DIRAS: Efficient LLM Annotation of Document Relevance for 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 when answering queries that need an integrated analysis of information (e.g., *Tell me good news in the stock market today.*)? To address these concerns, RAG developers need to annotate information retrieval (IR) data for their domain of interest, which is challenging because (1) domain-specific queries usually need nuanced definitions of relevance beyond shallow semantic relevance; and (2) human or GPT-4 annotation is costly and cannot cover all (query, document) pairs (i.e., annotation selection bias), thus harming the effectiveness in evaluating IR recall. To address these challenges, 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 consider nuanced relevance definition and annotate (partial) relevance labels with calibrated relevance scores. Extensive evaluation shows that DIRAS enables smaller (8B) LLMs to 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 one of the most popular paradigms 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. However, IR can be a performance bottleneck for RAG (Chen et al., 2023; 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 decrease in performance (Ni et al., 2023; Cuconasu et al., 2024; Niu et al., 2024; Schimanski et al., 2024a). Furthermore, 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 (Schimanski et al., 2024b; Bailey et al., 2008). To address these concerns, Saad-Falcon et al. (2023a) propose ARES to fine-tune an in-domain LM judge to evaluate context relevance. Although showing effectiveness in evaluating RAG systems for QA datasets like HotpotQA (Yang et al., 2018), ARES does not address two significant **C**hallenges that real-world RAGs are faced with:

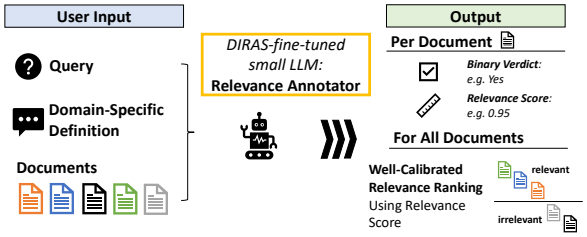

Figure 1: Overview of the functionality of DIRAS taking (query, relevance definition, document) triplets as input and output a binary verdict and a well-calibrated relevance score, which is sensitive to the grey-scale of partial relevance.

C1. IR Recall: RAG systems are more generally purposed than QA models, where many user queries require an integrative information analysis. For example, *Write an overview of major events in World War 2*; or *What is good news for the stock market today?* For such integrative queries, a good IR recall is necessary for comprehensive responses. Saad-Falcon et al.'s (2023a) approach evaluates the relevance of retrieved contexts (precision), but is agnostic to other important information that the RAG retriever might leave out (recall). Besides, current RAG literature (Yan et al., 2024; Wang

[1] All code, LLM generations, and human annotations in https://anonymous.4open.science/r/DIRAS-6D49.

et al., 2024) mostly relies on QA datasets (Joshi et al., 2017; Yang et al., 2018; Dinan et al., 2019; Trivedi et al., 2022) for evaluation, where the questions are less integrative and can mostly be answered by specific facts from one or few sources. For such questions, IR precision is more important than recall. As a result, the challenges of IR recall for integrative queries are heavily under-explored.

C2. Relevance Definitions and Partial Relevance: To thoroughly gather relevant information for integrative queries, the IR model should go beyond shallow semantic relationships and consider domain-specific relevance definitions. Furthermore, domain-specific requirements and subjectivity in IR annotation create a rich gray scale of *partial relevance* between *relevant* and *irrelevant* (Bailey et al., 2008; Saracevic, 2008; Thomas et al., 2024; also see App. A). However, partial relevance is neglected entirely in RAG context relevance evaluation (Saad-Falcon et al., 2023a; Es et al., 2024).

As a combined solution for these challenges, we propose **DIRAS**, a framework for efficient and effective relevance annotation. To address **C1**, DIRAS distills relevance annotation ability from SOTA generic teacher LLMs to small student LLMs, which can cost-efficiently annotate broad (query, document) pairs for IR recall evaluation. For better efficiency, student LLMs conduct pointwise annotation – annotating (query, document) pairs one-by-one – which is under-explored in related work (Sun et al., 2023a; Qin et al., 2024) but achieves good performance for relevance annotation. To address **C2**, DIRAS student LLMs are trained to comprehend nuanced relevance definitions, thus handling queries with various requirements. The student LLMs annotate binary relevance labels with well-calibrated relevance scores. Thus, the relevance scores can be used for relevance ranking and to calibrate the annotation accuracy (Ni et al., 2024). Thereby, these continuous relevance scores also address the grayscale of partial relevance (see Fig. 1 for an illustration of DIRAS functionality).

We evaluate DIRAS in three steps. First, we annotate ChatReportRetrieve to evaluate the design decisions for making DIRAS models optimized relevance annotators (§ 3.1). The evaluation shows that the fine-tuned student ($\leq$ 8B) LLMs effectively understand nuanced relevance definitions – achieving GPT-4-level performance (§ 3.3). Second, we showcase how DIRAS assists in real-life IR annotation by re-annotating ClimRetrieve (Schimanski et al., 2024b). Results show that DIRAS student LLMs can effectively capture partial relevance, leverage improved relevance definitions, mitigate annotation selection bias, and annotate benchmarking datasets for IR algorithms (§ 4.1). Third, we re-annotate document relevance for general QA and RAG datasets, showing DIRAS's potential in generic domains (§ 4.2). Collectively, our contributions include:

1. We propose DIRAS, a framework fine-tuning open-sourced LLMs into efficient and effective IR annotators, taking domain expertise into account.
2. We annotate ChatReportRetrieve, the first IR benchmark addressing integrative queries in RAG, and providing explicit relevance definition as annotation guidelines.
3. We showcase how to apply DIRAS in real-world IR annotation by re-annotating ClimRetrieve and general QA datasets.

2 DIRAS Pipeline

The DIRAS pipeline is illustrated in Fig. 2. It comprises three steps: sampling a subset of (query, document) pairs for training data creation, obtaining relevance definitions for queries, and distilling relevance annotation ability from teacher to student LLMs.

Sampling (Query, Document) Pairs: DIRAS takes domain-specific queries and documents as input. To sample representative (query, document) pairs as training data, it first ranks documents for each query using a small dense retriever. Then, it samples an equal number of documents within and outside of top-k (a pre-defined hyperparameter) to obtain representative documents for each query. While sampling in top-k aims at covering some relevant documents, sampling outside of top-k ensures covering the broader distribution of (query, document) pairs.

Obtaining Relevance Definitions: Each query in the sampled (query, document) pairs needs to be accompanied by an explicit definition indicating what is relevant or irrelevant to the question. The relevance definition can be generated by human experts, LLMs, or in a collaboration of both. In our experiments, we find GPT-4 generates suitable relevance definitions using the prompt in App. B.

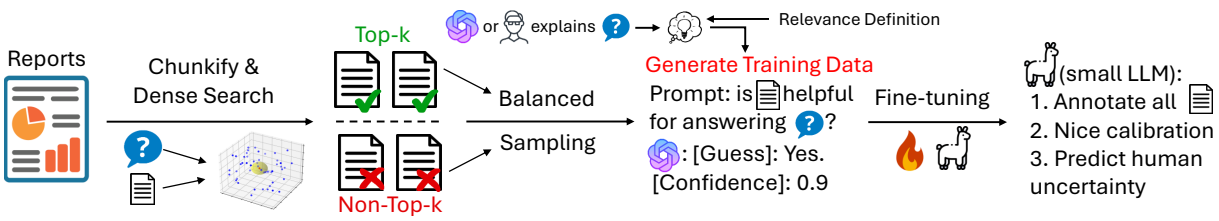

Figure 2: DIRAS pipeline. Domain-specific queries, and documents as inputs; calibrated student LLM annotators as outputs.

Distilling Relevance Annotations from Teacher to Student: With the sampled (query, definition, document) triplets, DIRAS creates relevance-annotation data with a SOTA generic teacher LLM $\mathcal{M}_t$ and the prompt template $\mathcal{P}$ (illustrated in Fig. 3). Finally, the created data is used to fine-tune student LLMs $\mathcal{M}_s$, which will be used to conduct broad binary relevance annotation with confidence scores for calibration (Tian et al., 2023).

3 Optimizing Relevance Annotation

To train a highly performant student LLM $\mathcal{M}_s$ for relevance annotation, we proceed in three steps. First, we annotate a novel task-specific dataset for our evaluation, ChatReportRetrieve (§ 3.1). Second, we assess design choices and related work baselines to find the best-performing strategy for our relevance-annotation data creation with the teacher model $\mathcal{M}_t$ (§ 3.2). Third, we optimize the fine-tuning of student LLMs $\mathcal{M}_s$ by investigating implementation variants (§ 3.3).

3.1 ChatReportRetrieve

To evaluate the teacher LLMs' ($\mathcal{M}_t$) and student LLMs' ($\mathcal{M}_s$) comprehension of nuanced relevance definitions, we need to provide them with the same annotation guidelines (i.e., relevance definitions) and compare their annotation performance. To the best of our knowledge, there is no existing IR dataset that provides a nuanced relevance definition for each query. Hence, we annotate ChatReportRetrieve for our evaluation.

Data and annotation guideline preparation: ChatReportRetrieve is based on real-user integrative queries from ChatReport – a chat tool for answering climate-related questions based on corporate reports[2] (Ni et al., 2023). We sample a wide range of climate reports and representative user queries about the reports to construct ChatReportRetrieve. Then, we conduct a train-test split, making sure no test set report or query is seen in the training set. Finally, we prompt GPT-4 to draft relevance definitions for all queries. GPT-4 drafted relevance definitions show a good understanding of the climate disclosure domain, according to a domain expert's feedback. See Appendices for details of data preprocessing (App. D) and relevance definition generation (App. B).

Test Data Annotation: We leverage relevance definitions as the annotation guidelines. *If and only if* a document addresses the relevance definition, it is deemed as (partially) relevant. We explicitly account for partial relevance when the document addresses the periphery of 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.

Evaluation Metrics: LLM predictions contain a binary relevance annotation and a confidence score. They will be evaluated against *relevance* or *uncertainty* of ChatReportRetrieve labels on four dimensions: (1) **Binary Relevance**: We compute the F1 Score of models' binary relevance prediction using *relevance labels*. Binary relevance labels are important for deciding which documents should be passed to LLMs. (2) **Calibration**: Confidence scores should calibrate the binary accuracy to indicate annotation quality. We use Expected Calibration Error (ECE), Brier Score, and AUROC to measure calibration performance, following Kadavath et al. (2022) and Tian et al. (2023). (3) **Information Retrieval**: The confidence scores also give a

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

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

Figure 3: Our prompt template $\mathcal{P}$ for distilling training data from $\mathcal{M}_t$. "[Reason]" is only used in the CoT setup. It is shortened for presentation. Full $\mathcal{P}$ is in App. Fig. 7.

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	-
RAGAs	-	68.15	-	37.13	-
ARES-0-Shot	25.48	52.63	79.35	77.67	58.79
ARES-2-Shot	17.38	3.85	63.49	44.97	32.42
ARES-4-Shot	18.16	5.13	69.51	49.68	35.62
ARES-8-Shot	16.99	28.81	66.77	48.75	40.33
ARES-16-Shot	20.65	24.20	64.76	44.85	38.61
Point-Ask	39.27	84.07	90.59	87.57	75.37
Point-Ask-Prob-D	44.74	84.52	**91.31**	88.39	77.24
Point-Tok-D	28.83	86.32	84.48	80.90	70.53
Point-Ask-D	**54.01**	**86.32**	91.10	**88.48**	**80.00**

Table 1: GPT-4's performance on ChatReportRetrieve test set with different ranking methods (Point- or List-wise), RAGAs and ARES-few-shot, with/without relevance definition (D), and calibration method (Ask or Tok). Bin., Cal., Info., and Unc. stand for evaluation dimensions in § 3.1.

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*. (4) **Uncertainty**: If the models understand the difficulty caused by partial relevance, they should have lower confidence scores on samples that humans found uncertain. Thus we compute average precision (AP) scores between confidence and *uncertainty labels*. Details of computing all metrics are in App. C.

3.2 Optimizing the Training Data Creation

We aim to train a highly performant student LLM $\mathcal{M}_s$ for relevance annotation. Thus, it is crucial to identify best-performing implementation choices that can be used to distill high-quality training data from the teacher LLM $\mathcal{M}_t$. Specifically, we compare the following four implementation choices:

ARES few-shot vs. relevance definitions: ARES (Saad-Falcon et al., 2023a) and DIRAS have different strategies to create target-domain training data: the former uses few-shot ICL[3] while the latter uses relevance definitions. We also include RAGAs (Es et al., 2024) relevance judgement as a baseline.

Pointwise vs. Listwise: The listwise method is popular in ranking data creation 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).

With vs. without relevance definition: As ChatReportRetrieve 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, relevance definition and CoT prompting (see Fig. 3 and Fig. 7, prompt without definition in Fig. 9). We use the listwise ranking prompt from Sun et al. (2023b) and Pradeep et al. (2023) (see prompt with-/without definition in Fig. 11/Fig. 10). 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. 8). 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. ARES and RAGAs settings are from the original papers. For few-shot ICL, we keep relevant/irrelevant samples balanced to avoid bias.

Takeaways: Results in Table 1 show that: (1) Few-shot ICL fails to teach domain-specific relevance. The ICL illustrations (even balanced) seem to bias GPT-4 to underperform the zero-shot setting. (2)

[3]Original ARES uses few-shot ICL to create synthetic queries instead of relevance labels, which is not applicable for ChatReportRetrieve. Thus we use their ChatGPT prompt for relevance judgement, while replacing GPT-3.5 with GPT-4.

Setting	Unc.	Bin.	Cal.	Info.	Avg.
Small-embed	-	-	-	66.34	-
Large-embed	-	-	-	69.36	-
BGE-Gemma	-	-	-	68.47	-
GPT-3.5	29.71	45.27	85.46	74.16	58.65
GPT-4	**54.01**	**86.32**	91.10	88.48	**80.00**
Llama3-CoT-Ask	36.57	76.58	89.30	86.15	72.15
Llama3-CoT-Tok	41.74	76.58	86.61	85.96	72.72
Llama3-Ask	40.18	82.11	90.14	86.02	74.61
Llama3-Tok	41.60	82.11	**91.35**	**89.19**	76.06†
Phi3-CoT-Ask	36.08	72.95	88.76	80.56	69.59
Phi3-CoT-Tok	35.49	72.95	84.20	80.64	68.32
Phi3-Ask	32.30	73.23	85.56	80.05	67.79
Phi3-Tok	38.00	73.23	89.52	86.94	71.92†
Gemma-CoT-Ask	31.60	72.38	86.38	81.39	67.94
Gemma-CoT-Tok	39.03	72.38	83.49	80.33	68.81
Gemma-Ask	25.74	67.13	81.80	77.43	63.02
Gemma-Tok	50.72	67.13	90.07	81.17	72.27†

Table 2: Comparison between the fine-tuned student $\mathcal{M}_s$ and different baselines on ChatReportRetrieve test data. The best scores are **bolded** and the second bests are underlined.† denotes the best score achieved by each backbone LLM.

With the proper calibration method (Ask), the pointwise method outperforms the listwise method. (3) The listwise method is sensitive to window size, while the pointwise method gives more consistent performance across prompts. (4) Adding relevance definition drops the listwise performance in 2 out of 3 cases, while that improves the pointwise performance. Thus we choose pointwise to be our distillation strategy.

3.3 Optimizing DIRAS student LLMs

DIRAS student LLMs $\mathcal{M}_s$ will be used to annotate all (query, document) combinations. Two methodological choices might influence the quality of fine-tuned student LLMs. First, we explore the role of Chain-of-Thought (CoT) reasoning. Second, we investigate the choice of calibration method (Tian et al., 2023). To explore these aspects, we fine-tune $\mathcal{M}_s$ in four settings: $\mathcal{M}_s$-CoT-Ask, $\mathcal{M}_s$-CoT-Tok, $\mathcal{M}_s$-Ask, $\mathcal{M}_s$-Tok, where CoT means $\mathcal{M}_s$ is tuned to generate [Reason], [Guess], and [Confidence]; without CoT denotes $\mathcal{M}_s$ 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. 3 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 student 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[4], a popular LLM-based reranker for general domain.

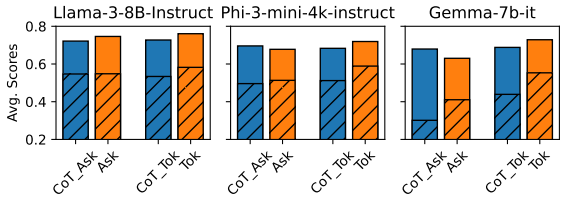

Figure 4: Shaded bars denote the performance of original models. Colored bars denote the improvement brought by fine-tuning.

As Fig. 4 shows, fine-tuning improves original models in all settings. Furthermore, Table 2 shows the results of all fine-tuned student models in comparison to all baselines. We observe that $\mathcal{M}_s$-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.

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, $\mathcal{M}_s$ should be fine-tuned without CoT for 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 Real-World Applications

Having established and benchmarked the design choices of DIRAS student LLMs, we now investigate their real-world application capabilities. First, we showcase how the DIRAS $\mathcal{M}_s$ can assist IR annotation in a real-world setting, leveraging the ClimRetrieve dataset (Schimanski et al., 2024b) (§ 4.1). Second, we show the DIRAS pipeline is also applicable for general domain QA (§ 4.2).

[4]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$_{improved}$	**30.89**	**29.31**
Llama3-Tok$_{generic}$	31.17	28.73
Llama3-Tok$_{improved}$	**32.53**	**32.65**

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

4.1 Applying to ClimRetrieve

ClimRetrieve (Schimanski et al., 2024b) records human analysts' real-life workflow of reading full reports, searching for relevant information, and annotating useful information for climate-related questions with relevance scores 1-3. ClimRetrieve contains 43K (query, document) pairs (8K unique documents but each document in a report is multiplied by the amount of analyzed queries per report) out of which 595 are 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). To succeed on this dataset, the IR model needs to (1) capture the analysts' mental model about useful information (i.e., relevance definition), and (2) understand fine-grained differences in degree of relevance (score 1-3).

Since ClimRetrieve is still in the climate domain, we re-annotate its data with the best $\mathcal{M}_s$ in § 3 (Llama3-Tok), and explore whether $\mathcal{M}_s$'s annotations can (1) **RQ1**: reflect fine-grained differences in degree of relevance; (2) **RQ2**: be improved through refining relevance definitions; (3) **RQ3**: mitigate annotation selection bias for better IR recall evaluation; and (4) **RQ4**: benchmark and select IR algorithms.

RQ1: Reflecting Fine-Grained Relevance Levels. We first evaluate Llama3-Tok's annotation 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[5] 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.

RQ2: Improving Performance through Improving Definitions. So far, we used GPT-4 to draft the relevance definitions. To investigate the effect of improved definitions, we compare two setups: (1) The *generic* relevance definition: the definition drafted by GPT-4. (2) The *improved* relevance definition: The only way to improve the definition is to align it closer to ClimRetrieve annotators' mental model of document relevance. We achieve this by adding relevant text samples to the prompt for generating the definition with GPT-4 (see App. I for details).

After creating the improved definitions, we repeat predicting the relevance scores with Llama3-Tok. Since we involve examples with various relevance scores to improve relevance definitions, these definitions might not help distinguish the granular level of partial relevance. However, the improved definitions might especially help to distinguish relevant documents from irrelevant ones. 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 improved definitions seems to enhance the performance (see App. J for details).

RQ3: Mitigating Annotation Selection 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,

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

	ClimRetrieve (62.00%)		
	All	**Rel.**	**Irr.**
Conf ≤ 0.95	36.84	27.90	48.48
Conf > 0.95	**85.48**	**78.18**	**91.30**

Table 5: Accuracy of student model Llama3-Tok annotations that disagree with original ClimRetrieve relevance labels. **All** denotes all sampled disagreed labels. **Rel.** (**Irr.**) denotes the subset where the original label is *relevant* (*irrelevant*). Conf ≤ 0.95 (> 0.95) denotes the subset where Llama3-Tok's confidence is lower (higher) than 0.95. **ClimRetrieve** (62.12%) means that 62.12% of data are annotated with > 0.95 confidence.

Setting	**Kendall's τ**
BGE-Base	35.71
BGE-Base-ft	**36.34**
BGE-Large	34.74
BGE-Large-ft	**36.55**

Table 6: Different embedding models' performance benchmarked by student model $\mathcal{M}_s$'s prediction on all 43K (query, document) pairs of ClimRetrieve. "ft" denotes the model is fine-tuned on in-domain data.

the dataset allows us to investigate our model's capabilities to counteract biases. For this purpose, we sample 200 disagreements between Llama3-Tok's annotation and the original ClimRetrieve labels. Then, we reannotate these samples with a human labler. To account for different confidence levels in Llama3-Tok's prediction, we differentiate prediction with confidence higher or lower than 0.95.

As Table 5 indicates, the model can be successfully used to overturn decisions of unseen, as irrelevant assumed documents (91.30% for confidence > 0.95). Strikingly, even samples annotated by humans, i.e., those labeled as relevant, can be overturned, though with a lower certainty. We attribute this to differences in the unknown mental model of the ClimRetrieve labeler and our explicit relevance definitions (for details, see App. L). However, for us, it is reaffirming to observe that the DIRAS model is consistent with its own definition. Thus, we conclude that DIRAS' labeling is effective and helps to mitigate annotation selection bias.

RQ4: Benchmarking IR. We use $\mathcal{M}_s$ 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 $\mathcal{M}_s$-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 ChatReportRetrieve test set[6] (fine-tuning details in App. K). 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 τ as the metric, which directly compares the correlation between two ranks. The results are shown in Table 6. We find the Llama3-Tok-annotated benchmark successfully picks out the fine-tuned checkpoints, showing a capability of benchmarking information retrieval algorithms.

4.2 Applying to QA Datasets

In this section, we apply the DIRAS pipeline to QA datasets that are widely used in RAG benchmarking. DIRAS addresses queries for broad information and IR recall. Thus, we include long-form QA datasets from ALCE (Gao et al., 2023), including ELI5 (Fan et al., 2019), ASQA (Stelmakh et al., 2023), and QAMPARI (Amouyal et al., 2023). We also include RAG-Bench (Fang et al., 2024) that consists of questions from TriviaQA (Joshi et al., 2017), WebQ (Berant et al., 2013), and Natural Questions (Kwiatkowski et al., 2019). RAG-Bench is chosen since it has labels for partial vs. full relevance, which is a focus of DIRAS. Importantly, the context relevance labels from ALCE and RAG-Bench are derived from reference answers to questions, using heuristics instead of manual annotation. Thus, they are to some extent **noisy**.

We first run the DIRAS pipeline[7] on each dataset to obtain corresponding Llama3-Tok models. Then we compare them with the teacher model – GPT-4. Results in Table 7 show that Llama3-Tok outperforms GPT-4 in IR. Then for each dataset, we repeat the annotation selection bias assessment of RQ3 in § 4.1. Again, we sample 200 disagreements between Llama3-Tok annotation and the original

[6]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 $\mathcal{M}_s$ is fine-tuned on the training set.

[7]Different from Climate change datasets, QA datasets do not require nuanced relevance definition. Thus we use a fixed relevance definition: *"the document is helpful only if its content answers the query"*. See full prompt in Fig. 14.

	ELI5			ASQA			QAMPARI			RAG-Bench		
	N	N@10	N@20	N	N@10	N@20	N	N@10	N@20	N	N@10	N@20
GPT-4	48.43	15.97	17.81	64.62	38.82	46.32	56.21	27.54	35.34	42.91	31.84	41.35
Llama3-Tok	**48.73**	**17.04**	**20.08**	**64.90**	**39.37**	**48.44**	**56.58**	**28.70**	**35.49**	**48.27**	**41.13**	**47.88**

Table 7: Applying DIRAS to QA datasets, the IR performance of student model Llama3-Tok and GPT-4. **N** denotes nDCG.

	ELI5 (85.05%)			**ASQA** (66.32%)			**QAMPARI** (72.66%)			**RAG-Bench** (60.33%)		
	All	**Rel.**	**Irr.**	**All**	**Rel.**	**Irr.**	**All**	**Rel.**	**Irr.**	**All**	**Rel.**	**Irr.**
Conf ≤ 0.95	62.35	48.48	71.15	67.27	71.43	66.67	74.11	73.33	74.23	58.41	38.10	63.04
Conf > 0.95	**84.71**	**83.12**	**100.0**	**90.09**	**95.83**	**88.51**	**90.00**	**81.25**	**91.49**	**96.36**	**95.24**	**96.63**

Table 8: Accuracy of student model Llama3-Tok annotations that disagree with original relevance labels. Same setup as Table 5.

(potentially noisy) relevance labels, and manually check whether Llama3 or original labels are correct. As shown in Table 8, Llama3-Tok's annotations are predominately correct with a confidence > 0.95 (e.g., 85.05% of ELI5). When there is a disagreement, relying on Llama3-Tok leads to less error (Acc. $> 50\%$), especially when the confidence > 0.95, thanks to the good calibration. For ASQA, QAMPARI, and RAG-Bench, the majority ($> 90\%$) of the disagreement lies in originally irrelevant labeled part of the dataset (**Irr.**), possibly due to (query, document) pairs are selectively annotated. DIRAS achieves high accuracy in **Irr.** disagreements. Therefore, we reaffirm the notion that applying DIRAS to annotate broader (query, document) pairs can effectively reduce annotation selection bias, and thus improve IR recall benchmarking. All implementation details are in App. M.

5 Background and Related Work

Prior work on IR in RAG has already explored the idea of using LLMs to judge relevance. The closest to our work are RAGAs and ARES. Es et al. (2024) develop RAGAs to evaluate the relevance of each sentence in a paragraph with a closed-source LLM and create an aggregated score by dividing the number of relevant sentences over all sentences. Saad-Falcon et al. (2023b) aim to bring a relevance judge to a target domain through few-shot in-context learning (ICL): they first generate synthetic questions to given target-domain passages, and then fine-tune small classifiers for relevance judgment. However, many real-world challenges remain unaddressed. Specifically, IR recall, partial relevance, and domain-specific relevance definitions are neglected. ARES appears to have target-domain IR evaluation, but the synthetic data approach focuses on less integrative queries: each question is generated given a single passage and hard negatives are passages sampled from the same document, which could be (partially) relevant for integrative queries asking for broader information. Furthermore, we also find that few-shot ICL fails to teach domain-specific relevance to LLMs (§ 3.2).

Besides, 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. These studies all focus on pairwise or listwise ranking methods, and discusses that pointwise methods may not work due to bad calibration (Qin et al., 2024). However, when it comes to relevance annotation instead of reranking, pointwise relevance prediction is more suitable because it: (1) analyzes and annotates (query, document) pairs one-by-one, thus is more efficient and may better consider relevance definitions; and (2) can annotate both document rank and binary relevance labels (relevant or irrelevant) which are important for RAG in order to decide what information should be passed to the generator. We also show pointwise annotation works better with proper calibration method (Tian et al., 2023).

6 Conclusion

In this work, we introduce the DIRAS pipeline to fine-tune open-source LLMs to calibrated annotators. 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. First, our results show that DIRAS fine-tuning grants small student LLMs GPT-4-level performance on specific domains, but GPT-4 is not guaranteed to be perfect in all domains and cases. In certain niche domains, it might be necessary to augment GPT-4 with domain knowledge or agentic designs to achieve human-level performance in relevance annotation, and then create reliable training data for DIRAS. Although performance not always guaranteed, LLM annotation for document relevance is still necessary due to the sheer volume of (query, document) pairs and selection bias of human annotation.

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 multi-modality, 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.

Samuel Joseph Amouyal, Tomer Wolfson, Ohad Rubin, Ori Yoran, Jonathan Herzig, and Jonathan Berant. 2023. Qampari: An open-domain question answering benchmark for questions with many answers from multiple paragraphs.

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.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.

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.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. 2023. 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].

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering.

Feiteng Fang, Yuelin Bai, Shiwen Ni, Min Yang, Xiaojun Chen, and Ruifeng Xu. 2024. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations.

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].

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension.

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].

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

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.

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. 2023a. Ares: An automated evaluation framework for retrieval-augmented generation systems.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2023b. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. ArXiv:2311.09476 [cs].

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.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2023. Asqa: Factoid questions meet long-form answers.

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 Tre-

bacz, 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, Renshen 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].

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Zilong Wang, Zifeng Wang, Long Le, Huaixiu Steven Zheng, Swaroop Mishra, Vincent Perot, Yuwei Zhang, Anush Mattapalli, Ankur Taly, Jingbo Shang, Chen-Yu Lee, and Tomas Pfister. 2024. Speculative rag: Enhancing retrieval augmented generation through drafting.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering.

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 query 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 query clearly looks 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 the Relevance Definition

Fig. 5 shows the prompt template for the creation of the query 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 11.

```
"""
An analyst posts a <question> about a climate report
     . Your task is to explain the <question> in the
      context of climate 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
     climate 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 5: Prompt for generating a query relevance definition.

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 τ. For nDCG, relevant scores 0.5 and 1 are assigned to partially relevant and relevant documents correspondingly. When averaging Calibration metrics, we average AUROC with $1-$ ECE and $1-$ Brier Score to keep

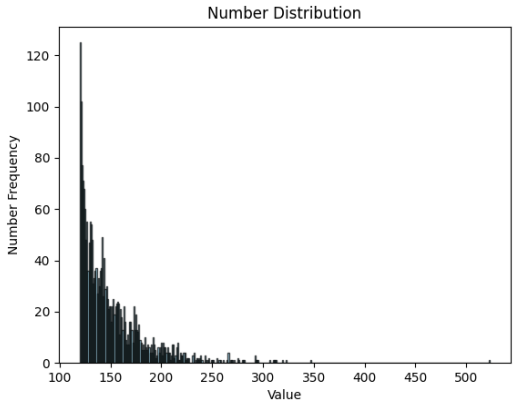

Figure 6: Distribution of chunk length after being extracted from climate reports and concatenation.

the trend consistent.

D ChatReportRetrieve Data Preprocessing

Data Sampling: We sample 31 questions and 80 reports from this application. Importantly, ChatReportRetrieve differs from general QA dataset and focuses on integrative queries which usually ask for broad relevant information. Therefore, with controlled annotation budget (i.e., number of (query, document) pairs), we are prone to have fewer (but representative) queries and more documents for each query. The queries are strategically sampled to ensure representativeness and diversity. Specifically, 11 queries are the core queries used in ChatReport, which cover essential topics of sustainability disclosure. 20 questions are selected from users' customized questions posed to the ChatReport tool. Climate reports are sampled randomly from openly accessible user submissions[8] Finally, we prompt GPT-4 to draft relevance definitions for all queries (see App. B).

PDF Parsing: We use IBM deepsearch parser (Team, 2022) to parse corporate reports into chunks. For chunks shorther than 120 tokens, we concatenate them with adjacent chunks to form chunks longer than 120. Figure Fig. 6 shows the formatted chunks length distribution.

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 queries and reports. For each query, we randomly sample 60 documents – 30 each from in the top-5 and 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 § 3.2). Data points in the test split are passed to human annotation. We use 31 climate-related queries for data sampling among 80 climate reports. For the separation of the train dataset and test dataset, these queries are classified into 3 categories: vague, specific, and TCFD. Within each category, queries are split randomly for training and testing, resulting in 20 queries for the train set and 11 queries for the test set. Meanwhile, the 80 climate reports are randomly split into 50 for train and 30 for test.

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

E Expert Annotation Process

As described in App. D, the data is obtained from real climate reports and split into chunks of around 150 words with the IBM deepsearch parser (Team, 2022). Table 9 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 falls 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 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.

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

Table 9: Statistical properties of the number of words in ChatReportRetrieve test set data.

Label	Occurance
Relevant	121
Partially	65
Not Relevant	474

Table 10: Label distribution in the ChatReportRetrieve testset.

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 10. 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 [9], 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.

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

G DIRAS Prompt Template $\mathcal{P}$

Fig. 7 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 climate 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 7: Full DIRAS Chain-of-Thought prompt for LLMs predicting relevance labels and calibrating.

H Alternative Prompts

Fig. 7, Fig. 8, and Fig. 9 show the alternative prompts with which we experimented.

I Creation of the Improved Relevance Definitions

Fig. 12 shows the prompt for the creation process of the improved 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.

```
{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 8: Output requirements for the alternative prompt setting $\mathcal{P}_{prob}$. Task description and input are the same as Fig. 7.

```
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 9: Task description and input part for the alternative prompt setting $\mathcal{P}_{w/o_e}$. Output requirements are the same as Fig. 7.

(2024b), we use the text samples with a score of 2 or higher to create the improved relevance definition. We include relevant text samples in the prompt for creating the relevance definitions to obtain improved definitions. The logic behind this definition creation is that we assume we know the mental model of ClimRetrieve human analysts, and thus know what information is relevant before annotation. This is common in corporate report analysis where experts will have fixed concepts in their heads, maybe even inspired by prior search processes.

Plugging the examples into the prompt results in a set of improved 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 11. While the meaning of the question remains rather static, there are nuanced differences in the examples that guide the relevance labeling.

```
<|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 10: 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 11: Listwise prompt with an extra input of explicit definition.

J MAP and nDCG Scores for Different Relevance Definitions

In the improved 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 **??**. Since we do not have non-relevant samples, we can only compare the nDCG. Table 12 shows the results. It becomes apparent that only for the general nDCG score, the improved query performs 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 improved definition with GPT-4. Thus, we implicitly equalize relevance 2 and 3 in importance. This means we are likely less effective in differentiat-

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 climate 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. Examples of information that the question is looking for: 1. A clear definition of climate change adaptation as understood and applied by the company in its operations and planning. 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. 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. 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). 5. Details on any collaborative efforts with governments, NGOs, or other corporations to address climate change adaptation collectively. 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 climate 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. Examples of information that the question is looking for: 1. Descriptions of how the company defines climate change adaptation in the context of its operations and strategic planning. 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. 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. 4. Details on how the company assesses and integrates climate change risks and opportunities into its investment decision-making processes, focusing on adaptation. 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 11: Example of a generic and expert-informed relevance definition for a question.

ing between 2 and 3. This could explain the lower results 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 13 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.

K Embedding Fine-Tuning

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

L Hand-Checking DIRAS Model Disagreements with ClimRetrieve

We want to check samples where our student model Llama3-Tok's prediction differs from the relevance label in ClimRetrieve. We sample equally from those samples where Llama3-Tok indicated rele-

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

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 12: 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 13: Comparison of using the generic and the expert-informed relevance definitions for ranking **all** ClimRetrieve (query, document) pairs.

vance and ClimRetrieve does not and vice versa. While our focus lies on those documents that were not annotated in ClimRetrieve (i.e. irrelevant ones), the dataset also allows us to investigate how our model performs in the edge case of when a ClimRetrieve annotator deems relevance and Llama3-Tok does not.

On the samples that were labeled as irrelevant in ClimRetrieve, we find that our student model Llama3-Tok model is effective in mitigating annotation selection bias and therefore incorporates the perspective of IR recall (see Table 5).

However, it is interesting that our student model Llama3-Tok successfully overrules human decisions. We attribute this to the fact that our created relevance definitions might differ from the mental model of the human annotator in ClimeRetrieve. Thus, humans in ClimRetrieve might have been consistent with their own mental model. For us, however, it is more important and reaffirming to observe that Llama3-Tok is consistent with its own, explicit relevance definitions.

We can view the (query, definition, document) pair in Fig. 13 as an example. When analyzing the query "Do the environmental/sustainability targets set by the company reference external climate change adaptation goals/targets?", the ClimRetrieve labeler interpreted the question broader, i.e., deeming this as relevant: "As a global technology leader, we are also committed to helping build the enabling societal conditions that will support a net zero economy.". However, for our student model, it is in line with the definition to assign a "not relevant" label. There is no explicit standard mentioned in the document.

M Implementation Details of Experiments on QA Datasets

ALCE Data: We obtain ELI5, ASQA, and QAMPARI from ALCE (Gao et al., 2023), where they parse the original open-domain QAs into RAG forms[11]. For each question, ALCE annotates 5 documents as oracle based on retrieval recall and reference answers, which are used as context relevance labels in our experiment. For each dataset, we randomly sample 100 queries to construct DIRAS training data, following the process in Fig. 2. Top-5 is selected for balanced sampling, thus resulting in 1000 (query, document) pairs for training. We sample 50 questions for test data, and include all (query, document) pairs for them, resulting in 5K (query, document) pairs for each dataset.

RAG-Bench Data: RAG-Bench classifies RAG sources into four types: (A) relevant and with answers, (B) relevant topic but without answers, (C) irrelevant topic, and (D) with counterfactual answers. We find (B) addresses partial relevance that DIRAS cares about. Therefore, we leverage its dev set for training and test set for testing, where (A) and (D) become relevant documents, and (B) and (C) are used as irrelevant ones.

Disagreement Sampling for Table 8: We sample 200 disagreed annotations four each dataset, 50 samples from each confidence range: Conf$<$ 90, 90 $<$Conf$<$ 95, 95 $<$Conf$<$ 98, and 98 $<$Conf$<$ 100 to balancedly cover different confidence scores (these bins are of similar size).

[11]Data files in https://github.com/princeton-nlp/ALCE

```
"""
An analyst posts a <question> about a climate report
     . Your task is to explain the <question> in the
      context of climate 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 climate
     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 12: Prompt Template enforcing structured output with the inclusion of examples.

```
QUERY: "Do the environmental/sustainability targets
     set by the company reference external climate
     change adaptation goals/targets?"

QUERY DEFINITION: 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) [...]

DOCUMENT: Enabling a more sustainable world
     Microsoft's actions alone will not solve the
     climate crisis. As a global technology leader,
     we are also committed to helping build the
     enabling societal conditions that will support
     a net zero economy. We're focused on
     accelerating the availability of new climate
     technologies, strengthening our climate policy
     agenda, helping to develop a more reliable and
     interoperable carbon accounting system,
     advocating for skilling programs to expand the
     green workforce, and working to enable a just
     energy transition.
```

Figure 13: Example for which DIRAS Llama3-Tok assigns "not relevant" and ClimRetrieve assigned "relevant". This example shows that while DIRAS Llama3-Tok might differ with the opinion of ClimRetrieve's annotator, it is consistent with its own definition.

```
You will be provided with a <question> and a <
     paragraph>.  Your job is to assess whether the
     <paragraph> is useful in answering the <
     question>, given the <background_information>
     defining what is useful.

<background_information>: "The <paragraph> is useful
      only if some of its content directly answer
     the <question> or at least a part of the <
     question>. Content with relevant topic but
     without direct answers are not useful."
<question>: "{question}"
<paragraph>: "{paragraph_chunk}"


Is <paragraph> useful for answering <question>?
     Provide your best guess 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 14: The DIRAS prompt for QA experiments.