Knowledge Graph aided LLM based ESG Question-Answering from News

Tanay Kumar Gupta[*,†], Tushar Goel[†], Ishan Verma, Lipika Dey[‡] and Sachit Bhardwaj[‡]

TCS Research, New Delhi, India

Abstract

Organizations around the globe have acknowledged the importance of sustainability. Sustainability performance has gained traction in investing and risk management and is now an integral part of business planning. The volume and velocity of information being published on the web have made use of natural language processing evident for insight generation. With the recent advancements in language modelling and the availability of Large Language Models (LLM), conversational insight generation is increasingly becoming popular. LLM combined with advanced retrieval techniques has eased the task of question-answering over large natural language datasets. In this work, we present a novel approach that leverages Knowledge Graph-based Retrieval Augmented Generation (KG-RAG) to facilitate question-answering in the context of sustainability news articles and corporate Environmental, Social, and Governance (ESG) performance. Our methodology encompasses the creation of an ESG Knowledge Graph, retrieval techniques that identify contextually relevant information, and an LLM-based answer-generation framework. We have experimented with multiple LLM models and have shown a comparative study of their performances against several baseline algorithms.

Keywords

Sustainability, ESG, Knowledge Graph, Large Language Models, Question Answering

1. Introduction

In the contemporary landscape, the imperative of sustainability for individuals, corporations, and government becomes increasingly evident [1]. Corporations embracing sustainability practices often find that they lead to not only cost savings but also increased profitability and long-term value. With increasing popularity amongst investors, it can facilitate access to capital and support business growth. Environmental, social, and governance, or ESG, is a concept that the United Nations Global Compact introduced for sustainability reporting in 2004 [2]. Since then, companies have used these to highlight their efforts and commitments towards sustainability. The Global Reporting Initiative (GRI)[1] is a leading international organization

The 2nd International Workshop on Knowledge Graphs for Sustainability (KG4S2024) – Co-located with the 21st Extended Semantic Web Conference (ESWC2024), May 27th, 2024, Hersonissos, Greece.

[*] Corresponding author.

[†] Authors contributed equally.

[‡] Authors contributed during their tenure with TCS Research.

gupta.tanay@tcs.com (T. K. Gupta); t.goel@tcs.com (T. Goel); ishan.verma@tcs.com (I. Verma); lipika.dey@tcs.com (L. Dey); sachit.bhardwaj3@tcs.com (S. Bhardwaj)

[1] https://www.globalreporting.org/about-gri/

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

that promotes sustainability reporting. It provides a comprehensive framework for companies and organizations to report their economic, environmental, and social performance.

Sustainability reports are published by organizations every year highlighting their efforts in the area of ESG. Sustainability news on the other hand stands out for its advantage over sustainability reports due to real-time event updates. It also offers a diverse array of perspectives and timely insights from experts and stakeholders, enabling a holistic understanding of sustainability challenges. Given the speed and volume of information flow, organizations have been adopting natural language processing techniques for news analysis for quite some time now [3].

Question-answering (QA) systems are pivotal in sustainability analysis due to their efficiency in data retrieval, contextual understanding, and ability to provide interdisciplinary insights. Sustainability encompasses diverse domains, and QA systems assist stakeholders in providing answers by extracting relevant data from various sources while understanding the context of their inquiries. These systems reduce the manual efforts required to read and review each story. QA systems aid in benchmarking, support data-driven decision-making, and tackle complex queries, ultimately contributing to more informed sustainability analysis and strategy development[4]. The emergence of large language models (LLM), trained on extensive text data, represents a groundbreaking advancement in natural language processing. These models excel at producing highly articulate and cohesive text with minimal input, unlocking new potentials in conversational AI, creative writing, and various other fields.

In this work, we present a system that facilitates ESG question-answering over news corpus using a sustainability knowledge graph-aided retrieval augmented generation through LLMs. Our methodology involves the construction of an ESG Knowledge Graph built on top of the GRI reporting structure that encapsulates the essential entities and relationships extracted from sustainability news articles. We show that KG-enhanced retrieval achieves better QA performance than simple embedding matches. We have experimented with multiple open-source LLMs and have shown their comparative performances. This framework ensures that the answers provided to investors are not only informative but also comprehensible.

The contributions of this work are highlighted below:

- The creation of Sustainability knowledge Graph using GRI reporting structure and other web sources.
- We propose the use of KG driven document retrieval for answering ESG questions.
- Human like answer generation using LLMs.
- Evaluation and comparison of results from our approach with several LLM based baseline approaches.
- Comparative evaluation of performances of multiple LLMs for ESG QA.

In the next section, we present details of the proposed system and its components.

2. ESG Question-Answering System

Figure 1 shows the architecture and components of the proposed system. The ESG knowledge graph creation is presented in section 2.1. The news processing module presented in section

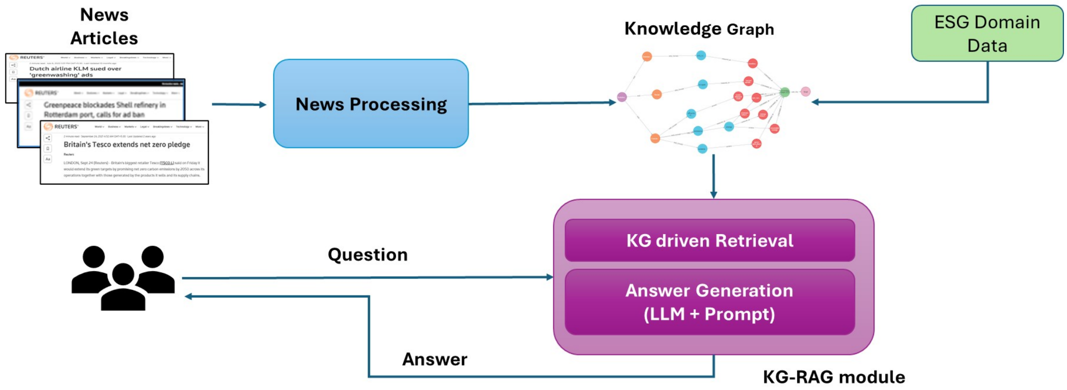
News Articles
Dutch airline KLM sued over 'greenwashing' ads
Greenpeace blockades Shell refinery in Rotterdam port, calls for ad ban
Britain's Tesco extends net zero pledge
News Processing
Knowledge Graph
ESG Domain Data
KG driven Retrieval
Answer Generation (LLM + Prompt)
Question
Answer
KG-RAG module

Figure 1: ESG Question-Answering system components

2.2 extracts information components from news articles and updates the existing Knowledge graph. Section 2.3 details the knowledge graph-driven news document retrieval and answer generation methodology.

2.1. ESG Knowledge Graph

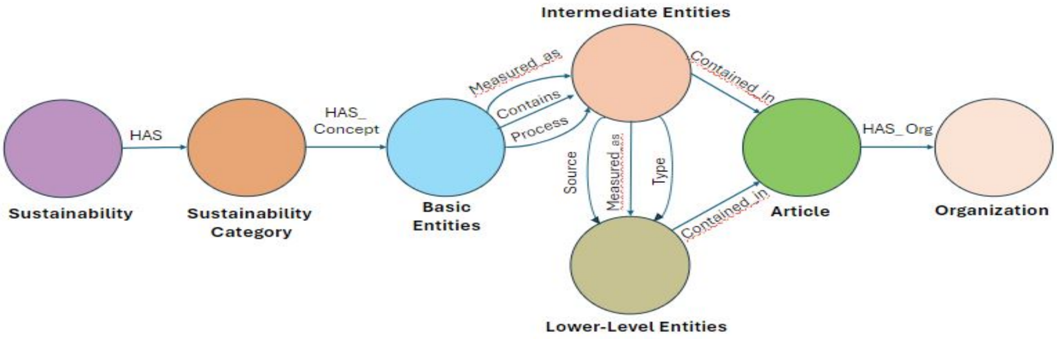

Figure 2: ESG knowledge graph schema

The ESG knowledge graph is created using elements taken from the widely recognized GRI sustainability reporting framework, FinSim4-ESG 2022 shared task, [5], and updated with information extracted from news articles. In GRI, sustainability reporting is considered from three perspectives: economic, environmental, and social [6] defined in the form of a series of indicators. The FinSim4-ESG 2022 task focused on the elaboration of an ESG taxonomy based on data like companies' sustainability reports, annual reports, and environment reports, and made use of them to analyze how an economic activity complies with the taxonomy.

Figure 2 represents the schema of the Sustainability knowledge Graph. The graph comprises multiple levels of entities defined as follows:

- **Sustainability**- This represents the root node of the graph.
- **Sustainability categories** – Three nodes branched under Sustainability represent the three aspects of ESG that are Environment, Social, and Governance.
- **Basic entities** – These include 32 fundamental terms representing basic indicators of GRIs which encompass the fundamental topics within the sustainability domain, such as water (GRI_303), emissions (GRI_305), and public policy (GRI_415), etc. Apart from GRI indicators, we have added additional basic entities taken from FinSim4-ESG shared task [5] data like Sustainable agriculture (GRI_none), executive compensation (GRI_none), climate change (GRI_none), etc. Each entity at this level contains two attributes, GRI ID and the definition of an indicator. We align GRI indicators with their corresponding sustainability categories by referencing the indicator definitions. In this work, GRI 300 series and GRI 400 series indicators are associated with Environment and Social category, respectively. However, for Governance, we have taken indicators from GRI 200 and 400 series, as GRI standards have not defined this category explicitly.
- **Intermediate Entities** – These are more specific than basic entities. For example, freshwater, greenhouse gas emission, bribery, tax fraud, lobbying, etc. fall into this category. The connections between basic entities and intermediate entities are represented by the relation '*Contains*' or '*Process*' or '*Measured_as*' depending on the nature of the relationships or interactions within the specified domain of the ontology. For instance, (biodiversity, *contains*, flora), (water, *measured_as*, water consumption).
- **Lower-Level Entities** – As we move down the hierarchy, entities become more specialized and granular, representing specific instances or sub-types. Examples include carbon-dioxide emission and methane emission are sources of greenhouse gas emissions. Similarly, physical hazard and chemical hazard are types of work-related injuries. The connections between Intermediate entities and Lower-level entities are represented by the relation '*Type*' or '*Source*' or '*Measured_as*'. For instance, (Renewable energy, *type*, solar energy), (greenhouse gas emission, *source*, carbon-dioxide emission), (emission reduction, *measured_as*, green transportation).

Intermediate and lower-level entities are derived from detailed descriptions of GRI indicators given in GRI standards glossary. Sample hierarchy of knowledge graph entities from basic to lower level is shown in Appendix A Table 3.

2.2. News processing

We have created a sustainability news corpus containing 4331 documents published during the year 2021-2022 from Reuters sustainability business news[2] to keep the collection sustainability focused. Each article is processed and attached to the knowledge graph in the form of entities and relations described below:

- **Article Nodes** - A news article within a knowledge graph is represented as an entity that encapsulates information obtained from news. Each article node comprises 5 attributes defined as follows:

[2] https://www.reuters.com/sustainability/

- Article_Id – a unique identifier for each article node.
- Publication_Date – It indicates the date when the article was published.
- Headline – It represents the title of the article.
- Article Summary – This attribute contains a summary of the content of the article obtained using a pre-trained transformer-based encoder-decoder model PEGASUS [7]. It helps in obtaining a fixed length representation of the article's content.
- All_Organizations – It represents a collection of all organizations mentioned within the article. We have used Named Entity Recognition module to obtain the list of organizations mentioned in the news article.

- **Organization Node** – This type of node in the knowledge graph represents the organizations that are frequently occurring in news articles. These nodes are connected to their corresponding article nodes through a *HAS_Org* relation. Creating a separate organization node aid in faster retrieval and organization-based filtering of articles.

Linking news to KG: For each article, we employ RAKE (Rapid Automatic Keyword Extraction) [8] algorithm to extract informative phrases that encapsulate the main themes and topics discussed in the articles. RAKE is a widely used unsupervised keyword extraction technique that leverages statistical measures, such as word frequency and co-occurrence, to identify important phrases within a document. Subsequently, we utilize the phrase-BERT algorithm [9] to generate embeddings for these extracted key phrases, capturing their semantic meanings in dense vector representations. Similarly, each entity in the KG is represented by 768-dimensional phrase-BERT embedding. For each RAKE key phrase, its maximally similar phrase in KG entities is found using cosine similarity. Each article phrase is mapped to the maximally matching entity and depicts a mapping of a single article to multiple intermediate or lower-level entities in the KG. Hence, an edge is built between an article node and all mapped intermediate or lower-level entities with "*Contained_in*" relation.

Figure 3 shows a subset of basic entities, intermediate entities, lower-level entities and article node that are interconnected based on the relationships described in the schema.

2.3. KG driven Retrieval augmented generation (KG-RAG)

In this module, we present a retrieval strategy that uses the knowledge graph created earlier for fetching relevant article nodes based on a user question, the content of which is further used for generating answer to the question using a large language model. Our QA module comprises two components: (i). KG Retriever, which takes a given question q within the context of a knowledge graph $G(N, E)$ and outputs the top k relevant article nodes; and (ii). an Answer Generator responsible for crafting the sequence of answers.

Knowledge Graph Retriever: Consider a knowledge graph $G(N, E)$, where N denotes the nodes and E represents the edges within the knowledge graph. Each node N contains a variable number of attributes. Let N_i represent the i^{th} node of the knowledge graph and N_{ij} is the j^{th} attribute of the node. It's important to note that nodes of the knowledge graph with different labels can have different numbers of attached attributes. The edges, denoted as E_{ik}, serve as connection between nodes N_i and N_k. Our task is to retrieve the top k most relevant article nodes to the question q.

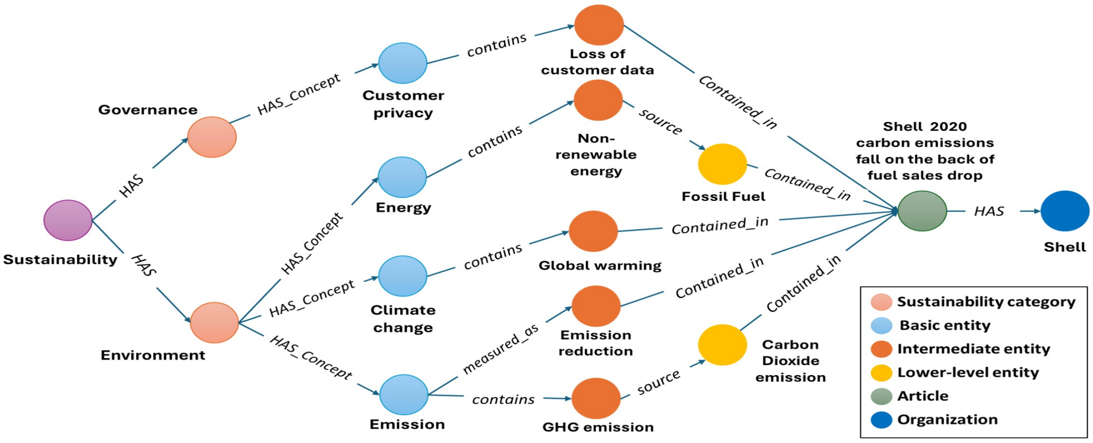

Figure 3: Subset of knowledge graph showing different entities and their relationships.

Graph Embedding: Given that each node is characterized by multiple attributes, we encode each attribute of the node as a dense vector using the Universal Sentence Encoder (USE). USE is a pre-trained deep learning model that encodes sentences into fixed-size vectors, capturing semantic meanings [10]. The USE produces a 512-dimension vector.

Retrieval Based on Traversed Sub-Graphs: In the process of creating embedding, the graph structure has not been utilized. To maximize the utility of the graph, it is essential to incorporate the path from article node to the root node. This path encompasses additional information present at the article level. Our approach involves traversing the graph from the article node through all possible paths to the sustainability (root) node, thereby creating a sub-graph that contains all paths leading to the article from the sustainability node(root). Paths leading to the root node are traced utilizing a depth-first algorithm [11]. The process initiates with all article nodes, identifying their first neighbors and proceeding in a singular direction to include all inter-related nodes from each level back to the root node as shown in algorithm 1. These traversed sub-graphs are denoted as the document representation Doc_rep, and they play a crucial role in generating scores for articles.

Document scoring: Each input question q is encoded using USE. Subsequently, we calculate the cosine similarity between the encoded question and all nodes N_i in the knowledge graph. Node similarity score is the maximum of similarity scores for attributes. To score an article node based on these similarity scores, we leverage the traversed sub-graph structure and $Doc\_rep\_Score_{ij}$ represents the score of j^{th} node of i^{th} document in Doc_rep in decreasing order. Our methodology involves determining the final score for each article by adding the maximum scored node from the traversed sub-graph and the average scores of the other top n-scored nodes. In cases where a question exhibits the highest similarity to a single path, that specific path singularly contributes to the overall score. Conversely, when multiple paths exist, the top n scores are distributed among these paths, highlighting diverse relevant paths. The scoring function utilized is outlined as follows:

$$Node\_score_i = max(cosine\_similarity(q, N_{ij}))\forall j$$

Algorithm 1 Document Representation or Traversed sub graph & Document Score

```
**Inputs:** $G(N, E)$, $Node\_Score$
**Output:** $Doc\_rep, Document\_Score$
$Doc_{id}, Neighbour\_list, Doc\_rep, Document\_Score \leftarrow []$
for $N = 1, 2, \ldots, n$ do
    if $N_i.label$ is Document then
        $Doc_{id} \leftarrow Node\_Id$
        $Neighbour\_list \leftarrow$ all nodes in the path from $N_i$ to sustainability
        $Doc\_rep \leftarrow [Doc_{id}, Neighbour\_list]$
    end if
end for
for $doc = 1, 2, \ldots, k$ in $Doc_{id}$ do
    $Document\_neighbour\_score \leftarrow [Node\_Score[doc]]$
    for $Neighbours = 1, 2, \ldots, n$ in $Neighbour\_list$ do
        $Document\_neighbour\_score \leftarrow Node\_Score[Neighbours]$
    end for
    $m = mean(sort(Document\_neighbour\_score)[1:4]$
    $Document\_Score \leftarrow max(Document\_neighbour\_score) + m$
end for
```

$$Document\_Score_i = Doc\_rep\_Score_{i0} + \frac{1}{n}\sum_{j=1}^{n} Doc\_rep\_Score_{ij}$$

Answer Generation: We have experimented with multiple LLM models for answer generation. LLM model takes as input the question, a fixed prompt, and the content of the retrieved article to generate the answer.

3. Experiments and Results

Within our Knowledge Graph, we have a comprehensive representation of 540 unique organizations. Notably, 28 organizations have more than 30 articles each. Among these, *Shell* holds the lead with a maximum of 62 articles. Additionally, 86 organizations are within the range of 10-30 articles, contributing to the richness of our graph. The majority of organizations, totaling 426, are associated with less than 10 articles each. This distribution reflects the diversity of information captured in the Knowledge Graph across various organizations and their respective article counts. We have curated a question dataset consisting of 144 questions generated with the help of multiple ESG domain experts. These questions are expert categorized into environment, social, and governance categories having 59, 41, and 44 questions, respectively. As validating QA on all companies entailed a significant computational load, we chose to run all 144 questions on 30 companies to evaluate the performance of our models. We have randomly picked 15 companies from greater than 30 articles and 10 from 10-30 articles and 5 from less than 5 articles. The answers generated by different methods are manually evaluated by domain experts. We have experimented with two variations of the Knowledge Graph referenced in this section as

KG and KG_Enhanced. KG_Enhanced includes additional attributes for article nodes such as summary and all organization which were omitted from KG.

For evaluating proposed KG-RAG architecture, we utilized the following baseline retrieval methods, returning top 3 news articles for each question considering news title and content:

- **BM25**- It calculates relevance scores based on term frequency, document length, and inverse document frequency. [12].
- **LangChain**- LangChain employs a retriever using a sentence transformer model. We created embeddings for complete articles separately using all-MiniLM-L6-v2 [13]. Based on similarity score with question we get top 3 articles.
- **MMR** - Maximal Marginal Relevance ensures the selection of diverse and relevant information by iteratively maximizing the similarity between retrieved items while minimizing redundancy [14]. It also uses all-MiniLM-L6-v2 for embeddings.

For the answer generation task, we have used the following LLMs:

- **Llama2** - Llama 2 7b chat-hf model, which is a transformer-based model with 7 billion parameters is used with 4-bit quantization [15]. It is a fine-tuned model of base Llama2 using RLHF. Its context length is 4096 tokens.
- **Mistral** - Mistral-7B-Instruct-v0.1 Large Language Model (LLM) is an instruction fine-tuned version of the Mistral-7B-v0.1 generative text model using a variety of publicly available conversation datasets [16]. This model also has context length is 4096 tokens and its 4-bit quantized version is used.
- **Phi2** - Phi-2 is a Transformer with 2.7 billion parameters [17]. It was trained using the same data sources as Phi-1, augmented with a new data source that consists of various NLP synthetic texts and filtered websites. Its context length is 2048 tokens.
- **TinyLlama** - TinyLlama-1.1B-Chat-v1.0, following the Llama 2 architecture, is a chat model fine-tuned on top of TinyLlama-1.1B base model [18].

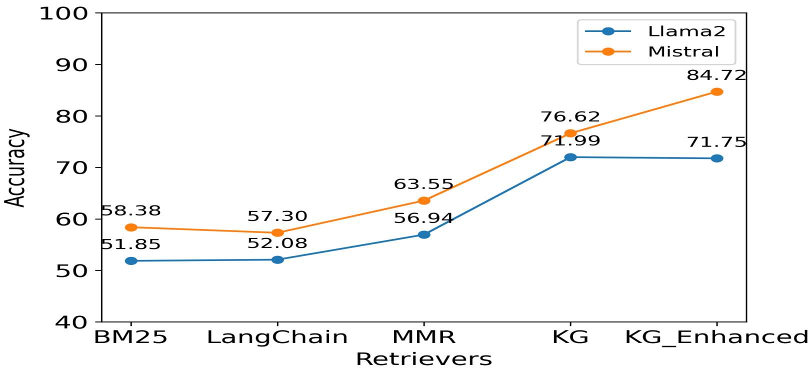

Figure 4: Retriever wise Accuracy (in percentage) plot for Llama2 & Mistral

Figure 4 illustrates the comparative performance of the two LLMs, Llama2 and Mistral, across different retrieval methods. Here accuracy represents percentage of correct answers generated. We observe a consistent trend where the accuracy of both Llama2 and Mistral tends to increase as we move from traditional models such as BM25, LangChain and MMR to more advanced KG based models, which is KG and KG_Enhanced. Notably, the KG_Enhanced model achieves the highest accuracy with Mistral and comparable accuracy to KG retriever with Llama2. Overall, the KG driven models achieves better accuracy than baselines with both LLMs. This observation underscores the effectiveness of leveraging KG in enhancing the accuracy of LLM based question-answering task.

Category	Environment	Social	Governance	Overall
KG_Llama2	70.61%	75.6%	70.45%	71.99%
KG_Mistral	70.72%	81.29%	80.3%	76.62%
KG_Phi2	52.07%	52.03%	59.84%	54.39%
KG_Enhanced_Llama2	74%	71.54%	68.93%	71.75%
KG_Enhanced_Mistral	**80.78%**	**82.54%**	**85.25%**	**84.72%**
KG_Enhanced_Phi2	49.71%	56.09%	63.63%	55.78%

Table 1: Overall and Category wise accuracy across LLMs

Table 1 presents the performance of different models in answering questions from different ESG categories. The numbers represent the percentage of questions answered correctly by each model in each category. KG_Enhanced_Mistral outperformed other models with an accuracy of 80.78%, showcasing its effectiveness in answering questions related to the environment category. In social category, KG_Enhanced_Mitral and KG_Mistral emerged as the top-performing models with 82.54% and 81.29% accuracy respectively whereas KG_Enhanced_Llama2 and KG_Llama2 demonstrated consistent performance, both exceeding 70% accuracy. Again, KG_Enhanced_Mistral showcased the highest accuracy in the governance category, achieving 85% accuracy. Overall, KG_Enhanced_Mistral and KG_Mistral maintained a competitive edge across all categories, contributing to an overall accuracy of 84.72% and 76.62%, respectively. Moreover, KG_Enhanced_Llama2 and KG_Llama2 consistently performed well, surpassing the 71% accuracy mark. This shows the ability of Knowledge Graph-based models to capture semantic relationships, showcased robust performance across diverse categories.

LLM Comparison: In general, Models with 7 billion parameters outperform smaller models. Mistral surpasses Llama2 and Phi2. In our experiments we found that Mistral tends to generate an answer only when it finds an exact match in the context; otherwise, it returns "Insufficient Information Available," as indicated in the prompt. Conversely, Llama2 tends to produce a form of summary followed by "Insufficient Information Available to answer or no mention about the question in the context". Llama2 exhibits a greater tendency to generate more incomplete answers than Mistral and is more prone to hallucination when there is no relevant article. In contrast, Mistral tends to extract more accurate answers from articles and has a lower tendency to hallucinate. On the other hand, if the correct articles are passed, Llama2 and Mistral both are comparably effective. Pointer-wise answers are better for Mistral. While Phi2 is not directly fine-tuned for instruction following tasks, it achieves comparable results. In cases of insufficient information in the articles, Phi2 may generate an answer based on its knowledge, there is no relevant information with an explanation and in very few cases, directly outputs "Insufficient Information Available" or "No mention in the text." Additionally, Phi2 tends to make more assumptions than larger models. Sample question answers are shown in Appendix A along with retrieved article headline.

TinyLlama with 1.1 billion parameters on the other hand exhibited comparatively weaker performance, consistently generating responses based on its own knowledge or producing random output. The answers generally fall short of expectations, and despite experimenting with various prompts, it consistently fails to generate satisfactory responses. As an illustration, when posed with the inquiry "Does Amazon have any initiatives or partnerships for Forest Restoration?" and provided with KG enhanced articles, other models deliver meaningful responses. In sharp contrast, TinyLlama offers an unhelpful reply, "# Given a text, generate response." Moreover, in certain instances, it tends to generate answers that lack coherence, following no discernible patterns and resembling random output.

4. Related Work

In the dynamic landscape of information retrieval, a task that is both crucial and continually evolving, recent research has made significant strides with the introduction of diverse retrieval models. Traditional BM25 has provided a robust foundation, relying on sparse representations to facilitate efficient document retrieval [12]. On the other hand, the Dense Passage Retrieval framework which leverages dense representations and neural networks is also achieving state-of-the-art performance in open-domain question answering and document retrieval [19]. Furthermore, innovative methodologies, such as Colbert, introduced by Ma et al.[20] emphasize the importance of context augmentation through text generation. These approaches have proven to be instrumental in enhancing query semantics and optimizing information retrieval processes. In [21], LLM-based diverse retrieval models and techniques collectively contribute to the ongoing evolution of information retrieval, offering valuable insights and solutions to address the multifaceted challenges in this field. However, with the advancement of Large Language models viz GPT3 [22], PaLM2 [23], Gemini [24], RAG stands out due to factual answer generation. The Retrieval-Augmented Generation (RAG) [25] model combines a retriever with a generator to answer questions over a knowledge base. A general-purpose fine-tuning recipe for RAG models which combines pre-trained parametric and non-parametric memory for language generation is used.

5. Conclusion & Future Work

In this work, we presented a Knowledge-Graph driven Retrieval Augmented Generation system to facilitate ESG question answering over news corpus. Our methodology encompasses the creation of an ESG Knowledge Graph, retrieval techniques that identify contextually relevant information, and an LLM-based answer-generation framework. We have experimented with multiple LLMs and have shown a comparative study of their performances against several baseline algorithms. LLMs with a higher number of parameters show promising results. In future, we are planning to explore LLM fine-tuning on ESG data. Also, we are yet to include temporal questions that span across multiple years of data. We are looking forward to advancements in both retrieval and LLM space to expand the current work.

References

[1] R. Lozano, R. von Haartman, Reinforcing the holistic perspective of sustainability: Analysis of the importance of sustainability drivers in organizations, Corporate Social Responsibility and Environmental Management 25 (2018) 508–522.

[2] A. Tsang, T. Frost, H. Cao, Environmental, social, and governance (esg) disclosure: A literature review, The British Accounting Review 55 (2023) 101149.

[3] S. Murakami, S. Muraoka, Exploring the potential of internet news for supply risk assessment of metals, Sustainability 14 (2022). URL: https://www.mdpi.com/2071-1050/14/1/409. doi:10.3390/su14010409.

[4] B. Bui, M. N. Houqe, M. Zaman, Climate governance effects on carbon disclosure and performance, The British Accounting Review 52 (2020) 100880.

[5] J. Kang, I. El Maarouf, FinSim4-ESG shared task: Learning semantic similarities for the financial domain. extended edition to ESG insights, in: C.-C. Chen, H.-H. Huang, H. Takamura, H.-H. Chen (Eds.), Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 211–217. URL: https://aclanthology.org/2022.finnlp-1.28. doi:10.18653/v1/2022.finnlp-1.28.

[6] J. Webber, A programmatic introduction to neo4j, in: Proceedings of the 3rd annual conference on Systems, programming, and applications: software for humanity, 2012, pp. 217–218.

[7] J. Zhang, Y. Zhao, M. Saleh, P. Liu, Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, in: International Conference on Machine Learning, PMLR, 2020, pp. 11328–11339.

[8] S. Rose, D. Engel, N. Cramer, W. Cowley, Automatic keyword extraction from individual documents, Text mining: applications and theory (2010) 1–20.

[9] S. Wang, L. Thompson, M. Iyyer, Phrase-bert: Improved phrase embeddings from bert with an application to corpus exploration, arXiv preprint arXiv:2109.06304 (2021).

[10] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. St. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, R. Kurzweil, Universal sentence encoder for English, in: E. Blanco, W. Lu (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 169–174. URL: https://aclanthology.org/D18-2029. doi:10.18653/v1/D18-2029.

[11] R. Tarjan, Depth-first search and linear graph algorithms, SIAM journal on computing 1 (1972) 146–160.

[12] S. Robertson, H. Zaragoza, et al., The probabilistic relevance framework: Bm25 and beyond, Foundations and Trends® in Information Retrieval 3 (2009) 333–389.

[13] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: https://arxiv.org/abs/1908.10084.

[14] S. Guo, S. Sanner, Probabilistic latent maximal marginal relevance, in: Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, 2010, pp. 833–834.

[15] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).

[16] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand, et al., Mixtral of experts, arXiv preprint arXiv:2401.04088 (2024).

[17] Y. Li, S. Bubeck, R. Eldan, A. Del Giorno, S. Gunasekar, Y. T. Lee, Textbooks are all you need ii: phi-1.5 technical report, arXiv preprint arXiv:2309.05463 (2023).

[18] P. Zhang, G. Zeng, T. Wang, W. Lu, Tinyllama: An open-source small language model, arXiv preprint arXiv:2401.02385 (2024).

[19] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih, Dense passage retrieval for open-domain question answering, arXiv preprint arXiv:2004.04906 (2020).

[20] O. Khattab, M. Zaharia, Colbert: Efficient and effective passage search via contextualized late interaction over bert, in: Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, 2020, pp. 39–48.

[21] Y. Zhu, H. Yuan, S. Wang, J. Liu, W. Liu, C. Deng, Z. Dou, J.-R. Wen, Large language models for information retrieval: A survey, arXiv preprint arXiv:2308.07107 (2023).

[22] L. Floridi, M. Chiriatti, Gpt-3: Its nature, scope, limits, and consequences, Minds and Machines 30 (2020) 681–694.

[23] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, et al., Palm 2 technical report, arXiv preprint arXiv:2305.10403 (2023).

[24] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al., Gemini: a family of highly capable multimodal models, arXiv preprint arXiv:2312.11805 (2023).

[25] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in Neural Information Processing Systems 33 (2020) 9459–9474.

A. Appendix

Table 2: Sample question and their answers obtained using the proposed framework. The ESG categorization of each question is indicated in brackets.

Question: Has Tesla permanently adopted a work-from-home policy? (S) **Answer:** No, Tesla has not permanently adopted a work-from-home policy. Elon Musk, the CEO of Tesla, has asked employees to return to the office or leave the company. The email sent to employees stated that everyone at Tesla is required to spend a minimum of 40 hours in the office per week. If an employee does not show up, they will be assumed to have resigned. The more senior the employee, the more visible their presence must be in the office. **Retrieved Articles**: Elon Musk tells Tesla staff: return to office or leave
Question: Does Amazon emphasize team diversity? (S) **Answer:** Yes, Amazon emphasizes diversity. According to the article, Amazon has hired Candi Castleberry Singleton as the new VP of Global Diversity, Equity, and Inclusion. This suggests that Amazon is making an effort to increase diversity within the company. Additionally, the article mentions that Amazon the number of aims to double Black leaders in the company and increase the percentage of women in senior tech roles. **Retrieved Articles:** 1.Amazon hires Twitter executive to tackle diversity challenges 2.Amazon aims to double U.S. Black employees in leadership this year
Question: Are there any reported cases of environmental violation or damage by Tesla or its supply chain activities? (E) **Answer:** Based on the article provided, there have been reports of environmental violations and damage related to Tesla's supply chain activities. According to the article, Tesla is under investigation by the U.S. Securities and Exchange Commission (SEC) regarding a whistle-blower complaint that the company failed to properly notify its shareholders and the public of fire risks associated with solar panel system defects over several years. Additionally, the article mentions that Tesla faced at least 10 lawsuits alleging widespread race discrimination or sexual harassment, including one by a California state civil rights agency. These reports suggest that Tesla may have violated environmental laws or regulations related to its supply chain activities. However, without further information, it is difficult to determine the extent of these violations or the potential environmental damage caused. **Retrieved Articles:** SEC probes Tesla over whistle-blower claims on solar panel defects
Question: How many women are employed by Amazon? (G) **Answer:** According to the provided text, as of 2020, women comprised 31.4% of Amazon's U.S. corporate employees. **Retrieved Articles:** Amazon aims to double U.S. Black employees in leadership this year
Question: Has Shell adopted Sustainable Aviation Fuel to increase decarbonization? (E) **Answer:** Yes, according to the provided text, Shell has announced plans to start producing low-carbon jet fuel at scale by 2025, with the goal of increasing its production of Sustainable Aviation Fuel (SAF) to 10% of its global aviation fuel sales by 2030. **Retrieved Articles:** 1.Oil giant Shell sets sights on sustainable aviation fuel take_off 2.Rolls_Royce partners with Shell in sustainable aviation fuel push

Table 3: ESG Knowledge Graph entities - examples

Basic Entities	Intermediate Entities	Lower-Level Entities
Energy (GRI_302)	renewable energy	solar energy, wind energy, hydropower, bio-energy, geothermal energy, renewable thermal energy, solar thermal energy, marine energy, bio-fuels, hydrogen, ocean
	Non-renewable energy	nuclear energy, coal, natural gas, petroleum, fossil fuel, gasoline, diesel fuel, jet fuel, liquified petroleum gas
	energy intensity	
	energy consumption	
	energy reduction	energy efficient vehicle
Occupational health and safety (GRI_403)	safe work environment	
	accidental spill	
	work-related injury or ill health	physical hazard, ergonomic hazard, chemical hazard, biological hazard, psycho social hazard, mental hazard
	hazard identification & risk assessment	
	employee fatalities	
Emission (GRI_305)	green house gas emission	carbon-dioxide emission, global warming potential, methane emission, nitrous oxide, sulphur oxide

Table 4: Sample of Prompt used with LLM for answer generation

```
"""
Only use the text given within context "##" to answer the question, don’t make up anything else based on your knowledge. If you don’t get the answer from context or no mention of answer in context, then generate "insufficient information available" and don’t add anything else. Answer in format
## Context: {context} ## Question: {question} Answer:
"""
```

Hardware: In our experiments, we utilized the MiGA100 GPU, which boasts 14 vCPUs, 60 GiB of RAM, and 20 GiB of GPU memory for our generation tasks. Additionally, we conducted experiments on the free version of Google Colab, utilizing the T4 GPU provided.