# DIRAS: Efficient LLM-Assisted Annotation of Document Relevance in Retrieval Augmented Generation

**Anonymous ACL submission**

## Abstract

Retrieval Augmented Generation (RAG) is widely employed to ground responses to queries on domain-specific documents. But do RAG implementations leave out important information or excessively include irrelevant information? To allay these concerns, it is necessary to annotate domain-specific benchmarks to evaluate information retrieval (IR) performance, as relevance definitions vary across queries and domains. Furthermore, such benchmarks should be cost-efficiently annotated to avoid annotation selection bias. In this paper, we propose DIRAS (**D**omain-specific **I**nformation **R**etrieval **A**nnotation with **S**calability), a manual-annotation-free schema that fine-tunes open-sourced LLMs to annotate relevance labels with calibrated relevance probabilities. Extensive evaluation shows that DIRAS fine-tuned models achieve GPT-4-level performance on annotating and ranking unseen (query, document) pairs, and is helpful for real-world RAG development.[1]

## 1 Introduction

RAG has become a popular paradigm for NLP applications (Gao et al., 2024). One core phase of RAG systems is Information Retrieval (IR), which leverages cheap retrievers to filter relevant information and thus save LLM inference costs. Given its cost-saving nature, IR might be a performance bottleneck for RAG (Chen et al., 2023b; Gao et al., 2024). Both leaving out important relevant information (*low recall*) as well as including excessively related but irrelevant information (*low precision*) may lead to severe performance down-grades (Ni et al., 2023; Cuconasu et al., 2024; Niu et al., 2024; Schimanski et al., 2024a). However, evaluation results on general-domain benchmarks (Thakur et al., 2021) may hardly indicate the IR performance on RAG systems, as the definition of relevance varies drastically across different domains and use cases (Bailey et al., 2008). See App. A for an example where the relevance judgment may vary widely with or without domain expertise. Therefore, domain-specific benchmarks need to be annotated to evaluate RAG systems in different domains.

To address this scarcity of domain-specific IR benchmarking, we propose **DIRAS**, a pipeline to annotate domain-specific IR data at scale (illustrated in Fig. 1). Specifically, users only need to input (1) some domain-specific queries and documents and (2) definitions for what is (ir)relevant for each query. Then, DIRAS distills relevance prediction data from GPT-4 to fine-tune open-sourced LLMs, which can then be used to annotate relevance on a large scale without extra API cost, avoiding annotation selection bias (Thakur et al., 2021) with limited expenditure.

Prior work distills open-sourced LLM-based rerankers using pairwise (Qin et al., 2024) or listwise (Sun et al., 2023b; Pradeep et al., 2023) as the ranking paradigm. However, we choose a pointwise approach (i.e., predicting a relevance score per document and then ranking according to the scores) for DIRAS, to fulfill our three **D**esiderata for a document relevance annotator:

**D1. Efficient and Effective**: When benchmarking IR of RAG systems, it is important to annotate all (query, paragraph) pairs to avoid annotation selection bias, since information thoroughness is critical for many queries (e.g., overall assessment, Ni et al., 2023). Therefore, the pointwise method is more favorable as it largely outperforms list- or pairwise method in efficiency (Sun et al., 2023a). Prior work worries that pointwise ranking is efficient but not effective due to the difficulty in calibration (Sun et al., 2023a; Qin et al., 2024). However, this method lacks empirical investigation in prior work.

[1] We will open-source all our codes, LLM generations, and human annotations.

In our work, we observe both GPT-4 and DIRAS fine-tuned LLMs achieve very competent pointwise ranking results thanks to good calibration.

**D2. Improved Leverage of the Relevance Definition**: Relevance and partial relevance judgments might be subjective without domain expertise or explicit annotation guidelines (Bailey et al., 2008; Saracevic, 2008; Thomas et al., 2024; see also App. A). Therefore, DIRAS explicitly puts relevance definition into relevance prediction prompts to achieve more objective and consistent results. Compared to the list- or pairwise method, the pointwise method is provided with only one document each time. Thus, it may analyze the document along the relevance definition in better detail, especially with CoT prompting (Wei et al., 2022).

**D3. Richer Predictions for RAG Requirements**: Listwise or pairwise ranking algorithms only predict a relative rank for documents. With only the rank, retrieving the same number of documents (top-k) for all questions is suboptimal as it is likely that different questions have different amounts of relevant information (details in § 5). In contrast, the pointwise method predicts not only ranks but also binary relevance and calibrated relevance probabilities. Both binary labels and relevance scores allow RAG systems to retrieve the actual amount of relevant information to a question. Furthermore, the calibrated relevance probability is also helpful in automatic annotation (Ni et al., 2024), indicating which annotations are partially relevant and/or check-worthy.

To evaluate DIRAS, we conduct experiments with two datasets. First, we annotate a high-quality dataset based on ChatReport[2] (Ni et al., 2023), a real-world RAG application analyzing lengthy corporate reports. This dataset also incorporates the ideas of partial relevance and labeling uncertainty. ChatReport is representative of RAG applications that are sensitive to retrieval results, as thoroughly analyzing disclosed information is crucial for assessing what is under-addressed in the reports. Evaluation on ChatReport data shows that the pointwise ranking of DIRAS achieves very satisfactory performance, even superior to the widely-adopted listwise method (Pradeep et al., 2023) (§ 3.1). The best DIRAS fine-tuned model also achieves GPT-4-level performance in terms of relevance ranking and calibration (§ 3.2).

Second, we use a DIRAS fine-tuned model to re-annotate all (query, document) combinations (43k in total) in ClimRetrieve (Schimanski et al., 2024b), a real-world record of experts' information seeking workflow. Experiments show that DIRAS fine-tuned models successfully understand fine-grained degree of relevance (§ 4.1) and enhance their performance through improved relevance definitions (§ 4.2). Furthermore, it mitigates IR annotation bias by identifying information ignored by experts (§ 4.3), and benchmarks IR algorithms' target domain performance upon all 43k (query, document) pairs (§ 4.4). Finally, we propose recommendations for future RAG designs based on our takeaways (§ 5).

Collectively, our contributions include:

1. We propose DIRAS, a framework tuning open-sourced LLMs into efficient and effective IR annotators, taking domain expertise into account.
2. We annotate a high-quality IR benchmark based on ChatReport with (partial) relevance labels and uncertainty labels, based on explicit relevance definitions.
3. We compare DIRAS fine-tuned models with a real-world information-seeking workflow by experts, showing the model accurately understands granular relevance definitions and helps mitigate annotation selection bias.

## 2 DIRAS

### 2.1 DIRAS Pipeline

To address the outlined desiderata (**D1**, **D2**, **D3**) in § 1, we design the DIRAS pipeline (illustrated in Fig. 1). This pipeline comprises the training data creation and fine-tuning of small LLMs to the calibrated annotators.

**Training Data Creation**: We create the training data from domain-specific sources like reports. As a result, the data will be composed of a set of domain-specific (query, document) pairs. Each query comprises a question and a definition indicating what is relevant or irrelevant to the question. The relevance definition can be designed by human experts or generated by LLMs following App. B. To obtain the documents for each question/query, we employ a sampling strategy using top-k relevant documents ranked by a small dense retriever. We set a top-k threshold and sample an equal number of documents within and outside of the threshold for each question. While sampling in top-k aims

[2] https://reports.chatclimate.ai/

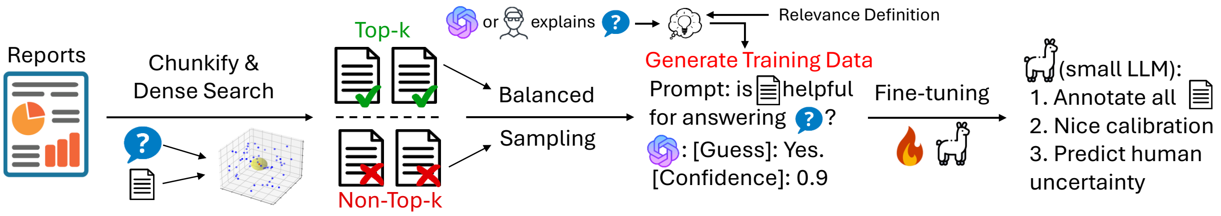


Figure 1: DIRAS pipeline. RAG-specific queries, documents as inputs; calibrated small-LLM annotators as outputs.

at covering some relevant documents, sampling outside of top-k ensures covering the broader distribution of (query, document) pairs.

**Open-Sourced LLM Fine-Tuning**: Once we sampled the domain-specific (query, document) pairs, we create instruction fine-tuning (IFT) data with a SOTA generic LLM $\mathcal{M}$ and the prompt template $\mathcal{P}$ (see e.g., Fig. 2). The prompt template incorporates a (query, document) pair and an instruction to predict a binary relevance label and its confidence score. Finally, the distilled IFT data is used to fine-tune open-sourced LLMs $\mathcal{M}_o$ ("o" stands for "open") to conduct binary relevance prediction with confidence.

### 2.2 Evaluation Metrics

Both $\mathcal{M}$'s and $\mathcal{M}_o$'s results should be evaluated against two types of human labels. (1) *Relevance labels*: whether a document is helpful for answering a query or not, disagreement between different annotators needs to be resolved to obtain the final relevance label. (2) *Uncertainty labels*: an annotation is uncertain, if the annotators have a strong disagreement, or the majority of them agree that the (query, document) is partially relevant. We use the following metrics for evaluation:

**Binary Relevance**: We compute the F1 Score of models' binary relevance prediction using *relevance labels*. For RAG systems, binary relevance labels are important for deciding which documents should be passed to LLMs.

**Calibration**: Both $\mathcal{M}$ and $\mathcal{M}_o$ give confidence scores, which should calibrate the binary accuracy to indicate whether the prediction is trustworthy. We use Expected Calibration Error (ECE), Brier Score, and AUROC to measure calibration performance, following Kadavath et al. (2022) and Tian et al. (2023).

**Information Retrieval**: The confidence scores also give a calibrated relevance probability which can be used to rank documents for each query. To directly evaluate the ranking performance, we measure nDCG and MAP upon *relevance labels*.

**Uncertainty**: If the models understand the difficulty and uncertainty caused by partial relevance, they should have lower confidence scores on samples that humans found to be uncertain. Thus we compute average precision (AP) scores between confidence and *uncertainty labels*. We chose AP because it is threshold-free by summarizing F1 scores of all thresholds of relevance scores. Details of computing all metrics are in App. C.

## 3 Experiments on ChatReport

To the best of our knowledge, there is no existing IR dataset that (1) accompanies each query with domain-specific relevance definition; and (2) provides annotations of human uncertainty caused by partial relevance. Therefore, we annotate a dataset based on ChatReport (Ni et al., 2023) to fulfill evaluation purposes in § 2.2. ChatReport is an online RAG application for corporate climate report analyses and Q&A about the reports.[3]

**Data Sampling**: We sample 31 questions and 80 reports from this application. Climate reports are sampled randomly from openly accessible user submissions[4] (PDF parsing details in App. D). The questions are strategically sampled to ensure representativeness and diversity. Specifically, 11 questions are the core questions used in ChatReport, which cover essential topics of sustainability disclosure. 20 questions are selected from users' customized questions posed to the ChatReport tool. Finally, we prompt GPT-4 to draft relevance definitions for all questions (see App. B).

**Train-Test Split**: We split the questions into 11 for testing and 20 for training. Similarly, we split the reports into 30 for testing and 50 for training. This ensures the evaluation on unseen questions and reports. For each question, we randomly sample 60 documents – 30 each from in the top-5 and

[3] See https://reports.chatclimate.ai/.

[4] See https://github.com/EdisonNi-hku/chatreport.

**Prompt:**
<question>: What is the firm's Scope 3 emission?
<question_definition>: This question is looking for information about the firm's emission in ...
<paragraph>: {one text chunk from a climate report}
Is <paragraph> helpful for answering <question>? Provide your best guess, and confidence score from 0 to 1.

**Teacher LLM $\mathcal{M}$:**
[Reason]: {Reason why the paragraph is (un)helpful.}
[Guess]: {Yes or No.}
[Confidence]: {confidence score between 0.0 and 1.0.}

Figure 2: Our prompt template $\mathcal{P}$ for distilling training data from $\mathcal{M}$. It is shortened for presentation. Full $\mathcal{P}$ is in Fig. 9.

outside the top-5 (using OpenAI text-embedding-3-small as the dense retriever). Ultimately, (query, document) pairs in training split are used to create training data with relevance label and confidence score predictions (details in § 3.1). Data points in the test split are passed to human annotation.

**Test Data Annotation**: We leverage relevance definitions as the annotation guidelines. A document is relevant if and only if it fulfills the definition. The data labeling process follows two steps. First, we employ two annotators who independently annotate all test data to be either relevant, irrelevant, or partially relevant. Second, we employ a subject-matter expert in corporate climate disclosure to resolve conflicts to obtain final relevance labels. Besides relevance labels, we also obtain uncertainty labels from human annotations: Whenever there is strong disagreement (co-existence of relevance and irrelevance labels) or agreement on partial relevance (two or more annotators agree on partial relevance), the data point is labeled as uncertain. Inter-annotator agreement and other details can be found in App. E.

### 3.1 How to Distill Training Data?

To train better open-sourced LLMs $\mathcal{M}_o$, it is crucial to distill high-quality training data from teacher LLM $\mathcal{M}$. Specifically, we compare the following three implementation choices:

**Pointwise vs. Listwise**: The listwise method is popular in ranking data distillation given its moderate cost and good performance (Sun et al., 2023b; Pradeep et al., 2023). However, the more efficient pointwise method is under-explored in prior work – majorly due to the concern about poor calibration (Sun et al., 2023a; Qin et al., 2024).

**Calibration method (Tok vs. Ask)**: One calibration method is to get the relevance confidence by probing the model's generation probability of the token Yes/No when predicting a document's relevance (Tok, Liang et al., 2023). An alternative way is directly asking LLMs to verbalize confidence score, which may work better for instruction following LLMs (Ask, Tian et al., 2023).

| Setting | Unc. | Bin. | Cal. | Info. | Avg. |
|---|---|---|---|---|---|
| List-2/1 | - | - | - | 76.86 | - |
| List-2/1-D | - | - | - | 74.72 | - |
| List-10/5 | - | - | - | 84.74 | - |
| List-10/5-D | - | - | - | 84.45 | - |
| List-20/10 | - | - | - | 78.05 | - |
| List-20/10-D | - | - | - | 82.54 | - |
| Point-Ask | 39.27 | 84.07 | **94.41** | 87.57 | 76.33 |
| Point-Ask-Prob-D | 44.74 | 84.52 | 93.72 | 88.39 | 77.84 |
| Point-Tok-D | 28.83 | 86.32 | 93.31 | 80.90 | 72.52 |
| Point-Ask-D | **54.01** | **86.32** | **94.41** | **88.48** | **80.80** |

Table 1: GPT-4's performance on ChatReport test set with different ranking methods (Point- or Listwise), with/without relevance definition (D), and calibration method (Ask or Tok). Unc. denotes the Average Precision (AP) of predicting uncertainty; Bin. denotes F1 score of binary relevance prediction; Cal. is the average of AUROC, ECE, and Brier Score; Info. is the average of nDCG and MAP; Avg. denotes the average of all metrics. Best scores of each column are **bolded**.

**With vs. without relevance definition**: As ChatReport test data is annotated based on the relevance definition, performance should increase if the model correctly takes the in-context relevance definition into consideration.

Following the takeaways of Thomas et al. (2024), we design the prompt $\mathcal{P}$ for the pointwise method with detailed task/role description, relevance definition and CoT prompting (see Fig. 2 and Fig. 9, prompt without definition in Fig. 11). We use the listwise ranking prompt from Sun et al. (2023b) and Pradeep et al. (2023) (see prompt with/without definition in Fig. 13/Fig. 12). For the pointwise method, we run one variation to test prompt sensitivity: directly asking for relevance probability instead of confidence for guess (prompt in Fig. 10). As the listwise ranking is sensitive to window/step size, we run three variations with window/step sizes of 2/1, 10/5, and 20/10. text-embedding-3-small is used for listwise methods' initial ranking. The results are shown in Table 1. We observe that: (1) With the proper calibration method (Ask), the pointwise method outperforms the listwise method. (2) The listwise method is sensitive to window size, while the pointwise method gives more consistent performance across prompts. (3) Adding relevance definition drops the listwise performance in 2 out

| Setting | Unc. | Bin. | Cal. | Info. | Avg. |
|---|---|---|---|---|---|
| Small-embed | - | - | 83.63 | 66.34 | - |
| Large-embed | - | - | 84.07 | 69.36 | - |
| BGE-Gemma | - | - | 78.21 | 68.47 | - |
| GPT-3.5 | 29.71 | 45.27 | 87.79 | 74.16 | 59.23 |
| GPT-4 | **54.01** | **86.32** | **94.41** | 88.48 | **80.80** |
| Llama3-CoT-Ask | 36.57 | 76.58 | 93.32 | 86.15 | 73.16 |
| Llama3-CoT-Tok | 41.74 | 76.58 | 90.82 | 85.96 | 73.78 |
| Llama3-Ask | 40.18 | 82.11 | 94.04 | 86.02 | 75.59 |
| Llama3-Tok | 41.60 | 82.11 | **94.41** | **89.19** | 76.83† |
| Phi3-CoT-Ask | 36.08 | 72.95 | 92.35 | 80.56 | 70.48 |
| Phi3-CoT-Tok | 35.49 | 72.95 | 88.15 | 80.64 | 69.31 |
| Phi3-Ask | 32.30 | 73.23 | 91.16 | 80.05 | 69.18 |
| Phi3-Tok | 38.00 | 73.23 | 92.05 | 86.94 | 72.55† |
| Gemma-CoT-Ask | 31.60 | 72.38 | 91.14 | 81.39 | 69.13 |
| Gemma-CoT-Tok | 39.03 | 72.38 | 88.58 | 80.33 | 70.08 |
| Gemma-Ask | 25.74 | 67.13 | 88.27 | 77.43 | 64.64 |
| Gemma-Tok | 50.72 | 67.13 | 92.44 | 81.17 | 72.87† |

Table 2: Comparison between the fine-tuned $\mathcal{M}_o$ and different baselines on ChatReport test data. Unc. denotes the Average Precision (AP) of predicting uncertainty; Bin. denotes F1 score of binary relevance prediction; Cal. is the average of AUROC, ECE, and Brier Score; Info. is the average of nDCG and MAP; Avg. denotes the average of all metrics. The best scores are **bolded** and the second bests are underlined.† denotes the best score achieved by each LLM architecture.

of 3 cases, while that improves the pointwise performance. Thus we choose pointwise to be our distillation strategy.

### 3.2 Open-Sourced LLM Fine-Tuning

DIRAS data fine-tuned models $\mathcal{M}_o$ will be used to predict all (query, document) combinations to mitigate annotation selection bias (Thakur et al., 2021). Therefore, shorter generations from $\mathcal{M}_o$ (e.g., without CoT) are favored as the inference cost for transformers increases quadratically with generation length. Additionally, the choice of calibration method matters for LLMs (Tian et al., 2023). To explore these aspects, we fine-tune $\mathcal{M}_o$ in four settings: $\mathcal{M}_o$-CoT-Ask, $\mathcal{M}_o$-CoT-Tok, $\mathcal{M}_o$-Ask, $\mathcal{M}_o$-Tok, where CoT means $\mathcal{M}_o$ is tuned to generate [Reason], [Guess], and [Confidence]; without CoT denotes $\mathcal{M}_o$ is tuned to only generate [Guess] and [Confidence]; "Ask" means the result is calibrated by the generated confidence score in [Confidence] field; and "Tok" means we take the token-level probability of "Yes/No" after "[Guess]:" as the confidence score for calibration. The prompt in Fig. 2 is used for fine-tuning. The "[Reason]:" line is removed in settings without CoT.

We fine-tune Llama-3-8B-instruct (AI@Meta, 2024), gemma-7b-it (Team et al., 2024b), and Phi-3-mini-4k-instruct (Abdin et al., 2024) (details in App. F). We compare these fine-tuned models with baselines including GPT-3.5 and GPT-4 using prompt $\mathcal{P}$; the OpenAI embedding models text-embedding-3-small, and text-embedding-3-large; and BGE Gemma reranker[5], a popular LLM-based reranker for general domain. As Fig. 3 shows, fine-tuning improves original models in all settings. Furthermore, Table 2 shows the results of all fine-tuned models in comparison to all baselines. We observe that $\mathcal{M}_o$-Tok outperforms other settings for all LLM architectures. The best setting Llama3-Tok achieves GPT-4 level performance in calibration and IR on unseen questions and reports.

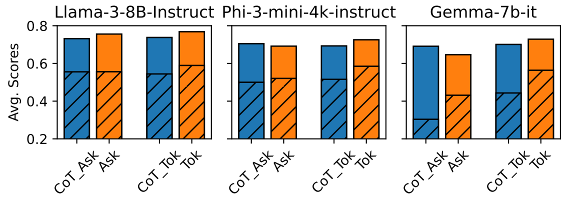


Figure 3: Shaded bars denote the performance of original models. Colored bars denote the improvement brought by fine-tuning.

Interestingly, we find that omitting the chain of thought usually leads to a performance increase for all three LLM architectures. CoT sometimes leads to a limited increase when asking for calibration (Ask), but constantly results in a performance drop when calibrated with token-level probability (Tok). Therefore, $\mathcal{M}_o$ should be fine-tuned without CoT for performance and inference efficiency. Moreover, Tok rarely underperforms Ask, different from Tian et al. (2023)'s finding and our observations in Table 1. Thus, future work may consider probabilities of important tokens (e.g., Yes/No in our prompt template) as a promising calibration tool.

## 4 Experiments on ClimRetrieve

In this section, we showcase how the DIRAS fine-tuned $\mathcal{M}_o$ can assist the IR annotation in a real-world setting, leveraging the ClimRetrieve dataset (Schimanski et al., 2024b). This dataset records analysts' real-life procedure of sustainability report analyses: raise questions about sustainability reports and then go through these reports to find relevant information to answer their questions. The dataset comprises 43k (query, document) pairs, among which 595 pairs are annotated as relevant. Relevant documents are annotated with a relevance score from 1 to 3 (1 translates to partially relevant

[5] https://huggingface.co/BAAI/bge-reranker-v2-gemma

| Setting | nDCG | nDCG@5 | nDCG@10 | nDCG@15 |
|---|---|---|---|---|
| Random | 71.04 | 50.88 | 52.77 | 54.45 |
| Small-embed | 74.52 | 61.28 | 60.36 | 61.69 |
| Large-embed | 76.30 | 63.13 | 63.36 | 64.67 |
| GPT-3.5 | 74.62 | 60.08 | 61.49 | 61.91 |
| GPT-4 | 75.55 | 60.89 | 63.23 | 65.26 |
| Llama3-Ask | **77.23** | **67.60** | 66.18 | **67.57** |
| Llama3-Tok | 76.55 | 67.20 | **66.23** | 65.83 |

Table 3: Performance on ranking the **relevant** (query, document) pairs in ClimRetrieve.

| Setting | nDCG | MAP |
|---|---|---|
| Llama3-Ask$_{generic}$ | 29.95 | 26.51 |
| Llama3-Ask$_{informed}$ | **30.89** | **29.31** |
| Llama3-Tok$_{generic}$ | 31.17 | 28.73 |
| Llama3-Tok$_{informed}$ | **32.53** | **32.65** |

Table 4: Comparison of using the generic and the expert-informed relevance definitions for ranking **all** ClimRetrieve (query, document) pairs.

and 2/3 means relevant). App. I gives a detailed overview of the dataset.

Compared to ChatReport data, ClimRetrieve is a more challenging and realistic setting because: (1) For ChatReport data, we draft explicit relevance definitions and annotate relevance dependent on them. But for ClimRetrieve, the analysts' mental model for what is (ir)relevant for their posed questions is unknown to us. (2) ClimRetrieve records human analysts' real-life workflow of reading the full reports and searching for relevant information, which differs from ChatReport data where annotators are presented with (query, document) pairs as separate data points. Thus, ClimRetrieve only has gold labels for relevant (query, document) pairs. Other not annotated (query, document) combinations might be either irrelevant or a part of annotation selection bias – a widely existing problem in IR annotation (Thakur et al., 2021).

This challenging and realistic setting of ClimRetrieve allows us to investigate the following research questions regarding DIRAS fine-tuned $\mathcal{M}_o$: (1) **RQ1**: Can $\mathcal{M}_o$ understand fine-grained differences in degree of relevance? (2) **RQ2**: Can we improve $\mathcal{M}_o$'s performance through improving relevance definitions? (3) **RQ3**: Can $\mathcal{M}_o$ assists in mitigating annotation selection bias (Thakur et al., 2021)? (4) **RQ4**: Can $\mathcal{M}_o$'s predictions help benchmarking IR algorithms? We use the best $\mathcal{M}_o$ in § 3 (Llama-3 without CoT) to study these RQs.

### 4.1 RQ1: Understanding Fine-Grained Relevance Levels

We first evaluate fine-tuned Llama-3 on 595 gold labels of ClimRetrieve to verify whether it can effectively recover analysts' ranking for relevant content by understanding which documents are more helpful than others. Relevance definitions are drafted with GPT-4 with the same procedure as § 3. We report nDCG[6] scores to measure the ranking performance on ClimRetrieve. Gold labels 1, 2, and 3 are assigned with relevance scores 1/3, 2/3, and 1. Besides OpenAI 3rd generation embedding models, we also have a random baseline where all (query, document) pairs are assigned a random relevance score between 0 and 1. The random baseline results are averaged over 5 random seeds (40 to 44). Importantly, all ClimRetrieve annotations are to some degree relevant, so improvement over the random baseline is challenging as the system needs to understand the trivial different degrees of relevance.

Table 3 presents different systems' performance. There is a clear trend of outperformance of the fine-tuned Llama-3 models in this challenging setting. They also exceed the random baseline by a significant margin, indicating the model correctly understands the fine-grained levels of relevance.

### 4.2 RQ2: Improving Performance through Improving Definitions

We then test whether we can improve $\mathcal{M}_o$'s performance by improving the relevance definition presented in prompts. In § 3 and § 4.1, we use GPT-4 to draft the relevance definitions. To investigate the role of the definitions, we compare two setups: (1) The *generic* relevance definition: the definition drafted by GPT-4. (2) The *informed* relevance definition: we use the labeled texts in the ClimRetrieve dataset as examples to create artificial expert- textitinformed relevance definitions. This way, we simulate the inclusion of expert knowledge to improve the relevance definition (see App. J for details).

After creating the definitions, we repeat predicting the relevance scores with the DIRAS fine-tuned Llama-3. First, we repeat the setup of previous § 4.1 with only human-annotated 595 relevant (query, document) pairs. We find the utilization of expert-informed definitions produces similar results, at most improving nDCG (see App. K). We argue that the definition might only make a significant difference when predicting all (query, document) pairs instead of only the human-annotated ones. The inclusion of specific nuances might espe-

[6]MAP can only measure binary relevance and since we only investigate relevant samples, it cannot be calculated.

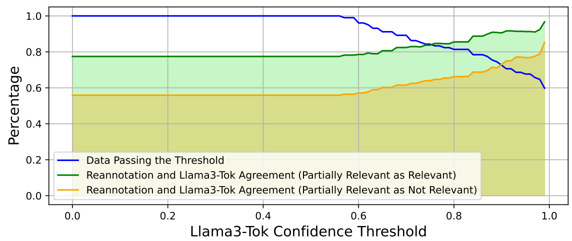


Figure 4: This table shows the percentage of agreement between Llama3-Tok and our human reannotation, and the amount of data remaining, when adjusting the confidence threshold for (query, document) pairs that were considered relevant by Llama3-Tok but not by the ClimRetrieve annotation.

cially help when compared with non-relevant documents. Calculating the nDCG and MAP score for all 43k (query, document) pairs, we find evidence for this notion (see Table 4). Thus, the inclusion of expert-informed definitions seems to improve the performance (for more details, see App. K).

### 4.3 RQ3: Mitigating Annotation Bias

ClimRetrieve employs a real-world analyst scenario. This entails that the human only selectively annotates documents that are likely to be relevant and assumes unannotated documents as irrelevant (see e.g., Thakur et al., 2021). Therefore, the dataset allows us to investigate our model's capabilities to counteract biases. For this purpose, we reannotated 288 strategically sampled (sampling details in App. M) (document, query) pairs to investigate Llama3-Tok's performance on ClimRetrieve's missing annotation. For samples not labeled as relevant in ClimRetrieve but by Llama3-Tok, the reannotations indicate that we can effectively overrule these blind spots. As Fig. 4 shows, increasing the threshold of Llama3-Tok confidence allows us to be increasingly sure that we indeed find a relevant document. For the around 60% of the samples with a Llama3-Tok confidence higher than 99%, almost all samples are reannotated as relevant. Besides annotation selection bias, there are also other reasons for overruling the initial annotations. ClimRetrieve annotators may rather have focused on hard relevant documents omitting partially relevant ones. Furthermore, our GPT-4 created definitions may broaden the scope of what is relevant. However, we find that our model is well-calibrated and consistent with the provided relevance definitions. Therefore, Llama3-Tok is helpful for mitigating annotation selection bias (for a detailed analysis, see App. M).

| Setting | Kendall's $\tau$ |
|---|---|
| BGE-Base | 35.71 |
| BGE-Base-ft | **36.34** |
| BGE-Large | 34.74 |
| BGE-Large-ft | **36.55** |

Table 5: Different embedding models' performance benchmarked by $\mathcal{M}_o$'s prediction on all 43K (query, document) pairs of ClimRetrieve. "ft" denotes the model is fine-tuned on in-domain data.

### 4.4 RQ4: Using $\mathcal{M}_o$ to Benchmark IR

After validating $\mathcal{M}_o$'s performance, we can use it to annotate all 43k ClimRetrieve datapoints and obtain a benchmarking dataset to select IR algorithms. This approach can be especially helpful when lacking human annotation and annotation selection bias is prevalent. Specifically, we compare the performance of embedding models before and after in-domain fine-tuning. If the $\mathcal{M}_o$-annotated benchmark gives higher scores to the fine-tuned checkpoints, that means it is capable of selecting a better model for this specific domain.

For this experiment, we first fine-tune open-sourced embedding models bge-large-en-v1.5 and bge-base-en-v1.5 (Chen et al., 2024) on ChatReport test set[7] (fine-tuning details in App. L). We then compare embedding models' relevance ranking with the predicted ranking of Llama3-Tok on all 43K (query, document) pairs in ClimRetrieve. We use Kendall's $\tau$ as the metric, which directly compares the correlation between two ranks. The results are shown in Table 5. We find the Llama3-Tok-annotated benchmark successfully picks out the fine-tuned checkpoints, showing a capability of benchmarking information retrieval algorithms.

Interestingly, the unfine-tuned BGE-Base correlates more to Llama3-Tok compared to BGE-Large, although the latter shows stronger performance on MTEB (Muennighoff et al., 2023). This indicates the necessity of domain-specific benchmarking to tell the in-domain performance.

## 5 Recommendation for Future RAG

**Avoiding Top-K Retrieval**: Naive RAG systems (Ni et al., 2023) usually retrieve top-k (a fixed number k) documents to augment LLM generation. However, different questions tend to have different

[7] We fine-tune on the test instead of the training set to (1) leverage high-quality human annotation for fine-tuning; and (2) avoid indirect data leakage as $\mathcal{M}_o$ is fine-tuned on the training set.

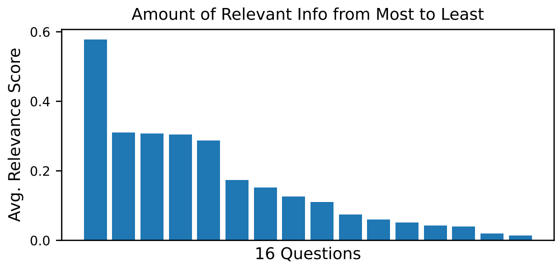


Figure 5: The proximate amount of relevant information for 16 questions in all ClimRetrieve reports, according to Llama3-Tok's relevance scores.

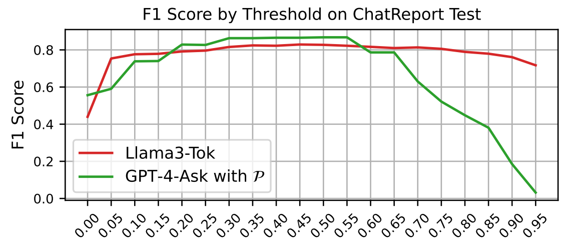


Figure 6: Instead of always retrieving top-k, we can retrieve documents if they have relevance scores higher than a threshold. This figure shows the change of F1 scores for obtaining relevant documents by thresholds.

amounts of relevant information.Advanced RAG employs query routers to pick retrieval strategies (Gao et al., 2024). However, choosing the proper k without access to full documents is still hard. To demonstrate this, we average the relevance score (predicted by Llama3-Tok) over all documents for each question in ClimRetrieve. The resulting average relevance score will be a proxy for the amount of relevant information on the question in all reports. As Fig. 5 shows, different questions vary considerably in the amount of relevant information. Therefore, we suggest not using top-k IR, avoiding the prior determined k that does not fit the actual amount of relevant information.

Given the calibrated prediction of DIRAS fine-tuned $\mathcal{M}_o$, an alternative way is to retrieve all documents whose relevance scores exceed a pre-defined threshold. Thus, different questions can retrieve different amounts of information depending on whether passing the relevance threshold. Advanced RAG designs can even strategically pick the calibrated threshold for different questions, for example, allowing more partial relevance for summary queries. Fig. 6 shows the F1 Scores of GPT-4 and Llama3-Tok with different relevance thresholds. Llama3-Tok achieves good F1 scores over a wide range of thresholds. Thanks to its compact size (8B), it can be efficiently deployed as a reranker in RAG systems.

**Optimizing Relevance Definitions**: Results in Table 2 and Table 3 are obtained with GPT-4-drafted relevance definitions (i.e., relevance definitions). Although this approach is useful in large-scale applications, there is still space for improvement by optimizing relevance definition, as shown in § 4.2. According to Bailey et al. (2008), the question originators are the gold standard for relevance definition. Hence, with the help of DIRAS, future RAG systems may allow users to customize their requirements for relevant information.

## 6 Related Literature

IR plays an important role in RAG but also becomes a performance bottleneck (Gao et al., 2024). Low precision in IR may cause LLMs to hallucinate or pick up irrelevant information (Cuconasu et al., 2024; Schimanski et al., 2024a). Low recall may leave out critical information for analysis (Ni et al., 2023). Prior work proposes various datasets and algorithms to benchmark RAG performance (Saad-Falcon et al., 2023; Chen et al., 2023a; Niu et al., 2024), but most of them focus on the final generation quality. No previous work studies how to efficiently generate domain-specific IR benchmarks for RAG. Sun et al. (2023b,a); Pradeep et al. (2023); Qin et al. (2024) find that SOTA generic LLMs are good rerankers and such ability can be distilled to open-sourced LLMs. However, they all focus on pairwise or listwise ranking methods, while our work shows that the pointwise method (1) better fulfills the need of annotating domain-specific IR; and (2) works better with proper calibration method (Tian et al., 2023). Our work follows the stream of research annotating document relevance with LLMs (Thomas et al., 2024), taking one step further by (1) showing how to annotate with small LLMs; (2) predicting relevance scores to address partial relevance; and (3) use them to benefit RAG development.

## 7 Conclusion

In this work, we introduce the DIRAS pipeline to fine-tune open-source LLMs to calibrated annotators. We prove the effectiveness of the approach on two dataset sets. The DIRAS approach has two significant advantages: (1) it is case-specialised allowing the incorporation of domain-specific knowledge into definitions, and (2) it helps to efficiently label a huge amount of documents with calibrated relevance scores.

## Limitations

As with every work, this has limitations. Our first limitation stems from the usage of two datasets focusing on a scenario of RAG report analyses. Given our expertise, this allows us to extend the deepness of our investigations: annotating domain-specific benchmarks and conducting error analyses. However, this limits the wideness of our research. While we argue that the results are representative of other knowledge-intensive RAG scenarios, it remains an open question for future work to generalize the DIRAS pipeline.

Second, this project focuses on text documents. This means we do not evaluate the performance of the DIRAS pipeline on graph and table content. While this also presents a general limitation of modern-day RAG systems, we believe it is a crucial future step to generalize DIRAS's idea of scalable information retrieval benchmarking to multi-modality.

Our third limitation, and also a viable option to address multimodality, lies in the recent introduction of long-context LLMs. These may make the role of information retrieval in RAG less crucial as entire documents can be used to answer a question. At the same time, we observe that long-context models are good in needle-in-a-haystack problems but not as good when multiplied needles exist (Team et al., 2024a). Thus, even for long-context LLMs, an efficient system like DIRAS could enable improving algorithms for finding and using multiple relevant pieces of information or help improve the model's ability to do so.

## Ethics Statement

**Human Annotation**: In this work, all human annotators are Graduate, Doctorate researchers, or Professors who have good knowledge about scientific communication and entailment. They are officially hired and have full knowledge of the context and utility of the collected data. We adhered strictly to ethical guidelines, respecting the dignity, rights, safety, and well-being of all participants.

**Data Privacy or Bias**: There are no data privacy issues or biases against certain demographics with regard to the data collected from real-world applications and LLM generations. All artifacts we use are under a creative commons license. We also notice no ethical risks associated with this work

**Reproducibility Statement**: To ensure full reproducibility, we will disclose all codes and data used in this project, as well as the LLM generations, GPT-4 and human annotations. For OpenAI models, we use "gpt-4-0125-preview" and "gpt-3.5-turbo-0125". We always fix the temperature to 0 when using APIs.

## References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Qin Cai, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Yen-Chun Chen, Yi-Ling Chen, Parul Chopra, Xiyang Dai, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Victor Fragoso, Dan Iter, Mei Gao, Min Gao, Jianfeng Gao, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Ce Liu, Mengchen Liu, Weishung Liu, Eric Lin, Zeqi Lin, Chong Luo, Piyush Madan, Matt Mazzola, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Xin Wang, Lijuan Wang, Chunyu Wang, Yu Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Haiping Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Jianwei Yang, Ziyi Yang, Yifan Yang, Donghan Yu, Lu Yuan, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone.

AI@Meta. 2024. Llama 3 model card.

Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, Arjen P. de Vries, and Emine Yilmaz. 2008. Relevance assessment: are judges exchangeable and does it matter. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, page 667–674, New York, NY, USA. Association for Computing Machinery.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2023a. Benchmarking Large Language Models in Retrieval-Augmented Generation. ArXiv:2309.01431 [cs].

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. 2023b. Dense X Retrieval: What Retrieval Granularity Should We Use?

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The Power of Noise: Redefining Retrieval for RAG Systems. ArXiv:2401.14887 [cs].

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. ArXiv:2312.10997 [cs].

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language Models (Mostly) Know What They Know. ArXiv:2207.05221 [cs].

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models.

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.

Jingwei Ni, Minjing Shi, Dominik Stammbach, Mrinmaya Sachan, Elliott Ash, and Markus Leippold. 2024. Afacta: Assisting the annotation of factual claim detection with reliable llm annotators.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. Rankzephyr: Effective and robust zero-shot listwise reranking is a breeze!

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. 2024. Large language models are effective text rankers with pairwise ranking prompting.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2023. Ares: An automated evaluation framework for retrieval-augmented generation systems.

Tefko Saracevic. 2008. Effects of inconsistent relevance judgments on information retrieval test results: A historical perspective. *Library Trends*, 56:763 – 783.

Tobias Schimanski, Jingwei Ni, Mathias Kraus, Elliott Ash, and Markus Leippold. 2024a. Towards faithful and robust llm specialists for evidence-based question-answering.

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

Weiwei Sun, Zheng Chen, Xinyu Ma, Lingyong Yan, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023a. Instruction distillation makes large language models efficient zero-shot rankers.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023b. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937, Singapore. Association for Computational Linguistics.

DeepSearch Team. 2022. Deep Search Toolkit.

Gemini Team, Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry, Lepikhin, Timothy Lillicrap, Jean baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, Luke Vilnis, Oscar Chang, Nobuyuki Morioka, George Tucker, Ce Zheng, Oliver Woodman, Nithya Attaluri, Tomas Kocisky, Evgenii Eltyshev, Xi Chen, Timothy Chung, Vittorio Selo, Siddhartha Brahma, Petko Georgiev, Ambrose Slone, Zhenkai Zhu, James Lottes, Siyuan Qiao, Ben Caine, Sebastian Riedel, Alex Tomala, Martin Chadwick, Juliette Love, Peter Choy, Sid Mittal, Neil Houlsby, Yunhao Tang, Matthew Lamm, Libin Bai, Qiao Zhang, Luheng He, Yong Cheng, Peter Humphreys, Yujia Li, Sergey Brin, Albin Cassirer, Yingjie Miao, Lukas Zilka, Taylor Tobin, Kelvin Xu, Lev Proleev, Daniel Sohn, Alberto Magni, Lisa Anne Hendricks, Isabel Gao, Santiago Ontanon, Oskar Bunyan, Nathan Byrd, Abhanshu Sharma, Biao Zhang, Mario Pinto, Rishika Sinha, Harsh Mehta, Dawei Jia, Sergi Caelles, Albert Webson, Alex Morris, Becca Roelofs, Yifan Ding, Robin Strudel, Xuehan Xiong, Marvin Ritter, Mostafa Dehghani, Rahma Chaabouni, Abhijit Karmarkar, Guangda Lai, Fabian Mentzer, Bibo Xu, YaGuang Li, Yujing Zhang, Tom Le Paine, Alex Goldin, Behnam Neyshabur, Kate Baumli, Anselm Levskaya, Michael Laskin, Wenhao Jia, Jack W. Rae, Kefan Xiao, Antoine He, Skye Giordano, Lakshman Yagati, Jean-Baptiste Lespiau, Paul Natsev, Sanjay Ganapathy, Fangyu Liu, Danilo Martins, Nanxin Chen, Yunhan Xu, Megan Barnes, Rhys May, Arpi Vezer, Junhyuk Oh, Ken Franko, Sophie Bridgers, Ruizhe Zhao, Boxi Wu, Basil Mustafa, Sean Sechrist, Emilio Parisotto, Thanumalayan Sankaranarayana Pillai, Chris Larkin, Chenjie Gu, Christina Sorokin, Maxim Krikun, Alexey Guseynov, Jessica Landon, Romina Datta, Alexander Pritzel, Phoebe Thacker, Fan Yang, Kevin Hui, Anja Hauth, Chih-Kuan Yeh, David Barker, Justin Mao-Jones, Sophia Austin, Hannah Sheahan, Parker Schuh, James Svensson, Rohan Jain, Vinay Ramasesh, Anton Briukhov, Da-Woon Chung, Tamara von Glehn, Christina Butterfield, Priya Jhakra, Matthew Wiethoff, Justin Frye, Jordan Grimstad, Beer Changpinyo, Charline Le Lan, Anna Bortsova, Yonghui Wu, Paul Voigtlaender, Tara Sainath, Shane Gu, Charlotte Smith, Will Hawkins, Kris Cao, James Besley, Srivatsan Srinivasan, Mark Omernick, Colin Gaffney, Gabriela Surita, Ryan Burnell, Bogdan Damoc, Junwhan Ahn, Andrew Brock, Mantas Pajarskas, Anastasia Petrushkina, Seb Noury, Lorenzo Blanco, Kevin Swersky, Arun Ahuja, Thi Avrahami, Vedant Misra, Raoul de Liedekerke, Mariko Iinuma, Alex Polozov, Sarah York, George van den Driessche, Paul Michel, Justin Chiu, Rory Blevins, Zach Gleicher, Adrià Recasens, Alban Rrustemi, Elena Gribovskaya, Aurko Roy, Wiktor Gworek, Sébastien M. R. Arnold, Lisa Lee, James Lee-Thorp, Marcello Maggioni, Enrique Piqueras, Kartikeya Badola, Sharad Vikram, Lucas Gonzalez, Anirudh Baddepudi, Evan Senter, Jacob Devlin, James Qin, Michael Azzam, Maja Trebacz, Martin Polacek, Kashyap Krishnakumar, Shuo yiin Chang, Matthew Tung, Ivo Penchev, Rishabh Joshi, Kate Olszewska, Carrie Muir, Mateo Wirth, Ale Jakse Hartman, Josh Newlan, Sheleem Kashem, Vijay Bolina, Elahe Dabir, Joost van Amersfoort, Zafarali Ahmed, James Cobon-Kerr, Aishwarya Kamath, Arnar Mar Hrafnkelsson, Le Hou, Ian Mackinnon, Alexandre Frechette, Eric Noland, Xiance Si, Emanuel Taropa, Dong Li, Phil Crone, Anmol Gulati, Sébastien Cevey, Jonas Adler, Ada Ma, David Silver, Simon Tokumine, Richard Powell, Stephan Lee, Kiran Vodrahalli, Samer Hassan, Diana Mincu, Antoine Yang, Nir Levine, Jenny Brennan, Mingqiu Wang, Sarah Hodkinson, Jeffrey Zhao, Josh Lipschultz, Aedan Pope, Michael B. Chang, Cheng Li, Laurent El Shafey, Michela Paganini, Sholto Douglas, Bernd Bohnet, Fabio Pardo, Seth Odoom, Mihaela Rosca, Cicero Nogueira dos Santos, Kedar Soparkar, Arthur Guez, Tom Hudson, Steven Hansen, Chulayuth Asawaroengchai, Ravi Addanki, Tianhe Yu, Wojciech Stokowiec, Mina Khan, Justin Gilmer, Jaehoon Lee, Carrie Grimes Bostock, Keran Rong, Jonathan Caton, Pedram Pejman, Filip Pavetic, Geoff Brown, Vivek Sharma, Mario Lučić, Rajkumar Samuel, Josip Djolonga, Amol Mandhane, Lars Lowe Sjösund, Elena Buchatskaya, Elspeth White, Natalie Clay, Jiepu Jiang, Hyeontaek Lim, Ross Hemsley, Zeyncep Cankara, Jane Labanowski, Nicola De Cao, David Steiner, Sayed Hadi Hashemi, Jacob Austin, Anita Gergely, Tim Blyth, Joe Stanton, Kaushik Shivakumar, Aditya Siddhant, Anders Andreassen, Carlos Araya, Nikhil Sethi, Rakesh Shivanna, Steven Hand, Ankur Bapna, Ali Khodaei, Antoine Miech, Garrett Tanzer, Andy Swing, Shantanu Thakoor, Lora Aroyo, Zhufeng Pan, Zachary Nado, Jakub Sygnowski, Stephanie Winkler, Dian Yu, Mohammad Saleh, Loren Maggiore, Yamini Bansal, Xavier Garcia, Mehran Kazemi, Piyush Patil, Ishita Dasgupta, Iain Barr, Minh Giang, Thais Kagohara, Ivo Danihelka, Amit Marathe, Vladimir Feinberg, Mohamed Elhawaty, Nimesh Ghelani, Dan Horgan, Helen Miller, Lexi Walker, Richard Tanburn, Mukarram Tariq, Disha Shrivastava, Fei Xia, Qingze Wang, Chung-Cheng Chiu, Zoe Ashwood, Khuslen Baatarsukh, Sina Samangooei, Raphaël Lopez Kaufman, Fred Alcober, Axel Stjerngren, Paul Komarek, Katerina Tsihlas, Anudhyan Boral, Ramona Comanescu, Jeremy Chen, Ruibo Liu, Chris Welty, Dawn Bloxwich, Charlie Chen, Yanhua Sun, Fangxiaoyu Feng, Matthew Mauger, Xerxes Dotiwalla, Vincent Hellendoorn, Michael Sharman, Ivy Zheng, Krishna Haridasan, Gabe Barth-Maron, Craig Swanson, Dominika Rogozińska, Alek Andreev, Paul Kishan Rubenstein, Ruoxin Sang, Dan Hurt, Gamaleldin Elsayed, Ren-shen Wang, Dave Lacey, Anastasija Ilić, Yao Zhao, Adam Iwanicki, Alejandro Lince, Alexander Chen, Christina Lyu, Carl Lebsack, Jordan Griffith, Meenu Gaba, Paramjit Sandhu, Phil Chen, Anna Koop, Ravi Rajwar, Soheil Hassas Yeganeh, Solomon Chang, Rui Zhu, Soroush Radpour, Elnaz Davoodi, Ving Ian Lei, Yang Xu, Daniel Toyama, Constant Segal, Martin Wicke, Hanzhao Lin, Anna Bulanova, Adrià Puigdomènech Badia, Nemanja Rakićević, Pablo Sprechmann, Angelos Filos, Shaobo Hou, Víctor Campos, Nora Kassner, Devendra Sachan, Meire Fortunato, Chimezie Iwuanyanwu, Vitaly Nikolaev, Balaji Lakshminarayanan, Sadegh Jazayeri, Mani Varadarajan, Chetan Tekur, Doug Fritz, Misha Khalman, David Reitter, Kingshuk Dasgupta, Shourya Sarcar, Tina Ornduff, Javier Snaider, Fantine Huot, Johnson Jia, Rupert Kemp, Nejc Trdin, Anitha Vijayakumar, Lucy Kim, Christof Angermueller, Li Lao, Tianqi Liu, Haibin Zhang, David Engel, Somer Greene, Anaïs White, Jessica Austin, Lilly Taylor, Shereen Ashraf, Dangyi Liu, Maria Georgaki, Irene Cai, Yana Kulizhskaya, Sonam Goenka, Brennan Saeta, Ying Xu, Christian Frank, Dario de Cesare, Brona Robenek, Harry Richardson, Mahmoud Alnahlawi, Christopher Yew, Priya Ponnapalli, Marco Tagliasacchi, Alex Korchemniy, Yelin Kim, Dinghua Li, Bill Rosgen, Kyle Levin, Jeremy Wiesner, Praseem Banzal, Praveen Srinivasan, Hongkun Yu, Çağlar Ünlü, David Reid, Zora Tung, Daniel Finchelstein, Ravin Kumar, Andre Elisseeff, Jin Huang, Ming Zhang, Ricardo Aguilar, Mai Giménez, Jiawei Xia, Olivier Dousse, Willi Gierke, Damion Yates, Komal Jalan, Lu Li, Eri Latorre-Chimoto, Duc Dung Nguyen, Ken Durden, Praveen Kallakuri, Yaxin Liu, Matthew Johnson, Tomy Tsai, Alice Talbert, Jasmine Liu, Alexander Neitz, Chen Elkind, Marco Selvi, Mimi Jasarevic, Livio Baldini Soares, Albert Cui, Pidong Wang, Alek Wenjiao Wang, Xinyu Ye, Krystal Kallarackal, Lucia Loher, Hoi Lam, Josef Broder, Dan Holtmann-Rice, Nina Martin, Bramandia Ramadhana, Mrinal Shukla, Sujoy Basu, Abhi Mohan, Nick Fernando, Noah Fiedel, Kim Paterson, Hui Li, Ankush Garg, Jane Park, DongHyun Choi, Diane Wu, Sankalp Singh, Zhishuai Zhang, Amir Globerson, Lily Yu, John Carpenter, Félix de Chaumont Quitry, Carey Radebaugh, Chu-Cheng Lin, Alex Tudor, Prakash Shroff, Drew Garmon, Dayou Du, Neera Vats, Han Lu, Shariq Iqbal, Alex Yakubovich, Nilesh Tripuraneni, James Manyika, Haroon Qureshi, Nan Hua, Christel Ngani, Maria Abi Raad, Hannah Forbes, Jeff Stanway, Mukund Sundararajan, Victor Ungureanu, Colton Bishop, Yunjie Li, Balaji Venkatraman, Bo Li, Chloe Thornton, Salvatore Scellato, Nishesh Gupta, Yicheng Wang, Ian Tenney, Xihui Wu, Ashish Shenoy, Gabriel Carvajal, Diana Gage Wright, Ben Bariach, Zhuyun Xiao, Peter Hawkins, Sid Dalmia, Clement Farabet, Pedro Valenzuela, Quan Yuan, Ananth Agarwal, Mia Chen, Wooyeol Kim, Brice Hulse, Nandita Dukkipati, Adam Paszke, Andrew Bolt, Kiam Choo, Jennifer Beattie, Jennifer Prendki, Harsha Vashisht, Rebeca Santamaria-Fernandez, Luis C. Cobo, Jarek Wilkiewicz, David Madras, Ali Elqursh, Grant Uy, Kevin Ramirez, Matt Harvey, Tyler Liechty, Heiga Zen, Jeff Seibert, Clara Huiyi Hu, Andrey Khorlin, Maigo Le, Asaf Aharoni, Megan Li, Lily Wang, Sandeep Kumar, Norman Casagrande, Jay Hoover, Dalia El Badawy, David Soergel, Denis Vnukov, Matt Miecnikowski, Jiri Simsa, Praveen Kumar, Thibault Sellam, Daniel Vlasic, Samira Daruki, Nir Shabat, John Zhang, Guolong Su, Jiageng Zhang, Jeremiah Liu, Yi Sun, Evan Palmer, Alireza Ghaffarkhah, Xi Xiong, Victor Cotruta, Michael Fink, Lucas Dixon, Ashwin Sreevatsa, Adrian Goedeckemeyer, Alek Dimitriev, Mohsen Jafari, Remi Crocker, Nicholas FitzGerald, Aviral Kumar, Sanjay Ghemawat, Ivan Philips, Frederick Liu, Yannie Liang, Rachel Sterneck, Alena Repina, Marcus Wu, Laura Knight, Marin Georgiev, Hyo Lee, Harry Askham, Abhishek Chakladar, Annie Louis, Carl Crous, Hardie Cate, Dessie Petrova, Michael Quinn, Denese Owusu-Afriyie, Achintya Singhal, Nan Wei, Solomon Kim, Damien Vincent, Milad Nasr, Christopher A. Choquette-Choo, Reiko Tojo, Shawn Lu, Diego de Las Casas, Yuchung Cheng, Tolga Bolukbasi, Katherine Lee, Saaber Fatehi, Rajagopal Ananthanarayanan, Miteyan Patel, Charbel Kaed, Jing Li, Shreyas Rammohan Belle, Zhe Chen, Jaclyn Konzelmann, Siim Põder, Roopal Garg, Vinod Koverkathu, Adam Brown, Chris Dyer, Rosanne Liu, Azade Nova, Jun Xu, Alanna Walton, Alicia Parrish, Mark Epstein, Sara McCarthy, Slav Petrov, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2024a. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024b. Gemma: Open models based on gemini research and technology.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. ArXiv:2104.08663 [cs].

Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. Large language models can accurately predict searcher preferences. ArXiv:2309.10621 [cs].

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. ArXiv:2305.14975 [cs].

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

# A Exemplifying Partial Relevance

When labeling whether a document is relevant for a question, there exists a large grey scale of relevance rather than a black-and-white relevant or irrelevant label. Humans can only consistently capture these nuances to a certain extent. The judgment of relevance also depends on the context and the annotator's domain expertise.

Consider for instance the following excerpt of a document:

```
"""
[...] Implement Risk Controls: Integrated Management
     System (IMS): The K&S Integrated Management
     System (IMS), which has been implemented at our
     six major design and manufacturing sites, is
     certified under the corporate ISO 9001:2015,
     ISO 14001:2015 and ISO 45001:2018
     certifications. Our integrated Quality,
     Environmental and Occupational Health & Safety
     (QEHS) Management System enables the
     achievement of harmonized K&S worldwide
     objectives.
"""
```

Furthermore, conclude as to whether this is relevant to answer the following question and definition:

```
"""
Meaning of the question: The question "What
     processes does the organization use to identify
     and assess climate-related risks?" is asking
     for information about the specific methods,
     tools, or strategies that a company employs to
     recognize and evaluate the potential risks to
     its operations, financial performance, and
     overall sustainability that are associated with
     climate change. This includes understanding
     how the organization anticipates, quantifies,
     and plans for the impacts of climate-related
     phenomena such as extreme weather events, long-
     term shifts in climate patterns, and regulatory
     changes aimed at mitigating climate change.

Examples of information that the question is looking
     for:
1. The use of climate risk assessment tools or
     software that helps in modeling and predicting
     potential impacts of climate change on the
     organization's operations.
2. Engagement with external consultants or experts
     specializing in climate science [...]
"""
```

The question is clearly looking for processes to identify and assess risks associated with climate change. Example 1. states that "climate risk assessment tools" are relevant. The paragraph states that the Integrated Management System serves to identify risks including environmental risks. In sustainability matters, climate change and environmental topics often fall under the same umbrella. Thus, yes, the paragraph is relevant for the question addressing a certified process to manage climate risks. However, also contrary arguments can be considered. We don't exactly know whether environmental and climate topics are viewed interchangeably. An expert may know clear differentiating factors between environmental and climate matters (e.g., not all environmental problems like water pollution affect the climate). Furthermore, the environmental management system is rather a minor note in this paragraph. Additionally, it seems that, although it is a general risk management system, the "Quality, Environmental and Occupational Health & Safety (QEHS) Management System" is rather used to achieve worldwide objectives for the company. Would you deem this relevant if it was the only information obtained for a company? And what if there are fifteen more documents that are clearly relevant? How would it be labeled then? It is possible to go to lengths and depending on which expert level or context a labeler holds. In a binary relevant/irrelevant setting, both labels would be partially wrong. The reason lies in the fact that when asking whether this document is relevant to the question, the answer is "partially right".

# B Creation of Question Relevance Definition

Fig. 7 shows the prompt template for the creation of the question relevance definition. We ask the model to produce a short definition on which the model should rely. Additionally, we ask the model to produce a list of examples. This structure should align with the manner an expert implicity or explicitly approaches the annotation task of labeling relevance. A definition alone would have the shortcoming that it only incorporated generic know-how. Complementing it with examples gives the expert the flexibility to extend the meaning of the terms in exemplified form. For a demonstration of the output, see Table 10.

```
"""
An analyst posts a <question> about a sustainability
      report. Your task is to explain the <question>
      in the context of sustainability reporting.
     Please first explain the meaning of the <
     question>, i.e., the meaning of the question
     itself and the concepts mentioned. And then
     give a list of examples, showing what
     information from the sustainability report the
     analyst is looking for by posting this <
     question>.

For <the question's meaning>, please start by
     repeating the question in the following format:
'''
The question "<question>" is asking for information
     about [...]
'''

For the <list of example information that the
     question is looking for>, follow the following
     example in terms of format:
---
[...]
3. Initiatives aimed at creating new job
     opportunities in the green economy within the
     company or in the broader community.
4. Policies or practices in place to ensure that the
      transition to sustainability is inclusive,
     considering gender, race, and economic status.
[...]
---

Here is the question:
<question>: ""{question}""

Format your reply in the following template and keep
      your answer concise:

Meaning of the question: <the question's meaning>
Examples of information that the question is looking
      for: <list of example information that the
     question is looking for>"""
```

Figure 7: RAG Prompt Template enforcing structured output.

## C Metrics Computation Details

In this project, we use Scikit-Learn (version 1.2.2) to compute AUROC, average precision scores, Brier scores, and F1 scores. We employ rank_eval (version 0.1.3) to compute nDCG and MAP scores, and Scipy.stats to compute Kendall's $\tau$. For nDCG, relevant scores 0.5 and 1 are assigned to partially relevant and relevant documents correspondingly.

## D PDF Parsing and Document Length

We use IBM deepsearch parser (Team, 2022) to parse corporate reports into chunks. For chunks shorter than 120 tokens, we concatenate them with adjacent chunks to form chunks longer than 120.

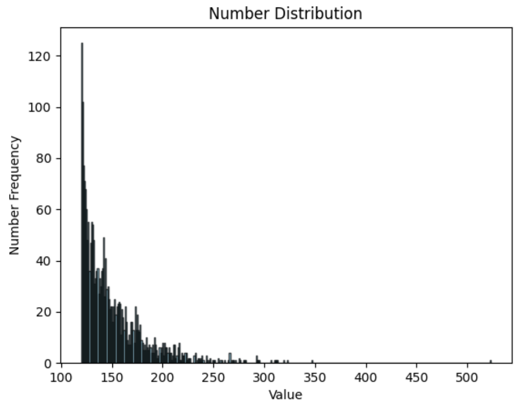


Figure 8: Distribution of chunk length after being extracted from sustainability reports and concatenation.

Figure Fig. 8 shows the formatted chunks length distribution.

## E Expert Annotation Process

To obtain an expert-annotated test set, we annotate the test data that contains equally 330 samples of each top-5 relevant and non-to-5 relevant documents to the 11 questions/queries, thus obtaining 660 (query, document) pairs including the corresponding expert definitions. As described in App. D, the data is obtained from real sustainability reports and split into chunks of around 150 words with the IBM deepsearch parser (Team, 2022). Table 6 shows an overview of statistical properties of the number of words in test set data.

Then, we form a group of three expert annotators. The expert annotators comprise one graduate and one PhD student working in NLP for climate change. These two experts label the entire dataset with three labels: the document is relevant, partially relevant, or not relevant for the query including the definition. Following a simple annotation guideline:

- Please first carefully read the provided relevance definition to understand what the question is looking for. The definition consists of a question explanation and examples of relevant information.

- If a paragraph clearly fall into the definition of relevance, i.e., explicitly mentioned by the question explanation or examples, please annotate relevant.

- If the paragraph is not explicitly covered by the definition but you think it somehow helps

| | Number of words per document | | | | | | |
|---|---|---|---|---|---|---|---|
| **Dataset Size** | **Mean** | **Std** | **Min** | **25%** | **50%** | **75%** | **Max** |
| 660 | 150 | 28.5 | 107 | 131 | 143 | 162 | 318 |

Table 6: Statistical properties of the number of words in ChatReport test set data.

| Label | Occurance |
|---|---|
| Relevant | 121 |
| Partially | 65 |
| Not Relevant | 474 |

Table 7: Label distribution in the ChatReport test dataset.

answering the question. Please annotate partially relevant.

- Otherwise please annotate irrelevant.

Additionally, one PhD student focusing on climate change and sustainability research serves as a subject-matter meta annotator to resolve conflicts or investigate cases where both labelers arise at the label partially.

Comparing the two base annotators in the setup, we can calculate inter-annotator agreement. The Cohen's kappa between the two labelers is 0.683 (substantial agreement). We also calculate annotators' agreement on partial relevance. The Cohen's Kappa turns out to be 0.129, suggesting that there are uncertainty and subjectivity associated with partial labels.

Besides the relevance, we also obtain an uncertainty label whenever there is strong disagreement (co-existence of relevance and irrelevance labels) or agreement on partial relevance (two or more annotators agree on partial relevance), the data point is labeled as uncertain. There are 103 (557) uncertain (certain) (query, document) pairs in the dataset.

Finally, the third expert annotator resolves the existing conflicts in the dataset. This results in a label distribution of Table 7. It becomes apparent that the majority of documents are not relevant while still a significant number is labeled as partially relevant and relevant.

# F LLM Fine-Tuning Settings

We use the default QLoRA hyperparameter settings [8], namely, an effective batch size of 32, a lora r of 64, a lora alpha of 16, a warmup ratio of 0.03, a constant learning rate scheduler, a learning rate of 0.0002, an Adam beta2 of 0.999, a max gradient norm of 0.3, a LoRA dropout of 0.1, 0 weight decay, a source max length of 2048, and a target max length of 512. We use LoRA module on all linear layers. All fine-tunings last 2 epochs.

All experiments are conducted on two clusters, one with 4 V100 GPUs and the other with 4 A100 (80G) GPUs. 1 GPU hour is used per fine-tuning.

[8]https://github.com/jondurbin/qlora

# G DIRAS Prompt Template $\mathcal{P}$

Fig. 9 shows the full prompt DIRAS prompt template for the Chain-of-Thought setup. The non-CoT setup just excludes the "[Reason]: ..."" part of the prompt.

```
You are a helpful assistant who assists human
     analysts in identifying useful information
     within sustainability reports for their
     analysis.

You will be provided with a <question> the analyst
     seeks to answer, a <paragraph> extracted from a
      lengthy report, and <background_information>
     that explains the <question>. <
     background_information> first explains the <
     question> and then raises examples to help you
     to better understand the <question>. Your job
     is to assess whether the <paragraph> is useful
     in answering the <question>.

<background_information>: "{background_information}"
<question>: "{question}"
<paragraph>: "{paragraph_chunk}"


Is <paragraph> helpful for answering <question>?
     Note that the <paragraph> can be helpful even
     it only addresses part of the <question>
     without fully answering it. Provide your best
     guess for this question and your confidence
     that the guess is correct. Reply in the
     following format:
[Reason]: <Reason why and how the paragraph is
     helpful or not helpful for answering the
     question. Clearly indicate your stance.>
[Guess]: <Your most likely guess, should be one of "
     Yes" or "No".>
[Confidence]: <Give your honest confidence score
     between 0.0 and 1.0 about the correctness of
     your guess. 0 means your previous guess is very
      likely to be wrong, and 1 means you are very
     confident about the guess.>
```

Figure 9: Full DIRAS Chain-of-Thought prompt for LLMs predicting relevance labels and calibrating.

# H Alternative Prompts

Fig. 9, Fig. 10, and Fig. 11 show the alternative prompts with which we experimented.

```
{Same task description and inputs}

Is <paragraph> helpful for answering <question>?
     Note that the <paragraph> can be helpful even
     it only addresses part of the <question>
     without fully answering it. Provide your best
     guess for this question and the probability
     that the <paragraph> is helpful. Reply in the
     following format:
[Reason]: <Reason why and how the paragraph is
     helpful or not helpful for answering the
     question. Clearly indicate your stance.>
[Guess]: <Your most likely guess, should be one of "
     Yes" or "No".>
[Probability Helpful]: <The probability between 0.0
     and 1.0 that the <paragraph> is helpful to the
     <question>. 0.0 is completely unhelpful and 1.0
      is completely helpful.>
```

Figure 10: Output requirements for the alternative prompt setting $\mathcal{P}_{prob}$. Task description and input are the same as Fig. 9.

```
You will be provided with a <question> the analyst
     seeks to answer, and a <paragraph> extracted
     from a lengthy report. Your job is to assess
     whether the <paragraph> is useful in answering
     the <question>.

<question>: "{question}"
<paragraph>: "{paragraph_chunk}"

{Same output requirements}
```

Figure 11: Task description and input part for the alternative prompt setting $\mathcal{P}_{w/o\_e}$. Output requirements are the same as Fig. 9.

# I ClimRetrieve Dataset Overview

The ClimRetrieve dataset simulates the typical tasks of a sustainability analyst. The annotators examine 30 sustainability reports. We use the report-level dataset of this paper which contains 43K (query, document) pairs labelled by relevance. This dataset is very long because every report is repeated per question to have unique (query, document) pairs. We also use the much shorter, only relevant (query, document) pairs containing 595 unique samples. Since the annotators only search for relevant information, these are the gold labels. There is no active labeling of the irrelevant (query, document) pairs. Table 8 offers a comparison of the statistical properties of the word count of the documents in ClimRetrieve. They are slightly longer than the ChatReport data (compare Table 6). Table 9 shows the label distribution in the relevant-only dataset. It becomes apparent that most of the (query, document) pairs are very relevant (label=3). This aligns with the report analyst setting where information is searched for until the question can effectively be answered.

```
<|system|>
You are RankLLM, an intelligent assistant that can
     rank passages based on their relevancy to the
     query.
<|user|>
I will provide you with {num} passages, each
     indicated by a numerical identifier [].
Rank the passages based on their relevance to the
     search query: {query}.
{passages}
Search Query: {query}.
Rank the {num} passages above based on their
     relevance to the search query. All the passages
      should be included and listed using
     identifiers, in descending order of relevance.
     The output format should be [] > [], e.g., [4]
     > [2]. Only respond with the ranking results,
     do not say any word or explain.
```

Figure 12: We use exactly the same listwise ranking prompt as Sun et al. (2023b) and Pradeep et al. (2023). Both system and user prompts are presented in this figure.

```
<|system|>
You are RankLLM, an intelligent assistant that can
     rank passages based on their relevancy to the
     query.
<|user|>
I will provide you with {num} passages, each
     indicated by a numerical identifier [].
Rank the passages based on their relevance to the
     search query: {query}.
{passages}
Search Query: {query}.

Here are some background information that explains
     the query: {relevance_definition}

Rank the {num} passages above based on their
     relevance to the search query. All the passages
      should be included and listed using
     identifiers, in descending order of relevance.
     The output format should be [] > [], e.g., [4]
     > [2]. Only respond with the ranking results,
     do not say any word or explain.
```

Figure 13: Listwise prompt with an extra input of explicit definition.

# J Creation of the Expert-Informed Relevance Definition

Fig. 14 shows the prompt for the creation process of the expert-informed relevance definitions. Following the procedure in Schimanski et al. (2024b), we make use of the text parts labeled as relevant. There exists a relevance score from 1-3 where 1 signals the least and 3 is most relevant. Similar to the base setup for the experiments in Schimanski et al. (2024b), we use the text samples with a score of 2 or higher to create the expert-informed relevance definition. We include every relevant text part as an example. Thus, the prompt includes a large set of examples per question that can be synthesized in the relevance definition. While this procedure creates an expert-informed definition, it also represents a data leakage. In a sense, the definition will retrofit with the documents it searches for. While

| | | Number of words per document | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Dataset** | **Dataset Size** | **Mean** | **Std** | **Min** | **25%** | **50%** | **75%** | **Max** |
| Report-Level | 43.445 | 172 | 60.0 | 1 | 135 | 186 | 220 | 499 |
| Only Relevant | 595 | 199 | 43.8 | 38 | 177 | 213 | 228 | 281 |

Table 8: Statistical properties of the number of words in ClimRetrieve data.

| Label | Occurance |
|---|---|
| 1 | 100 |
| 2 | 167 |
| 3 | 328 |

Table 9: Distribution of relevance labels over the relevant (query, document) pairs in ClimRetrieve.

this is a limitation, we argue that this is also the step an expert human would take when annotating. She will have fixed concepts in her head, maybe even inspired by prior search processes. Therefore, we argue that this data leakage experiment is still adequate to investigate whether an expert-informed definition helps align the search process.

Plugging the examples into the prompt results in a set of expert-informed relevance relevance definitions. When comparing these relevance definitions to the generic ones, it becomes apparent that GPT-4 already incorporated the majority of the concepts that the experts were looking for. Therefore, the adjustment of the relevance definition is visible but rather subtle. One example is displayed in Table 10. While the meaning of the question remains rather static, there are nuanced differences in the examples that guide the relevance labeling.

## K MAP and nDCG Scores for Different Relevance Definitions

In the expert-informed definition experiment, we compare two settings. First, we compare the predictions on the 595 relevant-only (query, document) pairs. This is a replication of the setting in § 4.1. Since we don't have non-relevant samples, we can only compare the nDCG. Table 11 shows the results. It becomes apparent that only for the general nDCG score, the expert-informed query rates better. For the nDCG@5, and nDCG@10, the best-performing model remains with the generic prompt. The picture turns again when widening to nDCG@15. This could be a result of the definition creation. We use examples of relevance labels 2 and 3 to create the expert-informed definition. Thus, we implicitly equalize relevance 2 and 3 in importance. This means we are likely less effective in differentiating between 2 and 3. This could explain the lower results for at lower k's where differentiating between 2 and 3 is important v.s. the overall nDCG where differentiating between 1 and 2/3 plays a more important role.

This intuition is reinforced by the second setting, comparing the predictions on all 43K (query, document) pairs. In this setting, we also calculate the relevance for a large amount of non-relevant pairs. As Table 12 shows, the expert-informed definition now seems effective, especially when comparing MAP. MAP is agnostic to the actual degree of relevance and rather just differentiates between relevant and not relevant. Thus, the clearly higher MAP scores show that the expert-informed definition helps in differentiating between the non-relevant pairs where the definition is not meant for vs. those the definition was created with and for. This indicates that our approach is indeed sensitive to adjusting the relevance definitions.

## L Embedding Fine-Tuning

We follow the official fine-tuning example[9] of (Chen et al., 2024) to fine-tune the embedding models. The models are fine-tuned on all annotated (query, document) pairs in ChatReport test set for 10 epochs, with a batch size of 4. Other hyperparameters are the same as the official example.

## M Hand-Checking of Model Results on ClimRetrieve

The ClimRetrieve dataset presents the opportunity to investigate the output of our best-performing fine-tuned model Llama3 without CoT from different angles. In this experiment, we hand-annotate the output of our model in different categories and compare it to the ClimRetrieve annotations. As outlined in § 4, one issue of our approach is that we can only approximate the mental model of the annotator to create the relevance definitions. For the analysis, we consider the predictions of the fine-tuned

[9] https://github.com/FlagOpen/FlagEmbedding/tree/master/examples/finetune

| Question | Generic Definition | Expert-informed Definition |
|---|---|---|
| Does the company provide definitions for climate change adaptation? | "Meaning of the question: The question ""Does the company provide definitions for climate change adaptation?"" is asking for information about whether the company has clearly outlined what it understands by the term ""climate change adaptation"" within its sustainability reporting. This involves understanding how the company perceives the process of adjusting to actual or expected climate change and its effects, as well as how it plans to manage or mitigate harm or exploit beneficial opportunities.<br><br>Examples of information that the question is looking for:<br>1. A clear definition of climate change adaptation as understood and applied by the company in its operations and planning.<br>2. Descriptions of the specific impacts of climate change that the company is adapting to, such as increased frequency of extreme weather events, rising sea levels, or changing agricultural conditions.<br>3. Examples of measures the company is taking to adapt to these impacts, such as investing in flood defenses, altering supply chain logistics, or developing drought-resistant crops.<br>4. Information on how the company's definition of climate change adaptation aligns with or differs from standard definitions provided by authoritative bodies like the Intergovernmental Panel on Climate Change (IPCC) or the United Nations Framework Convention on Climate Change (UNFCCC).<br>5. Details on any collaborative efforts with governments, NGOs, or other corporations to address climate change adaptation collectively.<br>6. Insights into how the company measures or evaluates the effectiveness of its adaptation strategies." | "Meaning of the question: The question ""Does the company provide definitions for climate change adaptation?"" is asking for information about whether the company has clearly defined what it means by climate change adaptation within its sustainability reporting. This involves understanding how the company interprets climate change adaptation, including any specific strategies, initiatives, or policies it has in place to adjust to current or expected future climate change impacts.<br><br>Examples of information that the question is looking for:<br>1. Descriptions of how the company defines climate change adaptation in the context of its operations and strategic planning.<br>2. Examples of specific adaptation measures the company has implemented or plans to implement, such as enhancing infrastructure resilience, diversifying water sources, or adjusting agricultural practices.<br>3. Information on how the company's definition of climate change adaptation aligns with or diverges from standard definitions provided by environmental organizations or regulatory bodies.<br>4. Details on how the company assesses and integrates climate change risks and opportunities into its investment decision-making processes, focusing on adaptation.<br>5. Statements on the company's involvement in partnerships or alliances aimed at promoting climate change adaptation and resilience, indicating a collaborative approach to defining and addressing adaptation needs." |

Table 10: Example of a generic and expert-informed relevance definition for a question.

| Setting | nDCG | nDCG@5 | nDCG@10 | nDCG@15 |
|---|---|---|---|---|
| Llama3-Ask$_{generic}$ | 77.23 | **67.60** | **66.18** | **67.57** |
| Llama3-Tok$_{generic}$ | 76.55 | 67.20 | 66.23 | 65.83 |
| Llama3-Ask$_{informed}$ | 76.52 | 63.24 | 65.69 | 66.39 |
| Llama3-Tok$_{informed}$ | **77.41** | 65.95 | 65.06 | 66.91 |

Table 11: Comparison of using the generic and the expert-informed relevance definitions for ranking **relevant only** ClimRetrieve (query, document) pairs.

| Setting | nDCG | nDCG@5 | nDCG@10 | nDCG@15 | MAP | MAP@5 | MAP@10 | MAP@15 |
|---|---|---|---|---|---|---|---|---|
| Llama3-Ask$_{generic}$ | 29.95 | 18.67 | 21.71 | 23.38 | 26.51 | 17.86 | 21.21 | 22.75 |
| Llama3-Tok$_{generic}$ | 31.17 | 20.35 | 23.21 | 25.17 | 28.73 | 19.58 | 23.15 | 25.05 |
| Llama3-Ask$_{informed}$ | 30.89 | 19.01 | 22.82 | 24.89 | 29.31 | 20.02 | 23.60 | 25.56 |
| Llama3-Tok$_{informed}$ | **32.53** | **21.47** | **24.99** | **26.92** | **32.65** | **22.97** | **27.20** | **28.77** |

Table 12: Comparison of using the generic and the expert-informed relevance definitions for ranking **all** ClimRetrieve (query, document) pairs.

Llama-3 model with the expert-informed relevance definitions created in App. J as they likely present the closest approximation to the mental model of the labelers. Furthermore, the ClimRetrieve dataset only has golden labels for the (partially) relevant documents. This has implications for our results. In this setup, we view our model as a calibrated annotator acting according to the given relevance definition – i.e., our model represents a fictional golden truth. This allows us to perform a qualitative edge case analysis on the ClimRetrieve data in different categories.

The first category is the true positive classifications (model says relevant, ClimRetrieve-human says relevant). Since they are golden annotations by the ClimRetrieve annotators and align with our model, they are not checked as they likely are error-free. A more nuanced view has to be adopted for the three categories false negatives (model says relevant, ClimRetrieve-human says not relevant), true negatives (model says not relevant, human says not ClimRetrieve-relevant), and false positives (model says not relevant, human says ClimRetrieve-relevant). To obtain a qualitative understanding, we investigate the appearances (see Table 13) and confidences (see Table 14) of the categories. For the confidences, we use the empirically best-performing model in information retrieval Llama3-Tok (see Table 2). To perform a qualitative investigation, we sample (document, query) pairs from each category and reannotate them as being "relevant", "partially relevant" or "not relevant". These reannotations do not serve to obtain a holistic view (as already done in § 3) but rather gain insights into the special cases of the model's predictions.

First, we investigate the false negatives (model says relevant, human says not relevant). Investigating these cases is of particular interest since we know that humans have a selection bias (Thakur et al., 2021). In the analyst scenario setup of ClimRetrieve, this circumstance is aggravated by the fact that analysts may only search for information until they deem that they can answer the question at hand. Looking at Table 13, it becomes apparent that there is a very large number of false negatives. However, Table 14 shows that the confidence in the true negatives is much higher than in the false negatives. This indicates that the false negatives contain a much larger spectrum of partial relevance. This can be an initial explanation for the inflated number of false negatives.

To perform a qualitative investigation, we sample (document, query) pairs from the false negatives and perform a reannotation with a single annotator. We have 16 unique questions/queries in the ClimRetrieve dataset and sample 6 (document, query) pairs per question/query. Furthermore, we want to investigate which role model confidence plays. Thus, the (document, query) pairs are sampled to contain two samples with a Llama3-Ask confidence above 0.9, two between 0.9 and 0.7, and two below 0.7. We choose Llama3-Ask confidence because it is easier to form thresholds and approximates the confidence well enough.

Since we know that our model and the ClimRetrieve annotation inherently disagree, we focus on our reannotation. Fig. 16 shows the resulting percentage of agreement between the model and our reannotation for increasing the confidence threshold. It becomes apparent that irrespective of the threshold, the agreement between our reannotator and the model is very high. Furthermore, increasing the threshold correlates with a high model-reannotation agreement. This indicates that the model's assignment of the "relevant" label indeed works. At the same time, it bears the question of whether the initial ClimRetrieve-human annotation is invalid. Why does it underestimate the true relevant documents by such a large magnitude? There are three avenues to explain these results. First, as already mentioned, humans may miss out on sources, suffer from selection bias, not taking duplicate information into account, and may stop annotating once they reach a sufficient level of information for a question. While these things are inherent in the analyst setting of ClimRetrieve and out of our control, we can observe that ClimRetrieve clearly misses

```
"""
An analyst posts a <question> about a sustainability
     report. Your task is to explain the <question>
     in the context of sustainability reporting.
    Please first explain the meaning of the <
    question>, i.e., meaning of the question itself
     and the concepts mentioned. And then give a
    list of examples, showing what information from
     the sustainability report the analyst is
    looking for by posting this <question>.

For <the question's meaning>, please start by
    repeating the question in the following format:
'''
The question "<question>" is asking for information
    about [...]
'''

For the <list of example information that the
    question is looking for>, following the
    following example in terms of format:
---
[...]
3. Initiatives aimed at creating new job
    opportunities in the green economy within the
    company or in the broader community.
4. Policies or practices in place to ensure that the
     transition to sustainability is inclusive,
    considering gender, race, and economic status.
[...]
---

Here is the question:
<question>: ""{question}""

Additionally, here is a <list of question-relevant
    example information> that an expert human
    labler annotated. Please keep these examples in
     mind when answering:
--- [BEGIN <list of question-relevant example
    information>]
{examples}
--- [END <list of question-relevant example
    information>]

Format your reply in the following template and keep
     your answer concise:

Meaning of the question: <the question's meaning>
Examples of information that the question is looking
     for: <list of example information that the
    question is looking for>"""
```

Figure 14: RAG Prompt Template enforcing structured output with the inclusion of examples.

out on some relevant information (see for instance Fig. 15). However and second, one further major reason for the inflated number of false negatives may be the change of definition we make by creating it with GPT-4. The original questions all concern the specific topic of climate change adaptation. This is a very narrow, specialized case of the general climate change domain. In our definitions, this generally looks different. Consider for instance the following example:

```
"""
Meaning of the question: The question "Do the
    environmental/sustainability targets set by the
     company reference external climate change
    adaptation goals/targets?" is asking for
    information about whether the company's stated
    goals or objectives for environmental
    sustainability or climate change mitigation are
     aligned with, or make reference to,
    established external goals or targets. These
    external references could include international
     agreements, national policies, or standards
    set by recognized organizations focused on
    climate change and sustainability.

Examples of information that the question is looking
     for:
1. In line with our commitment to the Net-Zero
    Banking Alliance (NZBA) [...]
"""
```

While the question itself only addresses "climate change adaptation", the relevance definition allows for contents that are in line with "climate change mitigation". Mitigation is much broader and much more discussed in sustainability reports. The examples in the definition additionally broaden the scope. This also hints at the third avenue to explain the many false negatives: they contain a lot of partially relevant (document, query) pairs. This can also explain the significant drop in average confidence between true and false negatives in Table 14. However, the important aspect of the DIRAS pipeline is that it is consistent with the provided definitions which Fig. 16 indicates.

This investigation unveils clear edge cases for partial relevance and the mismatch of different definitions, i.e. mental models of annotators. Nonetheless, Fig. 16 reveals that, even amongst these hard examples, the approach remains consistent with the provided definitions. Although looking at edge cases, our reannotations are largely aligned with our model when overturning the non-golden irrelevant documents of ClimRetrieve.

To further get an intuition on edge cases, we turn our attention to the false positives. These represent the most complicated cases where the ClimRetrieve human annotator deemed the (document, query) pair relevant and the model arose with the judgment of not relevant. Since only 148 (document, query) pairs are in the false positives, we randomly sample 96 (document, query) pairs and reannotate them. The first two quantitative observations remain similar to the false negatives. First, Table 14 shows that the average confidence for the false positives is significantly lower than those of the true positives indicating they present edge cases. Indeed, they have the lowest average confidence among all categories with 0.8871. However, Fig. 17 again shows that with a rising confidence threshold, the agreement between the prediction and our reannotation rises. This time, the level of partial relevance is higher than for the true negatives. These two cues towards higher complexity continue in the qualitative assessment of our reannotations. We repeat to ask the question why is there a discrepancy between our model and the

ClimRetrieve annotations? This time, we identify a multitude of edge cases. First, a technical edge case comes into play. ClimeRetrieve was created by matching span-labeled relevant texts with documents retrieved from the reports. A document is deemed relevant when one sentence of the span-labeled text appears in a document. However, one sentence alone cut off or in a different context may not be relevant. In 19% (18 of 96) of the annotated samples, this is the reason for a wrong label assigned. Second, the nature of partially relevant labels seems to be ambiguous. While the ClimRetrieve human annotator deemed many samples partially relevant, our reannotation overturned some of these cases in line with our model. Nonetheless, there remains a significant chunk of partially relevant samples (see Fig. 17). Third, using our model, we overturn some of the decisions of the ClimRetrieve annotators entirely. Even if they deemed it relevant, our reannotation and model suggest irrelevance given the definition. This can again be linked to the mismatch of the exact mental model of the ClimRetrieve annotators and our approximated reannotation guidelines. It has to be noted that these decisions may again be overruled by a second or third reannotator. The samples generally all have a connection with the topics in the question. Rather than presenting a gold standard, this comparison allows us to understand the fine-grained nuances of relevance, therefore reinforcing the very need of this project. Clear misannotations were only marked in 3 of 96 cases where our annotator could not identify any remote relevance.

Finally, we investigate the large chunk of the (document, query) pairs, the true negatives (model says not relevant, human says not relevant). Having a large basis of 37.409 (document, query) pairs, we again sample 6 pairs per question using two samples with a Llama3-Ask confidence above 0.9, two between 0.9 and 0.7, and two below 0.7. The true negatives are not at all comparable with the investigations before because we now sample from pairs with apriori agreement. Thus, we don't investigate edge cases in this category.

This is also confirmed by the results (see Fig. 18). Only two samples with a Llama3-Tok confidence threshold below 0.5 were labeled as partially relevant, none were labeled as relevant. This reinforces the general results in § 3 and shows that the previous two reannotations concerned edge cases.

Collectively, these observations prove the motivation for this project in developing a nuanced stand beyond binary and even partial relevance. They show that the models function consistently with themselves - even in edge-case scenarios. This can mean something different than consistent with human annotators who cannot share their mental model.

| | | "Golden" model with "golden" definition | |
|---|---|---|---|
| | | True | False |
| Human with own mental model | True | 447 (TP) | 148 (FP) |
| | False | 5.441 (FN) | 37.409 (TN) |

Table 13: Comparison of the Llama-3's relevance prediction vs. the ClimRetrieve human annotators on ClimRetrieve's 43K (document, query) pairs.

| Labels assigned by fine-tuned LLama3 | Average Confidence Llama3-Tok |
|---|---|
| all | 0.9681 |
| positives | 0.9049 |
| negatives | 0.978 |
| true positives | 0.9575 |
| false positives | 0.8871 |
| true negatives | 0.9784 |
| false negatives | 0.9006 |

Table 14: Average confidence scores of Llama3-Tok for the different classification categories.

```
relevance definition:
"""
Meaning of the question: The question "Has the
     company identified any synergies between its
     climate change adaptation goals and other
     business goals?" is asking for information
     about how the company's efforts to adapt to
     climate change are aligned with, or can
     complement, its other business objectives. This
      includes understanding if climate change
     initiatives also support broader strategic
     goals such as cost reduction, risk management,
     innovation, market expansion, or reputation
     enhancement.

Examples of information that the question is looking
      for:
1. Descriptions of [...]
"""

MISSED OUT DOCUMENT:
"""
[...] Along with these efforts, in each business
     segment, Sony develops and enhances risk
     management and business continuity plans (BCPs)
      from the perspective of improving risk
     management across supply chains, through the
     identification, analysis, and assessment of
     business continuity risks. Flood damage has
     grown in recent years due to the impact of
     climate change, prompting Sony to reassess the
     flood risk at its manufacturing sites in Japan
     and implement preventative measures that will
     mitigate flood damage and facilitate rapid
     recovery. Sony is collaborating with relevant
     companies and organizations, and conducts hands
     -on drills to address foreseeable risks, in an
     effort to enhance business continuity and
     accelerate flood recovery. Sony will continue
     to increase its resilience to climate change,
     based on its analyses and initiatives.* A
     global initiative in which participating
     corporations aim to operate on 100% renewable
     electricity. It is headed by an international
     non-governmental organization, the Climate
     Group, in partnership with the CDP.
"""
```

Figure 15: False Negative example of clearly relevant but missed out information in ClimRetrieve.

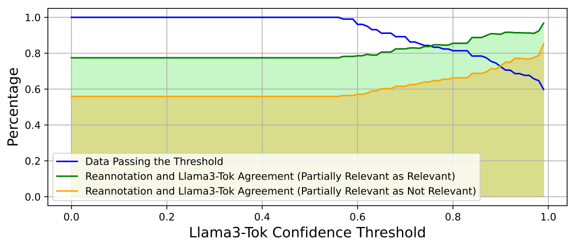


Figure 16: Percentage of agreement between model and our human annotation as well as the data remaining when adjusting the confidence threshold for false negatives.

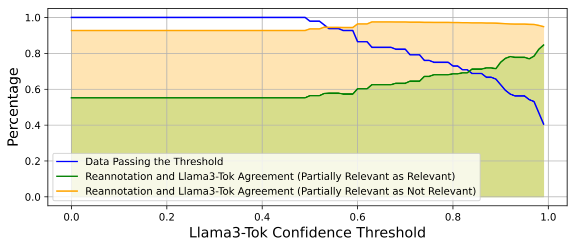


Figure 17: Percentage of agreement between model and our human annotation as well as the data remaining when adjusting the confidence threshold for false positives.

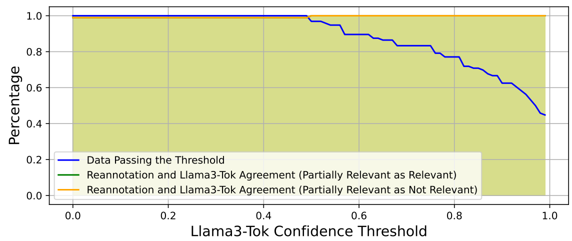


Figure 18: Percentage of agreement between model and our human annotation as well as the data remaining when adjusting the confidence threshold for true negatives.