What Makes Reading Comprehension Questions Difficult? Investigating Variation in Passage Sources and Question Types

Anonymous ACL submission

Abstract

In order for a natural language understanding benchmark to be useful in research, it has to consist of examples that are diverse and difficult enough to discriminate among current and near-future state-of-the-art systems. However, we do not yet know what kinds of passages and their sources help us collect a variety of challenging examples. In this study, we crowdsource multiple-choice reading comprehension questions for passages taken from seven qualitatively distinct sources, analyzing what attributes of passages contribute to the difficulty and question types of the collected examples. We find that passage source, length, and readability measures do not significantly affect question difficulty. Among seven question types we manually annotate, questions that require numerical reasoning and logical reasoning are relatively difficult but their frequencies depend on the passage sources. These results suggest that when creating a new benchmark dataset, we do not have to use difficult passages but select passage sources carefully so that it has questions that involve linguistic phenomena we are interested in.

1 Introduction

State-of-the-art systems have shown performance comparable with humans on many recent natural language understanding (NLU) datasets (Devlin et al., 2019; Sun et al., 2021), suggesting that these benchmarks will no longer be able to measure future progress. To move beyond this, we will need to find better ways of building difficult datasets, ideally without sacrificing diversity or coverage (Bowman and Dahl, 2021). To obtain such human-written examples at scale, there are active lines of crowdsourcing research on protocols of worker handling and feedback (Nangia et al., 2021) and the design of collection task (Ning et al., 2020; Rogers et al., 2020). However, we do not have clear information on what aspects of *text sources* affect the difficulty and diversity of examples.

MCTest: Tony walked home from school on his birthday. He was surprised to see a lot of cars in front of his house. When he opened the door and entered the house, he heard a lot of people yell, "Surprise!" It was a surprise party for his birthday. His parents called all his friends' parents and invited them to come to a party for Tony. [...]

Q: *Who were invited to the party and by who?*

□ *Tony's parents invited only his friends*

□ *Tony invited his friends and their parents*

□ *Tony's parents invited his friends' parents*

☑ *Tony's parents invited his friends and their parents*

ReClor: Humanitarian considerations aside, sheer economics dictates that country X should institute, as country Y has done, a nationwide system of air and ground transportation for conveying seriously injured persons to specialized trauma centers. Timely access to the kind of medical care that only specialized centers can provide could save the lives of many people. [...]

Q: *What is the economic argument supporting the idea of a transportation system across the nation of Country X?*

□ *Building the transportation system creates a substantial increase of jobs for the locals*

☑ *Increasing access to specialized medical centers can lower the chance of the workforce population dying*

□ *Transportation ticket prices directly contribute to the government's revenue*

□ *Country Y was successful with their attempts to potentially save lives so Country X should try it as well*

Figure 1: Example questions for passages from simple narratives (MCTest) and technical arguments (ReClor).

Crowdsourced datasets in reading comprehension use passages taken from a variety of sources such as news articles, exams, and blogs about which questions are written (Lai et al., 2017; Trischler et al., 2017; Rogers et al., 2020). The first example in Figure 1 is from MCTest (Richardson et al., 2013), whose passages are written in grade-school-level English. The second example is from ReClor (Yu et al., 2020), which consists of passages and questions written for graduate and law school admission examinations. We hypothesize that difficult passages such as in the second example are suitable for crowdsourcing challenging questions. Passages that are linguistically complex and have dense information could help facilitate writing questions that require a wide range of linguistic and world knowledge, following intricate events, and comprehending logical arguments. In contrast, easy passages as in children's stories likely talk about common situations and simple facts, which might prevent workers from writing such difficult questions.

In this work, we crowdsource multiple-choice reading comprehension questions to analyze how question difficulty and types are affected by the choice of source passage. Using passages extracted from seven different sources, we ask crowdworkers to write questions and four answer options about the given passages. We compute the difference between human and machine accuracy and use it as the difficulty of questions, investigating whether there is a correlation between the difficulty and linguistic aspects of passages such as their source, length, and readability measures.

In addition to a standard setting where we directly accept crowdworkers' submissions, we use an adversarial setting where they have to write questions that fool a strong reading comprehension model (Bartolo et al., 2020; Kiela et al., 2021). Although Kaushik et al. (2021) find questions that require numerical reasoning frequently appear in the adversarial data collection of the extractive QA task on Wikipedia articles, our aim is to see whether we observe a similar trend in multiple-choice questions written for different passage sources or if the adversarial setting is useful for collecting various types of questions there.

We find that the difficulty of collected questions does not significantly correlate with the differences of passages in linguistic aspects such as passage source, passage length, Flesch–Kincaid grade level (Kincaid et al., 1975), syntactic and lexical surprisal, elapsed time for answering, and the average word frequency in a passage. In contrast, we find that elapsed time for writing correlates with the question difficulty, though only weakly. Our main positive finding comes through our manual annotation of the types of reasoning that each question targets, where we observe that questions that require numerical reasoning and logical reasoning are relatively difficult and that their frequencies depend on the passage sources (e.g., numerical reasoning is found more often in MCTest and logical reasoning is in ReClor). These results suggest that when creating a new benchmark dataset or choosing one for evaluating NLU systems, choosing a diverse set of passages can help ensure a diverse range of question types, but that *difficulty* in passages need not be a priority. Our collected datasets could be useful for training reading comprehension models and further analysis of requisite knowledge and comprehension types in answering challenging multiple-choice questions.[1]

2 Related Work

Crowdsourcing NLU Datasets Crowdsourcing has been widely used to collect human-written examples at scale (Rajpurkar et al., 2016; Trischler et al., 2017). Crowdworkers are usually asked to write questions about a given text, sometimes with constraints imposed to obtain questions that require specific reasoning skills such as multi-hop reasoning (Yang et al., 2018) and understanding of temporal order, coreference, and causality (Rogers et al., 2020). In this work, to analyze examples naturally written by workers, we do not consider specific constraints on questions and answer options.

Current benchmark datasets constructed by crowdsourcing may not be of quality enough to precisely evaluate human-level NLU. For example, Jia and Liang (2017) point out that then-state-of-the-art models in SQuAD (Rajpurkar et al., 2016) are easily fooled by manually injected distracting sentences. Chen and Durrett (2019) and Min et al. (2019) show that questions in multi-hop reasoning datasets such as HotpotQA by Yang et al. (2018) do not necessarily require multi-hop reasoning across multiple paragraphs. Kaushik and Lipton (2018) find that baseline models with question-only and passage-only input often perform comparably well to full-input models in widely-used datasets.

To investigate how to collect high-quality, challenging questions in crowdsourcing, Nangia et al. (2021) compare different sourcing protocols and find that training workers and giving feedback about their submissions improve the difficulty and quality of questions in reading comprehension. To encourage workers to write difficult examples, Bartolo et al. (2020) propose to collect questions using a model-in-the-loop setting, where the requesters only accept workers' written questions that fool a strong reading comprehension model. Although this adversarial approach enables us to collect challenging questions efficiently, Gardner et al. (2020) point out that collected examples might be biased towards quirks of the adversary models. Bowman

[1] We will make our datasets, annotation instructions and results, and crowdsourcing scripts publicly available.

and Dahl (2021) extend this argument, and point out that adversarial methods can systematically eliminate coverage of some phenomena. This is also supported by Kaushik et al. (2021), but their findings are limited to the extractive QA setting for Wikipedia articles. Our motivation is to see if this argument is applicable to the multiple-choice format with a wide range of passage sources for which we expect crowdworkers to write linguistically diverse questions and answer options.

Sources of NLU Datasets Reading comprehension datasets are often constructed with a limited number of passage sources. Rajpurkar et al. (2016) sample about five hundred articles from the top 10,000 articles in PageRank of Wikipedia. Similarly, Dua et al. (2019) curate passages from Wikipedia articles containing numeric values to collect questions for mathematical and symbolic reasoning. Khashabi et al. (2018) construct a dataset in which questions are written for various passage sources such as news articles, science textbooks, and narratives. However, because their dataset is designed so that questions require multi-sentence reasoning (mainly about coreference resolution), we cannot use it for analyzing the variation of question types in general.

In a similar vein to our work, Sugawara et al. (2017) find that readability metrics and question difficulty do not correlate in reading comprehension datasets. Our study differs in the following two points which may cause different findings: First, their observational study of existing datasets has fundamental confounds because questions they examine are constructed by different sourcing methods (e.g., automatic generation, expert writing, and crowdsourcing), which could have an impact on the question difficulty. We aim to investigate uniformly crowdsourced examples across seven different sources to get insights for future data construction research using crowdsourcing. Second, they define question difficulty using human annotations alone, but it does not necessarily reflect the difficulty for current state-of-the-art models. In this study, we define the question difficulty as the human–machine performance gap using ten recent strong models, which enables more fine-grained analysis on the collected questions for a better benchmark of current models.

We adopt the multiple-choice format because, as Huang et al. (2019) discuss, it allows us to evaluate the human and machine performance easily.

3 Crowdsourcing Tasks

This study aims to analyze what kinds of passages make reading comprehension questions difficult in crowdsourcing. We use Amazon Mechanical Turk to access a large pool of workers. To collect difficult and quality examples, we require crowdworkers to take a qualification test for accepting our question writing tasks and validation tasks.

3.1 Worker Qualification

The qualification test has two parts, which we run in separate tasks: question answering and writing. To join the qualification test, workers have to meet the following minimum qualification: based in the United States, Canada, or United Kingdom, having an approval rate of at least 98%, and having at least 1,000 approved tasks.

The question answering task is used to identify workers who carefully answer reading comprehension questions. A single question answering task has five questions that are randomly sampled from the validation set of ReClor in which most of the questions are taken from actual exams. Those who correctly answer at least four out of five questions proceed to the next qualification phase.

The question writing task is used to familiarize workers with the writing task and select those who can carefully write multiple-choice reading comprehension questions. We ask workers to write two questions given two different passages randomly sampled from the validation set of RACE (Lai et al., 2017). This dataset consists of self-contained passages that are written for middle- and high-school exams in various subjects where we expect the passages to enable workers to write questions easily. Following Nangia et al. (2021), we then review workers' submissions and grade them using a rubric with four criteria: a question (1) is answerable without ambiguity (*yes* or *no*), (2) requires reading the whole passage (five-point scale), (3) is creative and non-obvious (five-point scale), and (4) has distractor answers that could look correct to someone who has not read carefully (*more than one*, *one*, or *no*). We rank workers using this rubric and allow about the top 50% workers to proceed to the main writing task. We make sure that these workers write two unambiguous and answerable questions.

3.2 Writing Task

In the main writing task, a worker is shown a single passage and asked to write a question about it along with four answer options, one of which is the correct answer. We provide instructions where we describe that questions have to be challenging but still answerable and unambiguous for humans, and we include good and bad examples.

Each worker who passes the qualification round is randomly assigned to either standard or adversarial data collection. In the standard collection, we accept workers' submissions without any filtering. In the adversarial collection, a written question is sent to a reading comprehension model immediately. If the model cannot answer that question correctly, we accept it. We allow workers to submit questions (i.e., get paid) after three attempts even if they keep failing to fool the model. We use UnifiedQA 3B v2 (Khashabi et al., 2020) for the adversary model, which is a T5-based transformer (Raffel et al., 2020) trained on a wide variety of question answering datasets such as MCTest, RACE, NarrativeQA (Kočiský et al., 2018), and SQuAD. While the source of training data that we use in our models will inevitably influence our findings, focusing on a model with very diverse pretraining and fine-tuning will minimize this effect.

Passage Sources We use passages from the following seven sources: (1) MCTest children's narratives, (2) Project Gutenberg narratives, (3) Slate online magazine articles from the 1990s sourced from the Open American National Corpus (Ide and Suderman, 2006), (4) middle- and high-school exams from RACE, (5) graduate-level exams from ReClor, and (6) science and (7) arts articles from Wikipedia. Details for Project Gutenberg and Wikipedia articles are presented in Appendix A. We use passages of the training sets of MCTest, RACE, and ReClor. For the other sources, we split available books and articles into passages. In the writing task, a passage is randomly taken from a passage pool in which there are the same number of passages extracted from each source.

3.3 Validation Task

We collect the votes of five workers for each of the collected questions. Those workers who passed the question answering task of the qualification round can accept the validation tasks. To incentivize workers, we include gold-labeled examples (Nangia et al., 2021) into the tasks (about 10% of the questions) and pay a bonus of $0.50 USD if a worker can answer those questions correctly at least 80% of the time. If a worker fails to answer them at least 60% of the time, we disqualify the worker from future rounds of data collection.

Worker Pay and Logistics For the writing tasks, the base pay is $2.00 per question, which we estimate to be about $15.00 per hour based on measurements from our pilot runs. If a worker succeeds in fooling the model in adversarial data collection, the worker gets an additional bonus of $1.00. For validation, a single task consisting of five questions pays $2.00, which we estimate to be about $15.00 per hour as well.

4 Crowdsourcing Results

4.1 Dataset Construction

We collected a total of 4,340 questions, with 620 in each of the seven sources, further divided into 310 each for the standard and adversarial methods. Each passage is paired with only one question. We randomly sample two out of five validation votes for validating collected examples and the remaining three votes for measuring human performance. In the validation, we regard a question as valid if at least one out of the two votes is the same as the writer's gold answer. If both votes are unanimously the same as the gold answer, the question is regarded as a high-agreement example. We find that 90.3% of collected questions are valid (92.0% for standard collection and 88.7% for adversarial collection). In addition, 65.7% of the collected questions are high-agreement (68.7% and 62.7% for standard and adversarial collection, respectively). We present the dataset and worker statistics in Appendices B and C.

4.2 Human Performance

Table 1 provides human and model performance. Because the questions are validated using two out of five human votes in the validation step above, we take an average accuracy of the remaining three votes (instead of taking the majority vote) to measure human performance. We observe 4.5% and 3.9% gaps between the standard and adversarial collection in the valid and high-agreement questions respectively.

4.3 Machine Performance

To establish a model performance that is not biased towards a single model, we compute the average accuracy (*M-avg.*) of ten different models from the following three classes: UnifiedQA large and 3B (v2, zero-shot; Khashabi et al., 2020), RoBERTa

		All valid examples					High-agreement portion				
Source	Method	Human	UniQA	DeBERTa	M-Avg.	Δ	Human	UniQA	DeBERTa	M-Avg.	Δ
MCTest	Dir.	84.6	68.3	84.5	74.2	10.4	90.6	71.5	88.2	77.4	**13.3**
	Adv.	85.0	26.5	75.3	58.9	26.1	88.1	27.9	78.6	60.5	27.6
	Total	84.8	47.4	79.9	66.6	**18.3**	89.3	49.3	83.3	68.8	**20.6**
Gutenberg	Dir.	81.9	70.7	84.5	76.7	5.1	88.9	75.0	88.5	80.3	8.7
	Adv.	77.5	26.4	80.1	61.4	16.1	81.9	28.3	82.6	64.3	17.5
	Total	79.7	48.8	82.3	69.2	10.5	85.6	53.1	85.7	72.8	12.8
Slate	Dir.	82.9	72.4	88.9	80.4	2.5	88.6	74.6	91.7	83.1	5.5
	Adv.	77.5	26.0	71.7	61.4	16.1	85.6	27.9	76.0	65.6	20.0
	Total	80.3	49.8	80.5	71.2	9.1	87.2	52.6	84.3	74.8	12.4
RACE	Dir.	85.9	70.4	85.0	77.2	8.8	91.0	74.8	90.4	80.9	10.1
	Adv.	84.4	28.9	69.4	58.1	**26.3**	90.5	31.0	73.8	60.0	**30.4**
	Total	85.2	50.0	77.3	67.8	17.4	90.7	53.3	82.2	70.7	20.1
ReClor	Dir.	87.8	72.6	88.5	77.2	**10.7**	90.5	79.6	91.1	81.5	9.1
	Adv.	78.3	29.2	71.5	59.0	19.3	82.4	32.4	74.5	63.6	18.9
	Total	83.2	51.7	80.4	68.5	14.8	86.8	58.1	83.5	73.3	13.5
Wiki. Sci.	Dir.	83.0	75.9	90.6	80.5	2.5	88.3	79.0	94.9	84.2	4.1
	Adv.	78.5	27.4	75.2	58.6	19.9	85.7	29.4	77.2	60.9	24.8
	Total	80.8	52.1	83.0	69.8	11.0	87.1	56.3	86.8	73.6	13.6
Wiki. Arts	Dir.	83.0	76.2	88.7	80.8	2.2	86.8	77.0	92.5	84.3	2.6
	Adv.	76.7	25.5	73.8	61.2	15.6	83.7	25.8	75.8	63.1	20.6
	Total	79.9	51.2	81.3	71.1	8.8	85.3	52.3	84.5	74.1	11.3
All sources	Dir.	84.2	72.4	87.2	78.1	6.0	89.3	75.9	91.0	81.7	7.6
	Adv.	79.7	27.1	73.8	59.8	19.9	85.4	29.0	76.9	62.6	22.8
	Total	82.0	50.2	80.7	69.2	12.8	87.5	53.6	84.3	72.6	14.9

Table 1: Accuracy of humans and models and their difference (Δ) between human accuracy and the average zero-shot performance of ten different models (*M-avg*) for all valid questions and the high-agreement portion of them. The highest and lowest gaps are highlighted in bold and underline. The questions are crowdsourced with (*Adv*) and without (*Dir*) adversarial feedback. *UniQA* is the zero-shot performance by the UnifiedQA 3B model and used in the adversarial data collection. *DeBERTa* is the performance by the xlarge model fine-tuned on RACE.

large (four models with different random seeds; Liu et al., 2019), and DeBERTa large and xlarge (v2, either fine-tuned on MNLI (Williams et al., 2018) first or not; He et al., 2020).

The RoBERTa and DeBERTa models are all fine-tuned on RACE. Among these models, DeBERTa xlarge (MNLI-fine-tuned) performs best on RACE, achieving 86.8% accuracy. Because UnifiedQA 3B (72.3% on RACE) is used in the adversarial data collection, it shows lower accuracies in the adversarially collected questions. We show the performance of these two models for comparison in Table 1. Except where noted, we do not train the models on any portion of our collected questions.

Supervised Performance For each dataset, we evaluate the performance of DeBERTa large trained on the datasets other than the target dataset in a leave-one-out manner. Our motivation is to see whether the accuracy values significantly improve by training (i.e., the human–model gaps decrease). If there is a large gain, it implies that the datasets have simple patterns among examples that the models can exploit. The result shows no significant gains in the adversarial datasets, while the standard datasets show some small gains (see Appendix D).

Partial-Input Performance As Kaushik and Lipton (2018) point out, reading comprehension datasets might have annotation artifacts that enable models to answer questions without passages or question sentences. To investigate such artifacts in our collected examples, we evaluate model performance with the ablation of questions (*P+A*), passages (*Q+A*), and both questions and passages (*A only*). We see large drops in the zero-shot performance by DeBERTa xlarge. In addition, we do not observe a big performance improvement in the supervised performance by DeBERTa large (MNLI-fine-tuned) as well. These results demonstrate that the collected questions and answer options do not have severe annotation artifacts in any passage sources (see Appendix E).

4.4 Human–Model Performance Gap

We compute the human–model performance gap (Δ) between the human and the average model accuracies. There is not a large variation in the gap across passage sources in the high-agreement portion ($\Delta = 14.9 \pm 3.5$). We observe the highest human performance for RACE questions and the lowest for Wikipedia arts, while seeing the highest model performance for MCTest and the lowest for Slate. Surprisingly, the questions sourced from MCTest, which consists of easy narrative passages, show the largest gap in the standard method and the total of both methods. The RACE questions give the largest gaps in the adversarial method, while MCTest shows the second-largest. Although ReClor consists of passages for graduate-level exams, it shows smaller gaps than RACE, which consists of passages for middle- and high-school English exams. Gutenberg passages are written for adults, but we do not observe that the examples written for those passages show larger gaps than the examples for MCTest. Rather, Gutenberg shows the lowest gaps among the adversarial questions. These observations are inconsistent with our initial hypothesis.

5 Linguistic Analysis

We analyze how linguistic aspects of the collected examples correlate with the human–model performance gap computed in the experiments. To get a better estimate of human performance, we use high-agreement examples (Nie et al., 2020). For ease of comparison, we split these examples into two subsets: easy ($\Delta \leq 20\%$) and hard ($\Delta \geq 40\%$). These subsets have 2,023 and 596 examples respectively.[2]

5.1 Readability Measures

We compute the correlation between the human–model performance gap and readability measures across all valid examples (Pearson's r and p-value) and independence between the distributions of the easy and hard subsets about the measures (p-value in Welch's t-test, which is a paired t-test where we do not know if two distributions have the same variance). We plot the density distributions of the easy and hard subsets here, while Appendix provides the instance-wise plots of all valid examples.

Passage Length We report the number of words (except for punctuation) as the passage length.[3] Across all examples, we observe Pearson's $r = 0.01$ ($p = 0.41$) (the full plot is in Appendix G). The t-test shows $p = 0.49$. Therefore, there does not appear to be any relationship between passage length and question difficulty.

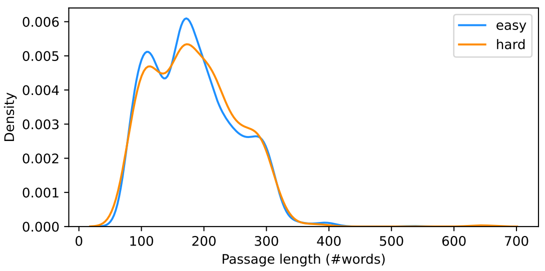


Figure 2: Passage length (number of words) of easy and hard examples.

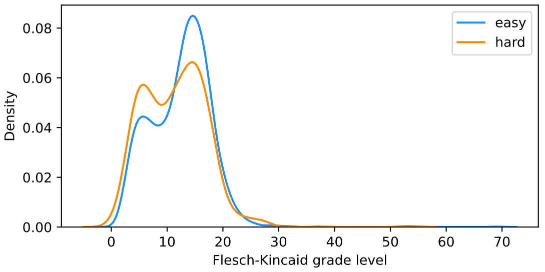


Figure 3: Flesch–Kincaid grade level of easy and hard examples.

Flesch–Kincaid Grade Level We use the Flesch–Kincaid grade level (Kincaid et al., 1975) as a basic metric of text readability. This metric defines readability based on approximate US grade level with no upper bound (higher is more difficult to read). It is computed for a passage using the average number of words that appear in a sentence and the average number of syllables in a word (See Appendix I for details). The correlation between the grade and human–model performance gap is $r = -0.09$ ($p < 0.001$) and the t-test shows $p < 0.001$. This result demonstrates that, surprisingly, passage readability has a small negative effect on the question difficulty, perhaps pointing to an interfering effect whereby our pre-qualified *human* validation annotators are more likely to make mistakes on more complex passages.

Syntactic and Lexical Surprisal The Flesch–Kincaid grade level only takes the sentence length and the number of syllables into account. To better estimate the passage difficulty in terms of the psycholinguistic modeling of human text processing, we use syntactic and lexical surprisal measures defined by Roark et al. (2009). These measures are computed using an incremental parsing and proved to be useful for predicting human reading time. We observe $r = 0.001$ ($p = 0.97$) for syntactic surprisal and $r = -0.008$ ($p = 0.63$) for lexical

[2] Appendix F provides the frequency of easy and hard examples across the passage sources and the collection methods.

[3] We analyze question and option length in Appendix H.

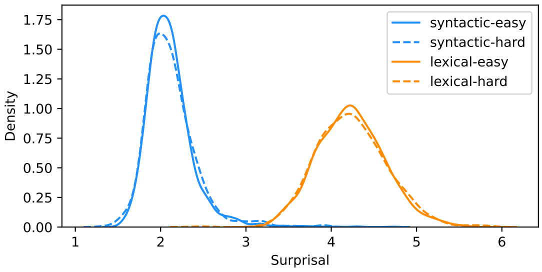


Figure 4: Syntactic and lexical surprisal for easy and hard examples.

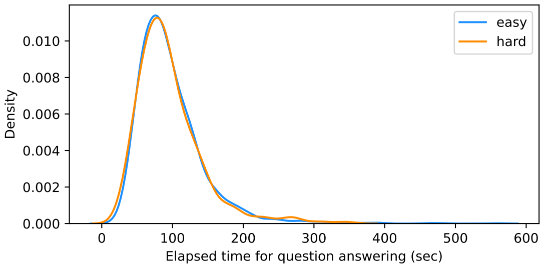


Figure 5: Elapsed time (seconds) for answering easy and hard examples.

surprisal in all examples. We do not observe any statistically significant difference between the easy and hard subsets (syntactic $p = 0.68$ and lexical $p = 0.89$ in the t-test; see Figure 4). Appendix J describes details of the calculation.

Annotation Speed Inspired by the psycholinguistic study of text complexity (Gibson, 1998; Lapata, 2006), we measure an average time of answering questions by crowdworkers in the validation tasks (Figure 5). This measures the elapsed time of both reading a given passage and thinking about its question, which is used as an approximation of reading time (as a proxy of text readability). The correlation coefficient ($r = -0.07$ with $p < 0.001$) and t-test ($p = 0.51$) show that there is only a small negative correlation. We also measure the elapsed time for writing questions as a reference (in Appendix K), observing that there is a weak positive correlation between writing time and question difficulty ($r = 0.05$ with $p = 0.001$).

Word Frequencies Chen and Meurers (2016) analyze the effect of word frequencies in text readability. Following their analysis, we use word frequencies per one million words reported in SUBTLEXus (Brysbaert and New, 2009) to calculate an average frequency of words appearing in a passage as a measure of passage difficulty in terms of vocabulary (a lower average frequency implies being more difficult for reading). We do not observe any statistically significant difference by the t-test $p = 0.40$ (Figure 6) and Pearson's $r = 0.02$ with $p = 0.29$ (see Appendix L for details).

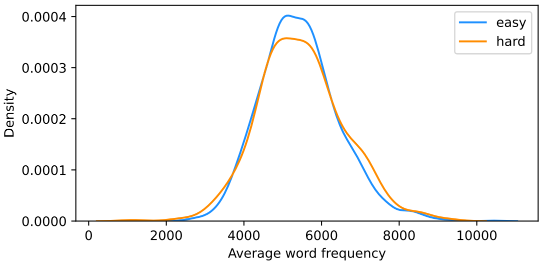


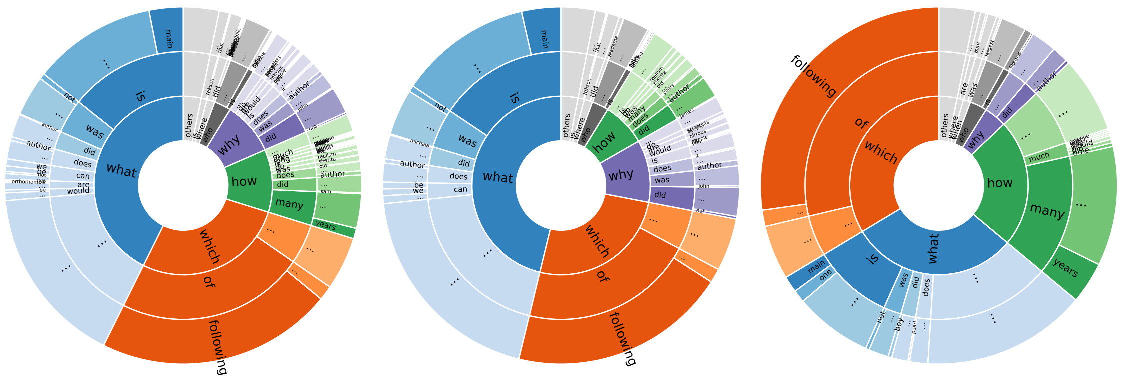
Figure 6: Average word frequency (per one million words) of a passage in easy and hard examples.

5.2 Question Types

We analyze how passage sources and collection methods affect question types in this section.

Question Words We automatically extract *wh*-words that first appear in the valid questions for analyzing the distribution of question types. If no *wh*-word is extracted, it is regarded as a polar question. Figure 7 plots the question words and their two subsequent words (except articles) in the easy and hard questions, where we observe that the hard questions are generic, not specific to given passages (e.g., *which of the following is correct?*) more often than the easy questions. This probably results from the difference between the standard and adversarial data collection. The workers in the adversarial collection tend to write generic questions, while those in the standard collection write questions that are a bit more balanced (e.g., there are more *why* and *how* questions). We also notice that the hard questions have more *how many* questions. This is likely due to the fact that it is easy for annotators to learn that numeric questions often fool the adversary model. These observations imply that adversarial data collection tends to concentrate the distribution of questions towards a few specific question types (e.g., generic and numeric). This is consistent with observations in Kaushik et al. (2021). See Appendix M for details.

Comprehension Types Following Bartolo et al. (2020) and Williams et al. (2020), we analyze what kind of comprehension is required for answering collected questions. We sample a total of 980 high-agreement questions, 70 from each of all passage sources and collection methods, and then manually



(a) Easy+hard questions

(b) Easy questions

(c) Hard questions

Figure 7: Question words and their two subsequent words in easy and hard examples.

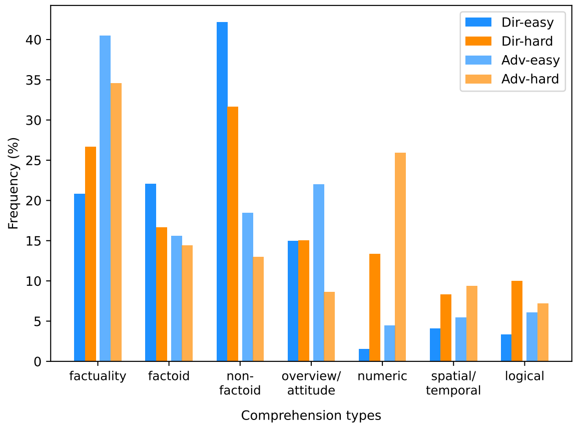


Figure 8: Frequency of comprehension types in easy and hard examples for each collection method.

annotate them with one or more labels of seven comprehension types. The definitions, examples, and detailed results are presented in Appendix M. Regarding the easy and hard subsets (708 and 199 examples respectively), Figure 8 shows the frequency of comprehension types across the question difficulty and the collection methods. We can see that *numeric*, *spatial/temporal*, and *logical* questions appear more often in the hard subset in both collection methods. Looking at the frequency across the passage sources, we find that MCTest has *numeric* questions more than the other sources. We also observe that the adversarial collection generally increases *numeric* questions and decreases *non-factoid* questions (e.g., *how* and *why* questions). *Spatial/temporal* and *logical* questions are less frequent than other comprehension types, but we find *logical* questions more often in ReClor. On the other hand, *spatial/temporal* questions are rarer in Slate, ReClor, and Wikipedia science articles. *Overview/attitude* questions are not relatively difficult, Slate passages have them most often. Although the definition of our comprehension types is coarse, these results show that there are some trends across the sources and collection methods.

6 Conclusion

To make an NLU benchmark useful, it has to consist of examples that are linguistically diverse and difficult enough to discriminate among state-of-the-art models. We crowdsource multiple-choice reading comprehension questions for passages extracted from seven different sources, analyzing what kinds of passages make questions difficult and diverse. Although we expect that the difficulty of passages affects the difficulty of questions, the collected questions do not show any strong correlation between the human–machine performance gap and passage source, length, and readability measures. In contrast, we find that there is a weak correlation between question difficulty and elapsed time for writing examples by workers. Our manual annotation of question types reveals that questions requiring numerical reasoning and logical reasoning are relatively difficult but their frequencies vary across the passage sources. These results suggest that when creating a new benchmark dataset, we need to select passage sources carefully, regardless of the length and difficulty of passages, so that the resulting dataset has questions that require understanding of linguistic phenomena we are interested in. We should take care of that especially in the adversarial setting because an adversary model could concentrate the distribution of questions towards a few specific question types.

Ethics Statement

We aim to accelerate scientific progress on robust general question answering, which could translate downstream to useful tools. We are not looking at possible sources of social bias, though this issue should be highly relevant to those considering sources to use as training data for applied systems. We are using Amazon Mechanical Turk despite its history of sometimes treating workers unfairly (Kummerfeld, 2021), especially in recourse for unfair rejections. We make sure that our own pay and rejection policies are comparable to in-person employment, but acknowledge that our study could encourage others to use Mechanical Turk, and that they might not be so careful.

References

Susan Bartlett, Grzegorz Kondrak, and Colin Cherry. 2009. On the syllabification of phonemes. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 308–316, Boulder, Colorado. Association for Computational Linguistics.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Samuel R. Bowman and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.

Marc Brysbaert and Boris New. 2009. Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41(4):977–990.

Jifan Chen and Greg Durrett. 2019. Understanding dataset design choices for multi-hop reasoning. 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 4026–4032, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiaobin Chen and Detmar Meurers. 2016. Characterizing text difficulty with word frequencies. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 84–94, San Diego, CA. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 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 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with disentangled attention. arXiv preprint 2006.03654.

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. 2006. Integrating linguistic resources: The American national corpus model. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Divyansh Kaushik, Douwe Kiela, Zachary C. Lipton, and Wen-tau Yih. 2021. On the efficacy of adversarial data collection for question answering: Results from a large-scale randomized study. 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 6618–6633, Online. Association for Computational Linguistics.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015, Brussels, Belgium. Association for Computational Linguistics.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. 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 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Research Branch Report 8-75. Chief of Naval Technical Training.

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.

Jonathan K. Kummerfeld. 2021. Quantifying and avoiding unfair qualification labour in crowdsourcing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 343–349, Online. Association for Computational Linguistics.

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.

Mirella Lapata. 2006. Automatic evaluation of information ordering: Kendall's tau. *Computational Linguistics*, 32(4):471–484.

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 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.

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.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. TORQUE: A reading comprehension dataset of temporal ordering questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1158–1172, 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, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

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.

Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 324–333, Singapore. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. Getting closer to AI complete question answering: A set of prerequisite real tasks. In *Proceedings of The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 8722–8731. AAAI Press.

Saku Sugawara, Yusuke Kido, Hikaru Yokono, and Akiko Aizawa. 2017. Evaluation metrics for machine reading comprehension: Prerequisite skills and readability. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 806–817, Vancouver, Canada. Association for Computational Linguistics.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. arXiv preprint 2107.02137.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

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.

Adina Williams, Tristan Thrush, and Douwe Kiela. 2020. ANLIzing the adversarial natural language inference dataset. arXiv preprint 2010.12729.

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.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. ReClor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations*.

A Passage Sources

For Project Gutenberg, we use books from the adventure, fiction, humor, novel, and story subjects.[4]

For Wikipedia articles, we use the articles listed in Vital articles Level 3.[5] For science, we include health, medicine and disease, science, technology, and mathematics categories. For arts, we include history, arts, philosophy and religion, and society and social sciences categories.

B Dataset Statistics

Table 2 presents the frequency of valid, high-agreement, and invalid examples across the passage sources and collection methods.

C Worker Statistics

1,050 workers joined the question-answering phase of the qualification round and 259 workers (24.7%) passed it. 157 workers submitted the question writing task, and 72 workers (36 each for the standard and adversarial collection) qualified for the writing batches, among which 49 workers joined. The workers were allowed to write up to 250 questions. 167 workers participated in the validation

[4] https://www.gutenberg.org/

[5] https://en.wikipedia.org/wiki/Wikipedia:Vital_articles

Source	Method	Valid	High
MCTest	Dir.	91.6	71.3
	Adv.	91.3	73.9
	Total	91.5	72.6
Gutenberg	Dir.	91.3	67.1
	Adv.	89.0	59.4
	Total	90.2	63.2
Slate	Dir.	90.0	66.1
	Adv.	85.5	59.0
	Total	87.7	62.6
RACE	Dir.	94.8	70.3
	Adv.	91.6	67.7
	Total	93.2	69.0
ReClor	Dir.	92.9	72.6
	Adv.	86.1	60.6
	Total	89.5	66.6
Wiki. Sci.	Dir.	92.3	69.0
	Adv.	88.4	58.1
	Total	90.3	63.5
Wiki. Arts	Dir.	91.0	64.5
	Adv.	88.7	60.0
	Total	89.8	62.3
All sources	Dir.	92.0	68.7
	Adv.	88.7	62.7
	Total	90.3	65.7

Table 2: Frequency of valid and high-agreement examples.

Source	Method	Valid	High
MCTest	Dir.	$70.7_{+6.9}$	$72.2_{+6.6}$
	Adv.	$65.6_{+1.8}$	$68.0_{+2.5}$
Gutenberg	Dir.	$79.2_{+5.6}$	$82.1_{+5.5}$
	Adv.	$76.0_{+2.4}$	$79.6_{+3.0}$
Slate	Dir.	$77.1_{+3.8}$	$79.1_{+3.1}$
	Adv.	$74.2_{+0.8}$	$77.0_{+1.0}$
RACE	Dir.	$78.2_{+8.6}$	$79.6_{+9.3}$
	Adv.	$71.8_{+2.3}$	$72.6_{+2.2}$
ReClor	Dir.	$74.6_{+1.6}$	$76.1_{+1.0}$
	Adv.	$72.6_{-0.4}$	$74.6_{-0.5}$
Wiki. Sci.	Dir.	$78.5_{+7.7}$	$79.4_{+8.5}$
	Adv.	$74.8_{+4.1}$	$74.9_{+4.0}$
Wiki. Arts	Dir.	$80.7_{+6.6}$	$79.7_{+5.4}$
	Adv.	$75.3_{+1.2}$	$75.2_{+1.0}$

Table 3: Supervised performance of DeBERTa large. The accuracy of each row is given by the model trained on the questions of the other rows (leave-one-out training). Subscript values show the difference from its zero-shot accuracy.

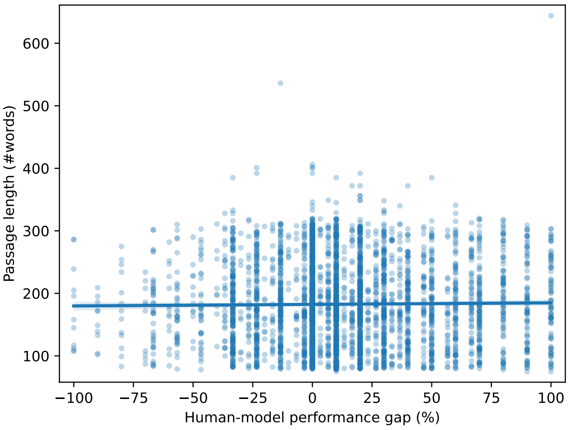


Figure 9: Passage length (number of words) and human–model performance gap. Pearson's $r = 0.01$ with $p = 0.41$.

batches. No worker answered more than 730 questions. Data collection took about a month including the qualification round and the validation batches.

D Supervised Model Performance

Table 3 shows that the supervised performance of the DeBERTa large model.

E Partial-Input Model Performance

Tables 4 and 5 report the zero-shot performance by DeBERTa xlarge and the supervised performance by DeBERTa large (MNLI).

F Easy and Hard Subsets

Table 6 presents the frequency of easy and hard examples across passage sources and collection methods.

G Passage Length

Figure 9 shows the plot between the passage length and the human–model performance gap.

H Question and Option Length

We plot the question and average option length (the number of words except for punctuation) in the high-agreement examples in Figure 10 across the collection methods and Figure 11 across the easy and hard subsets. The distributions of question and option length have slightly higher variances in the standard data collection than in the adversarial data collection. This result is consistent with Nangia et al. (2021).

I Readability Level

Figure 12 shows the plot between Flesch–Kincaid grade level (Kincaid et al., 1975) and the human–model performance gap. We compute the grade level (L) of a passage using the following formula:

$$L = 0.39 * m + 11.8 * n - 15.59 \qquad (1)$$

where m is the average length of the sentences and n is the average number of syllables of the words in

Source	Meth.	P+A	Q+A	A only
MCTest	Dir.	$73.3_{-14.9}$	$39.8_{-48.4}$	$29.4_{-58.8}$
	Adv.	$55.5_{-23.1}$	$41.5_{-37.1}$	$34.5_{-44.1}$
	Total	$64.2_{-19.1}$	$40.7_{-42.7}$	$32.0_{-51.3}$
Gutenberg	Dir.	$75.5_{-13.0}$	$40.9_{-47.6}$	$31.7_{-56.7}$
	Adv.	$55.4_{-27.2}$	$42.4_{-40.2}$	$34.2_{-48.4}$
	Total	$66.1_{-19.6}$	$41.6_{-44.1}$	$32.9_{-52.8}$
Slate	Dir.	$72.7_{-19.0}$	$45.9_{-45.9}$	$32.7_{-59.0}$
	Adv.	$54.1_{-21.9}$	$44.3_{-31.7}$	$33.9_{-42.1}$
	Total	$63.9_{-20.4}$	$45.1_{-39.2}$	$33.2_{-51.0}$
RACE	Dir.	$75.7_{-14.7}$	$49.5_{-40.8}$	$36.2_{-54.1}$
	Adv.	$49.0_{-24.8}$	$43.3_{-30.5}$	$31.9_{-41.9}$
	Total	$62.6_{-19.6}$	$46.5_{-35.7}$	$34.1_{-48.1}$
ReClor	Dir.	$78.7_{-12.4}$	$44.4_{-46.7}$	$35.1_{-56.0}$
	Adv.	$55.9_{-18.6}$	$41.5_{-33.0}$	$26.6_{-47.9}$
	Total	$68.3_{-15.3}$	$43.1_{-40.4}$	$31.2_{-52.3}$
Wiki. Sci.	Dir.	$76.2_{-18.7}$	$45.8_{-49.1}$	$33.2_{-61.7}$
	Adv.	$54.4_{-22.8}$	$35.6_{-41.7}$	$26.7_{-50.6}$
	Total	$66.2_{-20.6}$	$41.1_{-45.7}$	$30.2_{-56.6}$
Wiki. Arts	Dir.	$70.0_{-22.5}$	$49.0_{-43.5}$	$44.5_{-48.0}$
	Adv.	$53.8_{-22.0}$	$44.6_{-31.2}$	$26.3_{-49.5}$
	Total	$62.2_{-22.3}$	$46.9_{-37.6}$	$35.8_{-48.7}$
All src.	Dir.	$74.6_{-16.5}$	$45.0_{-46.0}$	$34.7_{-56.3}$
	Adv.	$54.0_{-22.9}$	$41.9_{-35.0}$	$30.6_{-46.3}$
	Total	$64.8_{-19.5}$	$43.6_{-40.8}$	$32.8_{-51.6}$

Table 4: Zero-shot performance of DeBERTa xlarge trained on RACE with ablation settings. From the input, we ablate questions (*P+A*), passages (*Q+A*), and both question and passages (*A only*). Subscripts show the difference from the full-input accuracy.

Method	P+A	Q+A	A only
Dir.	$71.6^{\pm0.8}_{+0.6}$	$46.0^{\pm2.2}_{+4.7}$	$38.6^{\pm1.5}_{+5.4}$
Adv.	$51.9^{\pm1.3}_{+1.2}$	$41.5^{\pm2.2}_{+1.5}$	$32.7^{\pm0.6}_{+3.3}$

Table 5: Supervised performance (three-fold cross validation) of DeBERTa large on the partial input. Superscripts are standard deviation and subscripts are gains from the zero-shot performance.

Source	Method	Easy	Hard
MCTest	Dir.	7.8	6.5
	Adv.	6.4	12.9
	Total	14.2	19.5
Gutenberg	Dir.	8.2	4.5
	Adv.	5.9	7.0
	Total	14.1	11.6
Slate	Dir.	8.6	2.9
	Adv.	6.1	7.9
	Total	14.6	10.7
RACE	Dir.	8.7	5.2
	Adv.	5.6	13.1
	Total	14.3	18.3
ReClor	Dir.	9.1	5.2
	Adv.	5.4	9.1
	Total	14.5	14.3
Wiki. Sci.	Dir.	9.3	3.2
	Adv.	4.9	10.1
	Total	14.2	13.3
Wiki. Arts	Dir.	8.5	3.2
	Adv.	5.5	9.2
	Total	14.0	12.4
#Questions		2,023	596

Table 6: Frequency (%) of easy and hard questions across the passage sources and collection methods.

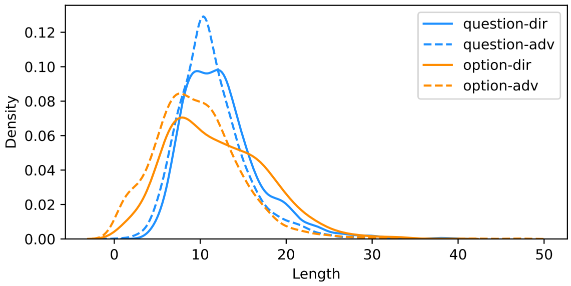


Figure 10: Question and option length (number of words) of examples collected in the standard and adversarial methods.

the passage. To estimate the number of syllables in a word, we use the implementation of the sonority sequencing principle (Bartlett et al., 2009) in NLTK (Bird et al., 2009).[6]

J Syntactic and Lexical Surprisal

Figures 13 and 14 show syntactic and lexical surprisal measures in all examples. Following Roark et al. (2009), we compute a surprisal value for each word, then take an average for each sentence, and finally take an average again over the passage. We use an incremental parser using a lexicalized probabilistic context-free grammar.[7]

K Elapsed Time for Answering Questions

Figure 15 shows the plot of elapsed time for answering questions by humans in the validation tasks. We measure the elapsed time from when a worker opens a task to when the worker submits their answer. In addition, we measure the elapsed time for writing questions as a reference (Figures 16 and 17). We observe that workers take a bit longer time for writing hard examples than for easy examples.

[6] https://www.nltk.org/_modules/nltk/tokenize/sonority_sequencing.html

[7] https://github.com/roarkbr/incremental-top-down-parser

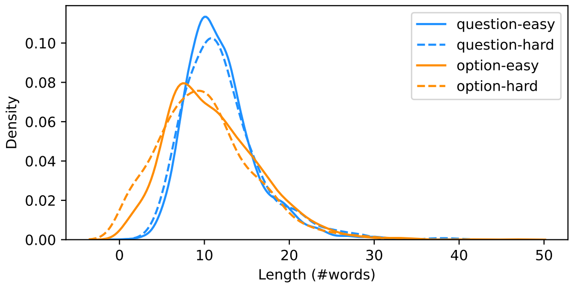


Figure 11: Question and option length (number of words) of easy and hard examples.

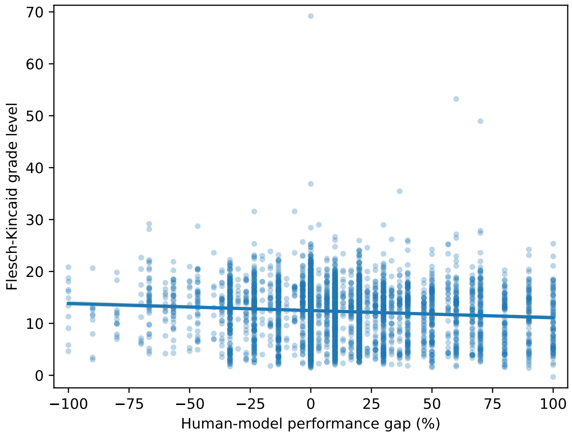


Figure 12: Flesch–Kincaid grade level and human–model performance gap. Pearson's $r = -0.09$ with $p < 0.001$.

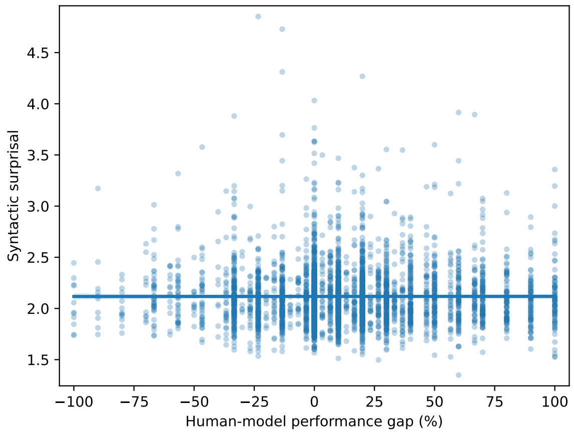


Figure 13: Syntactic surprisal for all valid examples. Pearson's $r = 0.001$ with $p = 0.97$.

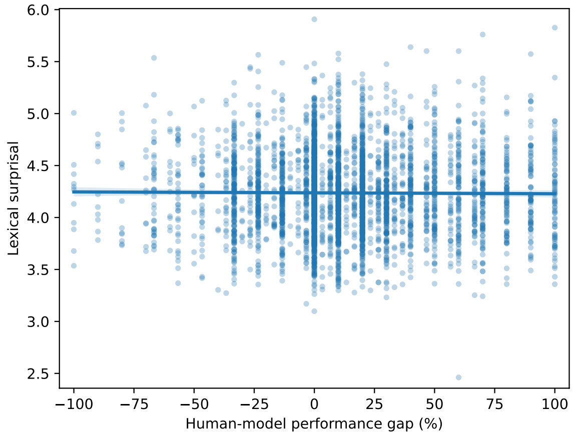


Figure 14: Lexical surprisal for all valid examples. Pearson's $r = -0.01$ with $p = 0.63$.

L Average Word Frequencies

Figure 18 plots the average word frequencies of all examples. We refer to SUBTLEXus (Brysbaert and New, 2009) for the word frequencies per one million words in a corpus of American English subtitles.

M Question and Comprehension Types

Figure 19 shows the frequency of the question words and their two subsequent words for each collection method. Figures 20 and 21 show the box plots between human–model performance gaps and questions words and comprehension types respectively. The triangle markers indicate mean values and black bars indicate medians. Figures 22 and 23 show the frequency of question types and comprehension types across the passage sources and collection methods. In the annotation of comprehension types, a question can have multiple labels. Therefore the sum of the frequencies may exceed 100%.

The definitions of comprehension types and their examples are as follows:

1. **Factuality** (*true/false/likely*) is reasoning of which answer option most (or least) describes facts or events in a given passage.

2. **Factoid** simply asks about described events or entities mainly with typical *what* questions.

3. **Non-factoid** is related to *why* and *how* questions such as ones asking about the causality, character's attitude, and the process of described events.

4. **Overview/Attitude** is for questions that ask about the summary, theme, and conclusion of the content of a given passage and the author's attitude and claim that readers can derive from it.

5. **Numeric** indicates questions that require arithmetic reasoning.

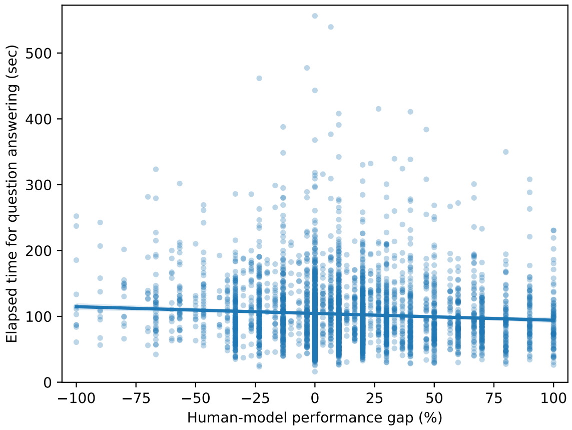


Figure 15: Elapsed time (seconds) for answering questions. Pearson's $r = -0.07$ with $p < 0.001$.

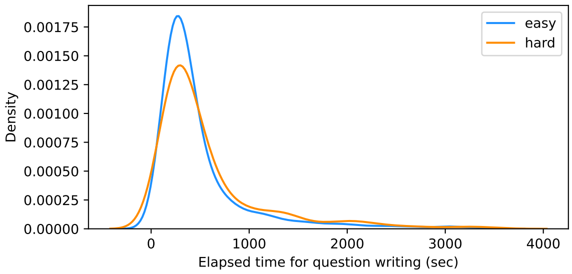


Figure 16: Elapsed time (seconds) for writing easy and hard examples. The t-test shows $p = 0.004$.

6. **Spatial/Temporal** is related to understanding of place and locations (spatial) or the temporal order or duration (temporal) of described events.

7. **Logical** is pertinent to logical reasoning and arguments described in a passage.

Table 7 shows examples of questions and options for the comprehension types. After extracting question words, we review about 100 questions to collect keywords that determine comprehension types (e.g., "reason" for *non-factoid*,"best summarize" for *overview/attitude* and "if" for *logical*). We then write simple rules that highlight these keywords, which help us manually annotate the remaining questions within about five hours.

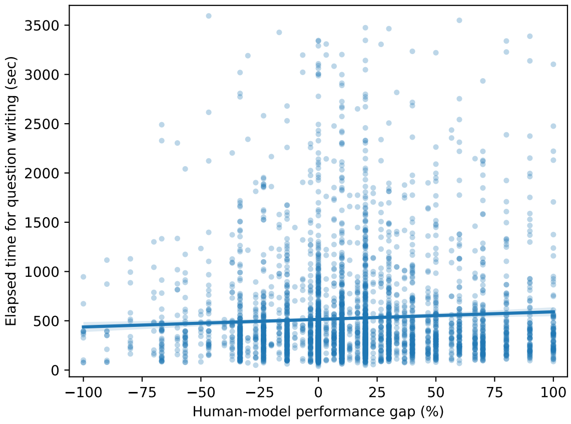


Figure 17: Elapsed time (seconds) for writing all examples. Pearson's $r = 0.05$ with $p = 0.001$.

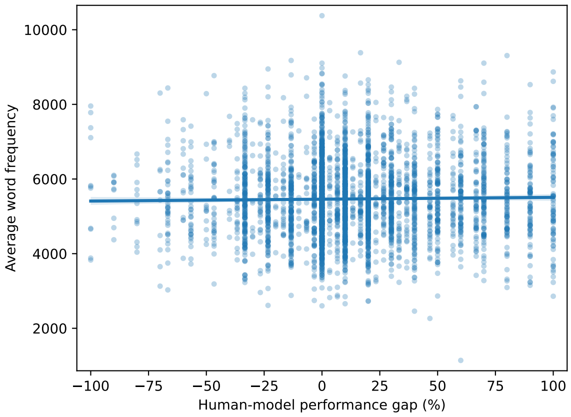
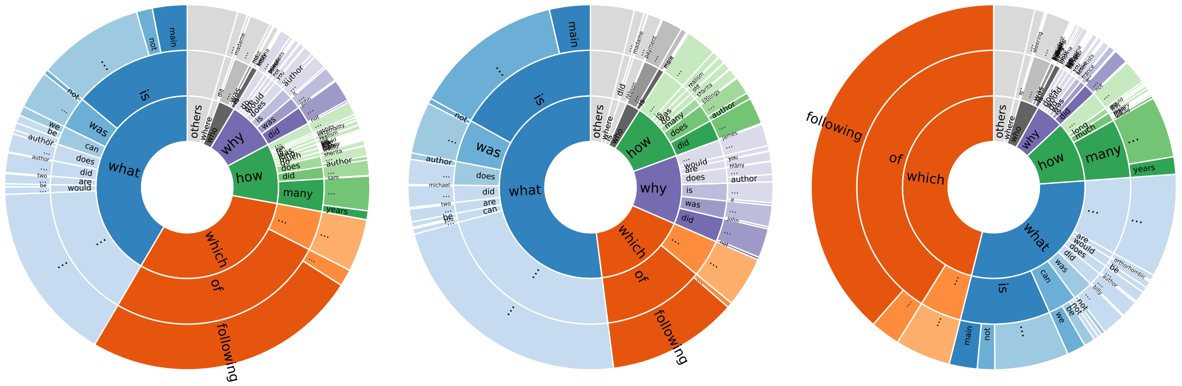


Figure 18: Average word frequencies using the values of SUBTLEXus. Pearson's $r = 0.02$ with $p = 0.29$.



(a) All questions

(b) Standard collection

(c) Adversarial collection

Figure 19: Question words and their two subsequent words in the standard and adversarial collection methods.

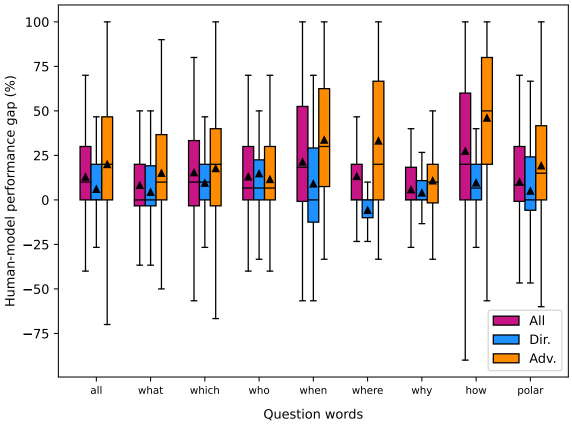


Figure 20: Question words and human–model performance gaps.

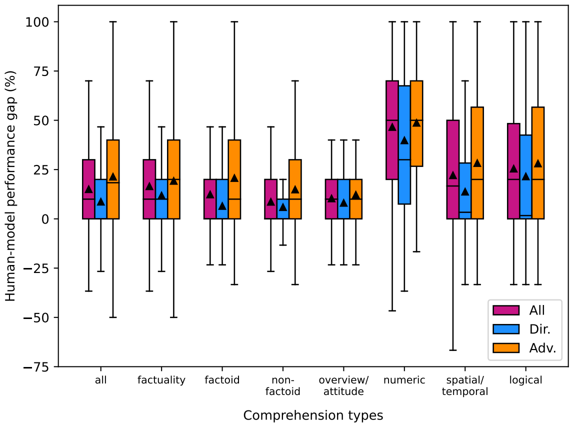


Figure 21: Comprehension types and human–model performance gaps.

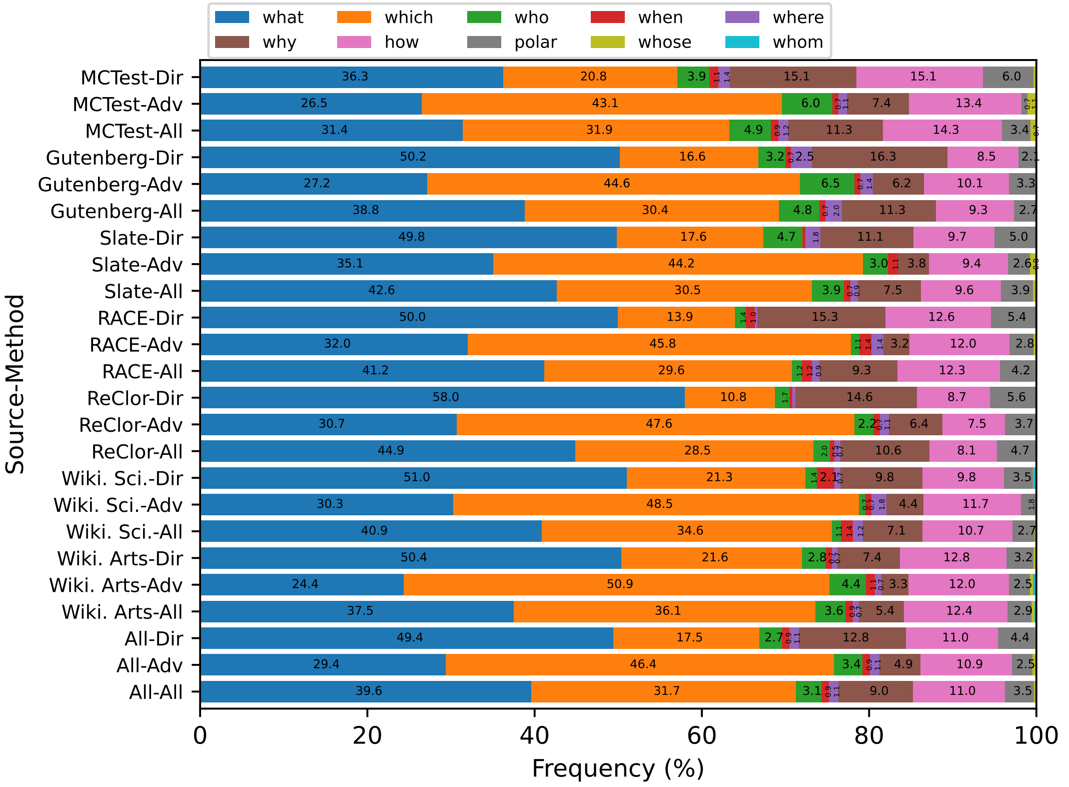


Figure 22: Frequency of question types (wh words) across passage sources and collection methods.

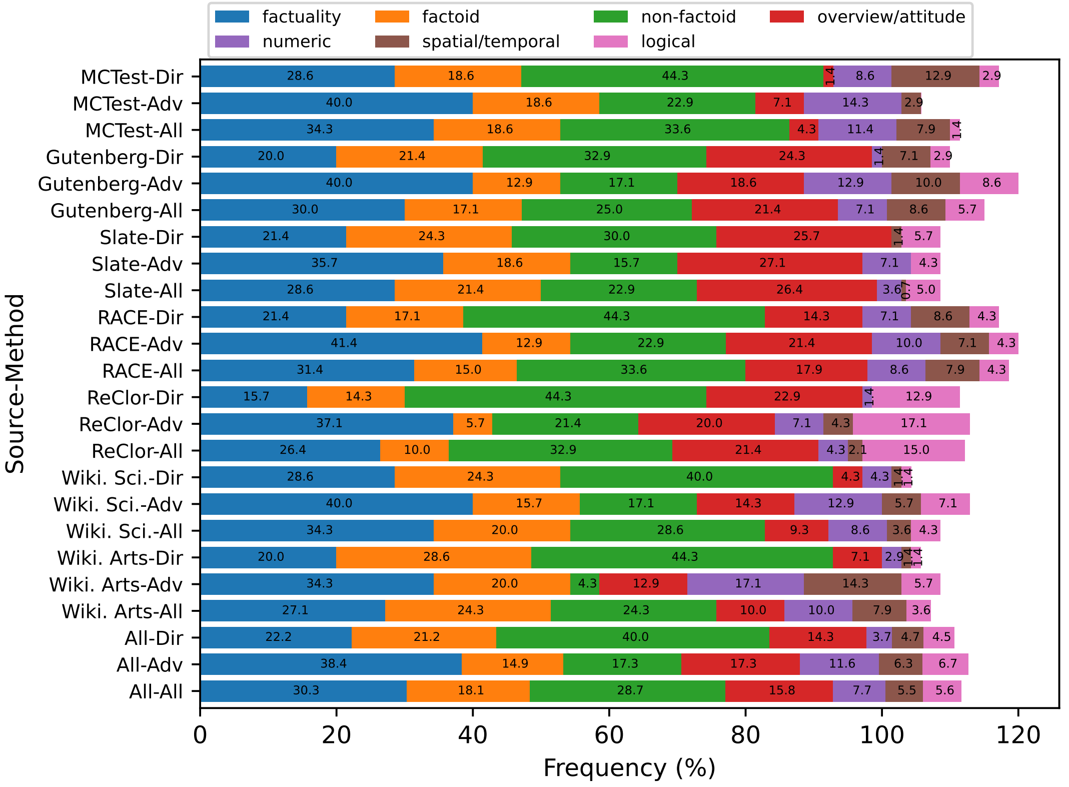


Figure 23: Frequency of comprehension types across passage sources and collection methods. Because a question can have multiple labels, the sum of the frequencies may exceed 100%.

Comprehension Type (source, difficulty)	Example
Factuality (Gutenberg, easy)	Q: Which of the following is not mentioned in the passage? A: □ An Earl lived in a house that had a relatively low profile. / □ There were some other buildings near the Manor. / □ Scroope is a village that is closely linked to an Earl's home. / ☑ Scroope Manor was sold to the village by the Earl.
Factoid (Wiki. science, easy)	Q: What helps many fish keep their buoyancy in water? A: □ muscles on either side of the backbone / □ fins / ☑ a swim bladder / □ a streamlined body
Non-factoid (Wiki. arts, hard)	Q: How did a major portion of English words enter the English language? A: □ French speakers can understand many English words without having to undergo any orthographical change. / □ Many words in Old English are from Old Norse. / ☑ About one-third of words in English entered the language from the long contact between French and English. / □ Romance languages have "Latinate" roots.
Overview/Attitude (Slate, easy)	Q: Which of the following is a criticism the author has about Dick Riordan? A: □ He's not transparent about his typical lunch looks like, which highlights his lack of wisdom. / ☑ He's okay syphoning resources from elsewhere to himself for personal gain. / □ Much like Hillary Clinton, he lacks any sort of coherent persona. / □ He is responsible for the vast swaths of one-story buildings that cover the entire landscape of L.A.
Numeric (RACE, hard)	Q: How old was Mary Shelley when she died? A: □ Mary Shelley was in her thirties when she died. / □ Mary Shelley died when she was forty four years old. / ☑ Mary Shelley died when she was in her fifties. / □ Mary Shelley lived well into her eighties before she died.
Spatial/Temporal (MCTest, easy)	Q: When did it start to rain? A: ☑ It started to rain after Will ate his biscuit and jam. / □ It started to rain after Will heard the thunder. / □ It started to rain while Will was at the store. / □ It started to rain on Will's walk home from the store.
Logical (ReClor, hard)	Q: Which statement, if true, would weaken the conclusion of the passage? A: □ Archaeologists have found remains of shipwrecks from 2000 BC between Crete and southern Greece. / ☑ The earliest bronze artifacts found in southern Greece date to 3000 BC. / □ The Minoans were far more accomplished in producing bronzeware than any other civilization in the area at the time. / □ The capacity of Minoan bronze furnaces was extraordinarily large compared to other societies in 2000 BC.

Table 7: Examples of the comprehension types taken from our collected data.