What Has Been Lost with Synthetic Evaluation?

Alexander Gill[†] **Abhilasha Ravichander**[‡] **Ana Marasović**[†]
[†]University of Utah
[‡]University of Washington
alex.gill@utah.edu aravicha@cs.washington.edu ana.marasovic@utah.edu

Abstract

Large language models (LLMs) are increasingly used for data generation. However, creating evaluation benchmarks raises the bar for this emerging paradigm. Benchmarks must target specific capabilities, penalize exploiting shortcuts, and be challenging. Through two case studies, we investigate whether LLMs can meet these demands by generating reasoning-over-text benchmarks and comparing them to those created through careful crowdsourcing. Specifically, we evaluate both the *validity* and *difficulty* of LLM-generated versions of two high-quality reading comprehension datasets: CondaQA, which evaluates reasoning about negation, and DROP, which targets reasoning about quantities. We find that prompting LLMs can produce variants of these datasets that are often valid according to the annotation guidelines, at a fraction of the cost of the original crowdsourcing effort. However, we show that they are *less challenging for LLMs* than their human-authored counterparts, do not preserve the same model rankings, and that these differences may be imperceptible to researchers inspecting the data. This finding sheds light on what may have been lost by generating evaluation data with LLMs, and calls for critically reassessing the immediate use of this increasingly prevalent approach to benchmark creation.

CondaQA	**Human**[†]	`o3-mini`[△]	`Llama-3.3`[°]
Original-Data Accuracy	91.9	72.4	78.8
Full Bundle Consistency	81.6	45.9	48.5
Paraphrase Consistency	93.6	69.6	76.5
Scope Consistency	86.5	55.5	62.8
Affirmative Consistency	88.2	56.9	67.3
DROP	**Human**[‡]	**o3-mini**[△]	`Llama-3.3`[°]
Original-Data Token F1	96.4	84.3	70.9
Consistency	N/A	53.5	35.5

Table 1: Performance on the CondaQA dev set and the DROP contrastive test set. The original accuracy is calculated on the data used to create contrastive instances. Consistency reflects the rate at which all minimally different examples within a bundle are answered correctly. *These results show that these reasoning benchmarks remain challenging. We study whether comparable variants can be created using LLMs.* [†]Cf. Ravichander et al. (2022). [‡]Cf. Dua et al. (2019). [△]`o3-mini-2025-01-31` [°]`meta-llama/Llama-3.3-70B-Instruct`

1 Introduction

The development of benchmarks that test frontier models in reasoning over text has been at the heart of NLP research (Rogers et al., 2023). Such benchmarks aim to cover a broad range of input variations and related skills, and prevent models from performing well by exploiting data shortcuts (Bowman and Dahl, 2021). They are created to be challenging, so that even the most capable emerging models cannot solve them immediately and they remain a meaningful yardstick for measuring progress. Yet these goals are often hindered by crowdsourcing annotation practices that favor speed to maximize hourly pay, such as crowdworkers reusing the same pattern(s) to create dataset examples. An alternative approach, prompting LLMs, does not suffer from the same misaligned incentives. LLMs are capable of creating content they will struggle to reason about (West et al., 2024, *the generative AI paradox*). However, a key challenge lies in getting LLMs to follow annotation instructions. The silver lining is that it is possible to review annotations and revise guidelines without delay because LLMs produce annotations instantly. Moreover, since their annotations come at little or no cost, one can iterate as many times as they are willing to review-and-revise.

This makes it tempting to substitute LLMs for human annotators, as is becoming increasingly common (§3). However, before adopting this approach to creating new challenging benchmarks, we must ask: *When created under the same annotation criteria, how do LLM-generated benchmarks*

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 9902–9945

compare to their challenging human-authored counterparts? If the data is not comparable in quality or as effective for assessing model limitations, this should give us pause in using LLMs as annotators in this context.

We address this question through the lens of reading comprehension, which is a common format in NLP for probing reasoning over text capabilities of frontier models (Rogers et al., 2023). Specifically, we study compositional reasoning over numbers as in DROP (Dua et al., 2019; Gardner et al., 2020) and reasoning about the implications of negated statements as in CondaQA (Ravichander et al., 2022). These reasoning types continue to challenge LLMs, as shown in Table 1.

What would LLMs-as-annotators need to be able to generate to create datasets like these? First, questions that are complex and require deeper understanding or inference, rather than being answerable by a phrase matching the paragraph's wording. Second, minimal edits to the passages or questions to test that the model has the skills related to a specific type of reasoning. For example, an edit that only changes what is negated in a passage enables testing whether the model has the skill of understanding negation scope. In turn, this skill enables reasoning about *both* the original and edited passage, and answering questions about the negated content in the context of *both* passages correctly.

We extensively prompt an LLM to generate complex questions and contrastive edits, as would likely be done by dataset creators opting to use LLMs as annotators. Prompt engineering we conduct (§4) illustrates that using LLMs as annotators does not eliminate the need for iterative instruction refinement and manual review, much like traditional human annotation. That is, although an LLM creates questions and edits faster and cheaper than people, the process of obtaining them may still require multiple rounds of manual review.

Our next findings are of practical importance. First, we manually analyze generations and find that LLM-generated data is moderately to highly valid, though rarely perfect (§4). Second, in a user study with NLP researchers, we compare human and generated edits and find that *generated edits are generally preferred to human-authored ones for better meeting the annotation guidelines* (§5). We attribute this result to their apparent simplicity, which allows them to adhere more closely to the guidelines. However, when benchmarking various models on human and generated versions, we find that human-authored edits still pose a greater challenge and rank models differently than generated versions, especially when bundle consistency is considered (§6).

These findings suggest that *the core challenge is not validity, but difficulty*. However, optimizing prompts for difficulty would require substantial human-in-the-loop effort because LLMs cannot yet reliably validate and answer their own generations. Moreover, this would effectively turn evaluation-data generation into a form of adversarial dataset construction whose "naturalness" has been questioned (Bowman and Dahl, 2021). Therefore, how to generate challenging instances without relying on excessive human effort and producing artificially-sounding text remains an open question. In light of this, using LLMs as annotators for any new benchmark does not seem advisable as of yet, especially if the validity of the generated data is the only measure of dataset quality.[1]

2 Background

Reading Comprehension to Probe Reasoning. SQuAD (Rajpurkar et al., 2016) popularized using reading comprehension to probe models. Weissenborn et al. (2017) show that SQuAD questions are often answerable with a span in the paragraph that matches the expected answer type and is lexically related to the key words in the question. This has led to the other datasets posing greater reasoning challenges (Yang et al., 2018; Kočiský et al., 2018; Yu et al., 2020, among others). We focus on two such datasets: DROP (Dua et al., 2019) and CondaQA (Ravichander et al., 2022), which still challenge current models (Table 1). DROP is widely used, as evident by its citation count, and CondaQA exemplifies the kind of high-quality, contrastive, and reasoning-focused dataset we aim to test LLMs' ability to generate.

DROP and CondaQA Questions. See Table 8 in the Appendix for examples of human-authored questions in both datasets. Questions in these datasets are posed over Wikipedia passages. In DROP, questions must "require complex linguistic understanding and discrete reasoning" (Dua et al., 2019). An extra condition is that the BiDAF model (Seo et al., 2017) cannot answer a question. We do not adversarially generate, as this may raise other issues (Bowman and Dahl, 2021). In CondaQA, questions should be "targeted towards the

[1]Our code is available at Github.

implications of the negated statement" (Ravichander et al., 2022), i.e, they should be about (1) the negated statement, rather than other information in the passage, and (2) about an implication of the negated statement.

Contrast Sets. Models can solve benchmarks by exploiting data shortcuts (Gururangan et al., 2018). Thus, they are better evaluated based on *consistency* — their ability to solve an entire set of slightly different examples with different ground truth, known as *contrast sets* (Gardner et al., 2020) or *counterfactual data* (Kaushik et al., 2020). Contrastive edits are valid if minimal changes give a different ground truth, while maintaining fluency. However, edits *targeting specific changes* help evaluate whether models have the necessary skills for a specific type of reasoning. For example, an edit that requires changing only what is negated tests models' understanding of the negation scope:

- "Nearly all of his possessions were destroyed *with the exception* of a guitar and an automobile"
- Edit: "Nearly all of his possessions were destroyed (including a guitar), *with the exception* of an automobile"

Only with this skill can models consistently reason about negation and answer questions as "Was Parsons able to use his guitar after the fire?" in both contexts. We introduce other edit types below.

CondaQA Edits. The example above is the CondaQA *scope edit*, which tests robustness to changing the part of the text that the negation cue affects. The other edit types are *paraphrase edits*,[2] which test a model's robustness to different ways of expressing negation, and *affirmative edits*, which reverse negation to assess whether models can process negation and whether they rely on subtle artifacts when they do.[3] CondaQA was designed to be contrastive, with all edit types systematically annotated. See Table 9 (Appendix) for human-authored edits. We aim to generate all of these edit types for CondaQA dev passages.

DROP Edits. Gardner et al. (2020) report three DROP edit types. *Compositional edits* add a reasoning step to questions (Table 8; Appendix). *Semantic edits* invert question meanings, such as changing "shortest" to "longest". *Temporal edits* modify event order by changing their order in the questions or the passage (Table 6; Appendix), or by changing the dates associated with these events in the passage. Unfortunately, the DROP contrast set does not explicitly mark the edit type.[4] Thus, one author of this paper has manually determined compositional and temporal edits that change event order in passages, but could not reliably detect semantic edits.[5] We generate compositional edits using pairs of passages and the generated questions. Because the contrast set contained too few temporal edits, we generate temporal edits using the original DROP dev passages that include dates.

3 Related Work

Question Generation (QG). This is a long-standing NLP task, with the first shared task in 2010 (Rus et al., 2010) and many approaches proposed since (Pan et al., 2019; Guo et al., 2024). QG still largely focuses on generating fluent, factual questions that closely mirror the context's wording (Ushio et al., 2022; Samuel et al., 2024). Yehudai et al. (2024) target long-form questions, but still require little composition or inference. Multi-hop QG, which links explicit facts across passages (Kulshreshtha and Rumshisky, 2023), is more relevant but differs from reasoning over *implicit* meaning within a single context as in CondaQA and DROP. Other related QG work focuses on text-to-SQL (Wu et al., 2021), pedagogical question generation (Tonga et al., 2025), and paraphrasing existing mathematical questions (Yu et al., 2024), none of which involve generating reasoning-intensive questions from scratch. Balepur et al. (2025) introduce RQA, where a question is generated from a given answer, showing that LLMs struggle particularly with numerical cases.

Edit Generation. A line of work has aimed to automate the creation of contrastive edits to improve their scale and diversity (Li et al., 2020; Ross et al., 2021; Shaikh et al., 2022; Ross et al., 2022; Dixit et al., 2022; Chen et al., 2023, among others); reviewed in Stepin et al. (2021); Nguyen et al. (2024); Wang et al. (2024). While these approaches accept any minor change that alters the answer, we focus on editing that is done along specific dimensions such as changing what is negated or adding a set of

[2]While not contrastive to the original, paraphrase edits combined with other edit types, like scope and affirmative edits, form sets where some variants yield different answers.

[3]Success on negated text but failure on its affirmative version suggests artifacts were used. Failure on negated, success on affirmative text shows an inability to handle negation.

[4]https://github.com/allenai/contrast-sets/blob/main/DROP/drop_contrast_sets_test.json

[5]We choose *passage-level* temporal edits because they, with compositional edits, cover at least one type of both question-level and passage-level DROP edits.

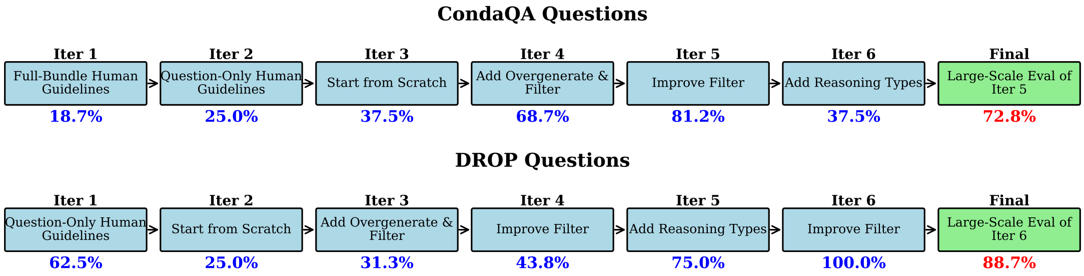


Figure 1: The percentage of **valid questions** in a sample of 16 for intermediate evaluations, and 100 or 114 for final evaluations. Questions are generated with `gpt-4-turbo-2024-04-09` and assessed by one author of this paper. The full prompt code and templates for each iteration can be found in our Github repository.

calculations. Ross et al. (2022) also steer edits toward targeted attributes, but their approach requires dataset creators to adjust semantic representation-based control codes, whereas we achieve targeted edits through direct natural language instructions.

Paraphrase Generation. This is another classic NLP task (McKeown, 1983). Our approach to generating CondaQA edits that change how negation is expressed, but keep the meaning intact, builds on work in prompting LLMs for paraphrasing (Cegin et al., 2023; Witteveen and Andrews, 2019; Pehlivanoglu et al., 2024) and controllable paraphrasing (Hosking and Lapata, 2021; Krishna et al., 2023).

Data Creation with LLMs. Generating data with LLMs is increasingly common (Long et al., 2024), from NLP tasks such as classification (Møller et al., 2024; Patwa et al., 2024), QA (Ye et al., 2022; Harsha et al., 2025), NLI (Liu et al., 2022; Hosseini et al., 2024), to instruction tuning and alignment (Wang et al., 2023; Bai et al., 2022). Our work differs from this line of work in three key ways. First, most prior efforts focus on generating *training*, not *evaluation*, data. Imperfect synthetic data may improve training (Amin et al., 2025), but is unacceptable for benchmarking. Second, much of the existing work targets the easier task of generating only labels or answers (Maheshwari et al., 2024), whereas we generate content: questions and edits. This is more challenging to execute due to more complex prompt engineering and validity review, as well as costly answer annotation. Third, our goal is to critically and systematically analyze generated data. While some studies generate evaluation datasets or replicate existing crowdsourcing pipelines (Gilardi et al., 2023; Qi et al., 2025; Roemmele and Gordon, 2024; Wu et al., 2025), they do not conduct such analysis. Das et al. (2024) compares LLM-generated data to human-authored counterparts, but focuses on training data and the downstream effects of its use. Our results both corroborate and go beyond Maheshwari et al. (2024)'s. Specifically, we show that synthetic data is not always a good predictor of relative performance, and that bias does not explain why synthetic benchmarks are easier.

4 Generating DROP and CondaQA

In this section, we outline our prompting workflow for generating questions (§4.1) and edits (§4.2). We also recap some of the prompting strategies we found most useful, and the data they were used to create in Table 10 (Appendix).

We give a Wikipedia passage from the original dataset to `gpt-4-turbo-2024-04-09` (the state-of-the-art LLM at the time) and refine prompt instruction iteratively, with one of the authors manually validating 16 generations for each promising candidate prompt. We consider an annotation valid if it adheres to the original annotation guidelines; see validity criteria in Table 11 in the Appendix. If validity is below 85%, we adjust the prompt and experiment using a chat interface until we reach a promising version, which we then re-evaluate on the 16 examples. Upon achieving either 100% validity, or if further prompt engineering suggests no prospects of progress, the same author evaluated the most valid intermediate prompt on a larger sample of 100+ instances. Each iteration in Figures 1–2 represents our prompt after exploring several minor variations around a central prompting idea in that iteration. We also show these results in Tables 12 and 13; Appendix. In total, we evaluate 421 paragraph-question pairs and 926 edits across both datasets in this section.

4.1 Prompting for Question Generation

We begin with prompts based on the instructions given to annotators.[6] For CondaQA, one such

[6]We obtain CondaQA crowdsourcing templates via personal correspondence. The DROP templates are available at:

prompt obtains a question and paragraph edits sequentially in a single chat (Iter 1; CondaQA) and the other collects questions or a specific edit alone (Iter 2; CondaQA). For DROP, the latter is the only option (Iter 1; DROP).[7] As shown in Fig. 1, these initial prompts result in low validity. *This shows that* `gpt-4-turbo-2024-04-09` *struggles to follow instructions designed for human crowdworkers.* Consequently, we switch to designing prompts from scratch.

For CondaQA, we prime the model to focus on useful information by guiding it through the following process: given a passage and a cue, it identifies what is negated, de-negates the cue, considers the implications of the negation, and then generates a question. The prompt requires that the generated question (1) targets the negation, (2) is about an implication of negation, not factual knowledge in the passage, (3) has a valid answer type (yes, no, don't know, or a span in the passage), and (4) avoids restating the negation. We try to ensure that the negation is targeted by requesting questions whose answers change depending on whether the negation cue or its de-negated version is in the passage.

For DROP, the initial prompts ask for questions that (i) require discrete reasoning over the passage and (ii) have a valid answer span (a span in the passage or question, date, or numbers).

The validity of these initial prompts is also low: only 37.5% for CondaQA (Iter 3) and 25% for DROP (Iter 2). *We improve these results with the overgenerate-then-filter protocol* (Yehudai et al., 2024). We generate 8 questions for each passage and select the first one that satisfies filtering. Each filter is done with another LLM call.

For CondaQA, we first introduce filters for (1)–(3) above, but we found that including a filter for (4) was crucial (see Iter 4 vs. 5 in Fig. 1). After this iteration, we are unable to further improve the validity of generated CondaQA questions, and ultimately, their validity in a sample of 114 passage-question pairs is 72.8%.

For DROP, we refine the filtering in a few iterations, each of which further improves the validity. The first filter checks that a question requires complex reasoning and looking over more than one part of the passage to answer it (Iter 4). The next filter checks that the question does not combine multiple questions into one and that it has a valid answer type. Before introducing the final filter, we enrich the prompt with an example of 9 types of discrete reasoning questions and request generating a question for each discrete reasoning type, then ranking them according to how well they adhere to the other guidelines (see Iter 5).[8] Given this improvement, we try a similar addition for CondaQA (Iter 6), but it drastically reduces validity.[9] *This demonstrates that a prompt engineering strategy crucial for producing one reasoning benchmark may not be universal.* The final filter checks that the passage is self-contained in terms of facts required to answer the generated question (e.g., it does not require external knowledge about a sport or player stats). With this final filter, we reach 100% validity on the small sample which leads to 88.7% valid generated DROP questions in a larger sample of 100 passage-question pairs. This performance drop from a small- to large-scale evaluation pinpoints a limitation of this protocol, which could also happen in traditional human evaluation.

4.2 Prompting for Editing

Unlike for question generation, initial prompts for certain edits, such as paraphrasing and affirmative ones, immediately give moderate to high validity (CondaQA Iter 3, Fig. 2). However, multiple iterations were required to refine the prompt for complex edits like changing the negation scope. In that case, we discover it is effective to have the model generate a question whose answer is the entity or event that is negated, and then edit the passage so the answer to the question changes (Iter 6; Scope):

```
System: Given the sentence and the negation cue, identify what
       is being negated.
User: Sentence: {sentence}\n\nNegation cue: {cue}
Assistant: {negation_subject}
User: Please generate a question that is answered by the
      subject of what is negated by \"{cue}\" in the original
      sentence. Just output the question and the answer in the
       form of \"Q: Generated question \nA: Answer to
      generated question\"
Assistant: Q: {generated_question} \n A: {generated_answer}
User: Now edit the original sentence so that the answer to the
       question is different, but make sure to keep the
      negation cue \"{cue}\" somewhere in the sentence.
Assistant: {edited_sentence}
User: Original Passage: {passage}\nNow rewrite the original
      passage with the rewritten sentence. Please make any
      edits necessary to make sure the whole passage is
      coherent. Please return just the edited passage.
Assistant: {edited_passage}
...
```

https://github.com/dDua/mturk-drop.

[7]The DROP contrast set authors shared that their instructions were loosely defined since they did the edits themselves.

[8]Types of discrete reasoning (Dua et al., 2019): subtraction, comparison, selection, addition, count and sort, coreference resolution, other arithmetic, set of spans, and other.

[9]Dimensions of commonsense reasoning (Ravichander et al., 2022): precondition, social norms, psychology, cause and effect, mutual exclusivity, synecdoche.

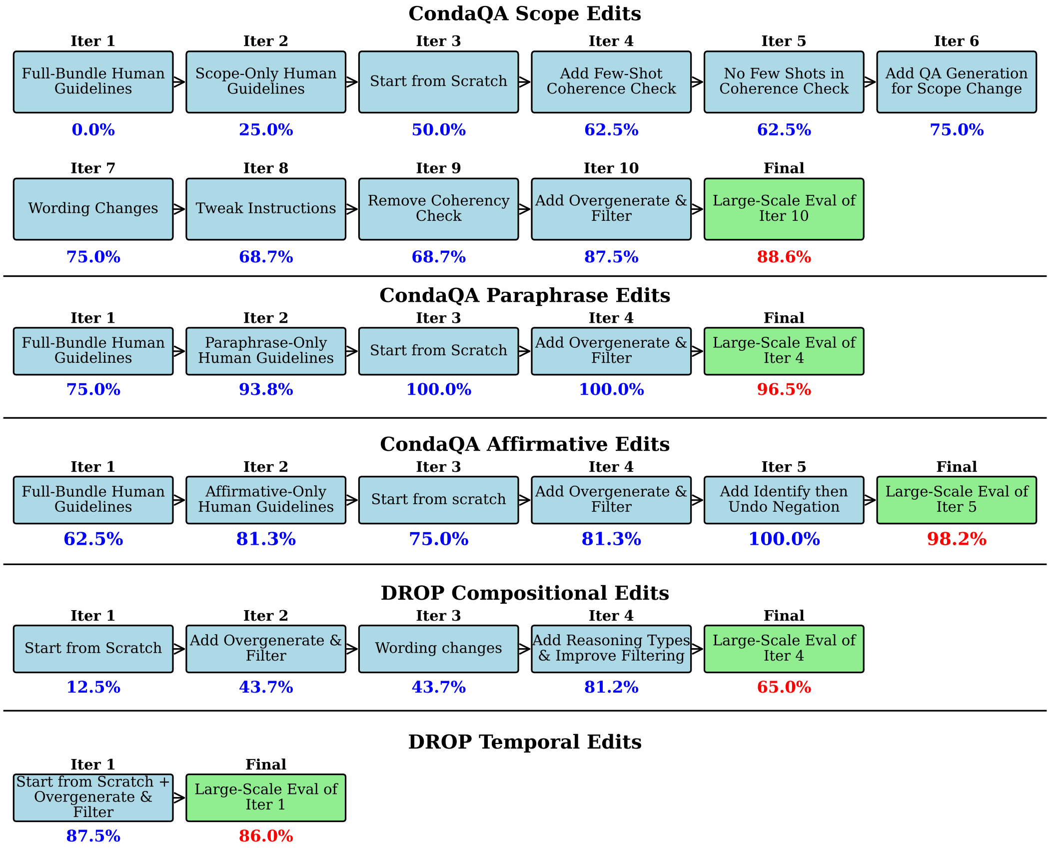


Figure 2: The percentage of **valid edits** in a sample of 16 for intermediate evaluations, and 100 or 114 for final evaluations. Edits are generated with `gpt-4-turbo-2024-04-09` and assessed by one author of this paper. The full prompt code and templates for each iteration can be found in our Github repository.

In most cases, the model needs to be primed to focus on relevant information and guided through a multi-step generation process. For instance, in affirmative edits, the initial prompt (Iter 3) involves removing the negation cue and then polishing the text. A more effective alternative (Iter 5) is to explicitly ask what is being negated, ask the model to de-negate the negation (e.g., change "rarely" to "often") while keeping the original negated content in focus, and then integrate the revised content into the passage. We often introduce overgenerating-and-filtering with other refinements, which makes its role in editing less immediately apparent. However, it is crucial (except for paraphrasing), as evident by its use in all final prompts.

Finally, while we achieve high validity for paraphrasing and affirmative edits, validity remains moderate for scope and temporal edits and poor for compositional edits.[10] In those cases, as with question generation, human verification would be necessary to ensure clean contrast sets.

5 Whose Annotations are Preferred?

We next study whether NLP researchers prefer human or generated questions and edits based on how closely they follow the original guidelines for producing questions and edits with desired properties. There is a concern about potential dataset contamination when doing such studies with synthetic data, as models may simply memorize and reproduce the human-authored dataset from their training data. To mitigate against the effects of straightforward reproduction, we analyze the generated data to eliminate the possibility that generations are paraphrased instances of their human-authored counterparts; see Appendix B.

Setup. We recruited 16 annotators who are NLP researchers and doctoral students to play the role of "dataset creators".[11] These annotators were selected to represent the typical demographic of researchers who develop evaluations for LLMs. We recruited

[10] To produce compositional edits, we do not edit the original questions because we simulate creating a dataset with complex questions and edits from scratch.

[11] We obtained approval from our institution's IRB.

them through social media platforms and personal networks and compensated them at a rate of $16 per hour. In total, 63 paragraph-question pairs and 125 edits across two datasets were evaluated in this study, at a total annotation cost of $256.

Screenshots of this study are shown in Figures 5–12 for CondaQA and Figures 19–26 for DROP (Appendix).[12] Annotators were asked to imagine themselves as dataset creators who were designing a benchmark to probe LLMs' capabilities to reason and were provided background about the format of the datasets (without mentioning their names). They were then asked to choose edits and questions that better meet the dataset specification. They were also asked if the question or edit was invalid. For CondaQA, annotators are shown one question pair and three edit pairs corresponding to the three edit types. For DROP, annotators are shown one question pair and one compositional-edit pair.[13] Since the compositional edits in §4 were made on generated questions, we created compositional edits for human-authored questions here to enable a valid comparison. The order of human and generated annotations for both studies is randomized. Other details are in Appendix §C.

Results. Tables 2–3 show the results. We observe that annotators assess both generated and human edits and questions as valid at a much lower rate than one author of this paper, with the exception of DROP compositional edits. They also rate the original CondaQA paraphrase/scope edits and questions, along with the DROP questions, as less valid than the generated ones. Moreover, annotators are often shown to prefer model-generated questions and edits to human-written questions and edits from the original dataset. We include example agreement and disagreement cases between annotators and one author of this paper for generated instances of CondaQA and DROP questions and edits in Tables 16–21 in the Appendix.

The annotators' higher validity and preference for generations likely stems from the fact that the generated edits tend to be simpler and, due to the way the prompts were designed, follow the guidelines for each type of edit more closely. In contrast, original annotators often interpret the question and edit definitions in slightly different ways and sometimes introduce extra flair to the annotations. While these human annotations are, in our opinion, still valid, they do not always strictly follow the edit guidelines. This is aligned with observations from prior work (Ruggeri et al., 2024). However, when juxtaposed with the generated edits and evaluated by NLP researchers based solely on the edit definitions, the generated annotations' stricter compliance with the guidelines highlights small flaws in the human edits that would not necessarily be evident if one only saw those edits in isolation.

	CondaQA	% Valid (Annotators)	% Valid (Author)
Gen.	Paraphrase	77.4	90.3
	Scope	64.5	87.1
	Affirmative	83.9	93.5
	Questions	64.5	77.4
Orig.	Paraphrase	64.5	-
	Scope	61.3	-
	Affirmative	90.3	-
	Questions	58.1	-
	DROP		
Gen.	Compositional	78.1	53.1
	Questions	81.2	84.4
Orig.	Compositional	84.4	-
	Questions	78.1	-

Table 2: The validity assessments for questions and edits from the human preferences study. Each row header states the dataset and edit type evaluated. The first column shows the percentage of total edits or questions evaluated which were marked as valid by annotators. All edits and questions evaluated by annotators were also evaluated by an author of this paper. Those results are shown in the second column.

6 Which Benchmark is More Difficult?

So far, we have observed that an LLM can often generate valid complex questions and contrastive edits along specific dimensions. Our goal was to assess the automation of creating *challenging* reasoning benchmarks. However, validity according to the instructions does not necessarily imply difficulty, as hardness often emerges organically from creative and diverse annotations. To this end, we compare model performance on datasets containing only generated annotations against those with only original human annotations.

[12]HTML templates for the study are available at our Github.
[13]We excluded temporal edits because we could identify only ≈20 human-authored ones.

	Win	Lose	% Tie	
			As good	As bad
CondaQA				
Paraphrase	32.3	19.3	35.5	12.9
(74.2%)	39.1	13.0	43.5	4.3
Scope	38.7	35.5	12.9	12.9
(61.3%)	57.9	21.0	21.0	0.0
Affirmative	20.0	19.3	48.4	3.2
(77.4%)	37.5	4.2	54.2	4.2
Questions	35.5	32.3	16.1	16.1
(48.4%)	53.3	20.0	6.7	20.0
DROP				
Compositional	18.8	34.4	43.8	3.1
(53.1%)	29.4	17.6	52.9	0.0
Questions	25.0	18.8	53.1	3.1
(75.0%)	25.0	8.3	63.6	0.0

Table 3: Human preference according to how well generations adhere to the definition of the CondaQA/DROP question/edit. Win/lose/tie refers to LLM-generated edits over human-authored. The percentages in the first column indicate the fraction of generations of that type for which the annotator and the author agree on validity. Every second row presents preferences considering only those unanimously valid generations where both the author and the annotator agree.

Setup. Generated questions may be ambiguous. To filter unambiguous questions for benchmarking, we collect answers from three annotators per question and keep those for which two annotators agree. As NLP expertise is not required here, we recruit 59 annotators through Prolific, all screened using examples with known answers. Screenshots of the answering interfaces are shown in Fig. 13 for CondaQA and Fig. 14–18 for DROP. In total, we collected 1242 answers across two datasets at a cost of $1744.

We use all CondaQA-generated bundles that one author finds valid across the question and all edit types. For benchmarking on original data, we use CondaQA dev questions corresponding to the same passages that have answers in all four passages of their bundle.[14] For DROP, we generate questions and compositional edits from passages in the contrast set with human compositional edits, and use question-edit pairs judged valid by one author. For benchmarking on original DROP, we first take compositional edits from the human-written dataset for all passages used to create the generated set. We then take the original test set questions from which those human-written compositional edits were created. We report question counts after ambiguity filtering in Table 7 in the Appendix.[15] We provide an illustration of our full synthetic benchmark creation process in Figure 3 in the Appendix.

LLMs can be biased when they are used both as a data generator and a task solver (Maheshwari et al., 2024), which causes inflated performance. Thus, we evaluate LLMs from 6 families on the original and generated data: `GPT-4-Turbo/GPT-4o`[16], `o3-mini`[17], `Claude-Opus-4.1`[18], `Gemini-2.5-Flash`[19], `Llama-3.3-70B`[20], `Qwen2.5-72b` (Team, 2024).

We evaluate CondaQA zero-shot and DROP in a 3-shot setting, following prior work. CondaQA is evaluated using accuracy, as most answers are “yes”, “no”, or “don’t know”. For DROP, we use token-level F1, the harmonic mean of precision and recall over overlapping tokens between the prediction and the gold answer. Both datasets are also evaluated using *consistency*, which is 1 if all instances in a bundle are answered correctly. For DROP, an instance is considered correct for consistency if its token-level F1 exceeds 0.8. CondaQA bundles consist of a question paired with four passages (“Full Bundle Consistency” in Table 4) or a single passage edited in a specific way. DROP bundles consist of a question and its compositional edit in the context of the same passage. We perform 5 runs of the original and generated datasets on each of the models tested. The reported data is the mean value across those 5 runs, and the standard deviation of the mean is shown in subscript. Statistical significance testing is performed to measure the significance of the differences between model performances on the generated and original datasets. We use a mixed effects model to account for variance in the scores across runs, and a one-tailed t-test with $p < .05$ as a threshold for significance.

Results. They are reported in Table 4, and show that LLMs’ accuracy/token-F1 and consistency tend to be higher in model-generated datasets than

[14] If multiple such questions exist, we select one at random.

[15] We manually check if 2 annotators refer to the same span with minor boundary disagreement. If so, the question is considered unambiguous, and one author sets the boundaries.

[16] https://openai.com/index/hello-gpt-4o/

[17] https://openai.com/index/openai-o3-mini/

[18] https://www.anthropic.com/claude/opus

[19] https://deepmind.google/models/gemini/flash/

[20] https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

	GPT-4-Turbo△		GPT-4o°		o3-mini†		Claude-Opus-4◇		Gemini-2.5-Flash○		Llama-3.3-70b‡		Qwen2.5-72b*	
	Orig.	*Gen.*	*Orig.*	*Gen.*	*Orig.*	*Gen.*	*Orig.*	*Gen.*	*Orig.*	*Gen.*	*Orig.*	*Gen.*	*Orig.*	*Gen.*
CondaQA														
Original-Data Accuracy	73.9$_{2.6}$	**79.7$_{1.2}$**	70.5$_{0.9}$	**78.3$_{1.4}$**	68.5$_{1.9}$	**83.4$_{1.4}$**	67.5$_{2.5}$	**80.3$_{0.9}$**	70.2$_{0.9}$	**81.7$_{1.4}$**	81.4$_{0.0}$	81.4$_{0.0}$	69.5$_{0.0}$	**78.3$_{0.8}$**
Full Bundle Consistency	40.5$_{2.7}$	**44.5$_{4.1}$**	32.6$_{1.6}$	**50.9$_{5.2}$**	40.5$_{2.1}$	**47.3$_{3.0}$**	41.9$_{2.8}$	**44.1$_{5.0}$**	**54.0$_{3.4}$**	39.5$_{2.6}$	**53.0$_{3.0}$**	40.9$_{2.3}$	40.9$_{2.1}$	**44.5$_{4.7}$**
Paraphrase Consistency	69.5$_{1.5}$	**73.9$_{1.0}$**	66.5$_{1.6}$	**72.5$_{1.0}$**	65.5$_{3.1}$	**78.6$_{1.8}$**	64.7$_{1.6}$	**72.1$_{3.0}$**	69.1$_{1.8}$	**72.5$_{2.7}$**	**78.9$_{1.6}$**	77.1$_{1.5}$	70.2$_{1.0}$	**74.6$_{1.5}$**
Scope Consistency	53.3$_{2.4}$	**67.8$_{1.7}$**	44.2$_{0.9}$	**64.5$_{3.1}$**	49.2$_{5.6}$	**65.3$_{2.0}$**	49.2$_{3.2}$	**62.0$_{4.0}$**	59.6$_{2.4}$	**64.5$_{1.1}$**	**60.0$_{2.7}$**	58.8$_{1.7}$	49.2$_{2.4}$	**59.6$_{3.4}$**
Affirmative Consistency	55.9$_{4.0}$	**56.2$_{2.1}$**	46.9$_{3.2}$	**63.8$_{1.6}$**	51.8$_{2.3}$	**61.9$_{3.4}$**	47.8$_{2.7}$	**56.2$_{3.4}$**	**59.6$_{3.4}$**	53.1$_{3.2}$	**68.6$_{1.8}$**	60.0$_{0.9}$	54.7$_{1.7}$	**58.1$_{2.1}$**
DROP														
Original-Data Token F1	61.7$_{2.1}$	**83.0$_{3.1}$**	62.8$_{2.6}$	**83.9$_{1.7}$**	75.2$_{4.1}$	**92.7$_{1.3}$**	73.2$_{0.8}$	**90.3$_{1.4}$**	71.9$_{1.8}$	**90.9$_{0.0}$**	52.8$_{2.0}$	**75.1$_{2.8}$**	57.1$_{1.8}$	**77.9$_{2.0}$**
Compositional Consistency	30.4$_{3.6}$	**58.8$_{2.7}$**	29.2$_{3.0}$	**63.6$_{0.9}$**	57.2$_{3.0}$	**80.8$_{3.0}$**	55.6$_{2.6}$	**82.0$_{2.4}$**	58.0$_{2.4}$	**86.4$_{1.7}$**	25.2$_{1.1}$	**35.2$_{3.3}$**	23.2$_{2.3}$	**48.4$_{3.6}$**

Table 4: Comparison of difficulty of the original CondaQA/DROP samples and their generated version measured by the performance with human-authored questions and edits against generated questions and edits. One author of this paper selected instances with valid generated questions and edits for this comparison. Human annotators provided answers for all of the valid generated questions and edits, and only those with at least 2/3 agreement were used for evaluation. Cells highlighted in blue indicate cases where scores on generated data were found to be significantly higher than the scores on the original data, and cells highlighted in yellow indicated cases where scores on original data were found to be significantly higher ($p < .05$). △`gpt-4-turbo-2024-04-09` °`gpt-4o-2024-11-20` ◇`claude-opus-4-1-20250805` ○`gemini-2.5-flash` †`o3-mini-2025-01-31` ‡`meta-llama/Llama-3.3-70B-Instruct` *`Qwen/Qwen2.5-72B-Instruct`

in human-authored counterparts. Four out of six model families show higher accuracy/token-F1 and consistency across the board on the generated dataset, with most scores showing a significant increase in model performance. `Llama-3.3-70B` and `Gemini-2.5-Flash` on the CondaQA dataset have higher scores on the human-authored datasets, with full bundle consistency and affirmative edit consistency scores on the original dataset being significantly higher than their generated counterparts. Both models have much lower accuracy in answering generated affirmative edits as compared to human-authored edits. Looking into the evaluation logs, we find no discernible pattern as to why generated affirmative edits are harder for these models. However, these are the exceptions to the general trend. Looking at DROP scores, the generated dataset and compositional edits are shown to be significantly easier than the original DROP dataset across all models. That is, generated versions are generally easier. We acknowledge that the results for the original DROP sample seem low. This is not due to a bug; we provide a detailed discussion in Appendix D.

Moreover, LLM-authored datasets do not preserve the relative ranking of models in CondaQA. That is, synthetic evaluations currently cannot reliably serve as a substitute for human-authored ones, even when used solely for model comparison.

Note that this occurs despite model-generated instances often being rated as more valid by humans. This finding suggests that even though generated datasets may closely follow data specifications, and appear to be high-quality to humans, they may feature systematic patterns or lack the creativity of human-authored datasets — thus making them more easily solvable by models. For example, we observe that the human-authored questions in CondaQA tend to be considerably longer (Figure 4).

7 Conclusion

We ask how benchmarks with LLM-generated content compare to human-authored versions. We summarize our main findings and conclusions here, and provide a more detailed discussion in Appendix E.

We find that while generated content is often valid (§4,§5), it tends to produce easier benchmarks (§6). Further, we find that these synthetically-generated datasets do not necessarily preserve model ordering, which could consequently lead practitioners to make different decisions around model development and deployment. Finally, we also find that these quality differences between synthetic and human-authored data, may not be perceptible to researchers inspecting the data. Thus, while LLMs are promising for data creation where complexity is less critical, reliable human annotators remain vital for new benchmarks assessing real-world generalization, corner cases, and testing models on nuanced, complex scenarios.

8 Limitations

Despite the contributions of our work, there are a few limitations to consider.

First, we focus only on two reasoning-over-text benchmarks, but seven distinct generation objectives: implications, paraphrasing, changing negation scope, identifying and undoing negation, discrete reasoning, adding intermediate reasoning steps, and altering the temporal order of events. We were constrained by: (i) needing to become deeply familiar with intricate annotation guidelines, (ii) identifying key prompting ideas that ended up being dataset-specific (e.g., reasoning step types for DROP and changing the negation scope through a QA instance generation), (iii) reading passages in these datasets over a thousand times ourselves, (iv) recruiting expert annotators who were either pursuing or had completed a doctoral degree in NLP, (v) designing annotation templates, and (vi) the annotation cost of thousands of dollars.

Second, despite extensive prompting, we may have missed strategies that yield more valid synthetic annotations. Likewise, another LLM might produce more valid generations than `gpt-4-turbo-2024-4-09`. However, greater validity does not imply greater difficulty, as our comparison of human and generated benchmarks shows. Critically, we first had to discover that valid generations (beyond labels) can still be notably easier. Now that this is known, improving difficulty would require further rounds of prompt review, costly answer annotation, and manual validity review. A related open question is whether our structured prompts — designed to decompose the task into subtasks with subtask-specific verifiers — improved validity at the cost of difficulty. These prompts were crucial for producing valid generations, yet may have also made the task easier for LLMs. This further highlights the challenge of balancing validity with difficulty.

Finally, in this work, we look at two key metrics for benchmark quality: validity and complexity. However, machine-generated and human-generated benchmarks could differ on several other tertiary aspects, which are nevertheless important for building high-quality evaluations, such as diversity, coverage, and representativeness. We leave it to future work to compare these data-creation paradigms on these additional axes.

Acknowledgments

We thank Jacob K. Johnson for the initial codebase and Dheeru Dua for sharing information about (contrastive) DROP. We are also grateful to the annotators for their contributions and to the UtahNLP lab for valuable feedback. Abhilasha Ravichander has been supported by NSF DMS-2134012 and ONR N00014-24-1-2207 while working on this project.

References

Kareem Amin, Sara Babakniya, Alex Bie, Weiwei Kong, Umar Syed, and Sergei Vassilvitskii. 2025. Escaping collapse: The strength of weak data for large language model training. *Preprint*, arXiv:2502.08924.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback. *Preprint*, arXiv:2212.08073.

Nishant Balepur, Feng Gu, Abhilasha Ravichander, Shi Feng, Jordan Lee Boyd-Graber, and Rachel Rudinger. 2025. Reverse question answering: Can an LLM write a question so hard (or bad) that it can‘t answer? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 44–64, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

Jan Cegin, Jakub Simko, and Peter Brusilovsky. 2023. ChatGPT to replace crowdsourcing of paraphrases for intent classification: Higher diversity and comparable model robustness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1889–1905, Singapore. Association for Computational Linguistics.

Zeming Chen, Qiyue Gao, Antoine Bosselut, Ashish Sabharwal, and Kyle Richardson. 2023. DISCO: Distilling counterfactuals with large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5514–5528, Toronto, Canada. Association for Computational Linguistics.

Debarati Das, Karin De Langis, Anna Martin-Boyle, Jaehyung Kim, Minhwa Lee, Zae Myung Kim, Shirley Anugrah Hayati, Risako Owan, Bin Hu, Ritik Parkar, Ryan Koo, Jonginn Park, Aahan Tyagi, Libby Ferland, Sanjali Roy, Vincent Liu, and Dongyeop Kang. 2024. Under the surface: Tracking the artifactuality of llm-generated data. *Preprint*, arXiv:2401.14698.

Tanay Dixit, Bhargavi Paranjape, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. CORE: A retrieve-then-edit framework for counterfactual data generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2964–2984, Abu Dhabi, United Arab Emirates. 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.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30).

Shasha Guo, Lizi Liao, Cuiping Li, and Tat-Seng Chua. 2024. A survey on neural question generation: Methods, applications, and prospects. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pages 8038–8047, Jeju, South Korea.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Chetan Harsha, Karmvir Singh Phogat, Sridhar Dasaratha, Sai Akhil Puranam, and Shashishekar Ramakrishna. 2025. Synthetic data generation using large language models for financial question answering. In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 76–95, Abu Dhabi, UAE. Association for Computational Linguistics.

Tom Hosking and Mirella Lapata. 2021. Factorising meaning and form for intent-preserving paraphrasing. 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 1405–1418, Online. Association for Computational Linguistics.

Mohammad Javad Hosseini, Andrey Petrov, Alex Fabrikant, and Annie Louis. 2024. A synthetic data approach for domain generalization of NLI models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2212–2226, Bangkok, Thailand. Association for Computational Linguistics.

Divyansh Kaushik, Eduard H. Hovy, and Zachary Chase Lipton. 2020. Learning the difference that makes A difference with counterfactually-augmented data. In *Proceedings of the 8th International Conference on Learning Representations, (ICLR)*.

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.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Saurabh Kulshreshtha and Anna Rumshisky. 2023. Reasoning circuits: Few-shot multi-hop question generation with structured rationales. In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pages 59–77, Toronto, Canada. Association for Computational Linguistics.

Chuanrong Li, Lin Shengshuo, Zeyu Liu, Xinyi Wu, Xuhui Zhou, and Shane Steinert-Threlkeld. 2020. Linguistically-informed transformations (LIT): A method for automatically generating contrast sets. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 126–135, Online. Association for Computational Linguistics.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11065–11082, Bangkok, Thailand. Association for Computational Linguistics.

Gaurav Maheshwari, Dmitry Ivanov, and Kevin El Haddad. 2024. Efficacy of synthetic data as a benchmark. *Preprint*, arXiv:2409.11968.

Kathleen R. McKeown. 1983. Paraphrasing questions using given and new information. *American Journal of Computational Linguistics*, 9(1):1–10.

Anders Giovanni Møller, Arianna Pera, Jacob Dalsgaard, and Luca Aiello. 2024. The parrot dilemma: Human-labeled vs. LLM-augmented data in classification tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 179–192, St. Julian's, Malta. Association for Computational Linguistics.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. 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.

Van Bach Nguyen, Christin Seifert, and Jörg Schlötterer. 2024. CEval: A benchmark for evaluating counterfactual text generation. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 55–69, Tokyo, Japan. Association for Computational Linguistics.

Liangming Pan, Wenqiang Lei, Tat-Seng Chua, and Min-Yen Kan. 2019. Recent advances in neural question generation. *Preprint*, arXiv:1905.08949.

Parth Patwa, Simone Filice, Zhiyu Chen, Giuseppe Castellucci, Oleg Rokhlenko, and Shervin Malmasi. 2024. Enhancing low-resource LLMs classification with PEFT and synthetic data. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6017–6023, Torino, Italia. ELRA and ICCL.

Meltem Kurt Pehlivanoglu, Robera Tadesse Gobosho, Muhammad Abdan Syakura, Vimal Shanmuganathan, and Luis de-la-Fuente-Valentín. 2024. Comparative analysis of paraphrasing performance of chatgpt, gpt-3, and T5 language models using a new chatgpt generated dataset: Paragpt. *Expert Syst. J. Knowl. Eng.*, 41(11).

Chengwen Qi, Ren Ma, Bowen Li, He Du, Binyuan Hui, Jinwang Wu, Yuanjun Laili, and Conghui He. 2025. Large language models meet symbolic provers for logical reasoning evaluation. In *Proceedings of The Thirteenth International Conference on Learning Representations (ICLR)*.

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.

Abhilasha Ravichander, Matt Gardner, and Ana Marasovic. 2022. CONDAQA: A contrastive reading comprehension dataset for reasoning about negation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8729–8755, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Melissa Roemmele and Andrew Gordon. 2024. From test-taking to test-making: Examining LLM authoring of commonsense assessment items. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5193–5203, Miami, Florida, USA. Association for Computational Linguistics.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. QA dataset explosion: A taxonomy of NLP resources for question answering and reading comprehension. *ACM Comput. Surv.*, 55(10):197:1–197:45.

Alexis Ross, Ana Marasović, and Matthew Peters. 2021. Explaining NLP models via minimal contrastive editing (MiCE). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3840–3852, Online. Association for Computational Linguistics.

Alexis Ross, Tongshuang Wu, Hao Peng, Matthew Peters, and Matt Gardner. 2022. Tailor: Generating and perturbing text with semantic controls. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3194–3213, Dublin, Ireland. Association for Computational Linguistics.

Federico Ruggeri, Eleonora Misino, Arianna Muti, Katerina Korre, Paolo Torroni, and Alberto Barrón-Cedeño. 2024. Let guidelines guide you: A prescriptive guideline-centered data annotation methodology. *Preprint*, arXiv:2406.14099.

Vasile Rus, Brendan Wyse, Paul Piwek, Mihai Lintean, Svetlana Stoyanchev, and Christian Moldovan. 2010. The first question generation shared task evaluation challenge. In *Proceedings of the 6th International Natural Language Generation Conference*. Association for Computational Linguistics.

Vinay Samuel, Houda Aynaou, Arijit Chowdhury, Karthik Venkat Ramanan, and Aman Chadha. 2024. Can LLMs augment low-resource reading comprehension datasets? opportunities and challenges. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 307–317, Bangkok, Thailand. Association for Computational Linguistics.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *Proceedings of the 5th International Conference on Learning Representations, (ICLR)*.

Samira Shaikh, Thiago Ferreira, and Amanda Stent, editors. 2022. *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*. Association for Computational Linguistics, Waterville, Maine, USA and virtual meeting.

Ilia Stepin, José Maria Alonso, Alejandro Catalá, and Martin Pereira-Fariña. 2021. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9:11974–12001.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Junior Cedric Tonga, Benjamin Clement, and Pierre-Yves Oudeyer. 2025. Automatic generation of question hints for mathematics problems using large language models in educational technology. In *Proceedings of Large Foundation Models for Educational Assessment*, volume 264 of *Proceedings of Machine Learning Research*, pages 61–102. PMLR.

Asahi Ushio, Fernando Alva-Manchego, and Jose Camacho-Collados. 2022. Generative language models for paragraph-level question generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 670–688, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Yongjie Wang, Xiaoqi Qiu, Yu Yue, Xu Guo, Zhiwei Zeng, Yuhong Feng, and Zhiqi Shen. 2024. A survey on natural language counterfactual generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4798–4818, Miami, Florida, USA. 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.

Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D. Hwang, Liwei Jiang, Jillian Fisher, Abhilasha Ravichander, Khyathi Chandu, Benjamin Newman, Pang Wei Koh, Allyson Ettinger, and Yejin Choi. 2024. The generative AI paradox: "what it can create, it may not understand". In *Proceedings of The Twelfth International Conference on Learning Representations (ICLR)*.

Sam Witteveen and Martin Andrews. 2019. Paraphrasing with large language models. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 215–220, Hong Kong. Association for Computational Linguistics.

Kun Wu, Lijie Wang, Zhenghua Li, Ao Zhang, Xinyan Xiao, Hua Wu, Min Zhang, and Haifeng Wang. 2021. Data augmentation with hierarchical SQL-to-question generation for cross-domain text-to-SQL parsing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8974–8983, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tongshuang Wu, Haiyi Zhu, Maya Albayrak, Alexis Axon, Amanda Bertsch, Wenxing Deng, Ziqi Ding, Boyuan Guo, Sireesh Gururaja, Tzu-Sheng Kuo, Jenny T. Liang, Ryan Liu, Ihita Mandal, Jeremiah Milbauer, Xiaolin Ni, Namrata Padmanabhan, Subhashini Ramkumar, Alexis Sudjianto, Jordan Taylor, Ying-Jui Tseng, Patricia Vaidos, Zhijin Wu, Wei Wu, and Chenyang Yang. 2025. Llms as workers in human-computational algorithms? replicating crowdsourcing pipelines with llms. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, CHI EA 2025, Yokohama, Japan, 26 April 2025- 1 May 2025*, pages 684:1–684:10. ACM.

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.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. ZeroGen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Asaf Yehudai, Boaz Carmeli, Yosi Mass, Ofir Arviv, Nathaniel Mills, Assaf Toledo, Eyal Shnarch, and Leshem Choshen. 2024. Genie: Achieving human parity in content-grounded datasets generation. In *Proceedings of The Twelfth International Conference on Learning Representations (ICLR)*.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024. Metamath: Bootstrap your own mathematical questions for large language models. In *Proceedings of The Twelfth International Conference on Learning Representations (ICLR)*.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. ReClor: A reading comprehension dataset requiring logical reasoning. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

Appendix Overview

In this Appendix, we provide:

- A description of prompting ideas that we found to be useful for generating valid benchmarks in Appendix A.
- A contamination analysis of generated content in Appendix B.
- Details of the preference study in Appendix C.
- A discussion of discrepancies between reported DROP results and our small-sample DROP results in Appendix D.
- A more detailed discussion of findings and takeaways in Appendix E.
- A version of Table 1 and Table 4, where instead of prompting the model to answer questions in the context of all related passages in a single chat, we use separate chats for each passage-question pair. See: Table 5.
- Examples of human-authored questions and edits in Tables 6, 8, and 9.
- Tables 12–13: Tabular versions of Fig. 1–2 (§4).
- The validity criteria for each annotation item in Table 11.
- Screenshots of the preference study (§5) in Fig. 5–12 for CondaQA and Fig. 19–26 for DROP.
- Screenshots of answer crowdsourcing (§6) in Fig. 13 for CondaQA and Fig. 14–18 for DROP.
- Table 7: Question counts after ambiguity filtering in §6.
- Figure 3: Illustration of our synthetic benchmark creating process. Question and edit generation discussed in §4, label creation discussed in §6.
- Figure 4: Distribution of lengths of model- vs. human-authored questions. Discussed in §6.
- Examples of the LLM-generated and human-authored questions in the two datasets: DROP (Table 14) and CondaQA (Table 15).
- Examples agreement and disagreement cases between annotators and one author of this paper for generated instances of CondaQA and DROP questions and edits in Tables 16-21.

A Overview of Key Prompting Ideas

While our main goal was not to develop a novel data generation strategy, our prompting yielded several practical insights:

- Human annotation guidelines make poor prompts.
- Improving filtering is promising.
- Prompting strategies that succeed for one dataset may fail to work on another.
- Decomposing the annotation into sub-tasks was crucial.

We outline some of the prompting strategies we found most useful, and the data they were used to create in Table 10.

B Contamination Analysis

Our goal is to compare benchmarks with LLM-generated content to their counterparts with human-authored content. But what if generations are actually memorized human instances from the original dataset — a problem known as *dataset contamination*. This could falsely lead us to conclude that synthetically-generated data is comparable to human-authored data, but these conclusions would then not generalize to new, unseeen datasets.

To assess this, we run a qualitative analysis of potential contamination in the generated questions by comparing them to the original human-authored datasets. We randomly sampled 50 generated CondaQA questions and 50 for DROP. We use the generation created by the corresponding best prompts. For each, we *manually* compared the generated question to all human-written questions associated with the same passage in the original dataset, and marked cases where the generated question was a paraphrase of an original. For the CondaQA dataset, we found that none of the 50 sampled questions were paraphrases of any of the questions from the original dataset. For DROP, we found that 9/50 of the sampled generated questions were paraphrases of questions from the original dataset.[21] From these results, we conclude that *the original datasets are unlikely to be a significant source of contamination for the model-generated data.*

C Details of the User Study in §5

Screenshots of the preference study are shown in Figures 5–12 for CondaQA and Figures 19–26 for DROP. All participants in the study first provided informed consent.

In the DROP preference study, annotators first choose between two questions written for a given passage, and then choose between two edits. However, these edits are based on a different passage. This mismatch arises because we reuse the generated questions from §4, but not the compositional edits. In §4, compositional edits are made

[21] One note with DROP is that many human-annotated questions were created for each passage, up to 70 for some.

on generated questions to simulate building a synthetic dataset from scratch. To enable a valid comparison between human and generated compositional edits in §5, we instead apply edits to human-authored questions that also have corresponding human-authored compositional edits.

For CondaQA, this complication does not occur. Annotators are shown pairs of questions and edits based on the same passage, and they express preferences directly over these pairs.

D DROP Sample Difficulty Sanity Checks

We observe that the results on the sample of the original, human-authored DROP dataset in Table 4 are notably lower than reported DROP results. To verify our evaluation setup for DROP, we benchmark `gpt-4o-2024-11-20` on a larger sample of 1,000 passage-question pairs from the original DROP test set. We obtain a token-F1 score of 83.1, which is in line with the reported token-F1 score for DROP for `gpt-4o-2024-11-20` (81.5)[22]. This confirms that the low performance on the smaller sample is not due to a bug in our evaluation scripts.

We reiterate here that we did not choose the DROP sample for Table 4, but use the sample of instances which have human compositional edits from the DROP contrastive set.

One explanation for the difficulty of this sample could be the high ratio of questions with numerical answers as compared to the original DROP test set (77% vs 61.6%). Token-F1 for numerical answers reduces to the exact match to the gold label, which is stricter than the token-F1 evaluation for spans or dates. However, the ratio of numerical questions in the generated data is actually higher than the original sample (96.6%), so we reject the hypothesis that the ratio of numerical questions alone is why the human-authored DROP benchmark in Table 4 is harder than the generated one.

Moreover, for CondaQA, we do not observe the same discrepancy between performance on smaller and larger samples. Specifically, the full dev set results in Table 1 for `o3-mini-2025-01-31` and `meta-llama/Llama-3.3-70B-Instruct` line up with the results for the original data in Table 4.

E Discussion and Practical Advice

Quality Tradeoffs of Automatic Evaluation. Using LLMs to generate datasets and validating them through sampling misses crucial quality metrics. Such assessments can miss out on other factors such as complexity and representativeness. Specifically, although generations may be valid, and may even appear preferable to human judges, we might not even realize what is missing: the nuances, complexity, and creativity of human-authored data.

Cost of Constructing Evaluations. For constructing benchmarks in this study, LLMs are orders of magnitude cheaper than crowdsourcing. We recommend synthetic evaluations for tasks where complexity and representativeness are not the primary criteria for measuring dataset quality. For example, models could be utilized to generate cheap and large-scale unit tests for models (Naik et al., 2018; Ribeiro et al., 2020), while human-constructed evaluations can be utilized complementarily for assessing real-world generalization, identifying corner cases, and validating model behavior on complex, nuanced scenarios.

Feedback Mechanisms. Yet another aspect of constructing high-quality evaluations is the ability to give feedback that influences the annotation process. In the crowdsourcing scenario, dataset creators often engage in two rounds of iterations — an initial "pilot study" to refine instructions based on annotator data, and often a second stage (once instructions are finalized) where crowdworkers are given feedback on their annotations (Nangia et al., 2021). Applying similar feedback loops to synthetic evaluations remains a challenge. While our prompt iterations resemble pilot studies, future work could explore feedback mechanisms — human or AI-driven — to help models refine their outputs.

On Human Effort while Seeking Difficulty. Our findings suggest that generating valid instances through prompting is not the main challenge — it is finding prompts that generate hard content. To find a better prompt for difficulty, in theory, one could iteratively prompt an LLM to generate candidate instances, have one or more people validate them, recruit three or more people per instance to get reliable answers, benchmark model performance on the resulting set, and repeat this until the difficulty level matches that of a human-authored dataset. Basically, repeat what we did in this paper many times. But, in practice, executing such a procedure would require substantial human-in-the-loop effort. More-

[22] https://github.com/openai/simple-evals

over, this turns evaluation-data generation into a dataset construction with an adversary in the loop that prior work has raised concerns about.

CondaQA	**Human**†	`o3-mini`$^{\triangle}$	`Llama-3.3`$^{\circ}$
Original-Data Accuracy	91.9	71.5	77.78
Bundle Consistency	81.6	40.3	36.2
Paraphrase Consistency	93.6	67.3	78.1
Scope Consistency	86.5	51.5	51.0
Affirmative Consistency	88.2	48.5	54.1
DROP	**Human**‡	**o3-mini**$^{\triangle}$	`Llama-3.3`$^{\circ}$
Original-Data Token F1	96.4	84.3	70.9
Consistency	N/A	53.5	35.5

Table 5: State-of-the-art performance on the CondaQA development set and the DROP contrastive test set. The original accuracy is calculated on the data used to create contrastive instances. Consistency reflects the rate at which all minimally different examples within a bundle are answered correctly. **Here, we use separate chats for each passage-question pair.** *These results show that these reasoning benchmarks remain challenging. We investigate whether they could be created using LLMs.* †Cf. Ravichander et al. (2022). ‡Cf. Dua et al. (2019). $^{\triangle}$`o3-mini-2025-01-31` $^{\circ}$`meta-llama/Llama-3.3-70B-Instruct`.

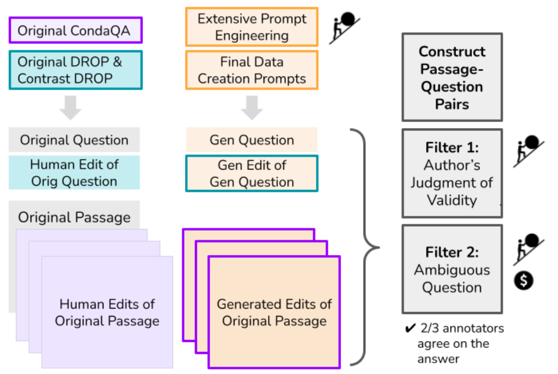


Figure 3: Illustration of our synthetic dataset creation process. Purple borders around a block indicates relevance to CondaQA, teal borders indicate relevance to DROP. All other blocks are relevant for both datasets.

Original	**Temporal Edit**
In the spring of 1625 the Spanish regained Bahia in Brazil and Breda in the Netherlands from the Dutch. **In the autumn** they repulsed the English at Cadiz.	**In the spring of 1625** the Spanish regained Bahia in Brazil and Breda in the Netherlands from the Dutch. **In winter the year earlier** they had repulsed the English at Cadiz.

Table 6: Example of human-authored temporal DROP-paragraph edit.

	Context	Count
CondaQA	is-valid(Gen Question, Original Passage) $\wedge$ is-valid(Gen Question, Gen x-Edit Passage) $\hookrightarrow \forall x \in$ {Paraphrase, Scope, Affirmative}	44
	is-valid(Gen Question, Original Passage) $\wedge$ is-valid(Gen Question, Gen Paraphrase-Edit Passage)	56
	is-valid(Gen Question, Original Passage) $\wedge$ is-valid(Gen Question, Gen Scope-Edit Passage)	49
	is-valid(Gen Question, Original Passage) $\wedge$ is-valid(Gen Question, Gen Affirmative-Edit Passage)	52
	is-valid(Gen Question, Original Passage)	59
DROP	is-valid(Original Passage, Gen Question) $\wedge$ is-valid(Original Passage, Gen Compositional Edit of Gen Question)	50
	is-valid(Original Passage, Gen Question)	66
	is-valid(Original Passage, Gen Compositional Edit of Generated Question)	52

Table 7: The count of generated instances that are unambiguous. The validity is determined by one of the authors.

Dataset	**Passage** (some parts shortened)	Question	Answer	Reasoning
DROP	In **2008**, at another auction at Christies, Ulrich sold a 1982 **Basquiat piece**, Untitled (Boxer), for **US $13,522,500** to an anonymous telephone bidder. **Another record price for a Basquiat painting was made in 2007**, when an untitled Basquiat work from 1981 **sold at Sothebys in New York for US$14.6 million**. In 2012, [...] That year, [...] was sold [...] for $16.3 million [...] A similar untitled piece [...] sold for £12.92 million [...] In 2013, [...] sold for $48.8 million [...] In 2016 an untitled piece sold at Christies for $57.3 million [...]	How many more dollars was the Basquiat sold in 2007 than in 2008? *Compositional Contrastive Question:* How many more dollars was the Basquiat sold in 2016 than in 2008 and 2007 combined?	$1,077,500 *Contrast Answer:* $29,177,500	Discrete Reasoning
CondaQA	The school song, ‘Fair Reed,’ is sung to the tune of the 1912 popular song ‘Believe Me, if All Those Endearing Young Charms.’ It may be imitative of the Harvard anthem "Fair Harvard," which is also sung to the tune of ‘Believe Me, if All Those Endearing Young Charms.’ **It was composed by former president William Trufant Foster shortly after Reed’s founding, and is rarely heard today.**	Assuming that President Trufant Foster had been able and willing to copyright his composition, would his estate be receiving any noticeable amount of royalties from the song today?	No	Implications of Negation

Table 8: Example questions and answers from DROP and CondaQA, showing examples of valid questions and the model capabilities they probe (discrete reasoning, and reasoning about the implications of negated statements).

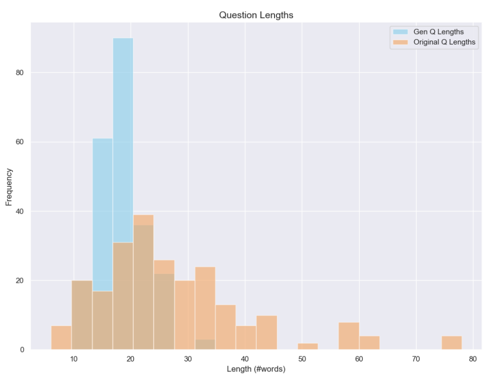


Figure 4: Distribution of lengths of model-generated questions vs human-authored questions. We find that on average generated questions are 18.35 words, and the human-authored questions were 27.17 words long. We also find that 26.3% of generated questions are more than 20 words long, but 67.68% of human-authored questions were more than 20 words long.

Original	Paraphrase Edit	Scope Edit	Affirmative Edit
It was composed by former president William Trufant Foster shortly after Reed's founding, and is **rarely** heard today.	Shortly after Reed College was founded, former president William Trufant Foster composed the song, which is **seldom** heard today.	It was composed by former president William Trufant Foster shortly after Reed's founding, and despite its current popularity, it was **rarely** heard in Foster's time.	It was composed by former president William Trufant Foster shortly after Reed's founding, and was an immediate success, remaining in use by the college to this day.

Table 9: Example of human-authored CondaQA edits. We omit the beginning of the paragraphs for space reasons.

Prompting Strategy	**Description**	**Data**
Overgeneration	Generating more examples than will be retained in the final dataset	CondaQA Questions, CondaQA Edits, DROP Questions, DROP Edits
Filters and Verifiers	Using LLMs or other automated methods to assess generated examples for specific quality criteria, and filter out examples that do not meet those criteria	CondaQA Questions, CondaQA Edits, DROP Questions, DROP Edits
Reasoning Step Types	Prompt includes examples that capture the reasoning type in question, model is asked to generate an example that encapsulates that type	DROP Questions, DROP Edits
Human Guidelines	Use annotation instructions given to human annotators	CondaQA Questions, CondaQA Edits, DROP Questions, DROP Edits
Decomposing	Breaking down the annotation task into sub-tasks an LLM can follow	CondaQA Questions, CondaQA Edits, DROP Questions, DROP Edits
QA Generation	Help an LLM understand whether a piece of information changed by having it generate questions for which answers should change if the change was made	CondaQA scope edit

Table 10: An overview of prompting strategies used in the study (Prompting Strategy), descriptions of the prompting ideas (Description), and the datasets they were used in generating (Data).

Annotation Item	Validity Criteria
CONDAQA	
Question	The question (i) targets the negated statement, rather than other information in the provided passage, and (ii) is about an implication of the negated statement.
Paraphrase Edit	The edit (i) yields a passage with the same meaning as the original passage, (ii) is coherent, and (iii) does not include the original negation cue.
Scope Edit	The edit (i) changes what is being negated by the negation cue, and (ii) is coherent.
Affirmative Edit	The edit (1) changes the passage such that what was being negated in the original passage is no longer negated, and (2) is coherent.
DROP	
Question	The question (i) requires discrete reasoning over the passage and (ii) is not directly answerable from the passage.
Compositional Edit	The edit adds additional reasoning step(s) to the original question.
Temporal Edit	The edit modifies either the order in which events appear in the original passage or the dates associated with each event to swap their temporal order.

Table 11: Validity criteria for each annotation item.

	Prompting Iteration Note	**% Valid**
CondaQA		
↳ Iteration 1 (16)	Full-bundle CondaQA instructions for people	18.7
↳ Iteration 2 (16)	Question-only CondaQA instructions for people	25.0
↳ Iteration 3 (16)	Started from scratch	37.5
↳ Iteration 4 (16)	+ Overgenerate & Filter questions that don't target negation, aren't about implication, & have no valid answer type	68.7
↳ Iteration 5 (16)	+ Filter questions that restate negation	81.2
↳ Iteration 6 (16)	+ Add reasoning type examples to the prompt	37.5
↳ Final Evaluation ⇒ Iteration 5 (114)		72.8
DROP		
↳ Iteration 1 (16)	Question-only DROP instructions for people	62.5
↳ Iteration 2 (16)	Started from scratch	25.0
↳ Iteration 3 (16)	+ Overgenerate & Filter simple questions	31.3
↳ Iteration 4 (16)	+ Filter compound question and those with a wrong answer type	43.8
↳ Iteration 5 (16)	+ Add discrete reasoning question types to the prompt	75.0
↳ Iteration 6 (16)	+ Filter question that are unanswerable without extra knowledge	100.0
↳ Final Evaluation ⇒ Iteration 6 (115)		88.7

Table 12: The percentage of valid questions in a sample of 16 for intermediate evaluations, and 100 or 114 for final evaluations. Questions are generated with `gpt-4-turbo-2024-04-09` and assessed by one author of this paper.

CondaQA	Prompting Iteration Note	% Valid
Paraphrase		
↳ Iteration 1 (16)	Full-bundle CondaQA instructions for people	75.0
↳ Iteration 2 (16)	Paraphrase-only CondaQA instructions for people	93.8
↳ Iteration 3 (16)	Started from scratch	100.0
↳ Iteration 4 (16)	+ Overgenerate & Filter meaning-changing and incoherent edits	100.0
↳ Final Evaluation ⇒ Iteration 4 (114)		96.5
Affirmative		
↳ Iteration 1 (16)	Full-bundle CondaQA instructions for people	62.5
↳ Iteration 2 (16)	Affirmative-only CondaQA instructions for people	81.3
↳ Iteration 3 (16)	Started from scratch	75.0
↳ Iteration 4 (16)	+ Overgenerate & Filter paraphrasing and incoherent edits	81.3
↳ Iteration 5 (16)	+ Identify then undo negation	100.0
↳ Final Evaluation ⇒ Iteration 5 (114)		98.2
Scope		
↳ Iteration 1 (16)	Full-bundle CondaQA instructions for people	0.0
↳ Iteration 2 (16)	Scope-only CondaQA instructions for people	25.0
↳ Iteration 3 (16)	Started from scratch	50.0
↳ Iteration 4 (16)	+ Few-shot check incoherent passages	62.5
↳ Iteration 5 (16)	- Remove fewshot prompting in coherency check	62.5
↳ Iteration 6 (16)	+ Change scope via QA generation	75.0
↳ Iteration 7 (16)	+ Wording changes	75.0
↳ Iteration 8 (16)	+ Tweak task instructions	68.7
↳ Iteration 9 (16)	- Remove coherency check	68.7
↳ Iteration 10 (16)	+ Overgenerate & Filter failed scope change checked with QA Filter incoherent passages	87.5
↳ Final Evaluation ⇒ Iteration 10 (114)		88.6
DROP	**Iteration Note**	**% Valid**
Compositional		
↳ Iteration 1 (16)	Started from scratch	12.5
↳ Iteration 2 (16)	+ Overgenerate & Filter edits that do not add steps Filter compound or incoherent edit-questions	43.7
↳ Iteration 3 (16)	+ Wording changes	43.7
↳ Iteration 4 (16)	+ Include reasoning step types in the prompt & improve the number of reasoning steps filter	81.2
↳ Final Evaluation ⇒ Iteration 4 (100)		65.0
Temporal		
↳ Iteration 1 (16)	Started from scratch + Overgenerate & Filter incoherent edits + equal event order edits	87.5
↳ Final Evaluation ⇒ Iteration (100)		86.0

Table 13: The percentage of valid edits in a sample of 16 for intermediate evaluations, and 100 or 114 for final evaluations. Edits are generated with `gpt-4-turbo-2024-04-09` and assessed by one author of this paper.

Dataset	Passage	Human-Authored Question	LLM-Generated Question
DROP	To open the regular season, the Lions hosted the St. Louis Rams. The Rams took an early lead when Greg Zuerlein kicked a 48-yard field goal, the only scoring drive of the first quarter. The Lions responded in the second quarter when Joique Bell ran in a touchdown from 1 yard out. St. Louis made it a 1-point game when Greg Zuerlein kicked a 29-yard field goal, and took the lead again when Cortland Finnegan intercepted a Kevin Smith pass and ran it back 31 yards for a touchdown. The final points of the half were scored when Detroit's Jason Hanson kicked a 41-yard field goal. The only points of the third quarter was a 45-yard field goal by Jason Hanson to tie the game. In the final quarter, the score continued to sea-saw. First the Rams' Brandon Gibson caught a 23-yard touchdown pass from Sam Bradford to take the lead back again, but the Lions soon tied it back up when Kevin Smith ran in a touchdown from 5 yards out. The Rams then went up by 3 points when kicked Greg Zuerlein his third field goal of the game, this one from 46 yards out. In the final 2 minutes, the Lions completed their third 80-yard scoring drive with a game-winning 5-yard touchdown catch by Kevin Smith, his second of the game. Greg Zuerlein's third field was the last of the game. The Lions were able to start 1-0 and capture their second straight regular season opening win.	How many more yards was the last touchdown than the first?	How many touchdowns did Kevin Smith score in total during this game?
DROP	Hoping to rebound from their road loss to the Jaguars, the Chargers went home for a Week 12 duel with the Baltimore Ravens. After a scoreless first quarter, San Diego struck first with kicker Nate Kaeding getting a 27-yard field goal. Afterwards, the Ravens took the lead with RB Willis McGahee getting a 1-yard TD run. Fortunately, the Chargers regained the lead with QB Philip Rivers completing a 35-yard TD pass to TE Antonio Gates (with a failed PAT), Kaeding kicking a 46-yard field goal, Rivers completing a 5-yard TD pass to WR Chris Chambers, and Kaeding kicking a 41-yard field goal. In the third quarter, San Diego increased its lead with Rivers and Gates hooking up with each other again on a 25-yard TD pass. Baltimore would manage to one final score as Ravens QB Kyle Boller completed a 13-yard TD pass to FB Le'Ron McClain. Fortunately, in the fourth quarter, the 'Bolts sealed the win with Kaeding nailing a 41-yard field goal. RB LaDainian Tomlinson (24 carries for 77 yards) became the fourth-fastest player and the 23rd player in NFL history to get 10,000 career rushing yards.	How many yards did Rivers complete?	What was the total distance of all field goals made by Nate Kaeding in the game?
DROP	Hoping to rebound from their road loss to the Jaguars, the Chargers went home for a Week 12 duel with the Baltimore Ravens. After a scoreless first quarter, San Diego struck first with kicker Nate Kaeding getting a 27-yard field goal. Afterwards, the Ravens took the lead with RB Willis McGahee getting a 1-yard TD run. Fortunately, the Chargers regained the lead with QB Philip Rivers completing a 35-yard TD pass to TE Antonio Gates (with a failed PAT), Kaeding kicking a 46-yard field goal, Rivers completing a 5-yard TD pass to WR Chris Chambers, and Kaeding kicking a 41-yard field goal. In the third quarter, San Diego increased its lead with Rivers and Gates hooking up with each other again on a 25-yard TD pass. Baltimore would manage to one final score as Ravens QB Kyle Boller completed a 13-yard TD pass to FB Le'Ron McClain. Fortunately, in the fourth quarter, the 'Bolts sealed the win with Kaeding nailing a 41-yard field goal. RB LaDainian Tomlinson (24 carries for 77 yards) became the fourth-fastest player and the 23rd player in NFL history to get 10,000 career rushing yards.	How many yards was the longest field goal of the game?	How many more yards did Nate Kaeding's second field goal in the game exceed his first field goal by?

Table 14: Examples of Human-Authored and LLM-Generated Questions for the DROP Dataset.

Passage	Human-Authored Question	LLM-Generated Question
The south celestial pole is visible only from the Southern Hemisphere. It lies in the dim constellation Octans, the Octant. **Sigma Octantis is identified as the south pole star, more than one degree away from the pole, but with a magnitude of 5.5 it is barely visible on a clear night.**	Would inexperienced sailors be able to easily use Sigma Octantis to determine their direction of travel within one compass degree of accuracy on a cloudless night?	Would Sigma Octantis be considered a reliable star for navigators to use under cloudy skies or in light-polluted areas?
It is difficult to distinguish an extinct volcano from a dormant (inactive) one. Dormant volcanoes are those that have not erupted for thousands of years, but are likely to erupt again in the future. Volcanoes are often considered to be extinct if there are no written records of its activity.,. **Before its catastrophic eruption of 1991, Pinatubo was an inconspicuous volcano, unknown to most people in the surrounding areas.** Two other examples are [...]	From the way Pinatubo is described in the passage, is it reasonable to assume that it was of fairly modest height, and/or lacking in the usual cone shape that is commonly associated with volcanoes?	Would Pinatubo likely have been a main point of interest on a geological tour before its eruption in 1991?
Much like elephants, manatees are polyphyodonts, and continuously replace their teeth from the back of the jaw. **Adults lack incisors, canines, and premolars, and instead have 8 to 10 cheek teeth in their mouth.** Manatees have an infinite supply of teeth moving in from the back and shedding in the front [...]	If the skull of an unknown animal is discovered to have premolars, is it possible that the skull belongs to an adult manatee?	If someone discovered a canine tooth, could it belong to an adult manatee?

Table 15: Examples of human-authored and LLM-generated questions for the **CondaQA** Dataset. The negated statement is highlighted in **bold** and the negation is highlighted in blue.

Validity Assessment	Passage	Original Question	LLM-Generated Compositional Edit	Author Note
Mutually Valid	The U.S. finance industry comprised only 10% of total non-farm business profits in 1947, but it grew to 50% by 2010. Over the same period, finance industry income as a proportion of GDP rose from 2.5% to 7.5%, and the finance industrys proportion of all corporate income rose from 10% to 20%. The mean earnings per employee hour in finance relative to all other sectors has closely mirrored the share of total U.S. income earned by the top 1% income earners since 1930. The mean salary in New York Citys finance industry rose from $80,000 in 1981 to $360,000 in 2011, while average New York City salaries rose from $40,000 to $70,000. In 1988, there were about 12,500 U.S. banks with less than $300 million in deposits, and about 900 with more deposits, but by 2012, there were only 4,200 banks with less than $300 million in deposits in the U.S., and over 1,800 with more.	How many more dollars was the average New York salaries in 2011 compared to average New York salaries in 1981?	What is the difference between the increase in the mean salary of New York City's finance sector and the average salary increase in New York City from 1981 to 2011?	The edit adds a reasoning step of calculating the mean salary increase of the finance sector and subtracting the average salary increase which was the required reasoning step in the original question.
Mutually Invalid	The Steelers traveled to Kansas City to take on the Chiefs. In the first quarter, the Chiefs scored first when Cairo Santos kicked a 30-yard field goal for a 3-0 lead. The Steelers would tie it up later on as Chris Boswell nailed a field goal from 24 yards out to make the game 3-3. In the second quarter, it was all Chiefs when Santos nailed 2 more field goals: from 27 and 22 yards out moving ahead 6-3 and the eventual halftime score of 9-3. After the break, the Chiefs moved ahead by double digits when Chircandrick West ran for a 1-yard touchdown for a 16-3 game. Backup QB Landry Jones was able to find Martavis Bryant on a 19-yard pass to make the score 16-10 for the Steelers later on in the third quarter. In the fourth quarter, the Steelers came within 3 when Boswell nailed another field goal from 36-yards out for a 16-13 game. However, the Chiefs eventually sealed the win when Alex Smith found Chris Conley on a 6-yard pass for a final score of 23-13. With the loss, the Steelers fell to 4-3 and also 2-2 without Roethlisberger as the starter. The defense failed to force any takeaways, but got inside the pocket and sacked Smith twice.	How many field goals did Cairo Santos make?	How many field goals did Cairo Santos make if the total number of field goals in the game was 5?	The edit adds in extra information but it does not require an additional reasoning step to answer the question. The reasoning steps are the same as the original question.
Author Valid + Annotator Invalid	A 2016 survey conducted by the Razumkov Centre found that 70% of the population declared themselves believers in any religion, while 6.3% declared themselves non-believers, and 2.7% declared to be atheists. Of the total Ukraian population, 81.9% declared to be Christians, comprising a 63.4% who declared to be Eastern Orthodox Church, 8.5% Ukrainian Greek Catholic Church, 7.1% simply christians, 1.9% Protestantism, and 1.0 Catholic Church. Islam comprises 1.1% of the population, while Judaism and Hinduism were the religions of 0.2% of the population each. A further 16.3% of the population believed in some other religion not identifying in one of those listed hitherto; it may comprise Rodnovery and other faiths.	How many percent of the total Ukraian population were not declared Christians?	What is the sum of the percentage of the total Ukrainian population that were declared non-Christians and those who declared themselves non-believers or atheists?	The edit, while adding reasoning steps, is no longer answerable. Since the percent of the population that are non-believers or atheists has some unknown overlap with the percent of the population that is non-christian, it is impossible to calculate with certainty the sum which the edit asks for. Annotators may have just noted the additional reasoning step and marked the edit as valid.

Table 16: Examples of LLM-generated compositional edits for DROP with various validity assessments from annotators and one of the authors of this paper. Compositional edits must add an additional reasoning step to the original question. The answer to a valid compositional edit must be either a span of text from the question or passage, a date, or a number.

Validity Assessment	Passage	LLM-Generated Question	Author Note
Mutually Valid	Coming off their impressive road win over the 49ers, the Falcons went home for a Week 6 Sunday night duel with the Chicago Bears. After a scoreless first quarter, Atlanta would trail early in the second quarter as Bears quarterback Jay Cutler found wide receiver Johnny Knox on a 23-yard touchdown pass. Afterwards, the Falcons took the lead as quarterback Matt Ryan completed a 40-yard touchdown pass to wide receiver Roddy White and a 10-yard touchdown pass to tight end Tony Gonzalez. After a scoreless third quarter, Chicago would tie the game in the fourth quarter with Cutler hooking up with tight end Greg Olsen on a 2-yard touchdown. Atlanta would regain the lead as running back Michael Turner got a 5-yard touchdown run. Afterwards, the defense would fend off a last-second Bears drive to lock up the victory.	By how many yards was Matt Ryan's touchdown pass to Roddy White longer than his touchdown pass to Tony Gonzalez?	The question is answerable numerically, requires complex reasoning (subtraction), and requires identifying multiple parts of the passage to answer. Since it meets all the requirements distinctly, this question is easily categorized as valid.
Mutually Invalid	Concerned by the Burmese consolidation of Shan states and Laotian states, China launched four invasions of Burma between 1765 and 1769. At first, the Qianlong Emperor envisaged an easy war, and sent in only the Green Standard troops stationed in Yunnan. The Qing invasion came as the majority of Burmese forces were deployed in their latest invasion of Siam. Nonetheless, battle-hardened Burmese troops defeated the first two invasions of 1765-1766 and 1766-1767 at the border. The regional conflict now escalated to a major war that involved military manoeuvres nationwide in both countries. The third invasion led by the elite Manchu Bannermen nearly succeeded, penetrating deep into central Burma within a few days' march from Ava. But the Bannermen of northern China could not cope with unfamiliar tropical terrains and lethal endemic diseases, and were driven back with heavy losses. After the close-call, Hsinbyushin redeployed his armies from Siam to the Chinese front. The fourth and largest invasion got bogged down at the frontier. With the Qing forces completely encircled, a truce was reached between the field commanders of the two sides in December 1769.	How many different invasions of Burma did China launch between 1765 and 1769?	This question is answered directly in the original passage in a single span *"China launched four invasions of Burma between 1765 and 1769."* so it does not meet the requirements of a DROP question and is invalid.
Author Invalid + Annotator Valid	Quarrels between Denmark and Sweden led to the Northern Seven Years' War in 1563, which ended in 1570 with the Treaty of Stettin. Primarily fought in western and southern Scandinavia, the war involved important naval battles fought in the Baltic. When Danish-held Varberg surrendered to Swedish forces in 1565, 150 Danish mercenaries escaped the subsequent massacre of the garrison by defecting to Sweden. Among them was Pontus de la Gardie, who thereafter became an important Swedish commander in the Livonian War. Livonia was also affected by the naval campaign of Danish admiral Peter or Per Munck, who bombarded Swedish Reval from sea in July 1569. The Treaty of Stettin made Denmark the supreme and dominating power in Northern Europe, yet failed to restore the Kalmar Union. Unfavourable conditions for Sweden led to a series of conflicts that only ended with the Great Northern War in 1720. Sweden agreed to turn over her possessions in Livonia in return for a payment by Holy Roman Emperor Maximilian II. Maximilian failed to pay the promised compensation, however, and thereby lost his influence on Baltic affairs. The terms of the treaty regarding Livonia were ignored, and thus the Livonian War continued. From Ivan's point of view, the treaty enabled the powers involved to form an alliance against him, now that they were no longer fighting each other.	How many years was it between the end of the Northern Seven Years' War and the beginning of the Great Northern War?	The authors interpretation of the information *only ended with the Great Northern War in 1720* is that the beginning year of the Great Northern War is not explicitly mentioned, so the question is too ambiguous to be answered with information from the passage. Annotators likely interpreted the same quote as referring to the year the war began, in which case the question would be valid.
Author Valid + Annotator Invalid	Coming off their season-sweeping home win over the 49ers, the Seahawks stayed at home for a Week 11 duel against the Chicago Bears, in the rematch of last year's NFC Divisional game (previously in Chicago). In the first quarter, Seattle trailed early as Bears RB Cedric Benson got a 43-yard TD run, along with kicker Robbie Gould getting a 31-yard field goal. The Seahawks would get on the board with QB Matt Hasselbeck completing a 19-yard TD pass to WR D. J. Hackett. In the second quarter, the Seahawks took the lead with RB Maurice Morris getting a 19-yard TD run. However, Chicago regained the lead with RB Adrian Peterson getting a 5-yard TD run. Seattle would tie the game kicker Josh Brown getting a 40-yard field goal. In the third quarter, the Seahwaks retook the lead as Hasselbeck completed a 4-yard TD pass to WR Nate Burleson for the only score of the period. In the fourth quarter, the Bears tried to retaliate as Gould kicked a 47-yard field goal. Afterwards, Seattle pulled away with Brown kicking a 23-yard and a 46-yard field goal. Chicago's final response would be Gould nailing a 48-yard field goal.	How many yards were the field goals that Josh Brown made in the fourth quarter?	An author of this paper marked the question as valid as it can be answered by summing up the two field goals from Brown. Annotators likely marked it as invalid as the question does not explicitly ask for the total yards of the field goals, so the question could be interpreted as asking for the yards of each field goal individually.

Table 17: Examples of LLM-Generated Questions for DROP with various validity assessments. Questions must involve complex reasoning and looking at more than one part of the passage to answer. The answer to a valid DROP question must be either a span of text from the question or passage, a date, or a number.

Validity Assessment	Passage	LLM-Generated Question	Author Note
Mutually Valid	The south celestial pole is visible only from the Southern Hemisphere. It lies in the dim constellation Octans, the Octant. **Sigma Octantis is identified as the south pole star, more than one degree away from the pole, but with a magnitude of 5.5 it is barely visible on a clear night**.	Would Sigma Octantis be considered a reliable star for navigators to use under cloudy skies or in light-polluted areas?	The question targets the negation that the star is barely visible, and asks about an implication of the star being barely visible.
Mutually Invalid	During the 1980s and early 1990s, Murdoch's publications were generally supportive of Britain's Prime Minister Margaret Thatcher. At the end of the Thatcher/Major era, Murdoch switched his support to the Labour Party and its leader, Tony Blair. The closeness of his relationship with Blair and their secret meetings to discuss national policies was to become a political issue in Britain. This later changed, with "The Sun", in its English editions, publicly renouncing the ruling Labour government and lending its support to David Cameron's Conservative Party, which soon afterwards formed a coalition government. **In Scotland, where the Conservatives had suffered a complete annihilation in 1997, the paper began to endorse the Scottish National Party (though not yet its flagship policy of independence), which soon after came to form the first ever outright majority in the proportionally elected Scottish Parliament.** Former Prime Minister Gordon Brown's official spokesman said in November 2009 that Brown and Murdoch "were in regular communication" and that "there is nothing unusual in the prime minister talking to Rupert Murdoch".	Did "The Sun" initially support the SNPś stance on independence when they first endorsed the party?	The question targets the negation, but not an implication of the negation. It asks for a direct factual interpretation of the negation, so the question is invalid.
Author Invalid + Annotator Valid	A Dominican friar, Orsini focused on his religious responsibilities as bishop rather than on papal administration. **Orsini's lack of political expertise led him to increasingly rely on an unscrupulous secretary (Cardinal Niccolò Coscia) whose financial abuses ruined the papal treasury, causing great damage to the Church in Rome.**	Did Orsini's choice of a close associate contribute positively to his leadership effectiveness?	One author concluded that this question did not target the implications of the negated statement of Niccolò Coscia being unscrupulous directly. His being unscrupulous is not directly attributable to Orsini's leadership effectiveness. Orsini could have been extremely effective despite Niccolò's unscrupulous behavior. Annotators likely felt that the direct connection did exist, and therefore rated the question as valid.
Author Valid + Annotator Invalid	The walls and sanctum's door frame of the both levels are intricately carved. These show themes such as "makaras" and other mythical creatures, apsaras, elephants in different stages of activity, females in waving or welcoming gesture. The upper level of Cave 6 is significant in that it shows a devotee in a kneeling posture at the Buddha's feet, an indication of devotional worship practices by the 5th century. The colossal Buddha of the shrine has an elaborate throne back, but was hastily finished in 477/478 CE, when king Harisena died. **The shrine antechamber of the cave features an unfinished sculptural group of the Six Buddhas of the Past, of which only five statues were carved.** This idea may have been influenced from those in Bagh Caves of Madhya Pradesh.	Is there more than one fully completed sculptural group in the antechamber of the cave?	An author concluded that this question was valid as the negation implies what the answer to this question would be. The answer should be "don't know" here because despite the finished or unfinished state of the sculptural group, there could potentially be more sculptural groups in the antechamber. Annotators marked the question invalid, either due to assuming that an answer of "don't know" made the question invalid, or deeming the connection between the implication of the negation and the question too weak.

Table 18: Examples of LLM-Generated Questions for CondaQA with various validity assessments. Questions must be about an implication of the negated statement. The negated statement is highlighted in **bold** and the negation is highlighted in blue.

Validity Assessment	Passage	LLM-Generated Paraphrase Edit	Author Note
Mutually Valid	The school song, "Fair Reed," is sung to the tune of the 1912 popular song "Believe Me, if All Those Endearing Young Charms." It may be imitative of the Harvard anthem "Fair Harvard," which is also sung to the tune of "Believe Me, if All Those Endearing Young Charms." **It was composed by former president William Trufant Foster shortly after Reed's founding, and is rarely heard today.**	The school song, "Fair Reed," is sung to the tune of the 1912 popular song "Believe Me, if All Those Endearing Young Charms." It may be imitative of the Harvard anthem "Fair Harvard," which is also sung to the tune of "Believe Me, if All Those Endearing Young Charms." **Former president William Trufant Foster composed it shortly after Reed's founding, and it is seldom heard today.**	The edited passage has the same meaning as the original, and the negation rarely is no longer present, so the edit is valid.
Mutually Invalid	Most of what is known about Anthony comes from the "Life of Anthony". **Written in Greek around 360 by Athanasius of Alexandria, it depicts Anthony as an illiterate and holy man who through his existence in a primordial landscape has an absolute connection to the divine truth, which always is in harmony with that of Athanasius as the biographer.**	Most of what is known about Anthony comes from the "Life of Anthony". **Written in Greek around 360 by Athanasius of Alexandria, it portrays Anthony as an uneducated yet holy man who, through his life in a primordial landscape, achieves a total connection to the divine truth, which is invariably aligned with that presented by Athanasius as the biographer.**	Although the negation is no longer present, the edited passage does not retain the same meaning as the original. *Uneducated* does not imply the same meaning as *illiterate* . Therefore the edit is invalid.
Author Invalid + Annotator Valid	In 2008, Narconon executive Marie Cecchini published, with Richard Lennox, a paper that claimed to show that the Narconon educational program reduced drug use among youths. **However, the study was funded by Narconon's parent organisation, ABLE, and subsequent correspondence in the same journal asserted that the study's conclusions were contradicted by its own data: that the control group "were more likely to resist pressures to take drugs" than the Narconon group.**	In 2008, Narconon executive Marie Cecchini published, with Richard Lennox, a paper that claimed to show that the Narconon educational program reduced drug use among youths. **However, the study was funded by Narconon's parent organisation, ABLE, and subsequent correspondence in the same journal asserted that the study's conclusions were contradicted by its own data: the control group "were more likely to yield to pressures to take drugs" less than the Narconon group.**	While the edited passage does remove the negation, the resulting passage is incoherent. *"were more likely to yield to pressure to take drugs" less than the Narconon group* doesn't make any sense, since the presence of "more" implies that there is another group the control group is more likely than to yield to pressure to take drugs. If the edit had been *"were less likely to yield to pressure to take drugs"*, then the edit would be valid. Annotators likely just saw that the double negation preserved the meaning of the original passage, and interpreted the sentence as coherent.
Author Valid + Annotator Invalid	In the early 20th century and into the Mexican Revolution, the production of coffee was particularly important but labor-intensive. This would lead to a practice called "enganche" (hook), where recruiters would lure workers with advanced pay and other incentives such as alcohol and then trap them with debts for travel and other items to be worked off. **This practice would lead to a kind of indentured servitude and uprisings in areas of the state, although they never led to large rebel armies as in other parts of Mexico.**	In the early 20th century and into the Mexican Revolution, the production of coffee was particularly important but labor-intensive. This would lead to a practice called "enganche" (hook), where recruiters would lure workers with advanced pay and other incentives such as alcohol and then trap them with debts for travel and other items to be worked off. **This practice would lead to a kind of indentured servitude and uprisings in areas of the state, albeit without resulting in the formation of large rebel armies as seen in other parts of Mexico.**	An author marked this edit as valid as it retained the meaning of the original passage (that the practice did not lead to the formation of large rebel armies). Annotators marked it as invalid, possibly because the exact meaning of the original passage is slightly changed in the paraphrase edit.

Table 19: Examples of LLM-Generated Paraphrase Edits for CondaQA with various validity assessments. Paraphrase edits must be a rewrite of the original negated statement such that the new passage has the same meaning as the original, but the negation is no longer included. The negated statement is highlighted in **bold** and the negation is highlighted in blue.

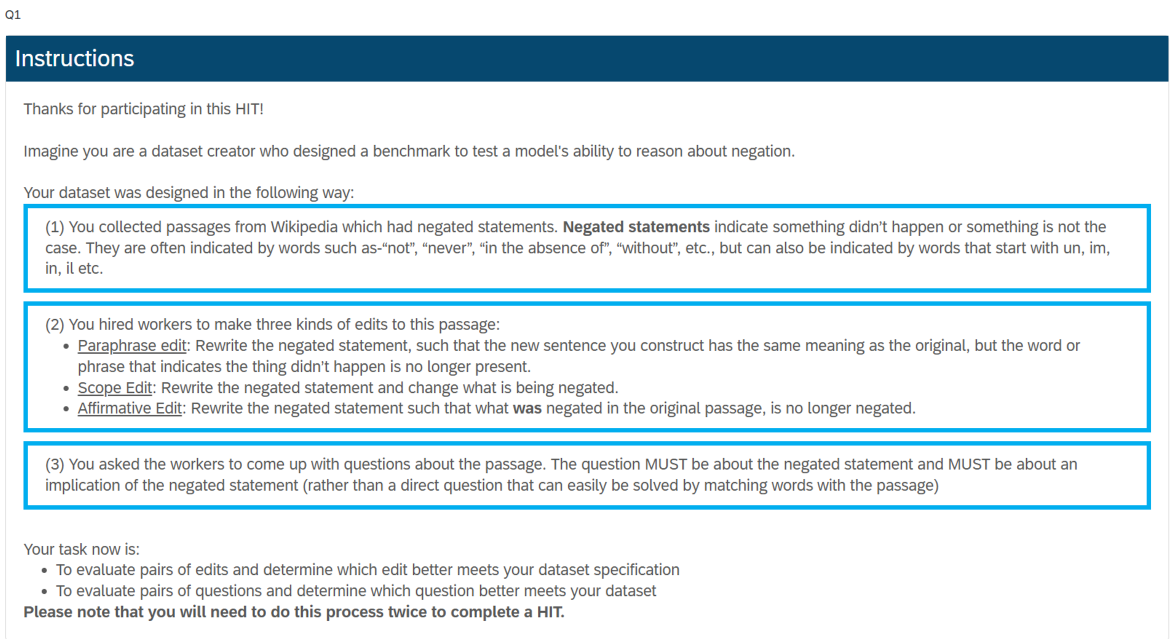
Validity Assessment	Passage	LLM-Generated Scope Edit	Author Note
Mutual Valid	Much like elephants, manatees are polyphyodonts, and continuously replace their teeth from the back of the jaw. **Adults lack incisors, canines, and premolars, and instead have 8 to 10 cheek teeth in their mouth.** Manatees have an infinite supply of teeth moving in from the back and shedding in the front, which are continuously formed by a dental capsule behind the tooth-row. These teeth are constantly worn down by the abrasive vascular plants they forage, particularly aquatic grasses. Unlike in manatees, the dugong's teeth do not continually grow back via horizontal tooth replacement. The dugong has two tusks which emerge in males during puberty, and sometime later in life for females after reaching the base of the premaxilla. The number of growth layer groups in a tusk indicates the age of a dugong.	Much like elephants, manatees are polyphyodonts, and continuously replace their teeth from the back of the jaw. **Adults lack molars and instead have 8 to 10 incisors, canines, and premolars in their mouth.** Manatees have an infinite supply of teeth moving in from the back and shedding in the front, which are continuously formed by a dental capsule behind the tooth-row. These teeth are constantly worn down by the abrasive vascular plants they forage, particularly aquatic grasses. Unlike in manatees, the dugong's teeth do not continually grow back via horizontal tooth replacement. The dugong has two tusks which emerge in males during puberty, and sometime later in life for females after reaching the base of the premaxilla. The number of growth layer groups in a tusk indicates the age of a dugong.	The edit changes what is being negated by lack from the incisors, canines and premolars to the molars, so the scope edit is valid.
Mutually Invalid	It is difficult to distinguish an extinct volcano from a dormant (inactive) one. Dormant volcanoes are those that have not erupted for thousands of years, but are likely to erupt again in the future. Volcanoes are often considered to be extinct if there are no written records of its activity. Nevertheless, volcanoes may remain dormant for a long period of time. For example, Yellowstone has a repose/recharge period of around 700,000 years, and Toba of around 380,000 years. Vesuvius was described by Roman writers as having been covered with gardens and vineyards before its eruption of 79 CE, which destroyed the towns of Herculaneum and Pompeii. **Before its catastrophic eruption of 1991, Pinatubo was an inconspicuous volcano, unknown to most people in the surrounding areas.** Two other examples are the long-dormant Soufrière Hills volcano on the island of Montserrat, thought to be extinct before activity resumed in 1995, and Fourpeaked Mountain in Alaska, which, before its September 2006 eruption, had not erupted since before 8000 BCE and had long been thought to be extinct.	It is difficult to distinguish an extinct volcano from a dormant (inactive) one. Dormant volcanoes are those that have not erupted for thousands of years, but are likely to erupt again in the future. Volcanoes are often considered to be extinct if there are no written records of its activity. Nevertheless, volcanoes may remain dormant for a long period of time. For example, Yellowstone has a repose/recharge period of around 700,000 years, and Toba of around 380,000 years. Vesuvius was described by Roman writers as having been covered with gardens and vineyards before its eruption of 79 CE, which destroyed the towns of Herculaneum and Pompeii. **Before its catastrophic eruption of 1991, Pinatubo was an inconspicuous volcano, except to a few local geologists who had studied its potential for activity.** Two other examples are the long-dormant Soufrière Hills volcano on the island of Montserrat, thought to be extinct before activity resumed in 1995, and Fourpeaked Mountain in Alaska, which, before its September 2006 eruption, had not erupted since before 8000 BCE and had long been thought to be extinct.	The negation in the edit is still negating the Pinatubo volcano. As the subject of the negation has not changed, the scope edit is therefore invalid.
Author Invalid + Annotator Valid	Hearts is an "evasion-type" trick-taking playing card game for four players, although most variations can accommodate between three and six players. It was first recorded in America in the 1880s and has many variants, some of which are also referred to as "Hearts"; especially the games of Black Lady and Black Maria. **The game is a member of the Whist group of trick-taking games (which also includes Bridge and Spades), but is unusual among Whist variants in that it is a trick-avoidance game; players avoid winning certain penalty cards in tricks, usually by avoiding winning tricks altogether.** The original game of Hearts is still current, but has been overtaken in popularity by Black Lady in the United States and Black Maria in Great Britain.	Hearts is an "evasion-type" trick-taking playing card game for four players, although most variations can accommodate between three and six players. It was first recorded in America in the 1880s and has many variants, some of which are also referred to as "Hearts"; especially the games of Black Lady and Black Maria. **The game is a member of the Whist group of trick-taking games (which also includes Bridge and Spades), but is unusual among Whist variants in that it involves using a unique scoring system where points are deducted for each trick won, encouraging players to win as few tricks as possible.** The original game of Hearts is still current, but has been overtaken in popularity by Black Lady in the United States and Black Maria in Great Britain.	An author interpreted the subject of negation in both the original and edited passages as Hearts being different from other Whist variants. The scope edit changes why Hearts is different from other Whist variants, but not the subject of what is unusual, which remains that Hearts is different from other Whist variants. An annotator likely interpreted the change in why Hearts is different as a sufficient change in scope of the negation.
Author Valid + Annotator Invalid	Gwen Stacy is recast as Paige Embry in "The Refrigerator Monologues". She is the unofficial leader of the Hell Hath Club, a group of women in the afterlife trying to cope with the brutal termination of their plot lines, and provides connecting narration for each of their stories. **She loved wearing her lab coat and feels unhappy about being stuck in the clothes she was buried in.**	Gwen Stacy is recast as Paige Embry in "The Refrigerator Monologues". She is the unofficial leader of the Hell Hath Club, a group of women in the afterlife trying to cope with the brutal termination of their plot lines, and provides connecting narration for each of their stories. **She loved wearing her lab coat and feels unhappy about not being able to wear it anymore, as she is stuck in the clothes she was buried in.**	The difference here is subtle, but the subject of the negation unhappy is on being stuck in the clothes she was wearing in the original passage, and is on not being able to wear her lab coat anymore in the scope edit. An author deemed this change in subject as sufficient. An annotator likely did not agree that the change in scope was different enough for the edit to be valid.

Table 20: Examples of LLM-Generated Scope Edits for CondaQA with various validity assessments. Scope edits must be a rewrite of the original passage such that what is being negated by the negation is changed. The negated statement is highlighted in **bold** and the negation is highlighted in blue.

Validity Assessment	Passage	LLM-Generated Affirmative Edit	Author Note
Mutual Valid	Johann Spurzheim was Gall's most important collaborator. **He worked as Gall's anatomist until 1813 when for unknown reasons they had a permanent falling out.** Publishing under his own name Spurzheim successfully disseminated phrenology throughout the United Kingdom during his lecture tours through 1814 and 1815 and the United States in 1832 where he would eventually die.	Johann Spurzheim was Gall's most important collaborator. **He worked as Gall's anatomist until 1813 when, due to professional disagreements over the methods and applications of phrenology, they had a permanent falling out.** Publishing under his own name, Spurzheim successfully disseminated phrenology throughout the United Kingdom during his lecture tours through 1814 and 1815 and the United States in 1832, where he would eventually die.	This edit is valid since the reasons for the falling out are no longer unknown in the Affirmative edit, and the new passage is coherent and consistent.
Author Invalid + Annotator Valid	In 2008, Narconon executive Marie Cecchini published, with Richard Lennox, a paper that claimed to show that the Narconon educational program reduced drug use among youths. **However, the study was funded by Narconon's parent organisation, ABLE, and subsequent correspondence in the same journal asserted that the study's conclusions were contradicted by its own data: that the control group "were more likely to resist pressures to take drugs" than the Narconon group.**	In 2008, Narconon executive Marie Cecchini published, with Richard Lennox, a paper that claimed to show that the Narconon educational program reduced drug use among youths. **However, the study was funded by Narconon's parent organisation, ABLE, and subsequent correspondence in the same journal asserted that the study's conclusions were contradicted by its own data: that the control group "were more likely to succumb to pressures to take drugs" than the Narconon group.**	While the affirmative edit does undo the negation, it also contradicts an earlier portion of the passage. The passage states that the studies conclusions contradict its own data, but the undone negation actually supports the studies conclusions. Annotators likely noted the reversal of the negation but failed to notice that it makes the passage as a whole inconsistent.
Author Valid + Annotator InValid	Cro-tatting combines needle tatting with crochet. The cro-tatting tool is a tatting needle with a crochet hook at the end. One can also cro-tat with a bullion crochet hook or a very straight crochet hook. In the 19th century, "crochet tatting" patterns were published which simply called for a crochet hook. One of the earliest patterns is for a crocheted afghan with tatted rings forming a raised design. Patterns are available in English and are equally divided between yarn and thread. In its most basic form, the rings are tatted with a length of plain thread between them, as in single-shuttle tatting. In modern patterns, beginning in the early 20th century, the rings are tatted and the arches or chains are crocheted. Many people consider cro-tatting more difficult than crochet or needle tatting. Some tatting instructors recommend using a tatting needle and a crochet hook to work cro-tatting patterns. **Stitches of cro-tatting (and needle tatting before a ring is closed) unravel easily, unlike tatting made with a shuttle.**	Cro-tatting combines needle tatting with crochet. The cro-tatting tool is a tatting needle with a crochet hook at the end. One can also cro-tat with a bullion crochet hook or a very straight crochet hook. In the 19th century, "crochet tatting" patterns were published which simply called for a crochet hook. One of the earliest patterns is for a crocheted afghan with tatted rings forming a raised design. Patterns are available in English and are equally divided between yarn and thread. In its most basic form, the rings are tatted with a length of plain thread between them, as in single-shuttle tatting. In modern patterns, beginning in the early 20th century, the rings are tatted and the arches or chains are crocheted. Many people consider cro-tatting more difficult than crochet or needle tatting. Some tatting instructors recommend using a tatting needle and a crochet hook to work cro-tatting patterns. **Stitches of cro-tatting (and needle tatting before a ring is closed) do not unravel easily, similar to tatting made with a shuttle.**	An author of this paper assessed the edit as valid, as it undoes the negation of unlike and changes it to similar. Although it changes another part of the passage, the negation of unlike compared to stitches of cro-tatting is removed, and the passage is coherent and consistent. Annotators likely noticed that the passage is still saying that tatting made with a shuttle does not unravel easily as in the original passage, and took that to mean the affirmative edit was not successful.

Table 21: Examples of LLM-Generated Affirmative Edits for CondaQA with various validity assessments. Affirmative edits must be a rewrite of the original passage such that was being negated is no longer negated, with minimal changes to the passage to keep it consistent and coherent. The negated statement is highlighted in **bold** and the negation is highlighted in blue.

Figure 5: The instructions for the **CondaQA** preference study (§5).



Q1

Instructions

Thanks for participating in this HIT!

Imagine you are a dataset creator who designed a benchmark to test a model's ability to reason about negation.

Your dataset was designed in the following way:

(1) You collected passages from Wikipedia which had negated statements. **Negated statements** indicate something didn't happen or something is not the case. They are often indicated by words such as-"not", "never", "in the absence of", "without", etc., but can also be indicated by words that start with un, im, in, il etc.

(2) You hired workers to make three kinds of edits to this passage:

- Paraphrase edit: Rewrite the negated statement, such that the new sentence you construct has the same meaning as the original, but the word or phrase that indicates the thing didn't happen is no longer present.
- Scope Edit: Rewrite the negated statement and change what is being negated.
- Affirmative Edit: Rewrite the negated statement such that what **was** negated in the original passage, is no longer negated.

(3) You asked the workers to come up with questions about the passage. The question MUST be about the negated statement and MUST be about an implication of the negated statement (rather than a direct question that can easily be solved by matching words with the passage)

Your task now is:

- To evaluate pairs of edits and determine which edit better meets your dataset specification
- To evaluate pairs of questions and determine which question better meets your dataset

Please note that you will need to do this process twice to complete a HIT.

Figure 6: This screenshot shows what comes after the instructions in the **CondaQA** preference study (§5): the original passage.

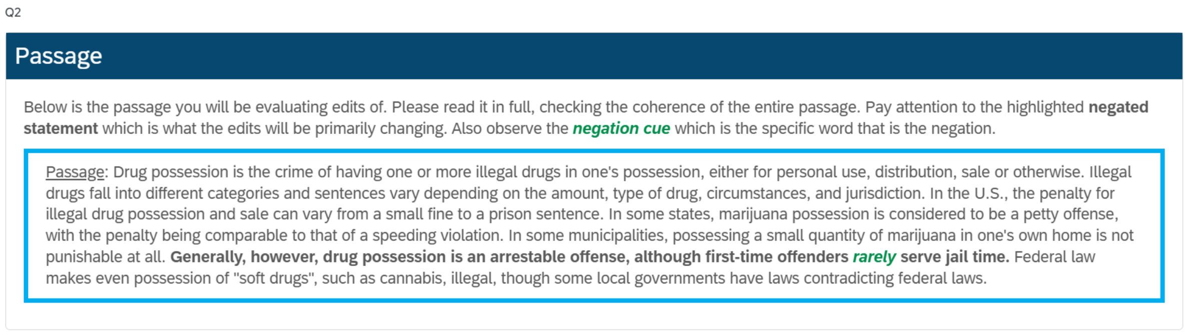


Figure 7: After the original passage (Fig. 6), annotators get a definition of a paraphrase edit and two paraphrase edits.

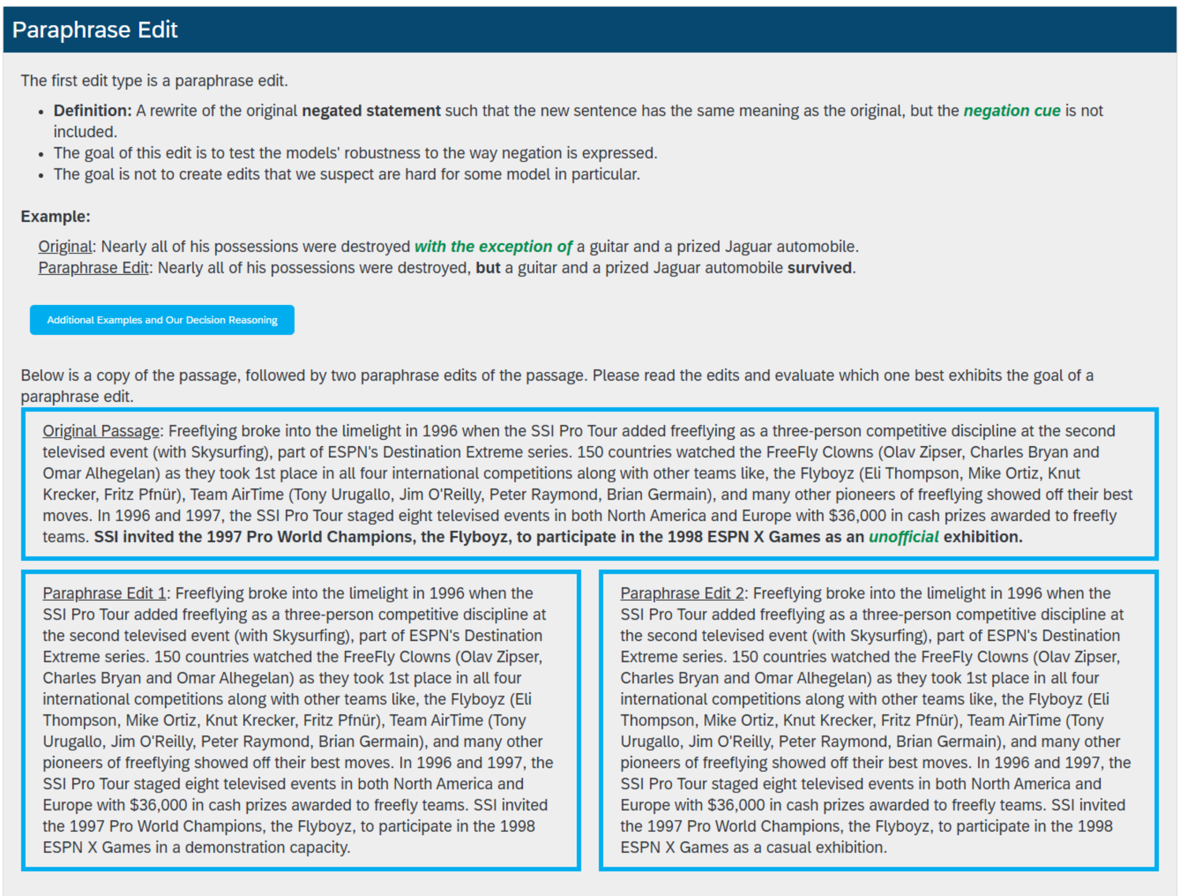


Figure 8: The options annotators select for edits in the preference study (§5). Here shown for paraphrase edits.

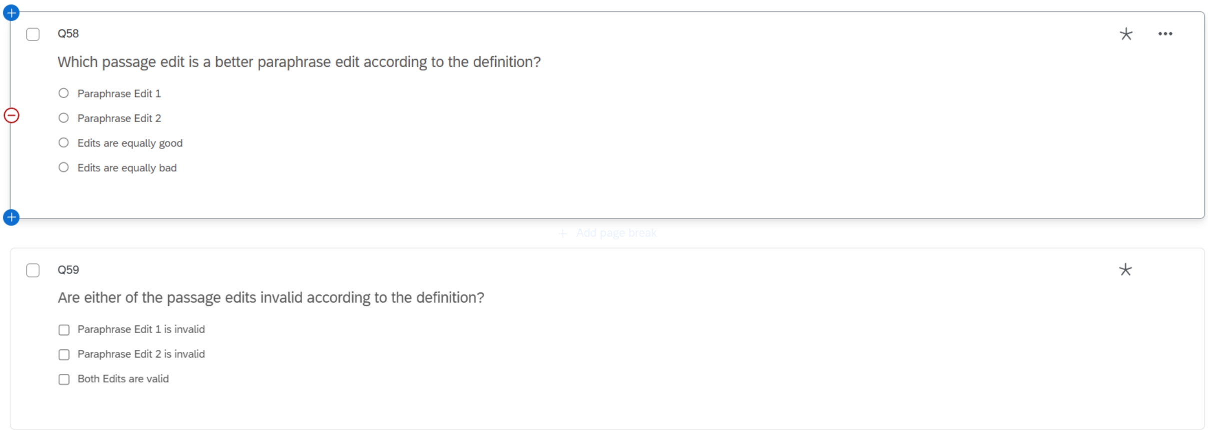


Figure 9: After the paraphrase edit selection (Figures 7–8), annotators get the scope edit definition and a pair of two scope edits.

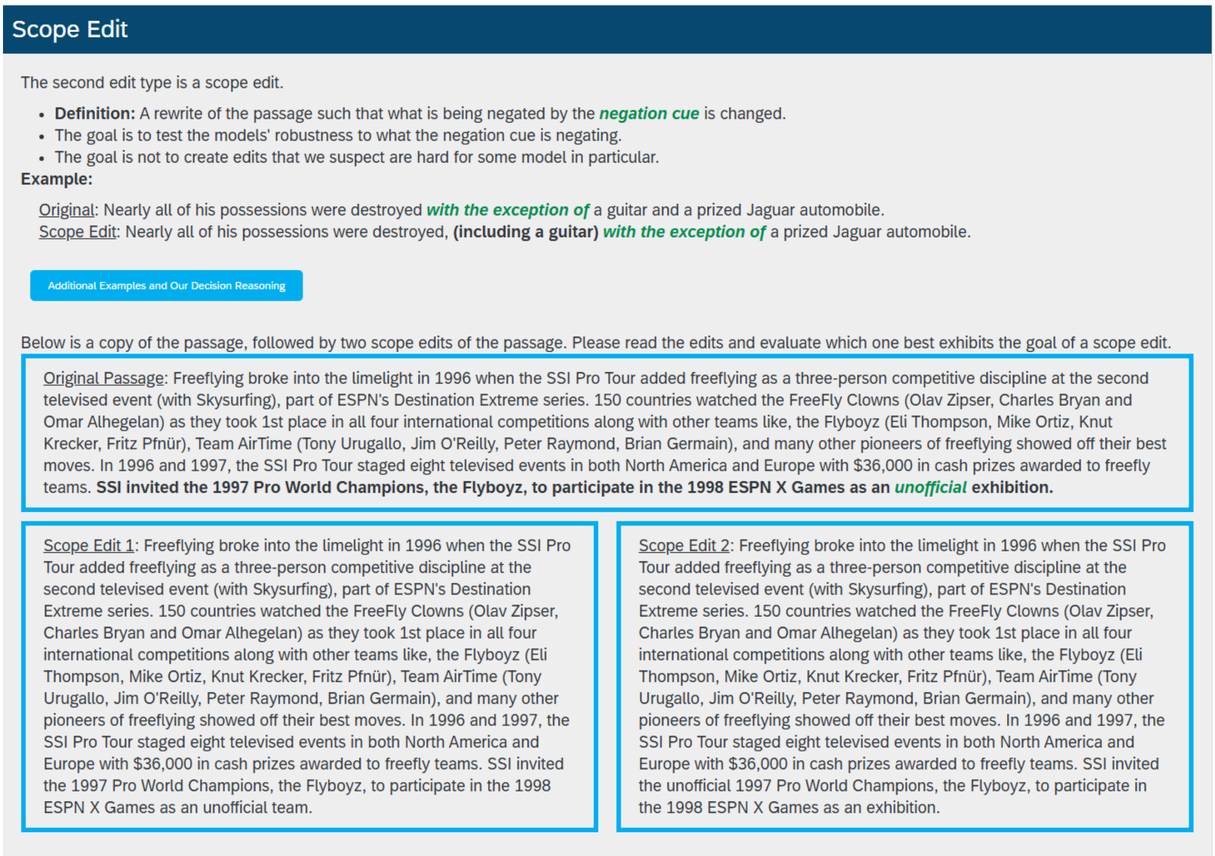


Figure 10: After the scope edits (Fig. 9), annotators get the affirmative edit definition and a pair of two affirmative edits.

Affirmative Edit

The third edit type is an affirmative edit.

- **Definition:** A rewrite of the passage such that what was being negated in the original is no longer negated, with minimal changes to the passage to keep it consistent and coherent.
- The goal is to test the models' robustness to the presence/absence of negation.
- The goal is not to create edits that we suspect are hard for some model in particular.

Example:

Original: Nearly all of his possessions were destroyed ***with the exception of*** a guitar and a prized Jaguar automobile.
Affirmative Edit: Nearly all of his possessions were destroyed, **including** a guitar and a prized Jaguar automobile.

Additional Examples and Our Decision Reasoning

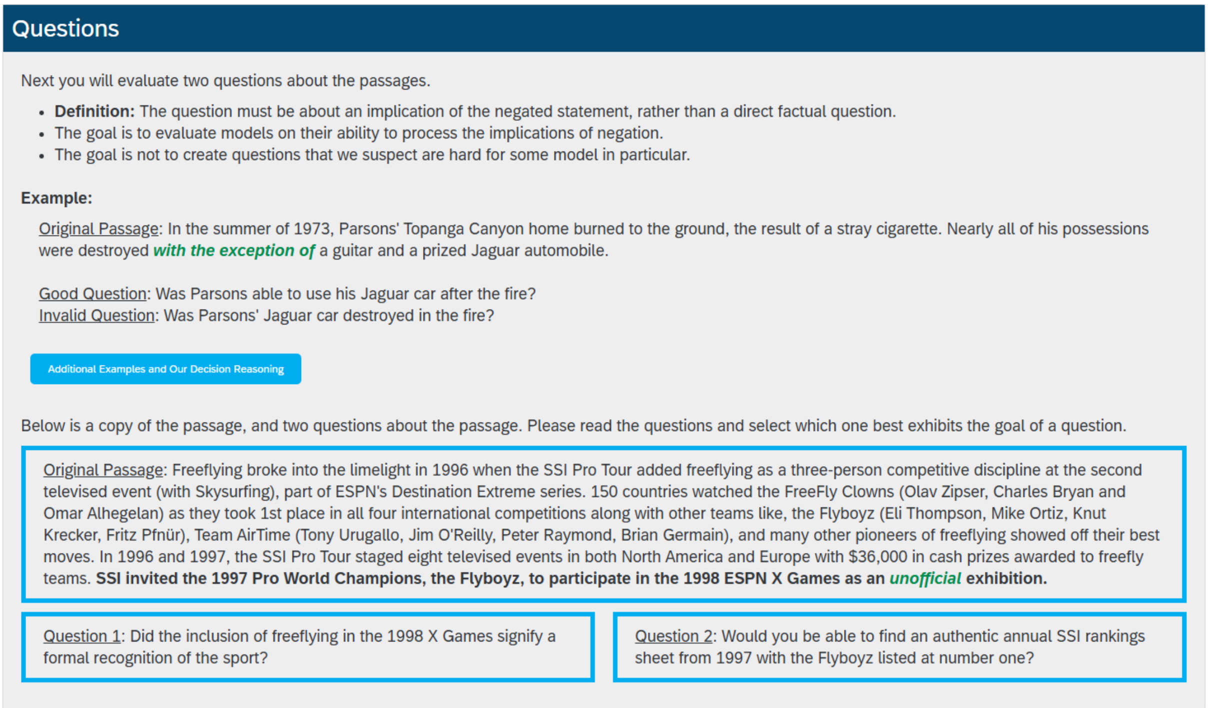
Below is a copy of the passage, followed by two affirmative edits of the passage. Please read the edits and select which one best exhibits the goal of an affirmative edit.

Original Passage: Freeflying broke into the limelight in 1996 when the SSI Pro Tour added freeflying as a three-person competitive discipline at the second televised event (with Skysurfing), part of ESPN's Destination Extreme series. 150 countries watched the FreeFly Clowns (Olav Zipser, Charles Bryan and Omar Alhegelan) as they took 1st place in all four international competitions along with other teams like, the Flyboyz (Eli Thompson, Mike Ortiz, Knut Krecker, Fritz Pfnür), Team AirTime (Tony Urugallo, Jim O'Reilly, Peter Raymond, Brian Germain), and many other pioneers of freeflying showed off their best moves. In 1996 and 1997, the SSI Pro Tour staged eight televised events in both North America and Europe with $36,000 in cash prizes awarded to freefly teams. **SSI invited the 1997 Pro World Champions, the Flyboyz, to participate in the 1998 ESPN X Games as an *unofficial* exhibition.**

Affirmative Edit 1: Freeflying broke into the limelight in 1996 when the SSI Pro Tour added freeflying as a three-person competitive discipline at the second televised event (with Skysurfing), part of ESPN's Destination Extreme series. 150 countries watched the FreeFly Clowns (Olav Zipser, Charles Bryan and Omar Alhegelan) as they took 1st place in all four international competitions along with other teams like, the Flyboyz (Eli Thompson, Mike Ortiz, Knut Krecker, Fritz Pfnür), Team AirTime (Tony Urugallo, Jim O'Reilly, Peter Raymond, Brian Germain), and many other pioneers of freeflying showed off their best moves. In 1996 and 1997, the SSI Pro Tour staged eight televised events in both North America and Europe with $36,000 in cash prizes awarded to freefly teams. SSI invited the 1997 Pro World Champions, the Flyboyz, to participate in the 1998 ESPN X Games as a sanctioned exhibition.

Affirmative Edit 2: Freeflying broke into the limelight in 1996 when the SSI Pro Tour added freeflying as a three-person competitive discipline at the second televised event (with Skysurfing), part of ESPN's Destination Extreme series. 150 countries watched the FreeFly Clowns (Olav Zipser, Charles Bryan and Omar Alhegelan) as they took 1st place in all four international competitions along with other teams like, the Flyboyz (Eli Thompson, Mike Ortiz, Knut Krecker, Fritz Pfnür), Team AirTime (Tony Urugallo, Jim O'Reilly, Peter Raymond, Brian Germain), and many other pioneers of freeflying showed off their best moves. In 1996 and 1997, the SSI Pro Tour staged eight televised events in both North America and Europe with $36,000 in cash prizes awarded to freefly teams. SSI invited the 1997 Pro World Champions, the Flyboyz, to participate in the 1998 ESPN X Games as an official exhibition.

Figure 11: After the affirmative edits (Fig. 10), annotators get the CondaQA-style question definition and a pair of two questions.



Questions

Next you will evaluate two questions about the passages.

- **Definition:** The question must be about an implication of the negated statement, rather than a direct factual question.
- The goal is to evaluate models on their ability to process the implications of negation.
- The goal is not to create questions that we suspect are hard for some model in particular.

Example:

Original Passage: In the summer of 1973, Parsons' Topanga Canyon home burned to the ground, the result of a stray cigarette. Nearly all of his possessions were destroyed ***with the exception of*** a guitar and a prized Jaguar automobile.

Good Question: Was Parsons able to use his Jaguar car after the fire?
Invalid Question: Was Parsons' Jaguar car destroyed in the fire?

Additional Examples and Our Decision Reasoning

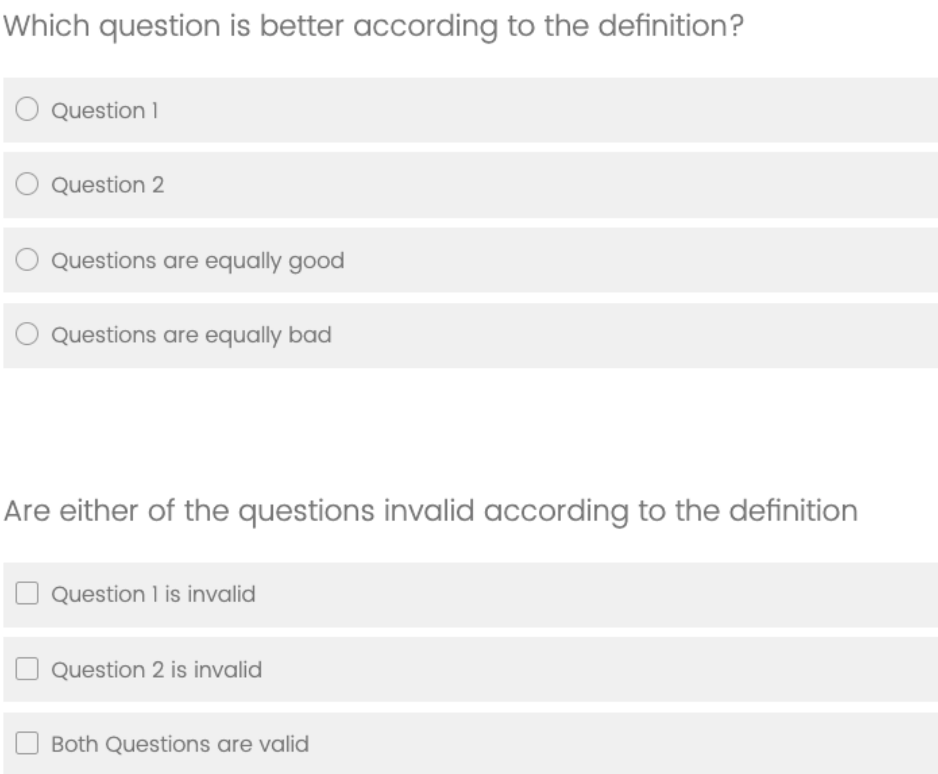
Below is a copy of the passage, and two questions about the passage. Please read the questions and select which one best exhibits the goal of a question.

Original Passage: Freeflying broke into the limelight in 1996 when the SSI Pro Tour added freeflying as a three-person competitive discipline at the second televised event (with Skysurfing), part of ESPN's Destination Extreme series. 150 countries watched the FreeFly Clowns (Olav Zipser, Charles Bryan and Omar Alhegelan) as they took 1st place in all four international competitions along with other teams like, the Flyboyz (Eli Thompson, Mike Ortiz, Knut Krecker, Fritz Pfnür), Team AirTime (Tony Urugallo, Jim O'Reilly, Peter Raymond, Brian Germain), and many other pioneers of freeflying showed off their best moves. In 1996 and 1997, the SSI Pro Tour staged eight televised events in both North America and Europe with $36,000 in cash prizes awarded to freefly teams. **SSI invited the 1997 Pro World Champions, the Flyboyz, to participate in the 1998 ESPN X Games as an *unofficial* exhibition.**

Question 1: Did the inclusion of freeflying in the 1998 X Games signify a formal recognition of the sport?

Question 2: Would you be able to find an authentic annual SSI rankings sheet from 1997 with the Flyboyz listed at number one?

Figure 12: The choices for stating a preference between two questions.



Which question is better according to the definition?

- Question 1
- Question 2
- Questions are equally good
- Questions are equally bad

Are either of the questions invalid according to the definition

- Question 1 is invalid
- Question 2 is invalid
- Both Questions are valid

Figure 13: **CondaQA** answer crowrdsourcing interface used in §6.

Questions

Finally, please re-read the question in the context of the passage, then answer the question.

- The answer to the question should be either "Yes", "No", "Don't Know" or a span of text directly from the passage/question.

Below is a copy of the passage, and two questions about the passage. Please read the questions and provide the answers for both.

Original Passage: Drug possession is the crime of having one or more illegal drugs in one's possession, either for personal use, distribution, sale or otherwise. Illegal drugs fall into different categories and sentences vary depending on the amount, type of drug, circumstances, and jurisdiction. In the U.S., the penalty for illegal drug possession and sale can vary from a small fine to a prison sentence. In some states, marijuana possession is considered to be a petty offense, with the penalty being comparable to that of a speeding violation. In some municipalities, possessing a small quantity of marijuana in one's own home is not punishable at all. **Generally, however, drug possession is an arrestable offense, although first-time offenders *rarely* serve jail time.** Federal law makes even possession of "soft drugs", such as cannabis, illegal, though some local governments have laws contradicting federal laws.

+ Add page break

Q16

Question 1: If a first-time drug possession offender in the U.S. does not go to jail, might they still need to attend court?

What is the answer for Question 1?

- Yes
- No
- Span of text from the Passage or Question
- Don't Know

Figure 14: The beginning of the instruction for the **DROP** answer crowdsourcing (§6). The instructions continues in the next figure.

Instructions

Thanks for participating in this HIT!

In this study, You will be provided with a passage and two questions about the passage. You will need to answer each of the questions using the given passage as context.

It is important that you DO NOT use any sort of AI or Language Model assistance (i.e. ChatGPT) for this task. This would render the annotations useless and the HIT invalid.

Tutorial:

The answer to the questions can be words from the passage (spans), numbers, or dates.

- A ***span*** is a continuous phrase taken directly from the passage or question. You can use the provided highlighting tool to select spans directly from the passage or question. If you find multiple spans, please select that number of spans when prompted. Please restrict each span to five words.
- A ***number*** type answer can include a digit specifying an actual value, along with a unit type that signifies the value type the digit represents.
- A ***date*** type answer can be used to specify a date with the help of a date selector.

NOTE: A date is exclusively a full date with a day, month and year. If a question is asking something like *"In what year did ..."* then the answer would be a **span** which selects the relevant year in the passage or question text. If a question is asking something like *"How many months did it take for..."*, then the answer would be a **number**. Please see the example passages and questions below for additional examples of answer types.

Passage Sample: Austria-Hungary precipitated the Bosnian crisis of 1908-1909 by officially annexing the former Ottoman territory of Bosnia and Herzegovina, which it had occupied since 1878. This angered the Kingdom of Serbia and its patron, the Pan-Slavic and Orthodox Russian Empire. Russian political manoeuvring in the region destabilised peace accords that were already fracturing in the Balkans, which came to be known as the "powder keg of Europe." In 1912 and 1913, the First Balkan War was fought between the Balkan League and the fracturing Ottoman Empire. The resulting Treaty of London further shrank the Ottoman Empire, creating an independent Albanian state while enlarging the territorial holdings of Bulgaria, Serbia, Montenegro, and Greece. When Bulgaria attacked Serbia and Greece on 16 June 1913, it lost most of Macedonia to Serbia and Greece, and Southern Dobruja to Romania in the 33-day Second Balkan War, further destabilising the region. The Great Powers were able to keep these Balkan conflicts contained, but the next one would spread throughout Europe and beyond.

Example Question 1: How many years did it take for Austria-Hungary to annex Bosnia and Herzegovina?
Answer: 31 Years *(number)*

Example Question 2: When did the Second Balkan war end?
Answer: 19 July 1913 *(date)*

Figure 15: The rest of the instruction for the **DROP** answer crowdsourcing (§6).

Example Question 3: Which two wars were ongoing in 1913?
Answer: "First Balkan war", "Second Balkan War" *(two separate spans)*

Example Question 4: Which country started the Second Balkan War?
Answer: Bulgaria *(span from passage)*

Question #1

Please read both the passage and the question below, then answer the question using the passage as context.

Passage: Coming off their impressive Monday night road win over the Broncos, the Steelers went home for a divisional rematch with the Cincinnati Bengals with first place in the division on the line. In the first quarter, Pittsburgh would deliver the game's first strike with a 28-yard field goal from kicker Jeff Reed. However, the Bengals would immediately answer as running back Bernard Scott returned a kickoff 96 yards for a touchdown (with a failed PAT). The Steelers would regain the lead in the second quarter as Reed got a 33-yard and a 35-yard field goal. Cincinnati would respond in the third quarter as kicker Shayne Graham made a 23-yard and a 32-yard field goal. Pittsburgh would tie the game in the fourth quarter with Reed nailing a 34-yard field goal. However, the Bengals would pull away as Graham booted a 32-yard and a 43-yard field goal.

Question: How many yards shorter was Shayne Graham's shortest field goal compared to Bernard Scott's kickoff return touchdown?

First please select the answer type for this question.

- Date
- Span(s)
- Number

Figure 16: Specifying the **DROP** *numerical* answer (§6).

Number Selection

The passage and the question are displayed here for reference.

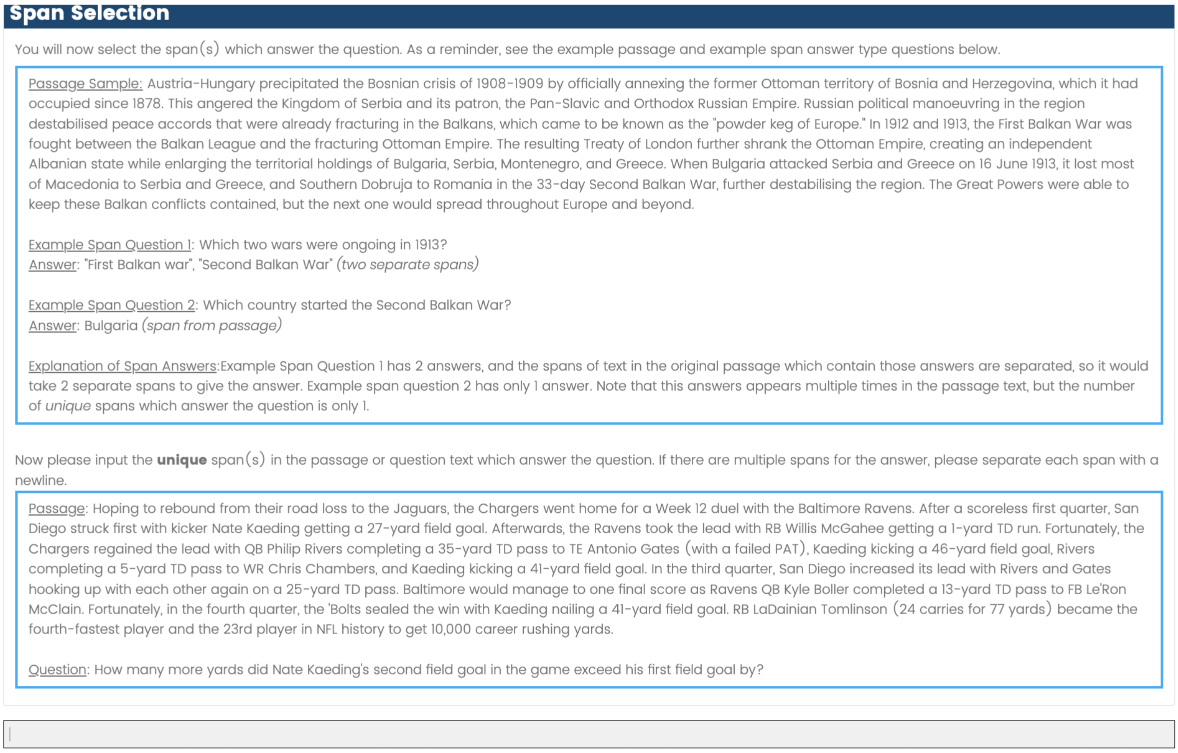
Passage: Coming off their impressive Monday night road win over the Broncos, the Steelers went home for a divisional rematch with the Cincinnati Bengals with first place in the division on the line. In the first quarter, Pittsburgh would deliver the game's first strike with a 28-yard field goal from kicker Jeff Reed. However, the Bengals would immediately answer as running back Bernard Scott returned a kickoff 96 yards for a touchdown (with a failed PAT). The Steelers would regain the lead in the second quarter as Reed got a 33-yard and a 35-yard field goal. Cincinnati would respond in the third quarter as kicker Shayne Graham made a 23-yard and a 32-yard field goal. Pittsburgh would tie the game in the fourth quarter with Reed nailing a 34-yard field goal. However, the Bengals would pull away as Graham booted a 32-yard and a 43-yard field goal.

Question: How many yards shorter was Shayne Graham's shortest field goal compared to Bernard Scott's kickoff return touchdown?

Enter the numerical value for the answer here

Now enter the unit of the number for the answer (e.g. "years", "yards", "people", etc.)

Figure 17: Specifying the **DROP** *span* answer (§6).



Span Selection

You will now select the span(s) which answer the question. As a reminder, see the example passage and example span answer type questions below.

Passage Sample: Austria-Hungary precipitated the Bosnian crisis of 1908-1909 by officially annexing the former Ottoman territory of Bosnia and Herzegovina, which it had occupied since 1878. This angered the Kingdom of Serbia and its patron, the Pan-Slavic and Orthodox Russian Empire. Russian political manoeuvring in the region destabilised peace accords that were already fracturing in the Balkans, which came to be known as the "powder keg of Europe." In 1912 and 1913, the First Balkan War was fought between the Balkan League and the fracturing Ottoman Empire. The resulting Treaty of London further shrank the Ottoman Empire, creating an independent Albanian state while enlarging the territorial holdings of Bulgaria, Serbia, Montenegro, and Greece. When Bulgaria attacked Serbia and Greece on 16 June 1913, it lost most of Macedonia to Serbia and Greece, and Southern Dobruja to Romania in the 33-day Second Balkan War, further destabilising the region. The Great Powers were able to keep these Balkan conflicts contained, but the next one would spread throughout Europe and beyond.

Example Span Question 1: Which two wars were ongoing in 1913?
Answer: "First Balkan war", "Second Balkan War" *(two separate spans)*

Example Span Question 2: Which country started the Second Balkan War?
Answer: Bulgaria *(span from passage)*

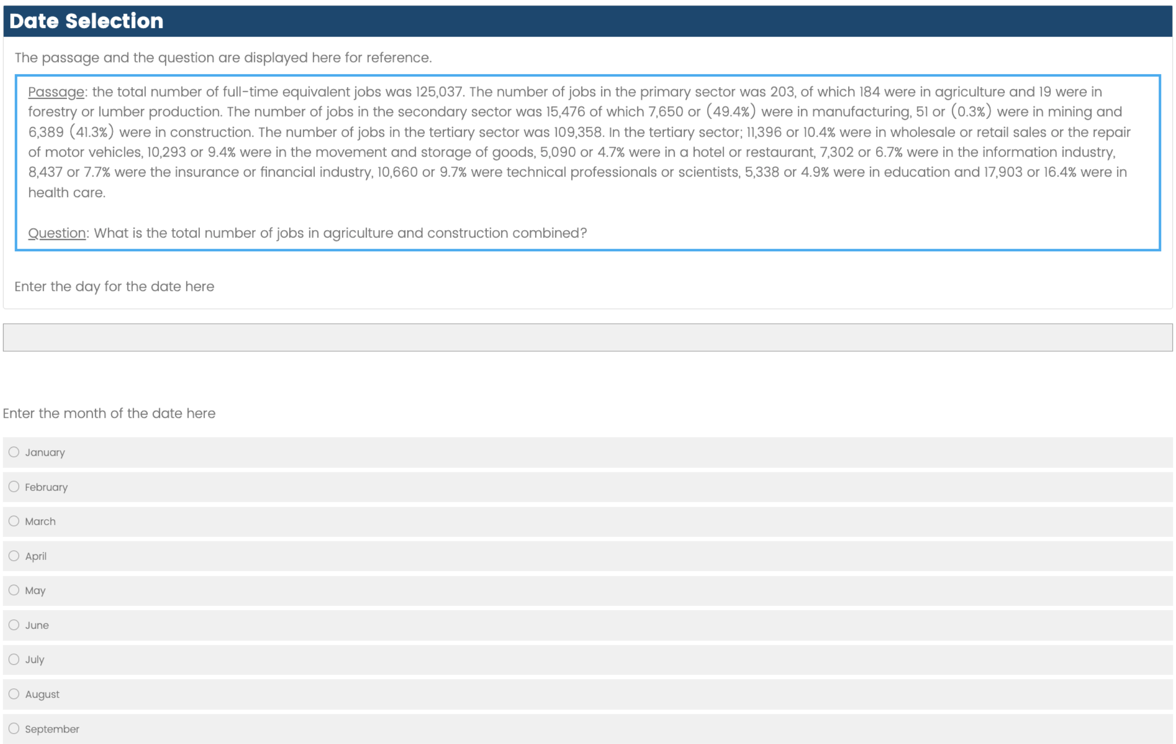
Explanation of Span Answers:Example Span Question 1 has 2 answers, and the spans of text in the original passage which contain those answers are separated, so it would take 2 separate spans to give the answer. Example span question 2 has only 1 answer. Note that this answers appears multiple times in the passage text, but the number of *unique* spans which answer the question is only 1.

Now please input the **unique** span(s) in the passage or question text which answer the question. If there are multiple spans for the answer, please separate each span with a newline.

Passage: Hoping to rebound from their road loss to the Jaguars, the Chargers went home for a Week 12 duel with the Baltimore Ravens. After a scoreless first quarter, San Diego struck first with kicker Nate Kaeding getting a 27-yard field goal. Afterwards, the Ravens took the lead with RB Willis McGahee getting a 1-yard TD run. Fortunately, the Chargers regained the lead with QB Philip Rivers completing a 35-yard TD pass to TE Antonio Gates (with a failed PAT), Kaeding kicking a 46-yard field goal, Rivers completing a 5-yard TD pass to WR Chris Chambers, and Kaeding kicking a 41-yard field goal. In the third quarter, San Diego increased its lead with Rivers and Gates hooking up with each other again on a 25-yard TD pass. Baltimore would manage to one final score as Ravens QB Kyle Boller completed a 13-yard TD pass to FB Le'Ron McClain. Fortunately, in the fourth quarter, the 'Bolts sealed the win with Kaeding nailing a 41-yard field goal. RB LaDainian Tomlinson (24 carries for 77 yards) became the fourth-fastest player and the 23rd player in NFL history to get 10,000 career rushing yards.

Question: How many more yards did Nate Kaeding's second field goal in the game exceed his first field goal by?

Figure 18: Specifying the **DROP** *date* answer (§6).



Date Selection

The passage and the question are displayed here for reference.

Passage: the total number of full-time equivalent jobs was 125,037. The number of jobs in the primary sector was 203, of which 184 were in agriculture and 19 were in forestry or lumber production. The number of jobs in the secondary sector was 15,476 of which 7,650 or (49.4%) were in manufacturing, 51 or (0.3%) were in mining and 6,389 (41.3%) were in construction. The number of jobs in the tertiary sector was 109,358. In the tertiary sector; 11,396 or 10.4% were in wholesale or retail sales or the repair of motor vehicles, 10,293 or 9.4% were in the movement and storage of goods, 5,090 or 4.7% were in a hotel or restaurant, 7,302 or 6.7% were in the information industry, 8,437 or 7.7% were the insurance or financial industry, 10,660 or 9.7% were technical professionals or scientists, 5,338 or 4.9% were in education and 17,903 or 16.4% were in health care.

Question: What is the total number of jobs in agriculture and construction combined?

Enter the day for the date here

Enter the month of the date here

- January
- February
- March
- April
- May
- June
- July
- August
- September

Figure 19: The beginning of the instruction for the **DROP** preference study (§5).

Instructions

Thanks for participating in this study!

Imagine you are a dataset creator who designed a benchmark to test a model's ability to perform discrete reasoning over passages of text.

Your dataset was designed in the following way:

(1) You collected passages from Wikipedia which had a narrative sequence of events and were amenable to complex questions.

(2) You hired workers to come up with questions about the passages, where each question could be answered in the context of a single Wikipedia passage. The questions elicited to require **complex linguistic understanding** and **discrete reasoning** over the passage to answer.

Examples of questions which require discrete reasoning

Passage	Question	Explanation
Before the UNPROFOR fully deployed, the HV clashed with an armed force of the RSK in the village of Nos Kalik, ... captured the village at 4:45 p.m. on **2 March 1992**. The JNA formed a battlegroup to counterattack the **next day**.	What date did the JNA form a battlegroup to counterattack after the village of Nos Kalik was captured?	Requires identification of multiple spans in the text and an **addition** step to compute the final answer.

Figure 20: The middle of the instruction for the **DROP** preference study.

In 1344, Momchil ... in **May 1345 the Turks ... devastated Bulgarian territories**. Soon after, on **7 July 1345**, Ottoman forces under Umur Beg defeated Momchil's army.	Which event happened later: the Turks marching from Asia Minor to devastate Bulgarian territories, or Ottoman forces defeating Momchil's army?	Requires **comparison** of two dates extracted from the passage.
In services, **Gurugram ranks number 1 in India in IT growth rate and existing technology infrastructure, and number 2 in startup ecosystem, innovation and livability** (Nov 2016).	What rank does Gurugram hold for livability?	Requires **selection** among multiple numeric options in the text.

Examples of questions which do NOT require discrete reasoning

Passage	**Question**	**Explanation**
Looking to stay in the playoff hunt, the Raiders welcomed the Dallas Cowboys to Oakland ... secure the 20-17 win.	What was the final score of the game?	Only a single numeric span needs to be identified.
The Eiffel Tower is located in Paris, France.	In which city is the Eiffel Tower	Only one textual span is needed to

Figure 21: The end of the instruction for the **DROP** preference study.

Examples of questions which do NOT require discrete reasoning

Passage	Question	Explanation
Looking to stay in the playoff hunt, the Raiders welcomed the Dallas Cowboys to Oakland ... secure the 20-17 win.	What was the final score of the game?	Only a single numeric span needs to be identified.
The Eiffel Tower is located in Paris, France.	In which city is the Eiffel Tower located?	Only one textual span is needed to answer.

(3) You then asked annotators to edit the questions in the following way:

- Compositional edit: Add an additional compositional reasoning step to the original question.

Your task now is:

- To evaluate pairs of questions and determine which question better meets your dataset
- To evaluate pairs of question edits and determine which edit better meets your dataset

Please note that you will need to do this process twice to complete a HIT.

Figure 22: After the general instructions, annotators are familiarized with DROP-like questions.

Questions

First you will evaluate two questions using a provided passage as context.

- **Definition:** The question must involve some kind of complex reasoning and must require looking at more than one part of the passage to answer.
- The question must be answerable using one of three answer types:
  - Spans of text from either the question or the passage
  - A Date
  - Numbers, for questions which explicitly state a specific unit of measurement (e.g. "How many *yards* did Brady run?")
- The goal of the questions are to test a models ability to perform discrete operations (such as addition, counting, sorting, etc) ovet the content of passages.
- The goal is **not** to create questions that we suspect are hard for some model in particular.

Example:

Passage: In ***1517, the seventeen-year-old King sailed to Castile***. There, his Flemish court . . . . ***In May 1518, Charles traveled to Barcelona in Aragon***.

Good Question: Where did Charles travel to first, Castile or Barcelona?
Invalid Question: In what year did Charles travel to Barcelona?

Explanation: The first question requires identification and association of dates with relevant events, and comparison of these dates to identify the answer. The second question requires only the identification and association of a single numerical value to an event.

Figure 23: The annotators are asked to give their preference between two DROP-like questions. One of the questions is generated, but that is not disclosed. The options are the same as for CondaQA; see Figure 12.

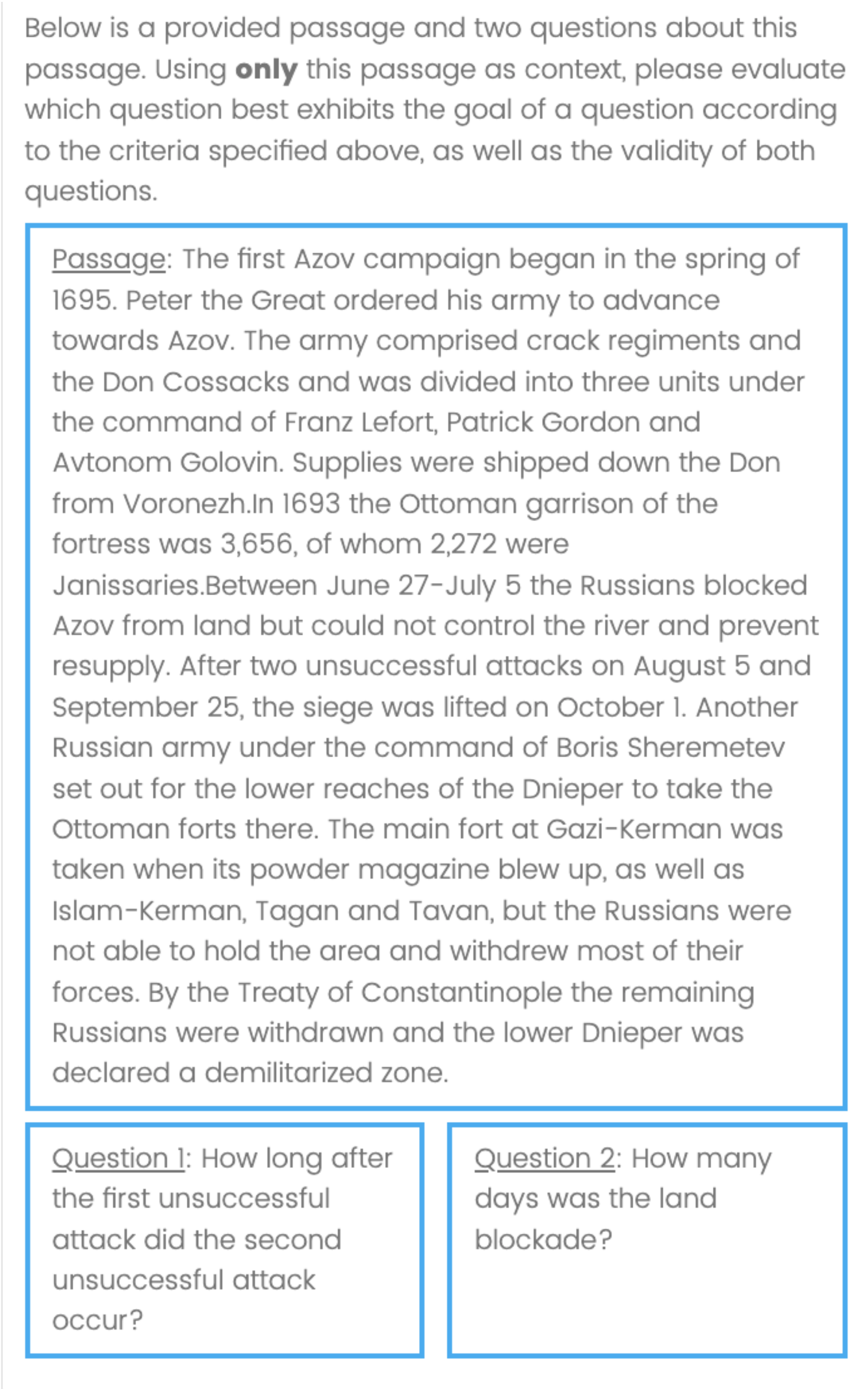


Figure 24: After selecting a preferred question (if any), annotators are familiarized with DROP compositional edits. The instruction continues in the next figure.

Compositional Edits

Now you will evaluate two edits of a question using a passage and the original question as context.

- **Definition:** The edit must add an additional reasoning step to the original question.
- The question edits must be answerable using one of three answer types:
  - Spans of text from either the question or the passage
  - A Date
  - Numbers, for questions which explicitly state a specific unit of measurement (e.g. "How many *yards* did Brady run?")
- The goal of this edit is to test a models robustness to the questions when compositional reasoning is introduced.
- The goal is **not** to create edits that we suspect are hard for some model in particular.

Example:

Passage: Coming off a bye week, the Bears battled rival Detroit Lions on Monday Night Football. The team kept the Lions from scoring until the fourth quarter, and forced four

Figure 25: The end of the instruction for the compositional edits.

Lions from scoring until the fourth quarter, and forced four takeaways. The first turnover forced was in the first half, when Lance Briggs stripped the ball from Mikel Leshoure, which was recovered by Julius Peppers. The second and third turnovers were forced in the third quarter on Zack Bowman's muffed punt recovery, and Brian Urlacher recovering Joique Bell's fumble. The final turnover occurred when Lions quarterback Matthew Stafford's pass was intercepted by D. J. Moore. The Bears struck first on Jay Cutler's touchdown pass to Brandon Marshall, and Robbie Gould kicked a field goal to increase the first half score to 10-0. During the game, Lions defensive tackle Ndamukong Suh threw Cutler to the ground, injuring his ribs. Cutler was eventually replaced by Jason Campbell for a play, before returning to the game. In the second half, Gould kicked another field goal, and prevented the Lions from scoring until the final 36 seconds of the game, when Stafford threw a 12-yard touchdown pass to Ryan Broyles to narrow the margin to six points, but the Bears sealed the victory by recovering the ensuring onside kick.

Original Question: How many touchdowns did the Lions score?

Good Edit: How many touchdowns did the Lions score in the second half?
Invalid Edit: How many touchdowns did the Bears score?

Explanation:The first edit adds an additional constraint to the counting problem in the original question. So now each touchdown is counted only after the event is associated with another event that identifies it as occuring within the second half of the game. The second edit is just changing the subject which the question is asking about. It does not alter the reasoning steps required to answer the question.

Figure 26: The annotators are asked to give their preference between two DROP-like compositional edits. One of the edits is generated, but that is not disclosed. The options are the same as for CondaQA; see Figure 8.

Below is a passage and the original question, followed by two compositional edits of the original question. Please read the passage, original question, and the compositional edits and evaluate which best exhibits the goals of a compositional edit.

Passage: Coming off their Monday night win over the Falcons, the Saints stayed at home for a Week 9 NFC South duel with the Carolina Panthers. New Orleans would trail in the first quarter as Panthers running back DeAngelo Williams got a 66-yard and a 7-yard touchdown run. In the second quarter, the Saints got on the board with a 23-yard field goal from kicker John Carney. Carolina would reply with kicker John Kasay getting a 32-yard field goal, yet New Orleans would close out the half with Carney's 25-yard field goal. In the third quarter, the Saints crept closer with a 10-yard touchdown by running back Pierre Thomas. The Panthers would reply with Kasay nailing a 25-yard field goal, yet New Orleans would close out the period with quarterback Drew Brees' 54-yard touchdown pass to wide receiver Robert Meachem. Afterwards, the Saints took command in the fourth quarter as Carney booted a 40-yard field goal, followed by defensive tackle Anthony Hargrove forcing Williams into a fumble and recovering it for a 1-yard touchdown run. With the win, the Saints improved to 8-0, which is the team's best start in franchise history.

Original Question: How long was the longest field goal?

Edit 1: What percentage of the total field goals in the game were made by John Carney?

Edit 2: How long was the longest field goal of the first half?