CLAPNQ: Cohesive Long-form Answers from Passages in Natural Questions for RAG systems

Sara Rosenthal, Avirup Sil, Radu Florian, Salim Roukos
IBM Research AI, USA
{sjrosenthal,avi,raduf,roukos}@us.ibm.com

Abstract

Retrieval Augmented Generation (RAG) has become a popular application for large language models. It is preferable that successful RAG systems provide accurate answers that are supported by being grounded in a passage without any hallucinations. While considerable work is required for building a full RAG pipeline, being able to benchmark performance is also necessary. We present CLAPNQ, a benchmark Long-form Question Answering dataset for the full RAG pipeline. CLAPNQ includes long answers with grounded gold passages from Natural Questions (NQ) and a corpus to perform either retrieval, generation, or the full RAG pipeline. The CLAPNQ answers are *concise*, 3x smaller than the full passage, and *cohesive*, meaning that the answer is composed fluently, often by integrating multiple pieces of the passage that are not contiguous. RAG models must adapt to these properties to be successful at CLAPNQ. We present baseline experiments and analysis for CLAPNQ that highlight areas where there is still significant room for improvement in grounded RAG. CLAPNQ is publicly available at https://github.com/primeqa/clapnq.

1 Introduction

Question answering (QA) has been a popular natural language processing task for many years. Large scale research in this area began with the tasks of Machine Reading Comprehension (Rajpurkar et al., 2016; Rogers et al., 2023; Fisch et al., 2021), and Information Retrieval (Manning et al., 2008; Voorhees and Harman, 2005; Thakur et al., 2021) and is more recently known as Retrieval Augmented Generation (Lewis et al., 2020; Guu et al., 2020) which encompasses both tasks. The recent popularity of generative AI with Large Language models (LLMs), such as GPT (Brown et al., 2020), Llama (Touvron et al., 2023), FLAN-T5 (Chung et al., 2024), and Mistral (Jiang et al., 2023) has shifted the focus to providing long and detailed answers for any user information need. An important challenge for responses produced by an LLM is ensuring that answers are faithful (being grounded in a supporting passage) to ensure that a user can be confident in the response provided to them.

CLAPNQ is a grounded long-form QA benchmark dataset for Retrieval Augmented Generation of LLMs. The answers are typically long, 2–3 sentences, in contrast to datasets based on machine reading comprehension such as Natural Questions (NQ) (Kwiatkowski et al., 2019) and SQuAD (Rajpurkar et al., 2016, 2018) which are just a few words. It is grounded on a single gold passage, in contrast to other long-form question answering (LFQA) datasets such as ELI5 (Fan et al., 2019), where gold passages are not available. It is built from a subset of the highly successful Natural Questions (Kwiatkowski et al., 2019) dataset for extractive QA from Wikipedia documents. The NQ questions are based on users real web search queries. Specifically, we explore the subset of NQ that has long answers (passages) but no short extractive answers. CLAPNQ is suitable for evaluating all parts of Retrieval Augmented Generation (RAG) systems: Retrieval, Generation and the full RAG pipeline (Figure 1):

Retrieval Retrieve *N* relevant passages for a question from the indexed CLAPNQ corpus.

Generation Generate a response/answer for the prompt which is the concatenation of the question, the gold passage, and the instruction for the model.

RAG Retrieve *N* passages for the question from the CLAPNQ corpus. Generate a response/answer for the prompt which is the concatenation of the question, *N* passages, and instruction for the model.

Transactions of the Association for Computational Linguistics, vol. 13, pp. 53–72, 2025. https://doi.org/10.1162/tacl_a_00729
Action Editor: Jimmy Lin. Submission batch: 7/2024; Revision batch: 9/2024; Published 1/2025.

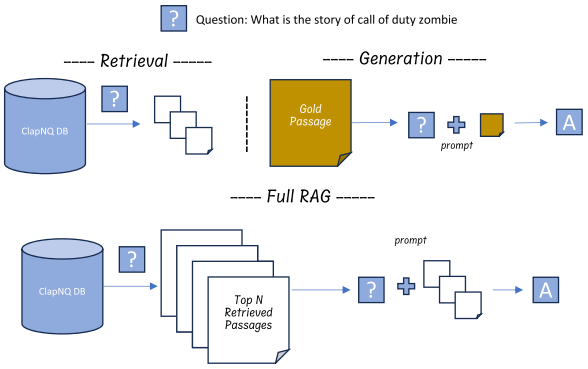

Figure 1: CLAPNQ is designed to test all parts of the RAG pipeline: Retrieval, Generation with gold passages, and the full RAG setup with generation on retrieved passages.

It is important to evaluate all RAG scenarios to measure retrieval and generation performance separately, as well as the full pipeline to illustrate how the retrieval performance and noisy passages impacts generation, making it a much more difficult and challenging task.

We present the CLAPNQ dataset of 4946 questions with gold passages for evaluating generation models on grounded LFQA with its corresponding corpus. The answers in CLAPNQ are faithful, concise, complete, and cohesive. An example of a question and grounded answer from CLAPNQ is shown in Table 1. We created CLAPNQ with the following properties in order to make it suitable for evaluating generative models:

Faithful The answer must be grounded in the gold passage. While the answers can be written differently than in the passage, they tend to be highly extractive due to the nature of the dataset creation.

Concise The answer must have all the information needed to answer the question but exclude information that is unrelated to the answer. In the original NQ dataset, the entire passage is considered the answer, but this has too much irrelevant information.

Complete A short answer (e.g., 2–3 words) commonly found using MRC systems is not sufficient for many types of questions that have a richer information need, or require clarity or an explanation. The response must include all information needed to answer the question.

Question: what is the story of call of duty zombie
Title: Call of Duty: Black Ops III **Passage: Black Ops III takes place in 2065, 40 years after the events of Black Ops II, in a world facing upheaval from climate change and new technologies .** Similar to its predecessors, the story follows a group of black ops soldiers . The game 's campaign is designed to support 4 - player cooperative gameplay, allowing for bigger, more open level design and less corridor shooting . As the player character is cybernetically enhanced, players have access to various special activities . **The game also features a standalone Zombies mode, and a ''Nightmares'' mode which replaces all enemies as zombies .**
Reference Answer: Call of duty: Black Ops III takes place in 2065 in a world facing upheaval from climate change and new technologies. The game features a standalone Zombies mode, and a ''Nightmares'' mode which replaces all enemies as zombies.

Table 1: An example of a CLAPNQ answerable question with the reference annotated answer. Sentences in **bold** were selected as relevant parts of the answer. The annotators combined them with modifications to make a cohesive and complete answer.

Cohesive The response contains the important concepts/facts of the passage assembled together to form a concise and complete answer. In many cases, forming a cohesive answer requires harmonizing multiple non-contiguous pieces of text in a fluent manner.

Unanswerable We retain a portion of NQ unanswerable questions that have similar properties to the answerable CLAPNQ questions. This has been largely overlooked by prior LFQA datasets, while expected for real-world RAG applications.

CLAPNQ is the first LFQA benchmark dataset to have grounded gold passages and a full corpus making it suitable for evaluating the full RAG pipeline. Our experiments and results in Section 4 show that LLMs still need considerable work in answering LFQA, remaining faithful to the document, performing the full RAG pipeline, and knowing when a question should not be answered. Our main contributions are:

1. The creation of CLAPNQ with non-consecutive relevant fragments, allowing us to test the ability of LLMs to extract

Dataset	Queries	A per Q	W in Q	W in A	S in A	IAA	Unanswerable
AquaMuse Abstractive	21042	1.0	9.2	106.7	3.7	–	–
AquaMuse Extractive	44217	1.0	9.2	106.7	3.7	–	–
ASQA	6316	1.3	10.1	80.7	3.2	0.48	–
ELI5	1507	12.0	19.6	116.9	5.7	0.16	–
ExpertQA	2169	1.0	21.2	174.8	6.1	–	–
TruthfulQA	817	3.2	12.4	9.0	1.0	0.37	11
WikiHowQA	1188189	1.0	7.0	70.1	7.6	–	–
CLAPNQ-R1	12657	1.1	9.2	39.0	1.6	–	–
CLAPNQ	4946	1.4	9.4	56.8	2.3	0.67	2493

Table 2: Comparison to existing Long-form QA datasets. Stats are shown for Answers (A), Queries (Q), Words (W), Sentences (S), Inter-Annotator Agreement (IAA) and Unanswerable. W in A of CLAPNQ is 1/3 of W in Passage (P) = 156.

just the relevant parts of the passage, while remaining faithful and concise.

2. A set of baseline experiments with State-of-the-Art (SOTA) models for both retrieval, generation, and the full RAG pipeline.
3. A human evaluation and discussion to highlight areas where there is room for improvement.

In the rest of this paper we present related work, the dataset creation and details, experiments and results on SOTA retrieval, generative models and the full RAG pipeline. We also present human evaluation, analysis, and areas of future research that the CLAPNQ benchmark can be used for advancing RAG research. CLAPNQ is publicly available in a Github repository.[1]

2 Related Work

Natural Questions (Kwiatkowski et al., 2019) is a large MRC QA dataset of 323k questions built using Wikipedia documents as the source for natural queries users inputted into Google. Each question was manually annotated given a provided Wikipedia document. There is also an open-retrieval version of NQ, OpenNQ (Lee et al., 2019), where the task is to find the answer to the question via retrieval, but it only focuses on the short extractive answers, and therefore does not include the same set of questions as CLAPNQ. This corpus is also considerably larger than our corpus as we just include the Wikipedia documents used in the CLAPNQ questions. Several datasets have been developed from NQ such as AmbigQA (Min et al., 2020), ASQA (Stelmakh et al., 2022), AquaMuse (Kulkarni et al., 2020), AttributedQA (Bohnet et al., 2023), MoQA (Yen et al., 2023), and now CLAPNQ.

Several RAG datasets exist for short extractive answers (e.g., Lee et al., 2019; Adlakha et al., 2022; Bohnet et al., 2023). MoQA (Yen et al., 2023) explores answers of varying length but the long answers are full paragraphs as in the original NQ. Current LFQA datasets include AquaMuse (Kulkarni et al., 2020), ASQA (Stelmakh et al., 2022), ELI5 (Fan et al., 2019), ExpertQA (Malaviya et al., 2024), TruthfulQA (Lin et al., 2022), and WikiHowQA (Deng et al., 2020). ASQA and ELI5 along with QAMPARI (Amouyal et al., 2023) are part of the Automatic LLMs' Citation Evaluation (ALCE) (Gao et al., 2023) benchmark. QAMPARI is not LFQA, but rather multiple short extractive answers. We compare all the LFQA datasets to CLAPNQ in Table 2. Most notably, CLAPNQ is the only dataset to include considerable unanswerable questions, manually annotated answers grounded on a single gold passage, and a corpus for the full RAG pipeline.

The Explain Like I'm 5 (ELI5) dataset consists of questions and responses from the Reddit thread. KILT-ELI5 (Petroni et al., 2021) provides Wikipedia documents that have been retrieved using the questions for benchmarking RAG. However, there are no gold passages and the KILT-ELI5 documents do not necessarily have the answer. The responses written for this sub-Reddit are by subject matter experts (SME) and are often not grounded on any text or passage. Each question is likely to have many responses and

[1] https://github.com/primeqa/clapnq.

they may not all be appropriate or relevant and inter-annotator agreement (IAA) is very low as shown in Table 2. IAA is measured as the mean RougeL F1 score between each pair of annotations for the same question.

TruthfulQA (Lin et al., 2022) has sets of true and false reference answers and a source that supports the reference answers for each question. It is a very small validation dataset as shown in Table 2 that was designed to be adversarial (the questions were intentionally picked to be ones that are answered incorrectly) to probe LLMs. The answers are also considerably shorter than the other LFQA datasets.

WikiHowQA (Deng et al., 2020) is ''How to'' instruction questions from the WikiHow website. For each page, the question is the title and the answer is the context. Only pages that have reference documents are kept. There can be many references for each question. The answers and references are long and have not been manually verified.

ExpertQA (Malaviya et al., 2024) consists of questions that are written by SMEs. They then use GPT-4 and various retriever setups (e.g., Closed-Book, and BM25) to generate several answers and retrieve relevant documents. The experts then evaluate the answers and evidence and can delete claims and evidence that are false and revise if they want to (it is optional). Only one answer was evaluated and revised for each question. Due to the approach of creating the dataset the answers are likely biased by the LLMs (Yu et al., 2023; Navigli et al., 2023).

AquaMuse (Kulkarni et al., 2020) is a summarization dataset using NQ questions that have a long answer (the passage) without a short answer similar to CLAPNQ. However, they use sentence-level matching (by encoding sentences for semantic similarity comparisons) to retrieve up to top 7 documents from Common Crawl while avoiding exact matches as the abstractive dataset. In the extractive version, the sentences in the original long answer are then replaced with the highly semantic similar sentences from the retrieved documents. This means the new summaries are as long as the original passage. The information in the original passage may not be in the retrieved documents.

ASQA (Stelmakh et al., 2022) has the distinctive characteristic that it uses ambiguous questions built from AmbiqQA (Min et al., 2020) which is derived from OpenNQ (Lee et al., 2019). These ambiguous questions tend to need longer answers to disambiguate the multiple aspects of the question. Each answer is generated from one or more passages that answer a specific instance of the question. The answers in the AmbigQA paper are all short and extractive, but in ASQA the explanation to disambiguate the different answers causes them to be long. ASQA is derived from the subset of NQ that has short answers with additional answers for the ambiguity from AmbigQA. Therefore, the gold passages for the multiple answers are not available for all ASQA questions and some of the evidence may not be part of OpenNQ. ASQA is perhaps the most similar to CLAPNQ, with the main differences being: 1) The ASQA answer comes from multiple passages while the CLAPNQ answer is contained in one passage. The gold passages are not maintained for ASQA. 2) The ASQA answers are considerably longer due to the ambiguity of the questions, but also indicating they may not be as concise. 3) We explore additional types of questions that tend to require a long answer such as boolean questions, conjunctive questions, descriptive questions, and questions requiring an explanation. 4) The IAA computed using RougeL for questions that were answered by multiple annotators is much lower than CLAPNQ at 0.48 compared to 0.67.

For a detailed survey of RAG approaches we direct the reader to the comprehensive RAG survey (Gao et al., 2024). It is worth noting that the benchmarks section in this survey is a short paragraph which refers to two datasets (Liu et al., 2023; Chen et al., 2024b) that focus on short extractive answers, attacks and robustness when the passages are purposely adversarial and unfaithful. Furthermore, the datasets questions and responses are created using ChatGPT which likely introduces biases (Yu et al., 2023; Navigli et al., 2023). The former (Liu et al., 2023) does not include retrieval and the latter (Chen et al., 2024b) has fixed retrieved passages instead of a corpus. We believe that this highlights the need for quality datasets (like CLAPNQ) focusing on faithfulness for the full RAG pipeline.

Recently, synthetically generated datasets such as Alpaca (Taori et al., 2023) and Vicuna (Chiang et al., 2023) that include LFQA have been created using LLMs. These datasets can be very large, containing 50k+ conversations, but they're built

to fine-tune LLMs and not applicable as evaluation benchmarks.

3 Dataset

CLAPNQ is created from the subset of Natural Questions (NQ) (Kwiatkowski et al., 2019) that have a long answer (passage) but no short answer. NQ consists of 323k examples. There are around 30,000 questions that are long answers without short answers excluding tables and lists. To increase the likelihood of longer answers we only explored ones that have more than 5 sentences. Each NQ train example is annotated by one person and each NQ dev example is annotated by 5 people. We only explore dev questions where the majority of the annotators agreed it was a long answer without a short answer. 12,657 training and 384 dev examples met our criteria for annotation.

3.1 Annotation Task

CLAPNQ was annotated by 7 skilled in-house annotators paid above minimum wage whose sole jobs are performing Natural Language Processing annotation tasks. The annotation task consisted of two rounds to provide high quality non-consecutive grounded answers to the question. Each task in both rounds took approximately 5 minutes. All annotations were performed on the Appen platform.[2] The details of each round are described below.

The main instruction provided to the annotators was: *Given a question and a passage, find the answer to the question in the passage. Check the boxes for the answer sentences and then copy/paste the relevant text into the answer box.* Finally, after creating an answer from the passage they were asked to *look over the question and answer and make sure it makes sense, is a concise answer, and is grammatically correct.* They had to confirm that they checked all of these things before completing the task. A screenshot of the task is provided in Appendix A, Figure 2.

After initial training and pilots with calibrating of instructions on around 100 questions, each of the NQ questions without a short answer was annotated by one trained annotator in Round 1.

In **Round 1**, the annotators were provided with the question, title, and long answer paragraph from NQ divided into sentences using a sentence tokenizer. The annotators had to select the sentences relevant to the answer and then write a concise answer in their own words with ''copy/pasting'' allowed. The annotators were instructed to write the answer using the selected sentences and that it should make sense, be concise, and grammatically correct. The question could also be skipped.

In **Round 2** of the annotation, all answers from Round 1 that were made up of two or more selected sentences that were *not consecutive* (meaning there was at least one non-selected sentence between them, see example in Table 1) were annotated a second time by a different annotator. These questions were selected as they are more likely to require harmonizing multiple non-contiguous pieces of text. The annotators saw the answer from the first round and could choose to keep the same answer or modify it. Therefore, the second round answers are likely to be of higher quality, however, due to human subjectivity both answers could still be good. In some cases, the Round 2 annotator skipped the question and it is also possible that they changed the answer to no longer be non-consecutive.

The final CLAPNQ dataset consists of all answers that have been annotated by more than one person. We provide the annotations from both rounds if they were different. The IAA using RougeL on the different Round 1 and 2 answers is 0.67, indicating the answers are usually similar. The selected sentences, information regarding the round, and whether the answer is not contiguous is included in the dataset.

3.2 Data Stats

The CLAPNQ dataset of 4,946 questions consists of both answerable and unanswerable questions as described below. The breakdown of the dataset is shown in Table 3. We also include the source of the questions within the original NQ dataset. Since NQ does not release the test set we only explored the train and development sets. Only 67 NQ dev questions qualified with the properties of our task so we use them and additional examples from NQ train as our test set. While the questions and passages are publicly available with NQ, the answers we provide are new. CLAPNQ questions have 1-2 reference answers. The questions are short at 9 words and the answers are long at around 57 words which is 1/3 of the average passage length of 156 words (See Table 2). In addition to the official dataset, we will release

[2] https://www.appen.com/.

Split	No. Questions	Answerable	NQ Source	Unanswerable	NQ Source
Train	3745	1954	Train	1791	Train
Dev	600	300	Train	300	Dev
Test	601	301	Train + 67 Dev	300	Dev
Total	4946	2555		2391	

Table 3: Data stats for CLAPNQ. In addition to providing the number of questions per split we also provide the original source from NQ as we used part of training for the dev and test set.

the round 1 data of 12k questions as training data, referred to as CLAPNQ-R1. Our initial experiments with training using CLAPNQ-R1 did not provide an improvement. We leave further exploration as future work.

3.2.1 Answerable

The answerable data contains the original question and gold passage (P) as well as the relevant sentences (RS) and answers (A) created by the annotators as described in the previous section. The Precision, Recall (R), and F1 scores for RougeL$_{(RS,P)}$ is 100/45/59 and for RougeL$_{(A,RS)}$ it is 92/72/79, respectively. The former scores are a sentence retrieval task, the latter a generative task. RougeL$_{(A,P)}$ is 94/32/46. The retrieval stage reduces the content by about 2x (R = 45) and the generation case reduces another 30% (R = 72) for a total reduction From P to A of approximately 3x (R = 32).

3.2.2 Unanswerable

A similar amount of unanswerable questions from NQ were extracted to complete the CLAPNQ dataset. In the NQ training set there is only one annotation, in the NQ dev s et all 5 annotators must have said it was unanswerable. The unanswerable questions were randomly chosen from examples that had more than 5 sentences in the passage by matching the first word distribution of the answerable questions. For example, in CLAPNQ, *What* and *Where* are the most common question types while *Who* is the most common question type for the NQ short answers. Since NQ does not have a gold passage for unanswerable questions, a random passage is chosen from the provided Wikipedia document. This passage is used in the generation experiments as the ‘‘gold’’ passage to indicate the question is unanswerable with this passage.

3.3 Retrieval Corpus

We provide a corpus that can be used to build an index for querying CLAPNQ in a retrieval setting. It is built using the passages[3] from the original Wikipedia NQ documents used in the CLAPNQ dataset including the answerable and unanswerable questions.

In some cases there were slightly different versions of the same document. We only kept one to avoid duplicate passage retrieval and ensure that all gold passages are present in the corpus.[4] The corpus includes 178,891 passages from 4,293 documents, of which 2,345 passages have questions associated with them across the 4,946 train, dev, and test answerable and unanswerable splits.[5]

4 Experiments and Results

We present baseline experiments on CLAPNQ for Retrieval, Generation, and the full RAG pipeline. An exhaustive implementation of methods and training setups is beyond the scope of this paper; we provide results to illustrate how CLAPNQ performs using common and SOTA approaches.

We report the commonly used retrieval metrics of nDCG@10 and Recall@10 for retrieval. We report several metrics to illustrate generation performance. Each of our metrics correlate with one of the CLAPNQ properties described in the introduction. The first two are the commonly used RougeL and Recall (this is the same as Rouge1). RougeL can be considered a good approximation for how *cohesive* the answer is as it will give more credit to longer spans. Recall is a good approximation for *completeness*. We also provide

[3]Very long (>3000 words) and short passages (<15 words) that are not gold answerable passages were discarded.

[4]We confirmed that the gold passage had very high overlap (RougeL > .90) to the alternative version of the passage or added it as an additional passage for the document (28 times).

[5]There is usually one gold passage, but 14 questions from the NQ dev set have two gold passages. Both are kept in retrieval, but only the more frequent one has a gold answer.

	DEV					TEST				
	nDCG				R	nDCG				R
Model	@1	@3	@5	@10	@10	@1	@3	@5	@10	@10
BM25	18	30	35	40	67	20	31	36	40	64
all-MiniLM-L6-v2	29	43	48	53	79	30	45	51	55	83
BGE-base	37	54	59	61	85	43	57	63	65	88
E5-base-v2	41	57	61	64	87	42	57	61	65	88

Table 4: Retrieval results on the answerable questions using nDCG @1, 3, 5, 10 and Recall@10 as metrics on the dev and test sets. We report several nDCG@k to illustrate the impact on the RAG task.

RougeL$_p$ which is an extractiveness metric that measures how *faithful* the response is. It computes the RougeL of the answer to the passage. Since CLAPNQ is extractive, we would expect a good system to have a high RougeL$_p$. In addition, we also provide the length (in characters) of the answer. We notice that length is a strong indicator of how well a model performs with answers that are close to the reference length being desirable, it is therefore a good approximating for how *concise* the answer is. Looking at all four of these metrics helps provide a comprehensive picture of model performance. Finally, we also provide the *unanswerable* accuracy. The output is considered unanswerable if its answer string indicates it is unanswerable, e.g., ‘‘I don’t know". The unanswerable strings differ per model.

4.1 Retrieval

We present retrieval results on popular public SOTA[6] base-size (768 embedding dimension) retrieval dense embedding models E5 (Wang et al., 2024), BGE (Chen et al., 2024a), and allMiniLM[7] (384 embedding dimension) in addition to BM25 (Robertson, 2009) by ingesting the CLAPNQ corpus described in Section 3.3. We ran the ingestion and evaluation for the embedding models using sentence transformers from the BEIR repository[8] keeping all default parameters, and we used ElasticSearch[9] for BM25 with a maximum passage length of 512 tokens. Passages that exceeded the length were divided with an overlap stride of 256. We provide nDCG results for 1, 3, and 5 in addition to 10 to illustrate the potential impact on the full RAG pipeline which we report in Section 4.3. The retrieval results are shown on the answerable questions from the CLAPNQ dev and test sets in Table 4. The E5-Base model performs best with nDCG@10 of 64 on the dev set and E5-base and BGE-base have the same performance of nDCG@10 of 65 on the test set. All these models include NQ as part of their training.

4.2 Generation

The generation task is: *Given a question and the gold relevant passage, generate an answer to the question*. The CLAPNQ dataset is designed to be faithful and concise so the generated response should have these properties.

We ran generation experiments with three families of models: Encoder-Decoder, Decoder LLMs, and Fine-Tuned Encoder Decoder. We also compare to a full passage baseline. The generation task is sent to the model using a prompt. Most models use an NQ prompt taken from FLAN-T5 (Chung et al., 2024). GPT and Llama have prompts based on their model suggestions, all prompts are provided in Appendix B. In our zero-shot setup the models were provided with the question, context, and prompt. In the 1-shot setup (1/0) the model was provided with the same answerable example from CLAPNQ training and in the 2-shot setup (1/1) the model was also provided with the same unanswerable question for the same passage. The generation results are shown in Table 5. A human evaluation and discussion is in Sections 5 and 7.

Encoder Decoder Models. We use FLAN-T5-Large and FLAN-T5-XXL for zero and few-shot experiments. We chose FLAN-T5 as it has already been trained on the NQ dataset and should therefore already be familiar with the task. The FLAN-T5 models, which are fine-tuned on short extractive tasks, like to provide short answers and therefore have poor Recall. The few-shot experiments outperform the zero-shot experiments,

[6]See the Retrieval tab of the MTEB leaderboard: https://huggingface.co/spaces/mteb/leaderboard.

[7]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.

[8]https://github.com/beir-cellar/beir/.

[9]https://www.elastic.co/elasticsearch.

		DEV					TEST				
		Answerable				Un-	Answerable				Un-
Model	FS	RougeL	R	RougeL$_p$	Len	ans%	RougeL	R	RougeL$_p$	Len	ans%
FLAN-T5-Large	–	18.6	11.8	7.1	33	79.9	13.8	8.5	5.0	27	83.6
FLAN-T5-Large	1/0	22.0	14.6	8.8	41	77.3	17.1	11.4	6.9	36	82.6
FLAN-T5-Large	1/1	20.3	13.4	8.1	38	81.7	16.3	10.4	6.1	34	85.3
FLAN-T5-XXL	–	22.1	15.0	10.0	45	84.0	22.0	15.6	9.7	56	91.5
FLAN-T5-XXL	1/0	31.9	23.6	15.0	75	78.1	28.9	21.1	14.3	76	84.9
FLAN-T5-XXL	1/1	28.3	21.1	13.0	63	84.8	24.0	17.2	11.4	63	89.2
Llama-13B-chat	–	35.5	64.3	34.0	491	25.0	35.0	61.3	34.0	491	27.4
GPT 4	–	35.9	67.7	30.0	759	18.0	33.4	65.1	30.3	797	22.2
Mistral-7B-Instruct	–	39.0	56.0	29.0	384	18.6	35.4	53.4	29.2	411	16.3
GPT 3.5	–	39.8	58.9	30.0	444	37.0	40.3	56.3	29.9	375	31.3
CLAPNQ-T5-LG-200	–	41.5	51.3	42.1	272	89.7	40.5	49.2	39.0	271	92.0
CLAPNQ-T5-LG	–	57.2	68.3	51.0	318	89.2	57.7	69.5	51.7	351	86.8
Full Passage	–	49.5	97.4	100.0	912	0.0	49.2	98.7	100.0	1039	0.0

Table 5: Generation results with the gold passage using RougeL, Recall, RougeL$_p$, Length, and Unanswerable accuracy as metrics. Experiments using pre-trained models, few-shot (1 answerable / 1 unanswerable examples), the fine-tuned model, CLAPNQ-T5-LG, and a full passage baseline.

but providing an unanswerable example has a trade-off of improving the unanswerable metrics while reducing the answerable metrics.

Decoder LLMs. We explored several SOTA Decoder models: LLama, Mistral, GPT 3.5 turbo, and GPT 4 turbo. The SOTA LLMs have poor unanswerable performance but better recall. They do not like to say ‘‘I don’t know" and almost always provide an answer. This is evident with all models but worst with Mistral and GPT 4. Interestingly, GPT 3.5 performed better than GPT 4, particularly for unanswerable. The LLMs tend to provide answers that are far too long, particularly for GPT 4 at an average of 759 /797 characters, and therefore are not concise. This is apparent from the high Recall but low RougeL. The low RougeL$_p$ indicates that the answers may not be faithful to the passage.

Fine-Tuned Encoder Decoder Model. We use FLAN-T5-Large for our fine-tuned (FT) experiment, which we call CLAPNQ-T5-LG (see implementation details in Appendix C). CLAPNQ-T5-LG has good unanswerable performance and good recall. It is clear that the answers are concise and it learns the appropriate answer length. It is closest to the average length of the reference responses which is 272 dev and 300 test characters. RougeL and Recall highlight that the answers are most cohesive and complete and RougeL$_p$ shows that it learns to extract the answer from the passage, while the other models are considerably less extractive.

We also explore a smaller training size to help measure whether performance can be improved when a small amount of labeled data is available. This is an important use case because labeling data in a new domain is costly. We call this experiment CLAPNQ-T5-LG-200 as it was trained using 200 examples (an equal amount of answerable and unanswerable questions) with 10 random samples and report the average. The RougeL and unanswerable metrics are better than the SOTA Decoder LLMs, but worse than training on the full dataset. The model tends to say unanswerable too much.

Full Passage Baseline. We compare to a baseline where the entire passage is taken as the answer. This performs very well in the automated metrics but it is clearly not concise as indicated by the length. The RougeL score highlights the difference of the LLMs to CLAPNQ-T5-LG which are considerably lower than providing the full passage. The difference between the average length of the generated answers, the reference answer, and the passage length are an indicator of how difficult the extraction task is. The answer must discard two thirds of the passage to be appropriately concise.

		DEV					TEST				
		Answerable				Un-	Answerable				Un-
Retriever	Generator	RougeL	R	RougeL$_p$	Len	ans%	RougeL	R	RougeL$_p$	Len	ans%
GOLD	GPT 3.5	39.8	58.9	30.0	444	37.0	40.3	56.3	29.9	375	31.3
E5-base-v2	GPT 3.5	34.0	52.8	30.0	459	27.3	35.0	48.9	31.4	373	20.2
GOLD	Mistral-7B-Instruct	39.0	56.0	29.0	384	18.6	35.4	53.4	29.2	411	16.3
E5-base-v2	Mistral-7B-Instruct	31.3	49.4	30.1	436	11.7	29.4	47.5	29.9	463	9.3
GOLD	CLAPNQ-T5-LG	57.3	68.3	51.0	317	89.5	57.8	69.5	51.7	351	86.8
all-MiniLM-L6v2	CLAPNQ-T5-LG	36.6	46.4	52.6	300	49.8	37.9	48.7	52.9	323	47.0
BGE-base	CLAPNQ-T5-LG	40.7	52.3	54.2	331	41.9	41.7	52.4	54.8	331	44.4
E5-base-v2	CLAPNQ-T5-LG	42.8	54.3	53.8	343	40.1	41.6	51.3	55.7	321	45.9
E5-base-v2	E5-CLAPNQ-T5-LG	30.4	37.5	34.3	204	82.7	26.7	32.9	33.0	195	84.6
E5-base-v2	E5-G-CLAPNQ-T5-LG	33.3	40.4	37.0	227	78.8	34.5	41.8	38.0	236	81.0

Table 6: Full RAG results with top 3 passages on CLAPNQ-T5-LG and LLMs using various retrievers. The metrics reported are RougeL, Recall, RougeL$_p$, Length, and Unanswerable accuracy. Each RAG setup can be compared to its GOLD setup where there is no retrieval.

4.3 Full RAG Pipeline

In our full RAG pipeline experiments we retrieve the top passages using the best performing retrieval model, E5-base-v2, and then perform generation on the same prompts as in Section 4.2, however instead of the gold passage, the top retrieved passages are included in the prompt. It is possible that the gold passage will not be in the top *N* passages, making the question unanswerable based on retrieval. The RAG task is far more difficult than the GOLD generation task as the model needs to learn which passages are irrelevant to the question. We experimented with including the top 3 and top 5 passages in the prompt. Based on the retrieval results in Table 4, 5 documents has a 4-point improvement over 3 documents. However, in our experiments including 5 passages in the prompt increased the noise and did not provide an improvement.

In the RAG experiments we explored each dense retriever with CLAPNQ-T5-LG, and the best retriever on the dev set, E5 Base, with the best performing generation models: GPT 3.5, Mistral-7b-Instruct, and CLAPNQ-T5-LG. Results are shown in Table 6 and we compare against the best GOLD generation baselines for each model from Table 5 to show the gap for RAG. GOLD can be considered as an upper bound as we would not expect the retriever to perform better than having only the grounded passage for the automated metrics. In all cases performance drops considerably for CLAPNQ-T5-LG with a very large drop in % unanswerable. Performance is also reduced for zero-shot GPT 3.5 and Mistral but not as much as CLAPNQ-T5-LG. A human evaluation and discussion that compares RAG to Gold is in Sections 5 and 7.

We also explored two fine-tuned models that incorporated RAG during training. They follow the same approach as CLAPNQ-T5-LG, but instead of the gold passage, the top 3 retrieval passages are included during training. In the second version, E5-G-CLAPNQ-T5-LG we ensure the gold passage is kept in the top 3 passages during training, at a randomly chosen position, even if it was not originally included. These models perform better on the unanswerable questions than CLAPNQ-T5-LG but much worse on the answerable questions. The RougeL score of E5-G-CLAPNQ-T5-LG (51.6/52.1) on the answerable questions that were answered is better than CLAPNQ-T5-LG (46.7/44.5) for the dev and test sets, but only a little more than half the answerable questions were answered. We suspect the discrepancy for unanswerables between the GOLD experiment and RAG trained models (89.5 vs 82.7/78.8) is because many questions are no longer unanswerable in the RAG setting as we will show in the human evaluation. We leave further experimentation on optimizing these models as future work.

5 Human Evaluation

In addition to reporting automated metrics we also performed a human evaluation on the GOLD and RAG setups to explore how appropriate and faithful users think the responses are as used in the literature (Es et al., 2024). For each question

	Model	Faithful	Approp	F+A	Win-Rate
Gold	CLAPNQ-T5-LG	3.7	3.7	3.7	66%
	GPT 3.5	3.3	3.6	3.4	34%
	Reference	3.9	3.8	3.8	57%
RAG	CLAPNQ-T5-LG	3.8	3.2	3.4	42%
	GPT 3.5	3.0	3.6	3.2	35%
	Reference	3.0	3.5	3.0	33%

Table 7: Human evaluation metrics on Faithful (F) and Appropriate (A) on a 4-point scale and win-rate. F+A is the harmonic mean of F and A.

and answer, we asked three annotators to indicate on a scale of 1 (No) - 4 (Yes) whether the answer looks appropriate (i.e., looks correct or answer relevance) and whether it is faithful to the passage. These metrics are only measured for the answerable questions. During the RAG evaluation we also asked the annotators to select which of the top 3 retrieved passages were relevant to the answering the question. If a question was marked faithful, we asked the annotators to select which passages were relevant to the answer. Finally, they performed a pair-wise comparison of the answers to indicate preference to compute win-rate. Ties were acceptable but they were asked to do so sparingly. The answers were shown to the annotators randomly and they did not know which model produced the answer. Instructions and a task screenshot are in Appendix A.

The human evaluation was for the GOLD and RAG setups. A total of 40 answerable and 10 unanswerable questions, with an equal amount of questions were randomly sampled from both the dev and test sets being included for each setup. The annotators that performed this task are the same annotators that worked on creating the dataset, however these annotations were done at a later time period. We compare CLAPNQ-T5-LG, GPT 3.5 (The best performing decoder LLM), and the reference answer. The evaluation is shown in Table 7.

In the GOLD setup, agreement was high for appropriateness (73%), faithfulness (88%), and win-rate (86%). The annotators preferred the CLAPNQ-T5-LG answers the most and GPT 3.5 answers the least. We investigated several examples where the CLAPNQ-T5-LG answers were preferred to the reference answer and both answers were good but the annotators preferred the direct copying by CLAPNQ-T5-LG. The reference and CLAPNQ-T5-LG answers were highly faithful and appropriate but GPT 3.5 was less faithful. This highlights the importance of being faithful to the passage as an answer can look correct but not be grounded in the passage which may indicate factually incorrect answers. The human evaluation shows that a model can successfully learn to generate faithful and appropriate responses, but the SOTA LLM models don't perform as well on this task.

In the RAG setup, agreement was very high for faithfulness (91%) and win-rate (90%) but much lower for appropriateness (68%). The annotators preferred the CLAPNQ-T5-LG answers the most, with little difference in preference between the reference and GPT 3.5 answers. The CLAPNQ-T5-LG answers were very faithful while GPT 3.5 and the reference were less faithful. The GPT 3.5 and reference answers were more appropriate while CLAPNQ-T5-LG was least appropriate. The changes from the GOLD setup highlight the importance of evaluating the RAG pipeline. The reference answers may not be in the retrieved passages even though they are correct. However, being faithful to the passages can provide an inappropriate answer if the retrieved passages are not relevant to the question. According to two or more annotators, 26/40 answerable questions had multiple relevant passages and 4/40 had no relevant passages. Additionally, 38, 39, and 32 of CLAPNQ-T5-LG, GPT 3.5, and reference responses were considered faithful to one or more passages. Fifty percent of the unanswerable questions had relevant passages.

6 Question Style Impact

We explore the differences between CLAPNQ and ASQA by evaluating how they compare to each other. As described in Section 2, the ASQA dataset is most similar to CLAPNQ, with a few key differences: The ASQA dataset was also created from a subset of NQ, but the questions come from the short extractive answers in NQ so there is no overlap with CLAPNQ. The ASQA questions are ambiguous and several passages are needed to generate the target answer while CLAPNQ questions are of different types and one passage is needed to generate the target answer. The answers in ASQA tend to be considerably longer on average, at 492 characters compared to 318 characters for CLAPNQ.

	CLAPNQ				ASQA			
	RougeL	R	RougeL$_p$	Len	RougeL	R	RougeL$_p$	Len
ASQA-T5-LG	51.3	77.5	67.7	551	54.9	69.5	89.5	542
CLAPNQ-T5-LG-ANS	62.8	74.1	54.8	364	44.9	42.1	72.4	217
CLAPNQ+ASQA-T5-LG-ANS	61.3	74.2	57.0	381	55.9	68.2	88.3	499

Table 8: A comparison of ASQA dev (there is no ASQA test) and CLAPNQ test using models trained on each dataset and the two datasets combined. Since ASQA does not have unanswerable questions, all models and results are on answerable questions only.

We compare the datasets by running generation experiments in the GOLD setup and a small human evaluation. Since ASQA only provides a dev set, we consider that to be the test set and only compare it to the CLAPNQ test set. Our human evaluation experiment was completed on 20 random questions from each dataset.

We fine-tuned an encoder-decoder model for each dataset, as well as a model on both datasets for generation experiments using Flan-T5-LG. Since ASQA only has answerable questions, we only explore the answerable questions for both datasets and train a CLAPNQ-T5-LG-ANS model on the CLAPNQ answerable subset. (Implementation details are described in Appendix C).

ASQA does not provide all of the context paragraphs needed to generate the answer. They provide the gold passages from the original NQ but only supply ‘‘knowledge’’ which is small snippets of text from any Wikipedia document that the annotators felt was needed to make the answer. In our experiments we mimic the oracle setting described in the ASQA paper (Stelmakh et al., 2022)[10] by including the context passages and the knowledge as the input context.

The generation results are shown in Table 8 and the human evaluation results are shown in Table 9. The generation experiments show that each model performs significantly better on its own dataset while the model fine-tuned on both datasets does well on both. The human evaluation shows that ASQA clearly does not perform well on CLAPNQ while the other models perform as well as the reference data (the ASQA answers are faithful but not appropriate). An investigation into some of these examples shows that they had extra information that was not needed to answer the question. On the other hand the human evaluation shows that CLAPNQ does perform well on ASQA but the reference data has a low win-rate (the reference answers are appropriate but not faithful). An investigation into some of these examples showed that there was missing evidence for parts of the answer which can encourage hallucination. We also looked at some cases where CLAPNQ did not do well and found that sometimes it only gave one answer instead of the two answers that needed to be distinguished due to ambiguity. These experiments show that the datasets compliment each other and that CLAPNQ is a more faithful and concise dataset.

[10]In the paper they only use the longer answer as the reference, while we compare to both gold answers as in our CLAPNQ experiments.

	Model	Faithful	Approp	F+A	Win-Rate
CLAPNQ	ASQA	3.9	2.7	3.4	38%
	CLAPNQ	4.0	3.1	3.5	68%
	CLAPNQ+ASQA	3.9	3.2	3.5	65%
	Reference	3.9	3.7	3.7	67%
ASQA	ASQA	3.7	3.1	3.3	63%
	CLAPNQ	3.8	3.2	3.4	65%
	CLAPNQ+ASQA	3.7	3.1	3.2	61%
	Reference	3.2	3.3	3.2	40%

Table 9: Human evaluation metrics comparing ASQA and CLAPNQ on Faithful (F) and Appropriate (A) on a 4-point scale and win-rate. F+A is the harmonic mean of F and A. All models are T5-LG fine-tuned models trained on answerable data only.

7 Discussion

In this section we describe some challenges we've encountered. We describe them here and provide examples in Appendix D.

Unanswerable Questions: While it is unlikely that the unanswerable questions have an answer in the randomly picked passage, we find that in some cases, there is actually an answer (Appendix D,

Table 10). There are other cases where the answer to an unanswerable question may appear correct when looking at the passage, but the passage may not be relevant (Appendix D, Table 11).

Generation: GPT 3.5 and Mistral will have answers that are correct but not faithful to the passage (Appendix D, Table 12). Since the prompts request that the answer use the passage, such an answer should not be provided, or the response should explain that the answer was found elsewhere. In many cases GPT 3.5 and Mistral give an answer that is considerably longer than CLAPNQ-T5-LG and the reference. The recall is high, but the answer is not concise and has extra irrelevant information. During the human evaluation the annotators tend to prefer the concise answers and will often mark long answers as less appropriate.

RAG: The answers can change considerably due to the multiple passages in RAG compared to GOLD (Appendix D, Table 13). In the RAG setting the automated metrics are much lower than the GOLD setting. However, the answers may be good but just have different information which was found only in the provided passages (Appendix D, Table 13).

If irrelevant passages are retrieved, the reference answer will have low extractiveness, but the other answers may still be incorrect while being grounded which is difficult to identify without human evaluation.

8 Future Directions

The automated evaluation, human evaluation, and discussion highlight several areas of future directions: 1) Unanswerable Questions: Many of the LLMs struggle with the unanswerable questions and often try to provide an answer. 2) Concise Answers: Many of the LLMs like to provide very long answers that are not concise, which is not preferred by humans. 3) Irrelevant Retrieval: The models will try to answer RAG questions even when the passages are irrelevant, either by being unfaithful or incorrect. 4) Multiple correct answers: It is harder to evaluate RAG correctly because the answers could be correct but different than the gold. 5) Dataset Enhancements: We hope to add more grounded reference answers, a multilingual version, and other domains.

9 Conclusion

We have presented CLAPNQ, a new benchmark dataset for evaluating the *full* RAG pipeline. CLAPNQ has the properties of being concise, complete, cohesive, faithful to the passage and unanswerable questions. A FT model can perform well when the correct passages are provided during retrieval, while SOTA LLMs are behind in faithfulness, conciseness and unanswerability. Finally, we've provided a human evaluation, discussion, and specific areas of future improvements. CLAPNQ is publicly available at `https://github.com/primeqa/clapnq`.

Ethics Statement

Limitations

As with any manually annotated dataset, there are likely to be some incorrect and unclear answers. We did out best to mitigate this as described in Section 3. We believe that, in general, the dataset quality is strong and can be used as is as a benchmark for RAG. CLAPNQ is built from Natural Questions (Kwiatkowski et al., 2019), therefore any limitations in Natural Questions and Wikipedia may also be present in CLAPNQ.

Intended Use

CLAPNQ and CLAPNQ-T5-LG are intended to be used to advance research in RAG. CLAPNQ is being released with an Apache 2.0 license. We do not approve of any adversarial or harmful uses of our work.

Biases

NQ train and dev have been included in training of most, if not all, LLMs which may lead to biases, particularly since CLAPNQ dev is part of NQ train. However, all models have this same advantage. While the questions and passages have been seen by all models the CLAPNQ answers are new and remain hidden. Any biases in NQ and Wikipedia may also be present in CLAPNQ.

Acknowledgments

We would like to thank our annotators for their high quality work creating and evaluating this dataset: Mohamed Nasr, Joekie Gurski, Hee Dong Lee, Roxana Passaro, Marina Variano, and Chie Ugumori. We also thank Arafat Sultan for initial annotations and analysis on the pilot task.

We would like to thank our TACL action editor, Jimmy Lin, and the anonymous reviewers for their helpful feedback.

References

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. TopiOCQA: Open-domain conversational question answering with topic switching. *Transactions of the Association for Computational Linguistics*, 10:468–483. https://doi.org/10.1162/tacl_a_00471

Samuel Amouyal, Tomer Wolfson, Ohad Rubin, Ori Yoran, Jonathan Herzig, and Jonathan Berant. 2023. QAMPARI: A benchmark for open-domain questions with many answers. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 97–110, Singapore. Association for Computational Linguistics.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Lierni Sestorain Saralegui, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. 2023. Attributed question answering: Evaluation and modeling for attributed large language models. *ArXiv preprint:2212.08037 v2*

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *ArXiv preprint:2402.03216 v4*.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024b. Benchmarking large language models in retrieval-augmented generation. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20–27, 2024, Vancouver, Canada*, pages 17754–17762. AAAI Press. https://doi.org/10.1609/aaai.v38i16.29728

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Yang Deng, Wai Lam, Yuexiang Xie, Daoyuan Chen, Yaliang Li, Min Yang, and Ying Shen. 2020. Joint learning of answer selection and answer summary generation in community question answering. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7–12, 2020*, pages 7651–7658. AAAI Press. https://doi.org/10.1609/aaai.v34i05.6266

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. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1346

Adam Fisch, Alon Talmor, Danqi Chen, Eunsol Choi, Minjoon Seo, Patrick Lewis, Robin Jia, and Sewon Min, editors. 2021. *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*. Association for Computational Linguistics, Punta Cana, Dominican Republic.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.398

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 preprint:2312.10997 v5*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-augmented language model pre-training. *arXiv preprint:2002.08909 v1*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint: 2310.06825 v1*.

Sayali Kulkarni, Sheide Chammas, Wan Zhu, Fei Sha, and Eugene Ie. 2020. AQuaMuSe: Automatically generating datasets for query-based multi-document summarization. *arXiv preprint: 2010.12694 v1*.

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. https://doi.org/10.1162/tacl_a_00276

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1612

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20. Red Hook, NY, USA. Curran Associates Inc.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.229

Yi Liu, Lianzhe Huang, Shicheng Li, Sishuo Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. RECALL: A benchmark for llms robustness against external counterfactual knowledge. *arXiv preprint: 2311.08147 v1*.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2024. ExpertQA: Expert-curated questions and attributed answers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3025–3045, Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.167

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK. https://doi.org/10.1017/CBO9780511809071

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.466

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: Origins, inventory, and discussion. *Journal of Data and Information Quality*, 15(2). https://doi.org/10.1145/3597307

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: A benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.200

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-2124

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1264

S. Robertson. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389. https://doi.org/10.1561/1500000019

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. QA dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Computing Surveys*, 55(10). https://doi.org/10.1145/3560260

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.566

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra,

Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. LLaMA 2: Open foundation and fine-tuned chat models. *arXiv preprint: 2307.09288 v1*.

Ellen M. Voorhees and Donna K. Harman. 2005. *TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing)*. MIT Press.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024. Text embeddings by weakly-supervised contrastive pre-training. *ArXiv preprint:2212.03533 v2*.

Howard Yen, Tianyu Gao, Jinhyuk Lee, and Danqi Chen. 2023. MoQA: Benchmarking multi-type open-domain question answering. In *Proceedings of the Third DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 8–29, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.dialdoc-1.2

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J. Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as attributed training data generator: A tale of diversity and bias. In *Advances in Neural Information Processing Systems*, volume 36, pages 55734–55784. Curran Associates, Inc.

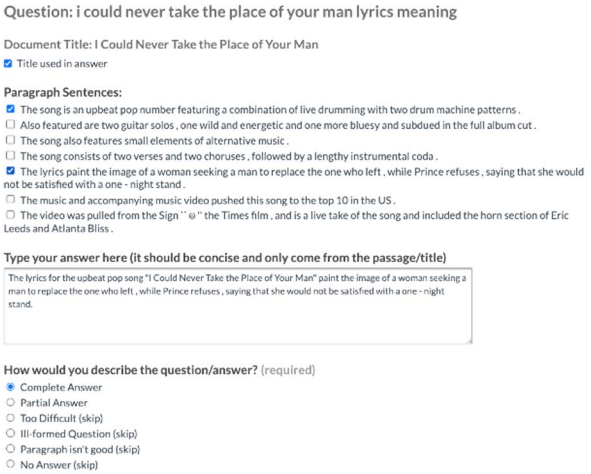

Figure 2: The Round 1 annotation task for CLAPNQ. The annotator had to select the title/sentences needed to answer the question, and then provide a concise answer.

A Annotation Tasks

All annotation tasks were performed using Appen. They are described in Sections 3 and 5 of the main paper. We provide screenshots and further instructions below.

A.1 Dataset Creation

The CLAPNQ dataset was created in two rounds. A screenshot of round 1 is shown in Figure 2. A small handful of the questions (1 in train, and 9 in dev) are high-quality annotations from the initial pilot rounds. These examples have several reference answers.

A.2 Human Evaluation

The human evaluation was performed a portion of the dev and test sets. Human evaluation on the GOLD generation task is shown in Figure 3. The RAG version had two additional questions regarding passage relevance as described in Section 5. We plan on releasing the human evaluation annotations as part of the dataset release. The general instructions to the annotator were as follows: *In this task, you will review the same question and passage and, for each one, rate the quality of the answer to the question. On each page, you will see 3 different answers to the same question. Read the question and passage and answer how well you are confident in the question, passage, and know the correct answer. For each model answer, (given the same context and passage): The answer to the model is in red. Please make your judgements on*

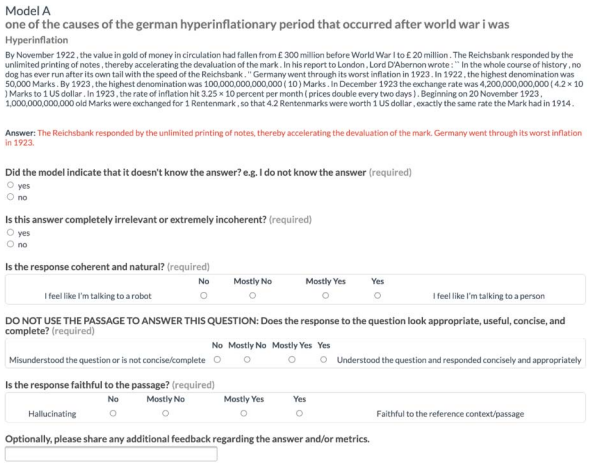

Figure 3: The human evaluation task used to compare the model answers in random order. The individual questions per answer are shown here for one model.

this red answer span. indicate if the answer is an ''I don't know'' or if the answer is completely incoherent. For each model response, answer the following questions on a scale of 1–4: 1) DO NOT USE THE PASSAGE TO ANSWER THIS QUESTION: Does the response to the question look appropriate, useful, concise, and complete? 2) Is the response faithful to the passage? Evaluate each metric independently. Finally, also perform a head to head comparison of the model responses by answering the following question for every pair of answers: Which response do you prefer in terms of faithfulness, appropriateness and naturalness?

The win-rate is computed per model per question for all annotators. If there are three models and three annotators being compared a model can win up to six times per question (or (# models - 1) × # annotators). The score per model per question is computed as wins/6 which is then averaged over all questions for the final model score.

B Prompts

The Flan-T5 (Chung et al., 2024) prompt which was used for most models is: {title}: {passage} Please answer a question about this article. If the question is unanswerable, say ''unanswerable''. user: {question}, answer:

The GPT Prompt is based on chat completion from OpenAI[11]: {'role': 'system', 'content': ''Generate next agent response, given the following document(s). If you cannot base your answer on the document, please state that you do not have an answer.'}, {'role': 'system', 'content': ''[title]: {title} [document]: {passage}, {'role': 'user', 'content': question}'}

The Llama Prompt is the default Llama 2 prompt (Touvron et al., 2023): <s>[INST] <<SYS>> You are a helpful, respectful and honest assistant. Always answer as helpfully as possible, while being safe. Your answers should not include any harmful, unethical, racist, sexist, toxic, dangerous, or illegal content. Please ensure that your responses are socially unbiased and positive in nature. If a question does not make any sense, or is not factually coherent, explain why instead of answering something not correct. If you don't know the answer to a question, please don't share false information. <</SYS>> [document]: {title} {passage}. [conversation]: question},. Answer with no more than 150 words. If you cannot base your answer on the above document(s), please state that you do not have an answer. [/INST].

C Implementation Details

We used HuggingFace transformers[12] for all training experiments. We experimented with several variations and our final CLAPNQ-T5-LG model was trained for 6 epochs, checking after each epoch against the development set to keep the best model. CLAPNQ-T5-LG has a learning rate of $1e-4$, batch size of 32, max input length of 412 tokens with an output length of 100 tokens. We kept the T5 context length during training to keep within GPU constraints and improve training speed. We cut off the end of the 368 passages (10% of the training data) that did not fit in the context. The prompts were not truncated during evaluation.

In the small sample size experiments, we accommodate for the smaller size by increasing the learning rate to $1e-3$. In the RAG CLAPNQ-T5-LG experiments, the context is considerably longer so we increase the context size to 1024. This maintains a similar amount of data that needs to be cut off as in the other experiments at 317 and 346 for the E5-CLAPNQ-T5-LG and E5-G-CLAPNQ-T5-LG models (<10%). We accommodate for these experiments with the longer context size by using

[11] https://learn.microsoft.com/en-us/azure/ai-services/openai/reference.

[12] https://huggingface.co/docs/transformers/model_doc/flan-t5.

Question: who are the characters in the black cat
Passage: Sephiria Arks (Sefiria Ākusu) , known as Number I, is the commander of the Numbers and is trusted enough by Chronos to be given total control over killing Creed . Sephiria was one of the few members who were ‘‘born for Chronos’’ , as she stated that when she was an infant she underwent surgery to enhance her recovery time from injuries . Sven once commented that her eyes allowed her to stare into the recesses of one ’s soul . This empathy is her greatest strength, which allows her to utilize her powers of manipulation to their utmost extent . She will do anything to finish her missions and will sacrifice whatever it takes in order to succeed in meeting an objective . However, she does show some inner conflict in her manipulation, even questioning Belze about her actions . She controls all of the Chronos Numbers ’ actions as well as many of the independent operatives in the hunt for Creed with little interference from the Elders .
Responses
GPT 3.5
In the Black Cat series, there are several characters. One of the main characters is Sephiria Arks, also known as Number I. She is the commander of the Numbers and has complete control over killing Creed. Sephiria underwent surgery as an infant to enhance her recovery time from injuries and has the ability to stare into the recesses of one’s soul. She is highly skilled in manipulation and will do whatever it takes to succeed in her missions. Sephiria controls the actions of the Chronos Numbers and many independent operatives in the hunt for Creed.
CLAPNQ-T5-LG
Sephiria Arks, known as Number I, is the commander of the Numbers and is trusted enough by Chronos to be given total control over killing Creed. Sven once commented that her eyes allowed her to stare into the recesses of one’s soul. This empathy is her greatest strength, which allows her to utilize her powers of manipulation to their utmost extent.
Reference
No answer

Table 10: This is an Unanswerable question that is actually partially answerable from the passage. It only mentions one character.

Question: what type of volcano is one tree hill
Passage: Maungakiekie One Tree Hill Maungakiekie / One Tree Hill after the removal of the ‘‘one tree’’ . Highest point Elevation 182 m (597 ft) Coordinates 36 °54 ’ 0’’ S 174 °46 ’ 59’’ E · / · 36.90000 °S 174.78306 °E · / - 36.90000; 174.78306 Coordinates: 36 °54 ’ 0’’ S 174 °46 ’ 59’’ E · / · 36.90000 °S 174.78306 °E · / - 36.90000; 174.78306 Geography Location North Island, New Zealand Geology Volcanic arc / belt Auckland volcanic field Location
Responses
GPT 3.5
One Tree Hill is part of the Auckland volcanic field, which is a volcanic arc/belt.
CLAPNQ-T5-LG
No answer
Reference
No answer

Table 11: This is a difficult example which confused most annotators during the human evaluation. It appears from the passage that the GPT answer may be correct but the answer is ‘‘scoria cone volcano’’ which is not in the passage.

a batch size of 8 and 10 epochs. In the experiments comparing CLAPNQ and ASQA we also used the longer context, epochs, and smaller batch size to accomodate for the longer length of ASQA questions and answers.

D Examples

We provide several examples of output comparing the various LLMs as described in Section 7. See Tables 10–13.

Question: who is given credit for inventing the printing press

Passage: The world 's first movable type printing press technology for printing paper books was made of porcelain materials and was invented around AD 1040 in China during the Northern Song Dynasty by the inventor Bi Sheng (990–1051) . Subsequently in 1377, the world 's oldest extant movable metal print book, Jikji, was printed in Korea during the Goryeo dynasty . Because of this, the diffusion of both movable - type systems was, to some degree, limited to primarily East Asia, although various sporadic reports of movable type technology were brought back to Europe by Christian missionaries, traders and business people who were returning to Europe after having worked in China for several years and influenced the development of printing technology in Europe . Some of these medieval European accounts are still preserved in the library archives of the Vatican and Oxford University among many others . Around 1450 Johannes Gutenberg, introduced the metal movable - type printing press in Europe, along with innovations in casting the type based on a matrix and hand mould . The small number of alphabetic characters needed for European languages was an important factor . Gutenberg was the first to create his type pieces from an alloy of lead, tin, and antimony – and these materials remained standard for 550 years .

Responses
GPT 3.5
The invention of the printing press is commonly credited to Johannes Gutenberg, who introduced the metal movable-type printing press in Europe around 1450. Gutenberg's innovations in casting the type based on a matrix and hand mould, along with the use of an alloy of lead, tin, and antimony for creating the type pieces, are considered significant advancements in printing technology.
Mistral
Johannes Gutenberg of Mainz, Germany is given credit for inventing the metal movable - type printing system in Europe and the printing press.
CLAPNQ-T5-LG
The world's first movable type printing press technology for printing paper books was invented around AD 1040 in China during the Northern Song Dynasty by the inventor Bi Sheng (990–1051). Around 1450 Johannes Gutenberg introduced the metal movable - type printing press in Europe.
Reference
The world 's first movable type printing press technology for printing paper books was made of porcelain materials and was invented around AD 1040 in China during the Northern Song Dynasty by the inventor Bi Sheng (990–1051) . Around 1450 Johannes Gutenberg, introduced the metal movable - type printing press in Europe.

Table 12: The GPT and Mistral answers look good but are only partially faithful to the passage, because they does not give any credit to the innovation in China that is also mentioned in the passage. The GPT answer is also too long.

Question: where are nimbus clouds found in the sky

Passage 1 (GOLD): Nimbostratus has a diffuse cloud base generally found anywhere from near surface in the low levels to about 3,000 m (9,800 ft) in the middle level of the troposphere . Although usually dark at its base, it often appears illuminated from within to a surface observer . Nimbostratus usually has a thickness of about 2000 to 4000 m . Though found worldwide, nimbostratus occurs more commonly in the middle latitudes . It is coded C 2 on the SYNOP report .

Passage 2: Nimbostratus occurs along a warm front or occluded front where the slowly rising warm air mass creates nimbostratus along with shallower stratus clouds producing less rain, these clouds being preceded by higher - level clouds such as cirrostratus and altostratus . Often, when an altostratus cloud thickens and descends into lower altitudes, it will become nimbostratus .

Passage 3: Multi-level nimbostratus is physically related to other stratiform genus - types by way of being non-convective in nature . However, the other sheet - like clouds usually each occupy only one or two levels at the same time . Stratus clouds are low - level and form from near ground level to 2,000 metres (6,600 ft) at all latitudes . In the middle level are the altostratus clouds that form from 2,000 metres (6,600 ft) to 7,000 metres (23,000 ft) in polar areas, 7,000 metres (23,000 ft) in temperate areas, and 7,600 metres (24,900 ft) in tropical areas . . .

Responses
GPT 3.5
Gold Nimbostratus clouds are generally found anywhere from near the surface in the low levels to about 3,000 m (9,800 ft) in the middle level of the troposphere. They can be found worldwide, but they occur more commonly in the middle latitudes.
RAG Nimbostratus clouds are found in the middle level of the troposphere, typically anywhere from near the surface to about 3,000 meters (9,800 feet) in altitude.
Mistral
Gold Nimbostratus clouds are found in the middle levels of the troposphere, at altitudes ranging from near the surface up to about 3,000 meters (9,800 ft).
RAG Nimbostratus clouds are found in the middle levels of the troposphere, from near surface to about 3,000 m (9,800 ft) in altitude. They are often preceded by higher-level clouds such as cirrostratus and altostratus and are characterized by their thick, dark, and uniform appearance. Nimbostratus clouds are usually associated with frontal systems and are responsible for producing continuous and heavy precipitation.
CLAPNQ-T5-LG
Gold Nimbostratus has a diffuse cloud base generally found anywhere from near surface in the low levels to about 3,000 m (9,800 ft) in the middle level of the troposphere. Though found worldwide, nimbostratus occurs more commonly in the middle latitudes.
RAG Nimbostratus has a diffuse cloud base generally found anywhere from near surface in the low levels to about 3,000 m (9,800 ft) in the middle level of the troposphere. Stratus clouds are low - level and form from near ground level to 2,000 metres (6,600 ft) at all latitudes. In the middle level are the altostratus clouds that form from 2,000 metres (6,600 ft) to 7,000 metres (23,000 ft) in polar areas, 7,000 metres (23,000 ft) in temperate areas, and 7,600 metres (24,900 ft) in tropical areas.
Reference
Nimbostratus has a diffuse cloud base generally found anywhere from near surface in the low levels to about 3,000 m in the middle level of the troposphere. Though found worldwide, nimbostratus occurs more commonly in the middle latitudes.

Table 13: In this example we compare the responses when just the GOLD answer is provided and the top 3 passages are provided when the first passage is the GOLD passage. All models provide good responses when only the gold answer is provided. In the non-RAG setting CLAPNQ-T5-LG and Mistral provided irrelevant information from the other passages about other clouds. All three missed the sentence that they can be found world-wide in the RAG setup.