CONDAQA: A *Contrastive* Reading Comprehension Dataset for Reasoning about Negation

Abhilasha Ravichander[*]
Carnegie Mellon University
aravicha@cs.cmu.edu

Matt Gardner
Microsoft Semantic Machines
mattgardner@microsoft.com

Ana Marasović[*]
University of Utah
ana.marasovic@utah.edu

Abstract

The full power of human language-based communication cannot be realized without negation. All human languages have some form of negation. Despite this, negation remains a challenging phenomenon for current natural language understanding systems. To facilitate the future development of models that can process negation effectively, we present CONDAQA, the first English reading comprehension dataset which requires reasoning about the implications of negated statements in paragraphs. We collect paragraphs with diverse negation cues, then have crowdworkers ask questions about the *implications* of the negated statement in the passage. We also have workers make three kinds of edits to the passage—paraphrasing the negated statement, changing the scope of the negation, and reversing the negation—resulting in clusters of question-answer pairs that are difficult for models to answer with spurious shortcuts. CONDAQA features 14,182 question-answer pairs with over 200 unique negation cues and is challenging for current state-of-the-art models. The best performing model on CONDAQA (UNIFIEDQA-V2-3B) achieves only 42% on our consistency metric, well below human performance which is 81%. We release our dataset, along with fully-finetuned, few-shot, and zero-shot evaluations, to facilitate the development of future NLP methods that work on negated language.

1 Introduction

Negation is fundamental to human communication. It is a phenomenon of semantic opposition, relating one expression to another whose meaning is in some way opposed. Negation supports key properties of human linguistic systems such as contradiction and denial (Horn, 1989).

Despite the prevalence of negation, processing it effectively continues to elude models. Here are just a few of the many recently reported failures: "The model [BERT-Large trained on SQuAD] does not seem capable of handling...simple examples of negation" (Ribeiro et al., 2020). "We find that indeed the presence of negation can significantly impact downstream quality [of machine translation systems]" (Hossain et al., 2020a). "State-of-the-art models answer questions from the VQA...correctly, but struggle when asked a logical composition including negation" (Gokhale et al., 2020). *How can NLU systems meet this long-standing challenge?*

To facilitate systems that can process negation effectively, it is crucial to have high-quality evaluations that accurately measure models' competency at processing and understanding negation. In this work, we take a step toward this goal by contributing the first large-scale reading comprehension dataset, CONDAQA, focused on reasoning about negated statements in language.[1]

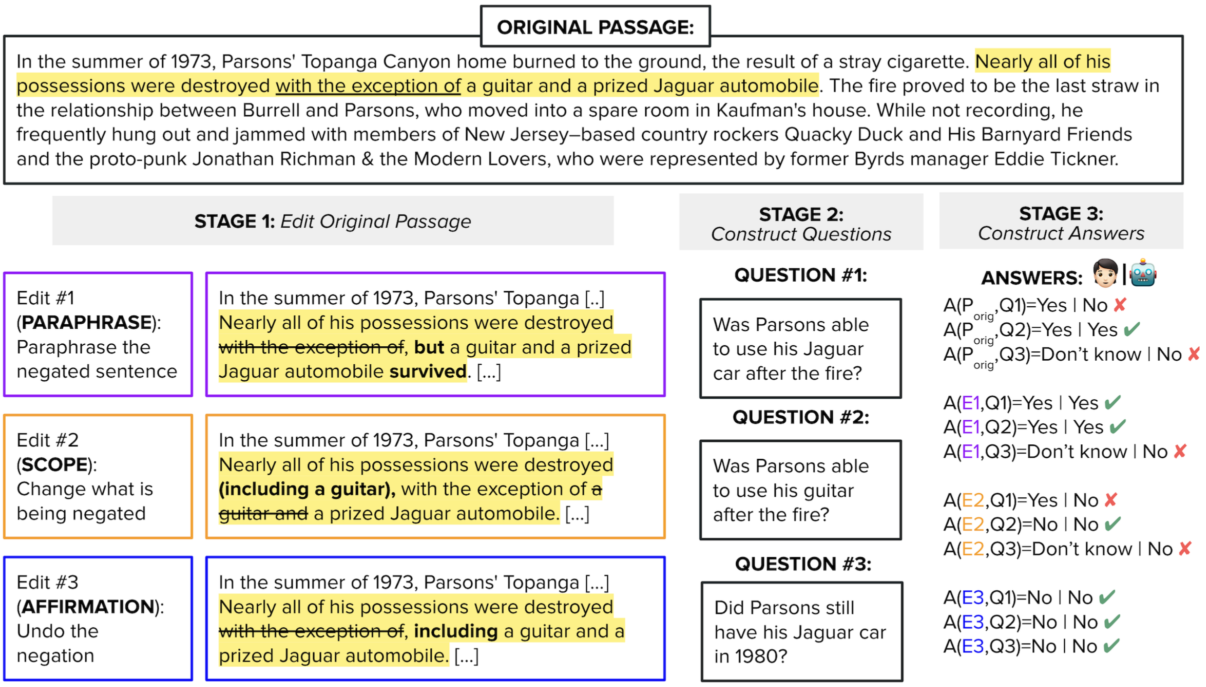
The three-stage annotation process we develop to construct CONDAQA is illustrated in Fig. 1. We first collect passages from English Wikipedia that contain negation cues, including single- and multi-word negation phrases, as well as affixal negation. In the first stage, crowdworkers make three types of modifications to the original passage: (1) they paraphrase the negated statement, (2) they modify the scope of the negated statement (while retaining the negation cue), and (3) they undo the negation. In the second stage, we instruct crowdworkers to ask challenging questions about the *implications* of the negated statement. The crowdworkers then answer the questions they wrote previously for the original and edited passages.

This process resulted in a dataset of 14,182 questions, covering a variety of negation cue types and over 200 unique negation cues, as well as a *con-*

[*] Work undertaken while Abhilasha Ravichander and Ana Marasović were at the Allen Institute for AI.

[1] **CO**ntrastively-annotated **N**egation **DA**taset of **Q**uestion-**A**nswer pairs

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 8729 - 8755
December 7-11, 2022 ©2022 Association for Computational Linguistics



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. The fire proved to be the last straw in the relationship between Burrell and Parsons, who moved into a spare room in Kaufman's house. While not recording, he frequently hung out and jammed with members of New Jersey–based country rockers Quacky Duck and His Barnyard Friends and the proto-punk Jonathan Richman & the Modern Lovers, who were represented by former Byrds manager Eddie Tickner.

STAGE 1: Edit Original Passage
STAGE 2: Construct Questions
STAGE 3: Construct Answers

Edit #1 (**PARAPHRASE**): Paraphrase the negated sentence

In the summer of 1973, Parsons' Topanga [..] Nearly all of his possessions were destroyed ~~with the exception of~~, **but** a guitar and a prized Jaguar automobile **survived**. [...]

Edit #2 (**SCOPE**): Change what is being negated

In the summer of 1973, Parsons' Topanga [...] Nearly all of his possessions were destroyed **(including a guitar),** with the exception of ~~a guitar and~~ a prized Jaguar automobile. [...]

Edit #3 (**AFFIRMATION**): Undo the negation

In the summer of 1973, Parsons' Topanga [...] Nearly all of his possessions were destroyed ~~with the exception of~~, **including** a guitar and a prized Jaguar automobile. [...]

QUESTION #1: Was Parsons able to use his Jaguar car after the fire?

QUESTION #2: Was Parsons able to use his guitar after the fire?

QUESTION #3: Did Parsons still have his Jaguar car in 1980?

ANSWERS: 🧑 | 🤖

A(P_{orig},Q1)=Yes | No ✗
A(P_{orig},Q2)=Yes | Yes ✓
A(P_{orig},Q3)=Don't know | No ✗

A(E1,Q1)=Yes | Yes ✓
A(E1,Q2)=Yes | Yes ✓
A(E1,Q3)=Don't know | No ✗

A(E2,Q1)=Yes | No ✗
A(E2,Q2)=No | No ✓
A(E2,Q3)=Don't know | No ✗

A(E3,Q1)=No | No ✓
A(E3,Q2)=No | No ✓
A(E3,Q3)=No | No ✓

Figure 1: CONDAQA three-stage collection procedure. The original passage is selected by a crowdworker from a given set of 10 passages. 🧑 Gold answers given by crowdworkers; 🤖 Answers predicted by InstructGPT (`text-davinci-002`) prompted with 8 shots. See §2 for more details about each stage.

trastive dataset, with passages that are lexically similar to each other but that may induce different answers for the same questions. To perform well on CONDAQA, models must be able to reason about the implications of negated statements in text. In addition to accuracy, the contrastive nature of CONDAQA enables us to measure the *consistency* of models—i.e., the extent to which models make correct predictions on closely-related inputs.

We extensively benchmark baseline models on CONDAQA in three training data regimes: using all training examples, using only a small fraction (few-shot), or not using any examples (zero-shot). We show that CONDAQA is challenging for current models. Finetuning UNIFIED-QA-3B (Khashabi et al., 2022)—which was trained on 20 QA datasets—on CONDAQA, achieves the best result of 73.26% compared to human accuracy of 91.49%. Further, we find that models are largely inconsistent; the best model achieves a consistency score of only 42.18% (40% below human consistency). This very low consistency score demonstrates that handling negation phenomena is still a major unresolved issue in NLP, along with sensitivity to contrasting data more generally. The dataset and baselines are available at `https://github.com/AbhilashaRavichander/CondaQA`.

2 CONDAQA Data Collection

This section describes our goals in constructing CONDAQA and our data collection procedure.

Design Considerations Our goal is to evaluate models on their ability to process the contextual implications of negation. We have the following four desiderata for our question-answering dataset:

1. The dataset should include a wide variety of ways negation can be expressed.
2. Questions should be targeted towards the *implications* of a negated statement, rather than the factual content of what was or wasn't negated, to remove common sources of spurious cues in QA datasets (Kaushik and Lipton, 2018; Naik et al., 2018; McCoy et al., 2019).
3. The dataset should feature contrastive groups: passages that are closely-related, but that may admit different answers to questions, in order to reduce models' reliance on potential spurious cues in the data and to enable more robust evaluation (Kaushik et al., 2019; Gardner et al., 2020).
4. Questions should probe the extent to which models are sensitive to how the negation is expressed. In order to do this, there should be contrasting passages that differ only in their negation cue or its scope.

Dataset Construction Overview We generate questions through a process that consists of the following steps, as shown in Figure 1:

1. We extract passages from Wikipedia which contain negated phrases.
2. We show ten passages to crowdworkers, and allow them to choose a passage they would like to work on.
3. Crowdworkers make three kinds of edits to the passage: (i) paraphrasing the negated statement, (ii) changing the scope of the negation, (iii) rewriting the passage to include an affirmative statement in place of the negated statement. For all three kinds of edits, the crowdworkers modified the passage as appropriate for internal consistency.
4. Crowdworkers ask questions that target the implications of a negated statement in the passage, taking passage context into account.
5. Crowdworkers provide answers to the constructed questions for the Wikipedia passage, as well as the three edited passages.

Further, we validate the development and test portions of CONDAQA to ensure their quality.

Passage Selection We extract passages from a July 2021 version of Wikipedia that contain either single-word negation cues (e.g., 'no') or multi-word negation cues (e.g., 'in the absence of'). We use negation cues from (Morante et al., 2011; van Son et al., 2016) as a starting point which we extend. Our single-word negation cues include affixal negation cues (e.g., '*il*-legal'), and span several grammatical categories including:

1. **Verbs**: In this novel, he took pains to avoid the scientific impossibilities which had bothered some readers of the "Skylark" novels.
2. **Nouns**: In the absence of oxygen, the citric acid cycle ceases.
3. **Adjectives**: Turning the club over to managers, later revealed to be honest people, still left Wills in desperate financial straits with heavy debts to the dishonest IRS for taxes.
4. **Adverbs**: Nasheed reportedly resigned involuntarily to forestall an escalation of violence;
5. **Prepositions**: Nearly half a century later, after Fort Laramie had been built without permission on Lakota land.
6. **Pronouns**: I mean, nobody retires anymore.
7. **Complementizers**: Leave the door ajar, lest any latecomers should find themselves shut out.
8. **Conjunctions**: Virginia has no 'pocket veto' and bills will become law if the governor chooses to neither approve nor veto legislation.
9. **Particles**: Botham did not bat again.

Crowdworker Recruitment We use the Crowdaq platform (Ning et al., 2020) to recruit a small pool of qualified workers to contribute to CONDAQA. We provide instructions, a tutorial and a qualification task. Workers were asked to read the instructions, and optionally to also go through the tutorial. Workers then took a qualification exam which consisted of 12 multiple-choice questions that evaluated comprehension of the instructions. We recruit crowdworkers who answer $>$70% of the questions correctly for the next stage of the dataset construction task. In total, 36 crowdworkers contributed to CONDAQA. We paid 8 USD/HIT, which could on average be completed in less than 30 minutes. Each HIT consisted of choosing a passage, making edits to the passage, creating questions, and answering those questions.

Contrastive Dataset Construction We use Amazon Mechanical Turk to crowdsource question-answer pairs about negated statements. Each question is asked in the context of a negated statement in a Wikipedia passage.

In the first stage of the task, we show crowdworkers ten selected passages of approximately the same length and let them choose which to work on. This allows crowdworkers the flexibility to choose passages which are easy to understand, as well as to choose passages which are conducive to making contrastive edits (for example, it may be difficult to reverse the negation in a passage about 'Gödel's *incompleteness* theorems').

After selecting a passage, crowdworkers make three kinds of edits to the original Wikipedia passage (Fig. 1): (1) they rewrite the negated statement such that the sentence's meaning is preserved (PARAPHRASE EDIT); (2) they rewrite the negated statement, changing the scope of the negation (SCOPE EDIT); and (3) they reverse the negated event (AFFIRMATIVE EDIT). We ask crowdworkers to additionally make edits outside of the negated statement where necessary to ensure that the passage remains internally consistent.

In the second stage of the task, the crowdworker asks at least three questions about the implications of the negated statement in the original Wikipedia passage. We encourage the construction of good questions about implications by providing several examples of such questions, as well as by sending bonuses to creative crowdworkers, ranging from 10-15. Crowdworkers can group these questions, to indicate questions that are very similar to each other, but admit different answers.

In the final stage of this task, crowdworkers provide answers to the questions, in context of the Wikipedia passages as well as for the three edited passages. The answers to the questions are required to be either Yes/No/Don't Know, a span in the question, or a span in the passage. Following best practices for crowdsourcing protocols described in the literature (Nangia et al., 2021), we provide personalized feedback to each crowdworker based on their previous round of submissions, describing where their submission was incorrect, why their submission was incorrect, and what they could have submitted instead. In all, we provide over 15 iterations of expert feedback on the annotations. We collect this data over a period of ∼seven months.

Data Cleaning and Validation In order to estimate human performance, and to construct a high-quality evaluation with fewer ambiguous examples, we have five verifiers provide answers for each question in the development and test sets. Crowdworkers were given passages, as well as the passage edits and questions contributed in the previous stage of our task. In each HIT, crowdworkers answered 60 questions in total, spanning five passage sets. We found there was substantial inter-annotator agreement; for the test set we observed a Fleiss' κ of 63.27 for examples whose answers are Yes/No/Don't know (97% of examples), 62.75 when answers are a span in the question (2% of examples), and 48.54 when answers were indicated to be a span in the passage (1% of examples). We only retain examples in the test and development sets where *at least four annotators* agreed on the answer. However, since this procedure results in few questions with 'don't know' as the answer, we include an additional stage where we (the authors) manually verify and include questions where 'don't know' was the answer provided by the question author. As a result, we discard 1,160 instances from the test set, and 270 from the development set.

3 CondaQA Data Analysis

In this section, we provide an analysis of the passages, questions, edits, and answers in CondaQA.

	Train	Dev	Test
# Passages	474	115	700
Average passage length	130.02	131.24	131.0
Negated statement length	28.12	29.96	28.0
# Unique negation cues	134	62	171
# Unseen negation cues	-	18	75
# Questions	5832	1110	7240
Average Question Length	24.2	26.38	24.35
# Questions w/ >20 tokens	2836	650	3616
# Distinct question words	6045	2235	7603

Table 1: Dataset statistics of CondaQA. Passage statistics are computed on Wiki passages but not on edits.

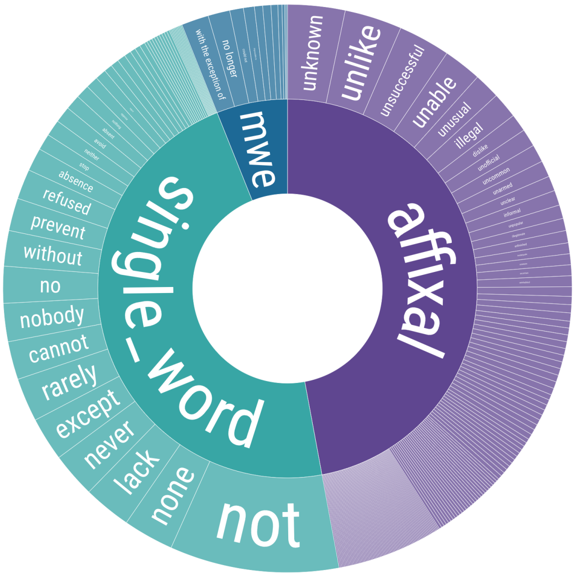


Figure 2: Distribution of negation cues in CondaQA. Inner circle represents distribution of negation cue types by their frequency and the outer circle represents cues.

Descriptive statistics are provided in Table 1.

Negation Cues Negation is expressed in many complex and varied ways in language (Horn, 1989). To characterize the distribution of types of negated statements in CondaQA, we analyze the negation cues in Wikipedia passages that annotators could select. Figures 2 and 4 (Appendix) show that the distribution over these cues and their grammatical roles is considerably diverse. Moreover, there are 219 unique cues in CondaQA and 75 novel cues in the test set that are unseen in the training data. This is a substantially wider range of negation cues than what is included in prior work; see Appendix A for a detailed comparison.

Reasoning Type	Passage Snippet	Question	Answer	Explanation
Social Norms (10%)	On October 8, 1883, the US patent office ruled that Edison's patent was based on the work of William E. Sawyer and was, therefore, **invalid** . Litigation continued for nearly six years. In 1885, Latimer switched camps and started working with Edison.	From the information given in the passage, would you say that coincidence is the most charitable explanation for what was essentially the same innovation, in much the same way that Newton and Leibniz seemingly discovered calculus independently, without knowing of the other's progress?	YES	Plagarism is frowned upon in society, more so than accidentally reaching the same conclusions as someone else.
Psychology (9%)	[...] Disraeli later romanticised his origins, claiming his father's family was of grand Iberian and Venetian descent; in fact Isaac's family was of no great distinction [...] Historians differ on Disraeli's motives for rewriting his family history: [...] Sarah Bradford believes “his **dislike** of the commonplace would not allow him to accept the facts of his birth as being as middle-class and undramatic as they really were”.	Would Disraeli have been flattered by a biography that explored his middle class upbringing, according to Bradford?	NO	A person such as Disraeli who wants to project a grandiose image of themselves is likely to be unhappy when people discuss mundane aspects about his upbringing.
Cause and Effect (7%)	Oil produced from palm fruit is called ‘red palm oil’ or just ‘palm oil’... In its **unprocessed** state, red palm oil has an intense deep red color because of its abundant carotene content. [...]	Would a consumer who was primarily interested in the eye-health benefits of carotenes and lycopene want to shop for palm oils by their color, rather than listening to marketing slogans such as “extra virgin” or “minimally processed”?	YES	A high carotene content causes a deep red color, so a person searching for things with high carotene content can look at their color.

Table 2: Examples of questions that target the implications of negated statements in CondaQA and reasoning steps to correctly answer the questions. Negated statements are in blue. Categories inspired by LoBue and Yates (2011). Expanded analysis is shown in the Appendix (Table 12).

Commonsense inferences We assess commonsense inferences types required to answer CondaQA questions. We sample 100 questions from the test set and manually annotate the dimensions of commonsense reasoning required to answer them. Table 2 shows some of these reasoning types (the full version in Table 12 in the Appendix).

Editing Strategies Recall that the passages with negated statements are sourced from Wikipedia and crowdworkers make three kinds of edits (Fig. 1). Through a qualitative analysis of the data, we identify commonly employed edit strategies (Tables 3 and 13). We also analyze to what extent edits cause an answer to change. We find that the affirmative edits change the answers of 77.7% of questions from the original Wikipedia passage, and the scope edits change the answer of 70.6% of questions.

Potential edit artifacts Because we had crowdworkers edit Wikipedia paragraphs, a potential concern is that the edited text could be unnatural and give spurious cues to a model about the correct answer. We ran two tests to try to quantify potential bias in this edited data. First, we trained a BERT model (Devlin et al., 2019) to predict the edit type given just the passage. The model performs only a little better than random chance (34.4%), most of the improvement coming from the ability to sometimes detect affirmative edits (where the negation cue has been removed). Second, we compared the perplexity of the original paragraphs to the perplexity of the edited paragraphs, according to the GPT language model (Radford et al., 2018), finding that they are largely similar. Details for both of these experiments are in Appendix B.

4 Baseline Performance on CondaQA

We now evaluate state-of-the-art models' abilities to solve instances of CondaQA. We evaluate models that we train either on the entire CondaQA training data or few examples, as well as zero-shot models. We use two classes of metrics:

Accuracy The percentage of predictions which match the ground truth answer. If the answer is a span, this metric measures whether the prediction

Revision Strategy	Edited Passage
	Paraphrase Edit
Complement substitution	Though Philby claimed publicly in January 1988 that he did not regret his decisions and that ~~he missed nothing about England except~~ the only things he missed about England were some friends, Colman's mustard, and Lea & Perrins Worcestershire sauce...
Synonym substitution	Local tetanus is ~~an uncommon~~a rare form of the disease and it causes persistent contractions of muscles in the same area of the sufferer's body as where the original injury was made.
Antonym substitution	The population of the Thirteen States was ~~not homogeneous~~ heterogeneous in political views and attitudes.
Ellipsis	~~Sunni scholars put trust in narrators such as Aisha, whom Shia reject~~While the Shia tend to reject narrators such as Aisha, Sunni scholars tend to trust them.
	Scope Edit
Complement inversion	~~Sunni~~Shia scholars put trust in narrators such as Aisha, whom ~~Shia~~Sunni reject.
Superset-subset	During the coronavirus outbreak of 2020, alcohol sales~~, and even the~~ were made illegal, but the transportation of alcohol outside of one's home~~, was made illegal~~ remained legal.
Temporal shift	As the new Emperor could not exert his constitutional powers ~~until~~once he came of age, a regency was set up by the National Assembly.
Veridicality	Contrary to assumptions that he was illiterate, on arrival he was given aptitude tests which determined that ~~he was illiterate~~not only could he read the questions and respond in writing, but he also had an above-average IQ of 109.

Table 3: Examples of revision strategies employed by crowdworkers for paraphrase and scope edits. Categories for paraphrases are inspired by Bhagat and Hovy (2013). The negation cue is shown in blue and newly-inserted text is in red. Expanded analysis is shown in the Appendix (Table 13).

matches the ground truth answer exactly.

Group Consistency CondaQA's dense annotations enable us to study model robustness through group consistency. We wish to measure whether a model correctly captures how the presence of negated phrases influences what can be inferred from a paragraph. Measuring this requires varying (and sometimes removing) the negated phrases and seeing how the model responds (see Table 14 in the Appendix); it is only by looking at consistency across these perturbations that we can tell whether a model understands the phenomena in question (Gardner et al., 2020). Specifically, for a group of minimally-different instances, consistency measures whether the prediction matches the ground truth answer for every element in that group. We consider two types of groups: (a) *Question-level consistency*: each group is formed around a question and the answers to that question for the original Wikipedia passage, as well as the three edited passage instances (All), (b) *Edit-level consistency*: each group is formed around a question, the answers to that question for the original Wikipedia passage, and only one of the edited passages (Paraphrase consistency, Scope consistency, and Affirmative consistency). To compute consistency, we use the 5,608 questions in the test set that have (passage, answer) pairs for all four edit types (excluding any question where at least one passage was removed during validation).

4.1 Models and Controls

The baseline models that we benchmark on CondaQA are listed in Table 4. We categorize them based on whether they use (a) all of the training data we provide (full finetuned), (b) a small fraction of the available training data (few-shot), (c) no training data (zero-shot), and on (d) whether they measure dataset artifacts (controls).

For **full finetuning**, we train and evaluate three BERT-like models: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2021b,a), in addition to UnifiedQA-v2 (Khashabi et al., 2022), a T5 variant (Raffel et al., 2020) that was further specialized for QA by training the model on 20 QA datasets. More information about these models is given in Appendix C.1. We study Base, Large, and 3B sizes of UnifiedQA-v2. Each fully-finetuned model is trained with 5 random seeds, and we report the average performance across seeds on the entire test set.

In the **few-shot** setting with 8–9 shots, we evaluate UnifiedQA-v2-{Base, Large, 3B} (Khashabi et al., 2022), GPT-3 (`davinci`; Brown et al., 2020), and a version of InstructGPT$_{orig}$ (Ouyang et al., 2022) known as `text-davinci-002`; henceforth referred to as 🪄InstructGPT. We additionally prompt 🪄InstructGPT with chain of thoughts (CoT; Wei et al., 2022) as this should be beneficial for reasoning tasks. We do prompt-based finetuning of UnifiedQA-v2 (i.e., change its parameters) and in-context learning of the GPT models (i.e., we do not change their parameters). Besides these models, in the **zero-shot** setting, we also evaluate UnifiedQA-v2-11B and FLAN-T5-11B (Chung et al., 2022), a T5 variant that was further trained with instruction finetuning and CoT data. Details of few- and zero-shot settings are given in Appendix C.2. Due to the cost of the OpenAI API and sensitivity of few-shot learning to the choice of few examples (Zhao et al., 2021; Logan IV et al., 2022; Perez et al., 2021), we evaluate few- and zero-shot models as follows. We split the train/test sets into five disjoint sets, sample 9 shots from each train subset, evaluate models on such five train-test splits, and report the average performance across them. On average each test split contains 1448 instances.

We evaluate **heuristic** baselines to measure the extent to which models can use data artifacts to answer CondaQA questions. These baselines can answer questions correctly only if there is bias in the answer distribution given a question or other metadata since they do not get paragraphs. We train UnifiedQA-v2-large on just: (i) (question, answer) pairs, (ii) (question, edit type, answer) triples where the edit type denotes whether the passage was a paraphrase, scope edit, etc., and (iii) (question, negation cue, answer) triples. We find these baselines do little better than just answering “No”.

Human Performance We measure human performance on CondaQA development and test sets. Every question was answered by five crowdworkers. To evaluate human performance, we treat each answer to a question as the human prediction in turn, and compare it with the most frequent answer amongst the remaining answers. For questions where the gold answer was decided by experts (§2), we treat each answer as the human prediction and compare it to the gold answer. Human accuracy is 91.94%, with a consistency score of 81.58%.

5 Results

Model performance on CondaQA is given in Table 4. The best performing model is fully finetuned UnifiedQA-v2-3B with an accuracy of 73.26% and overall consistency of 42.18%, where the estimated human accuracy is 91.94% and consistency 81.58%. This gap shows that CondaQA questions are both answerable by humans, and challenging for state-of-the-art models.

We create a contrastive dataset to be able to measure consistency as measuring models’ ability to *robustly* predict answers across small input perturbations can provide a more accurate view of linguistic capabilities (Gardner et al., 2020). Here, there is a gap of ∼40% in consistency between humans and the best model. Models are most robust to paraphrase edits: if a model answers a question correctly for the original passage, it is likely to be robust to changes in how that negation is expressed. We observe that the heuristic-based baselines exhibit low consistency, suggesting the consistency metric may be a more reliable measure than accuracy to evaluate models’ ability to process negation. Thus, mainstream benchmarks should consider including consistency as a metric to more reliably measure progress on language understanding.

Few- and zero-shot baselines do not match fully finetuned models’ performance, but considerably improve over the majority baseline. For UnifiedQA-v2 in particular, this suggests that some reasoning about implications of negation is acquired during pretraining. Surprisingly, UnifiedQA-v2 few-shot performance is worse than zero-shot. While this behavior has been reported for in-context learning with GPT-3 (Brown et al., 2020; Xie et al., 2022), we did not expect to observe this for a finetuned model.[2] UnifiedQA-v2-3B finetuned with a few examples is comparable to 🪄InstructGPT (`text-davinci-002`; at least 175B parameters) with in-context learning. Chain-of-thought prompting (CoT) notably improves the performance of 🪄InstructGPT, especially in terms of the most challenging metrics: scope and affirmative consistency. In the zero-shot setting, the 11B version of UnifiedQA-v2 performs the best, while the base version of only 220M parameters is comparable to 🪄InstructGPT. UnifiedQA-v2-11B is also better than FLAN-T5-XXL (a 11B-parameter model as well). Given that UnifiedQA-v1 (Khashabi et al., 2020) has been effective for tasks beyond QA (Bragg et al., 2021; Marasović et al., 2022), this result suggests that UnifiedQA

[2] A lower learning rate or less training steps do not help improve UnifiedQA-v2 few-shot performance.

Model	# Param	Accuracy	Consistency	Paraphrase Consistency	Scope Consistency	Affirmative Consistency
Heuristics						
Majority	-	47.75	1.35	51.50	16.48	8.71
Question-Only	770M	52.32	11.80	48.15	24.42	24.02
Edit-Type Only	770M	53.85	12.44	50.54	25.83	25.26
Negation-Cue Only	770M	56.79	14.89	55.96	29.17	27.89
Fully Supervised						
BERT-Large	340M	46.3	2.20	44.21	14.76	12.35
RoBERTa-Large	355M	54.08	13.64	51.64	26.53	27.18
DeBERTa-v2-XLarge	710M	54.01	13.37	52.72	25.61	25.69
DeBERTa-v3-Large	304M	57.09	18.02	56.50	30.13	30.93
UnifiedQA-v2-Base	220M	57.94	17.49	54.62	30.39	32.98
UnifiedQA-v2-large	770M	66.72	30.20	63.98	43.68	46.45
UnifiedQA-v2-3B	3B	**73.26**	**42.18**	**72.80**	**55.68**	**57.22**
Few-Shot						
UnifiedQA-v2-Base	220M	52.58	11.97	50.11	24.19	25.03
UnifiedQA-v2-Large	770M	55.84	16.80	56.05	30.25	29.93
UnifiedQA-v2-3B	3B	61.14	22.52	62.05	35.71	35.41
GPT-3	175B	52.42	5.22	48.94	23.31	20.31
🪄InstructGPT	N/A	60.88	20.30	63.92	36.40	33.98
🪄InstructGPT + CoT	N/A	**66.28**	**27.28**	**64.27**	**45.08**	**44.91**
Zero-Shot						
UnifiedQA-v2-Base	220M	55.65	16.20	52.47	29.23	30.83
UnifiedQA-v2-Large	770M	61.74	23.07	61.16	37.14	37.14
UnifiedQA-v2-3B	3B	69.41	34.91	70.71	47.94	49.74
UnifiedQA-v2-11B	11B	**73.11**	**40.02**	**75.48**	**53.72**	**54.12**
FLAN-T5-XXL	11B	67.53	31.61	67.43	45.45	47.86
GPT-3	175B	43.72	1.28	41.33	10.67	10.89
🪄InstructGPT	N/A	54.00	16.32	55.54	29.87	27.81
Human Performance						
Human	-	**91.94**	**81.58**	**93.65**	**86.49**	**88.22**

Table 4: Model performance on CondaQA. All heuristics are built on top of UnifiedQA-Large. **Boldface** indicates the best model on each metric for every training setup (*Supervised*, *Few-Shot*, *Zero-Shot*). Supervised models are trained and evaluated across five random seeds using the full train and test sets. Due to the cost of OpenAI API, for few- and zero-shot models we report the average performance across five train-test splits. For more details and description of metrics see §4. GPT-3 version: `davinci`; 🪄InstructGPT version: `text-davinci-002`.

models are strong but overlooked baselines in recent works on large-scale models.

6 Analysis

While examining model errors, we find UnifiedQA-v2-large has a negative correlation with question length (Figure 7 in Appendix D). Humans can still reliably answer such long questions that are frequent in CondaQA. We also analyze the performance of UnifiedQA-v2-large across answer types, finding that: (i) the model performs best when the answer is “No”, (ii) it almost never predicts “Don’t know”, and (iii) its performance on span extraction questions is in-between those two extremes (Figure 8 in Appendix D). UnifiedQA-v2-3B exhibits similar behavior, with improved performance on questions which admit “Don’t know” as an answer.

We analyze questions across the Wikipedia passages and the passages with edited scopes, with the focus on: (i) instances where the true answer does not change with the edited scope and the model should be stable, and (ii) instances where the true answer does change and the model should be sensitive to the edit. We find that when the fully-finetuned UnifiedQA-v2-3B (the best-performing model) answers the question correctly for the Wikipedia passage, it only produces the answer correctly for 63.23% of questions where the scope edit induces a different answer. In contrast, the model answers questions correctly for 91.03% of the instances where the answer does not change with the scope edit.[3] This suggests the model is not sensitive to changes of the scope of negated statements.

We also analyze to what extent UnifiedQA-v2-

[3] Computed over the subset of questions which had high agreement for all four passages.

3B distinguishes between negated statements and their affirmative counterparts. We examine model predictions for 1080 sample pairs where the answer changes when the negation is undone. For 43.52% of these, the model changes its predictions. This suggests, in contrast to previous work (Kassner and Schütze, 2020; Ettinger, 2020), that models are sensitive to negated contexts to some extent.

7 Related Work

In Aristotle's *de Interpretatione*, all declarative statements are classified as either affirmations or negations used to affirm or contradict the occurrence of events (Ackrill et al., 1975). Negation is expressed through a variety of formulations (Horn, 1989) and is prevalent in English corpora (Hossain et al., 2020c). Despite that, evidence from multiple tasks that require language understanding capabilities—such as NLI (Naik et al., 2018), sentiment analysis (Li and Huang, 2009; Zhu et al., 2014; Barnes et al., 2019), paraphrase identification (Kovatchev et al., 2019), machine translation (Fancellu and Webber, 2015; Hossain et al., 2020a), and QA (Ribeiro et al., 2020; Sen and Saffari, 2020)—identify negation as a challenging semantic phenomenon for models. Hossain et al. (2022) analyze negation in 8 NLU datasets and conclude: "new corpora accounting for negation are needed to solve NLU tasks when negation is present". We expect CondaQA will help.

Negation Annotations Jiménez-Zafra et al. (2020) overview datasets with negation as the main phenomenon and mention the following: BioScope (Vincze et al., 2008), ProbBank Focus (Blanco and Moldovan, 2011), ConanDoyle-neg (Morante and Daelemans, 2012), SFU Review$_{\text{EN}}$ (Konstantinova et al., 2012), NEG-DrugDDI (Bokharaeian and Díaz, 2013), NegDDI-Drug (Bokharaeian et al., 2014), and DT-Neg (Banjade and Rus, 2016). These datasets are small (<4K) and annotated with different schemes and guidelines as there is no established formalism for negation due to its complexity—the case when the QA format is useful (Gardner et al., 2019a). There are datasets focused on negation cue/scope/focus detection, or negated event recognition (Morante and Blanco, 2012; Reitan et al., 2015; Fancellu et al., 2017; He et al., 2017; Li and Lu, 2018; Hossain et al., 2020b). Jiménez-Zafra et al. (2020) assert that the lack of large datasets remains a major obstacle.

Probing Negation Ettinger (2020) introduces a dataset of 72 sentences for probing understanding of negation. Kassner and Schütze (2020) analyze factual knowledge in the presence of negation. Several works have recently constructed challenge sets that focus on negation for existing NLI datasets (Cooper et al., 1996; Dagan et al., 2005; Giampiccolo et al., 2007). Hartmann et al. (2021) introduce a multilingual dataset for probing negation based on XNLI/MNLI (Conneau et al., 2018; Williams et al., 2018). Hossain et al. (2020c) analyze negation in three existing NLI datasets and find they are unsuitable for studying how NLI models handle negation. They introduce a new benchmark of 4.5K instances based on 1.5K seed instances from the three NLI datasets. Geiger et al. (2020) construct a dataset targeting the interaction between lexical entailment and negation, finding that models trained on general-purpose NLI datasets do not perform well, but finetuning with their dataset is sufficient to address this failure. In contrast to several of these works, we contribute training data and find that simply finetuning on these examples is not sufficient to address the challenges in CondaQA. See Appendix §A for a detailed comparison.

Improving Negation Understanding Efforts to improve models' negation abilities that can be studied on CondaQA are: unlikelihood training (Hosseini et al., 2021), NLI data (Kim et al., 2019), commonsense knowledge (Jiang et al., 2021), multitasking (Moore and Barnes, 2021), extra MLM (Khandelwal and Sawant, 2020; Truong et al., 2022).

8 Conclusion

Negation supports key properties of human linguistic systems such as the ability to distinguish between truth and falsity. We present CondaQA, a QA dataset that contains 14,182 examples to evaluate models' ability to reason about the implication of negated statements. We describe a procedure for contrastive dataset collection that results in challenging questions, present a detailed analysis of the dataset, and evaluate a suite of strong baselines in fully-finetuned, few-shot, and zero-shot settings. We evaluate models on both their accuracy and consistency, and find that this dataset is highly challenging—even the best-performing model is 18 points lower in accuracy than our human baseline, and about 40 points lower in consistency. We expect that CondaQA will facilitate NLU systems that can handle negation.

Limitations

In this work, we contribute CONDAQA, a dataset to facilitate the development of models that can process negation. Though CONDAQA currently represents the largest NLU dataset that evaluates a model's ability to process the implications of negation statements, it is possible to construct a larger dataset, with more examples spanning different answer types. Further, CONDAQA is an English dataset, and it would be interesting to extend our data collection procedures to build high-quality resources in non-English languages. Finally, while we attempt to extensively measure and control for artifacts in CONDAQA, it is possible that the dataset has hidden artifacts that we did not study.

Acknowledgements

The authors are grateful to Aakanksha Naik, Aditya Potukuchi, Rajat Kulshreshtha, Shruti Rijhwani, and Ashu Ravichander for feedback on the crowdsourcing task, and to Noah Smith and Jena Hwang for valuable discussions related to this project. The authors would also like to thank all members of the AllenNLP team and the three anonymous reviewers for their valuable feedback.

References

John L Ackrill et al. 1975. *Categories and De interpretatione*. Clarendon Press.

Rajendra Banjade and Vasile Rus. 2016. DT-neg: Tutorial dialogues annotated for negation scope and focus in context. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3768–3771, Portorož, Slovenia. European Language Resources Association (ELRA).

Jeremy Barnes, Lilja Øvrelid, and Erik Velldal. 2019. Sentiment analysis is not solved! assessing and probing sentiment classification. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 12–23, Florence, Italy. Association for Computational Linguistics.

Rahul Bhagat and Eduard Hovy. 2013. Squibs: What is a paraphrase? *Computational Linguistics*, 39(3):463–472.

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

Eduardo Blanco and Dan Moldovan. 2011. Semantic representation of negation using focus detection. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 581–589, Portland, Oregon, USA. Association for Computational Linguistics.

Behrouz Bokharaeian and Alberto Díaz. 2013. NIL_UCM: Extracting drug-drug interactions from text through combination of sequence and tree kernels. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 644–650, Atlanta, Georgia, USA. Association for Computational Linguistics.

Behrouz Bokharaeian, Alberto Díaz, Mariana Neves, and Virginia Francisco. 2014. Exploring negation annotations in the drugddi corpus. In *Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing (BIOTxtM),*.

Jonathan Bragg, Arman Cohan, Kyle Lo, and Iz Beltagy. 2021. FLEX: unifying evaluation for few-shot NLP. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 15787–15800.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. arXiv:2210.11416.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Johan Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, and Massimo Poesio. 1996. Using the framework. Technical report.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In Joaquin Quiñonero Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché-Buc, editors, *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190. Springer.

Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5925–5932, Hong Kong, China. Association for Computational Linguistics.

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

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Federico Fancellu, Adam Lopez, Bonnie Webber, and Hangfeng He. 2017. Detecting negation scope is easy, except when it isn't. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 58–63, Valencia, Spain. Association for Computational Linguistics.

Federico Fancellu and Bonnie Webber. 2015. Translating negation: A manual error analysis. In *Proceedings of the Second Workshop on Extra-Propositional Aspects of Meaning in Computational Semantics (ExProM 2015)*, pages 2–11, Denver, Colorado. Association for Computational Linguistics.

Christiane Fellbaum. 1998. A semantic network of english verbs. *WordNet: An electronic lexical database*, 3:153–178.

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.

Matt Gardner, Jonathan Berant, Hannaneh Hajishirzi, Alon Talmor, and Sewon Min. 2019a. Question answering is a format; when is it useful? arXiv:1909.11291.

Matt Gardner, Jonathan Berant, Hannaneh Hajishirzi, Alon Talmor, and Sewon Min. 2019b. Question answering is a format; when is it useful? *CoRR*, abs/1909.11291.

Atticus Geiger, Kyle Richardson, and Christopher Potts. 2020. Neural natural language inference models partially embed theories of lexical entailment and negation. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 163–173, Online. Association for Computational Linguistics.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague. Association for Computational Linguistics.

Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. VQA-LOL: visual question answering under the lens of logic. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXI*, volume 12366 of *Lecture Notes in Computer Science*, pages 379–396. Springer.

Mareike Hartmann, Miryam de Lhoneux, Daniel Hershcovich, Yova Kementchedjhieva, Lukas Nielsen, Chen Qiu, and Anders Søgaard. 2021. A multilingual benchmark for probing negation-awareness with minimal pairs. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 244–257, Online. Association for Computational Linguistics.

Hangfeng He, Federico Fancellu, and Bonnie Webber. 2017. Neural networks for negation cue detection in Chinese. In *Proceedings of the Workshop Computational Semantics Beyond Events and Roles*, pages 59–63, Valencia, Spain. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. Deberta: decoding-enhanced bert with disentangled attention.

Laurence Horn. 1989. *A natural history of negation*. The University of Chicago Press.

Md Mosharaf Hossain, Antonios Anastasopoulos, Eduardo Blanco, and Alexis Palmer. 2020a. It's not a non-issue: Negation as a source of error in machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3869–3885, Online. Association for Computational Linguistics.

Md Mosharaf Hossain, Dhivya Chinnappa, and Eduardo Blanco. 2022. An analysis of negation in natural language understanding corpora. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 716–723, Dublin, Ireland. Association for Computational Linguistics.

Md Mosharaf Hossain, Kathleen Hamilton, Alexis Palmer, and Eduardo Blanco. 2020b. Predicting the focus of negation: Model and error analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8389–8401, Online. Association for Computational Linguistics.

Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020c. An analysis of natural language inference benchmarks through the lens of negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9106–9118, Online. Association for Computational Linguistics.

Arian Hosseini, Siva Reddy, Dzmitry Bahdanau, R Devon Hjelm, Alessandro Sordoni, and Aaron Courville. 2021. Understanding by understanding not: Modeling negation in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1301–1312, Online. Association for Computational Linguistics.

Liwei Jiang, Antoine Bosselut, Chandra Bhagavatula, and Yejin Choi. 2021. "I'm not mad": Commonsense implications of negation and contradiction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4380–4397, Online. Association for Computational Linguistics.

Salud María Jiménez-Zafra, Roser Morante, María Teresa Martín-Valdivia, and L. Alfonso Ureña-López. 2020. Corpora annotated with negation: An overview. *Computational Linguistics*, 46(1):1–52.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

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

Aditya Khandelwal and Suraj Sawant. 2020. NegBERT: A transfer learning approach for negation detection and scope resolution. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5739–5748, Marseille, France. European Language Resources Association.

Daniel Khashabi, Yeganeh Kordi, and Hannaneh Hajishirzi. 2022. Unifiedqa-v2: Stronger generalization via broader cross-format training. *CoRR*, abs/2202.12359.

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

Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 235–249, Minneapolis, Minnesota. Association for Computational Linguistics.

Natalia Konstantinova, Sheila C.M. de Sousa, Noa P. Cruz, Manuel J. Maña, Maite Taboada, and Ruslan Mitkov. 2012. A review corpus annotated for negation, speculation and their scope. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3190–3195, Istanbul, Turkey. European Language Resources Association (ELRA).

Venelin Kovatchev, M. Antonia Marti, Maria Salamo, and Javier Beltran. 2019. A qualitative evaluation framework for paraphrase identification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 568–577, Varna, Bulgaria. INCOMA Ltd.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical*

Methods in Natural Language Processing, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Hao Li and Wei Lu. 2018. Learning with structured representations for negation scope extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 533–539, Melbourne, Australia. Association for Computational Linguistics.

Shoushan Li and Chu-Ren Huang. 2009. Sentiment classification considering negation and contrast transition. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 1*, pages 307–316, Hong Kong. City University of Hong Kong.

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

Peter LoBue and Alexander Yates. 2011. Types of common-sense knowledge needed for recognizing textual entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 329–334, Portland, Oregon, USA. Association for Computational Linguistics.

Robert Logan IV, Ivana Balazevic, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel. 2022. Cutting down on prompts and parameters: Simple few-shot learning with language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2824–2835, Dublin, Ireland. Association for Computational Linguistics.

Ana Marasović, Iz Beltagy, Doug Downey, and Matthew Peters. 2022. Few-shot self-rationalization with natural language prompts. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 410–424, Seattle, United States. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Andrew Moore and Jeremy Barnes. 2021. Multi-task learning of negation and speculation for targeted sentiment classification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2838–2869, Online. Association for Computational Linguistics.

Roser Morante and Eduardo Blanco. 2012. *SEM 2012 shared task: Resolving the scope and focus of negation. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 265–274, Montréal, Canada. Association for Computational Linguistics.

Roser Morante and Walter Daelemans. 2012. ConanDoyle-neg: Annotation of negation cues and their scope in conan doyle stories. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1563–1568, Istanbul, Turkey. European Language Resources Association (ELRA).

Roser Morante, Sarah Schrauwen, and Walter Daelemans. 2011. Annotation of negation cues and their scope: Guidelines v1. *Computational linguistics and psycholinguistics technical report series, CTRS-003*, pages 1–42.

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.

Qiang Ning, Hao Wu, Pradeep Dasigi, Dheeru Dua, Matt Gardner, Robert L. Logan IV, Ana Marasović, and Zhen Nie. 2020. Easy, reproducible and quality-controlled data collection with CROWDAQ. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 127–134, Online. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 11054–11070.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

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

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

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.

Johan Reitan, Jørgen Faret, Björn Gambäck, and Lars Bungum. 2015. Negation scope detection for Twitter sentiment analysis. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 99–108, Lisboa, Portugal. 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.

Priyanka Sen and Amir Saffari. 2020. What do models learn from question answering datasets? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2429–2438, Online. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew K. Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, and et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *CoRR*, abs/2206.04615.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. Learning to summarize from human feedback. *CoRR*, abs/2009.01325.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *CoRR*, abs/2210.09261.

Thinh Hung Truong, Timothy Baldwin, Trevor Cohn, and Karin Verspoor. 2022. Improving negation detection with negation-focused pre-training. arXiv:2205.04012.

Chantal van Son, Emiel van Miltenburg, and Roser Morante. 2016. Building a dictionary of affixal negations. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM)*, pages 49–56, Osaka, Japan. The COLING 2016 Organizing Committee.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinform.*, 9(S-11).

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Hannaneh Hajishirzi, Noah A. Smith, and Daniel Khashabi. 2022. Benchmarking generalization via in-context instructions on 1, 600+ language tasks. *CoRR*, abs/2204.07705.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

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.

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

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. In *Proceedings of NeurIPS*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

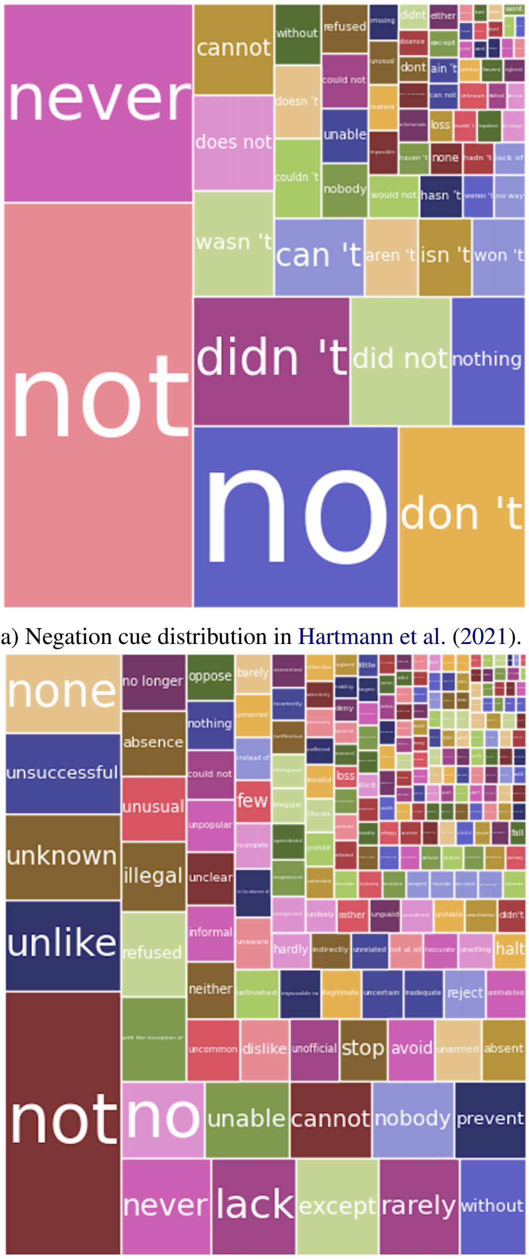
Xiaodan Zhu, Hongyu Guo, Saif Mohammad, and Svetlana Kiritchenko. 2014. An empirical study on the effect of negation words on sentiment. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 304–313, Baltimore, Maryland. Association for Computational Linguistics.

A Extended Comparison to Prior Negation Datasets

In this section, we complement the discussion in §7 on how CONDAQA differs from existing datasets focused on negation. A detailed comparison is given in Table 5.

Our goal with constructing CONDAQA is to contribute a high-quality and systematic evaluation that will facilitate future models that can adequately process negation. To this end, we aim to construct a benchmark where artifacts are carefully mitigated (Gardner et al., 2020), that is large enough to support robust evaluation, and that covers competencies any NLU system needs for adequate processing of negation. For example, the ability to recognize the implications of negated statements, distinguish them from their affirmative counterparts, and identify their scope. As such, main properties that CONDAQA has compared to prior datasets focused on negation are:

1. It is the first English reading-comprehension dataset that targets how models process negated statements in paragraphs (Gardner et al., 2019b).
2. It features three types of contrastive inputs to test a model's sensitivity to the presence of negation, its exact scope, and the way it is phrased. As such, it is the first contrastive dataset for studying negation.
3. It is substantially larger in size to facilitate robust evaluation.
4. It contains diverse forms of negation. Prior work constructing negation-based challenge sets for NLI models have largely constructed instances by using 'not' as the only negation cue (Hossain et al., 2020c; Naik et al., 2018). Hartmann et al. (2021) extend this and include 66 English negation cues in their NLI challenge set. Our dataset consists of over 200 negation cues. Figures 3a and 3b illustrate the distribution of negation cues in the dataset by Hartmann et al. (2021) and CONDAQA, respectively. CONDAQA is less skewed toward a few negation cues such as "not", "never", "no", etc.
5. All examples are manually constructed by well-trained crowdworkers rather than by using rules and templates.
6. It includes a rigorous validation procedure by several crowdworkers to mitigate examples being incorrect or ambiguous.



(a) Negation cue distribution in Hartmann et al. (2021).

(b) Negation cue distribution in CONDAQA.

Figure 3: Visualization of negation cues distributions.

B Analysis of CONDAQA

Commonsense Inferences We provide a categorization of the types of commonsense inferences required to answer CONDAQA questions. These categories are presented in Table 12.

Edit Strategies We provide a set of edit strategies that were employed by crowdworkers to make paraphrase and scope edits. These edits are given in Table 13.

Question/Passage Overlap An issue with some NLU datasets is that simple heuristics based on

Dataset	Task	Size	Contrastive Training Data	Passage /Premise /Prompt Length	Question /Hypothesis Length	# Negation Cues	Data Creation	Answer Exists
CONDAQA	RC	**14,182**	✓	**132.50**	**24.44**	**219**	Trained crowdworkers (a) paraphrase negation, (b) change negation scope, (c) remove the negation, (d) ask questions about implications of negation, (e) provide answers, (f) verify answers	✓
Hossain et al. (2020c)	NLI	1500 (MNLI) 1500 (SNLI) 1500 (RTE)	✗	16.71 (MNLI) 12.82 (SNLI) 23.73 (RTE)	11.27 (MNLI) 8.69 (SNLI) 11.04 (RTE)	1	Insert negation cue automatically	✓
Geiger et al. (2020)	NLI	2678	✗	9.27	9.27	1	Fill template automatically using Wordnet (Fellbaum, 1998)	✓
Hartmann et al. (2021)	NLI	1960	✗	19.35	9.95	66	Remove negation	✓
Ettinger (2020)	Cloze task	72 (NEG-136-SIMP) 64 (NEG-136-NAT)	✗	5.5 / 7.5	-	1	Psycholinguistic stimuli	✗

Table 5: Comparison between CONDAQA and prior datasets focusing on probing negation. We examine the English data in Hartmann et al. (2021), the MNLI/SNLI/RTE splits in Hossain et al. (2020c), NMoNLI (Geiger et al., 2020), as well as the NEG-136-SIMP and NEG-136-NAT datasets (Ettinger, 2020). CONDAQA is a reading comprehension dataset (RC), tasks in Hartmann et al. (2021) and Hossain et al. (2020c) are stress tests for existing general-purpose NLI datasets such as MNLI. NMoNLI is used both as a challenge (evaluation) set and to train models on a subset of the data. NEG-136-SIMP/NEG-136-NAT are datasets of cloze-style prompts. Passage/Premise/Prompt length and Question/Hypothesis length are described using the average number of words in the input. "Answer exists" describes whether a correct answer exists for the negated statement in the dataset, or if the evaluation relies on negated and affirmative statements requiring different predictions.

lexical overlap are sufficient to achieve high performance (Weissenborn et al., 2017; Naik et al., 2018). We measure the lexical overlap between CONDAQA questions and passages and find that is considerably lower than many prior QA datasets. Specifically, the average overlap between questions words and passage words is 0.52, which is lower compared to SQuAD 1.0 (Rajpurkar et al., 2016) (0.63), SQuAD 2.0 (Rajpurkar et al., 2018) (0.63), RACE (Lai et al., 2017) (0.67), and Quoref (Dasigi et al., 2019) (0.72).

Distribution of grammatical categories of negation cues We analyze the distribution over grammatical categories for single-word negation cues in CONDAQA. We use the NLTK library (Bird et al., 2009) to identify part-of-speech tags for these cues. These results are shown in Figure 4.

Model sensitivity to edits One potential issue with the dataset may be that models find it trivial to distinguish between edited passages and leverage this information to answer questions. To evaluate whether models can easily distinguish between the original passages and edited versions, we train BERT (Devlin et al., 2019) on the task of identifying whether a passage is sourced from Wikipedia or is an edited passage produced by a crowdworker. We expect it should be simple for these models to distinguish between the Wikipedia passages and

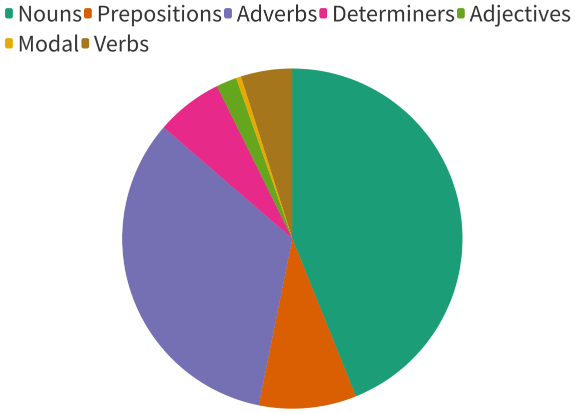


Figure 4: Distribution of grammatical categories of negation cues in CONDAQA.

the affirmative edits, as the model can simply rely on the presence or absence of a negation cue. We observe that as expected, models are somewhat able to distinguish the original Wikipedia passages from affirmative edits, but are largely unable to discriminate between the original passage and the paraphrase and scope edits (Table 6).

Naturalness of edits New edits made by crowdworkers may contain unnatural sentences or linguistic constructs. To quantify this and to exclude the possibility that model performance degrades only due to the unnaturalness of the edits, we compare the perplexity assigned by the OpenAI-GPT lan-

Model	All	Original-Pa.	Original-Sc.	Original-Aff.
Majority	25.65%	50%	51.86%	50.68%
BERT	34.4%	53.95%	55.27%	63.25%

Table 6: Performance of models trained to distinguish Wikipedia text from edits made by crowdworkers. We used Bert-base, averaged over three random seeds.

Split	Original	Paraphrase	Scope	Affirmative
Train	77.29	76.75	77.64	78.27
Dev	71.23	70.85	72.79	71.60
Test	74.38	74.88	75.63	76.05

Table 7: Average perplexities of original and (paraphrase, scope, affirmative) edited passages calculated with OpenAI-GPT (Radford et al., 2018).

guage model (Radford et al., 2018) to the edited passages and the original Wikipedia passages, finding that they are largely similar (Table 7).

Consistency Groups We provide data statistics on the instances that are used to compute consistency metrics on the dataset. There are 5,608 instances in the dataset that are included in consistency groups, and thus there are 1,402 "groups" to compute question-level consistency. and each edit-level consistency metric.

C Model Training Details

All models we evaluate on CondaQA are pretrained transformer-based language models. We test them in three training settings: (ii) finetuned on the entire training data (§C.1), (ii) finetuned on a few examples (few-shot; §C.2), and (iii) without training (zero-shot; §C.2).

C.1 Fully Finetuned

We train all fully-finetuned model with five seeds and report the average performance across them. For every seed, we evaluate the model with the best validation accuracy on the entire test set.

BERT (Devlin et al., 2019) BERT is pretrained with masked language modeling (MLM) and a next-sentence prediction objective. Since a majority of the questions have Yes/No/Don't know as the answer, we finetune BERT and other BERT-like models (see below) in a multi-class classification setting. We train all BERT-like models in this fashion. In our experiments, we BERT-Large. We train with a learning rate of 1e-5 for 10 epochs.

RoBERTa (Liu et al., 2019) RoBERTa is a more robustly pretrained version of BERT. In our experiments, we use RoBERTa-Large.

DeBERTa (He et al., 2021b,a) DeBERTa has a disentangled attention mechanism and it is pretrained with a version of MLM objective that uses the content and position of the context words. In our experiments, we use DeBERTa-v2-XLarge and DeBERTa-v3-Large.

UnifiedQA (Khashabi et al., 2020, 2022) UnifiedQA is built on top of the T5 architecture (Raffel et al., 2020) by further training it on 20 QA datasets. We use UnifiedQA-v2 and finetune it with a learning rate of 5e-5 for 5 epochs. In the fully-finetuned setting, we study Base, Large, and 3B versions of UnifiedQA-v2.

C.2 Few-shot and Zero-Shot

Unlike fully-finetuned models, we evaluate few- and zero-shot models on 5 train-test splits due to the cost of the OpenAI API. Evaluation on multiple disjoint splits of test data (that in union form the entire test set) with different choices of shots allows us to consider in our evaluation the sensitivity of few-shot learning to the choice of few examples. If the cost was not a concern, we would use five sets of few training examples and the *entire* test set.

GPT-3 (`davinci`; Brown et al., 2020) This is the original GPT-3 model trained using only the standard LM objective. Its maximum input sequence length is ∼2K tokens which allows to fit on average 8–9 CondaQA training examples. Thus, we use this number of shots for few-shot experiments. To benchmark GPT models, we use the OpenAI API (in October 2022). We show one prompt for few-shot GPT models in Fig. 5.

🪄InstructGPT (`text-davinci-002`; Ouyang et al., 2022) This GPT variant does not come with a corresponding paper and little is known about it. It has recently been confirmed that it is an Instruct model, but unlike the original InstructGPT$_{orig}$ (`text-davinci-001`; Ouyang et al., 2022) it is not derived from GPT-3 (`davinci`).[4] InstructGPT$_{orig}$ has been trained on the data that includes "prompts submitted to earlier versions of the InstructGPT models on the OpenAI API Playground". InstructGPT$_{orig}$ is finetuned with

[4] `https://twitter.com/janleike/status/1584681562318458880`

Sampling Strategy	Accuracy	Consistency	Paraphrase Consistency	Scope Consistency	Affirmative Consistency
1	**52.81**	**6.62**	**49.83**	**21.95**	**21.95**
2	51.42	5.57	44.95	21.25	26.48
3	50.31	4.88	40.07	18.12	25.78

Table 8: Few-shot results of GPT-3 (`davinci`) on one split of the test data (1/5 of the entire test set, ∼1440 examples) using different strategies for sampling few shots. See §C.2 for descriptions of the sampling strategies.

Max Seq Len	Accuracy	Consistency	Paraphrase Consistency	Scope Consistency	Affirmative Consistency
2045	60.88	20.30	63.92	36.40	33.98
4000	59.70	20.42	62.94	36.04	34.38

Table 9: "InstructGPT" (`text-davinci-002`) performance on one split of the test data (1/5 of the entire test set, ∼1440 examples) with more and less examples in the context. The average number of shots that fit in 2045 tokens (`davinci` max. input length) is 8–9, and 17-18 if the context is 4000 tokens (`text-davinci-002` max. input length).

reinforcement learning from human feedback (Stiennon et al., 2020). `text-davinci-002` has two times longer maximum input sequence length than `davinci` suggesting that the overall model size is notably larger too. This also means we can fit more examples in the context, but we do not find that to improve `text-davinci-002`'s performance; see Table 9. It has been reported on social media that `text-davinci-002` has notably stronger performance than `text-davinci-001`, but where do these improvements come from is publicly unknown.[5]

Chain-of-Thoughts (CoT) prompting (Wei et al., 2022) This type of prompting makes the model explain its prediction before providing it. When it was introduced, CoT prompting demonstrated benefits for math and commonsense reasoning. Since then, Suzgun et al. (2022) report that CoT prompting gives substantial improvements for a hard subset of the BIG-Bench tasks (Srivastava et al., 2022).[6] This makes it a promising prompt for our proposed task of reasoning about implications of negation. The suggested way to conduct CoT prompting (and how we use it in this paper) is as follows:

- **Input:** `{task_description}` `{task_examples}` `{test_instance}` *Answer: Let's think step by step.*
- **Output:** `{explanation}` *So the answer is* `{answer}`

One of the authors wrote explanations for all shots in each split (45 explanations in total) in few hours. In Figure 6, we show an example of a CoT prompt we use for "InstructGPT" (`text-davinci-002`).

FLAN-T5 (Chung et al., 2022) FLAN-T5 is a T5 variant that is further trained with instruction finetuning that includes CoT prompting, on over 1.8K tasks. We prompt FLAN-T5 in the zero-shot setting by constructing each test instance as follows:

- **Input:** *Passage:* `{passage}`\n*Question:* `{question}`\n*Give the rationale before answering.*
- **Output:** `{explanation}` *So the answer is* `{answer}`.

This output form is the most common, but the model sometimes generates "(final) answer is", "(final) answer:", etc., instead of "So the answer is".

UnifiedQA-v2 (Khashabi et al., 2022) We also evaluate UnifiedQA-v2 in a few- and zero-shot settings in addition to fully training it. We construct instances following how they are constructed for training UnifiedQA-v2:

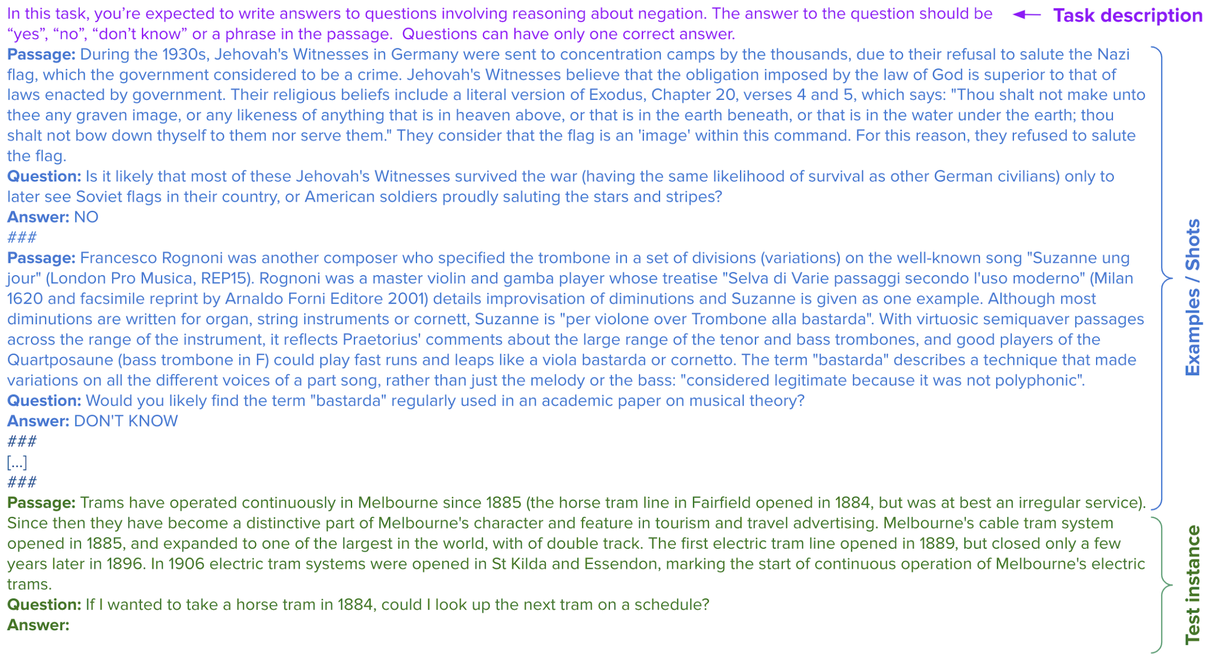
- **Input:** `{passage}\n{question}`
- **Output:** `{answer}`

We normalize and lowercase passages, questions, and answers. We manually choose hyperparameters following Bragg et al. (2021) and keep them fixed.

Which few examples to select? CONDAQA's unique structure raises the question of which 8–9

[5] `https://twitter.com/ben_bogin/status/1532022804886978568`

[6] Another work shows limitations of prompting with explanations (Ye and Durrett, 2022).



In this task, you're expected to write answers to questions involving reasoning about negation. The answer to the question should be "yes", "no", "don't know" or a phrase in the passage. Questions can have only one correct answer.

Passage: During the 1930s, Jehovah's Witnesses in Germany were sent to concentration camps by the thousands, due to their refusal to salute the Nazi flag, which the government considered to be a crime. Jehovah's Witnesses believe that the obligation imposed by the law of God is superior to that of laws enacted by government. Their religious beliefs include a literal version of Exodus, Chapter 20, verses 4 and 5, which says: "Thou shalt not make unto thee any graven image, or any likeness of anything that is in heaven above, or that is in the earth beneath, or that is in the water under the earth; thou shalt not bow down thyself to them nor serve them." They consider that the flag is an 'image' within this command. For this reason, they refused to salute the flag.
Question: Is it likely that most of these Jehovah's Witnesses survived the war (having the same likelihood of survival as other German civilians) only to later see Soviet flags in their country, or American soldiers proudly saluting the stars and stripes?
Answer: NO
###
Passage: Francesco Rognoni was another composer who specified the trombone in a set of divisions (variations) on the well-known song "Suzanne ung jour" (London Pro Musica, REP15). Rognoni was a master violin and gamba player whose treatise "Selva di Varie passaggi secondo l'uso moderno" (Milan 1620 and facsimile reprint by Arnaldo Forni Editore 2001) details improvisation of diminutions and Suzanne is given as one example. Although most diminutions are written for organ, string instruments or cornett, Suzanne is "per violone over Trombone alla bastarda". With virtuosic semiquaver passages across the range of the instrument, it reflects Praetorius' comments about the large range of the tenor and bass trombones, and good players of the Quartposaune (bass trombone in F) could play fast runs and leaps like a viola bastarda or cornetto. The term "bastarda" describes a technique that made variations on all the different voices of a part song, rather than just the melody or the bass: "considered legitimate because it was not polyphonic".
Question: Would you likely find the term "bastarda" regularly used in an academic paper on musical theory?
Answer: DON'T KNOW
###
[...]
###
Passage: Trams have operated continuously in Melbourne since 1885 (the horse tram line in Fairfield opened in 1884, but was at best an irregular service). Since then they have become a distinctive part of Melbourne's character and feature in tourism and travel advertising. Melbourne's cable tram system opened in 1885, and expanded to one of the largest in the world, with of double track. The first electric tram line opened in 1889, but closed only a few years later in 1896. In 1906 electric tram systems were opened in St Kilda and Essendon, marking the start of continuous operation of Melbourne's electric trams.
Question: If I wanted to take a horse tram in 1884, could I look up the next tram on a schedule?
Answer:

Figure 5: A prompt used to get generations from GPT-3 (`davinci`) and "InstructGPT" (`text-davinci-002`). We designed the task description following Wang et al. (2022). The zero-shot prompt is the same except that there are no examples.

examples to use for few-shot learning:

1. Randomly selected,
2. Random without affirmative paragraphs to include more paragraphs with negation cues,
3. Two groups of two questions and corresponding 4 paragraphs (original and three edited),
4. Three groups of two questions and corresponding 3 paragraphs (original, scope- and paraphrase-edited; no affirmative).

We hypothesize that the last two options could be beneficial for consistency of few-shot models. We prompt `davinci` with 1st and 3rd options, and depending which is better we evaluate 2nd or 4th (i.e., the better option without affirmative paragraphs). Contrary to our expectations, we find that the 1st option works better than 3rd, as well as better than the 2nd option; see Table 8. Therefore, for each training split, we sample 9 paragraph-question pairs randomly (sometimes only 8 fit in the context) and use these samples for all few-shot experiments.

D Model analysis

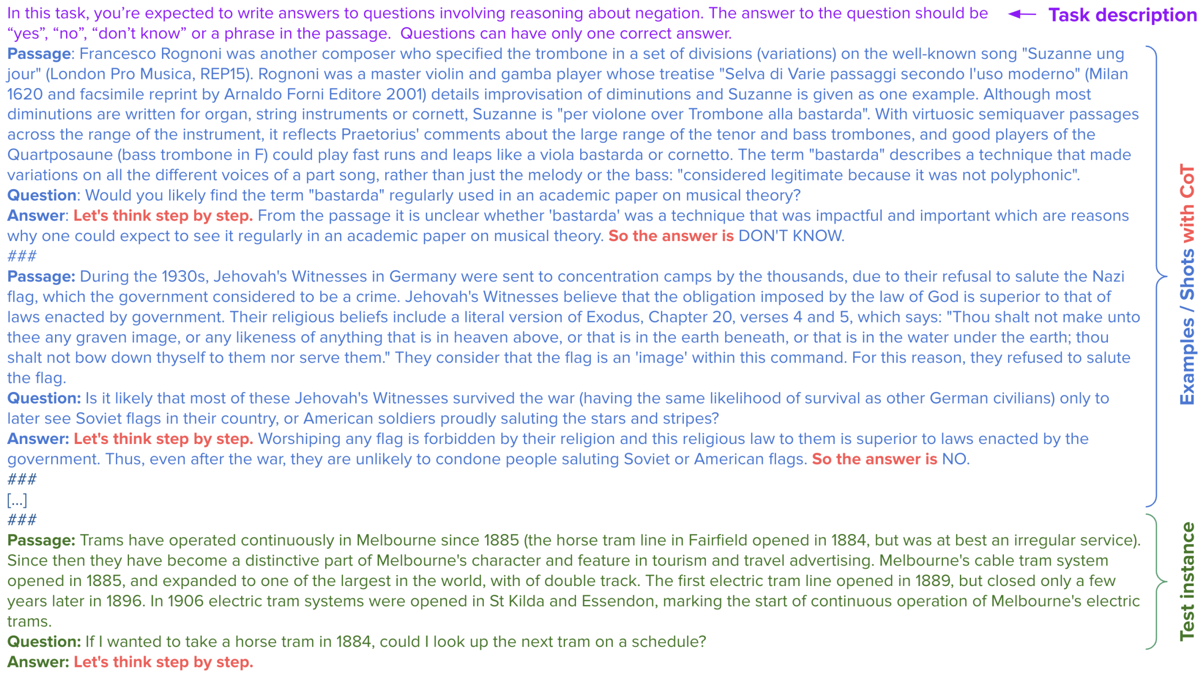
Model performance stratified by passage type In Table 10, we report the accuracy of model predictions corresponding to the type of passage: i.e whether the question was asked on the original Wikipedia passage, its paraphrase edit, its scope edit or the affirmative edit. When we compare those results with those in Table 4, we observe that UnifiedQA-v2 shows largely similar QA performance in terms of accuracy on these different passage types, despite having very different consistency scores with the original passage. In contrast, GPT-3 and 🪄Instruct-GPT in the few-shot setting perform better on the original Wikipedia passages and their paraphrased versions than on the scope and affirmative edits, possibly suggesting that these models work best on passages that are available online.

Model	Original	Paraphrase	Scope	Affirmative
UnifiedQA-V2-3B	75.53	74.23	69.42	71.43
UnifiedQA-V2-Large	68.35	68.13	63.25	67.22
GPT-3	57.67	59.79	51.32	43.91
🪄InstructGPT	67.99	70.63	53.37	51.84

Table 10: Accuracy of UnifiedQA-V2, GPT-3, and 🪄Instruct-GPT stratified by the type of passage.

Model performance by question length In Figure 7, we show model performance stratified by question length. We observe that longer questions are more difficult for UnifiedQA-v2-Large but UnifiedQA-v2-3B appears to exhibit similar QA performance on some of these long questions.

Model performance by answer type In Figure 8, we show results of model performance stratified



Task description
Examples / Shots with CoT
Test instance

In this task, you're expected to write answers to questions involving reasoning about negation. The answer to the question should be "yes", "no", "don't know" or a phrase in the passage. Questions can have only one correct answer.

Passage: Francesco Rognoni was another composer who specified the trombone in a set of divisions (variations) on the well-known song "Suzanne ung jour" (London Pro Musica, REP15). Rognoni was a master violin and gamba player whose treatise "Selva di Varie passaggi secondo l'uso moderno" (Milan 1620 and facsimile reprint by Arnaldo Forni Editore 2001) details improvisation of diminutions and Suzanne is given as one example. Although most diminutions are written for organ, string instruments or cornett, Suzanne is "per violone over Trombone alla bastarda". With virtuosic semiquaver passages across the range of the instrument, it reflects Praetorius' comments about the large range of the tenor and bass trombones, and good players of the Quartposaune (bass trombone in F) could play fast runs and leaps like a viola bastarda or cornetto. The term "bastarda" describes a technique that made variations on all the different voices of a part song, rather than just the melody or the bass: "considered legitimate because it was not polyphonic".
Question: Would you likely find the term "bastarda" regularly used in an academic paper on musical theory?
Answer: **Let's think step by step.** From the passage it is unclear whether 'bastarda' was a technique that was impactful and important which are reasons why one could expect to see it regularly in an academic paper on musical theory. **So the answer is** DON'T KNOW.
###
Passage: During the 1930s, Jehovah's Witnesses in Germany were sent to concentration camps by the thousands, due to their refusal to salute the Nazi flag, which the government considered to be a crime. Jehovah's Witnesses believe that the obligation imposed by the law of God is superior to that of laws enacted by government. Their religious beliefs include a literal version of Exodus, Chapter 20, verses 4 and 5, which says: "Thou shalt not make unto thee any graven image, or any likeness of anything that is in heaven above, or that is in the earth beneath, or that is in the water under the earth; thou shalt not bow down thyself to them nor serve them." They consider that the flag is an 'image' within this command. For this reason, they refused to salute the flag.
Question: Is it likely that most of these Jehovah's Witnesses survived the war (having the same likelihood of survival as other German civilians) only to later see Soviet flags in their country, or American soldiers proudly saluting the stars and stripes?
Answer: **Let's think step by step.** Worshiping any flag is forbidden by their religion and this religious law to them is superior to laws enacted by the government. Thus, even after the war, they are unlikely to condone people saluting Soviet or American flags. **So the answer is** NO.
###
[...]
###
Passage: Trams have operated continuously in Melbourne since 1885 (the horse tram line in Fairfield opened in 1884, but was at best an irregular service). Since then they have become a distinctive part of Melbourne's character and feature in tourism and travel advertising. Melbourne's cable tram system opened in 1885, and expanded to one of the largest in the world, with of double track. The first electric tram line opened in 1889, but closed only a few years later in 1896. In 1906 electric tram systems were opened in St Kilda and Essendon, marking the start of continuous operation of Melbourne's electric trams.
Question: If I wanted to take a horse tram in 1884, could I look up the next tram on a schedule?
Answer: **Let's think step by step.**

Figure 6: A *chain-of-thought* prompt (includes "Let's think step by step. `{explanation}`. So the answer is") used to get generations from "InstructGPT" (`text-davinci-002`). We designed the task description following Wang et al. (2022).

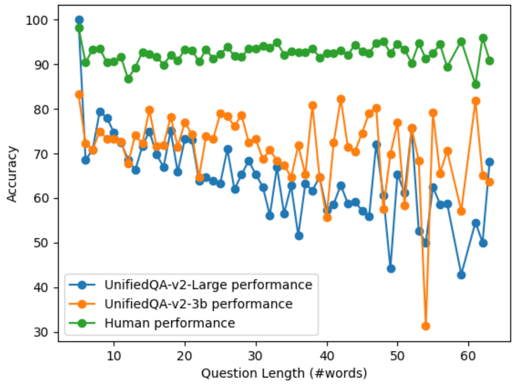


Figure 7: UnifiedQA-v2-Large and UnifiedQA-v2-3b performance stratified by the length of the question.

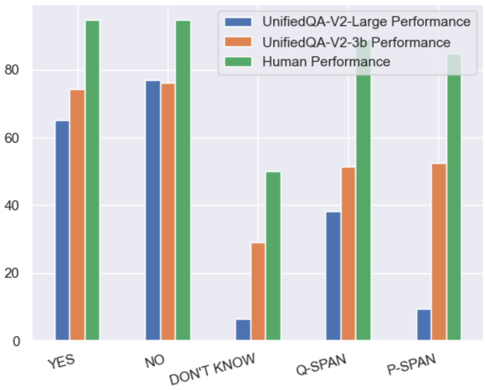


Figure 8: Model accuracy for UnifiedQA-v2-Large and UnifiedQA-v2-3b based on answer type.

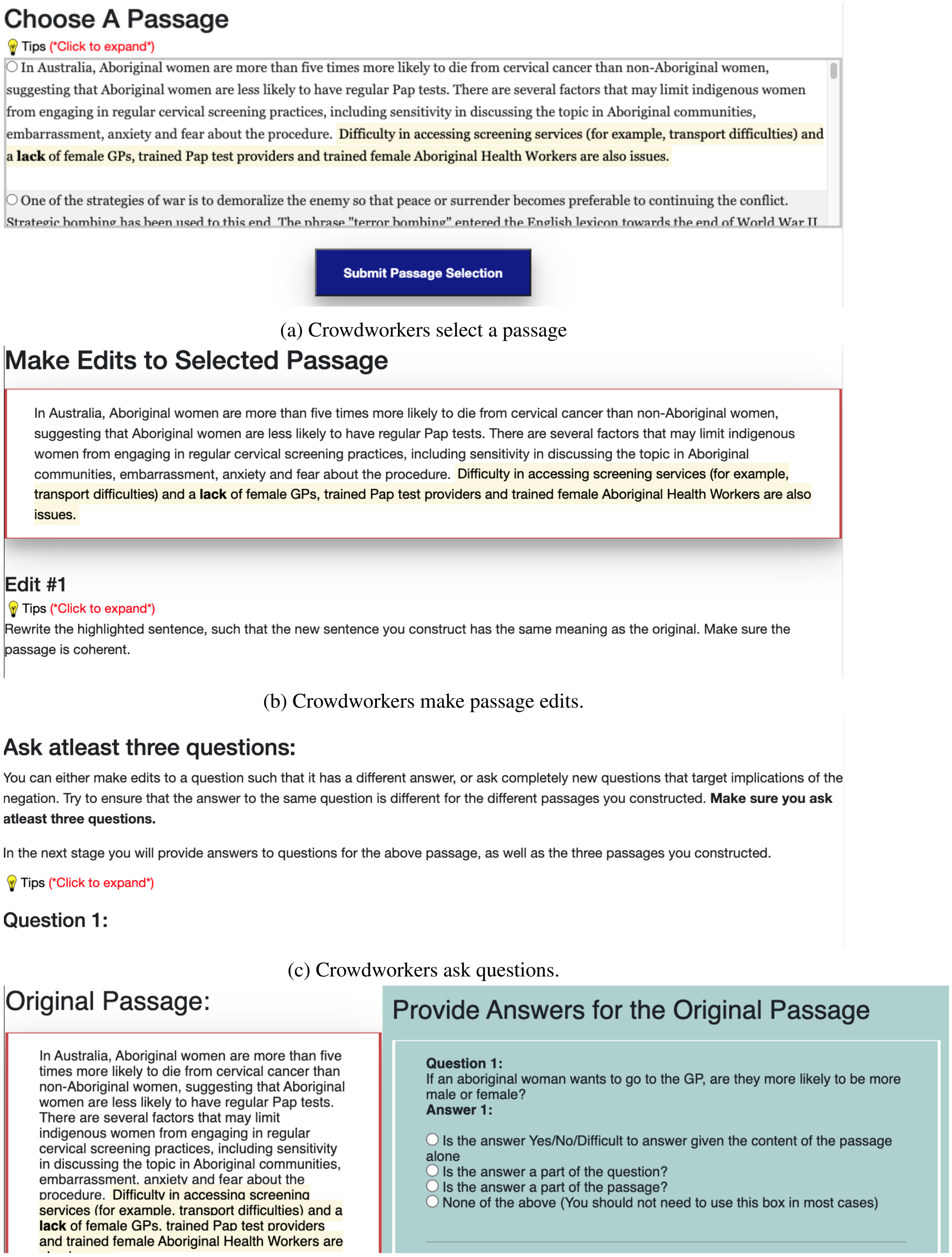
by answer type (Figure 8).

Variance in model performance We report the standard deviation of UnifiedQA-V2 models computed over the results from five *seeds*, as well as the standard deviation of GPT-3 and 🪄Instruct-GPT in few-shot and zero-shot settings computed over five *splits*. These are shown in Table 11.

Novelty of negation cues We compare the performance of fully-finetuned UnifiedQA-v2 Large/3B on Wikipedia passages where the negation cue has occurred in the training data, with the performance for novel negation cues. We find that model accuracy for UnifiedQA-V2-Large is 68.03 when the negation cue is unseen (has not been the cue in the negated statement that crowdworkers construct questions around in the training data), and 70.45 when it has appeared before in the training data. Similarly, UnifiedQA-V2-3B's accuracy is 74.38 and 73.73 for unseen and seen cues respectively. This suggests that the novelty of the negation cue is not a major factor of difficulty for UnifiedQA-v2 once it has been finetuned on the entire training data.

E Crowdsourcing Interface Templates

We include an example of the annotation interface we showed to crowdworkers. Figure 9 shows a sample of each stage of our task.



Choose A Passage

💡 Tips (*Click to expand*)

○ In Australia, Aboriginal women are more than five times more likely to die from cervical cancer than non-Aboriginal women, suggesting that Aboriginal women are less likely to have regular Pap tests. There are several factors that may limit indigenous women from engaging in regular cervical screening practices, including sensitivity in discussing the topic in Aboriginal communities, embarrassment, anxiety and fear about the procedure. Difficulty in accessing screening services (for example, transport difficulties) and a **lack** of female GPs, trained Pap test providers and trained female Aboriginal Health Workers are also issues.

○ One of the strategies of war is to demoralize the enemy so that peace or surrender becomes preferable to continuing the conflict. Strategic bombing has been used to this end. The phrase "terror bombing" entered the English lexicon towards the end of World War II

(a) Crowdworkers select a passage

Make Edits to Selected Passage

In Australia, Aboriginal women are more than five times more likely to die from cervical cancer than non-Aboriginal women, suggesting that Aboriginal women are less likely to have regular Pap tests. There are several factors that may limit indigenous women from engaging in regular cervical screening practices, including sensitivity in discussing the topic in Aboriginal communities, embarrassment, anxiety and fear about the procedure. Difficulty in accessing screening services (for example, transport difficulties) and a **lack** of female GPs, trained Pap test providers and trained female Aboriginal Health Workers are also issues.

Edit #1

💡 Tips (*Click to expand*)

Rewrite the highlighted sentence, such that the new sentence you construct has the same meaning as the original. Make sure the passage is coherent.

(b) Crowdworkers make passage edits.

Ask atleast three questions:

You can either make edits to a question such that it has a different answer, or ask completely new questions that target implications of the negation. Try to ensure that the answer to the same question is different for the different passages you constructed. **Make sure you ask atleast three questions.**

In the next stage you will provide answers to questions for the above passage, as well as the three passages you constructed.

💡 Tips (*Click to expand*)

Question 1:

(c) Crowdworkers ask questions.

Original Passage:

In Australia, Aboriginal women are more than five times more likely to die from cervical cancer than non-Aboriginal women, suggesting that Aboriginal women are less likely to have regular Pap tests. There are several factors that may limit indigenous women from engaging in regular cervical screening practices, including sensitivity in discussing the topic in Aboriginal communities, embarrassment, anxiety and fear about the procedure. Difficulty in accessing screening services (for example, transport difficulties) and a **lack** of female GPs, trained Pap test providers and trained female Aboriginal Health Workers are

Provide Answers for the Original Passage

Question 1:
If an aboriginal woman wants to go to the GP, are they more likely to be more male or female?
Answer 1:

○ Is the answer Yes/No/Difficult to answer given the content of the passage alone
○ Is the answer a part of the question?
○ Is the answer a part of the passage?
○ None of the above (You should not need to use this box in most cases)

(d) Crowdworkers answer questions.

Figure 9: Sample of our Question-Answering HIT, where crowdworkers can choose a passage, make edits to that passage, ask questions about that passage and then answer those questions.

Model	# Param	Accuracy	Consistency	Paraphrase Consistency	Scope Consistency	Affirmative Consistency
Fully Finetuned						
UnifiedQA-v2-Base	220M	57.94$_{\pm 0.25}$	17.49$_{\pm 0.47}$	54.62$_{\pm 0.47}$	30.39$_{\pm 0.49}$	32.98$_{\pm 0.48}$
UnifiedQA-v2-large	770M	66.72$_{\pm 0.13}$	30.20$_{\pm 0.10}$	63.98$_{\pm 0.31}$	43.68$_{\pm 0.25}$	46.45$_{\pm 0.38}$
UnifiedQA-v2-3B	3B	**73.26**$_{\pm 0.46}$	**42.18**$_{\pm 0.72}$	**72.80**$_{\pm 0.68}$	**55.68**$_{\pm 0.58}$	**57.22**$_{\pm 0.77}$
Few-Shot						
UnifiedQA-v2-Base	220M	52.58$_{\pm 1.57}$	11.97$_{\pm 1.57}$	50.11$_{\pm 3.32}$	24.19$_{\pm 2.83}$	25.03$_{\pm 3.81}$
UnifiedQA-v2-Large	770M	55.84$_{\pm 2.04}$	16.80$_{\pm 2.28}$	56.05$_{\pm 2.96}$	30.25$_{\pm 2.01}$	29.93$_{\pm 3.14}$
UnifiedQA-v2-3B	3B	61.14$_{\pm 3.45}$	22.52$_{\pm 5.2}$	62.05$_{\pm 2.82}$	35.71$_{\pm 3.66}$	35.41$_{\pm 5.46}$
GPT-3*	175B	52.42$_{\pm 2.04}$	5.22$_{\pm 2.48}$	48.94$_{\pm 1.11}$	23.31$_{\pm 3.24}$	20.31$_{\pm 5.35}$
🪄InstructGPT**	N/A	60.88$_{\pm 1.44}$	20.30$_{\pm 1.38}$	63.92$_{\pm 1.48}$	36.40$_{\pm 3.10}$	33.98$_{\pm 1.53}$
🪄InstructGPT**+ CoT	N/A	**66.28**$_{\pm 2.49}$	**27.28**$_{\pm 3.85}$	**64.27**$_{\pm 3.36}$	**45.08**$_{\pm 2.82}$	**44.91**$_{\pm 3.05}$
Zero-Shot						
UnifiedQA-v2-Base	220M	55.65$_{\pm 1.44}$	16.20$_{\pm 1.74}$	52.47$_{\pm 2.366}$	29.23$_{\pm 1.27}$	30.83$_{\pm 1.95}$
UnifiedQA-v2-Large	770M	61.74$_{\pm 0.8}$	23.07$_{\pm 2.39}$	61.16$_{\pm 2.58}$	37.14$_{\pm 1.3}$	37.14$_{\pm 2.93}$
UnifiedQA-v2-3B	3B	69.41$_{\pm 0.99}$	34.91$_{\pm 1.81}$	70.71$_{\pm 1.87}$	47.94$_{\pm 2.39}$	49.74$_{\pm 2.22}$
UnifiedQA-v2-11B	11B	**73.11**$_{\pm 1.74}$	**40.02**$_{\pm 2.84}$	**75.48**$_{\pm 2.98}$	**53.72**$_{\pm 2.32}$	**54.12**$_{\pm 3.84}$
FLAN-T5-XXL	11B	67.53$_{\pm 1.25}$	31.61$_{\pm 3.37}$	67.43$_{\pm 2.54}$	45.45$_{\pm 3.27}$	47.86$_{\pm 2.58}$
GPT-3*	175B	43.72$_{\pm 0.86}$	1.28$_{\pm 0.35}$	41.33$_{\pm 2.60}$	10.67$_{\pm 1.90}$	10.89$_{\pm 1.282}$
🪄InstructGPT**	N/A	54.00$_{\pm 2.05}$	16.32$_{\pm 2.95}$	55.54$_{\pm 2.56}$	29.87$_{\pm 3.54}$	27.81 $_{\pm 2.44}$

Table 11: Model performance on CondaQA with standard deviation. **Boldface** indicates the best model on each metric for every training setup (*Supervised*, *Few-Shot*, *Zero-Shot*). Supervised models are trained and evaluated across five random seeds using the full train and test sets. Due to the cost of OpenAI API, for few- and zero-shot models we report the average performance across five train-test splits. For more details and description of metrics see §4. GPT-3 version: `davinci`; 🪄InstructGPT version: `text-davinci-002`.

Reasoning Type	**Passage Snippet**	**Question**	**Answer**	**Explanation**
Precondition (12%)	At first reluctantly but then with increasing vigour, Galen promoted Hippocratic teaching, including venesection and bloodletting, then **unknown** in Rome [...]	Would doctors in Rome regularly have performed venesection?	NO	People can't do a complicated procedure that they don't know.
Social Norms (10%)	On October 8, 1883, the US patent office ruled that Edison's patent was based on the work of William E. Sawyer and was, therefore, **invalid** . Litigation continued for nearly six years. In 1885, Latimer switched camps and started working with Edison.	From the information given in the passage, would you say that coincidence is the most charitable explanation for what was essentially the same innovation, in much the same way that Newton and Leibniz seemingly discovered calculus independently, without knowing of the other's progress?	YES	Plagarism is frowned upon in society, more so than accidentally reaching the same conclusions as someone else.
Psychology (9%)	[...] Disraeli later romanticised his origins, claiming his father's family was of grand Iberian and Venetian descent; in fact Isaac's family was of no great distinction [...] Historians differ on Disraeli's motives for rewriting his family history: [...] Sarah Bradford believes "his **dislike** of the commonplace would not allow him to accept the facts of his birth as being as middle-class and undramatic as they really were".	Would Disraeli have been flattered by a biography that explored his middle class upbringing, according to Bradford?	NO	A person such as Disraeli who wants to project a grandiose image of themselves is likely to be unhappy when people discuss mundane aspects about his upbringing.
Cause and Effect (7%)	Oil produced from palm fruit is called 'red palm oil' or just 'palm oil'... In its **unprocessed** state, red palm oil has an intense deep red color because of its abundant carotene content. [...]	Would a consumer who was primarily interested in the eye-health benefits of carotenes and lycopene want to shop for palm oils by their color, rather than listening to marketing slogans such as "extra virgin" or "minimally processed"?	YES	A high carotene content causes a deep red color, so a person searching for things with high carotene content can look at their color.
Mutual Exclusivity (5%)	[...] The waterway system covered much of the country, and by the 1980s Finland had extended roadways and railroads to areas **not** served by waterways, effectively opening up all of the country's forest reserves to commercial use.	Would a person in 1990 taking a nap near a river in Finland be likely to be woken up by a train horn?	NO	It is likely that the government prioritized building roads and railways in places not near waterways
Synecdoche (2%)	Al-Libi told the interrogators details about Richard Reid, a British citizen who had joined al-Qaeda and trained to carry out a suicide bombing of an airliner, which he **unsuccessfully** attempted on December 22, 2001. [...]	Would al-Qaeda take responsibility for Richard Reid's suicide bombing attempt?	YES	Richard Reid was a member of the Al Qaeda.

Table 12: Examples of types of questions that target the implications of negated statements in CONDAQA, and reasoning steps to correctly answer the questions. Negated statements are in blue. Relevant categories derived from LoBue and Yates (2011) when appropriate.

Revision Strategy	**Edited Passage**
	PARAPHRASE EDIT
Complement substitution	Though Philby claimed publicly in January 1988 that he did not regret his decisions and that ~~he missed nothing about England except~~ the only things he missed about England were some friends, Colman's mustard, and Lea & Perrins Worcestershire sauce, his wife Rufina Ivanovna Pukhova later described Philby as "disappointed in many ways" by what he found in Moscow.
Synonym substitution	Local tetanus is ~~an uncommon~~a rare form of the disease and it causes persistent contractions of muscles in the same area of the sufferer's body as where the original injury was made.
Antonym substitution	The population of the Thirteen States was ~~not homogeneous~~ heterogeneous in political views and attitudes.
Numerical equivalence	The period before 1920 is known as the dead-ball era, during which players would ~~rarely~~ hit home runs at a low frequency.
Ellipsis	~~Sunni scholars put trust in narrators such as Aisha, whom Shia reject~~While the Shia tend to reject narrators such as Aisha, Sunni scholars tend to trust them.
Noun-adjective conversion	~~While~~ Longships were used by the Norse in ~~warfare~~a military capacity, ~~they were mostly used as~~ but mostly for ~~troop transports~~transporting troops, ~~not~~ rather than as true warships.
	SCOPE EDIT
Complement inversion	~~Sunni~~Shia scholars put trust in narrators such as Aisha, whom ~~Shia~~Sunni reject.
Superset-subset replacement	During the coronavirus outbreak of 2020, alcohol sales~~, and even the~~ were made illegal, but the transportation of alcohol outside of one's home~~, was made illegal~~ remained legal.
Attribute change	Moocher's look is very ~~similar to~~unlike Scrooge's, except for the fact that ~~he wears~~they both wear tattered clothes, ~~but unlike~~and just like his very rich cousin, Moocher is also a sweetheart.
Temporal shift	As the new Emperor could not exert his constitutional powers ~~until~~once he came of age, a regency was set up by the National Assembly.
Veridicality	Contrary to assumptions that he was illiterate, on arrival he was given aptitude tests which determined that ~~he was illiterate~~not only could he read the questions and respond in writing, but he also had an above-average IQ of 109.

Table 13: Examples of revision strategies employed by crowdworkers for paraphrase and scope edits. Categories for paraphrases are inspired by Bhagat and Hovy (2013). The negation cue is in blue and newly-inserted text is in red.

Paragraph #1: Scorsese was initially reluctant to develop the project, though he eventually came to relate to LaMotta's story. Schrader re-wrote Martin's first screenplay, and Scorsese and De Niro together made uncredited contributions thereafter. *Pesci was a famous actor prior to appearing in this role, but Moriarty was unknown to the producers before he suggested her for her role.* During principal photography, each of the boxing scenes was choreographed for a specific visual style and De Niro gained approximately to portray LaMotta in his later post-boxing years. Scorsese was exacting in the process of editing and mixing the film, expecting it to be his last major feature.
Question: Is it possible that the writers of this movie had specifically tailored the character to Joe Pesci's unique on-screen charisma, with the hopes that he would accept the role?
Answer: Yes
Paragraph #2: Scorsese was initially reluctant to develop the project, though he eventually came to relate to LaMotta's story. Schrader re-wrote Martin's first screenplay, and Scorsese and De Niro together made uncredited contributions thereafter. *Before appearing in this movie, Pesci had not achieved fame as an actor, and neither had Moriarty, who he suggested for her role.* During principal photography, each of the boxing scenes was choreographed for a specific visual style and De Niro gained approximately to portray LaMotta in his later post-boxing years. Scorsese was exacting in the process of editing and mixing the film, expecting it to be his last major feature.
Question: Is it possible that the writers of this movie had specifically tailored the character to Joe Pesci's unique on-screen charisma, with the hopes that he would accept the role?
Answer: No

Table 14: Presumably, answering this question in the context of the second paragraph requires reasoning about negation, while if the question is answered in the context of the first paragraph it does not. However, if the model is only ever presented instances like the second paragraph, it is possible that there would be subtle artifacts that lead to a model's good performance without ever needing to fully process the negation. By making minimal changes to the paragraph that intervene on the negation, we can increase our confidence that the model is able to correctly process the negation in the second paragraph. The question-paragraph pairs must be considered jointly to accurately characterize a model's ability handle negation, hence our focus on group consistency as our preferred performance metric.