripping are discarded. This produces a knowledge graph of 258,590 chunks across the 498 reports. From this pool we randomly sample 300 chunks (seed 42) for QA generation; personas (Section III-C3) are rotated cyclically across the sample so each ESG pillar receives roughly one third of the questions.
- 3) **Generate Questions**: An LLM analyzes each sampled chunk and generates a natural-language question that is (a) answerable solely from the chunk’s content and (b) representative of a genuine ESG information need. Questions are constrained to avoid compound queries that conflate multiple retrieval targets.
- 4) **Generate Answers**: Using the same chunk as the sole context, the LLM generates a concise reference answer. This serves as the ground truth against which the RAG system’s output is compared by RAGAS metrics.
- 5) **Quality Filter**: Degenerate outputs are removed using minimum length thresholds. Source chunks shorter than 50 characters are discarded before generation. Generated QA pairs are retained only if both the question and the reference answer each exceed 10 characters, filtering out empty or truncated outputs.

3) *Persona-Driven Query Specialization*: To ensure domain coverage across all three ESG pillars, we layer the RAGAS persona framework [15] on top of the generation protocol. Each persona directs the question generation LLM toward a specific stakeholder perspective:

```
1 from ragas.testset.persona import Persona
2 persona_environmental_kpi_extractor = Persona(
3     name="Environmental KPI Extractor",
4     role_description="Focuses on extracting key environmental
5     performance
6     indicators (KPIs) from documents, such as carbon
7     emissions, water
8     consumption, energy usage, waste generation, and
9     renewable energy adoption.")
10 persona_social_kpi_extractor = Persona(
11     name="Social KPI Extractor",
12     role_description="Focuses on extracting key social
13     performance indicators
14     (KPIs) from documents, such as employee diversity, labor
15     practices,
```

```
16     community engagement, human rights, and health and safety
17     metrics.")
18 persona_governance_kpi_extractor = Persona(
19     name="Governance KPI Extractor",
20     role_description="Focuses on extracting key governance
21     performance
22     indicators (KPIs) from documents, such as board diversity
23     , executive
24     compensation, anti-corruption policies, shareholder
25     rights, and data
26     privacy measures.")
```

Listing 1. ESG Persona Configuration

The combination of the structured generation protocol, length-based filtering, and persona-driven specialization produces an evaluation dataset that (i) is fully grounded in real ESG source documents, (ii) covers all three ESG pillars, (iii) contains traceable evidence links for retrieval validation, and (iv) captures persona-guided ESG information needs—making it a rigorous benchmark for RAG evaluation in the sustainability reporting domain.

After length filtering, the protocol yields 284 QA pairs (Environmental: 97, Social: 94, Governance: 93) drawn from 181 unique source documents out of the 498 in the corpus.

D. Open-Source Model Selection

Our evaluation focuses on open-source models ranging from 2B to 30B parameters. This band spans the five evaluated checkpoints (plus the Gemma 4 models in the evaluation framework) and remains feasible in our local deployment setting, allowing direct comparison across compact 2B–4B-class models, 8B-class models, and 30B MoE models under the same 48 GB GPU budget.

1) *Selection Criteria and Rationale*: The model selection strategy employed a multi-criteria approach to provide practical coverage of several widely used open-source model families under a common local-deployment setup:

- **Industry Adoption**: Models selected based on widespread adoption in enterprise and research settings, as evidenced by leaderboard rankings and community usage statistics
- **Architectural Diversity**: Models from different families (GLM, Qwen, Gemma, Ministral, Nemotron) to evaluate architectural impacts and training methodology differences
- **Geographic Representation**: Models from major AI research hubs (US, Europe, China) to ensure global perspective
- **Size Range**: 2B–30B parameter range covering edge (gemma4:e2b) to workstation-class MoE models (glm-4.7-flash), while remaining feasible on the available hardware
- **Platform Accessibility**: All models accessible through Ollama platform for consistent evaluation environment and reproducibility

2) *Evaluated Models*: The evaluation included seven open-source models selected based on their performance in industry benchmarks and research literature:

- **glm-4.7-flash:q4_K_M**: 30 billion parameter Mixture-of-Experts model from Zhipu AI (3B active), demonstrating strong performance in Chinese language tasks and ranking among top models in the Open LLM Leaderboard [5]. The q4_K_M quantization format has been reported to reduce memory footprint substantially while preserving most of the original model performance [16].
- **qwen3:4b-instruct**: 4 billion parameter model from Alibaba, representing the latest generation of the Qwen series which has shown exceptional performance in multilingual tasks [30]. Selected for its balanced performance across English and Chinese ESG reporting contexts.
- **gemma3:4b**: 4 billion parameter model from Google, representing the latest generation architecture with significant improvements over previous Gemma models [18]. Chosen for its strong performance in reasoning tasks and Google’s extensive research backing in language model development.
- **gemma4:e4b**: 8 billion parameter model from Google DeepMind (4.5B effective), designed for frontier intelligence locally with configurable thinking modes [18]. Selected to represent the latest generation of Gemma architecture with enhanced reasoning and agentic capabilities.
- **gemma4:e2b**: 2 billion parameter model from Google DeepMind (2.3B effective), optimized for edge device deployment [18]. Selected to evaluate efficiency-focused deployment scenarios.
- **ministral-3:8b**: 8 billion parameter model from Mistral AI, the latest generation dense model with state-of-the-art performance in edge deployment scenarios [24]. Selected as the European representative and for its strong performance in multilingual contexts.
- **nemotron-3-nano:4b**: 4 billion parameter model from NVIDIA, featuring a hybrid Mamba-Transformer MoE architecture optimized for enterprise agentic applications [27]. Selected for its enterprise-grade design and NVIDIA’s extensive ecosystem support.

3) *Exclusion Criteria*: Several popular models were deliberately excluded to preserve a controlled and computationally feasible comparison. The study design intentionally evaluates representative checkpoints per model family within the practical 2B–30B local-deployment range, under a fixed hardware budget and a single shared pipeline configuration:

- **Llama Series**: Excluded because the study evaluates one representative checkpoint per family rather than multiple checkpoints from the same lineage; broader family coverage, including Llama, is planned as future work [38]
- **Larger Models (>30B)**: Models exceeding 30B total parameters excluded to keep the comparison feasible under the available local hardware budget and to avoid runtime disparities dominating the experimental design
- **Specialized Models**: Domain-specific models (e.g., medical, legal) excluded to keep the study focused on general-purpose generators for ESG RAG rather than task-specific

adaptation

- **Older Generations**: Models predating 2023 excluded to keep the comparison centered on recent checkpoints available in the current deployment stack

This selection strategy ensures broad coverage of current open-source LLM capabilities while maintaining experimental feasibility and reproducibility standards essential for academic research [8].

E. Experimental Setup

1) *Hardware Configuration*: All experiments were conducted on a system with the following hardware specifications to ensure reproducibility and provide context for computational requirements:

- **CPU**: 64-core Intel QEMU Virtual CPU (x86_64 architecture)
- **Memory**: 125 GB RAM with 118 GB available memory
- **GPU**: NVIDIA RTX 6000 Ada Generation with 48 GB VRAM
- **GPU Driver**: NVIDIA-SMI 580.126.09 with CUDA 13.0 support
- **Virtualization**: KVM full virtualization environment

The GPU configuration provided substantial computational resources for model inference, with the RTX 6000 Ada Generation offering 48 GB of VRAM sufficient for loading and evaluating multiple models up to 30B parameters simultaneously. End-to-end runtime was approximately 21 hours and 30 minutes for the original five models: about 9 hours 34 minutes for Step 2 (RAG generation across five models on 100 QA pairs over the 498-document corpus) and about 11 hours 56 minutes for Step 3 (RAGAS LLM-judged scoring of seven metrics over the same five model×100 QA matrix). The two additional Gemma 4 models were evaluated in a separate batch under the same pipeline configuration, extending the total runtime accordingly.

2) *Evaluation Subset*: To bound computational cost, RAG evaluation was run on a deterministic 100-item prefix of the 284-pair synthetic testset. The remaining 184 QA pairs were not used in the experiments reported here and are reserved for future sensitivity analyses. Across the seven evaluated models this yields 700 model–query evaluations.

3) *Prompt Format*: All five evaluated models receive identical single-turn prompting. The system message instructs the generator to ground its answer in the retrieved context rather than draw on parametric knowledge: “You are a helpful assistant that answers questions based on given documents only.” The human-turn template follows the Python f-string

```
"question: {query}\n\nDocuments: {relevant_doc}"
```

where `relevant_doc` is the single top-1 retrieved chunk wrapped in a one-element Python list. No few-shot exemplars, chain-of-thought scaffolding, or persona conditioning are applied at inference time. Because the system message, user template, retrieved context, and retriever are held constant, the main experimental variable is the generator model itself. However, decoding parameters were not explicitly normalized

across models in Step 2, so generation behavior may also reflect model-specific default sampling settings exposed through Ollama.

4) *Embedding Standardization*: To ensure consistent evaluation across all models, we standardized the embedding component using qwen3-embedding:4b for both dataset generation and RAG evaluation. This approach eliminates embedding model variability, allowing us to focus evaluation on generation model performance differences.

5) *Retrieval Configuration*: Retrieval is configured at top- $k=1$: for each query the system computes cosine similarity against all 258,590 corpus-chunk embeddings and returns only the single highest-scoring chunk as the entire context window for the generator. There is no reranker, no MMR diversification, no similarity threshold, and no top- $k>1$ fusion. This deliberately minimal-context configuration isolates the generator’s behaviour on a fixed evidence window and eliminates cross-chunk fusion as a confound when comparing model families, at the cost of imposing a retrieval ceiling that we quantify in Section IV-B and acknowledge in Section V. Larger top- k , hybrid retrievers (BM25 + dense), and reranker variants are left as future work.

6) *Evaluation Metrics*: Our evaluation employs the RAGAS framework with seven key metrics:

- **LLM Context Precision With Reference**: LLM-judged variant of context precision. Given the reference answer and the retrieved contexts, the judge decides which retrieved chunks contain information useful for deriving the reference, then averages the chunk-level verdicts into a precision score.
- **Context Precision**: Proportion of retrieved context that is actually relevant.
- **Context Recall**: Ability to retrieve all relevant context from the document corpus.
- **Context Relevance**: Relevance of retrieved context to the query.
- **Faithfulness**: Consistency of generated answers with retrieved context.
- **Answer Relevancy**: Relevance of generated answers to the original query.
- **Factual Correctness**: F1-style accuracy of information provided in generated answers, scored against the reference.

7) *Evaluation Process*: The evaluation process followed these steps:

- 1) **Dataset Generation**: Use persona-driven approach to generate evaluation queries from the 498-document dataset.
- 2) **RAG Pipeline Setup**: Configure each model with standardized embedding and retrieval components.
- 3) **Query Execution**: Process evaluation queries through each RAG system.
- 4) **Metrics Computation**: Calculate RAGAS scores for each model-query combination.
- 5) **Performance Analysis**: Aggregate and analyze results across models and metrics.

8) *Implementation Details*: Table II summarises all hyperparameters and configuration choices used in both the dataset generation and RAG evaluation pipelines, providing the information needed to reproduce the experimental setup.

TABLE II
IMPLEMENTATION HYPERPARAMETERS FOR DATASET GENERATION AND RAG EVALUATION.

Parameter	Value
<i>Dataset Generation</i>	
Generator LLM	qwen3:4b-instruct
Generator context window	12,000 tokens
Embedding model	qwen3-embedding:4b
Chunk size	1,000 characters
Chunk overlap	200 characters
Min. chunk length	50 characters
Min. QA pair length	10 characters (question and answer)
<i>RAG Evaluation (Step 2: retrieval + generation)</i>	
Embedding model	qwen3-embedding:4b
Retrieval method	Cosine similarity
Top-k	1
Generator temperature	Not overridden in Step 2; model-specific Ollama defaults therefore apply unless defined otherwise by the served checkpoint
System prompt	“You are a helpful assistant that answers questions based on given documents only.”
User prompt template	"question: {query}\n\nDocuments: {relevant_doc}" (see Section III-E3)
<i>RAGAS Scoring (Step 3: LLM-as-judge)</i>	
Judge LLM	qwen3:4b-instruct
Judge embedding	qwen3-embedding:4b
Judge temperature	Controlled by the RAGAS LLM wrapper for single-completion scoring (near-deterministic default behavior)
Max workers	1 (serial execution)
Per-call timeout	3600 s

IV. RESULTS: OPEN-SOURCE LLM PERFORMANCE ANALYSIS IN RAG APPLICATION FOR ESG DOMAIN

This section presents the evaluation results of seven open-source large language models on our curated ESG dataset, using the methodology described in the previous section. Our evaluation includes models from different families and parameter sizes: glm-4.7-flash (30B MoE, q4_K_M), nemotron-3-nano:4b (4B), qwen3:4b-instruct (4B), gemma3:4b (4B), gemma4:e4b (8B, 4.5B effective), gemma4:e2b (2B, 2.3B effective), and ministral-3:8b (8B). This diverse model selection enables systematic analysis of performance characteristics across different architectures and sizes in the open-source ecosystem. The evaluation encompasses both retrieval and generation performance across environmental, social, and governance domains. To ensure consistency across all experiments, we use qwen3-embedding:4b for both dataset generation and RAG evaluation processes.

A. Overall Performance Comparison

Table III presents the performance evaluation of seven open-source LLMs across seven RAGAS metrics. The results reveal clear differences in model capabilities across different aspects

of the RAG pipeline, with notable performance differences emerging between model families and sizes.

B. Retrieval Performance Analysis

Retrieval performance is strong but no longer perfect once judged with the RAGAS LLM-based metrics. Context relevance reaches 0.965–0.985 across all models, indicating the top-1 retrieved chunk is almost always topically aligned with the query. LLM-judged context precision with reference is also high (0.830–0.878), confirming that the retrieved chunk usually contains the information needed to derive the reference answer. Context precision and context recall sit at 0.783–0.805 and 0.578–0.605 respectively.

The shared retriever (qwen3-embedding:4b, top- $k=1$) means CtxRel would be identical across models in a deterministic retrieval setup; the small differences in LLMCP(Ref), CP, CR, and CtxRel arise from the LLM judge’s per-sample interpretation of whether the retrieved chunk fully supports the reference. The Gemma 4 models (e4b and e2b) show slightly lower retrieval-side scores compared to the other five models, which may reflect differences in how their responses interact with the LLM judge’s scoring criteria. Recall below unity reflects a realistic ceiling for top- $k=1$ retrieval: when the reference answer spans information that overlaps multiple chunks, a single retrieved chunk cannot cover all of it. These scores are not driven by an embedding-model mismatch between retrieval and judging — both stages use `qwen3-embedding:4b` — but rather by genuine retrieval gaps that the LLM judge surfaces.

C. Generation Performance Analysis

Generator performance reveals substantial differences between models, particularly in faithfulness. Faithfulness shows the widest spread, with `glm-4.7-flash:q4_K_M` leading (0.822), followed by `gemma4:e4b` (0.796), `gemma4:e2b` (0.784), `gemma3:4b` (0.776), `qwen3:4b-instruct` (0.719), `nemotron-3-nano:4b` (0.625), and `ministral-3:8b` (0.607). The 0.215 spread between best and worst on this metric is the dominant generation-quality differentiator. Notably, both Gemma 4 models outperform `gemma3:4b` on faithfulness, with `gemma4:e4b` (0.796) ranking second overall, despite having lower retrieval-side scores. Since all seven models receive the same anchoring system prompt and the same retrieved context (Section III-E3), this spread is consistent with model-level differences in how strictly each generator follows the “answer from documents only” instruction, though some portion of the variation may also reflect unnormalized default decoding settings in Step 2.

Answer relevancy shows a wider spread when Gemma 4 models are included (0.760–0.881). The Gemma 4 models score notably lower on this metric (`gemma4:e4b` 0.762, `gemma4:e2b` 0.760), while the remaining five models cluster tightly (0.818–0.881) with `ministral-3:8b` first (0.881), followed by `gemma3:4b` (0.877), `qwen3:4b-instruct` (0.863), `nemotron-3-nano:4b` (0.851), and `glm-4.7-flash:q4_K_M` (0.818). The lower AR scores for Gemma 4

models suggest their responses may be less directly aligned with the user query, possibly due to differences in instruction-following behavior or decoding defaults.

D. Factual Correctness and Domain-Specific Performance

Factual correctness scores range from 0.387 to 0.449 across evaluated models. `Qwen3:4b-instruct` leads (0.449), followed by `gemma3:4b` (0.427), `gemma4:e2b` (0.423), `ministral-3:8b` (0.405), `gemma4:e4b` (0.405), `nemotron-3-nano:4b` (0.402), and `glm-4.7-flash:q4_K_M` (0.387). The 0.062 spread is much narrower than the faithfulness spread, suggesting all seven models face similar headwinds when extracting specific ESG facts from a single retrieved chunk. Even the top performer leaves substantial headroom for improvement.

1) *ESG Domain Complexity Challenges*: The factual correctness challenges in ESG applications stem from several domain-specific complexities that distinguish sustainability reporting from general knowledge domains [6]:

- **Technical Terminology Evolution**: ESG reporting terminology evolves rapidly with new frameworks (GRI, SASB, TCFD, ESRS) and regulatory requirements, creating challenges for models trained on historical data [13].
- **Quantitative Data Precision**: ESG metrics require high numerical precision (e.g., carbon emissions in tons, percentage reductions, monetary values) where small errors can have significant regulatory implications [9].
- **Contextual Dependencies**: ESG facts often depend on specific regulatory contexts, industry standards, and geographic variations that general-purpose models may not capture accurately [6, 14, 12].
- **Forward-Looking Statements**: ESG reports contain numerous forward-looking commitments and targets that require careful distinction between historical facts and projected goals [34].

2) *Model Family Performance Patterns*: The variation in factual correctness across model families may reflect differences in training emphasis and data composition, though the precise causes cannot be determined from evaluation scores alone:

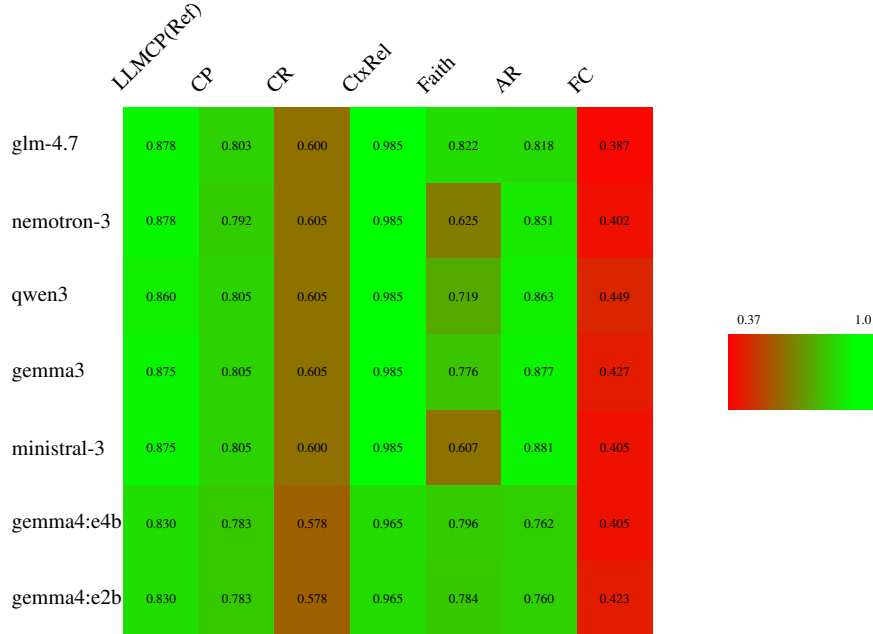
- **Qwen3:4b-instruct (0.449)**: Top factual correctness; given the self-enhancement caveat (Section V), this leadership should be interpreted cautiously, as the dataset generator and RAGAS judge are also instances of `qwen3:4b-instruct` [30].
- **Gemma3:4b (0.427)**: Strong factual performance; the underlying reasons are unclear from this evaluation alone [18].
- **Gemma4:e2b (0.423)**: Strong factual correctness for a 2B model, nearly matching `gemma3:4b` despite having fewer parameters [18].
- **Gemma4:e4b (0.405)**: Mid-range factual correctness; performs similarly to `ministral-3:8b` despite being a comparable effective parameter count [18].
- **Ministral-3:8b (0.405)**: Mid-range factual correctness paired with the highest answer relevancy (0.881) [24].

TABLE III

RAG PERFORMANCE EVALUATION RESULTS ACROSS SEVEN OPEN-SOURCE LLMs FOR ESG DOMAIN. LLMCP(REF) = LLM CONTEXT PRECISION WITH REFERENCE, CP = CONTEXT PRECISION, CR = CONTEXT RECALL, CTXREL = CONTEXT RELEVANCE, FAITH = FAITHFULNESS, AR = ANSWER RELEVANCY, FC = FACTUAL CORRECTNESS (F1). BEST SCORE PER COLUMN IS IN BOLD; TIES HIGHLIGHT ALL TIED MODELS.

Model	Retriever Metrics				Generator Metrics		Answer Quality
	LLMCP(Ref)	CP	CR	CtxRel	Faith	AR	FC (F1)
glm-4.7-flash	0.878	0.803	0.600	0.985	0.822	0.818	0.387
nemotron-3-nano	0.878	0.792	0.605	0.985	0.625	0.851	0.402
qwen3	0.860	0.805	0.605	0.985	0.719	0.863	0.449
gemma3	0.875	0.805	0.605	0.985	0.776	0.877	0.427
gemma4:e4b	0.830	0.783	0.578	0.965	0.796	0.762	0.405
gemma4:e2b	0.830	0.783	0.578	0.965	0.784	0.760	0.423
ministral-3	0.875	0.805	0.600	0.985	0.607	0.881	0.405

Fig. 5. RAG performance heatmap across seven models and metrics. Darker green indicates higher scores; lighter/redder indicates lower scores. Retrieval metrics (LLMCP(Ref), CP, CR, CtxRel) cluster tightly, while generation metrics (Faith, AR, FC) show wider variance.



- **Nemotron-3-nano:4b (0.402):** Mid-range factual correctness; near-tied with ministral-3 and gemma4:e4b [27].
- **GLM-4.7-flash:q4_K_M (0.387):** Lowest factual correctness; the 4-bit quantization is a plausible contributor, though the cause cannot be isolated from this evaluation alone [36, 16].

3) *Practical Implications and Improvement Strategies:* The factual correctness results have important implications for ESG RAG deployment:

- **Human-in-the-Loop Validation:** Current performance levels suggest that critical ESG applications should incorporate human validation mechanisms for high-stakes decisions [3].
- **Domain-Specific Fine-Tuning:** The performance variation across models indicates significant potential for

improvement through ESG-specific fine-tuning using curated sustainability datasets [39, 1].

- **Hybrid Retrieval Architectures:** Larger top- k , rerankers, and modular retrieval pipelines could leverage complementary retrieval signals to improve overall factual accuracy [17].
- **Retrieval Enhancement:** Improved retrieval systems with better ESG-specific indexing could provide more accurate context, reducing factual errors [7].

The narrow 0.387–0.449 band across all seven models, all sitting well below 0.5, indicates that factual correctness remains the weakest link of the pipeline in this evaluation. This pattern persists under a fixed retrieval configuration with relatively strong retrieval-side scores (LLMCP(Ref) up to 0.878) and high answer relevancy (≥ 0.760 for every model),

indicating that factual correctness remains a major bottleneck in the current pipeline configuration. Because retrieval was fixed at top- $k=1$ and no retrieval ablation was performed, these results should not be interpreted as ruling out retrieval improvements as a path to better factual accuracy. ESG applications that demand high precision for regulatory compliance and stakeholder trust would therefore benefit substantially from domain-specific fine-tuning or human-in-the-loop validation, regardless of which generator is chosen from this set.

E. Model-Specific Insights

glm-4.7-flash:q4_K_M leads faithfulness (0.822) and ties for the highest LLM-judged context precision with reference (0.878). Its trade-off is the lowest answer relevancy (0.818) and lowest factual correctness (0.387) of the seven, plausibly influenced by 4-bit quantization. Within this evaluation setting, it appears best suited to faithfulness-critical applications where staying anchored to retrieved evidence outweighs phrasing fluency.

nemotron-3-nano:4b ties for the highest LLMCP(Ref) (0.878) and ranks mid-pack on AR (0.851) and FC (0.402), but posts the lowest faithfulness (0.625) of the set. Within this evaluation setting, it appears to be a solid generalist for retrieval-side use cases, but weaker when strict context grounding is required.

qwen3:4b-instruct leads factual correctness (0.449) and posts competitive faithfulness (0.719) and answer relevancy (0.863). Because the dataset generator and RAGAS judge are also qwen3 family, this leadership is qualified by the self-enhancement caveat (Section V).

gemma3:4b is the most balanced model: second on faithfulness (0.776), second on answer relevancy (0.877), and second on factual correctness (0.427) — no clear weakness on any generation metric. Within this evaluation setting, it is a reasonable default when the deployment requirements are not yet known.

gemma4:e4b (8B, 4.5B effective) ranks second on faithfulness (0.796) but posts lower answer relevancy (0.762) than the other five models, with mid-range factual correctness (0.405). Its strong faithfulness suggests effective instruction-following for context grounding, but the lower AR indicates room for improvement in query alignment. Within this evaluation setting, it is a strong candidate for faithfulness-prioritizing applications.

gemma4:e2b (2B, 2.3B effective) is noteworthy for its efficiency: despite being the smallest model in the set, it achieves competitive faithfulness (0.784) and factual correctness (0.423). It trails only on answer relevancy (0.760), where it is the lowest of the set. Within this evaluation setting, it is a compelling choice for resource-constrained deployments that still require strong context grounding.

ministral-3:8b leads answer relevancy (0.881) but posts the lowest faithfulness (0.607) of the set, with mid-range factual correctness (0.405). Within this evaluation setting, it appears best suited to user-facing applications where on-topic phrasing

matters most and the consequence of mild context drift is acceptable.

Overall, no single model dominates every metric. Model selection should be driven by which generation property the application weighs most heavily, with the self-enhancement caveat in mind when comparing qwen3 against the others.

F. Implications for ESG RAG Applications

Within the reported ESG RAG evaluation setting, glm-4.7-flash:q4_K_M is the strongest choice for faithfulness-critical tasks (0.822), where the system must stay anchored to the retrieved evidence rather than draw on parametric knowledge. For applications prioritizing factual accuracy, qwen3:4b-instruct offers the highest factual correctness (0.449) and competitive answer relevancy (0.863), making it well suited for compliance-focused ESG systems — subject to the self-enhancement caveat in Section V. For applications prioritizing query-response alignment, ministral-3:8b delivers the highest answer relevancy (0.881), making it suitable for user-facing ESG information systems. Gemma3:4b offers the most balanced profile across the three generation metrics in this evaluation and is a reasonable default when deployment requirements are not yet known. Among the Gemma 4 family, gemma4:e4b (8B) is a strong faithfulness-oriented alternative (0.796) with the edge-deployable gemma4:e2b (2B) achieving surprisingly competitive factual correctness (0.423) at minimal computational cost, making it suited for resource-constrained settings.

V. LIMITATIONS

Several limitations should be considered when interpreting the findings of this study.

Corpus scope: The evaluation corpus comprises 498 ESG reports from EU-listed companies, collected from publicly available corporate websites. Findings may not generalize to ESG reporting in other geographic regions, languages, or regulatory frameworks.

Synthetic evaluation set: The 100 QA pairs used for evaluation were generated synthetically by the same LLM family used for dataset construction (qwen3:4b-instruct). Real user queries may differ substantially in phrasing, complexity, and information needs, which could affect the observed performance rankings.

Subset selection: The reported results are based on a deterministic prefix of the 284-pair synthetic testset rather than a randomized or stratified evaluation subset. The observed rankings may therefore be sensitive to subset selection.

Top- $k=1$ retrieval ceiling: All retrieval results are bounded by single-chunk top-1 retrieval. Context recall around 0.60 reflects cases where the reference answer’s supporting evidence spans information beyond a single 1000-character chunk. Retrieval performance under higher top- k , larger chunk sizes, or independent (non-synthetic) query sets is unknown.

No non-RAG baseline: This study evaluates RAG pipelines only; direct LLM inference without retrieval was not included.

The absolute benefit of the retrieval component therefore cannot be quantified from these results alone.

No hyperparameter ablation: Chunk size (1000 characters), overlap (200 characters), and top-k (1) were fixed throughout. The sensitivity of results to these choices was not studied.

LLM-as-judge variability: RAGAS metrics rely on LLM-based scoring. In our setup, scoring is mediated by the RAGAS LLM wrapper rather than a raw direct call to the judge model, which applies near-deterministic single-completion behavior. Even so, judge-model choice and LLM-as-judge effects may still introduce uncertainty not captured by point estimates alone.

Generator decoding control: Step 2 generation used direct `ChatOllama(model=...)` calls without explicitly normalizing temperature or related sampling parameters across models. As a result, some observed differences may reflect both checkpoint differences and model-specific default decoding behavior.

Sampling fraction: Only 300 of 258,590 corpus chunks (0.12%) were drawn for QA generation, covering 181 of the 498 source documents. Results reflect performance on this sample rather than full-corpus coverage.

Self-enhancement risk: The dataset generator, one of the evaluated models, and the RAGAS judge are all instances of `qwen3:4b-instruct`. Although the generator/judge call paths are separate from inference, `qwen3` scores on faithfulness, answer relevancy, and factual correctness may be inflated relative to other models due to stylistic and lexical alignment between generation, response, and scoring [2]. Cross-validation with an independent judge model is left to future work.

VI. CONCLUSION AND FUTURE WORK

This study presents an empirical evaluation framework for open-source Large Language Models (LLMs) in ESG (Environmental, Social, and Governance) KPI extraction using Retrieval-Augmented Generation (RAG). Our analysis evaluates seven open-source models ranging from 2B to 30B parameters on a dataset of 498 ESG documents, revealing marked differences in performance characteristics across model architectures and sizes.

The performance evaluation reveals genuine variation across both retrieval and generation. Retrieval is strong but not perfect (context recall ≈ 0.58 – 0.61 , context precision ≈ 0.78 – 0.81 , context relevance 0.965 – 0.985 , LLM-judged context precision with reference ≈ 0.83 – 0.88), bounded by the top- $k=1$ design. Generation diverges most on faithfulness (0.607 – 0.822) and least on answer relevancy (0.760 – 0.881). Different models lead different metrics: `glm-4.7-flash:q4_K_M` leads faithfulness (0.822), `qwen3:4b-instruct` leads factual correctness (0.449), and `ministral-3:8b` leads answer relevancy (0.881). Factual correctness across all models (0.387 – 0.449) indicates ongoing opportunities for domain-specific fine-tuning and further enhancement.

Strategically, this evaluation framework offers ESG practitioners directional guidance for open-source model selection based on their specific requirements. Under the reported configuration, organizations can use these results as directional guidance: `glm-4.7-flash:q4_K_M` for faithfulness-critical applications, `qwen3:4b-instruct` for factual-accuracy-critical tasks (subject to the self-enhancement caveat), `ministral-3:8b` for query-focused applications, `gemma4:e4b` as a faithfulness-oriented alternative, `gemma4:e2b` for resource-constrained deployments, and `gemma3:4b` as a balanced default. The narrow factual-correctness band across all seven models indicates that domain-specific fine-tuning and human-in-the-loop validation remain valuable directions regardless of generator choice.

The study establishes a transparent and reproducible methodology for evaluating open-source LLMs in sustainability reporting, providing essential guidance for organizations seeking open-source AI solutions for ESG analytics.

Beyond the model comparison itself, the study contributes an ESG-RAG evaluation resource that combines a processed report corpus, grounded synthetic QA pairs, and reproducible evaluation artifacts for future benchmarking and ablation studies.

A. Future Work

Several avenues for future research emerge from this study:

Extended Open-Source Model Benchmarking: We plan to expand evaluation to include additional open-source models such as Llama 3, Mixtral, and emerging models from different parameter ranges (1B–70B). This broader comparison will provide deeper insights into the evolving landscape of open-source LLM capabilities for ESG applications.

Domain-Specific Fine-Tuning: Given the improved but still moderate factual correctness scores across all models, future work will investigate domain-specific fine-tuning approaches for ESG applications. This includes training models on specialized ESG datasets and developing adaptation techniques to further enhance factual accuracy in sustainability reporting contexts, building on the promising progress demonstrated in this evaluation.

Broader Corpus Validation: While our 498-document dataset provides coverage of EU companies’ ESG reporting, validating our findings across diverse geographies, industries, and reporting frameworks will enhance the generalizability of the evaluation framework.

Cost-Performance Optimization: Future research will explore optimization techniques for open-source model deployment, including quantization strategies, model compression, and efficient serving architectures to maximize performance efficiency for ESG applications.

Ablation Studies and Baselines: This study fixed chunk size (1000 characters), overlap (200 characters), and top-k (1) without ablating their individual effects. Future work should vary these hyperparameters systematically and include a non-RAG baseline (direct LLM inference without retrieval) to quantify the retrieval component’s contribution to overall performance.

Hybrid Architecture Evaluation: Future work will investigate hybrid approaches that combine different open-source models for different pipeline components, potentially optimizing both accuracy and computational efficiency for ESG RAG systems.

ACKNOWLEDGEMENTS

This research was partially supported by the Italian Research Center on High Performance Computing, Big Data and Quantum Computing (ICSC), funded by the European Union through NextGenerationEU (PNRR-HPC, CUP: C83C22000560007).

REFERENCES

- [1] Shafiuddin Rehan Ahmed et al. “Enhancing Retrieval for ESGLLM via ESG-CID—A Disclosure Content Index Finetuning Dataset for Mapping GRI and ESRS”. In: *arXiv preprint arXiv:2503.10674* (2025).
- [2] Amazon Web Services. *Generate Synthetic Data for Evaluating RAG Systems Using Amazon Bedrock*. <https://aws.amazon.com/blogs/machine-learning/generate-synthetic-data-for-evaluating-rag-systems-using-amazon-bedrock/>. AWS Machine Learning Blog. 2024.
- [3] Dario Amodi et al. “Concrete problems in AI safety”. In: *arXiv preprint arXiv:1606.06565* (2016).
- [4] Ozgur Ardic et al. “Information Extraction from Sustainability Reports in Turkish through RAG Approach”. In: *2024 32nd Signal Processing and Communications Applications Conference (SIU)*. 2024, pp. 1–4. DOI: 10.1109/SIU61531.2024.10600994.
- [5] Edward Beeching et al. *Open LLM Leaderboard*. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard. HuggingFace. 2023.
- [6] Florian Berg, Julian F Kölbel, and Roberto Rigobon. “Aggregate confusion: The divergence of ESG ratings”. In: *Review of Finance* 26.6 (2022), pp. 1315–1344.
- [7] Mattia Birti, Francesco Osborne, and Andrea Maurino. “Optimizing Large Language Models for ESG Activity Detection in Financial Texts”. In: *Proceedings of the 6th ACM International Conference on AI in Finance*. 2025. DOI: 10.1145/3768292.3770371.
- [8] Tom Brown et al. “Language models are few-shot learners”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 1877–1901.
- [9] Mark Carney. *Breaking the tragedy of the horizon: Climate change and financial stability*. Speech at Lloyd’s of London, Bank of England. 2015.
- [10] CDP (Carbon Disclosure Project). *CDP: Turning Transparency to Action*. <https://www.cdp.net/>. Accessed: 2025-07-08. 2025.
- [11] Confident AI. *DeepEval Documentation: RAG Evaluation*. <https://deepeval.com/docs/getting-started-rag>. Accessed: 2026-05-18. 2026.
- [12] European Commission. *Commission Delegated Regulation (EU) 2023/2772 supplementing Directive 2013/34/EU as regards sustainability reporting standards (ESRS)*. Official Journal of the European Union, L series. European Sustainability Reporting Standards, developed by EFRAG. 2023.
- [13] European Commission. *Corporate Sustainability Reporting Directive (CSRD)*. https://finance.ec.europa.eu/capital-markets-union-and-financial-markets/company-reporting-and-auditing/company-reporting/corporate-sustainability-reporting_en. 2023.
- [14] European Union, European Parliament & Council. *Directive (EU) 2022/2464 – Corporate Sustainability Reporting Directive (CSRD)*. <https://eur-lex.europa.eu/eli/dir/2022/2464/oj>. Adopted 16 December 2022; in force 5 January 2023; phased reporting from January 2024 onwards. 2022. (Visited on 07/08/2025).
- [15] ExplodingGradients. *Ragas: Supercharge Your LLM Application Evaluations*. <https://github.com/explodinggradients/ragas>. 2024.
- [16] Elias Frantar et al. “GPTQ: Accurate post-training quantization for generative pre-trained transformers”. In: *arXiv preprint arXiv:2210.17323* (2022).
- [17] Yunfan Gao et al. “Retrieval-Augmented Generation for Large Language Models: A Survey”. In: *arXiv preprint arXiv:2312.10997* (2023).
- [18] Gemma Team, Google DeepMind. *Gemma: Open models based on Gemini research and technology*. *arXiv preprint arXiv:2403.08295*. 2024.
- [19] Global Reporting Initiative (GRI). *GRI Standards: Universal, Topic and Sector Standards*. <https://www.globalreporting.org/>. Framework first published October 2021, effective for reports in January 2023. 2021. (Visited on 07/08/2025).
- [20] Yuxian Gu, Xiaochuang Han, Zheng Li, et al. “A Survey on LLM-as-a-Judge”. In: *arXiv preprint arXiv:2411.15594* (2024).
- [21] Chaoyue He et al. “ESGenius: Benchmarking LLMs on Environmental, Social, and Governance (ESG) and Sustainability Knowledge”. In: *arXiv preprint arXiv:2506.01646* (2025).
- [22] Seungone Kim et al. “Prometheus: Inducing Fine-Grained Evaluation Capability in Language Models”. In: *arXiv preprint arXiv:2310.08491* (2024).
- [23] Patrick Lewis et al. “Retrieval-augmented generation for knowledge-intensive NLP tasks”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 9459–9474.
- [24] Alexander H Liu et al. “Ministral 3”. In: *arXiv preprint arXiv:2601.08584* (2026).
- [25] Yang Liu et al. “G-Eval: NLG Evaluation Using GPT-4 with Better Human Alignment”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2023, pp. 2511–2522. DOI: 10.18653/v1/2023.emnlp-main.153.

- [26] Rafik Mankour et al. “Climate Finance Bench: A Benchmark for RAG over Corporate Climate Disclosures”. In: *arXiv preprint arXiv:2505.22752* (2025).
- [27] NVIDIA. *Nemotron 3 Nano: Open, Efficient Mixture-of-Experts Hybrid Mamba-Transformer Model for Agentic Reasoning*. arXiv preprint arXiv:2512.20848. 2025.
- [28] Angel Ontiveros et al. “ESG-Consultant: Developing of an ESG Compliance Consulting Tool for Companies Using RAG”. In: *Natural Language Processing and Information Systems*. Ed. by Ryutaro Ichise. Cham: Springer Nature Switzerland, 2026, pp. 223–229. ISBN: 978-3-031-97144-0.
- [29] Augusto Londero Orsolin et al. “Analysis of Indicators for the Integration, Implementation, and Development of the Environmental, Social, and Governance – ESG Report in ISE/B3 Companies”. In: *Revista Catarinense da Ciência Contábil* (2024). DOI: 10.16930/2237-7662202435152.
- [30] Qwen Team. *Qwen3 Technical Report*. arXiv preprint arXiv:2505.09388. 2025.
- [31] Teodora Maria Rusu et al. “Sustainability Performance Reporting.” In: *Sustainability (2071-1050)* 16.19 (2024).
- [32] Jon Saad-Falcon et al. “ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems”. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2024, pp. 338–354. DOI: 10.18653/v1/2024.naacl-long.20.
- [33] Sustainability Accounting Standards Board (SASB). *Sustainability Accounting Standards Board (SASB) Conceptual Framework*. <https://www.sasb.org/>. Standards launched for 77–79 industries; stewardship transferred to IFRS Foundation’s ISSB in August 2022. 2017. (Visited on 07/08/2025).
- [34] Task Force on Climate-related Financial Disclosures. *Recommendations of the Task Force on Climate-related Financial Disclosures*. <https://www.fsb-tcfd.org/recommendations/>. 2017.
- [35] Task Force on Climate-related Financial Disclosures, Financial Stability Board. *Recommendations of the Task Force on Climate-related Financial Disclosures*. <https://www.fsb-tcfd.org/>. Final recommendations released June 2017; framework active until October 12, 2023; site maintained as resource; monitoring of progress handed to IFRS Foundation as of November 2023. 2017. (Visited on 07/08/2025).
- [36] Team GLM et al. *ChatGLM: A family of large language models from GLM-130B to GLM-4 All Tools*. arXiv preprint arXiv:2406.12793. 2024.
- [37] Rafael Teixeira de Lima et al. “Dataset Taxonomy and Generation Strategies for Evaluating RAG Systems”. In: *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*. 2025. URL: <https://aclanthology.org/2025.coling-industry.4>.
- [38] Hugo Touvron et al. “LLaMA: Open and efficient foundation language models”. In: *arXiv preprint arXiv:2302.13971* (2023).
- [39] Qilong Wu 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*. 2025, pp. 1184–1203. DOI: 10.18653/v1/2025.findings-naacl.66.
- [40] Tsung-Yu Yang and Meng-Chi Chen. “ESGLens: An LLM-Based RAG Framework for Interactive ESG Report Analysis and Score Prediction”. In: *arXiv preprint arXiv:2604.19779* (2026). URL: <https://arxiv.org/abs/2604.19779>.
- [41] Lei Zhang et al. “Benchmarking Multimodal Understanding and Complex Reasoning for ESG Tasks”. In: *arXiv preprint arXiv:2507.18932* (2025).
- [42] Chi Zhu et al. “JudgeLM: Fine-Tuned Large Language Models Are Scalable Judges”. In: *International Conference on Learning Representations*. 2025. URL: <https://openreview.net/forum?id=xsELpEPn4A>.
- [43] Yi Zou et al. “ESGReveal: An LLM-based Approach for Extracting Structured Data from ESG Reports”. In: *Journal of Cleaner Production* 489 (2025). DOI: 10.1016/j.jclepro.2024.144572.