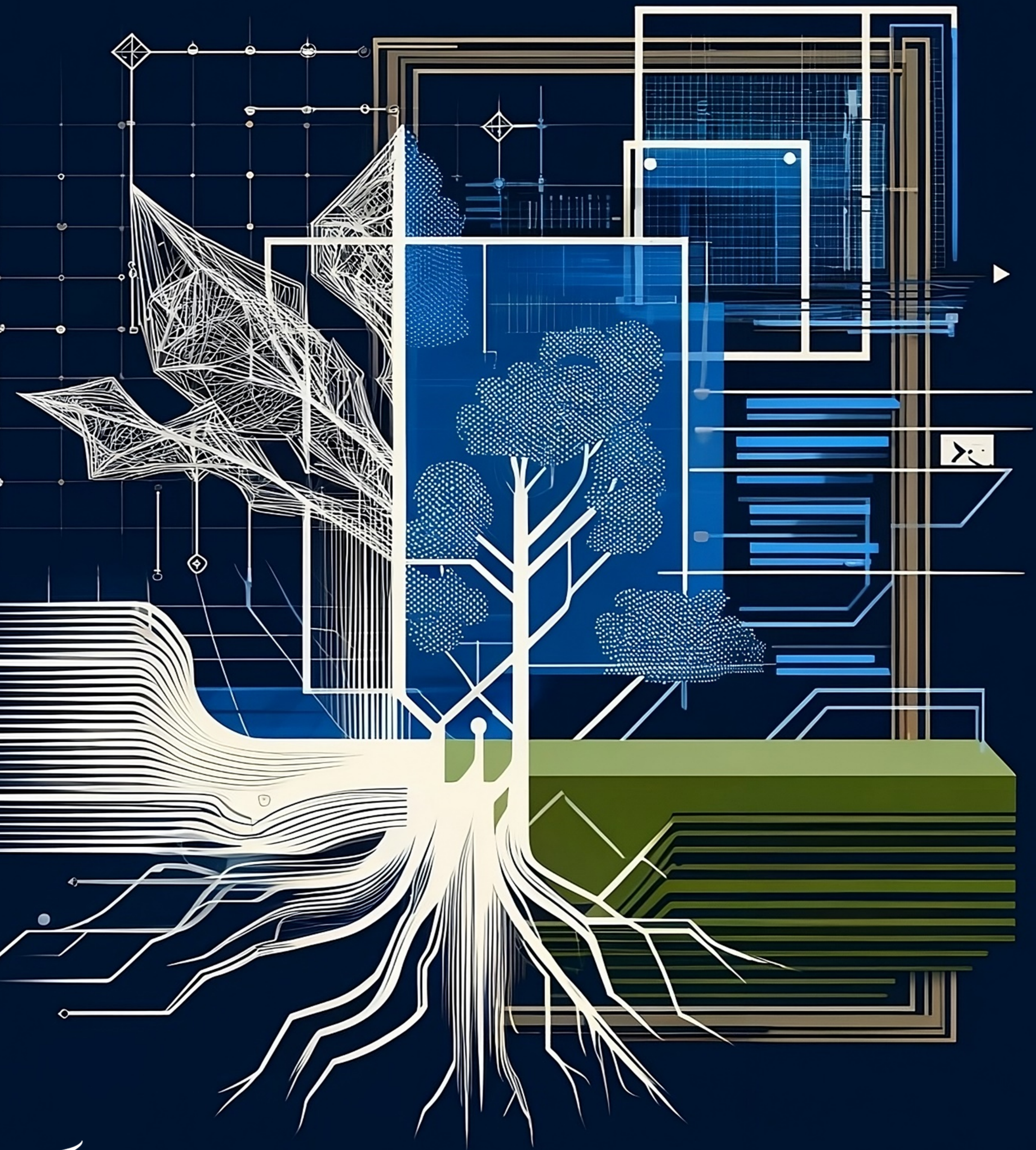
Corporate Sustainability Analysis via Retrieval-Augmented Generation

Joost Jansen



Corporate Sustainability Analysis via Retrieval-Augmented Generation

Joost Jansen
Delft University of Technology | MN
j.j.jansen-2@student.tudelft.nl

Supervisors:

Shubhalaxmi Mukherjee
Delft University of Technology
S.Mukherjee-2@tudelft.nl

Pradeep Murukannaiah
Delft University of Technology
P.K.Murukannaiah@tudelft.nl

Abstract

We investigate the application of Retrieval-Augmented Generation (RAG) for enhancing the analysis of corporate sustainability disclosures. We introduce CorSus, a novel dataset for evaluating RAG models in verifying corporate sustainability-related claims, using data from the Transition Pathway Initiative for over 100 companies. We further develop a subset of this dataset with reference documentation and fully explained answers. Finally, in a systematic framework, we optimise and benchmark state-of-the-art RAG approaches using the CorSus dataset. With this work, we aim to empower stakeholders with a tool for informed evaluations of corporate sustainability practices, thereby encouraging a greater commitment to environmental responsibility.

1 Introduction

Given the urgency of climate change, it is crucial to ensure that companies are transparent about their sustainability practices. This transparency enables stakeholders, including investors, policymakers and the general public, to hold companies accountable for their impact on the environment.

Unfortunately, most large corporations provide lengthy sustainability reports that are difficult to digest for the majority of stakeholders. In addition, as of January 2024, EU legislation requires all companies, except listed micro-enterprises, to report on social and environmental risks (European Commission, 2023), leading to a significant rise in sustainability reports. Thus, stakeholders largely rely on third-party rating agencies to analyse such reports. However, the services of rating agencies can be expensive, lack transparency, and vary due to differing sustainability evaluation criteria (Berg et al., 2022). Therefore, the automation of analysing sustainability reports is a key problem for improving accessibility, efficiency and transparency of companies' sustainability performance.

The analysis of a sustainability report can effectively be framed as a claim-verification problem, where a number of claims made by a company need to be evaluated based on that company's provided documents. This type of analysis is increasingly being automated through innovative techniques (Katranidis and Barany, 2024; Thorne et al., 2018; Diggelmann et al., 2020; Chen et al., 2022, 2023).

A claim-verification system consists of two key processes: evidence retrieval and veracity determination. The goal of evidence retrieval phase is to accurately identify and gather relevant evidence. Following this, the veracity determination phase assesses the accuracy of the claim by examining its relationship with the retrieved evidence.

Retrieval Augmented Generation (RAG) (Lewis et al., 2020) is a promising method for claim-verification. RAG integrates external knowledge directly into a pre-trained Large Language Model (LLM). This integration allows the LLM to access up-to-date, verifiable, domain-specific knowledge, which is crucial for accurately assessing claims. Due to the nature of incorporating external knowledge no further training is needed on a specific domain such as corporate sustainability and data can be easily updated without retraining.

ChatReport (Ni et al., 2023) incorporates a RAG design to automatically analyse sustainability reports based on the Task Force on Climate-related Financial Disclosures (TCFD) questions. It produces the answers, sources, and conformity scores of the report on how well it answers each question. Although this method claims a hallucination-free rate of 83% and an accuracy of 75%, these metrics are hard to verify. As answers need to be manually evaluated by an expert within the corporate sustainability domain, this becomes an expensive and time-consuming task.

To bridge this gap, we produce CorSus, a corporate sustainability benchmark, encompassing 100 companies, utilising the Transition Pathway Ini-

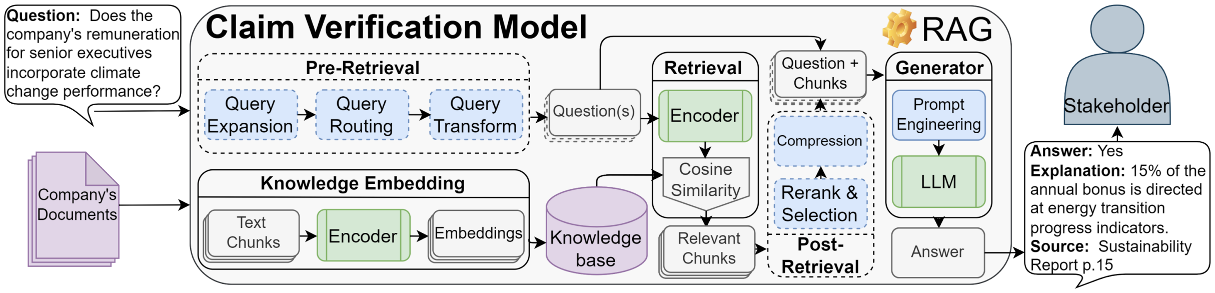


Figure 1: Overview of a claim verification model utilising RAG to address a question from the CorSus dataset. The Pre-Retrieval and Post-Retrieval stages are optional additions to the In-Context RAG architecture. Green blocks indicate elements that can be fine-tuned, while blue blocks denote modules where various methods can be applied to determine the optimal approach. Purple blocks represent the CorSus dataset.

tiative (TPI) assessments of these companies and manually collecting their disclosures from 2020 to 2022. CorSus includes two parts: the quantitative part enables the assessment of RAG models' ability to verify sustainability claims in financial disclosures, and the qualitative part to examine the documentation reference and the explainability of these claims. Additionally, we develop a framework to enhance RAG models for domain-specific claim verification. By integrating this framework with our dataset, we evaluate the ChatReport method and explore extensions to optimise the architecture for analysing disclosures on climate-related issues.

Contributions Our contribution is three fold.

- The CorSus benchmark with (1) a quantitative part including 8 claims with binary answers, utilising disclosures from 100 companies across three years, yielding 2400 labelled data points; and (2) a qualitative part comprising 8 claims along with fully explained answers and references from the provided documents of 12 companies for the year 2022, producing 96 labelled data points.
- A systematic framework designed to optimise and evaluate domain-specific RAG models.
- A comprehensive assessment and enhancement of the ChatReport method, resulting in an improved version named ChatReport+.

2 Related Work

We review related works on claim verification benchmarks and RAG.

2.1 Claim Verification Benchmarks

A multitude of claim/fact-verification benchmarks rely on crowd-authored claims from Wikipedia (Diggelmann et al., 2020; Thorne et al., 2018; Ma et al., 2023a). These benchmarks are mainly used to asses and improve NLP models on how well they verify the facts without the use of a knowledge base. Given a claim and context, the answer of the model is to be Supported, Refuted or Not Enough Information. A common limitation of these benchmarks is that there is no explanation for the answer provided. As displaying the veracity prediction along with the corresponding textual explanations can make the fact-checking system more credible to human users (Atanasova, 2024), EX-FEVER (Ma et al., 2023a) and E-FEVER (Stammbach and Ash, 2020) extend FEVER to include explanations, and showed how RAG can be used to provide explanations.

A number of domain-specific claim-verification benchmarks exists, e.g., for health (Kotonya and Toni, 2020) and for political claims (Chen et al., 2023). A closely related benchmark to ours is Climate-FEVER (Diggelmann et al., 2020), which consists of climate claims based on Wikipedia. While similar in structure, our benchmark differs by focusing specifically on corporate sustainability.

2.2 Retrieval Augmented Generation

Finetuned-RAG: The first work of integrating knowledge bases into generative LLMs was achieved with the Retrieval Augmented Generation (RAG) model (Lewis et al., 2020). This model features an end-to-end trainable architecture where both an encoder and a generative LLM are refined through backpropagation, enabling the retrieval and integration of relevant knowledge as a latent variable to enhance precision and truthfulness. After RAG more methods focusing on finetuning the retriever or the generator began arising(Asai et al., 2024; Borgeaud et al., 2022; Guu et al., 2020; Karpukhin et al., 2020).

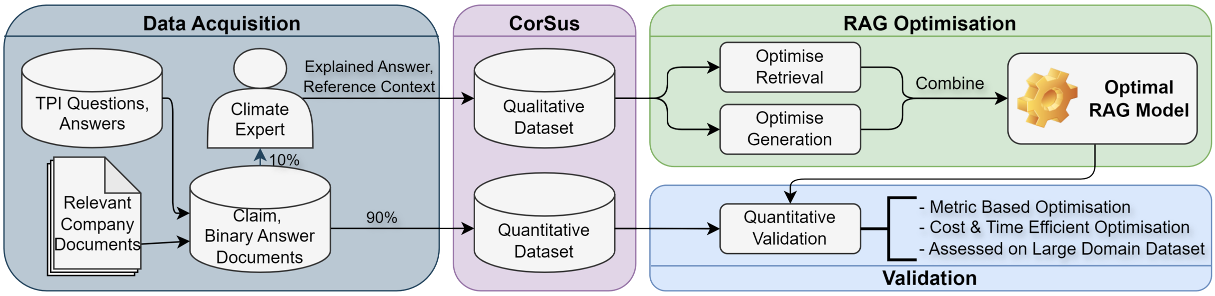


Figure 2: Our optimisation framework for Corporate Sustainability Claim Verification using RAG, integrating the Transition Pathway Initaitve (TPI) Questions and Answers with company documents to produces CorSus. CorSus is then used to optimise and validate RAG models.

In-Context-RAG: Subsequent developments of releases of well pre-trained models like OpenAI's chatGPT-3 (Floridi and Chiriatti, 2020) and Google's T5 (Raffel et al., 2020) led to the In-Context RAG (Ram et al., 2023) approach, which leverages pretrained models and encoders without further combined finetuning to achieve performance comparable to more extensively finetuned RAG models. In-Context RAG involves three phases: knowledge embedding, retrieval, and generation. In the embedding phase, unstructured knowledge is segmented, encoded, and stored with embeddings in a knowledge base. In the retrieval phase, the posed question is encoded, and similar chunks are retrieved using cosine similarity. These chunks and the question are then passed to a generative model to produce an answer.

Advanced-RAG: Recent advancements in In-Context RAG, hereafter called RAG, can be seen as modular improvements as illustrated by Gao et al. (2024). These advanced RAG methods can be divided into the pre- or post-retrieval stage of the RAG pipeline, as illustrated in Figure 1. In the pre-retrieval phase, research has been done on query expansion (Zhou et al., 2023; Dhuliawala et al., 2024), query routing (Singh et al., 2021) and query transformation (Gao et al., 2023; Ma et al., 2023b). In the post-retrieval phase, research has been done on reranking & selection (Liu et al., 2024; Cohere, 2023; Xiao et al., 2023; Liu, 2023) and compression (Chen et al., 2023; Yang et al., 2023; Xu et al., 2024) of retrieved documents. Aside from pre- and post-retrieval improvements, the flow of the whole RAG pipeline can also be adjusted by retrieving documentation at multiple stages through recursive (Trivedi et al., 2023), iterative (Shao et al., 2023) or adaptive (Jiang et al., 2023; Asai et al., 2024) retrieval. At last, domain-specific prompt engineering helps guide the generative model in interpreting the retrieved text and query. The authors of ChatReport developed a prompt that instructs the model to respond as a senior equity analyst with climate science expertise (Ni et al., 2023). Although they evaluate their method, they do not compare their method against a baseline, resulting in an incomplete assessment of its relative effectiveness. A more extensive background on RAG and explanation of these advanced methods is provided in Appendix A.

Combination-RAG: Despite growing interest in RAG techniques, the literature is primarily dominated by systematic reviews (Gao et al., 2024; Li et al., 2022) and pairwise comparisons of recent methods (Gao et al., 2023; Dhuliawala et al., 2024; Shao et al., 2023; Ma et al., 2023b). These studies often overlook potential synergies between different RAG techniques, revealing a gap in comprehensive experimental analyses of advanced RAG methods. Our study addresses this gap by evaluating multiple RAG techniques and their combinations. Specifically, we focus on various retrieval and generative models, incorporating pre- and post-retrieval methods, and the prompt-engineered method from ChatReport, providing insights into their effectiveness and real-world applicability. Additional related-work on climate-focused benchmarks and natural language processing can be found in Appendix B.

3 Methodology

To improve the RAG method within corporate sustainability, we have devised a framework that consists of data acquisition, optimisation, and evaluation components, as depicted in Figure 2. Our framework introduces a streamlined RAG optimisation using both qualitative and quantitative assessments. While this paper focuses on applying our approach to corporate sustainability, the method can be equivalently applied to other domains of claim verification tasks using RAG.

3.1 Data Acquisition

In this section, we provide a detailed description of the specific steps involved in manufacturing our dataset CorSus.

3.1.1 Claim Collection

To develop our dataset we use the 4th version of the Management Quality Evaluation of the Transition Pathway Initiative (TPI)[1]. This binary dataset consists of an evaluation of 1051 companies on 19sustainability-related questions in the years 2021 to 2023. TPI, developed by an international group of asset owners, assesses companies based on their management of greenhouse gas emissions and alignment with the UN Paris Climate Agreement goals. To address the imbalance in the TPI dataset and achieve a balanced distribution, we selected 8 out of the 19 questions where the proportion of "Yes" to "No" answers was within 85% for each question. This selection resulted in 64.9% of "Yes" answers in our dataset. Aside of a question, TPI also provides an assessment to this question. We'll be viewing the question with the assessment as one claim, due to the underlying complexity of certain questions. Appendix F provives more information on the TPI questions, assessments and claims.

3.1.2 Document Collection

For our knowledge base, we utilised documents from company websites made available one year before their TPI assessment. Due to inconsistent website layouts and document formats, automated scraping was unfeasible, requiring manual downloads. This limitation restricted our study to 100 companies from 2021 to 2023, all listed in the Climate Action 100+[2] list, which includes the world's largest corporate greenhouse gas emitters. These companies were chosen for their significant relevance in climate change action.

3.1.3 CorSus-Qualitative

To improve the CorSus dataset's utility and interpretability, we added a qualitative subset. This is crucial for enhancing the retrieval and generational phases, detailed in Section 3.2. In addition to the binary responses (Yes/No) to the sustainability questions, the qualitative subset includes detailed explanations and referenced text excerpts from company documents selected by sustainability experts. These additions allow for comprehensive comparison of the model's explainability and optimisation of the RAG stages. We selected companies with similar distribution and sectors as the quantitative dataset to ensure representativeness. Appendix F.2 provides detailed examples and construction methods of the qualitative dataset.

CorSus	Quantitative	Qualitative
Number of companies	100	12
Years	2020, 2021, 2022	2022
Number of questions	8	8
Number of documents	782	48
Percentage "Yes" answers	64.9%	64.6%
Total datapoints	2400	96

Table 1: Overview of attributes in the Quantitative and Qualitative subset of the CorSus dataset.

3.2 RAG Optimisation

We employ metric-driven insights to separately refine the retrieval and generation phases. This approach enhances the model's accuracy and reliability by incorporating expert-verified data, helping to identify and address its limitations. A more detailed background and formulation of the various RAG techniques employed are provided in Appendix A.

3.2.1 Retrieval Optimisation

To optimise retrieval, we utilise the references linked to each claim in our qualitative dataset to evaluate the performance of the retriever, including enhancements made during the pre-retrieval phase. The effectiveness of the retrieval process is measured using F1@k, where @k represents the number of document chunks retrieved.

3.2.2 Generational Optimisation

To determine the optimal configuration for the generational model with post-retrieval and prompt-engineering techniques, we use a fixed set of information from CorSus-qualitative reference documentation, categorised into four types:

Only relevant: Uses references entirely relevant to the claim, discarding any irrelevant data.
Partially relevant: Provides half of the relevant information, simulating incomplete data condition.

[1] www.transitionpathwayinitiative.org

[2] www.climateaction100.org/whos-involved/companies

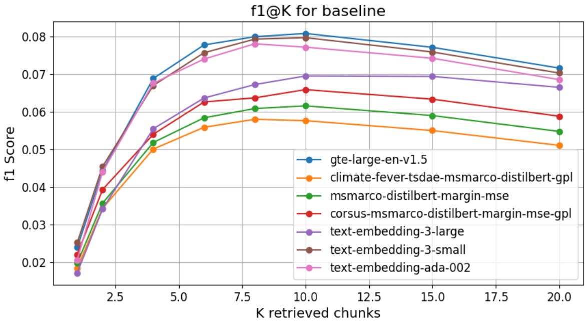


Figure 3: The F1 scores of different retrieval models across a range of k chunks retrieved.

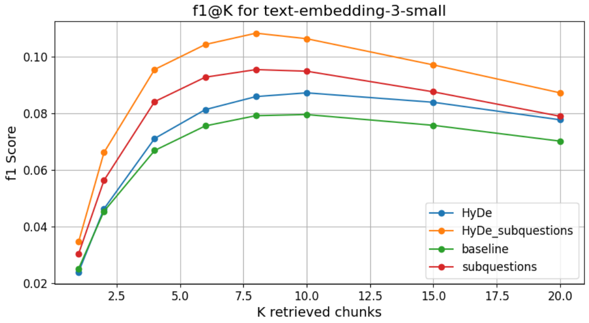


Figure 4: The F1 scores of different retrieval models across a range of k chunks retrieved.

No relevancy: Includes irrelevant information from the same company, testing the model's ability to ignore misleading information and abstain from answering "Yes".
Noise: Contains irrelevant information from a non-company noise document, testing the model's ability to ignore non-pertinent information.

The generational model is assessed using accuracy and balanced accuracy. Balanced accuracy, the mean of specificity and sensitivity, ensures equal weighting for "No" and "Yes" answers in our imbalanced dataset. Additionally, we analyse the confusion matrix and compare the generated explanations with those in our qualitative dataset to measure the explainability's correctness.

3.3 Quantitative Validation

After finding the optimal combination of retrieval and generational models and methods separately, it is essential to evaluate their overall effectiveness by conducting an accuracy assessment on the entire system. Since only a limited subset of data is used for optimisation, testing the model's performance on a broader quantitative dataset ensures generalisability and effectiveness beyond the initial set. This involves comparing the optimised model's performance against a baseline, providing a clear measure of the enhancements achieved. As a result, the model's improvement and validation are guided by metric-driven insights, ensuring cost and time-efficient optimisation. We will use the same metrics as for generational optimisation, excluding correctness of explanations due to the absence of explanations in the CorSus-Quantitative dataset.

4 Experiments: CorSus-Qualitative

Using our CorSus-Qualitative model, we separately test our retrieval and generational models, as outlined in Section 3. We evaluate various existing models and methods to find the optimal solution for corporate sustainability analysis using our dataset.

4.1 Setup

We use the LLamaIndex library with a PostgreSQL database for vector storage and AWS S3 for disclosure management. Following ChatReport (Ni et al., 2023), we split documents into 500-character chunks with a 20-character overlap and use a temperature of 0 for our generational models.

4.2 Retrieval Optimisation

This section evaluates retrieval performance across various models and methods to identify the most effective strategies for optimising retrieval accuracy.

4.2.1 Model Selection

Building on GPL creators' work emphasising the refinement of text similarity tools within specific subdomains (Wang et al., 2022), we compare retrieval models. We evaluate the baseline *msmarco-distilbert-margin-mse* model against the *climate-fever-tsdae-msmarco-distilbert-gpl* model, specifically tuned for climate-related discourse in the Climate-FEVER dataset (Diggelmann et al., 2020). This comparison is relevant given the thematic overlap with our dataset, as Climate-FEVER focuses on climate-related claims from Wikipedia. Additionally, we introduce *corsus-msmarco-distilbert-margin-mse-gpl*, further finetuned on our CorSus-quantitative dataset using the GPL method. Further finetuning details are in Appendix D. We also examine whether State-Of-The-Art (SOTA) retrieval models provide significant performance improvements. We use Alibaba's white-box model *gte-large-en-v1.5*, the top-performing model on the MTEB leaderboard[3] for retrieval tasks (Muen-

[3] huggingface.co/mteb, accessed on 02-04-2024

Information	Prompt	Reranker	Accuracy	Balanced Accuracy	TP	FP	TN	FN
Noise	Baseline	-	1.000	1.000	-	0	8	-
	ChatReport	-	0.625	0.625	-	3	5	-
No Relevancy	Baseline	-	0.344	0.344	-	63	33	-
	ChatReport	-	0.667	0.667	-	32	64	-
Partially Relevant	Baseline	LostInTheMiddleReranker	0.573	0.435	53	28	2	13
		-	0.635	0.508	56	25	5	10
	ChatReport	LostInTheMiddleReranker	0.604	0.558	45	17	13	21
		-	0.615	0.583	44	15	15	22
Only Relevant	Baseline	-	0.688	0.518	64	28	2	2
	ChatReport	-	0.740	0.656	58	17	13	8

Table 2: Comparison of the baseline prompt (Appendix C.1) and the ChatReport prompt (Appendix C.2) with and without the LostInTheMiddleReranker (Liu et al., 2024) in assessing generation quality using the *gpt-3.5-turbo-0125* model under varying levels of information. Metrics include Accuracy, Balanced Accuracy, True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). Underlined values highlight the highest Accuracy, TP and FN, and the lowest FP and FN per information type.

nighoff et al., 2023), and compare it with OpenAI's black-box text-embedding models *ada-002*, *3-small*, and *3-large*. The latter two are newer versions of the *ada-002* model used in ChatReport (Ni et al., 2023). This evaluation aims to determine the efficacy of SOTA technology and finetuned models in enhancing retrieval accuracy within our dataset. The results of different models are presented in Figure 3.

4.3 Method Selection

We explore various pre-retrieval methods, such as the Subquestions method (Chen et al., 2022), the HyDE method (Gao et al., 2023), and a combination of both. These methods improve retrieval by generating subquestions or comparable text related to the retrieval passage. Additional details on these methods are provided in Appendix A. As a baseline retrieval method, we use the original question to retrieve the most relevant information, similar to the approach used in ChatReport. The performance of these methods is shown using various retrieval models, with results for the *text-embedding-3-small* model displayed in Figure 4. Similar graphs for other models can be found in Appendix E.

4.3.1 Retrieval Models & Methods Matter

Models: Figure 3 compares the performance of various retrieval models using the baseline method. The *gte-large-en-v1.5*, OpenAI's text-embedding *ada-002*, and *3-small* models show the highest performance. Interestingly, the *text-embedding-3-large* model, despite its larger dimension size, underperforms compared to the *3-small* model, indicating poorer generalisation in climate data. Performance varies slightly by method, highlighting the importance of combining state-of-the-art models with effective methods. The climate-tuned model underperforms relative to its base model, suggesting domain differences between Climate-FEVER and corporate sustainability. The finetuned *corsus-msmarco-distilbert-margin-mse-gpl* model surpasses its baseline but falls short of SOTA models, underscoring the critical role of advanced models in optimising retrieval performance.

Methods: Figure 4 presents F1 scores for various methods at different K chunks retrieved. The key observations are that the HyDE/Subquestions method consistently outperforms others across all K values for all models. The Subquestions method alone also shows strong performance, highlighting the value of breaking down queries. In contrast, the baseline method performs the worst, demonstrating significant improvements from advanced methods like HyDE and Subquestions. Performance peaks around $K = 8$ to $K = 10$ chunks, suggesting an optimal retrieval chunk range for our dataset. Comparisons of all results in Appendix E show that the best combination is the *text-embedding-3-small* model with the HyDE/Subquestions method.

4.4 Generational Optimisation

As depicted in our framework, we determine the optimal generation settings under full, partial, no relevant information, or noise conditions. We compare the performance of the ChatReport engineered prompt against a baseline prompt, as detailed in Appendix C. We also evaluate the *LostInTheMiddleReranker* method (Liu et al., 2024), which improves generational accuracy by putting the most

Generational Models	Prompt	Accuracy	Balanced Accuracy	Correct Explanation	TP	FP	TN	FN
gpt-3.5-turbo-0125	Baseline	0.688	0.518	0.667	64	28	2	2
gpt-3.5-turbo-0125	ChatReport	0.74	0.656	0.74	58	17	13	8
gpt-4o-2024-05-13	Baseline	0.667	0.539	0.667	58	24	6	8
gpt-4o-2024-05-13	ChatReport	0.771	0.706	0.771	58	14	16	8
llama3:70b	Baseline	0.656	0.568	0.625	53	20	10	13
llama3:70b	ChatReport	0.695	0.626	0.674	53	16	13	13
llama3:8b	Baseline	0.646	0.515	0.635	57	25	5	9
llama3:8b	ChatReport	0.677	0.629	0.646	50	15	15	16
Finetuned Models	**Prompt**	**Accuracy**	**Balanced Accuracy**	**Correct Explanation**	**TP**	**FP**	**TN**	**FN**
llama2:7b	Baseline	0.688	0.5	0.531	66	30	0	0
llama2:7b	ChatReport	0.688	0.5	0.656	66	30	0	0
adapt-finance-llama2-7b	Baseline	0.604	0.512	0.167	50	22	8	16
adapt-finance-llama2-7b	ChatReport	0.646	0.533	0.50	55	23	7	11
eci-io-climategpt-7b	Baseline	0.649	0.536	0.299	44	16	6	11
eci-io-climategpt-7b	ChatReport	0.632	0.524	0.411	53	23	7	12

Table 3: **Relevant information:** Comparison of Baseline and *ChatReport* prompts using only relevant information for different models. Metrics include Accuracy, Balanced accuracy, Percentage of correctly explained answers, True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). Underlined values highlight the highest Accuracy, TP and FN, and the lowest FP and FN for generational and finetuned models separately.

relevant information at the outer edges instead of chronologically. For "Noise" and "No Relevancy" conditions, true positives and false negatives are omitted as the correct information is never provided. , focusing instead on false positives and true negatives. We use the SEC-10K[4] example as a noise document, asking the 8 questions only once due to the identical underlying document. The SEC-10K filings contains financial and operational information that is irrelevant to sustainability claims. We compare SOTA white-box models *llama3:8b-instruct* and *llama3:70b-instruct* (Meta, 2024) with black-box models *gpt-3.5-turbo-0125* (Floridi and Chiriatti, 2020) and *gpt-4o-2024-05-125* (OpenAI, 2024) . Additionally, we assess how further fine-tuned models on finance [*adapt-finance* (Cheng et al., 2024)] and climate [*climategpt* (Thulke et al., 2024)] data perform compared to their base model *llama2:7B-chat* (Touvron et al., 2023). Due to time and budget constraints, the further fine tuning of generational models on the CorSus dataset is left for future research. The results are presented in Table 2 and Table 3, with additional results in Appendix E.

4.4.1 Information Relevancy & Reranking

Table 2 shows that the accuracy of generational outputs varies significantly based on the relevance of the provided information. When only relevant information is given, the accuracy is 0.740, indicating a 26% failure rate even with optimal information. Accuracy drops further to 0.635 with only partially relevant information. In the "No Relevancy" scenario, the generative model often answers "Yes" to irrelevant company information, showing a bias in affirming when given irrelevant sustainable information. However, the ChatReport prompt reduces this error, making the model more critical. With pure noise, the ChatReport prompt increases errors, indicating that the model is confused by sustainble-related prompting when no sustainable information provided. These findings underscore the importance of relevant information and prompt optimisation for generative models. In the case of LostInTheMiddleReranker, accuracy drops with both the baseline and ChatReport prompt, suggesting that placing the most relevant information at the beginning of a prompt is more beneficial than at the outer edges, contrary to previous work.

4.4.2 Generational Model Selection

Table 3, shows the performance of various generative models using only relevant information. The ChatReport prompt consistently outperformed the baseline prompt, achieving higher accuracy, balanced accuracy, and correct explanation rates across models. The *gpt-4o* model with the ChatReport prompt had the best overall performance. Despite its smaller size, *llama3:8b* performed comparably to larger models like *llama3:70b* and *gpt3.5*. Fine-tuned models showed mixed results, with slight improvements over the base model. The *llama2:7b* base model failed to answer "No" to any question but had a higher explanation rate. However, fine-tuned versions of *llama2:7b* improved in balanced accuracy but had lower correct explanation rates. Both *llama2:7b* and its fine-tuned ver-

[4] www.sec.gov/files/form10-k.pdf

sions underperformed compared to newer models like *llama3:8b*, highlighting the need for advanced training rather than merely increasing model size. The fine-tuned models' low balanced accuracy and correct explanation rates indicate they struggle with these questions, emphasising the necessity for more advanced models. Overall, these results emphasise the importance of prompt engineering and model selection in improving generative models for corporate sustainability analysis.

5 Validation: CorSus-Quantitative

In previous experiments, we assessed different models and methods for the retrieval and generation components of a RAG pipeline. Here, we compare the combined optimal method against a baseline and the ChatReport method.

5.1 Baseline

We adopt the original ChatReport configuration, utilising OpenAI's *text-embedding-ada-002* for retrieval and *gpt-3.5-turbo-0125* for generation with $top_k = 20$, termed ChatReport3. We use the same configuration for the Baseline RAG but with the baseline prompt, as depicted in Appendix C.1.

5.2 Optimal Model

Based on our results in Section 4, we compare the Baseline RAG and ChatReport methods with the optimised models ChatReport3+ and ChatReport4+. ChatReport3+ extends ChatReport3 with the HyDe/Subquestion method and a $top_k = 10$. ChatReport4+ employs the *text-embedding-3-small* model for retrieval with the HyDe-Subquestion method and the *gpt-4o-2024-05-13* model with the ChatReport prompt for generation. Table 4 displays all the results.

5.3 Qualitative Optimisation Boosts Quantitative Assessment

Table 4 shows that improved versions ChatReport+ significantly increases accuracy compared to the baseline and the original ChatReport. The baseline architecture achieved 0.589 accuracy, with the highest true positives (1351) but also the highest false positives (923) and lowest true negatives (64), indicating a bias toward positives. The ChatReport3 model improved accuracy to 0.617 by reducing false positives to 697 and increasing true negatives to 286, though true positives slightly dropped to 1195. This setup better balances true positives and false positives. ChatReport3+ further improved accuracy to 0.643, increasing true positives to 1276 and reducing false negatives to 139, indicating better information retrieval and overall performance. The best performance was from ChatReport4+, with an accuracy of 0.693 and balanced accuracy of 0.660. This model significantly reduced false positives to 514 and increased true negatives to 470, demonstrating superior capability in identifying false claims. The increase in false negatives to 222 indicates more caution in accepting information.

Experiment	Accuracy	Balanced	TP	FP	TN	FN
Baseline	0.589	0.509	1351	923	62	64
ChatReport3	0.617*	0.567	1195	697	286	222
ChatReport3+	0.643*†	0.587	1276	717	268	139
ChatReport4+	0.693*†	0.660	1194	514	470	222

Table 4: Performance metrics for Baseline RAG, ChatReport3 and the improved ChatReport+ designs on the CorSus-Quantitative dataset. Asterisks (*) and the cross (†) indicate a significant increase in accuracy based on the McNemar test ($p < 0.05$) compared to Baseline RAG and ChatReport, respectively. Metrics include Accuracy, Balanced Accuracy (Balanced), True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). Underlined values indicate the highest Accuracy, TP, and FN, and the lowest FP and FN.

6 Conclusion

This research has demonstrated the potential of Retrieval-Augmented Generation (RAG) models to enhance the analysis of corporate sustainability disclosures. By creating the CorSus-Quantitative and CorSus-Qualitative datasets, this study provided a robust framework for evaluating and optimising RAG models within the domain of corporate sustainability. The results indicate that the improved ChatReport+ models outperform the baseline and the original ChatReport prompt method, achieving the highest accuracy with 69.3%. This signifies the importance of an optimised retrieval process in reducing false negatives and the need for better generational models for enhancing the overall model performance. Despite these improvements, the optimal model fails to provide the correct answer one-third of the time, indicating that RAG models cannot yet be used reliably without human validation. This study lays the groundwork for more accurate, efficient, and transparent analysis tools, aiding stakeholders in making informed decisions about corporate sustainability practices.

Limitations & Future Work

Subjectivity in Corporate Sustianability: During the qualitative assessment, sustainability experts reviewed the same documents as the Transition Pathway Initiative (TPI) but often reached different conclusions. Out of 96 datapoints, they disagreed on 16. This discrepancy arises from the subjective nature of some sustainability questions, leading to varied interpretations. Questions requiring judgment calls showed the highest divergence, reflecting personal or collective values. We adjusted the TPI dataset to include these expert insights for the qualitative subset. Subjectivity in responses poses a challenge for RAG models, as there is often no single "correct" answer. This variability complicates the model's ability to generate appropriate responses and highlights a fundamental limitation in handling subjective questions.

Diversity of Questions & Companies: Although the questions selected from TPI are relevant to corporate sustainability, we devised only eight different questions over various companies and years. This lack of question diversity might lead to overfitting to specific questions rather than the general domain of corporate sustainability. Apart from the questions, the selected companies were all part of the Climate Action 100+ group, comprising the largest greenhouse gas emitters. Despite being from different sectors, the size and nature of these companies may lead to reporting formats that differ from less polluting companies or those in other industries. Nonetheless, the CorSus benchmark highlights the difficulty of these questions and underscores the need for further research in this area.

Models & Method Selection: We limited our research on a selection of models and methods from current literature to demonstrate potential improvements without extensive qualitative evaluation. Future research could explore synergies with techniques like iterative retrieval, compression, and further reranking.

Use of Blackbox Models: For this research, we utilised blackbox models from OpenAI due to their strong performance, time efficiency, and cost-effectiveness. However, these models do not offer true transparency into their underlying mechanisms. Additionally, updates to these models can lead to variations in experimental results. Future research should consider solely using whitebox models, which provide greater insight and consistency, to enhance transparency and reproducibility.

False Evidence and Answers: Despite optimisation, our ChatReport+ model fails to answer correctly one-third of the time, making it essential for users to exercise caution. We therefore advocate for a human-in-the-loop approach, where generated answers are cross-checked with sources by a human annotator to ensure reliability. Future research could look into the performance of these models with such a system.

Reliance on Company-provided Data: The CorSus dataset is designed to evaluate the performance of RAG models against corporate sustainability reports, which often reflect the firm's perspective and may include greenwashing. The current approach may lack objectivity due to reliance on company-provided data. Future work could incorporate external perspectives and additional sources to reduce bias and enhance accuracy.

Ethical Considerations

Impact on Rating Agencies: While the main goal of this research is to improve the accessibility and accuracy of sustainability disclosures for stakeholders, automating the analysis of these reports may have unintended consequences. It could lead to job displacement in sectors that rely on manual analysis, such as ESG rating agencies. However, this shift from fact verification and summarisation could enable rating agencies to focus more on critical assessments and providing deeper insights, enhancing the overall evaluation of information disclosed by companies.

Environmental Impact: The increasing use of AI technologies requires substantial computational resources, which can have a negative impact on the climate, especially if non-renewable energy sources are used. This research acknowledges the environmental footprint of running large-scale AI models and emphasises the need for optimising computational processes and utilising green energy whenever possible to minimise adverse environmental effects.

Acknowledgements

We express our sincere gratitude to MN for their computational resources. Special thanks to Michael Kurz and Xander Urbach for their insights and guidance in corporate sustainability.

The cover page image was generated using OpenAI's ChatGPT-4o and is © 2024 OpenAI.

Disclaimer

The contents of this research are entirely the responsibility of the authors and do not necessarily reflect the views or policies of MN.

No part of the contribution that may have been provided by MN should be construed as an official position or policy statement of the organisation. Information provided by MN employees is intended solely to support academic research and should not be construed as a binding statement of the organisation. All statements made by MN personnel in the context of this research have been made solely to support the authors in conducting academic research. These statements do not represent the formal opinion or official position of MN.

MN accepts no responsibility for the accuracy, completeness or use of the information as contained in this research. Any interpretations, findings, conclusions or recommendations presented in this research are solely those of the authors and do not imply any endorsement by MN.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Pepa Atanasova. 2024. Generating fact checking explanations. In *Accountable and Explainable Methods for Complex Reasoning over Text*, pages 83–103. Springer.

Florian Berg, Julian F Koelbel, and Roberto Rigobon. 2022. Aggregate confusion: The divergence of esg ratings. *Review of Finance*, 26(6):1315–1344.

Julia Anna Bingler, Mathias Kraus, Markus Leippold, and Nicolas Webersinke. 2022. Cheap talk and cherry-picking: What climatebert has to say on corporate climate risk disclosures. *Finance Research Letters*, 47:102776.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2023. Complex claim verification with evidence retrieved in the wild. *arXiv preprint arXiv:2305.11859*.

Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022. Generating literal and implied subquestions to fact-check complex claims. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3495–3516, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Daixuan Cheng, Shaohan Huang, and Furu Wei. 2024. Adapting large language models via reading comprehension. In *The Twelfth International Conference on Learning Representations*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sashank Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2024. Palm: scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24(1).

Cohere. 2023. Say goodbye to irrelevant search results: Cohere rerank is here. `https://txt.cohere.com/rerank/`. Accessed: 02-02-2024.

Shehzaad Zuzar Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason E Weston. 2024. Chain-of-verification reduces hallucination in large language models.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. In *In NeurIPS 2020 workshop Tackling Climate Change with Machine Learning*.

Matouš Eibich, Shivay Nagpal, and Alexander Fred-Ojala. 2024. Aragog: Advanced rag output grading. *arXiv preprint arXiv:2404.01037*.

Ekimetric. 2023. Climateqa. https://www.climateqa.com/. Accessed: 2023-12-01.

European Commission. 2023. The European Green Deal. https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal_en. Accessed: 2023-12-01.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, Toronto, Canada. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Vasileios Katranidis and Gabor Barany. 2024. Faaf: Facts as a function for the evaluation of rag systems. *arXiv preprint arXiv:2403.03888*.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Markus Leippold and Francesco Saverio Varini. 2020. Climatext: A dataset for climate change topic detection. In *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. 2022. A survey on retrieval-augmented text generation. *arXiv preprint arXiv:2202.01110*.

J. Liu. 2023. Using llms for retrieval and reranking. Accessed: 2024-04-20.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 11:157–173.

Huanhuan Ma, Weizhi Xu, Yifan Wei, Liuji Chen, Liang Wang, Qiang Liu, and Shu Wu. 2023a. Ex-fever: A dataset for multi-hop explainable fact verification. *arXiv preprint arXiv:2310.09754*.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023b. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5315, Singapore. Association for Computational Linguistics.

Srishti Mehra, Robert Louka, and Yixun Zhang. 2022. Esgbert: Language model to help with classification tasks related to companies environmental, social, and governance practices. *arXiv preprint arXiv:2203.16788*.

Meta. 2024. Llama 3 model card. Accessed: 2024-06-23.

Prakamya Mishra and Rohan Mittal. 2021. Neuralnere: Neural named entity relationship extraction for end-to-end climate change knowledge graph construction. In *ICML 2021 Workshop on Tackling Climate Change with Machine Learning*.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Jingwei Ni, Julia Bingler, Chiara Colesanti-Senni, Mathias Kraus, Glen Gostlow, Tobias Schimanski, Dominik Stammbach, Saeid Ashraf Vaghefi, Qian Wang, Nicolas Webersinke, Tobias Wekhof, Tingyu Yu, and Markus Leippold. 2023. CHATREPORT: Democratizing sustainability disclosure analysis through LLM-based tools. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 21–51, Singapore. Association for Computational Linguistics.

Tim Nugent, Nicole Stelea, and Jochen L. Leidner. 2021. Detecting environmental, social and governance (esg) topics using domain-specific language models and data augmentation. In *Flexible Query Answering Systems: 14th International Conference, FQAS 2021, Bratislava, Slovakia, September 19–24, 2021, Proceedings*, page 157–169, Berlin, Heidelberg. Springer-Verlag.

OpenAI. 2024. Gpt-4o: The comprehensive guide and explanation. https://blog.roboflow.com/gpt-4o-guide/. Accessed: 2024-06-23.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9248–9274, Singapore. Association for Computational Linguistics.

Devendra Singh, Siva Reddy, Will Hamilton, Chris Dyer, and Dani Yogatama. 2021. End-to-end training of multi-document reader and retriever for open-domain question answering. *Advances in Neural Information Processing Systems*, 34:25968–25981.

Daniel Spokoyny, Tanmay Laud, Tom Corringham, and Taylor Berg-Kirkpatrick. 2023. Towards answering climate questionnaires from unstructured climate reports.

Dominik Stammbach and Elliott Ash. 2020. e-fever: Explanations and summaries for automated fact checking. In *Proceedings of the 2020 Truth and Trust Online (TTO 2020)*, pages 32 – 43, Arlington, VA. Hacks Hackers. Conference for Truth and Trust Online (TTO 2020) (virtual); Conference Location: online; Conference Date: October 16-17, 2020; Due to the Coronavirus (COVID-19) the conference was conducted virtually.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

David Thulke, Yingbo Gao, Petrus Pelser, Rein Brune, Rricha Jalota, Floris Fok, Michael Ramos, Ian van Wyk, Abdallah Nasir, Hayden Goldstein, Taylor Tragemann, Katie Nguyen, Ariana Fowler, Andrew Stanco, Jon Gabriel, Jordan Taylor, Dean Moro, Evgenii Tsymbalov, Juliette de Waal, Evgeny Matusov, Mudar Yaghi, Mohammad Shihadah, Hermann Ney, Christian Dugast, Jonathan Dotan, and Daniel Erasmus. 2024. Climategpt: Towards ai synthesizing interdisciplinary research on climate change.

Jinfang Tian, Qian Cheng, Rui Xue, Yilong Han, and Yuli Shan. 2023. A dataset on corporate sustainability disclosure. *Scientific Data*, 10(1).

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.

Saeid Ashraf Vaghefi, Qian Wang, Veruska Muccione, Jingwei Ni, Mathias Kraus, Julia Bingler, Tobias Schimanski, Chiara Colesanti-Senni, Dominik Stammbach, Nicolas Webersinke, et al. 2023. Chatclimate: Grounding conversational ai in climate science.

Roopal Vaid, Kartikey Pant, and Manish Shrivastava. 2022. Towards fine-grained classification of climate change related social media text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 434–443, Dublin, Ireland. Association for Computational Linguistics.

Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. 2022. GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2345–2360, Seattle, United States. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2024. RECOMP: Improving retrieval-augmented LMs with context compression and selective augmentation. In *The Twelfth International Conference on Learning Representations*.

Haoyan Yang, Zhitao Li, Yong Zhang, Jianzong Wang, Ning Cheng, Ming Li, and Jing Xiao. 2023. PRCA: Fitting black-box large language models for retrieval question answering via pluggable reward-driven contextual adapter. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5364–5375, Singapore. Association for Computational Linguistics.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

A Background & Formulation

In this section, we introduce the formulation of claim verification and provide the necessary background on Retrieval Augmented Generation (RAG) relevant to our research.

A.1 Repeated Claim Verification

Considering a domain with C distinct claims c and O objects o for analysis. Each object o is evaluated against all C claims using the data D, where d_o represents the data corresponding to object o. In this setup, a fact-checking system named V generates a binary answer a and an explanation e for each claim-object pair as follows:

$$V(c_i, d_j) = a_{i,j}, e_{i,j} \quad \forall i \in C, j \in O \tag{1}$$

It is important to highlight that the same claims are verified across all objects, leading to repetitive evaluations. A response is considered accurate if it matches the true answer $a_{i,j} = a^T_{i,j}$ and the explanation is deemed correct if it conveys the same meaning, $e_{i,j} \approx e^T_{i,j}$, where T indicates the true values.

A.2 In-Context Retrieval Augmented Generation

The In-Context RAG system (Ram et al., 2023), as illustrated in Figure 1, operates through three primary stages, enhanced by additional pre- and post-retrieval phases. Initially, the Knowledge Embedding phase processes a corpus of unstructured documents D, segmented into chunks d_k where each $d_k \in d_o \in D$. These chunks are encoded by a model f, producing vector representations stored in a database. During the Retrieval phase, the system selects data chunks d_k for a given query q using the following criterion:

$$R(q, d_o) = \underset{d_k}{\operatorname{argmax}} \operatorname{sim}(f(q), f(d_k)) \quad \forall d_k \in d_o \tag{2}$$

Here, sim represents the cosine similarity between vector embeddings of the query and the chunk. The top K chunks with the highest similarity are retrieved and provided to the Generator G, which then generates a response based on the query augmented with the selected document chunks:

$$G(q, R(q, d_o)) = e_q \tag{3}$$

A.2.1 Adapting RAG for Claim Verification

To tailor the RAG framework for claim verification, we modify the generative model's prompt to produce responses in a structured JSON format, containing both a binary answer a (True or False) and a detailed explanation e of the reasoning behind the answer:

$$\begin{aligned} V_{RAG}(c_i, d_j) &= G(JSON(c_i, R(c_i, d_j))) \\ &= a_{i,j}, e_{i,j} \quad \forall i \in C, j \in O \end{aligned} \tag{4}$$

This format ensures easy system evaluation and provides a clear answer along with the reasoning, enhancing transparency and understandability. The example prompt used for generating these responses is provided in Appendix C.1. This approach establishes a baseline for subsequent comparisons.

A.3 Advanced RAG Modules

The basic RAG architecture supports modular extensions, as detailed by Gao et. al (2024). We will examine several RAG techniques corresponding to the various stages shown in Figure 1. Based on the literature, we will determine if these techniques offer potential improvements for enhancing claim verification in our corporate sustainability setting.

A.3.1 Pre-Retrieval

In the pre-retrieval stage the original query, in our case claim, can be transformed, expanded and directed to the right documents to retrieve the correct knowledge.

Query Routing: Although an RAG application can handle multiple tasks and sources, such as structured data, unstructured content, and APIs, a specific query may only be relevant to one use case. Thus, various techniques exist to identify the most pertinent source for a given query (Jiang et al., 2023; Dhuliawala et al., 2024). The query is then processed by another LLM, often referred to as an agent, specialised in that task. In our case, since all data is in PDF format, we limit our query routing to direct specific claims to the documents of the relevant company. We group all vector representations of chunks from the same company and year into a single knowledge base, avoiding further subdivisions of the information.

Query Expansion: Current literature indicates that generating sub-questions from the original claim is an effective strategy for verifying and retrieving relevant information (Chen et al., 2023, 2022). For a given query q, an LLM generates several subquestions q_s. These subquestions are then used in the retrieval process, with each subquestion retrieving $k/|q_s|$ chunks to maintain the same total number of retrieved chunks. As a result, this approach can retrieve diverse information. The prompt for subquestion generation is depicted in Appendix C.3.

Query Transformation: In addition to expanding the original query, it can also be transformed. The authors of HyDE generate a hypothetical document chunk based on the original query (Gao et al., 2023). This hypothetical document is then provided to the retriever to obtain k relevant chunks. This method, known as Hypothetical Document Embedding (HyDE), yields higher similarity to the relevant chunks, thus enhancing the overall retrieval performance (Gao et al., 2023). The prompt for HyDe generation is depicted in Appendix C.4.

A.4 Post-Retrieval

In the post-retrieval stage numerous work has been done in reranking, compression and selection.

Reranking & Selection: The sequence in which retrieved chunks are presented significantly impacts the generator's interpretation (Liu et al., 2024; Eibich et al., 2024). The LostInTheMiddleReranker, which prioritises important texts based on the retrieval score (cosine similarity) at the document's outer stages rather than in chronological order, has been demonstrated to enhance the generative capabilities of RAG models (Liu et al., 2024). In addition to chunk ordering, several reranking techniques filter retrieved chunks to a smaller subset (Cohere, 2023; Xiao et al., 2023; Liu, 2023). These methods, which are based on cross-sentence or generative LLMs, are too computationally expensive to apply to the entire dataset but perform well on smaller portions and are therefore applied as post-retrieval selection.

Compression: Due to the context limitations of LLMs, summarising or compressing retrieved chunks is common (Yang et al., 2023; Xu et al., 2024). However, we avoid these methods in corporate sustainability and fact-checking contexts, as they can generate untruthful content (Ouyang et al., 2022; Bommasani et al., 2021; Chowdhery et al., 2024), risking information loss.

A.4.1 Generator

Beyond simply appending the retrieved chunks to the query, additional prompt engineering offers the advantage of guiding the generative model on how to interpret the retrieved text and query. For instance, ChatReport's engineered prompt instructs the model to respond as a senior equity analyst with expertise in climate science (Ni et al., 2023), which aligns with our study's needs. Their prompt was developed through automatic prompt generation in collaboration with climate experts. Given the prompt's effectiveness in meeting our requirements, we have decided not to pursue further optimisation of different prompt construction methods. Instead, we will evaluate the performance of this established format in our experiments. The sole modification we implement is substituting the TCFD assessment guidelines with the TPI assessment guidelines, primarily to align our analysis more closely with TPI's comprehensive approach in evaluating corporate sustainability. All prompts used in our study are detailed in Appendix C.

B Additional Related Work

B.1 Climate Benchmarks:

Currently, there are several benchmarks for evaluating NLP models in the climate subdomain. The authors of Climate-FEVER built a dataset for claim verification of climate-related questions (Diggelmann et al., 2020). They gathered climate claims from scientifically informed and climate change sceptic sources, then combined them with relevant evidence from Wikipedia, labelling the evidence as supporting, refuting, or lacking sufficient information. In addition, several datasets are created to categorise climate-related texts for binary or multi-class classification, including stance detection and climate topic classification (Vaid et al., 2022; Leippold and Varini, 2020; Mishra and Mittal, 2021). These datasets rely on information found on social media and Wikipedia. Spokoyny et al. (2023) incorporated the above-named works with their own benchmark to develop ClimaBench. They developed a dataset about climate-related insurance policies which are based on climate-accessible questionnaires. Currently, existing climate benchmarks mainly focus on text classification, except for Climate-FEVER which can be used for RAG model evaluation, but does not specifically address corporate sustainability, a gap present in all benchmarks.

B.2 Advancements in Climate-Focused Natural Language Processing

A significant portion of current climate-related Natural Language Processing (NLP) research centers on classification tasks. Multiple studies demonstrate that further pretraining on climate-related text improves text classification (Nugent et al., 2021; Bingler et al., 2022; Mehra et al., 2022; Ekimetric, 2023). Additionally, some researchers have utilised financial disclosures to develop the Corporate Sustainability Disclosure Index (CSDI) (Tian et al., 2023). In this approach, firms listed in China were analysed using the tf-idf weighting scheme to assign sustainability scores, resulting in the creation of a sustainability index.

In the scope of Q&A tools within the climate subdomain, several chat tools leveraging LLMs have been developed. Climategpt is a Llama2 model that is further finetuned on climate specific data (Thulke et al., 2024). The authors of ChatClimate (Vaghefi et al., 2023) employed a RAG design, using encoded chunks of IPCC reports as a knowledge base. The top-k semantically similar chunks to the questions are included in the prompt, which is then answered by GPT-4, providing clear responses with cited sources on climate change topics. Similarly, ChatReport (Ni et al., 2023) offers a tool that automatically analyses sustainability reports based on TCFD questions, responding in a manner akin to ChatClimate.

The primary focus of these tools is optimising the augmentation phase through prompt engineering. A common limitation, however, is their dependence on qualitative evaluations by climate experts, often involving only a limited number of reports, or in some cases, none at all. This raises concerns about the effectiveness of these tools and indicates a lack of comprehensive performance evaluation. Consequently, this limitation hinders future studies from improving their designs.

C Prompts

C.1 Baseline Prompt

This prompt includes no extensive prompt engineering:

```
Given the following question: {question}
and the following extracted sources:

{summaries}

Answer the question and format your answer in JSON format with the two keys: BINARY_ANSWER (this should be a yes or no), ANSWER (this should contain your fully explained answer string without sources)
Keep your ANSWER within {answer_length} words.
```

C.2 ChatReports Adapted Prompt

In this prompt, we replaced the TCFD Assessment guidelines with the TPI Assessment guidelines referenced as {guidelines}. The prompt is displayed on the next page.

```
As a senior equity analyst with expertise in climate science evaluating a company’s sustainability report, you are presented with the following background information: {basic_info} With the above information and the following extracted components (which may have incomplete sentences at the beginnings and the ends) of the sustainability report at hand, please respond to the posed question, ensuring to reference the relevant parts ("SOURCES").

Format your answer in JSON format with the three keys: BINARY_ANSWER (this should be a yes or no), ANSWER (this should contain your fully explained answer string without sources), and SOURCES (this should be a list of the source numbers that were referenced in your answer).

QUESTION: {question}
=========
{summaries}
=========

Please adhere to the following guidelines in your answer: 1. Your response must be precise, thorough, and grounded on specific extracts from the report to verify its authenticity. 2. If you are unsure, simply acknowledge the lack of knowledge, rather than fabricating an answer. 3. Keep your ANSWER within {answer_length} words. 4. Be skeptical to the information disclosed in the report as there might be greenwashing (exaggerating the firm’s environmental responsibility). Always answer in a critical tone. 5. cheap talks are statements that are costless to make and may not necessarily reflect the true intentions or future actions of the company. Be critical for all cheap talks you discovered in the report. 6. Always acknowledge that the information provided is representing the company’s view based on its report. 7. Scrutinize whether the report is grounded in quantifiable, concrete data or vague, unverifiable statements, and communicate your findings. {guidelines}

Your FINAL_ANSWER in JSON (ensure there’s no format error):
```

C.3 Subquestion Generation Prompt

In this prompt, we use the prompt as stated in the LLamaIndex library. As tools we’ll use the library document of the company in question and as question the CorSus claim.

```
Given a user question, and a list of tools, output a list of relevant sub-questions
in json markdown that when composed can help answer the full user question:
# Example 2
<Tools>
“‘json
{tools_str}
‘‘‘

<User Question>
{query_str}

<Output>
```

C.4 Hypothetical Document (HyDe) Generation Prompt

In the prompt on the next page we use the prompt of the authors of HyDE (Gao et al., 2023). We make a small adjustment to sustainability disclosures and add two few-shot examples from our CorSus-Qualitative dataset. For the HyDe-Subquesiton version, we use the following prompt on the generated subquestions:.

Please write a passage of a sustainability disclosure to answer the question and assessment.
Try to include as many key details as possible.

Example 1:
Question: Does the company disclose an internal price of carbon?
Companies are assessed as Yes if they report on Scope 3 emissions separately, either in total or in one or more categories, or if they provide a total for Scope 1, 2 and 3 emissions.
Passage:
We currently plan to use external prices as the basis for CSC's internal carbon pricing. External prices include domestic carbon tax and overseas carbon tariffs. Besides calculating carbon emission related costs and conducting sensitivity analysis, we will also be able to effectively evaluate the benefits of carbon reduction related capital expenditures or R&D expenses. The internal carbon price will be able to effectively control the Company's overall carbon emissions, and drive the development of production processes and technologies with lower carbon emissions, or readjustment of internal operations and production processes.
Example 2:
Question: Has the company had its operational (Scope 1 and/or 2) greenhouse gas emissions data verified? Companies are assessed as Yes if their operational greenhouse gas emissions have been independently verified by a third party, or if they state they are international assurance standard they have used and the level of assurance.
Passage:
We undertake external verification of our greenhouse gas emissions annually. Our Scope 1 and 2 greenhouse gas emissions from assets and activities under our operational control and emissions associated with the use of our energy products (Scope 3) included in our net carbon intensity have been verified to a level of limited assurance.

{question}

Passage:

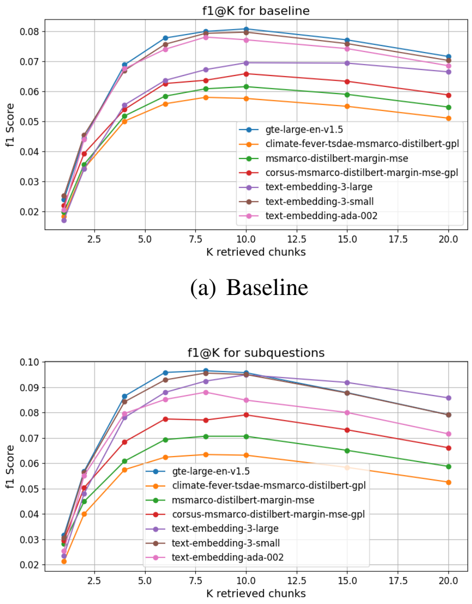
D Finetuning Retrieval Model

As part of our research, we further finetuned the 'msmarco-distilbert-margin-mse' model on the CorSus-Quantitative dataset using the GPL method (Wang et al., 2022). We adhered to the same parameters as those used in the original paper except for adjusting the GPL Steps from 140000 to 3000 due to limited computational resources. Below are the detailed parameters and configurations we used for finetuning:

- **Generator Model for Text-Generation**: `BeIR/query-gen-msmarco-t5-base-v1`
- **Base Model**: `GPL/msmarco-distilbert-margin-mse` (the starting checkpoint of the experiments in the paper)
- **Retrievers**:
  - `msmarco-distilbert-base-v3`
  - `msmarco-MiniLM-L-6-v3`
- **Retriever Score Functions**:
  - `cos_sim`
  - `cos_sim`
- **Cross-Encoder**: `cross-encoder/ms-marco-MiniLM-L-6-v2`
- **GPL Score Function**: `dot`
- **Batch Size for GPL**: 32
- **GPL Steps**: 3000
- **New Size**: -1
- **Queries Per Passage**: -1

This fine-tuning process aims to improve retrieval performance by generating synthetic queries for passages and refining the base model by using these query, passage tuples. We used all documents in the CorSus-Quantitative dataset, excluding those in the CorSus-Qualitative dataset, to ensure a fair train-validation split.

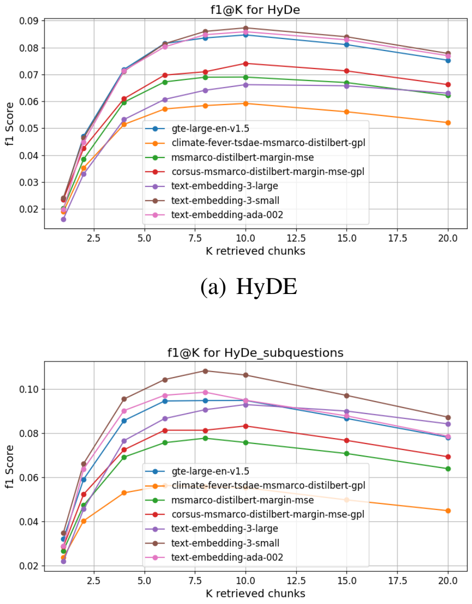
E Additional Results



(a) Baseline

(b) Subquestions

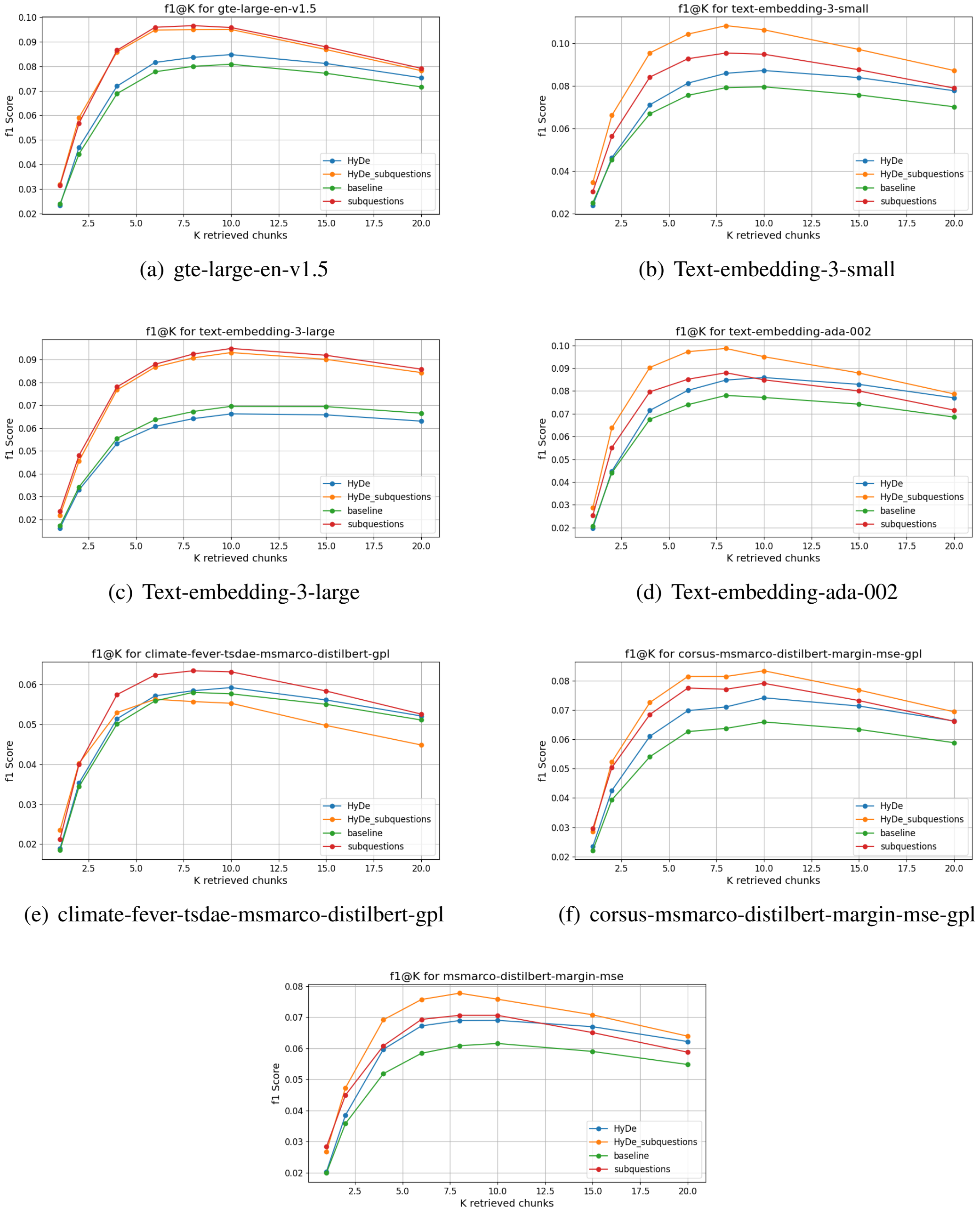
Figure 5: Retrieval method performance with different models



(a) HyDE

(b) HyDe Subquestions

Figure 6: Retrieval method performance with different models



(a) gte-large-en-v1.5

(b) Text-embedding-3-small

(c) Text-embedding-3-large

(d) Text-embedding-ada-002

(e) climate-fever-tsdae-msmarco-distilbert-gpl

(f) corsus-msmarco-distilbert-margin-mse-gpl

(g) Baseline distilbert-gpl

Figure 7: Retrieval model performance with different methods

Information Type	Generational Model	Prompt	Reranker	Accuracy	Balanced Accuracy	Correct Explanation	TP	FP	TN	FN
Noise	gpt-3.5-turbo-0125	Baseline	-	**1.000**	**1.000**	-	-	0	8	-
		chatReport	-	0.625	0.625	-	-	3	5	-
	gpt-4o-2024-05-13	Baseline	-	**1.000**	**1.000**	-	-	0	8	-
		chatReport	-	**1.000**	**1.000**	-	-	0	8	-
	llama3:70b	Baseline	-	**1.000**	**1.000**	-	-	0	8	-
		chatReport	-	**1.000**	**1.000**	-	-	0	8	-
	llama3:7b	Baseline	-	**1.000**	**1.000**	-	-	0	8	-
		chatReport	-	**1.000**	**1.000**	-	-	0	8	-
No Relevancy	gpt-3.5-turbo-0125	Baseline	-	0.344	0.344	-	-	63	33	-
		chatReport	-	0.667	0.667	-	-	32	64	-
	gpt-4o-2024-05-13	Baseline	-	**0.750**	**0.750**	-	-	24	72	-
		chatReport	-	**0.750**	**0.750**	-	-	24	72	-
	llama3:70b	Baseline	-	0.719	0.719	-	-	27	69	-
		chatReport	-	0.794	0.794	-	-	20	77	-
	llama3:7b	Baseline	-	0.458	0.458	-	-	52	44	-
		chatReport	-	0.604	0.604	-	-	38	58	-
Partially Relevant	gpt-3.5-turbo-0125	Baseline	LostInTheMiddleReranker	0.573	0.435	-	53	28	2	13
			-	0.635	0.508	-	56	25	5	10
		chatReport	LostInTheMiddleReranker	0.604	0.558	-	45	17	13	21
			-	0.615	0.583	-	44	15	15	22
	gpt-4o-2024-05-13	Baseline	LostInTheMiddleReranker	0.594	0.495	-	50	23	7	16
			-	0.604	0.521	-	49	21	9	17
		chatReport	LostInTheMiddleReranker	0.667	**0.648**	-	46	12	18	20
			-	0.667	**0.648**	-	46	12	18	20
	llama3:70b	Baseline	LostInTheMiddleReranker	**0.674**	0.564	-	54	19	8	11
			-	0.613	0.52	-	49	20	8	16
		chatReport	LostInTheMiddleReranker	**0.674**	0.608	-	48	16	12	13
			-	0.636	0.571	-	45	17	11	15
	llama3:7b	Baseline	LostInTheMiddleReranker	0.625	0.509	-	54	24	6	12
			-	0.594	0.505	-	49	22	8	17
		chatReport	LostInTheMiddleReranker	0.583	0.552	-	42	16	14	24
			-	0.583	0.552	-	42	16	14	24
Only Relevant	gpt-3.5-turbo-0125	Baseline	-	0.688	0.518	0.667	64	28	2	2
		chatReport	-	0.740	0.656	0.74	58	17	13	8
	gpt-4o-2024-05-13	Baseline	-	0.667	0.539	0.667	58	24	6	8
		chatReport	-	**0.771**	**0.706**	**0.771**	58	14	16	8
	llama3:70b	Baseline	-	0.656	0.568	0.625	53	20	10	13
		chatReport	-	0.695	0.626	0.674	53	16	13	13
	llama3:8b	Baseline	-	0.646	0.515	0.635	57	25	5	9
		chatReport	-	0.677	0.629	0.646	50	15	15	16
	llama2:7b	Baseline	-	0.688	0.5	0.531	66	30	0	0
		chatReport	-	0.688	0.5	0.656	66	30	0	0
	adapt-finance-llama2-7b	Baseline	-	0.604	0.512	0.167	50	22	8	16
		chatReport	-	0.646	0.533	0.50	55	23	7	11
	eci-io-climategpt-7b	Baseline	-	0.649	0.536	0.299	44	16	6	11
		chatReport	-	0.632	0.524	0.411	53	23	7	12

Table 5: Comparison of the ChatReport prompt (Appendix C.2) and the baseline prompt (Appendix C.1) with and without the LostInTheMiddleReranker (Liu et al., 2024) in assessing generation quality using different models. The evaluation is conducted under varying levels of relevant information: "Noise" includes irrelevant SEC-10K text, "No Relevancy" includes irrelevant information from the same company, "Partial Relevancy" includes half of the relevant information, and "Only Relevant" uses only entirely relevant references. Metrics include Accuracy, Balanced Accuracy (mean of sensitivity and specificity), Correct Explanation, True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). For "Noise" and "No Relevancy" conditions, TP and FN are omitted as the correct information is never provided, focusing instead on FP and TN. The percentage of positive and negative answers is 64.6% and 35.4%, respectively for all cases where information is provided. Bold values indicate top performance per Information Type.

F Corporate Sustainability Benchmark

F.1 Assumptions

Several critical assumptions were established in the course of developing the benchmark:

1. All publicly available data used by the Transition Pathway Initiative (TPI) is available solely in PDF format on their website. No other sources are used as the website of the company itself or other publicly available data not present on their website.

2. As reference documents to answer the TPI questions of a particular year, we took the closest available documents provided.

3. We excluded any documents that cover no sustainable insights, such as financial 10-k forms.

As stated in the TPI methodology all information used during their assessment is solely based on public information provided by the company in the assessed year. We excluded the websites of companies as publicly available data to ensure consistent and a standardised approach. In the case of TPI, all assessment dates are around April / May. Therefore, we took the annual / sustainability disclosures of the year before that as the most recent documentation.

F.2 Development CorSus-qualitative

To construct the qualitative subset, we enlisted sustainability experts familiar with the TPI framework and the specific sustainability criteria of each assessed company. The process involved several steps:

1. **Document Review:** Experts were provided with the TPI question, the recorded answer, and the relevant documents from the company for the corresponding year.

2. **Text Retrieval:** Experts were tasked with identifying relevant sections of text within these documents that directly related to the TPI question.

3. **Explanation Formulation:** Based on the identified text and the context of the TPI assessment, experts then crafted a detailed explanation that justified the given answer, linking the documentation explicitly to the TPI's evaluation criteria.

F.2.1 Misallignment TPI and Sustainability Experts Answers

During the data collection phase, sustainability experts reviewed the same documents used by TPI to assess corporate sustainability. These experts, equipped with the TPI question, the corresponding answer, and the relevant documents, often arrived at different conclusions than the expert's of TPI. In total, the sustainability experts disagreed with that of the TPI answers on 16 of the 96 datapoints. This discrepancy can be attributed to the subjective nature of some sustainability questions, which may lead to varied interpretations and opinions.

The questions that tend to show the highest levels of divergence are those that inherently require a judgment call, reflecting personal or collective values and perspectives on sustainability. This subjectivity underscores the variability in responses between the TPI and the sustainability experts. In light of these discrepancies, we adjusted the original TPI dataset to include insights from the sustainability experts for the qualitative subset.

Document Type	**Quantitative**	**Qualitative**
Annual Report	275	11
Sustainability Report	196	9
Climate Report	74	4
CDP Report	46	2
ESG Report	28	1
Lobbying Report	24	6
TCFD Report	16	0
Proxy Report	16	0
Governance Report	13	0
ESG Datasheet	13	2
Climate Review	12	2
Bond Report	9	0
Industry Associations Report	9	4
Remuneration Report	7	3
Economic Contribution Report	6	1
SASB Report	6	0
Universal Registration Document	6	1
GRI Report	5	0
Resources Report	3	1
Financial Statements	3	0
Board of Directors Report	3	0
UN Commitment Report	3	0
Assessment Report	3	0
Compensation Report	2	1
Accountability Report	1	0
Third Party Review	1	0
Modern Slavery Report	1	0
Biodiversity Report	1	0
Total	**782**	**48**

Table 6: Document types within CorSus-Quantitative and CorSus-Qualitative dataset.

www.transitionpathwayinitiative.org/methodology

F.3 Listed Companies in CorSus

The companies included in the CorSus benchmark:
Quantitative: ADBRI, AGL Energy, Air France KLM, Air Liquide, Airbus, American Electric Power, Anglo American, Anhui Conch Cement, Arcelor Mittal, BASF, BHP, BMW, BP, Bayer, Berkshire Hathaway, BlueScope Steel, Boeing, Boral, CRH, Cemex, Centrica, China Steel, Coca-Cola, Danone, Dominion Energy, Dow, E.ON, EDF, Enbridge, Enel, Engie, Eni, Equinor, Ford, Fortum, General Motors, Glencore, Grupo Argos, HeidelbergCement, Hitachi, Hon Hai Precision Industry, Honda, Iberdrola, Incitec Pivot, Kinder Morgan, LyondellBasell Industries, Marathon Petroleum, Mercedes-Benz, NRG Energy, NTPC, National Grid, Nestle, Nippon Steel, Nissan, OMV, Orica, Origin Energy, PPL, PTT, Panasonic, Pepsico, Petrobras, Phillips 66, Posco, Procter & Gamble, Qantas, RWE, Reliance Industries, Renault, Repsol, Rio Tinto, Rolls-Royce, SK Innovation, SSAB, SSE, Sasol, Shell, South32, St Gobain, Suncor Energy, Suzano, Suzuki, TC Energy, ThyssenKrupp, Toray Industries, TotalEnergies, Toyota, Trane Technologies, UltraTech Cement, Unilever, Uniper, United Tractors, Vale, Volkswagen, Volvo, WEC Energy Group, Walmart, Weyerhaeuser, Woolworths, Xcel Energy.
Qualitative: Air Liquide, Anglo American, BASF, BP, Berkshire Hathaway, China Steel, Enel, Engie, Equinor, OMV, Phillips 66, Shell.

Question ID	CorSus	
	Quantitative	Qualitative
9	0.833	0.833
10	0.806	0.833
13	0.278	0.167
15	0.75	0.75
16	0.75	0.75
17	0.667	0.75
18	0.75	0.75
19	0.361	0.333
Average	0.649	0.646

Table 7: Quantitative and Qualitative percentage of "Yes" answers for each question ID in the CorSus assessment.

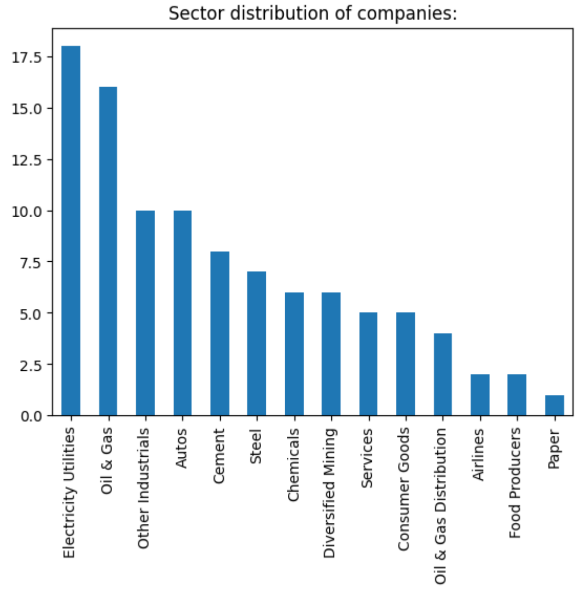


Figure 8: This bar plot displays the sector distribution of companies in the CorSus benchmark. The benchmark includes companies from the CA100 dataset, primarily originating from polluting sectors such as Electricity Utilities and Oil & Gas.

Question	Yes	No	Not applicable
1	100.00	0.00	0.00
2	98.72	1.28	0.00
3	99.62	0.38	0.00
4	98.47	1.53	0.00
5	98.85	1.15	0.00
6	96.04	3.96	0.00
7	97.95	2.05	0.00
8	95.14	4.86	0.00
9	84.97	15.03	0.00
10	80.82	19.18	0.00
11	96.80	3.20	0.00
12	29.67	7.16	63.17
13	17.90	82.10	0.00
14	97.44	2.56	0.00
15	78.64	21.36	0.00
16	69.82	30.18	0.00
17	78.77	21.23	0.00
18	69.18	30.82	0.00
19	15.32	86.68	0.00

Table 8: Distribution of original Transition Pathway Initiative (TPI) questions. Only questions within the 85% distribution range were selected for further analysis, omitting question 14 due to its non-binary response. The underlined questions (9, 10, 13, 15, 16, 17, 18, and 19) meet this criterion, ensuring a more representative and balanced evaluation of model performance.

TPI ID	Question	Assessment Criteria	Example Documents	Example Answer	Example Explanation	References
9	Has the company had its operational (Scope 1 and/or 2) greenhouse gas emissions data verified?	Companies are assessed as Yes if their operational greenhouse gas emissions have been independently verified by a third party, or if they state they are international assurance standard they have used and the level of assurance.	Annual Report, Climate Report, ESG Datasheet	Yes	Yes, Deloitte and Ernst & Young, the Statutory Auditors of ENGIE, have given reasonable assurance that the scope 1 and 2 emissions data of ENGIE as presented in its Annual Report for 2022, have been prepared, in all material aspects, in accordance with the criteria used by the company.	Annual Report p.52,p.43,p.54
10	Does the company support domestic and international efforts to mitigate climate change?	Companies are assessed as Yes if they demonstrate support for mitigating climate change through membership of business associations that are supportive, and if they have a clear company position on public policy and regulation.	Annual Report, Climate Report, ESG Datasheet	No	No, The documents reviewed only refer to the company's report on transparent lobby practices (Climate report, page 7), without describing the outcome of this evaluation on whether its lobby practices support the climate goals of the Paris Agreement. As such the conclusion that the company supports climate mitigation through its business association memberships cannot be drawn on the basis of these materials.	Climate Report p.7
13	Does the company disclose its membership and involvement in organisations or coalitions dedicated specifically to climate issues?	Companies are assessed as Yes if the company discloses: (a) Its memberships in specific organisations dedicated to climate-related issues (b) Its involvement in these organisations	Annual Report, Climate Report, ESG Datasheet	Yes	Yes, ENGIE conducts a review of its membership of professional and industry associations including a detailed evaluation of the alignment of these associations with the objectives of the Paris Agreement. The documents reviewed do not disclose the outcome of this review.	Climate Report p.7
15	Does the company's remuneration for senior executives incorporate climate change performance?	Companies are assessed as Yes if executive remuneration incorporates climate change performance.	Annual Report, Climate Report, ESG Datasheet	Yes	Yes, ENGIE's compensation policy includes climate change performance criteria in both the annual variable and long term incentives. Climate change is part of the 35% non-financial criteria of the annual variable, where it is included in the CSR criteria (30%). It is also included in the CSR objectives (20%) of the long-term incentives.	Annual Report p.49
16	Does the company incorporate climate change risks and opportunities in their strategy?	Companies are assessed as Yes if they detail how they incorporate climate change risks and opportunities in their strategy (mitigation, new products, R&D, etc.), and if they disclose the impact of climate change risks and opportunities on financial planning (OPEX, CAPEX, M&A, debt).	Annual Report, Climate Report, ESG Datasheet	Yes	Yes, ENGIE identifies climate risks and opportunities as per the TCFD requirements, and describes how the risk is evaluated in the company's governance structure. The company also describes risks it anticipates across the entire supply chain, and physical risks. Adaptation to the physical risks of climate change is embedded in the Group2̌019s investment process.ENGIE also identifies a number of opportunities related to renewables and energy efficiency. These are linked to the company's strategy, which is in turn linked to CAPEX.	Annual Report P.8 P.18 and Climate Report p.5

Table 9: **Part 1:** Selected questions and assessment requirements from the Transition Pathway Initiative, including an example answer and reference from the CorSus-Qualitative dataset for the company Engie.

TPI ID	Question	Assessment Criteria	Example Documents	Example Answer	Example Explanation	References
17	Does the company undertake climate scenario planning?	Companies are assessed as Yes if they mention the 2 degrees scenario in relation to business planning or confirm they have conducted climate related scenario analysis, and if they describe the business impact of one or more climate scenario analysis.	Annual Report, Climate Report, ESG Datasheet	No	No, Documents reviewed do not explicitly list climate scenario's used for strategic planning. The Climate Report only mentions company is using scenario planning, and the company's own convictions regarding decarbonization scenarios.	Climate Report p.3
18	Does the company disclose an internal price of carbon?	Companies are assessed as Yes if they disclose their internal carbon price.	Annual Report, Climate Report, ESG Datasheet	No	No, the company says it implements an internal carbon price but documents reviewed do not disclose internal carbon price applied.	No Reference
19	Does the company ensure consistency between its climate change policy and the positions taken by trade associations of which it is a member?	Companies are assessed as Yes if they have a stated policy or commitment to ensure consistency between their climate change policy and the position taken by the trade associations of which they are members, and for responding appropriately in those instances where the trade association positions is significantly weaker than or contradicts that of the company.	Annual Report, Climate Report, ESG Datasheet	No	No, In the documents that are part of the review, ENGIE has no policy for ensuring consistency and no policy for responding appropriately. They do describe a policy for ensuring consistency. However, their policy for responding appropriately is also insufficient because they remain a member of 'partially aligned' industry associations.	Climate Report p.7

Table 10: **Part 2:** Selected questions and assessment requirements from the Transition Pathway Initiative, including an example answer and reference from the CorSus-Qualitative dataset for the company Engie.

ID	Generated Answer	True Answer	Generated Explanation	Sources Used
9	Yes	Yes	Yes, Engie has had its operational (Scope 1 and 2) greenhouse gas emissions data verified. The report mentions that the company has undergone a rigorous process of independent verification by third parties, including setting up analytical procedures, conducting interviews, and carrying out substantive tests on a representative sample of entities. This level of assurance provides credibility to the reported emissions data (SOURCE: Engie_2022_Annual_Report.pdf, p.5).	ESG Datasheet (p.5), Annual Report (p.52, 22, 23, 54, 28), Climate Report.pdf (p.8, 7), ESG Datasheet (p.3)
10	Yes	No	Yes, ENGIE supports domestic and international efforts to mitigate climate change. The company has demonstrated its commitment through various initiatives such as being a founding member of the First Movers coalition, supporting renewable hydrogen projects, signing pledges with public and private partners, and setting ambitious decarbonization goals. Additionally, ENGIE has specific governance structures in place to address climate-related issues and regularly reports on its climate performance, aligning with international benchmarks. However, further scrutiny is needed to ensure the company's actions align with its stated commitments and are not merely greenwashing tactics. The information provided is based on the company's report.	Annual Report (p.25, 5, 29, 50, 30, 16), Climate Report (p.7, 6, 3), ESG Datasheet (p.3)
13	Yes	Yes	Yes, ENGIE discloses its membership and involvement in organizations dedicated to climate issues. The company is a founding member of the First Movers coalition, aiming to accelerate the development of a carbon-neutral supply chain. Additionally, ENGIE has signed pledges and commitments with public and private partners to advance the energy transition. These actions demonstrate the company's active engagement in organizations focused on climate-related issues (SOURCE: Engie_2022_Annual_Report.pdf, p.25).	Annual Report (p.25, 50, 15, 29), Climate Report (p.6, 7)
15	Yes	Yes	Yes, the company's remuneration for senior executives, including the Chief Executive Officer, does incorporate climate change performance. The variable compensation for senior managers and the CEO is partly dependent on meeting objectives related to reducing CO2 emissions and increasing the share of renewables in the company's electricity mix capacities. This demonstrates a tangible link between executive compensation and climate performance, aligning financial incentives with sustainability goals (SOURCES: Engie_2022_Climate_Report.pdf, p.7; Engie_2022_Annual_Report.pdf, p.50).	Climate_Report (p.7, 6), Annual Report (p.50, 49, 23, 47, 54), ESG Datasheet (p.5)
16	Yes	Yes	Yes, Engie incorporates climate change risks and opportunities in their strategy. They analyze physical risks associated with climate change, adapt their activities, and have a partnership to model future energy production trends. Engie assesses the impact of climate change on their strategy, investment process, and financial planning. They have specific governance structures to address climate issues and set a Net Zero Carbon target for 2045. However, further verification is needed to ensure the actual implementation and effectiveness of these strategies (SOURCES: Engie_2022_Climate_Report.pdf).	Climate Report (p.5, 6, 2, 7), Annual Report (p.29, 18, 50), ESG Datasheet (p.7)
17	Yes	No	Yes, Engie undertakes climate scenario planning as outlined in their sustainability report. They have partnered with the Pierre Simon Laplace Institute to model future energy production trends and the impact of extreme events on their technologies. They analyze the impact of climate change on their strategy using a country-by-country approach and study major climate regions of interest. Engie aims to develop plans for adapting all its sites and activities based on the climate impacts identified, demonstrating a proactive approach to climate scenario planning (SOURCES: Engie_2022_Climate_Report.pdf, p.5; Engie_2022_ESG_Datasheet.pdf, p.7).	Climate Report (p.5, 3), ESG Datasheet (p.7), Annual Report (p.23, 31, 30, 18, 29)
18	Yes	No	Yes, Engie discloses an internal price of carbon as part of its sustainability report. The report mentions the consideration of carbon prices in investment decisions and long-term forecasts of energy prices, indicating a commitment to incorporating carbon pricing into its operations (SOURCE: Engie_2022_Climate_Report.pdf, p.7). This demonstrates a proactive approach to addressing climate change, although further scrutiny is needed to ensure the actual implementation and impact of this internal carbon price.	Climate Report (p.7, 8, 2, 6), ESG Datasheet (p.3), Annual Report (p.30, 16, 22, 39)
19	No	No	Based on the extracted components from Engie's sustainability report, the company does not explicitly ensure consistency between its climate change policy and the positions taken by trade associations of which it is a member. While Engie outlines ambitious decarbonization goals and emphasizes its commitment to the energy transition, there is a lack of specific mention or commitment to addressing potential conflicts between its climate policy and trade association positions. Without clear statements or actions addressing this issue, Engie's alignment with trade associations on climate change remains uncertain and potentially lacking. The report contains some cheap talks that emphasize broad commitments without concrete actions or accountability measures.	Climate Report (p.8, 3, 6), AnnualReport (p.5, 15, 14, 2)

Table 11: Example answers generated by chatReport3+ model with the correlated true answer for ENGIE in the year 2022 on the CorSus-Qualitative benchmark. A clear distinction is visible that the model still struggles to identify the true negative answers.