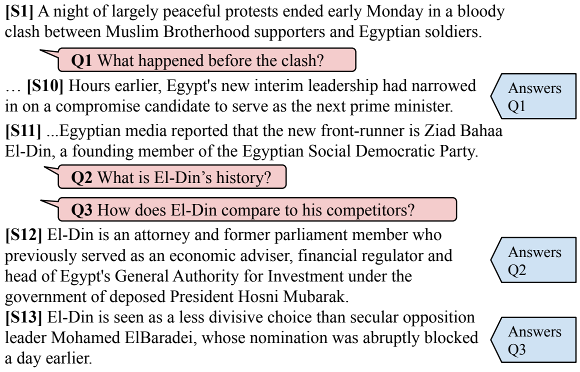
Discourse Comprehension: A Question Answering Framework to Represent Sentence Connections

Wei-Jen Ko[1] **Cutter Dalton**[2] **Mark Simmons**[3] **Eliza Fisher**[4]
Greg Durrett[1] **Junyi Jessy Li**[4]
[1] Computer Science, [4] Linguistics, The University of Texas at Austin
[2] Linguistics, University of Colorado Boulder
[3] Linguistics, University of California San Diego
wjko@utexas.edu, cutter.dalton@colorado.edu, mjsimmons@ucsd.edu, eliza.fisher@utexas.edu,
gdurrett@cs.utexas.edu, jessy@utexas.edu

Abstract

While there has been substantial progress in text comprehension through simple factoid question answering, more holistic comprehension of a discourse still presents a major challenge (Dunietz et al., 2020). Someone critically reflecting on a text as they read it will pose curiosity-driven, often open-ended questions, which reflect deep understanding of the content and require complex reasoning to answer (Ko et al., 2020; Westera et al., 2020). A key challenge in building and evaluating models for this type of *discourse comprehension* is the lack of annotated data, especially since collecting answers to such questions requires high cognitive load for annotators. This paper presents a novel paradigm that enables scalable data collection targeting the comprehension of news documents, viewing these questions through the lens of discourse. The resulting corpus, DCQA (**D**iscourse **C**omprehension by **Q**uestion **A**nswering), captures both discourse and semantic links between sentences in the form of free-form, open-ended questions. On an evaluation set that we annotated on questions from Ko et al. (2020), we show that DCQA provides valuable supervision for answering open-ended questions. We additionally design pre-training methods utilizing existing question-answering resources, and use synthetic data to accommodate unanswerable questions.



[S1] A night of largely peaceful protests ended early Monday in a bloody clash between Muslim Brotherhood supporters and Egyptian soldiers.

Q1 What happened before the clash?

… **[S10]** Hours earlier, Egypt's new interim leadership had narrowed in on a compromise candidate to serve as the next prime minister.

[S11] ...Egyptian media reported that the new front-runner is Ziad Bahaa El-Din, a founding member of the Egyptian Social Democratic Party.

Q2 What is El-Din's history?

Q3 How does El-Din compare to his competitors?

[S12] El-Din is an attorney and former parliament member who previously served as an economic adviser, financial regulator and head of Egypt's General Authority for Investment under the government of deposed President Hosni Mubarak.

[S13] El-Din is seen as a less divisive choice than secular opposition leader Mohamed ElBaradei, whose nomination was abruptly blocked a day earlier.

Figure 1: Discourse comprehension involves making high level inferences across sentences, often in the form of open-ended questions. This is an example from our dataset DCQA.

1 Introduction

Research in question answering has pushed machine comprehension to new heights, especially in answering factoid questions (e.g., SQuAD (Rajpurkar et al., 2018) and Natural Questions (Kwiatkowski et al., 2019)) and even questions that require multiple steps of reasoning (Yang et al., 2018). Yet while existing systems are effective at sifting through a long document to find a fact, they fail to achieve what we consider *discourse comprehension*. Deep comprehension of a discourse requires establishing temporal and semantic relationships across abstract concepts in various parts of a document (Hobbs, 1985), going beyond understanding the individual pieces of knowledge conveyed or the details of events (Wegner et al., 1984). The gap between existing QA frameworks and discourse comprehension has been highlighted in recent work. Dunietz et al. (2020) emphasized the lack of narrative understanding capability in existing machine reading comprehension systems and benchmark datasets. Recent research (Ko et al., 2020; Westera et al., 2020) demonstrated that human reading comprehension is marked by the spontaneous generation of *open-ended* questions *anchored* in one part of an article; some of these questions are later answered in the article itself, forming a connection in the discourse (Figure 1). Ko et al. (2020) showed that compared to existing QA datasets, these questions are products of higher-level (semantic and discourse) processing (e.g.,"how" and "why" questions) whose answers are typically complex linguistic units like whole sentences. Such reader-generated questions are far out of reach from the capabilities of systems trained

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 11752–11764
December 7-11, 2022 ©2022 Association for Computational Linguistics

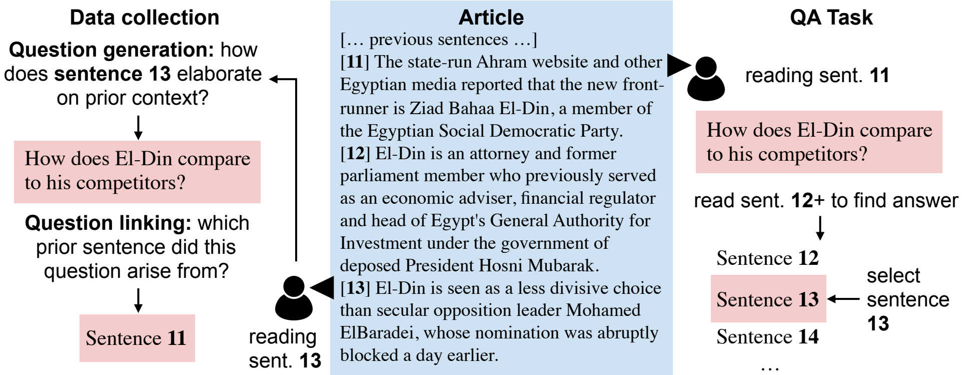

Data collection
Question generation: how does sentence 13 elaborate on prior context?
How does El-Din compare to his competitors?
Question linking: which prior sentence did this question arise from?
Sentence 11
reading sent. 13
Article
[… previous sentences …]
[11] The state-run Ahram website and other Egyptian media reported that the new front-runner is Ziad Bahaa El-Din, a member of the Egyptian Social Democratic Party.
[12] El-Din is an attorney and former parliament member who previously served as an economic adviser, financial regulator and head of Egypt's General Authority for Investment under the government of deposed President Hosni Mubarak.
[13] El-Din is seen as a less divisive choice than secular opposition leader Mohamed ElBaradei, whose nomination was abruptly blocked a day earlier.
QA Task
reading sent. 11
How does El-Din compare to his competitors?
read sent. 12+ to find answer
Sentence 12
Sentence 13
Sentence 14
…
select sentence 13

Figure 2: DCQA framework. During data collection, an annotator writes down what question a particular sentence answers and which previous sentence it arose from (i.e., its *anchor point*). During QA, an annotator or an automatic system sees that question attached to the prior sentence, and has to find the answer in the remainder of the document.

on current QA datasets.

This view of discourse comprehension falls within the linguistic framework of Questions Under Discussion (QUD) (Velleman and Beaver, 2016; De Kuthy et al., 2018), where discourse progresses by continuously evoking implicit or explicit questions and answering them. The open-endedness of these questions is triggered by readers' psychological mechanisms including corrections of knowledge deficits and active monitoring of the common ground (Graesser et al., 1992). These fundamental differences in how questions are constructed from existing, mostly factoid QA datasets lead to difficulty spotting answers with keyword or paraphrastic overlaps. Additionally, the questions are contextualized, as they rely on an established common ground and the anchor point.

Collecting question-answer pairs that target discourse comprehension entails a high cognitive load for human annotators. Consequently, existing resources designed to train models to answer such questions are scarce. Westera et al. (2020) introduced TED-Q, a smaller dataset covering 6 TED talks, including 1102 answered questions. Ko et al. (2020)'s INQUISITIVE dataset consists of more (19k) questions evoked under the QUD paradigm, yet they did not annotate answers.

We present DCQA (**D**iscourse **C**omprehension by **Q**uestion **A**nswering), a dataset of 22,394 English question-answer pairs distributed across 606 English news articles (Figure 2). We view DCQA as a key resource to train QA systems to answer discourse-driven, contextual, and open-ended questions as those in INQUISITIVE. On its own, DCQA is a discourse framework that represents connections (or relationships) between sentences via free-form questions and their answers.

DCQA uses a novel crowdsourcing paradigm inspired by QUD recovery (De Kuthy et al., 2018; Riester, 2019): instead of collecting the answer labels given the question, we start from an answer sentence and collect the question. Specifically, for each sentence in an article, annotators ask a question that reflects how the main purpose of the sentence connects to an anchor sentence in prior context, such that the sentence is the answer to the question. This paradigm is scalable while maintaining a diverse range of open-ended question types, a varied distribution of distance between question anchor and answers, and the ability to capture interesting linguistic phenomena. We further demonstrate that annotating answers given the questions is much more subjective and challenging.

We present two experiments using DCQA. We first evaluate a model's ability to extract the correct sentence to answer questions in DCQA. Our Longformer (Beltagy et al., 2020) baseline achieves 67% accuracy, doing well in light of human performance of 72.2%. However, the model struggles much more than humans do when the answer sentence is further away from the question anchor point.

Our main experiment evaluates whether DCQA can be used to train systems to answer questions such as those in INQUISITIVE, where the questions are asked without seeing any upcoming context that may contain answers. To enable evaluation over INQUISITIVE, we collect a human-annotated test set. We find that existing datasets, including SQuAD, QuAC, ELI5, and TED-Q, yield poor accuracy (1.4% to 25%) on extracting answer sentences for the answerable questions in INQUISITIVE. In contrast, training on DCQA leads to a performance of 40.8%. We further design ways to enable pre-training using SQuAD and ELI5, which leads to significant performance improvements. Finally, we present a pipeline system for question-answering that handles unanswerable questions utilizing synthetic data. [1]

2 Background and related work

Discourse and question answering. Discourse theories profile the relationships between linguistic units in a document that make it coherent and meaningful; these relations could be temporal, causal, elaborative, etc. (Mann and Thompson, 1988; Lascarides and Asher, 2008; Prasad et al., 2008). Most of these theories define fixed relation taxonomies, which have been viewed as templatic question (Prasad and Joshi, 2008; Pyatkin et al., 2020). In contrast, QUD makes use of free-form questions to represent discursive relationships (Riester, 2019); each utterance in a text builds on the reader's common ground and can evoke questions; each utterance is also an answer to an implicit (or explicit, if present) question.

INQUISITIVE (Ko et al., 2020) is a collection of such curiosity-driven, QUD-style questions generated by readers as they read through a document (i.e., asking questions on-the-fly without seeing any upcoming context); specifically, 19K questions for the first 5 sentences across 1500 news documents from three sources: Newsela (Xu et al., 2015), WSJ articles from the Penn Treebank (Marcus et al., 1993), and Associated Press (Harman and Liberman, 1993). Each question is attached to a span in the current sentence the reader is reading, e.g.,

> ... It's an obvious source of excitement for Doyle, who is 27 years old and has <u>severe autism.</u>
> **Q:** Why is this important?

One major motivation of DCQA is to train models to answer these questions. Thus *as opposed to* creating a "challenge dataset", we are mainly exploring more scalable ways for data collection.

TED-Q (Westera et al., 2020) is also a QUD dataset annotating questions and answers simultaneously over 6 TED talk transcripts. In contrast, we propose a data collection paradigm that works from the answers back to the context for lower cognitive load and higher scalability. Also, all answers from TED-Q are within 5 sentences after the question anchor, which is not the case in our dataset.

Open-ended question answering. While there are many datasets for question answering, most of them are factoid in nature (Rajpurkar et al., 2018; Kwiatkowski et al., 2019). Ko et al. (2020) found that the question types are so different that pre-training on factoid questions *hurts* the performance of *generating* open-ended questions. Fan et al. (2019) collected human-written answers for their open-ended "ELI5" questions, but the questions and answers are not grounded in a document context. QuAC (Choi et al., 2018) introduces information-seeking question-answer pairs given context in dialogue; in contrast DCQA aims to represent discursive and semantic links that cover the full document. Also QuAC questions are conditioned on previous dialogue turns, which is not the case in DCQA. Soleimani et al. (2021) collected paragraphs with a subheading that is a question from news articles, so the whole paragraph can be seen as the answer to the question. However, more than 86% of the time in INQUISITIVE (Section 6.1), and 99% for TED-Q, the answers to naturally generated open-ended questions consist of one sentence in the text. By more precisely locating the answers, DCQA also establishes finer-grained connections within a document.

3 The DCQA dataset

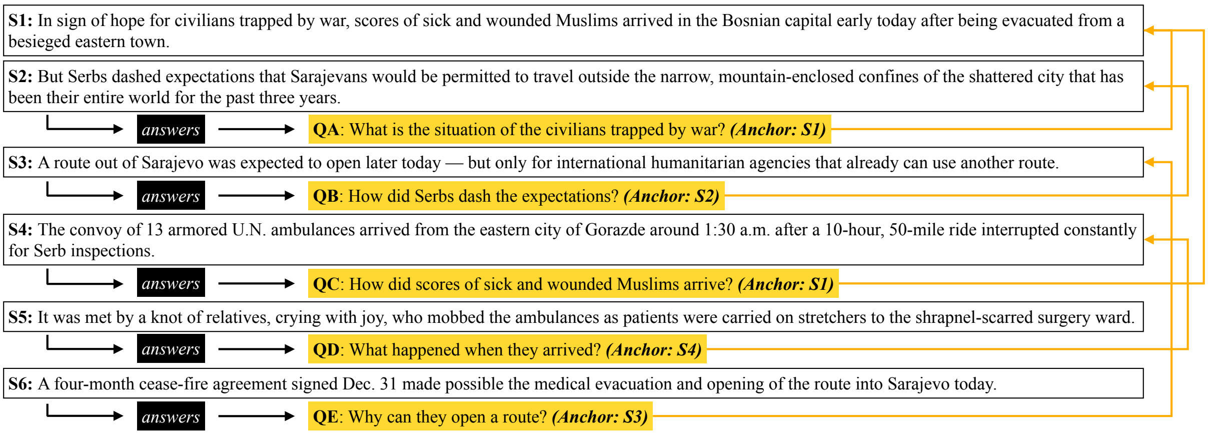
We present DCQA using a new data collection paradigm that describes how sentences in an article relate to prior context; specifically, what *question* is the current sentence answering given its prior context, inspired by literature in QUD discovery (De Kuthy et al., 2018; Riester, 2019).

We annotate news documents due to their rich linguistic structure that mixes narratives, different lines of events, and perspectives from parties involved (Van Dijk, 2013; Choubey et al., 2020).

3.1 Annotation paradigm

We design a crowdsourcing task, illustrated in Figure 2. At a high level, data collection consists of two *simultaneous* tasks: *question generation* given an answer sentence, and *question linking* that places the question in a precise anchor point in the document. Annotators start from the beginning of the article. For each new sentence, they write a question that reflects the main purpose of the new sentence and indicates how the new sentence elaborates on prior context. The new sentence is thus the answer to the question. Because the question should arise from one of the previous sentences in the article, the annotator also chooses which earlier

[1] We release DCQA at https://github.com/wjko2/DCQA-Discourse-Comprehension-by-Question-Answering.



S1: In sign of hope for civilians trapped by war, scores of sick and wounded Muslims arrived in the Bosnian capital early today after being evacuated from a besieged eastern town.

S2: But Serbs dashed expectations that Sarajevans would be permitted to travel outside the narrow, mountain-enclosed confines of the shattered city that has been their entire world for the past three years.

answers **QA**: What is the situation of the civilians trapped by war? ***(Anchor: S1)***

S3: A route out of Sarajevo was expected to open later today — but only for international humanitarian agencies that already can use another route.

answers **QB**: How did Serbs dash the expectations? ***(Anchor: S2)***

S4: The convoy of 13 armored U.N. ambulances arrived from the eastern city of Gorazde around 1:30 a.m. after a 10-hour, 50-mile ride interrupted constantly for Serb inspections.

answers **QC**: How did scores of sick and wounded Muslims arrive? ***(Anchor: S1)***

S5: It was met by a knot of relatives, crying with joy, who mobbed the ambulances as patients were carried on stretchers to the shrapnel-scarred surgery ward.

answers **QD**: What happened when they arrived? ***(Anchor: S4)***

S6: A four-month cease-fire agreement signed Dec. 31 made possible the medical evacuation and opening of the route into Sarajevo today.

answers **QE**: Why can they open a route? ***(Anchor: S3)***

Figure 3: Snippet of an annotated article: each sentence after the first is annotated with a question linked to an earlier sentence where it arises from (which we call the question's *anchor sentence*). Black arrows represent question generation; orange arrows represent question linking.

sentence most naturally evokes the question (which we refer to as the question's *"anchor"* sentence).

We instructed annotators to ask open-ended questions that require some explanation as an answer, instead of questions that could be answered simply by a place/person/time or yes/no. They are also told that when writing the question, they should assume that the question could be asked and understood by people only reading the earlier sentences, following the QUD framework and Ko et al. (2020); Westera et al. (2020). This means avoiding reference to any information first introduced in the new sentence and avoiding copying phrases only used in the new sentence. Reusing phrases from or referencing previous sentences is allowed.

An example of the first 6 sentences of a fully-annotated article is shown in Figure 3, where we can see each sentence after the first is annotated with a question. For instance, sentence 4 mainly describes the trip of U.N. ambulances arriving, and it provides more detail on how the sick and wounded Muslims described in sentence 1 arrive.

We illustrate the **Question Answering task** in the right pane of Figure 2. Given the anchor sentence and the question, the task is to extract a sentence in the upcoming context that would answer this question. Because the questions are contextual, they typically need to be associated with the anchor to be comprehensible (we illustrate the importance of identifying the anchor through model analysis in Section 5). Additionally, since we seek to answer questions that are generated during the natural progress of discourse, the questions presuppose the common ground established by the reader, which consists of all context up to the anchor sentence. (Because of this question generation paradigm, in DCQA, as well as INQUISITIVE and TED-Q, all answers to the questions are after the anchor sentence.) Thus the QA task is that given the context, anchor, and question, find the sentence that contains the answer. For example, S1 in Figure 3 evokes QA and QC, and the task is to find the answer sentences 2 and 4, respectively:

> **[Context+Anchor]**: In sign of hope for civilians trapped by war, scores of sick and wounded Muslims arrived in the Bosnian capital early today after being evacuated from a besieged eastern town.
> **[QA]**: *What is the situation of the civilians trapped by war?*
> **[Answer to QA]**: sentence 2.
> **[QC]**: *How did scores of sick and wounded Muslims arrive?*
> **[Answer to QC]**: sentence 4.

3.2 Annotators and corpus

The data we aim to collect consists of free-form, mostly open-ended questions. The nature of the data is inherently subjective, and the free text annotation poses challenges for automatic evaluation of quality. Therefore, we take measures to both recruit good annotators and ensure annotation quality.

Our annotation team consists of three expert annotators, as well as a small number of *qualified and trained* crowd workers. We first piloted the task among the expert annotators, who were undergraduate linguistics researchers familiar with linguistics literature on discourse coherence. The crowdsourcing task was then collectively developed to be accessible to crowdworkers and to provide a list of constraints for the questions; we show in Appendix C.2 the key instructions.

We use Amazon Mechanical Turk as our crowd-

Similarity	1	2	3	4	5
Percentage	26.1	14.6	17.6	16.9	24.9
Same anchor	35.8	23.7	34.8	43.1	69.8

Table 1: Human rating of question similarity for questions with the same answer sentence (top row), and fraction of questions with the same anchor sentence in each bucket (bottom row).

Distance	1	2-4	≥ 5
DCQA	48.8	24.2	27.0
TED-Q	58.2	41.8	0.0
INQUISITIVE	18.1	27.8	54.1

Table 2: The distribution of the distance (in terms of sentence ID differences) between the answer and anchor sentences.

sourcing platform. To ensure quality, we designed a qualification task consisting of one article. We manually inspected the quality of the responses of each candidate, and only workers who asked satisfactory questions were invited to continue with the task. Throughout the process, we also monitored the responses of the workers and reminded them about specific guidelines when necessary. The workers are paid around $10/hr. No demographic information was collected.

Corpus We annotate a subset of the articles used in the INQUISITIVE dataset (Ko et al., 2020) (c.f. Section 2). In contrast to INQUISITIVE which only contains curiosity-driven questions about the first five sentences for each article, we annotate up to the first 20 sentences. DCQA consists of 22,394 questions distributed across 606 articles, 51 of which are annotated by experts. Each article is annotated by 2 annotators.

4 Analyzing DCQA

This section presents a series of analytical questions we pose in order to understand DCQA questions, answers, and potential challenges.

Are questions similar given the same answer sentence? We first quantify the level of subjectivity in our question-answering paradigm via two angles: (1) Given the same answer sentence, how similar are the questions asked by different annotators? (2) When questions are similar, do annotators agree on which sentence is the anchor?

Since the questions are free-text in form, we use human judgments to assess their similarity. We asked our expert annotators to rate the similarity of a sample of 261 pairs of questions asked with the same answer sentence, on a scale of 1-5. Table 1 (row "percentage") shows that the questions different people ask may differ: while 41.8% of the questions are highly similar with ratings of 4-5, 40.7% of the questions are semantically different with ratings 1-2.

Table 1 (row "same anchor") shows how often annotators agree on the anchor sentence. When people ask the same questions (similarity 5), the percentage of agreement on the anchor sentence is high. Similar questions can have different anchors but answered in the same sentence; Appendix A.1 shows an example. Distinct questions can also share the same anchor sentence, as shown in Figure 1, questions 2 and 3.

How well can humans do the QA task? Our annotation paradigm produces open-ended questions given the answer sentence, while ultimately the question-answering task seeks to find the answer given the question. Thus we explore to what extent humans can, given a question, find the original answer sentence where the question was generated. We asked an expert annotator to answer questions in a randomly sampled subset of 1175 questions from the validation and test sets, without informing them of the "gold" answer labels. If the annotator thinks there are multiple answers, he could annotate multiple sentences. After the answers are annotated, the gold answer is shown and the annotator adjudicates the two versions of answers.

The agreement between the annotator and the gold labels is 72.2%. In cases where the annotator selected multiple answer sentences, we report the expected agreement when randomly choosing one of them. For 69.5% of the questions, the annotator exactly chose the gold answer; for 5.3%, the annotator selected multiple answers including the gold. On 2.3% of the data, the annotator thinks the gold answer cannot be used to answer the question without a stretch, thus should be regarded as noise. For the rest, the annotator thinks that both answers are reasonable as there are multiple possible answers, although they did not originally choose the gold. In these cases, the annotator tended to choose an earlier sentence than the gold. An example for multiple answer sentences is in Appendix A.1.

How far away are answer sentences from question anchors? Table 2 shows the distances between the question anchor and the answer sentence, by difference of sentence IDs. While for a large proportion of questions, the answer sentence could be found just a few sentences after the question anchor sentence, there are also some questions with answers that are far away later in the article.

Compared to TED-Q, DCQA contains fewer questions that are answered by the immediate next sentence. DCQA also has answers ≥5 sentences away from the question, which TED-Q does not have. This means that fewer questions in DCQA can be trivially found in the immediate next sentence. Additionally, this also shows that the sentence connections captured by DCQA naturally result in a different structure from QADiscourse (Pyatkin et al., 2020) that expresses discourse relations across adjacent sentences in templatic questions.

We also show the figures for INQUISITIVE over a subset that we annotate for answers (Section 6.1). Most of the INQUISITIVE answers are far away from where the question was asked; this is in stark contrast with TED-Q, highlighting the distributional differences between TED talks and news.

Linguistic characteristics of DCQA. An interesting effect of the question generation process is that annotators often find semantic links between two parts of the article that are not pragmatically or discursively connected. For example:

> **[Context]**: This small Dallas suburb's got trouble. Trouble with a capital T and that rhymes with P and that stands for pool. More than 30 years ago, Prof. Harold Hill, the con man in Meredith Willson's 'The Music Man,' warned the citizens of River City, Iowa, against the game.
> **[Sentence 6]**: Now kindred spirits on Addison's town council have barred the town's fanciest hotel, the Grand Kempinski, from installing three free pool tables in its new lounge.
> **[Question]**: Just how fancy is the Grand Kempinski?
> **[Sentence 13]**: At the lounge [of the Grand Kempinski], manager Elizabeth Dyer won't admit patrons in jeans, T-shirts or tennis shoes.

This question arises from the adjective *fanciest* used in Sentence 6. The discursive intent of Sentence 6 is to describe the events that make the game of pool relevant to the Dallas suburb mentioned; the assertion that the Grand Kempinski hotel is the "town's fanciest" is not discursively salient at this point, though it is relevant later in the article, as indicated by Sentence 13. This shows that a conscious reader was able to connect two pragmatically unrelated sentences by their semantic content. Such links have been deemed important in discourse literature, e.g., the Discourse Graphbank (Wolf and Gibson, 2005) and the Entity Relation between two adjacent sentences in the Penn Discourse Treebank (Prasad et al., 2008).

	Train	Val.	Test
# questions	20942	718	734
# docs	555	22	30

Table 3: Train/test/validation splits of DCQA.

Finally, we present an analysis about the type of the questions asked in Appendix A.2, using the schema in Cao and Wang (2021). DCQA has a good coverage of the key question types in existing high-level question answering datasets (TED-Q and INQUISITIVE); the most frequent question types are concept, cause, procedural, and example.

5 Question answering on DCQA

As each of the questions already has a designated answer sentence during data collection, DCQA can be used to learn models for this type of QA, or as a testbed for existing models.

Model Because our task involves reasoning over long documents, we choose Longformer (Beltagy et al., 2020) as our model. Longformer is a transformer model with an attention pattern that combines a local windowed attention and global attention on pre-selected, task-dependent tokens only. This makes the computation time scale linearly with the sequence length instead of quadratic (as in conventional transformers), allowing faster computation on longer sequences, as well as better performance on QA tasks with long contexts.

Our model operates similarly to the BERT QA model for SQuAD (Devlin et al., 2019). For the model input, we concatenate the question, the anchor sentence, and the article, separated by delimiters. Note that by passing in the full article, the model has access to all context prior to the anchor. Since the goal of our model is to predict a sentence instead of the answer span, we add two tokens in front of each sentence in the article, the start of sentence token `[sos]` and the sentence ID. The model is trained to predict the span of the two added tokens in front of the answer sentence.

Settings We use our expert annotated question-answer pairs for the validation and test sets, and crowd annotation for the training set. Table 3 shows the distribution of articles and questions across training, testing and validation sets.

Type	Model	Human
1	85.4	81.8
2-4	46.3	69.3
≥5	51.0	57.1

Table 4: Model and human accuracy for answer sentence extraction, statified by distance between question anchor and answer.

Because we will later experiment on INQUISITIVE (Section 6), we exclude articles overlapping with the INQUISITIVE test set (10 articles containing 376 questions) during training.

We use HuggingFace (Wolf et al., 2020) for our implementation. We use the Adam (Kingma and Ba, 2015) optimizer with $(\beta_1, \beta_2) = (0.9, 0.999)$ and a learning rate of 5e-5.

Results The fine-tuned Longformer model achieves 67.2% accuracy. Note that a baseline that predicts the immediate next sentence after the anchor would achieve an accuracy of 40%. We also experimented training with TED-Q, which achieved an inferior result of 39.5%.

We observe the overall performance is about 5% lower than a human's (72.2%, Section 4). However, the model is doing substantially worse on answers that do not immediately follow the anchor (Table 4; human performance is reported on the 1,175 annotated examples in Section 4). We observe that human performance for answers further away from the anchor is also lower, reflecting that establishing longer-distance discursive connections is also more subjective for humans. Table 9 in Appendix B.1 stratifies model performance per question type.

How important are the context and anchor? To evaluate the influence of the context and anchor sentences, we experiment on 3 ablation settings. The results are shown in Table 5.

(1) ***No prior context:*** We remove all the sentences prior to the anchor sentence from model inputs (while keeping the anchor itself). This setting resulted in an accuracy of 66.4%, which is slightly worse yet on par with the context present, showing that the model does not use much information from the sentences before the anchor.

(2) ***Unspecified anchor:*** For the Longformer model, we do not concatenate the anchor sentence with the question, while keeping the full article intact (which includes the prior context and the anchor). We observed an accuracy of 43.5%, which is a substantial 23.7% drop from the version with the anchor sentence present. This shows that the model is struggling to locate the anchor if it is not given as supervision.

	accuracy
Original model	67.2
No prior context	66.4
Unspecified anchor	43.5
No context and no anchor	42.6

Table 5: Results for ablating context and source sentence

(3) ***No prior context and no anchor:*** In this setting, in addition to applying both modification from above: no prior context in the input article and the anchor sentence is not concatenated with the question for the model, the anchor is also removed from the input article. This yields an accuracy of 42.6%, slightly worse but on par with (2).

We can see from (2) and (3) that specifying which sentence the question arises from is very important for the model performance, while the presupposed common ground does not play such a significant role.

As a small experiment to assess whether humans can find the anchor points, we asked one of our expert annotators to try answering a set of 90 questions under those ablation settings. We found that for (1) and (2), the human performed almost the same as providing all the information, while (3) is 35% worse. This shows that human are better at finding the anchor from the article while it is not specified, provided that the full article is present. Additionally, the questions are contextual, and requires the anchor to comprehend.

6 Question answering on external datasets

We first describe our collection of the Inquisitive test set (Section 6.1). Then we explore only answerable questions (Section 6.2), and design pretraining methods to allow the use of other QA datasets. Finally, we present a pipeline system that also handles question answerability (Section 6.3), by predicting whether a question is answerable using synthesized data from DCQA.

6.1 INQUISITIVE answer collection

To construct the evaluation set, for a subset of INQUISITIVE questions in their test set, two of our expert annotators independently chose the sentence that contains the answer. Then they met to adjudicate a final version of the answers.

This newly annotated data consists of answers to 719 INQUISITIVE questions from 60 articles. Before adjudication, the annotators agree on whether there is an answer 78.8% of the time. Out of those questions at least one annotator had answers, 42.7% have total agreement, and another 7.3% have partial agreement. This shows that answering the INQUISITIVE questions is challenging, even for trained linguistics senior students.

After adjudication, 424 of the questions have answers. 57 of the questions have multiple valid answers (for 30 cases, related information spreads across multiple sentences; for 27 cases, the questions can be answered from multiple angles).

During testing, we judge the model as correct if it predicts any one of the answers.

6.2 Answering answerable questions

We use Longformer as discussed in Section 5 as our base model. Additionally, we explore the utility of other existing QA datasets via pre-training: (1) we synthetically augment SQuAD 2.0 (Rajpurkar et al., 2018) for better compatibility of data format between DCQA and INQUISITIVE; (2) we utilize ELI5 (Fan et al., 2019) to provide more supervision for answering open-ended questions. We also explore using TED-Q for training. We also experiment on training with QuAC (Choi et al., 2018) for comparison, since a portion QuAC questions are open-ended questions, though its domain and the dependency of the questions on previous dialog turns make the format different from ours. We do not use previous dialog turns in our experiments.

SQuAD pretraining While DCQA and INQUISITIVE contain mainly high-level questions, their format is different: INQUISITIVE questions are asked about a sub-sentential span in the document, and the questions are more dependent on the span.

To pre-train the model to bridge this formatting gap, we create a synthetic dataset from SQuAD that simulates the input format with an annotated question span utilizing coreference substitution. Preprocessing details are descibed in Appendix C.1.

ELI5 pretraining We also sought existing data for additional supervision to answer open-ended questions. The most related dataset we found is ELI5 (Fan et al., 2019) for long-form question answering. While the questions in the dataset comes with a human-written answer, they do not have context to form a question answering text. Preprocessing details are described in Appendix C.2.

	DCQA	INQUISITIVE
SQuAD	34.7	9.9
ELI5	1.3	1.4
QuAC	35.9	17.2
TED-Q	39.5	25.0
+SQuAD+ELI5	46.0	38.2
DCQA	**67.2**	40.8
+SQuAD	66.7	41.3
+ELI5	64.0	42.9
+SQuAD+ELI5	66.0	**45.5**
+TED-Q	64.0	40.6

Table 6: Question answering results. Systems trained on DCQA (includes pre-training) can get the strongest performance on INQUISITIVE.

Settings When using either SQuAD or ELI5 pretraining, we pretrain for 6 epochs. When using both types of pretraining, we directly mix the data together, and pretrain for 3 epochs.[2] The paremeters are tuned on discourse questions validation set. All other settings are the same as in Section 5.

Results Table 6 shows for both DCQA and all INQUISITIVE questions with an answer, the accuracy values of answer sentence extraction. Training only on TED-Q, synthesized SQuAD, QuAC, and ELI5 examples produces poor results, showing that existing datasets cannot be directly used to train model to answer the open-ended questions meant to facilitate discourse comprehension.

When tested on INQUISITIVE, training on DCQA achieves the best performance across all settings, showing its ability to provide supervision for open-ended question answering. Although pretraining did not help for answering DCQA questions, for INQUISITIVE, both types of pretraining are helpful individually, and using them together yields the best result. We conducted a binomial significance test between DCQA and S+E+DCQA, and found that the improvement is statistically significant. Additionally, pre-training also improved performance when used with TED-Q (rows 3 vs. 4). This shows that our design to adapt the two datasets is successful. TED-Q used on top of the other two pre-training did not improve performance.

6.3 Handling question answerability

We experiment on the full open-ended question answering task: given context and question, answer

[2] Also, using all SQuAD questions performed better when both types of pre-training are performed, while only using the answered questions performed the best when pre-training with SQuAD-only.

System	DCQA	S+E+DCQA	S+E+TED
F1	0.260	0.358	0.242

Table 7: Pipeline system question answering results

the question only if there is an answer to be found in the article.

While DCQA does not come with unanswerable questions, we generate them by truncating parts of the articles that contain answers.[3] Specifically, we truncate each article in the training set of DCQA to the first 12 sentences, and label questions with answers after sentence 12 as unanswered. Using this data, we train a model to predict whether a question is answerable given the text, following the setup of models in Sections 5 and 6.2. We then combine this model with models in Section 6.2 into a pipeline: first predict whether an answer exists, then provide the answer. For the full task, we split the collected INQUISITIVE answers into a validation (223 examples) and test (496 examples) sets. We adjust the threshold of predicting on the validation set of discourse questions.

Results show that the model could predict if there is an answer correctly 84% of the time on the synthetic discourse questions test set, but only 67% of the time when tested on INQUISITIVE. In both datasets, about 59% of the questions are answered. This shows that although our way of generating synthetic data is useful, answerability prediction is challenging.

We report the result of the pipeline system in Table 7, applying first our answerability model, then the corresponding question answering model. We use F1 score of correctly predicted questions following Rajpurkar et al. (2018). If the model predicts no answer, we treat it as not making a prediction when calculating the F1 score.

While DCQA and pre-training is useful, the numbers clearly shows that a full open-ended question answering system, that includes answerability prediction, is an extremely challenging task. Answerability prediction presents a bottleneck here, and we leave for future work to design better strata-gies and models.

7 Conclusion

We present DCQA that connects pieces in a document via open-ended questions and full-sentence answers. DCQA is collected via a new paradigm that regards the main purpose of a new sentence as an answer to a free-form question evoked earlier in the context. Consequently, this paradigm yields both discourse and semantic links across all sentences in a document. DCQA is introduced with the goal of providing a more scalable data collection paradigm, also as initial resource, for answering open-ended questions for discourse comprehension. Our experiments showed that DCQA provides valuable supervision for such tasks.

Limitations

DCQA collects questions in a reactive manner: the answer is first observed before the question is generated. This is, by design, different from methods where questions are elicited as a person reads (i.e., without seeing upcoming context, as in INQUISITIVE and TED-Q). Seeing the answer before asking the question inevitably results in a slight distributional shift from datasets such as INQUISITIVE, as seen in Table 8 (Appendix A.2). Qualitatively, we observe that the questions tend to be a bit more specific than INQUISITIVE, and answers are more easily associated with a particular sentence.

Another notable difference is that DCQA does not address unanswerable questions. While we designed synthetic data augmentation methods to train models to handle such questions, this is challenging, as discussed in Section 6.3. We hope future work could find better solutions.

Multi-sentence answers exist much more frequently in high-level question answering than factoid QA; in DCQA, it happens if questions elicited from different answer sentences share an anchor sentence (Appendix A.1). We leave multi-sentence answers for future work.

Finally, although DCQA is designed as a general paradigm for data collection, the dataset presented in this paper is collected on English news articles. Thus the distribution of questions and answers may change by genre and/or language, which should be explored in future work.

Acknowledgments

We are grateful for the feedback provided by anonymous reviewers. This work was partially supported by NSF grants IIS-2145479, IIS-2145280, and IIS-1850153. We acknowledge the Texas Advanced Computing Center at UT Austin for many of the results within this paper.

[3] While we could in theory simply “take out” the answer sentence, it would leave the text incoherent and unnatural.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Shuyang Cao and Lu Wang. 2021. Controllable open-ended question generation with a new question type ontology. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6424–6439.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184.

Prafulla Kumar Choubey, Aaron Lee, Ruihong Huang, and Lu Wang. 2020. Discourse as a function of event: Profiling discourse structure in news articles around the main event. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5374–5386.

Kordula De Kuthy, Nils Reiter, and Arndt Riester. 2018. Qud-based annotation of discourse structure and information structure: Tool and evaluation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Jesse Dunietz, Greg Burnham, Akash Bharadwaj, Owen Rambow, Jennifer Chu-Carroll, and Dave Ferrucci. 2020. To test machine comprehension, start by defining comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7839–7859.

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.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6.

Arthur C Graesser, Natalie Person, and John Huber. 1992. Mechanisms that generate questions. *Questions and information systems*, 2:167–187.

Donna Harman and Mark Liberman. 1993. TIPSTER Complete LDC93T3A. *Linguistic Data Consortium*.

Jerry R Hobbs. 1985. On the coherence and structure of discourse.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations*.

Wei-Jen Ko, Te-yuan Chen, Yiyan Huang, Greg Durrett, and Junyi Jessy Li. 2020. Inquisitive question generation for high level text comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6544–6555.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Alex Lascarides and Nicholas Asher. 2008. Segmented discourse representation theory: Dynamic semantics with discourse structure. In *Computing meaning*, pages 87–124. Springer.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Mitch Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, 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.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The Penn Discourse TreeBank 2.0. In *Language Resources and Evaluation Conference*.

Rashmi Prasad and Aravind Joshi. 2008. A discourse-based approach to generating why-questions from texts. In *Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge*, pages 1–3. Citeseer.

Valentina Pyatkin, Ayal Klein, Reut Tsarfaty, and Ido Dagan. 2020. Qadiscourse-discourse relations as qa pairs: Representation, crowdsourcing and baselines. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2804–2819.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Arndt Riester. 2019. Constructing qud trees. In *Questions in discourse*, pages 164–193. Brill.

Amir Soleimani, Christof Monz, and Marcel Worring. 2021. Nlquad: A non-factoid long question answering data set. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1245–1255.

Teun A Van Dijk. 2013. *News as discourse*.

Leah Velleman and David Beaver. 2016. Question-based models of information structure. In *The Oxford handbook of information structure*.

Martha L Wegner, Robert H Brookshire, and LE Nicholas. 1984. Comprehension of main ideas and details in coherent and noncoherent discourse by aphasic and nonaphasic listeners. *Brain and language*, 21(1):37–51.

Matthijs Westera, Laia Mayol, and Hannah Rohde. 2020. TED-Q: Ted talks and the questions they evoke. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC'2020)*.

Florian Wolf and Edward Gibson. 2005. Representing discourse coherence: A corpus-based study. *Computational linguistics*, 31(2):249–287.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

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. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

A Additional analysis of DCQA

A.1 Additional examples

An example showing similar questions with different anchors but the same answer sentence:

> **[Sentence 11]** Now, those animals, once just visitors, have established resident populations – and they are spreading. ***[Q1]*** *How far have the Wolverines spread?*
> **[Sentence 12]** "We have growing evidence of them using larger and larger areas over time," Aubry said. ***[Q2]*** *How far have the Wolverines gone in their repopulation?*
> **[Sentence 13]** So far, scientists have confirmed resident wolverine populations from the North Cascades to as far south as this bait lure south of Highway 2 west of Leavenworth. **Answers both Q1 and Q2.**

In the example below, the expert annotator found multiple answer sentences for the question whose gold answer was Sentence 12.

> **[Sentence 1]**: Amid skepticism that Russia's war in Chechnya can be ended across a negotiating table, peace talks were set to resume Wednesday in neighboring Ingushetia. ***[Question]****: What has been the fallout of the war?*
> **[Sentence 12 (original and expert answer)]**: The Russian offensive has turned Grozny into a wasteland littered with rotting bodies, twisted metal and debris.
> **[Sentence 13 (expert answer)]**: Hardly a building is untouched.
> **[Sentence 14 (expert answer)]**: The war has also cost Russia dearly – in lives, prestige and rubles.

A.2 What questions are asked?

We further examine what types of questions are asked. We fine-tune a classifier based on pretrained BERT (Devlin et al., 2019) using the data and classification scheme from Cao and Wang (2021).[4] Table 8 shows the distribution of the types of question. We also show distributions of other open-ended question datasets for comparison.

In all datasets, concept questions are the most frequent; those questions ask about the definition or background knowledge. Compared to other datasets, ours contain many more causal questions (e.g., *why did Joyce Benes stop feeding the horses?*), reflecting that annotators frequently make causal inferences across events. In contrast we see fewer procedural and example questions. Our dataset also tend to contain few judgmental questions, i.e., question about opinions, which may be a reflection of news articles trying to stay objective.

B Additional model analysis

B.1 Model accuracy by question type

Table 9 shows the accuracy stratified by different types of questions as classified using the model from Cao and Wang (2021). The QA model performs well on extent and consequence questions, followed by concept, verification, and disjunct

[4] Human evaluation (with one of our expert annotators) of this system on a random set of 100 DCQA questions shows 64.9 F1.

	INQ.	Ours	TED-Q
verification	4.0	7.9	15.5
disjunctive	0.1	1.0	1.3
concept	31.3	32.5	23.3
extent	7.7	5.7	4.9
example	13.7	6.9	15.0
comparison	0.6	0.5	0.8
cause	14.1	31.8	13.8
consequence	4.2	0.6	1.5
procedural	14.3	10.8	14.7
judgmental	9.9	2.4	9.2

Table 8: Question types in each dataset, classified using the model from Cao and Wang (2021). Our dataset has good coverage of the key question types in the other two datasets.

questions, while performing relatively worse on comparison or cause questions.

Type	Accuracy
concept	66.8
verification	67.8
procedural	58.4
comparison	50.0
cause	55.7
judgmental	62.6
extent	71.3
example	64.3
disjunct	66.7
consequence	69.0

Table 9: Model performance statified by question type, trained and tested on DCQA.

C Pretraining details

C.1 SQuAD pretraining

At a high level, certain phrases in the SQuAD questions may be referred to by pronouns or demonstratives in their corresponding article, and we can replace such phrases in the questions by these more context-dependent forms, and sample an element in the chain that precedes the answer as the highlighted span. Specifically, we: **(1)** combine 5 consecutive paragraphs from SQuAD articles to create longer text and extract coreference chains;[5] **(2)** for each question whose answer span is in the combined text, we look for exact matches between ngrams in the question and expressions in coreference chains; **(3)** if there is a match, substitute the question phrase with a random reference in the matched chain, or the demonstrative "this"; When choosing the random reference we also consider whether the it is in possessive form. **(4)** designate a random chain element preceding the answer as the highlighted span.

For example, for the SQuAD question "*Of what group in the periodic table is oxygen a member?*", we found "oxygen" in the corresponding article referred to as "it", "the element", etc. Thus we change the question to *"Of what group in the periodic table is **it** a member?*", and generate the following highlight using one of the sentences before the answer: *"At standard temperature and pressure, two atoms of the element bind to form dioxygen."*

This method synthesizes questions with noun spans only, which is the most frequent span category in INQUISITIVE; we did not handle verbs or adjectives this way because of their variability.

We synthesized 45437 questions, including 42021 questions with an answer. We use special tokens to denote the start and end of the synthesized span.

C.2 ELI5 pretraining

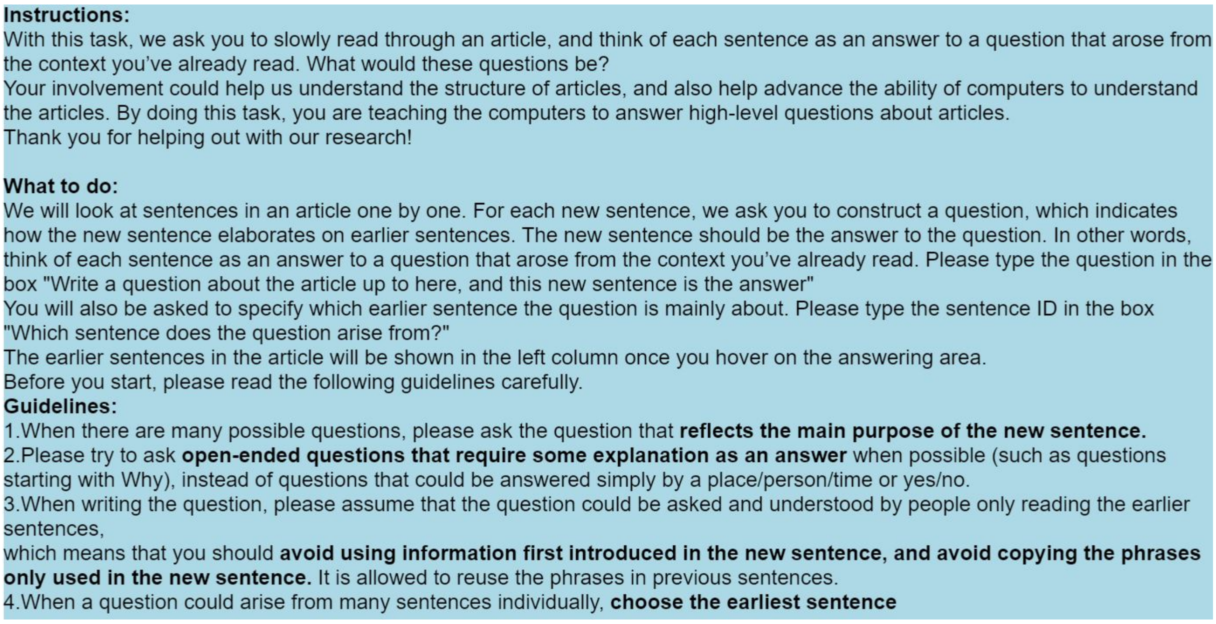
We retrieve the sentence in the supporting documents with the highest BM25 score as the approximate answer sentence. Following Petroni et al. (2021), we use the whole English Wikipedia as the supporting corpus instead of the original supporting documents. We combine sentences before and after the answer sentence with the answer itself to form the "article" as the input to the question answering model. The number of context sentences are randomly chosen so that the length of the synthesized article is comparable to INQUISITIVE articles and the answers are evenly distributed among different positions in the synthesized article.

To prevent low quality answers, we have an additional filtering step that keeps only the examples whose cosine similarity between the sentence embeddings (embedded using distilled SRoBERTa (Reimers and Gurevych, 2019)) of the retrieved answer and the gold answer is larger than 0.55. This resulted in 55740 examples.

C.3 Additional experimental details

We run our experiments on NVIDIA Tesla V100 SXM2 GPUs. We use 2 GPUs when training the model and it takes about 10 hours for the QA model. The longformer-base-4096 we use has roughly 148M parameters. Our hyperparameters are tuned on the validation set; we mainly tune the learning rate and the number of epochs. Our reported results are from a single run.

[5] We use the AllenNLP tool (Gardner et al., 2018).



Instructions:
With this task, we ask you to slowly read through an article, and think of each sentence as an answer to a question that arose from the context you've already read. What would these questions be?
Your involvement could help us understand the structure of articles, and also help advance the ability of computers to understand the articles. By doing this task, you are teaching the computers to answer high-level questions about articles.
Thank you for helping out with our research!

What to do:
We will look at sentences in an article one by one. For each new sentence, we ask you to construct a question, which indicates how the new sentence elaborates on earlier sentences. The new sentence should be the answer to the question. In other words, think of each sentence as an answer to a question that arose from the context you've already read. Please type the question in the box "Write a question about the article up to here, and this new sentence is the answer"
You will also be asked to specify which earlier sentence the question is mainly about. Please type the sentence ID in the box "Which sentence does the question arise from?"
The earlier sentences in the article will be shown in the left column once you hover on the answering area.
Before you start, please read the following guidelines carefully.
Guidelines:
1.When there are many possible questions, please ask the question that **reflects the main purpose of the new sentence.**
2.Please try to ask **open-ended questions that require some explanation as an answer** when possible (such as questions starting with Why), instead of questions that could be answered simply by a place/person/time or yes/no.
3.When writing the question, please assume that the question could be asked and understood by people only reading the earlier sentences,
which means that you should **avoid using information first introduced in the new sentence, and avoid copying the phrases only used in the new sentence.** It is allowed to reuse the phrases in previous sentences.
4.When a question could arise from many sentences individually, **choose the earliest sentence**

Figure 4: Key instructions of our crowdsourcing task for question collection.

D Instructions for question collection

Instructions for DCQA's crowdsourcing interface is shown in Figure 4.

E Copyright information related to DCQA

DCQA's annotated data is a subset of the articles used in the INQUISITIVE dataset (Ko et al., 2020). This data is sourced from three existing datasets: Newsela (Xu et al., 2015), WSJ articles from the Penn Treebank (Marcus et al., 1993), and Associated Press (Harman and Liberman, 1993). The Newsela dataset can be requested free of charge for researchers at https://newsela.com/data, and the authors have obtained permission to perform research on this data. The Penn Treebank is one of the most widely used resource in NLP, and is available via the LDC at https://catalog.ldc.upenn.edu/LDC99T42. The Associated Press data is also available from the LDC at https://catalog.ldc.upenn.edu/LDC93T3A.