# Paraphrasing in Affirmative Terms Improves Negation Understanding

**MohammadHossein Rezaei** and **Eduardo Blanco**
Department of Computer Science, University of Arizona
{mhrezaei,eduardoblanco}@arizona.edu

## Abstract

Negation is a common linguistic phenomenon. Yet language models face challenges with negation in many natural language understanding tasks such as question answering and natural language inference. In this paper, we experiment with seamless strategies that incorporate affirmative interpretations (i.e., paraphrases without negation) to make models more robust against negation. Crucially, our affirmative interpretations are obtained automatically. We show improvements with CondaQA, a large corpus requiring reasoning with negation, and five natural language understanding tasks.

## 1 Introduction

Negation is a fundamental linguistic phenomenon present in all human languages (Horn, 1989). Language models underperform in various natural language understanding (NLU) tasks when the input includes negation. For example, Ettinger (2020) and Kassner and Schütze (2020) show that BERT (Devlin et al., 2019) fails to distinguish between negated and non-negated cloze questions. Researchers have also shown that large language models such as GPT-3 (Brown et al., 2020) and InstructGPT (Ouyang et al., 2022) are insensitive to negation and fail to reason under negation (Truong et al., 2023). Jang et al. (2022) point out that language models violate the logical negation property ($p$ is true iff $\neg p$ is false). Hossain et al. (2022a) analyze negation in eight popular corpora for six NLU tasks. They conclude that (a) NLU corpora have few negations compared to general-purpose texts and (b) the few negations in them are often unimportant. To our knowledge, CondaQA (Ravichander et al., 2022) is the largest benchmark (14,182 question-answer pairs from Wikipedia) requiring reasoning over the implications of negations.

In this paper, we paraphrase sentences with negation *without using negation* to make models for natural language understanding more robust when negation is present in the input. We will use the term *affirmative interpretation* to refer to paraphrases without negation (e.g., *I am not sad*: *I am just ok*, *I am happy*, etc.). Appendix A provides examples of how affirmative interpretations differ from simple paraphrases.

The main contributions of this paper are (a) strategies to generate and incorporate affirmative interpretations and (b) experimental results demonstrating that doing so yields better results.[1] In addition to CondaQA, we experiment with five of the eight corpora analyzed by Hossain et al. (2022a): CommonsenseQA (Talmor et al., 2019), STS-B (Cer et al., 2017), QNLI (Rajpurkar et al., 2016), WiC (Pilehvar and Camacho-Collados, 2019), and WSC (Levesque et al., 2012).[2] We do not experiment with the other three corpora because they do not contain any negation (Roemmele et al., 2011, COPA), there is no difference in results when negation is present (Cer et al., 2017, QQP; 0.01 in macro F1), or has already been shown (Hossain and Blanco, 2022) to benefit from affirmative interpretations (Socher et al., 2013, SST-2). The corpora we experiment with are in English.

**Related Work** Early research on negation targeted detecting negating cues and generating semantic representations, usually by identifying the scope and focus (Morante et al., 2011; Morante and Daelemans, 2012; van Son et al., 2016; Khandelwal and Sawant, 2020; Truong et al., 2022).

More recent works bypass formal representations. Instead, they make neural models robust when the input contains negation. Hosseini et al. (2021) combine unlikelihood training and syntactic data augmentation to enhance the ability of BERT to understand negation with negated LAMA (Kass-

[1] Code available at https://github.com/mhrezaei1/paraphrase-affirmative under Apache 2.0 license.
[2] See examples from these corpora in Appendix B.



*Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 602–615
August 11-16, 2024 ©2024 Association for Computational Linguistics

ner and Schütze, 2020). Singh et al. (2023) present a pretraining strategy designed for negation. Unlike these works, we couple original inputs containing negation with affirmative interpretations.

The first work on affirmative interpretations was by Sarabi et al. (2019). Hossain et al. (2022b) present AFIN, a corpus of ≈3,000 sentences with negations and their affirmative interpretations. These two previous works are limited to generating affirmative interpretations from negations; they do not provide extrinsic evaluations. More recently, Hossain and Blanco (2022) present Large-AFIN, over 153,000 pairs of sentences with negation and their affirmative interpretations obtained from parallel corpora via backtranslation. In this paper, we present strategies to generate affirmative interpretations that do not require parallel corpora or a machine translation system. Moreover, we demonstrate that incorporating affirmative interpretations yields better results with CondaQA and five other natural language understanding tasks.

## 2 Generating Affirmative Interpretations

An affirmative interpretation generator is a system that takes a sentence with negation as its input and outputs an affirmative interpretation. The task is similar to paraphrase generation with an additional constraint: the output must not contain negation.

We use two approaches to generate affirmative interpretations. The first one is an off-the-shelf T5 (Raffel et al., 2020) fine-tuned by Hossain and Blanco (2022) with Large-AFIN (Section 1) to generate affirmative interpretations. We refer to this model as `T5-HB`, and to the affirmative interpretations generated by `T5-HB` as $A_{HB}$.

The second approach bypasses the need for a large collection of pairs of sentences with negation and their affirmative interpretations. It is based on the work by Vorobev and Kuznetsov (2023), who fine-tuned T5 on a paraphrase dataset obtained with ChatGPT (419,197 sentences and five paraphrases per sentence). We refer to this model as `T5-CG`. Note that it is trained to generate paraphrases—not affirmative interpretations. We obtain affirmative interpretations with `T5-CG` by generating five paraphrases and selecting the first one that does not contain negation. We refer to these affirmative interpretations as $A_{CG}$.[3] For examples of $A_{HB}$ and $A_{CG}$, see Appendix D.

[3]At the time of writing, ChatGPT cannot reliably paraphrase without negation. See an example in Appendix C

We use all negation cues in CondaQA to identify negation cues in our experiments. CondaQA contains over 200 unique cues, including single words (e.g., inaction, unassisted, unknown), affixal negations (e.g., dislike, unmyelinated, unconnected, inadequate, impartial), and multiword expressions (e.g., a lack of, in the absence of, no longer, not at all, rather than). They also include multiple part-of-speech tags such as nouns (e.g., absence, nobody, inability), adverbs (e.g., indirectly, involuntarily, unexpectedly), determiners (e.g., neither, no, none), and verbs (e.g., cannot, refuse, exclude).

## 3 Experimental Results

We use RoBERTa-Large (Liu et al., 2019) as the base model. In addition to experimenting with the original inputs for a task (e.g., passage and question from CondaQA), we couple the original input with one affirmative interpretation of the sentence with negation (if any; no change otherwise). Affirmative interpretations are concatenated to the original input after the `<sep>` special token. Our approach is the same regardless of the type of negation. For implementation details, see Appendix E and F.

### 3.1 CondaQA

CondaQA (Ravichander et al., 2022) is a question-answering dataset that requires reasoning over negation. It was created by asking crowdworkers to write questions about a negated sentence within a paragraph retrieved from Wikipedia. Crowdworkers also made three edits to the original paragraph:

1. *Paraphrase Edit*: Paraphrase the negation.
2. *Scope Edit*: Change the scope of the negation.
3. *Affirmative Edit*: Remove the negation.

Additionally, they answered the question based on the original passage and all three edited passages. (see examples in Appendix G). Note that *paraphrase* edits preserve meaning thus answers remain unchanged. On the other hand, *scope* edits change meaning but the answer may or may not remain the same. Finally, *affirmative* edits reverse meaning thus answers are also reversed.

*Paraphrase* edits are not the same as our affirmative interpretations—crowdworkers were not asked to paraphrase *without using negation*. We discovered, however, that 40.5% of these edits satisfy our definition of affirmative interpretation. We believe crowdworkers simply found it intuitive to paraphrase the negation without using negation. We refer to these affirmative interpretations as $A_G$ (Gold)

| | # Pars. | Input Representation: Training | Input Representation: Testing | Acc. | Group Consistency: All | Group Consistency: Par. | Group Consistency: Sco. | Group Consistency: Aff. |
|---|---|---|---|---|---|---|---|---|
| From Ravichander et al. (2022) | | | | | | | | |
| RoBERTa-Large | 355M | P+Q | P+Q | 54.1 | 13.6 | 51.6 | 26.5 | 27.2 |
| UnifiedQA-v2-Base | 220M | P+Q | P+Q | 58.0 | 17.5 | 54.6 | 30.4 | 33.0 |
| UnifiedQA-v2-Large | 770M | P+Q | P+Q | 66.7 | 30.2 | 64.0 | 43.7 | 46.5 |
| UnifiedQA-v2-3B | 3B | P+Q | P+Q | 73.3 | 42.2 | 72.8 | 55.7 | 57.2 |
| Our Implementation | | | | | | | | |
| RoBERTa-Large | 355M | P+Q | P+Q | 64.9 | 29.6 | 61.3 | 42.3 | 48.3 |
| w/ sentence with neg. from P (S) | | P+Q+S | P+Q+S | 65.2 | 31.1 | 58.4 | 44.1 | 49.2 |
| w/ 1st par. of S by T5-CG ($S_{CG}$) | | P+Q+$S_{CG}$ | P+Q+$S_{CG}$ | 65.7 | 28.4 | 60.8 | 42.4 | 48.6 |
| w/ Affirmative Interpretations | | | | | | | | |
| | | P+Q+$A_{HB}$ | P+Q | 62.8 | 26.3 | 60.5 | 39.2 | 43.3 |
| | | P+Q+$A_{HB}$ | P+Q+$A_{HB}$ | 67.1* | 31.4 | 61.9 | 43.8 | 50.7 |
| | | P+Q+$A_{CG}$ | P+Q | 61.3 | 23.4 | 59.6 | 37.8 | 37.8 |
| | | P+Q+$A_{CG}$ | P+Q+$A_{CG}$ | 66.4* | 31.7 | 62.6 | 44.6 | 49.4 |
| | | P+Q+$A_{HB}$+$A_{CG}$ | P+Q+$A_{HB}$+$A_{CG}$ | 65.6 | 30.1 | 60.9 | 43.7 | 49.9 |
| | | P+Q+$A_G$ | P+Q | 63.6 | 26.7 | 61.4 | 38.8 | 43.9 |
| | | P+Q+$A_G$ | P+Q+$A_{HB}$ | 64.4 | 28.3 | 57.2 | 40.7 | 46.2 |
| | | P+Q+$A_G$ | P+Q+$A_{CG}$ | 65.6 | 30.3 | 61.3 | 42.4 | 49.0 |
| | | P+Q+$A_G$ or $A_{HB}$ | P+Q | 62.5 | 25.7 | 60.1 | 38.6 | 42.4 |
| | | P+Q+$A_G$ or $A_{HB}$ | P+Q+$A_{HB}$ | 65.7 | 30.2 | 61.1 | 41.3 | 48.9 |
| | | P+Q+$A_G$ or $A_{CG}$ | P+Q | 60.6 | 22.0 | 57.9 | 35.2 | 36.8 |
| | | P+Q+$A_G$ or $A_{CG}$ | P+Q+$A_{CG}$ | 66.7* | 32.2 | 62.2 | 44.9 | 50.9 |

Table 1: Results on the CondaQA test set. Q, P and S stand for question, passage and sentence with negation from P. $S_{CG}$ stands for the first paraphrase of S obtained with T5-CG, without avoiding negations. An asterisk ('*') indicates statistically significant improvements (McNemar's test (McNemar, 1947), $p < 0.05$) with respect to not using affirmative interpretations (P+Q). UnifiedQA is fine-tuned with ≈1M question-answer pairs from 20 corpora yet it does not outperform our best approach to incorporate affirmative interpretations (Accuracy: 66.7 vs. 67.1) unless it uses an order of magnitude more parameters (3B vs. 355M). The negated sentence (S) or a paraphrase that is not an affirmative interpretation ($S_{CG}$) bring minor improvements compared to $A_{HB}$ and $A_{CG}$ affirmative interpretations.

and only use them for training purposes, as using them at prediction time would be unrealistic.

Our evaluation reuses the metrics proposed by the authors of CondaQA: accuracy and group consistency. Group consistency is the percentage of questions answered correctly for all the passages in a group. The groups include the original passage and either all three or one of the edited passages.

Table 1 summarizes the experimental results (see Appendix H for additional results). Our implementation of RoBERTa-Large obtains substantially better results than those by Ravichander et al. (2022, Acc.: 64.9 vs. 54.1). Reviewing the training details revealed that the difference is that they stop training after ten epochs while we use early stopping and stop after 18 epochs.

The best-performing model in terms of accuracy is UnifiedQA-v2-3B (Khashabi et al., 2022), which is a 3B-parameter T5 model pre-trained on 20 question-answering corpora data (≈1M, Appendix I). Smaller versions of UnifiedQA (220M and 770M parameters) obtain substantially lower results despite being trained with the same corpora (Acc.: 58.0 and 66.7). Our implementation of RoBERTa-Large using the question and passage as input almost rivals UnifiedQA-v2-Large (64.9 vs. 66.7) despite the latter having twice the size and being fine-tuned with ≈1M question-answering pairs.

Coupling the original input (passage and question) with either the sentence that contains negation (S) or the first paraphrase obtained with T5-CG with no effort to avoid negation ($S_{CG}$) brings minor improvements (64.9 vs. 65.2, 65.7). More interestingly, incorporating affirmative interpretations brings statistically significantly better results (64.9 vs. 67.1 ($A_{HB}$), 66.4 ($A_{CG}$) and 66.7 ($A_G$ or $A_{CG}$/$A_{CG}$)). We conclude the following from the results:

- The benefits of affirmative interpretations are not due to pinpointing the sentence within the passage that is most relevant to answer the question (P+Q+S vs. P+Q+$S_{CG}$ vs. P+Q+$A_{HB}$).
- Training with affirmative interpretations is always beneficial as long as they are also used at prediction time. Note that we only use automatically obtained affirmative interpretations (all but $A_G$) at testing time. However,

| | Negated sentence | Affirmative interpretation |
|---|---|---|
| Adjective (48%) | The island became *completely uninhabited* by 1980 with the automation of the lighthouse. | The island became *vacant* by the 1980s because of the automation of the lighthouse. |
| | They are also made to work the company *unpaid* as a form of "training". | They are made to work the company *free* as a form of "training". |
| Verb (28%) | Early Negro leagues were able to attract top talent but *were unable* to retain them due to financial, logistical and contractual difficulties. | Early Negro Leagues were able to attract top talent but *failed* to retain them due to financial, logistical and contractual difficulties. |
| | Although the original date is *not used in modern times*, it has become an official holiday. | Although the original date was *used in the ancient times*, it has become an official holiday. |
| Quantity (24%) | But *nobody outside of the Muslim world* made daily use of them before Stevin. | *Muslim groups were the only ones* to made daily use of them before Stevin. |
| | However, he enjoyed it but *not at that age*. | He enjoyed it at *another age*. |
| Drop negation without further modifications (10%) | The *unpopular* central government found itself in the difficult position of trying to gain support for spending cuts from the recalcitrant regional governments. | The central government found itself in a difficult position trying to get support for spending cuts from recalcitrant regional governments. |
| | Approximately 30% of the acellular component of bone consists of organic matter, while roughly 70% by mass is attributed to the *inorganic* phase. | Around 30% of the acellular component of bone is made up by organic matter. |

Table 2: Qualitative analysis of $A_{HB}$ affirmative interpretations that result in fixing errors made by the system not using affirmative interpretations with CondaQA (P+Q vs. P+Q+$A_{HB}$, Table 1). The affirmative interpretations rephrase in affirmative terms an adjective (48%), a verb (28%), or a quantity (24%). We also observe that 10% are erroneous as they simply drop the negated content.

| | % w/ negation | % meaning-preserving |
|---|---|---|
| $A_{HB}$ | 23 | 64 |
| $A_{CG}$ | 46 | 83 |
| $S_{CG}$ | 60 | 90 |

Table 3: Qualitative analysis (100 samples from CondaQA) of affirmative interpretations ($A_{HB}$ and $A_{CG}$) and the first paraphrase by `T5-CG` without avoiding negation ($S_{CG}$). Affirmative interpretations are less meaning-preserving, but the experimental results demonstrate that they are more beneficial (Table 1).

using both of them together does not yield better results (Acc.: 65.6 vs. 66.4 and 67.1).

- At training time, complementing $A_G$ (available for ≈40% of paraphrase edits) with $A_{HB}$ or $A_{CG}$ is beneficial (last and second-to-last block).

**Qualitative and Error Analysis** Manual analysis of 100 samples from CondaQA reveals that $A_{CG}$ contains less negations than $A_{HB}$ (46% vs. 23%). $A_{CG}$, however, contains less meaning-preserving paraphrases (36% vs. 17%). On the other hand, paraphrases in $S_{CG}$ rarely do not preserve meaning (10%) but often include negation (60%). (Table 3). Sometimes it is not natural to rewrite a sentence without negation (e.g., *The inner membrane is rich in an unusual phospholipid, cardiolipin.*) Out of the 23 samples where $A_{HB}$ contains negation, a human was able to rewrite 15 of them without negation. Combined with the results from Table 1, this analysis leads to the conclusion that affirmative interpretations are beneficial despite being noisy.

We also analyzed 50 samples of the errors made representing the input with P+Q that are fixed using affirmative interpretations from $A_{HB}$. A negated adjective is replaced by its affirmative counterpart (e.g., *not happy* → *sad*) in 48% of cases. Table 2 shows the analysis and examples of negated sentences and their $A_{HB}$ affirmative interpretations.

### 3.2 Other NLU Tasks

We experiment with five additional NLU tasks to evaluate the benefits of affirmative interpretations. We access these corpora through the GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) benchmarks. We report results on the development set of each corpus, given that the test sets are not publicly available. In addition, we report the results for important and non-important instances as identified by Hossain et al. (2022a). They consider a negation *unimportant* if one can disregard it and still make the correct prediction. For example, *John didn't eat the steak with gusto* (most likely) entails *John ate meat* even if one disregards the negation.

Table 4 presents the results. Incorporating affirmative interpretations ($A_{HB}$ or $A_{CG}$) improves perfor-

| | CmnsnsQA | | STS-B | | QNLI | | WiC | | WSC | |
|---|---|---|---|---|---|---|---|---|---|---|
| | F1 | | Prsn | Sprmn | F1 | | F1 | | F1 | |
| RoBERTa | 0.70 | | 0.92 | 0.92 | 0.93 | | 0.71 | | 0.69 | |
| instances without negation | 0.69 | | 0.92 | 0.92 | 0.93 | | 0.71 | | 0.67 | |
| instances with negation | 0.73 | | 0.88 | 0.88 | 0.92 | | 0.66 | | 0.71 | |
| Important | 0.67 | | 0.82 | 0.85 | 0.78 | | n/a | | n/a | |
| Unimportant | 0.80 | | 0.88 | 0.88 | 0.92 | | 0.66 | | 0.71 | |
| RoBERTa w/ Affirmative Interpret. | | | | | | | | | | |
| obtained using T5-HB ($A_{HB}$) | 0.72 | (+2.9%) | 0.92 | 0.91 | 0.94 | (+1.1%) | 0.70 | (-1.4%) | 0.68 | (-1.4%) |
| instances without negation | 0.72 | (+4.3%) | 0.92 | 0.92 | 0.94 | (+1.1%) | 0.71 | (+0.0%) | 0.62 | (-7.5%) |
| instances with negation | 0.74 | (+1.4%) | 0.88 | 0.88 | 0.92 | (+0.0%) | 0.70 | (+6.1%) | 0.74 | (+4.2%) |
| Important | 0.70 | (+4.5%) | 0.83 | 0.84 | 0.89 | (+14.1%) | n/a | | n/a | |
| Unimportant | 0.80 | (+0.0%) | 0.87 | 0.88 | 0.92 | (+0.0%) | 0.70 | (+6.1%) | 0.74 | (+4.2%) |
| obtained using T5-CG ($A_{CG}$) | 0.71 | (+1.4%) | 0.92 | 0.92 | 0.94 | (+1.1%) | 0.73 | (+2.8%) | 0.71 | (+2.9%) |
| instances without negation | 0.71 | (+2.9%) | 0.93 | 0.92 | 0.94 | (+1.1%) | 0.73 | (+2.8%) | 0.68 | (+1.5%) |
| instances with negation | 0.74 | (+1.4%) | 0.88 | 0.88 | 0.92 | (+0.0%) | 0.70 | (+6.1%) | 0.75 | (+5.6%) |
| Important | 0.69 | (+3.0%) | 0.82 | 0.87 | 0.89 | (+14.1%) | n/a | | n/a | |
| Unimportant | 0.80 | (+0.0%) | 0.88 | 0.88 | 0.92 | (+0.0%) | 0.70 | (+6.1%) | 0.75 | (+5.6%) |

Table 4: Results on additional NLU tasks (macro F1 except with STS-B (Pearson and Spearman correlations)). Percentages between parentheses indicate improvements compared to models not using affirmative interpretations. Affirmative interpretations yield better results, and $A_{CG}$ outperforms $A_{HB}$. The largest gains are with important negations, although we observe gains with instances without negation (up to 4.3%) except with WSC (-7.5%).

mance across all corpora with instances containing important negations; the only exception is STS-B with $A_{HB}$ (Spearman: -1.2%) and $A_{CG}$ (Pearson: no difference). It is worth noting that WiC and WSC have no important negations, yet either $A_{HB}$ or $A_{CG}$ yield substantial improvements with unimportant negations (4.2–6.1%). Surprisingly, we found that incorporating affirmative interpretations is beneficial for instances *without* negation across all corpora except WSC with $A_{HB}$.

These experiments demonstrate that incorporating affirmative interpretations not only obtains higher or comparable results with instances containing important negations, but also often improves results with instances not containing negation.

## 4 Conclusion

We have presented two strategies to generate and incorporate affirmative interpretations into models for natural language understanding. The idea is simple yet effective: complement inputs that contain negation with a paraphrase that does not contain negation. Crucially, we have demonstrated that automatically obtained (noisy) affirmative interpretations yield improvements with (a) CondaQA compared with a model with twice as many parameters pre-trained with ≈1M question-answer pairs from 20 existing corpora and (b) five NLU tasks. Our methodology is architecture- and task-agnostic. In fact, the model to generate affirmative interpretations was tuned with out-of-domain corpora.

**Future Work.** The methods we have presented are simple and effective, but they are not the only way to incorporate or generate affirmative interpretations. For example, one might be able to use LLMs such as GPT-4 or Llama to generate affirmative interpretations. Another interesting direction is to investigate the effect of affirmative interpretations on other NLU tasks, such as sentiment analysis or text classification. Finally, it would be interesting to investigate the effect of affirmative interpretations on other languages, especially those with different word order or negation structures.

## Limitations

The scope of this paper is limited to question answering (CondaQA) and natural language understanding (five tasks and corpora) in English with an emphasis on negation. We leave for future work the task of exploring whether affirmative interpretations are beneficial in other languages. We acknowledge that this strategy might not generalize to other languages.

We also acknowledge that we did not conduct experiments with the latest GPT models or spend substantial amounts of time engineering prompts. We note, however, that good faith efforts using prompts showed that ChatGPT may not be well suited for generating affirmative interpretations at this time (Appendix C).

It is worth pointing out that writing affirmative interpretations for negated sentences might not be straightforward or even possible in some cases. In this paper, we did not focus on the task of determining whether a sentence can be paraphrased without negation. We leave this for future work.

None of the corpora that we work with include information about the scope and focus of negation. Therefore, we do not have any insight into the relation between affirmative interpretations and the scope and focus of a negation.

## Ethics Statement

The work in this paper does not involve human subjects. We only use publicly available datasets and models. We do not collect any personal information. Therefore, this work does not raise any ethical concerns.

## Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 2310334. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

We used computational resources available at the Chameleon testbed to run our experiments (Keahey et al., 2020). We are also grateful to the anonymous reviewers for their valuable comments.

## References

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. 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 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge.

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.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Laurence R. Horn. 1989. *A Natural History of Negation*. University of Chicago Press.

Md Mosharaf Hossain and Eduardo Blanco. 2022. Leveraging affirmative interpretations from negation improves natural language understanding. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5833–5847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Md Mosharaf Hossain, Dhivya Chinnappa, and Eduardo Blanco. 2022a. 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, Luke Holman, Anusha Kakileti, Tiffany Kao, Nathan Brito, Aaron Mathews, and Eduardo Blanco. 2022b. A question-answer driven approach to reveal affirmative interpretations from verbal negations. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 490–503, Seattle, United States. 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.

Myeongjun Jang, Frank Mtumbuka, and Thomas Lukasiewicz. 2022. Beyond distributional hypothesis: Let language models learn meaning-text correspondence. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2030–2042, Seattle, United States. Association for Computational Linguistics.

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.

Kate Keahey, Jason Anderson, Zhuo Zhen, Pierre Riteau, Paul Ruth, Dan Stanzione, Mert Cevik, Jacob Colleran, Haryadi S. Gunawi, Cody Hammock, Joe Mambretti, Alexander Barnes, François Halbach, Alex Rocha, and Joe Stubbs. 2020. Lessons learned from the chameleon testbed. In *Proceedings of the 2020 USENIX Annual Technical Conference (USENIX ATC '20)*. USENIX Association.

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, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Daniel Khashabi, Tushar Khot, and Ashish Sabharwal. 2020. More bang for your buck: Natural perturbation for robust question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 163–170, Online. Association for Computational Linguistics.

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

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8082–8090.

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

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

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.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd Schema Challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, pages 552–561. AAAI Press, Rome, Italy.

Kevin Lin, Oyvind Tafjord, Peter Clark, and Matt Gardner. 2019. Reasoning over paragraph effects in situations. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 58–62, Hong Kong, China. Association for Computational Linguistics.

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.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. 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. Technical Report CTRS-003, Computational Linguistics and Psycholinguistics Technical Report Series.

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 Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Jason Phang, Phil Yeres, Jesse Swanson, Haokun Liu, Ian F. Tenney, Phu Mon Htut, Clara Vania, Alex Wang, and Samuel R. Bowman. 2020. `jiant` 2.0: A software toolkit for research on general-purpose text understanding models. `http://jiant.info/`.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. 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 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

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

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

Melissa Roemmele, Cosmin Bejan, and Andrew Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8732–8740.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. 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 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Zahra Sarabi, Erin Killian, Eduardo Blanco, and Alexis Palmer. 2019. A corpus of negations and their underlying positive interpretations. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 158–167, Minneapolis, Minnesota. Association for Computational Linguistics.

Rituraj Singh, Rahul Kumar, and Vivek Sridhar. 2023. NLMs: Augmenting negation in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13104–13116, Singapore. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Thinh Truong, Timothy Baldwin, Trevor Cohn, and Karin Verspoor. 2022. Improving negation detection with negation-focused pre-training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4188–4193, Seattle, United States. Association for Computational Linguistics.

Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. 2023. Language models are not naysayers: an analysis of language models on negation benchmarks. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 101–114, Toronto, Canada. Association for Computational Linguistics.

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.

Vladimir Vorobev and Maxim Kuznetsov. 2023. A paraphrasing model based on chatgpt paraphrases.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *CoRR*, abs/1905.00537.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

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

## A Paraphrases vs. Affirmative Interpretations

Affirmative interpretations are paraphrases without negation. Table 5 shows examples of automatically generated paraphrases from a negated sentence. Not all of them are correct affirmative interpretations: some (a) contain negation or (b) do not preserve the meaning of the original sentence with negation (and thus they are not actual paraphrases to begin with). The definition of affirmative interpretation is a paraphrase (i.e., rewording that preserves meaning) not containing negation.

Note that an automatically obtained paraphrase that does not preserve the full meaning (and thus does not satisfy the definition of affirmative interpretation) does not necessarily contradict the meaning of the original sentence with negation. For example, *I stayed home today* is not a true paraphrase of *I didn't go shopping today* but is not a contradiction either. In this example, obtaining *I stayed home today*, despite being only plausible and not a paraphrase of *I didn't go shopping today*, could be useful to answer questions such as "Did I go shopping today?" as *staying home* contradicts *going shopping*.

## B NLU Corpora

Table 6 shows examples from the five NLU corpora that we experiment with. The examples are from the development set of each corpus. In our experiments, we append the affirmative interpretation of the negated sentence in the input to the end of the input after a special token.

## C Attempting to Generate Affirmative Interpretations with ChatGPT

At the time of writing, ChatGPT cannot reliably generate affirmative interpretations (i.e., paraphrase without using negation). In the example in Figure 1, it appears convinced to be able to do so, yet it clearly fails: *unhappy* and *lack* are negations. Perhaps surprisingly, ChatGPT appears to know that the generated output does contain negation.

| | Negation? | Same Meaning? |
|---|---|---|
| Original Sentence with Negation: | | |
| The lightning strikes caused no serious permanent damage. | Yes | n/a |
| Automatically Generated Paraphrases (unfiltered): | | |
| The lightning did not cause any damage. | Yes | No |
| The lightning did not cause any significant and permanent damage. | Yes | Yes |
| The lightning strikes caused serious permanent damage. | No | No |
| Lightning strikes caused short-term damage. | No | Yes |

Table 5: Examples of automatically generated paraphrases from a negated sentence. The first two paraphrases contain negation, and only the second one preserves meaning. The next two paraphrases do not contain negation, and only the fourth one preserves meaning. Only the fourth automatically obtained paraphrase is an affirmative interpretation: it does not contain negation and it is a true paraphrase of the original sentence with negation—*not causing serious permanent damage* carries roughly the same meaning than *causing short-term damage*.

| | Input | Output |
|---|---|---|
| Question Answering | | |
| CommonsenseQA | What are you waiting alongside with when you're in a reception area?<br>A) Motel, B) Chair, C) Hospital, D) People, E) Hotel | D |
| Similarity and Paraphrasing | | |
| STS-B | Three men are playing guitars.<br>Three men are on stage playing guitars. | 3.75 (out of 5) |
| Inference | | |
| QNLI | What happened to Dane?<br>Dane was killed in a horse-riding accident when Nikola was five. | Entailment (i.e., question is answered) |
| Word Sense Disambiguation | | |
| WiC | Room and *board*.<br>He nailed *boards* across the windows. | Not same meaning |
| Coreference Resolution | | |
| WSC | Mark told *Pete* many lies about himself, which Pete included in his book. *He* should have been more truthful. | Not coreferent |

Table 6: Examples of instances from the NLU tasks used in our experiments. The first column indicates the task and the corpus. The second column shows the input to the system. The third column shows the expected output.

| Negated Sentence and Affirmative Interpretations | | Correct? |
|---|---|---|
| Negated Sentence | The National Palace is one of Managua's oldest buildings, undamaged by the 1972 earthquake. | n/a |
| $A_{HB}$ | The National Palace, one of Managua's oldest buildings, survived the 1972 earthquake. | Yes |
| $A_{CG}$ | The National Palace, which was one of the oldest structures in Managua, remained intact following the 1972 earthquake. | Yes |
| Negated Sentence | It is not rare to find pearls that measure as much as 14mm across. | n/a |
| $A_{HB}$ | It is not uncommon to find pearls that measure as much as 14mm across. | No |
| $A_{CG}$ | The size of 14mm pearls is not uncommon. | No |
| Human | It is common to find pearls that measure as much as 14mm across. | Yes |

Table 7: Examples of negated sentences and affirmative interpretations generated by T5-HB ($A_{HB}$) and T5-CG ($A_{CG}$). The last column indicates whether affirmative interpretation are correct (i.e., meaning preserving and without negation). *Human* is a human-generated affirmative interpretation.

| |
|---|
| - *Negated Sentence*: An increasing minority of young people cannot understand Japanese and instead use the Ryukyuan languages only.<br>- *(wrong) Affirmative Interpretation*: An increasing minority of young people understand only Ryukyuan languages instead.<br>- *Observation*: The affirmative interpretations drops an important part of the meaning of the negated sentence: not understanding Japanese. We note, however, that the affirmative interpretation is factual given the negated sentence and it is likely to be useful for downstream tasks. |
| - *Negated Sentence*: By war's end, no other nation formally recognized the Confederacy.<br>- *(wrong) Affirmative Interpretation*: Only one nation recognized the Confederacy at the end of war.<br>- *Observation*: This error seems to be due to lack of context of the negated sentence. The affirmative interpretation does not have negation and is plausible, but it is impossible to know how many nations recognized the Confederacy without additional context. Indeed, *no other nation did X* could mean that *any number of nations did X*. |

Table 8: A couple examples of negated sentence from CondaQA and automatically generated affirmative interpretations that are wrong. We also provide our observations.

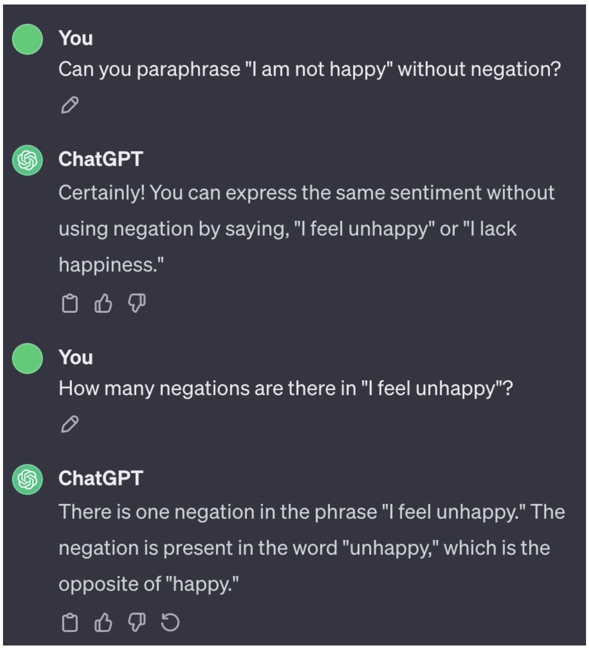


Figure 1: Attempting to generate affirmative interpretations with ChatGPT results in a nonsensical conversation. ChatGPT appears to be able to identify negations yet uses them when instructed to not do so

| Instance Representation | | |
|---|---|---|
| Training | Testing | Learning Rate |
| P+Q | P+Q | 1e-5 |
| P+Q+S | P+Q+S | 5e-6 |
| P+Q+$P_{CG}$ | P+Q+$P_{CG}$ | 1e-5 |
| P+Q+$A_{HB}$ | P+Q | 1e-5 |
| P+Q+$A_{HB}$ | P+Q+$A_{HB}$ | 1e-4 |
| P+Q+$A_{CG}$ | P+Q | 1e-5 |
| P+Q+$A_{CG}$ | P+Q+$A_{CG}$ | 1e-4 |
| P+Q+$A_{HB}$+$A_{CG}$ | P+Q+$A_{HB}$+$A_{CG}$ | 1e-5 |
| P+Q+$A_G$ | P+Q | 1e-5 |
| P+Q+$A_G$ | P+Q+$A_{HB}$ | 1e-5 |
| P+Q+$A_G$ | P+Q+$A_{CG}$ | 1e-5 |
| P+Q+$A_G$ or $A_{HB}$ | P+Q | 1e-5 |
| P+Q+$A_G$ or $A_{HB}$ | P+Q+$A_{HB}$ | 5e-5 |
| P+Q+$A_G$ or $A_{CG}$ | P+Q | 1e-5 |
| P+Q+$A_G$ or $A_{CG}$ | P+Q+$A_{CG}$ | 5e-5 |

Table 9: Learning rates used in our experiments with CondaQA. Note that $A_G$ affirmative interpretations are not available at testing time.

## D Affirmative Interpretations Examples

Table 7 shows two negated sentences and their automatically obtained affirmative interpretations. The bottom half of the table includes errors, as the automatically generated affirmative interpretations contain negations. Table 8 contains a couple examples from CondaQA in which the process to generate affirmative interpretations made mistakes along with our observations.

## E Training Details with CondaQA

We use the RoBERTa-Large model (Liu et al., 2019) for our experiments with CondaQA. We use the implementation of RoBERTa-Large in the HuggingFace Transformers library (Wolf et al., 2020). The model is trained using early stopping with a patience of 3 epochs and batch size 16. Table 9 shows the learning rates that we used for our experiments with CondaQA. We use the default values for the other hyperparameters.

## F Training Details with Additional NLU Tasks

We use the implementation by Phang et al. (2020) with RoBERTa-Large as the base model. We use the default values for the hyperparameters, with the exception of the learning rate, batch size and maximum number of epochs for early stopping.

Table 10 shows the learning rates and batch sizes that we used for our experiments on each corpus.

## G CondaQA Dataset

Figure 2 shows an example from CondaQA. Note that CondaQA highlights the original negated sentences from the original passages but not the edited sentences. However, we use the available information in the dataset such as the original sentence, the original passage and the edited passage to extract the edited sentences. Specifically, we identify sentence boundaries in the original passage and pinpoint the index of the sentence that contains negation. Then, we identify sentence boundaries in the edited passage and use the same index to extract the edited sentence. We use the extracted edited sentence to generate affirmative interpretations. The authors manually analyzed 100 samples of the extracted edited sentences and confirmed that in 96% of the cases, the extracted edited sentences are the same as the edited sentences in the passage.

Additionally, Table 11 shows the basic properties of the edits made by crowdworkers.

## H Additional Results with CondaQA

Table 12 shows additional results with RoBERTa-Large and CondaQA for each edit type. The results show that incorporating affirmative interpretations with RoBERTa-Large improves results not only with the entire test set, but also with each edit type individually. However, not all of the improvements are statistically significant. The only statistically significant improvements are with (1) the scope edit type when trained with P+Q+$\mathsf{A}_{\mathsf{CG}}$ or $\mathsf{A}_{\mathsf{CG}}$ and tested with P+Q+$\mathsf{A}_{\mathsf{CG}}$, and (2) the affirmative edit type when trained with P+Q+$\mathsf{A}_{\mathsf{HB}}$ and tested with P+Q+$\mathsf{A}_{\mathsf{HB}}$.

## I UnifiedQA-v2 Training Corpora

Table 13 shows the QA corpora that Khashabi et al. (2022) used to train UnifiedQA-v2. These corpora span the following QA formats: extractive, abstractive, multiple-choice, and yes-no questions.

| | CmmnsnsQA | STS-B | QNLI | WiC | WSC |
|---|---|---|---|---|---|
| RoBERTa | 1e-5 (16) | 1e-5 (16) | 1e-5 (8) | 1e-5 (16) | 1e-6 (16) |
| RoBERTa w/ Affirmative Interpret. | | | | | |
| obtained using T5-HB ($A_{HB}$) | 5e-6 (16) | 5e-6 (8) | 5e-6 (16) | 1e-5 (16) | 5e-6 (16) |
| obtained using T5-CG ($A_{CG}$) | 5e-6 (16) | 5e-6 (16) | 1e-5 (16) | 5e-6 (16) | 5e-6 (16) |

Table 10: The learning rates (and batch sizes) used in our experiments with each corpus.

| | |
|---|---|
| Original Passage: | A semiconductor diode is a device typically made from a single p-n junction. At the junction of a p-type and an n-type semiconductor, there forms a depletion region where current conduction is inhibited by the lack of mobile charge carriers. When the device is "forward biased" (connected with the p-side at higher electric potential than the n-side), this depletion region is diminished, allowing for significant conduction, while only very small current can be achieved when the diode is "reverse biased" and thus the depletion region expanded. |
| Original Sentence (with Negation): | At the junction of a p-type and an n-type semiconductor, there forms a depletion region where current conduction is inhibited by the lack of mobile charge carriers. |
| Negation Cue: | lack |
| Edited Passage: | A semiconductor diode is a device typically made from a single p-n junction. At the junction of a p-type and an n-type semiconductor there forms a depletion region where current conduction is inhibited by the absence of mobile charge carriers. When the device is "forward biased" (connected with the p-side at higher electric potential than the n-side), this depletion region is diminished, allowing for significant conduction, while only very small current can be achieved when the diode is "reverse biased" and thus the depletion region expanded. |
| Edit Type: | Paraphrase |
| Question: | Is the current conduction negatively affected by the amount of mobile charge carriers? |
| Answer: | Yes |
| Extracted Edited Sentence: | At the junction of a p-type and an n-type semiconductor there forms a depletion region where current conduction is inhibited by the absence of mobile charge carriers. |

Figure 2: An example from CondaQA. The negation in the original sentence is *lack*. The crowdworkers wrote a paraphrase of the original sentence, which is included in the edited passage (*[. . . ] by the absence of mobile charge carriers*). The question is written based on the original paragraph and answered based on the original and all three edited passages (only paraphrase edit shown). The answer to the question (for the edited passage) is *Yes*. The dataset does not explicitly indicate the edited sentence. However, we extract it as explained in Appendix G.

| Edit | % Negated | Meaning | Answer |
|---|---|---|---|
| Paraphrase | 59.5 | Same | Unchanged |
| Scope | 97.7 | Changed | Unchanged or changed |
| Affirmative | 43.6 | Reversed | Reversed |

Table 11: Basic properties of the edits made by crowdworkers in the process of creating CondaQA. The *Negated* column shows the percentage of edits that have negation. The *Meaning* and *Answer* columns indicate the differences in meaning (if any) between (1) the original and edited passage and (2) answers to the same question according to the original and edited passage. *Changed* does not necessarily mean *reversed*.

| | # Params. | Input Representation: Training | Input Representation: Testing | Accuracy: All | Ori. | Par. | Sco. | Aff. |
|---|---|---|---|---|---|---|---|---|
| RoBERTa-Large | 355M | Q | Q | 47.4 | 52.1 | 52.3 | 47.4 | 39.0 |
| | | P | P | 45.4 | 46.5 | 46.1 | 45.2 | 43.9 |
| | | P+Q | P+Q | 64.9 | 67.2 | 66.0 | 59.5 | 66.0 |
| w/ sentence with neg. from P (S) | | P+Q+S | P+Q+S | 65.2 | 66.0 | 64.6 | 61.8 | 68.3 |
| w/ 1st par. of S by T5-CG ($S_{CG}$) | | P+Q+$S_{CG}$ | P+Q+$S_{CG}$ | 65.7 | 68.3 | 67.1 | 60.2 | 67.0 |
| w/ Affirmative Interpretations | | | | | | | | |
| | | P+Q+$A_{HB}$ | P+Q | 62.8 | 64.6 | 62.9 | 58.6 | 64.9 |
| | | P+Q+$A_{HB}$ | P+Q+$A_{HB}$ | 67.1* | 68.5 | 68.0 | 61.8 | 69.7* |
| | | P+Q+$A_{CG}$ | P+Q | 61.3 | 64.7 | 62.3 | 58.2 | 59.8 |
| | | P+Q+$A_{CG}$ | P+Q+$A_{CG}$ | 66.4* | 68.6 | 67.2 | 61.7 | 67.8 |
| | | P+Q+$A_{HB}$+$A_{CG}$ | P+Q+$A_{HB}$+$A_{CG}$ | 65.6 | 68.4 | 66.6 | 59.4 | 67.6 |
| | | P+Q+$A_{G}$ | P+Q | 63.6 | 65.2 | 64.8 | 58.6 | 65.5 |
| | | P+Q+$A_{G}$ | P+Q+$A_{HB}$ | 64.4 | 65.5 | 65.3 | 60.3 | 66.2 |
| | | P+Q+$A_{G}$ | P+Q+$A_{CG}$ | 65.6 | 67.2 | 66.8 | 59.7 | 68.2 |
| | | P+Q+$A_{G}$ or $A_{HB}$ | P+Q | 62.5 | 64.2 | 63.4 | 58.5 | 63.6 |
| | | P+Q+$A_{G}$ or $A_{HB}$ | P+Q+$A_{HB}$ | 65.7 | 67.2 | 67.2 | 59.6 | 68.2 |
| | | P+Q+$A_{G}$ or $A_{CG}$ | P+Q | 60.6 | 62.6 | 61.7 | 57.6 | 60.3 |
| | | P+Q+$A_{G}$ or $A_{CG}$ | P+Q+$A_{CG}$ | 66.7* | 69.0 | 67.2 | 62.4* | 67.8 |

Table 12: The accuracy of RoBERTa-Large on the CondaQA test set for each edit type. We indicate statistically significant improvements (McNemar's test (McNemar, 1947), $p < 0.05$) with respect to the model trained without affirmative interpretations (P+Q during training and testing) on each edit type with an asterisk (*).

| Corpus | # Train Inst. | Reference |
|---|---|---|
| Squad 1.1 | 87,599 | Rajpurkar et al. (2016) |
| Squad 2 | 130,319 | Rajpurkar et al. (2018) |
| Newsqa | 92,549 | Trischler et al. (2017) |
| Quoref | 19,399 | Dasigi et al. (2019) |
| Ropes | 10,924 | Lin et al. (2019) |
| NarrativeQA | 32,747 | Kočiský et al. (2018) |
| DROP | 77,409 | Dua et al. (2019) |
| NaturalQuestions | 307,373 | Kwiatkowski et al. (2019) |
| MCTest | 1,480 | Richardson et al. (2013) |
| RACE | 87,866 | Lai et al. (2017) |
| OpenBookQA | 4,957 | Mihaylov et al. (2018) |
| ARC | 2,590 | Clark et al. (2018) |
| CommonsenseQA | 9,741 | Talmor et al. (2019) |
| QASC | 8,134 | Khot et al. (2020) |
| PhysicalIQA | 16,000 | Bisk et al. (2019) |
| SocialIQA | 33,410 | Sap et al. (2019) |
| Winogrande | 40,398 | Sakaguchi et al. (2020) |
| BoolQ | 9,427 | Clark et al. (2019) |
| MultiRC (yes/no) | 6,000 | Khashabi et al. (2018) |
| BoolQ-NP | 9,727 | Khashabi et al. (2020) |

Table 13: The corpora that Khashabi et al. (2022) used to train UnifiedQA-v2, and the number of training instances in each corpus.