QuALITY: Question Answering with Long Input Texts, Yes!

Richard Yuanzhe Pang* **Alicia Parrish*** **Nitish Joshi*** **Nikita Nangia** **Jason Phang**
Angelica Chen **Vishakh Padmakumar** **Johnny Ma** **Jana Thompson** **He He**
Samuel R. Bowman
New York University
{yzpang,alicia.v.parrish}@nyu.edu

Abstract

To enable building and testing models on long-document comprehension, we introduce QuALITY, a multiple-choice QA dataset with context passages in English that have an average length of about 5,000 tokens, much longer than typical current models can process. Unlike in prior work with passages, our questions are written and validated by contributors who have read the entire passage, rather than relying on summaries or excerpts. In addition, only half of the questions are answerable by annotators working under tight time constraints, indicating that skimming and simple search aren't enough to consistently perform well. Current models perform poorly on this task (55.4%) and significantly lag behind human performance (93.5%).

1 Introduction

Most of the best models for natural language understanding are restricted to processing only a few hundred words of text at a time, preventing them from solving tasks that require a holistic understanding of an entire article or story. Moving past this limitation would open up new applications in areas like news comprehension, summarization, or applied question answering.

We think that new benchmark datasets will help us do this. Most existing datasets (Rajpurkar et al., 2018; Fan et al., 2019; Lelkes et al., 2021) use shorter contexts that humans can read within a few minutes. While there are open-domain QA datasets that require longer contexts (Joshi et al., 2017; Zhu et al., 2021), the challenge is often finding the short excerpt that answers the questions at hand; however, long-document QA requires reading a long context and understanding it as a whole to correctly answer questions.

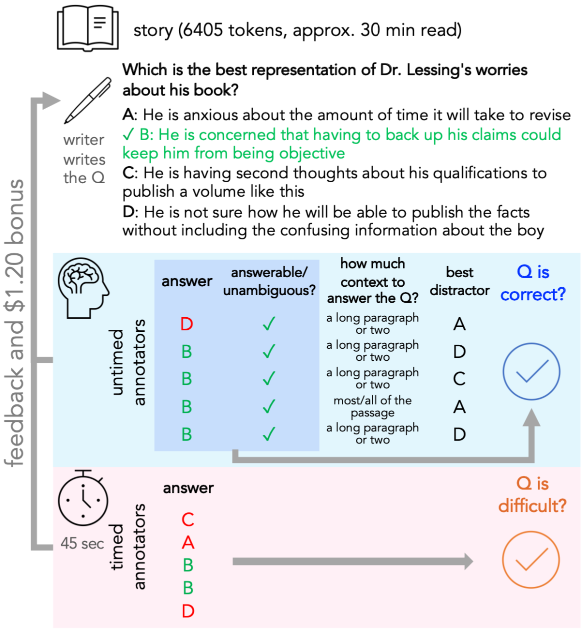

story (6405 tokens, approx. 30 min read)
Which is the best representation of Dr. Lessing's worries about his book?
A: He is anxious about the amount of time it will take to revise
✓ B: He is concerned that having to back up his claims could keep him from being objective
C: He is having second thoughts about his qualifications to publish a volume like this
D: He is not sure how he will be able to publish the facts without including the confusing information about the boy
writer writes the Q
feedback and $1.20 bonus
untimed annotators
answer
answerable/ unambiguous?
how much context to answer the Q?
best distractor
Q is correct?
D ✓ a long paragraph or two A
B ✓ a long paragraph or two D
B ✓ a long paragraph or two C
B ✓ most/all of the passage A
B ✓ a long paragraph or two D
45 sec
timed annotators
answer
C
A
B
B
D
Q is difficult?

Figure 1: The crowdsourcing pipeline with an example from QuALITY. One of our writers reads the passage and writes 10 questions. Each question is then validated with the help of five annotators who read the full article, plus five more who are given only 45 seconds per question. Writers receive feedback from both validations between batches of writing. If a majority of timed annotators get the question wrong, but the untimed annotators get it right, we classify the example as HARD and give an additional bonus to the writer.

NarrativeQA (Kočiský et al., 2018) is the most established existing long-text benchmark for language undesrtanding. It's a free-text-response QA dataset built around movie scripts and books, with an average of about 63k tokens of input per question. The authors creatively use *summaries* of the texts as the basis for their questions to make data collection relatively efficient. This protocol leads to short answers (avg. 4.7 tokens), and few questions require more complex explanation-based reasoning: >60% are what/who questions and <10% are why/reason questions. Further, the sources

[*] Equal contribution.

Source	Dif.	Question	Answer Options	Label
Gutenberg	Hard	Why was the Volpla vocabulary limited when the narrator took a few into the valley?	(a) They had not been alive long enough to learn enough English to communicate well (b) They were encountering concepts that were unfamiliar from the lab environment (c) They are not smart enough to have a fully developed language, no matter how hard they try (d) They were confusing their own language with English, having trouble keeping the languages separate	b
	Easy	What is Russell's greatest fear?	(a) Being disappointed (b) Losing his mind (c) Being lost and alone (d) Living forever	c
Slate	Hard	Which is NOT a reason why the narrator is concerned with the antichrist?	(a) Evangelical Christians are preaching that the end of the world is coming soon. (b) He is concerned that Christians will become violent toward Jews. (c) He thinks his life will be more important and influential than the average person. (d) He is conducting research for his dissertation.	d
	Easy	Why does the author tell a story about his vehicle?	(a) To talk about how fast he drives (b) To make a point about what has the most impact on the economy (c) To talk about safe driving speeds (d) To make a point about how many different things impact the unemployment rate	b
Misc.	Hard	How does Sara feel about the Chevrolet ad?	(a) She thinks it's a final chance to bond with her father (b) She is sorry she did not watch the whole ad before she reacted to it (c) She is upset at the glorification of the military (d) he is frustrated that it tokenized a Mexican family	b
	Easy	Why did Birmingham build over the Victorian era relics?	(a) To create space for a Maglev train (b) To erase their history (c) They were running out of room (d) To make technological progress	d

Table 1: Representative examples from the training and dev sets in QuALITY. Examples are selected randomly.

are usually famous, such that they are analyzed and discussed quite widely in the training data used by large language models. Additionally, the generation-based format comes with the additional hurdle of determining how to fairly assess accuracy, since metrics like BLUE, ROUGE, or BERTScore may not accurately convey the quality of generations (Wang et al., 2020; Durmus et al., 2020). To ease the burden of evaluation, we opt for a multiple-choice format to evaluate a model's long-document understanding ability.

We introduce our dataset QuALITY, Question Answering with Long Input Text, Yes![2] This is a multiple-choice QA dataset that uses source articles in English that are 2k–8k tokens.[3] We collect this dataset using a creative crowdsourcing pipeline that ensures that the examples have unambiguous answers but are still challenging. Example writers for QuALITY are instructed to carefully read the full source article before writing questions. They are also explicitly instructed to write questions that are unambiguous and require consolidating information from multiple parts of the text. Then, to ensure our questions require readers to understand the larger context from the passage, in addition to running standard validation where annotators read the source text and answer the questions, we also run speed validation (§2.3). In speed validation, annotators only have access to the article for 45 seconds, so they can only skim or search for phrases to answer the question. If a question is unanswerable in this setting but unambiguous and answerable in the standard untimed setting, we use it as a signal for question difficulty. This crowdsourcing process is slow and expensive ($9.10/question),[4] but we successfully collect a challenging, high-quality, long-document multiple-choice QA dataset. QuALITY has 6,737 questions in total, of which 3,360 questions are in the difficult subset, QuALITY-HARD. Table 1 shows representative EASY and HARD examples from different types of source texts.

We test Longformer, RoBERTa, DeBERTaV3, and T5 models, using as much of the full source text as possible. We also test two-step systems with an extractive model that passes shorter contexts to the QA model. We use fastText, DPR, or ROUGE-1 recall based matching with questions for text extraction. The best model performance is achieved by DeBERTaV3-large with DPR-based extraction, with an accuracy of 55.4%. The best model's accuracy on QuALITY-HARD is 46.7%.

[2] The data is available, along with links to code and additional resources, at https://github.com/nyu-mll/quality.

[3] 1.5k–6k words, not counting punctuation.

[4] This includes the cost of question writing and validation for questions that were discarded after validation.

Model accuracy is far below human accuracy on QuALITY, where human accuracy is 93.5% on the full dataset and 89.1% on QuALITY-HARD.

2 Data Collection

2.1 Overview

Sources In order to create a dataset that is both broadly usable and meets the goal of containing long input texts, we need to use only sources that are licensed under CC-BY-SA (or more permissive licenses) and contain a sufficient number of articles of at least 2k tokens that are likely to allow for complex questions. The three sources that we include in QuALITY are Project Gutenberg fiction stories (mostly science fiction),[5] Slate magazine articles from the Open American National Corpus (Fillmore et al., 1998; Ide and Suderman, 2004), and other nonfiction articles taken from The Long+Short,[6] Freesouls,[7] and the book Open Access (Suber, 2012). Table 3 shows how many articles and questions come from each. Most of the Gutenberg texts are from the 1950s–1970s, while the Slate and misc. texts are mostly from the 1990s and post-2000.

Texts are provided with the original HTML tags indicating paragraph breaks and basic formatting (e.g., italics), and it is in this format, with images removed, that we present the texts to our writers and annotators. In our dataset release, we also include a version of each file with this information stripped away, as current models, including our baselines, are not trained to consume these tags.

We set a maximum length for the texts at 6k words using word-level tokenization without counting HTML tags.[8] For around 40% of the Gutenberg articles, the full text data is much longer; in these cases we truncate the texts and manually check to make sure the truncation happens at a reasonable location (i.e., not in the middle of a paragraph).

Stages of Data Collection We collect data over several rounds with each group of writers in order to provide them with feedback throughout the process. We iterate through the following pipeline in each round: (i) we assign writers a set of passages, and they write 10 unique questions for each one (§2.2), (ii) annotators complete the speed validation task (§2.3.1), (iii) annotators complete the untimed validation task (§2.3.2), and (iv) writers are awarded bonuses and feedback based on the annotations.

2.2 Question Writing

We hire 22 experienced writers—most with degrees or professional experience related to literature or teaching—from the freelancing platform Upwork and design a multi-part incentive structure to encourage difficult yet answerable questions. Details about the hiring process and writer qualifications can be found in Appendix A.1.1.

The Writing Task We design a feedback and incentive structure to encourage writers towards questions that are answerable, unambiguous, and difficult. Writers construct examples over multiple rounds, and they receive (i) detailed feedback following each round based on the two validation tasks and (ii) bonuses based on how many of their questions met our criteria for HARD questions. Each writer constructs 10 questions with four answer options for a given passage, and they complete 6–30 such passages each round. Writers earn an average rate of $21.05/hr, after bonuses. Details about this process and the timeline can be found in Appendix A.1.2.

2.3 Data Validation

We use two separate validation tasks to evaluate if (i) the questions are difficult by testing if they are answerable under strict time constraints (speed validation) and (ii) the questions have a single correct answer (untimed validation). We recruit 45 annotators via Amazon Mechanical Turk (MTurk); details on the qualification process are in Appendix A.2.1.

2.3.1 Speed Validation

We want to ensure that the questions require understanding of the full text to answer correctly. If a person can quickly identify the answer to a question, such as through skimming or ctrl-F-style in-browser search, then the question does not require broader understanding of the passage in order to answer it correctly, and a model is likely to be able to identify the correct answer via extractive methods. More precisely, we aim to collect questions for which annotators, in the aggregate, are unable to select the correct answer under strict time constraints, and we construct a speed validation task

[5] http://www.gutenberg.org

[6] http://thelongandshort.org

[7] http://freesouls.cc

[8] But if we use spaCy tokenization, the maximum number of tokens is larger, as shown in Figure 2.

to test this. Questions that pass this bar make up the HARD subset of QuALITY.

Procedure We collect five annotations per task; within each task, questions are presented one at a time to ensure the time limit is consistent for each question. The worker first reads the question and the four answer options without access to the passage. Then they press a button to reveal the entire passage, and they have 40 seconds to skim or search for keywords (e.g., with ctrl+F) to determine the correct answer. After the timer runs out, the passage disappears, and they have 5 more seconds to select an answer. The user interface can be found in Appendix A. Each HIT consists of 10 questions, each from different passages, and the order of the answer options is randomized. Within each HIT, there are nine questions written by the Upwork writers and one question written by the authors as a catch question. We pay workers $2.25 per HIT and award a bonus of $0.20 for each correct answer. On average, workers earn a bonus of $1.03 per HIT, and we estimate based on workers' survey responses that each HIT takes 11-12 minutes, for an effective rate of just over $17/hr. We use the catch questions to track annotator performance and ensure that all workers are performing well above chance on these examples, indicating that they are consistently making a faithful effort to find the answer in the text (see Appendix A.2.2 for additional details on the task, catch questions, and annotator performance).

2.3.2 Untimed Validation

To ensure all questions in QuALITY are correct and unambiguous, we conduct a validation task without a time limit, but with strong incentives towards accuracy. We collect three annotations for each example in the training set, and five annotations for each example in the dev and test sets.

Procedure Each task consists of one passage with all 20 questions created by the writers. Each of the 20 reading comprehension questions has three evaluation questions immediately below it. Workers are instructed to first read the passage carefully, then answer all the questions. Each HIT pays $6.50, with a $0.50 bonus for each question in which both the reading comprehension question and evaluation question 1 (see below) agree with the majority vote label. [9] We estimate based on survey responses that workers spend about 50–60 minutes on this task, and the average bonus rate is $8.13 per task, for an average effective pay of $15.96/hr.

Evaluation Questions We include evaluation questions to help assess question quality. We ask the three evaluation questions shown in Table 2 immediately following each reading comprehension question. The first evaluation question is used to determine inclusion into the final dataset, as we exclude any questions for which the majority of annotators marked that the question was either ambiguous or unanswerable. The second and third evaluation questions are used for feedback to the writers.

Question	Answer Options
Q1. Is the question answerable and unambiguous?	○ Yes, there is a single answer choice that is the most correct. ○ No, two or more answer choices are equally correct. ○ No, it is unclear what the question is asking, or the question or answer choices are unrelated to the passage.
Q2. How much of the passage/text is needed as context to answer this question correctly?	○ Only a sentence or two of context ○ At least a long paragraph or two of context ○ At least a third of the passage for context ○ Most or all of the passage for context
Q3. Which of the options that you did not select was the best "distractor" item?	○ Option 1 ○ Option 2 ○ Option 3 ○ Option 4

Table 2: Evaluation questions asked following each reading comprehension question during the untimed validation task. Q3 additionally defined distractor items as "an answer choice that you might be tempted to select if you hadn't read the text very closely."

We find that responses to the first two evaluation questions slightly differ between the HARD and EASY subsets of QuALITY. For Q1, individual raters are less likely to rate a HARD question as answerable and unambiguous (92.8%) compared to an EASY question (95.1%). For Q2, in the HARD subset, 26.1% of the time, the question is rated as needing at least a third of the context or more (the 3rd and 4th options), compared to 21.7% of the time in the EASY subset.

[9] Bonuses were $0.40 during the first round and were calculated based only on accuracy on the reading comprehension questions. We updated The bonus criteria and amount as a result of our evaluation of the results and worker feedback on a short survey.

Annotator Performance We track annotator performance throughout data collection and remove any workers whose average accuracy falls below 75% in any given round. Annotator agreement on the reading comprehension questions for each passage is high, with a median Krippendorff's alpha of 0.71. Agreement on evaluation question 1 is also high, with 92.6% individual agreement with the majority vote annotationwith the two 'No' options collapsed for analysis). As evaluation questions 2 and 3 are much more subjective, responses are quite noisy, with median alpha values of 0.12 and 0.21, respectively. Additional details on annotator performance and our protocol for reannotating data can be found in Appendix A.2.3.

3 Dataset Information and Analysis

After aggregating the labels assigned via untimed validations with the original writer's label, we calculate the gold label via majority vote of annotators.[10] We only keep the questions for which (i) a majority vote label (strictly larger than 50%) can be assigned and (ii) the majority of annotators rate the questions as answerable and unambiguous. We are left with 6,737 out of 7,620 (88.4%) questions that meet this inclusion criteria for QuALITY.[11] We label a subset of QuALITY as HARD. These correspond to questions for which the majority of the annotators answer the questions incorrectly in the speed validation setting, and this constitutes 49.9% of the final dataset.

3.1 Human Accuracy

We estimate human accuracy on QuALITY on a random sample of 20 passages (367 questions). Each question is annotated by 3 new validation annotators who had not previously annotated that passage, and whose labels do not contribute to the assignment of the gold label. We calculate the majority vote of these three annotators, which yields an accuracy of 93.5% relative to the gold label. This breaks down to 89.1% on the HARD subset and 97.0% on the EASY subset. In 37.5% of the items that human evaluators got wrong, all three predicted answers differed. These annotators marked the questions as answerable and unambiguous 98.4% of the time.

3.2 Size and Splits

We split the data into train/dev/test sets such that there is minimal overlap in question writers between the training set and the dev/test sets. This ensures that a model will not be rewarded for overfitting to any idiosyncrasies of a single writer's style.

Table 3 shows the number of articles in QuALITY and HARD questions for each of the split. Gutenberg sources result in the highest proportion of HARD questions, and miscellaneous sources result in the lowest proportion.

3.3 Length

Article Length Figure 2a shows the article lengths in QuALITY. The two peaks in the histogram correspond to articles from Slate and Project Gutenberg. The average context length in QuALITY is 5,159 tokens, which is much longer than other existing challenging QA datasets — CosmosQA (Huang et al., 2019) and RACE (Lai et al., 2017) contain an average context length of 70 and 322 tokens, respectively.

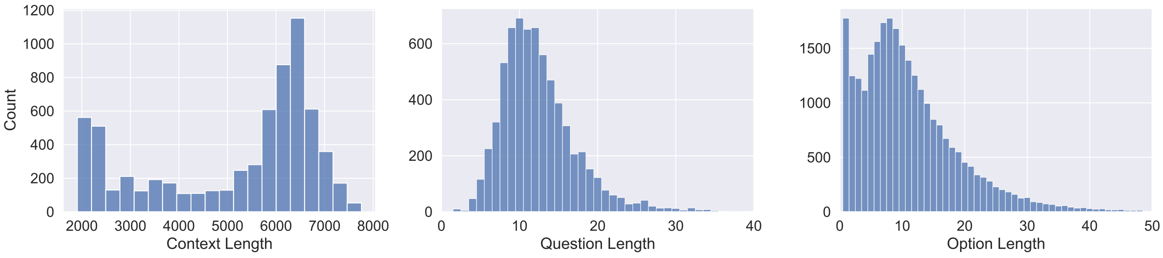
Question and Option Lengths We similarly plot the question length and option length in QuALITY in Figure 2b and Figure 2c, respectively. The average question length is 12.5 tokens, whereas the average length of an option is 11.2 tokens.

3.4 Lexical Overlap

Prior work has shown that a lot of questions in existing datasets such as SQuAD can be answered by exploiting lexical overlap of the question with the article (Weissenborn et al., 2017). To understand how effective this heuristic is in QuALITY, we compute the lexical overlap between the options and the article in QuALITY. The lexical overlap is computed as the fraction of the tokens in the option which are present in the article. Figure 3 plots the distribution of lexical overlap for the correct options and the incorrect options — since each question has three incorrect options, we plot the maximum lexical overlap among the three. We observe that correct options do not have a higher

[10] This definition of gold label follows the protocol in MNLI (Williams et al., 2018). For each question, both the writer's label and the gold label are provided in the dataset. We hypothesize that writers sometimes mislabel the correct option, because many say that they usually work on a separate document before filling in our data collection UI. We analyze a random subset of the examples where the writer's label and the gold label do not match, and confirm that the assigned gold labels are correct. The gold label differs from the writer's label for ∼4% (274/7620) of the questions.

[11] We release the discarded questions corresponding to passages in the train and dev sets as part of a supplemental dataset.



(a) (b) (c)

Figure 2: Article length, question length, and option length in QuALITY. The average article, question, and option is 5,159 tokens, 12.5 tokens, and 11.2 tokens, respectively. The maximum length of an article, question, and option is 7,759 tokens, 103 tokens, and 75 tokens, respectively. The histograms are truncated to only keep visible mass.

	Gutenberg				Slate				misc				all			
Split	Articles	All Qs	HARD Qs	% HARD	Articles	All Qs	HARD Qs	% HARD	Articles	All Qs	HARD Qs	% HARD	Articles	All Qs	HARD Qs	% HARD
Train	118	2000	1056	52.8	22	355	142	40.0	10	168	53	31.5	150	2523	1251	49.5
Dev	86	1552	873	56.2	19	351	149	42.5	10	183	43	23.5	115	2086	1065	51.1
Test	81	1486	828	55.7	25	450	170	37.8	10	192	46	24.0	116	2128	1044	49.1
All	285	5038	2757	54.7	66	1156	461	40.0	30	543	142	26.2	381	6737	3360	49.9

Table 3: Data splits within QuALITY by different sources. Items that did not pass validation are excluded from this table. 'HARD Qs' are the number of questions in the QuALITY-HARD subset, these are the questions that annotators could not correctly answer in speed-validation, but annotators did correctly answer in the standard untimed validation.

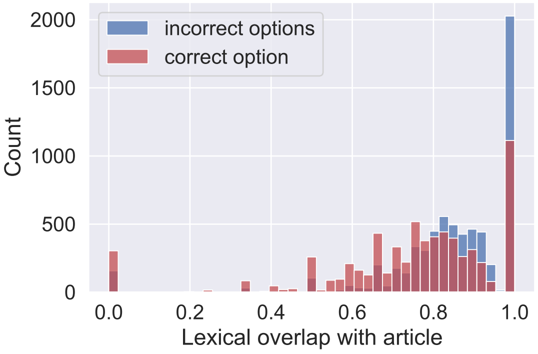


Figure 3: Lexical overlap of the correct and incorrect options with the context. Since each question has three incorrect options, we use the option with the highest lexical overlap. Simply predicting the option with the highest lexical overlap achieves only 26.6% accuracy.

lexical overlap than the incorrect options, thus making it difficult for models to rely on this heuristic.

3.5 Question Types

We analyze the proportion of question types that occur in QuALITY by automatically categorizing each question based on the first question word that it contains.[12] The results of this analysis are provided in Table 4, along with examples from each question type. QuALITY contains many questions that require complex responses about "how" and "why" an event happened in a greater proportion of cases compared to similar datasets such as NarrativeQA. However, we do not observe that our measure of question difficulty is affected by question type.

3.6 Reasoning Strategies

As part of a qualitative analysis, we manually annotate the reasoning strategy needed in each question, and we present these results in Table 5. We take a subset of 500 questions randomly selected from the full dataset and manually annotate them. Each question is annotated by two of the authors; any disagreements in categorization are resolved via discussion. As we do not read the full passages, it is not always possible to completely tell what reasoning strategy will be needed, but we consider both the question and answer options in categorizing each item. We find that many of the questions rely on (i) reasoning about the best way to describe something from the text, (ii) determining

[12] In cases where the question starts with an auxiliary verb, or where there is no question word, but an auxiliary verb appears after a comma, we categorize the question as a yes/no question.

Question Type	# EASY	# HARD	% total	Example from the Training Set
what	1361	1471	42.2	What is the immediate significance of Ed defending the ads on his Facebook?
why	832	825	24.6	Why does Howell not want Linton to approach Snead in the restaurant?
how	385	416	11.9	How does Barker view his own film?
which	253	244	7.4	Which word least describes McGill?
who	151	132	4.2	Who is the most hated celebrity of 1999?
how + meas.	51	75	1.9	How many caves had Garmon and Rolf traveled through before their crash?
yes/no	53	55	1.6	Was it Nelson's decision to become part of the military?
where	43	42	1.3	Where was the space craft heading in the end?
when	35	34	1.0	When did the Hanseatic League begin?
other	155	124	4.1	Dole's quote would have been perceived as ______ if it had included included the exclamation points from his tone?

Table 4: Different question types in QuALITY, split by the number that are categorized as HARD and EASY based on the speed validation. Most of the questions in the 'other' category are finish-the-phrase style questions, and for the example in the table, the answer options are different ways that sentence could be completed. 'How + meas.' collapses multiple questions with 'how' plus some measurement, such as 'how long' or 'how many.' Note that in most of the yes/no questions, the answer options include the necessary reasoning to support the yes/no answer. Examples shown in the table are randomly selected from the training set, with the caveat that we selected the 'when' question by hand, since about half of the questions categorized as 'when' are actually of the form 'when X happened, what did ...'

the correct explanation for something in the text, or (iii) the reader making an interpretation or using symbolism from the passage. All three of these reasoning types are likely to rely on broader context from the passage, compared to questions about who did something or where something happened. We also find that, despite questions using 'what' being the most frequent in the question-types analysis, very few of the questions in QuALITY depend on reasoning about objects or entities. Rather, most of these 'what' questions are asking about the description of a person or situation, or they are asking for an interpretation from the reader. Further details about this analysis, the categories used, and examples of each reasoning type are provided in Appendix C.

Reasoning Type	# HARD/ 251	# EASY/ 249	% of total
Description	89	77	33.2
Why/reason	73	83	31.2
Symbolism/interpretation	76	63	27.8
How/method	25	19	8.8
Event	17	18	7.0
Person	11	17	5.6
Not/except	13	6	3.8
Relation	12	7	3.8
Entity	7	9	3.2
Finish the Phrase	3	12	3.0
Location	5	7	2.4
Numeric	5	6	2.2
Object	5	4	1.8
What if	3	4	1.4
Duration	1	2	0.6

Table 5: Qualitative assessment on a 500 example subset of QuALITY, split by difficulty, and categorizing the different kinds of things that need to be reasoned about to answer questions in the dataset. These categories are inspired by those used in NarrativeQA and adapted for the reasoning types observed in our data. Questions can require multiple reasoning types, so values do not add up to 100%. The HARD questions are slightly more likely to have multiple reasoning types.

4 Baseline Experiments

4.1 Models

Long-Context Models We experiment with using the Longformer model (Beltagy et al., 2020), which uses a combination of sliding-window local attention and global attention to encode long inputs. The Longformer encoder models support up to a maximum of 4,096 tokens.[13] We test Longformer because it is likely to fit most or all of the context needed to answer the questions for the majority of examples in QuALITY.[14] The maximum token limit is still a challenge, though, and we expect models that can take longer inputs to perform better.

Extractive Models As an alternative to feeding the whole input context into a transformer model or truncating, we also test retrieval methods to score and extract relevant sentences from the passage and

[13] Longformer Encoder-Decoder (LED) supports up to 16,384 encoder input tokens, but due to difficulty tuning the model, our results are still pending.

[14] We always make sure that the entire question and all options are visible to models, but the article is sometimes truncated.

feed only the selected sentences as inputs to a given model. This extractive approach allows us to use a wider range of higher-performing short-sequence transformer models, at the cost of missing some context that those models may need.

Using the question as a retrieval query, we score each sentence in the passage relative to the query. We then select sentences in order of descending relevance until we reach a maximum of 300 words.[15] We then sort the selected sentences based on the original passage order and use the concatenated sentences as the 'passage' for that example.

We consider three scoring methods. First, we use ROUGE-1 recall relative to the query. Second, we use cosine similarity based on bag-of-words of fastText (Bojanowski et al., 2017) embeddings. Third, we use DPR (Karpukhin et al., 2020), a model trained for open-domain retrieval for QA. Because DPR tackles span-based question-answering, their *reader* model is unsuitable for our multiple-choice dataset. However, we are able to use the *retriever* model for extraction, using the separate question- and context-encoders to encode our question and context sentences to vector representations. We then score similarity based on the negative Euclidean (L^2) distance.

After extraction, we apply standard models for multiple-choice question-answering: RoBERTa (Liu et al., 2019) and DeBERTaV3 (He et al., 2021) encoder models, and the T5(Raffel et al., 2020) encoder-decoder model. To establish an upper bound of how well extractive models can do, we also introduce an oracle baseline in which we apply the same extraction strategy described above, but we use the correct answer as the extraction query.

Question-Only Baselines To test for dataset artifacts, we consider a baseline where we only give the models the questions and answer options, leaving out the passage. We test this question-only baseline for each model type.

Supplementary Training Data To supplement the training examples in QuALITY, we incorporate additional training examples from the RACE task dataset (Lai et al., 2017). Like QuALITY, RACE is a passage-based, four-way multiple-choice question-answering dataset. Although the passages are much shorter (321.9 words on average), the training set is large ($\sim$88k questions), so we can expect reasonable knowledge transfer from RACE to QuALITY. We use the full RACE dataset, including both middle-school and high-school questions, for our intermediate training.

We consider three fine-tuning formats: (1) fine-tuning on QuALITY data, (2) fine-tuning on RACE and zero-shot evaluating on QuALITY, and (3) applying intermediate training (Phang et al., 2018; Pruksachatkun et al., 2020) by first fine-tuning on RACE and then QuALITY.

4.2 Results and Analysis

Table 6 shows model performance on the test set. The results on the development set can be found in Appendix E.3.

All results in Table 6 fall well below human performance. There is a gap of 38.1 points between our current best-performing model (DeBERTaV3-large trained on RACE→QuALITY, using DPR-based extraction) and human performance on the full test set. On QuALITY-HARD, this gap increases to 42.4 points.

Comparing models using different training data, we see that the RACE→QuALITY results outperform RACE results in most cases. Fine-tuning on QuALITY contributes to a small performance gain. Both RACE and RACE→QuALITY significantly outperform the QuALITY only results, likely because of the small size of the QuALITY training set, though this suggests that knowledge transfer from RACE is useful.

Using the same training data, DeBERTaV3-large consistently outperforms other models. In terms of extraction strategies, within each approach, DPR-based extraction almost always produces the best result. Compared to the RoBERTa and DeBERTa models fined-tuned on short contexts, the Longformer models appear to struggle to learn to tackle the task from the long inputs, underperforming even the RoBERTa-base extraction-based models. We speculate that a combination of more long-context training data and better long-context models may improve performance beyond the extraction-based models. As with the other models, intermediate training on RACE improves performance on QuALITY.

Question-Only Baselines The best-performing question-only baseline is DeBERTaV3-large using the RACE→QuALITY setting for training, achieving an accuracy of 43.3%. The corresponding performance is only 12.1 percentage points lower than the DeBERTaV3-large's performance with text ex-

[15] Punctuation is not counted towards this limit.

			Extraction Based on Qs			
Training Data	Model	Full	R-1	fastText	DPR	Question-Only
QuALITY	Longformer-base	30.7/29.3	–	–	–	–
	RoBERTa-base	–	33.4/30.7	39.7/36.1	39.9/34.0	36.6/34.8
	RoBERTa-large	–	29.4/28.0	42.7/35.7	26.2/25.1	26.4/25.7
	DeBERTaV3-base	–	36.7/35.7	38.9/35.9	44.1/38.5	38.2/35.6
	DeBERTaV3-large	–	46.5/39.3	45.5/40.2	49.0/41.2	39.7/35.2
	T5-base	–	28.0/28.0	28.9/27.4	29.3/29.1	30.1/29.9
RACE	Longformer-base	35.2/30.8	–	–	–	–
	RoBERTa-base	–	42.4/36.8	43.2/37.2	44.2/36.1	33.8/29.7
	RoBERTa-large	–	47.0/37.5	47.9/40.2	48.7/40.2	36.6/33.1
	DeBERTaV3-base	–	45.3/36.1	46.1/39.0	47.8/39.4	34.7/30.5
	DeBERTaV3-large	–	52.9/43.4	51.2/42.4	53.0/44.4	36.5/30.0
	T5-base	–	41.5/38.6	42.3/39.9	43.4/41.0	36.5/34.8
RACE	Longformer-base	39.5/35.3	–	–	–	–
↓	RoBERTa-base	–	42.1/38.3	43.0/40.1	44.3/39.8	38.1/37.5
QuALITY	RoBERTa-large	–	48.0/40.8	50.4/43.7	51.4/44.7	40.4/37.1
	DeBERTaV3-base	–	46.8/38.7	49.8/43.2	51.2/42.4	41.4/37.9
	DeBERTaV3-large	–	53.8/**46.3**	54.7/**46.7**	**55.4**/46.1	43.3/38.2
–	Human Annotators	93.5/89.1	–	–	–	–

Table 6: Accuracy on the full QuALITY test set and accuracy on the QuALITY-HARD subset (formatted as full/HARD). The "Full" columns shows results from training with the source inputs from QuALITY, truncated to fit into memory, without using an extractive model to selects portions of text. R-1 (ROUGE-1), fastText, DPR are three extraction methods (§4.1) used to select relevant portions of the source text. In the "Training data" column, QuALITY means the models are fine-tuned on QuALITY data only; RACE means that models are fine-tuned on RACE only; RACE→QuALITY means that we first do intermediate training on RACE and then fine-tune on QuALITY. Poor RoBERTa-large performance (for training on QuALITY only) is likely due to unstable training given our small training set (Mosbach et al., 2021).

cerpts from DPR. This small margin of improvement may indicate that current models are not effectively using the input contexts.

QuALITY vs. QuALITY-HARD Model performance on QuALITY-HARD is always lower than on the full QuALITY test-set, including on the question-only baselines. This suggests that speed-validation filtering yields questions that are more challenging for human annotators *and* models.

Extraction by Oracle Answer We show in Appendix E.3, Table 9 the results of the oracle-answer-based extraction on the development set. Compared to Table 8, we see that using the oracle answers for extraction improves performance significantly (topping out at 78.3%), but is still below human performance by 15 points. Thus, even in the unrealistic scenario of an oracle extractor with models trained on oracle-extracted sentences, the extraction-based models still underperform humans. This demonstrates that extracting relevant excerpts alone is insufficient to solve QuALITY questions, and that QuALITY questions require reasoning over the full passage to answer.

5 Related Work

Rogers et al. (2021) surveys the QA dataset explosion of recent years and the many formats and types of QA datasets. The QA datasets closest to our work either have long contexts or require consolidating information from multiple parts of the source text through multi-hop reasoning.

TriviaQA (Joshi et al., 2017) and SearchQA (Dunn et al., 2017) contain questions with more than one document as the context, but since the supporting documents are collected after writing the question-answer pairs, most questions can be answered after retrieving a short context. HotpotQA (Yang et al., 2018), QAngaroo (Welbl et al., 2018) and ComplexWebQuestions (Talmor and Berant, 2018) were constructed to have more challenging questions which require multi-hop reasoning across multiple paragraphs. However, there has been recent work (Jiang and Bansal, 2019; Min et al., 2019) showing that these datasets contain reasoning shortcuts and a large fraction of the questions can be answered with single-hop reasoning.

NarrativeQA (Kočiský et al., 2018) is the most similar work to ours. NarrativeQA uses entire Gutenberg books and film scripts as contexts, with an average length of 60k tokens. The authors creatively make the data-collection task tractable by using Wikipedia summaries for the books as context when crowdsourcing questions. Unlike QuALITY, NarrativeQA is a freeform generation-based task. And while there are many existing multiple-choice QA datasets (Richardson et al., 2013; Hill et al., 2015; Lai et al., 2017; Bajgar et al., 2016; Huang et al., 2019), these datasets use much shorter context passages (<500 tokens) than our dataset.

One of the primary challenges of building a long-document QA dataset like QuALITY or NarrativeQA is building a tractable crowdsourcing pipeline that enables collecting high-quality examples. Roit et al. (2020) collect a challenging QA-SRL dataset by carefully hiring and training crowdworkers, with a strict qualification followed by 2 hours of training that includes giving workers extensive feedback. Nangia et al. (2021) compare crowdsourcing methods for collecting high-quality QA data and find that a long training process with iterative feedback and qualifications is an effective crowdsourcing strategy.

6 Conclusion

We introduce the long-document QA dataset QuALITY. This dataset was crowdsourced and validated by humans to ensure that the questions are answerable, unambiguous, and challenging. The QuALITY-HARD subset, comprising half the dataset, consists of questions that are unanswerable by annotators working under tight time constraints, helping ensure that skimming and simple search don't yield high performance.

We find that current models significantly lag behind human performance on QuALITY, with a 38.1 percentage point gap between human annotators and the best performing model. The gap is even wider on QuALITY-HARD, at 42.3 points. We hope that research that aims at this gap will contribute to expanding the scope of texts on which effective NLU systems can be applied.

Ethical Considerations

Both the authors of our source texts and the authors of our questions are based primarily in the US, and represent a relatively privileged, educated population. A system that performs well on our dataset is, thus, only demonstrating its effectiveness on mainstream US English, and should not be presumed to be effective on text from other populations.

Acknowledgements

This project has benefited from financial support to SB by Eric and Wendy Schmidt (made by recommendation of the Schmidt Futures program), Samsung Research (under the project *Improving Deep Learning using Latent Structure*), Samsung Advanced Institute of Technology (under the project *Next Generation Deep Learning: From Pattern Recognition to AI*), and Apple. This material is based upon work supported by the National Science Foundation under Grant Nos. 1922658 and 2046556. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

We thank Jon Ander Campos, Alex Wang, and Omer Levy for valuable discussion. We also thank the writers who wrote our source texts (credited in the data itself) and the writers who wrote our questions: Megan Barbee, Bridget Barrett, Kourtney Bradley, Kyle J. Brown, Alicia Chatten, Christine D., Leah Dorschner-Karim, Bobbie Dunn, Charisse Hake, Javier Hernandez, Molly Montgomery, Carilee Moran, Tracy M. Snyder, Lorna Stevenson, Isaiah Swanson, Kyla Thiel, Lisa V., Ryan Warrick, Julia Williamson, and others who chose to remain anonymous.

Author Contributions

- Locating appropriate passage sources: AP, JM, VP, AC, NN, NJ, JT
- Preprocessing passages: AC, NN, NJ, JP
- Data collection protocol design: NN, AP, RP, HH
- Data collection user interface: RP, JT
- Data collection backend infrastructure: AC
- Data collection management and postprocessing: RP, AP, NJ
- Writing catch questions: VP, JM, JT, AP, NN, NJ, RP
- Data analysis: AP, NJ, RP, VP
- Modeling: JP, AC
- Writing: AP, RP, NN, NJ, JP, HH, SB
- Project management: RP, AP

- Advising: SB

References

Ondrej Bajgar, Rudolf Kadlec, and Jan Kleindienst. 2016. Embracing data abundance: Booktest dataset for reading comprehension. *arXiv preprint arXiv:1610.00956*.

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

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. SearchQA: A new Q&A dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. 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.

Charles Fillmore, Nancy Ide, Daniel Jurafsky, and Catherine Macleod. 1998. An American national corpus: A proposal. In *Proceedings of the First Annual Conference on Language Resources and Evaluation*, pages 965–969. Citeseer.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The Goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. 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 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Nancy Ide and Keith Suderman. 2004. The American national corpus first release. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Yichen Jiang and Mohit Bansal. 2019. Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop QA. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2726–2736, Florence, Italy. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

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

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Adam D. Lelkes, Vinh Q. Tran, and Cong Yu. 2021. Quiz-style question generation for news stories. In *Proceedings of the the Web Conference 2021*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. Compositional questions do not necessitate multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4249–4257, Florence, Italy. Association for Computational Linguistics.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning {bert}: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations*.

Nikita Nangia, Saku Sugawara, Harsh Trivedi, Alex Warstadt, Clara Vania, and Samuel R. Bowman. 2021. What ingredients make for an effective crowdsourcing protocol for difficult NLU data collection tasks? 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 1221–1235, Online. Association for Computational Linguistics.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on STILTs: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA. Association for Computational Linguistics.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2021. QA dataset explosion: A taxonomy of NLP resources for question answering and reading comprehension. *arXiv preprint arXiv:2107.12708*.

Paul Roit, Ayal Klein, Daniela Stepanov, Jonathan Mamou, Julian Michael, Gabriel Stanovsky, Luke Zettlemoyer, and Ido Dagan. 2020. Controlled crowdsourcing for high-quality QA-SRL annotation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7008–7013, Online. Association for Computational Linguistics.

Peter Suber. 2012. *Open Access*. MIT Press.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 641–651, New Orleans, Louisiana. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Dirk Weissenborn, Georg Wiese, and Laura Seiffe. 2017. Making neural QA as simple as possible but not simpler. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 271–280, Vancouver, Canada. Association for Computational Linguistics.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 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, Online. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William 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, Brussels, Belgium. Association for Computational Linguistics.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*.

A Details on Writing, Speed Validation, and Untimed Validation

A.1 Writing

A.1.1 Writer Recruitment

We need writers who have good reading comprehension skills and/or writers who have experience constructing reading comprehension questions (e.g., literature teachers who have experience writing tests for their students). We hire two groups of writers on the freelancing platform Upwork (the second group two months after the first group, after we decided to increase the final size of the dataset). For each group, we advertise a task titled *Writing college-level reading comprehension questions*. The job post is visible to all U.S. Upwork freelancers, and we specifically send out job invitations to promising freelancers who are writers or teachers, or who have college-level degrees in English, Literature, Creative Writing, Philosophy, Education, or similar fields.

For the first group, we received 104 applications in the span of two weeks. Of those, we selected 26 people to complete a qualification task as a paid interview. For the second group, we received 65 applications and interviewed 11. The interview task consists of (i) reading through detailed instructions, (ii) reading through a tutorial example passage with 10 example questions, each with an explanation of what made it a good or a bad question, and (iii) writing 10 reading comprehension questions for a new passage; regardless of whether we eventually hire them, we pay workers $30 and estimate that this task takes 2 hours to complete. Three authors (of this paper) then assess each writer's work using the following criteria: (i) whether the writer-provided correct answers are actually the correct answers, (ii) whether the questions are answerable and unambiguous, and (iii) whether more than just a few sentences of context are needed to correctly answer the questions. Based on these criteria, we select the top performing 15 writers to continue on to the main task in the first group, and 7 in the second group.

For the 22 writers we hire after the interview, 15 have a college degree in English, literature, philosophy, creative writing, or education; 4 of these 15 writers are Ph.D. students or graduates. 11 of the 22 writers have taught high-school or college-level English or literature classes; among these 11 writers, 7 have 5+ years of teaching experience. 2 of the 22 writers mention that they write novels.

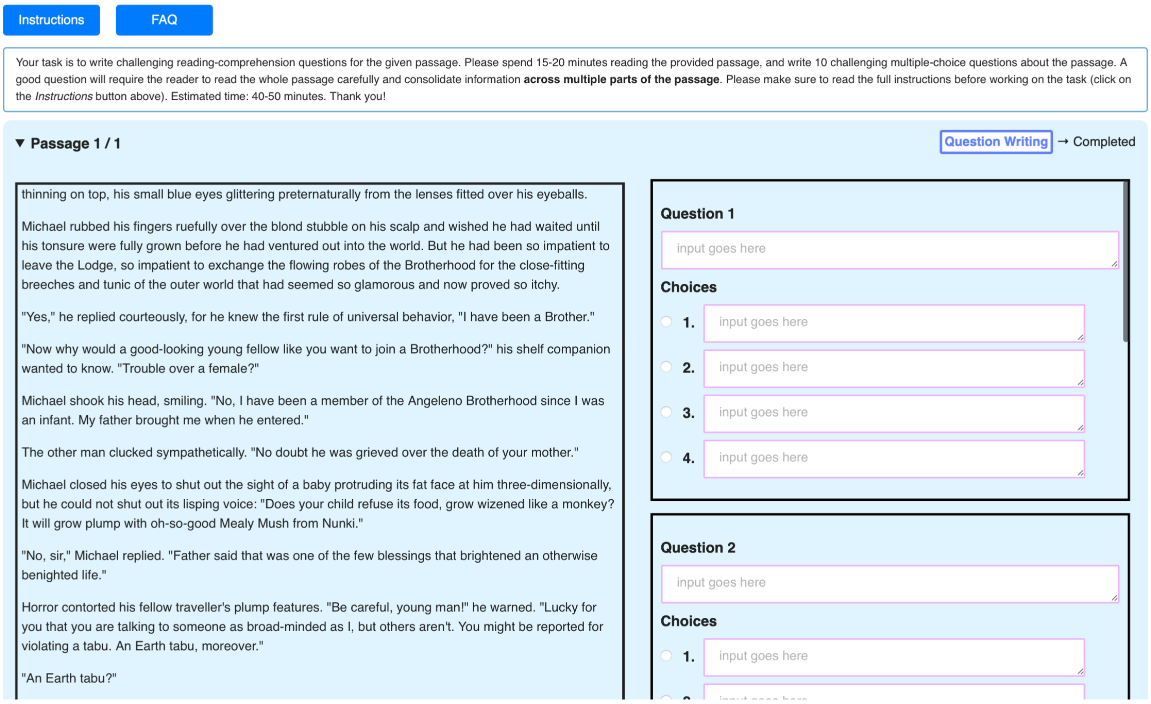
A.1.2 Writing Task

Bonuses on top of the base rates for a single project are a feature of Upwork. Freelancers also value public project testimonials which are shown to future clients. We use these two media to incentivize workers to write high-quality questions.

Each writer constructs 10 questions for a given passage, completing 6-30 passages in a given round and continuing for three complete rounds.[16] Each round is followed by feedback about how many questions validators decide are difficult and well-written, allowing writers to improve for the next round. Writers earn $12.50 per passage and receive a bonus of $1.20 for each question that meets the following criteria: (i) the majority of validators agree with the writer's original label, (ii) a majority of validators rated the question as answerable and unambiguous, and (iii) the majority of validators answered the question *incorrectly* in the speed validation task (§2.3.1). On average, writers receive bonuses on 4.2 questions per passage, resulting in average earnings of $17.54 per passage. Based on writer self-reports, the median time to complete one writing task is about 50 minutes, for an effective rate of $21.05/hr. Upwork charges fees on the workers' end. We account for this by adding an extra 20% to their pay, bringing our final cost to $2.10 per question.

Besides using the monetary bonus as an incentive for writing answerable, unambiguous, and difficult questions, we also instruct writers that their questions should use the entire context. Throughout the course of data collection, we provide writers with detailed feedback based on validations (detailed in §2.3.2) and this feedback includes information about how much of the passage needed to be read in order to answer the question. We monitor the proportion of questions that require more than a few paragraphs of context to answer correctly; if this rate significantly lags behind other writers, we inform the writers that their work is falling below expectations and ask them to be more careful with this issue in the next round. We also encourage writers to write difficult distractors, and the feedback we provide also contains what anno-

[16] On average, group 1 writers complete 6, 14, 30 articles for the three batches, respectively; group 2 writers complete 6, 14, 20 articles for the three batches, respectively. The writing time limits for batches 1, 2, 3 are around 1, 2, 3 weeks, respectively. Validation for any batch takes less than a week.



Your task is to write challenging reading-comprehension questions for the given passage. Please spend 15-20 minutes reading the provided passage, and write 10 challenging multiple-choice questions about the passage. A good question will require the reader to read the whole passage carefully and consolidate information **across multiple parts of the passage**. Please make sure to read the full instructions before working on the task (click on the *Instructions* button above). Estimated time: 40-50 minutes. Thank you!

▼ Passage 1 / 1

Question Writing → Completed

thinning on top, his small blue eyes glittering preternaturally from the lenses fitted over his eyeballs.

Michael rubbed his fingers ruefully over the blond stubble on his scalp and wished he had waited until his tonsure were fully grown before he had ventured out into the world. But he had been so impatient to leave the Lodge, so impatient to exchange the flowing robes of the Brotherhood for the close-fitting breeches and tunic of the outer world that had seemed so glamorous and now proved so itchy.

"Yes," he replied courteously, for he knew the first rule of universal behavior, "I have been a Brother."

"Now why would a good-looking young fellow like you want to join a Brotherhood?" his shelf companion wanted to know. "Trouble over a female?"

Michael shook his head, smiling. "No, I have been a member of the Angeleno Brotherhood since I was an infant. My father brought me when he entered."

The other man clucked sympathetically. "No doubt he was grieved over the death of your mother."

Michael closed his eyes to shut out the sight of a baby protruding its fat face at him three-dimensionally, but he could not shut out its lisping voice: "Does your child refuse its food, grow wizened like a monkey? It will grow plump with oh-so-good Mealy Mush from Nunki."

"No, sir," Michael replied. "Father said that was one of the few blessings that brightened an otherwise benighted life."

Horror contorted his fellow traveller's plump features. "Be careful, young man!" he warned. "Lucky for you that you are talking to someone as broad-minded as I, but others aren't. You might be reported for violating a tabu. An Earth tabu, moreover."

"An Earth tabu?"

Question 1

input goes here

Choices

○ **1.** input goes here

○ **2.** input goes here

○ **3.** input goes here

○ **4.** input goes here

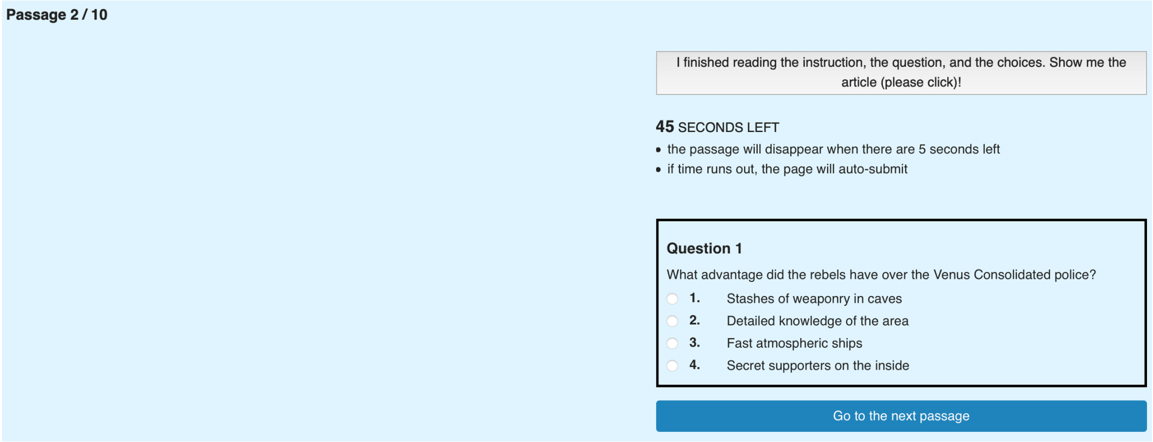
Question 2

input goes here

Choices

○ **1.** input goes here

Figure 4: The writing UI.



Passage 2 / 10

I finished reading the instruction, the question, and the choices. Show me the article (please click)!

45 SECONDS LEFT
- the passage will disappear when there are 5 seconds left
- if time runs out, the page will auto-submit

Question 1

What advantage did the rebels have over the Venus Consolidated police?

○ **1.** Stashes of weaponry in caves

○ **2.** Detailed knowledge of the area

○ **3.** Fast atmospheric ships

○ **4.** Secret supporters on the inside

Go to the next passage

Figure 5: The speed validation UI (part 1).

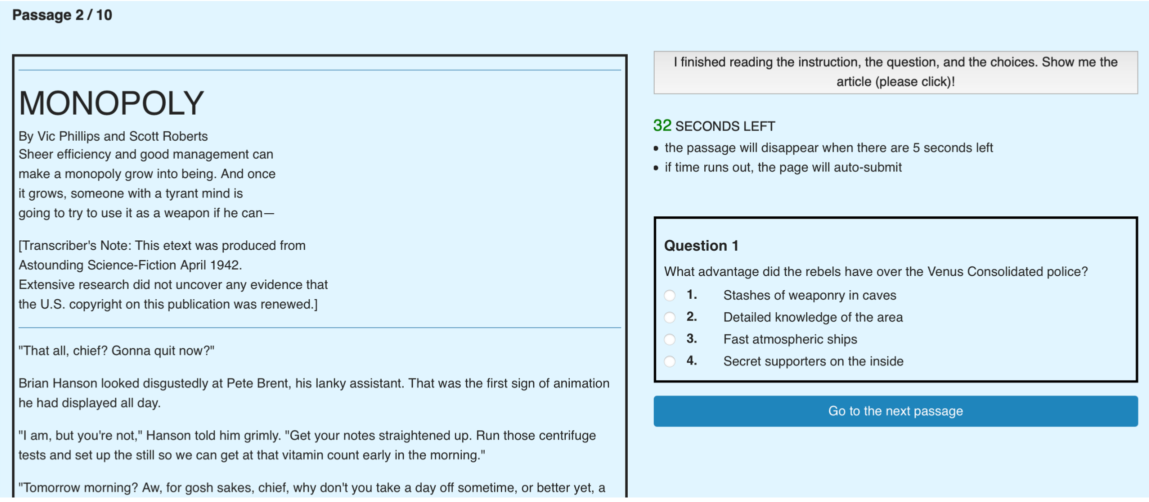
tators think is the most difficult distractor for each question (§2.3.2).

If writers have fewer than 40% of questions meet the above three bonus criteria *and* fewer than 75% of questions meet criteria (i) and (ii), we exclude them from future writing rounds: One writer was excluded after batch 1, and one writer was excluded after batch 2, for this reason. We also exclude two writers who missed deadlines by significant margins. Two other writers voluntarily left the project before finishing all three batches.

A.1.3 The Writing UI

Figure 4 shows our writing UI. A writer creates 10 multiple-choice questions with four answer options each on each page. Before the interview task and each batch of data collection, we explain our bonus structure to the writers. In order to encourage writers towards writing the types of questions that require understanding of the general context from the passage, we provide the following examples of themes questions can target in order to spur writers' creativity and provide suggestions if they have trouble coming up with difficult questions; however, they do not have to follow our suggestions.

- Characters' feelings and motivations
- Causes and consequences of described events
- Definitions, properties, and processes ex-



Passage 2 / 10

MONOPOLY

By Vic Phillips and Scott Roberts
Sheer efficiency and good management can make a monopoly grow into being. And once it grows, someone with a tyrant mind is going to try to use it as a weapon if he can—

[Transcriber's Note: This etext was produced from Astounding Science-Fiction April 1942. Extensive research did not uncover any evidence that the U.S. copyright on this publication was renewed.]

"That all, chief? Gonna quit now?"

Brian Hanson looked disgustedly at Pete Brent, his lanky assistant. That was the first sign of animation he had displayed all day.

"I am, but you're not," Hanson told him grimly. "Get your notes straightened up. Run those centrifuge tests and set up the still so we can get at that vitamin count early in the morning."

"Tomorrow morning? Aw, for gosh sakes, chief, why don't you take a day off sometime, or better yet, a

I finished reading the instruction, the question, and the choices. Show me the article (please click)!

32 SECONDS LEFT

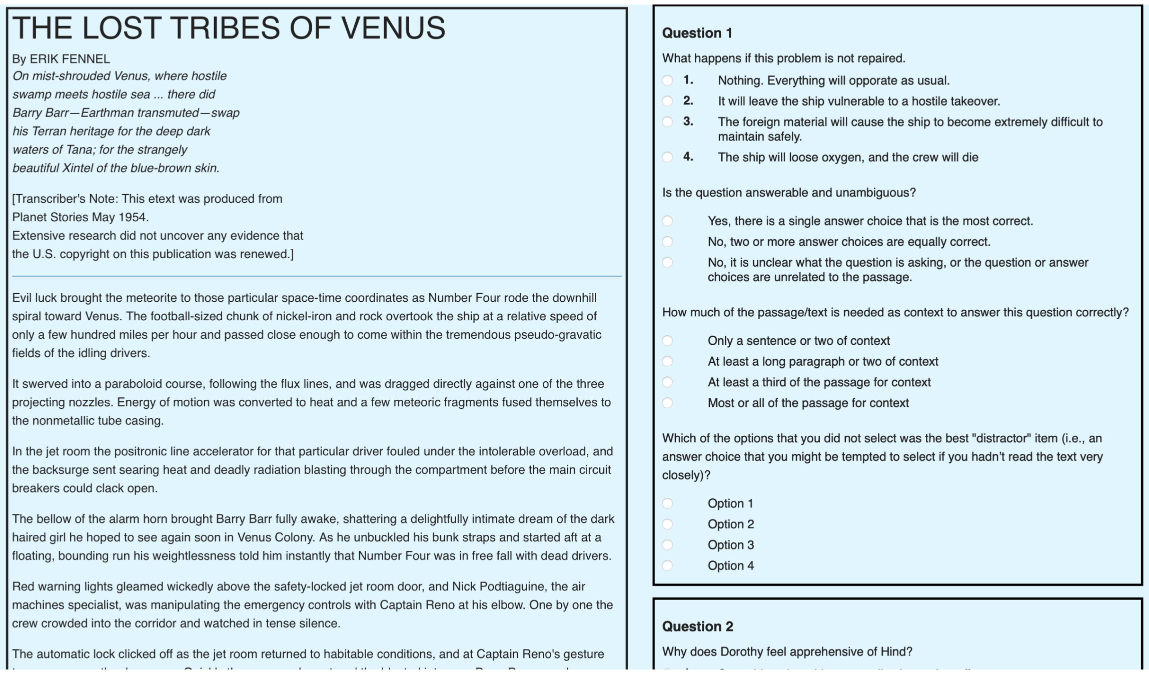
- the passage will disappear when there are 5 seconds left
- if time runs out, the page will auto-submit

Question 1

What advantage did the rebels have over the Venus Consolidated police?

○ **1.** Stashes of weaponry in caves
○ **2.** Detailed knowledge of the area
○ **3.** Fast atmospheric ships
○ **4.** Secret supporters on the inside

Figure 6: The speed validation UI (part 2).



THE LOST TRIBES OF VENUS

By ERIK FENNEL
On mist-shrouded Venus, where hostile swamp meets hostile sea ... there did Barry Barr—Earthman transmuted—swap his Terran heritage for the deep dark waters of Tana; for the strangely beautiful Xintel of the blue-brown skin.

[Transcriber's Note: This etext was produced from Planet Stories May 1954. Extensive research did not uncover any evidence that the U.S. copyright on this publication was renewed.]

Evil luck brought the meteorite to those particular space-time coordinates as Number Four rode the downhill spiral toward Venus. The football-sized chunk of nickel-iron and rock overtook the ship at a relative speed of only a few hundred miles per hour and passed close enough to come within the tremendous pseudo-gravatic fields of the idling drivers.

It swerved into a paraboloid course, following the flux lines, and was dragged directly against one of the three projecting nozzles. Energy of motion was converted to heat and a few meteoric fragments fused themselves to the nonmetallic tube casing.

In the jet room the positronic line accelerator for that particular driver fouled under the intolerable overload, and the backsurge sent searing heat and deadly radiation blasting through the compartment before the main circuit breakers could clack open.

The bellow of the alarm horn brought Barry Barr fully awake, shattering a delightfully intimate dream of the dark haired girl he hoped to see again soon in Venus Colony. As he unbuckled his bunk straps and started aft at a floating, bounding run his weightlessness told him instantly that Number Four was in free fall with dead drivers.

Red warning lights gleamed wickedly above the safety-locked jet room door, and Nick Podtiaguine, the air machines specialist, was manipulating the emergency controls with Captain Reno at his elbow. One by one the crew crowded into the corridor and watched in tense silence.

The automatic lock clicked off as the jet room returned to habitable conditions, and at Captain Reno's gesture

Question 1

What happens if this problem is not repaired.

○ **1.** Nothing. Everything will opporate as usual.
○ **2.** It will leave the ship vulnerable to a hostile takeover.
○ **3.** The foreign material will cause the ship to become extremely difficult to maintain safely.
○ **4.** The ship will loose oxygen, and the crew will die

Is the question answerable and unambiguous?

○ Yes, there is a single answer choice that is the most correct.
○ No, two or more answer choices are equally correct.
○ No, it is unclear what the question is asking, or the question or answer choices are unrelated to the passage.

How much of the passage/text is needed as context to answer this question correctly?

○ Only a sentence or two of context
○ At least a long paragraph or two of context
○ At least a third of the passage for context
○ Most or all of the passage for context

Which of the options that you did not select was the best "distractor" item (i.e., an answer choice that you might be tempted to select if you hadn't read the text very closely)?

○ Option 1
○ Option 2
○ Option 3
○ Option 4

Question 2

Why does Dorothy feel apprehensive of Hind?

Figure 7: The (untimed) validation UI.

plained in a passage
- The summary and lesson of a passage
- What would have happened had a character made a different choice

We also allow writers to skip a given passage in case they find that they would be unable to write high-quality questions for that passage. Specifically, we tell writers the following.

> If a passage is too difficult to write questions for, you can skip the article by choosing another URL to work on. We recommend that you do this if: (1) The text is hard to read due to major formatting issues. (2) The text is very technical or relies on cultural knowledge that you're unfamiliar with. (3) You think the passage is much too boring. We ideally want you to write questions for passages you find interesting!

A.2 Validation

A.2.1 Annotator Recruitment

We recruit annotators via Amazon Mechanical Turk (MTurk). We use a qualification task to identify annotators with good reading comprehension skills. This task is open to all workers with more than 1000 tasks (HITs) accepted and a HIT accept rate of at least 98%. We pay $5 for completing the qualification task, plus a $5 bonus for passing it. The task consists of a ∼3000-word passage with 10 multiple choice questions written or reviewed by the authors, each with a series of evaluation questions asking about the quality of that question. Of the 10 questions, 2 are intentionally ambiguous[17] in order to test if workers can accurately identify poorer quality questions.

In order to pass the qualification, workers need to (i) get at least 6/7 or 7/8 of the unambiguous questions correct, (ii) correctly identify at least one of the two ambiguous questions as ambiguous/unanswerable, and (iii) correctly identify at least half of the unambiguous questions as unambiguous. A total of 148 crowdworkers completed the task, and 45 of them passed (30.4%). All workers who pass the qualification are invited to complete tasks as part of both the speed validation and the untimed validation.

A.2.2 Speed Validation

Catch Questions We expect accuracy in the speeded task to be fairly low, so we construct catch questions to ensure that workers are not randomly guessing without attempting to find the correct answer. These trials are written by the authors and are designed to be answerable with only 45 seconds of access to the passage. For example, a catch question may ask who spoke a quote, like "Who said 'You're a wizard Harry!'?", where a single ctrl+F search of the quote gives the annotator the answer. Another catch question may have four options, three of which are clearly improbable. We do not validate the catch questions for correctness in the untimed validation, and so we do not include them in the final dataset, but we release them as a supplemental file for reproducibility.

Payment For most of the task, we pay workers $2.25 per HIT and award a bonus of $0.20 for each correct answer. However, during the first of six rounds, we paid $2.00 per HIT with a $0.18 bonus for each correct answer. After asking workers for feedback about the task via a survey, we decided to increase the rate of pay because workers reported spending slightly longer on the task than we originally estimated.

Task Procedure Each MTurk task consists of 10 speed validation questions (with randomly chosen articles). In each task, once the annotator clicks into the page, they have unlimited time to read the question and the options, but the article is not shown (Figure 5). Then, the annotator clicks the button that says "I finished reading the instruction, the question, and the choices. Show me the article (please click)!" As soon as the annotator clicks the button, the countdown clock of 45 seconds starts, and the article appears (Figure 6). The annotator can make the choice and submit at any time.

When there are only 5 seconds left, the article hides itself. The annotator has 5 seconds to make the choice. If the time expires, the page auto-submits, and the choice "-1" (indicating that the annotator did not make a choice) is recorded, which we mark as incorrect.

Annotator Performance Individual annotators consistently scored well above the chance rate of 25% on the catch questions. In all cases where an annotator's accuracy fell below 50% in a round, they were removed from future rounds.[18] Average overall accuracy on the catch questions was 83.8%, indicating that most workers were able to develop a strategy for finding a correct answer when it could be found.

Accuracy on the questions written by Upwork writers was 48.2% overall, but annotators got better at this task over time, likely by developing new strategies to search for answers. Average accuracy was 39.5% in the first round, rising to 58.4% in the final round of data collection. When the majority of annotators (at least 3/5) are able to answer questions correctly in this setting, we exclude that question from the difficult subset.

A.2.3 Untimed Validation

Figure 7 shows the UI for untimed validation. As two writers each write 10 questions for the same article, there are 20 unique questions per article. Each validation UI page contains all 20 sets of questions, and each set of questions contains the reading comprehension question and the three additional evaluation prompts. Therefore, in total, there are 80 prompts on each page. The annotator has to complete all 80 before they can submit the page and complete the task.

[17] We later found that one question was *unintentionally* ambiguous; we do not use this question in assessing whether workers pass the qualification.

[18] Two annotators fell below this threshold, though in that round they had also performed below threshold in the untimed validation. No annotators needed to be removed *solely* based on performance in the speed validation task.

Annotator Performance When a label can be assigned through majority vote of just the annotators, individual annotator agreement with the majority vote answer is 89.7% for all items identified as answerable and unambiguous, and 91.2% for across all examples that make it into the clean version of the dataset after taking the original writer's label into consideration. Throughout the course of the study, workers need to maintain at least 75% accuracy each round to maintain the qualification and continue to the next round. In a few cases, we identify passages that are themselves ambiguous or especially difficult. In these cases, we do not use those passages in computing by-round accuracy for the annotators. We exclude a total of 11 workers throughout the course of data collection for low accuracy, most of them after the first or second round.

Data Reannotation During each untimed validation round, we keep track of the rate at which each worker agrees with the original writers' labels for each question in order to quickly identify cases where either (i) a worker has misunderstood the passage, or (ii) a worker is putting insufficient effort towards the task. For any tasks where the individual annotator disagrees with the writer's labels on at least 40% of questions, we automatically re-post that passage for reannotation and replace the data, with the assumption that the annotator may have misunderstood something crucial in the passage.[19] After all the annotations are complete, we calculate the gold label answer via majority vote of annotators plus the original writer's label, and assess individual annotator accuracy. If any worker is excluded in a round for low accuracy (i.e., below 75% accuracy), we discard all of their responses from that round, reannotate and replace their data, and re-calculate the gold label and accuracy scores.

B Data

In order to increase the diversity of genres we use as contexts, we attempted to include Switchboard conversations. However, after presenting just 12 such conversations to our writers, we decided to discard all the Switchboard questions because many writers informed us that it was very difficult to come up with difficult questions for the Switchboard conversations. The writers indicated they found the Switchboard articles more difficult because the conversations were relatively short and usually involve very simple everyday topics, without the kinds of plot twists that are more common in short stories or complex details that are more common in long-form articles.

C Reasoning Types

For the purposes of this analysis, we define reasoning type as the category that needs to be reasoned about in selecting the correct answer option (e.g., 'person' is usually a 'who' question and corresponds to answer options that are characters or people) or the type of strategy that must be used in answering (e.g., 'symbolism/interpretation' requires the reader to extrapolate from the context or identify something not stated in a passage, like its theme). We identify 15 categories of reasoning types to include in our analysis. These categories are initially inspired by those used in NarrativeQA, but we adapt them to our dataset, as we find that many questions do not fit their categorization. These categories are not mutually exclusive, and nearly a third of the questions are categorized as two or more types.

Reasoning Type Definitions The following includes definitions of all the categories used, along with at least one hand-selected example to demonstrate a question belonging to that category. All in-text examples are selected out of the training set.

- **Description**: The question relies on the reader reasoning about which description is correct. Often these questions are about describing a character's feelings ('How do Lowry and the Exec feel about the Venusians?') or point of view ('How is the book "Living a Normal Sex Life" seen by these people?'), describing a feature of the story ('What makes Grannie Annie's writing remarkable?'), or describing an individual ('Which word least describes Don?')
- **Why/reason**: The question relies on the reader reasoning about the cause or explanation for something in the story. Most of these questions begin with 'why' and ask about the cause of an event ('Why does the crew get off the ship with Moran?'), causes of characters' feelings ('Why does Ben take

[19] We identified 10 passages that, after multiple rounds of reannotation, were not passing this threshold. This may be due to one of the writers misunderstanding the passage and thereby creating several ambiguous questions, so for these 10 passage we chose to include *all* annotations collected.

Question	Answer Options	Reasoning Types
What would have happened if Click's camera broke in the crash?	(a) Irish would have died on impact. (b) They would have returned immediately to Luna Base. (c) They would have caught Gunther faster. (d) They would have continued to believe the monsters were real.	What if; Event
What isn't a reason for narrator to be so skeptical of Gorb?	(a) Gorb looked just like an Earthling (b) Gorb was asking for too much money (c) Gorb had no proof to back up his claims (d) he had never heard of Wazzenazz	Why/reason; Not/except
How many caves had Garmon and Rolf traveled through before their crash?	(a) thirty seven (b) forty seven (c) thirty (d) forty	Numeric
How do you think Meredith feels about the rest of the crew?	(a) She has a close bond of respect and (platonic) love for the rest of the members (b) She respects and loves one person the most (c) She's become friends with them slowly over time and appreciates them all (d) She respects one person the most and loves another person the most	Description; Symbolism/ interpretation; Relation
The less you share...	(a) ...the more privacy you have. (b) ...the more your intellectual property is protected. (c) ...the less power you have. (d) ...the less your cultural goods will be appropriated.	Symbolism/ interpretation; Finish the phrase
How did Meryl Streep prepare for the role of Roberta?	(a) She learned to play the violin without any former instrument training. (b) She began to act very helplessly and feeble around the rest of the cast. (c) She is a method actor and became very vulnerable. (d) She made herself look dumpy and thick-waisted.	How/method

Table 7: Full examples of the annotations from our analysis of reasoning types on a subset of questions from QuALITY. Examples are taken from analyzed examples from the training set. Examples are selected non-randomly and are intended to demonstrate a range of reasoning types observed.

offence to Cobb's comments about spacemen?'), or characters' internal motivations ('Why does Joseph lie about the water supply?'), though other questions formulate this differently while still asking for the underlying reason ('What makes Gubelin an outlier in the present day?' and What is the purpose of a comanalysis?).

- **Symbolism/interpretation**: The question relies on the reader making an interpretation that goes beyond what's explicitly said in the story, or it asks about symbolism or themes from the story. Many questions explicitly ask the reader to interpret what message the author was trying to convey ('What point is being made by comparing Fight Club to the UFC?') or what tone the story takes ('What is the tone like throughout the story?'). Other questions require the reader to predict what will happen next ('What will happen next to Jery?') or ask about the use of literary cues like irony ('What is ironic about Earth's customer service policy?').

- **How/method**: The question relies on reasoning about how something happened or the method that was used. Most of these questions rely on the question word 'how' to ask about a process ('How did Meryl Streep prepare for the role of Roberta?'), the manner in which something happens ('How did Templin find about about Pendleton's death?'), or a method by which something happens ('How does the shape of Starre's ship benefit them?').

- **Event**: The question relies on reasoning about an event, or asks for an event as the answer option. The majority of these questions focus on what someone did/plans to do ('What did Joe and Glmpauszn plan to do?') or what happened to someone ('What happened to Morgan Brockman by the end of the passage?').

- **Person**: The question relies on reasoning about which person or people are involved. Most ask about a specific person ('Who is Owen Fiss and what did he do?'), though many questions of this type still require reasoning about the entire passage to answer ('Who seems to have the least to hide in the text?').

- **Not/except**: The question requires the reader to select the answer option that *least* answers the question, flipping the typical way a multiple-choice task is performed. All of these questions use some word to indicate this flipping, such as 'least' ('Which word least describes McGill?'), 'not' ('What word doesn't describe the natives from Tunpesh?'), or 'except' ('Dole makes all of the following charges against the New York Times EXCEPT for: with the NYT?').

- **Relation**: The question relies on reasoning about the relationship between two or more characters, as in 'Who is Sporr and what is his authority in calling the narrator Yandro?' or questions that ask about how one character feels about another ('How does Jakdane feel about Trella?').

- **Entity**: The question relies on reasoning about a non-human entity or a group, as in 'We can assume that Saladin's army represents which group?'.

- **Finish the phrase**: The form of the question requires either a fill-in-the-blank style response or is a partial phrase that must be completed by selecting the correct answer option. Often, these questions do not include an explicit question word. Some of them have a blank written in ('The film reviewer is generally _____ the actors in "Princess Mononoke," and _______ the actors in "The Limey," respectively:') and others are just a partial sentence ('The less you share...').

- **Location**: The question relies on reasoning about a place, as in 'What city is Temple-Tracy in?'.

- **Numeric**: The question relies on finding or computing the correct numeric option, as in 'How many caves had Garmon and Rolf traveled through before their crash?'.

- **Object**: The question relies on reasoning about an object, as in 'What does Captain Hannah use as an organic processor?'.

- **What if**: The question requires the reader to make an inference about what *would have been true* if some fact from the story were changed, and most of these questions explicitly set up the counterfactual scenario ('What would have happened if the Peace State had not crash landed?').

- **Duration**: The question relies on reasoning about how long something happened for or how much time passed between two events, as in 'How long did Maggie care for Ben before he finally awoke after rescuing him?'.

Annotation Details Three authors of this paper analyze a set of 500 randomly selected questions. One author annotates all 500, and the other two annotators analyze 250 each, such that each example is annotated by two unique individuals. Following annotation, the authors discuss any disagreements and adjust their original coding once consensus is reached. Using this consensus approach allows for clarification of the categories during and after annotation, which leads to an internally consistent coding scheme.

Sample Annotations Table 7 shows a set of representative example annotations from this analysis, demonstrating several sentences that were categorized as more than one reasoning type.

D More Details on Analysis

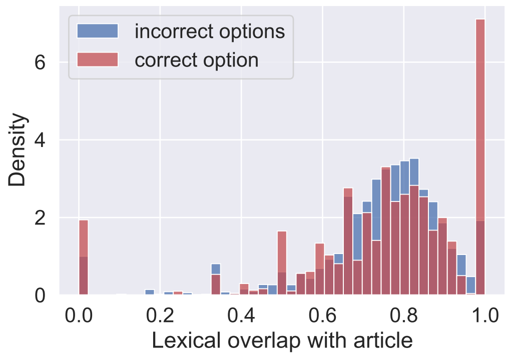


Figure 8: Lexical overlap of all the correct and incorrect options with the article. The distribution is normalized since there are thrice as many incorrect options as there are correct options.

Lexical Overlap In addition to comparing lexical overlap of the correct option and the maximum lexical overlap of the incorrect option with the article (Section 3.4), we also plot a normalized distribution of lexical overlap for all the correct and incorrect options in Figure 8. Despite a higher fraction of the correct options having complete overlap with the article, models would not be able to exploit this heuristic, since other incorrect options for the same question may have complete overlap. This is demonstrated by the plot in Figure 3 and the fact that a baseline which relies on the lexical overlap heuristic only achieves 26.6% accuracy.

Training Data	Model	Full	Extraction Based on Qs			Question-Only
			R-1	fastText	DPR	
QuALITY	Longformer-base	33.7/32.6	–	–	–	–
	RoBERTa-base	–	33.7/32.0	39.2/37.2	40.0/36.4	36.0/35.6
	RoBERTa-large	–	30.0/28.7	42.7/38.3	26.7/24.0	26.7/23.0
	DeBERTaV3-base	–	34.7/33.0	38.0/35.3	41.8/37.4	36.9/34.3
	DeBERTaV3-large	–	44.0/37.6	44.3/37.9	45.1/39.2	38.1/33.7
	T5-base	–	27.3/26.6	27.6/25.5	28.3/29.4	27.9/28.7
RACE	Longformer-base	34.5/31.6	–	–	–	–
	RoBERTa-base	–	43.7/38.2	43.3/38.9	44.1/37.3	36.8/34.6
	RoBERTa-large	–	48.6/41.7	48.4/42.3	50.9/45.3	37.2/35.3
	DeBERTaV3-base	–	46.5/38.7	44.8/37.9	48.8/41.7	35.8/31.6
	DeBERTaV3-large	–	51.2/43.9	50.5/43.8	53.5/47.3	38.3/34.3
	T5-base	–	39.0/37.7	39.7/39.2	39.9/38.5	37.2/35.6
RACE ↓ QuALITY	Longformer-base	38.1/32.8	–	–	–	–
	RoBERTa-base	–	43.7/38.2	41.7/36.2	43.8/37.2	37.4/36.6
	RoBERTa-large	–	47.7/42.5	46.8/43.1	50.8/46.2	39.1/37.8
	DeBERTaV3-base	–	45.5/40.0	46.6/40.1	46.7/40.9	39.6/35.2
	DeBERTaV3-large	–	51.7/44.7	50.7/43.3	53.6/47.4	41.4/39.2

Table 8: Accuracy on QuALITY development set (full/difficult).

E More Details on Modeling

E.1 Extraction

For ROUGE-1 scoring, we use the `rouge-score` Python package.[20]

For fastText scoring, we use SpaCy with the `en_core_web_sm` model for tokenization, and use embeddings trained on Common Crawl, [21] using the top 300k words in the vocabulary.

For DPR, we use the implementation in the Transformers package (Wolf et al., 2020), using the `facebook/dpr-ctx_encoder-multiset-base` and `facebook/dpr-question_encoder-multiset-base` models for encoding the context and query respectively.

E.2 Training

For training on QuALITY, we train for 20 epochs. For RoBERTa and DeBERTaV3, we use a maximum sequence length of 512. For Longformer, we train use the maximum sequence length of 4,096.

For training on RACE, we train for 3 epochs with a maximum sequence length of 512 (including Longformer). The same models used for supplementary training are used for zero-shot RACE evaluation on QuALITY.

All models are trained with a learning rate of 1e-5 and warm-up ratio of 0.1. RoBERTa and DeBERTaV3 are trained with batch size 16, while Longformer is trained with batch size 4. All models are trained with the AdamW optimizer.

E.3 Results

Table 8 shows the results on development set. Table 9 shows the results using oracle-answer-based extraction. Please refer to the discussion in Section 4.2.

[20] `https://github.com/google-research/google-research/tree/master/rouge`

[21] `https://dl.fbaipublicfiles.com/fasttext/vectors-english/crawl-300d-2M.vec.zip`

		Extraction Based on *Oracle Answers*		
Training Data	Model	R-1	fastText	DPR
QuALITY	RoBERTa-base	69.1/67.3	61.3/57.2	78.3/77.7
	RoBERTa-large	67.5/64.2	63.9/60.0	75.3/73.7
	DeBERTaV3-base	70.8/68.5	65.5/60.8	76.4/74.9
	DeBERTaV3-large	68.9/65.2	66.6/60.8	77.8/75.1
RACE	RoBERTa-base	54.5/49.9	56.1/49.7	53.4/47.9
	RoBERTa-large	59.2/53.7	58.2/52.1	56.9/51.5
	DeBERTaV3-base	56.5/51.5	55.9/48.7	52.0/45.0
	DeBERTaV3-large	59.8/54.1	59.8/53.5	57.5/49.6
RACE	RoBERTa-base	67.6/62.8	64.6/58.8	70.2/67.2
↓	RoBERTa-large	68.9/63.7	66.6/60.5	64.1/60.0
QuALITY	DeBERTaV3-base	69.6/64.4	68.1/62.5	66.9/61.2
	DeBERTaV3-large	71.0/66.5	68.2/62.9	71.9/67.1

Table 9: Oracle accuracy on the full QuALITY development set and on the QuALITY-HARD subset (full/hard) with models using the *correct answers* as queries to retrieve relevant excerpts. These results are meant to show the relative contribution of the retrieval and reading components of the two-stage models. Caution: These results rely on answers at test time, which are not available to any model during a conventional deployment or test set evaluation, and so are of very limited value in conventional comparisons.