

# Paraphrasing in Affirmative Terms Improves Negation Understanding

MohammadHossein Rezaei   Eduardo Blanco  
University of Arizona





Muslim mathematicians were the first to utilize decimals instead of fractions on a large scale...But *nobody* outside of the Muslim world make daily use of them before Stevin.



Muslim mathematicians were the first to utilize decimals instead of fractions on a large scale...But *nobody* outside of the Muslim world make daily use of them before Stevin.



If you were able to find accounting records from the Middle East that slightly predated Stevin's publication, would you likely see decimals being used for transactions, even if they were in an Arabic script?



Muslim mathematicians were the first to utilize decimals instead of fractions on a large scale...But *nobody* outside of the Muslim world make daily use of them before Stevin.



If you were able to find accounting records from the Middle East that slightly predated Stevin's publication, would you likely see decimals being used for transactions, even if they were in an Arabic script?



No



# Affirmative Interpretations

# Affirmative Interpretations

- *Definition:* paraphrases without negation

# Affirmative Interpretations

- *Definition:* paraphrases without negation
- Example: I am *not* sad.

# Affirmative Interpretations

- *Definition:* paraphrases without negation
- Example: I am *not* sad.
- Affirmative Interpretation: I am happy.



# Affirmative Interpretations

- *Definition:* paraphrases without negation
- Example: I am *not* sad.
- Affirmative Interpretation: I am happy. I am just ok.



Muslim mathematicians were the first to utilize decimals instead of fractions on a large scale...But *nobody* outside of the Muslim world make daily use of them before Stevin.



If you were able to find accounting records from the Middle East that slightly predated Stevin's publication, would you likely see decimals being used for transactions, even if they were in an Arabic script?



Muslim mathematicians were the first to utilize decimals instead of fractions on a large scale...But *nobody* outside of the Muslim world make daily use of them before Stevin. [SEP] Muslim groups were the *only ones* to made daily use of them before Stevin.



If you were able to find accounting records from the Middle East that slightly predated Stevin's publication, would you likely see decimals being used for transactions, even if they were in an Arabic script?



Muslim mathematicians were the first to utilize decimals instead of fractions on a large scale...But *nobody* outside of the Muslim world make daily use of them before Stevin. [SEP] Muslim groups were the *only ones* to made daily use of them before Stevin.



If you were able to find accounting records from the Middle East that slightly predated Stevin's publication, would you likely see decimals being used for transactions, even if they were in an Arabic script?



Yes



# Generating Affirmative Interpretations

# Generating Affirmative Interpretations

- T5-HB (Hossain and Blanco, 2022)

# Generating Affirmative Interpretations

- T5-HB (Hossain and Blanco, 2022)
  - Off-the-shelf affirmative interpretation generator

# Generating Affirmative Interpretations

- T5-HB (Hossain and Blanco, 2022)
  - Off-the-shelf affirmative interpretation generator
  - Trained on Large-AFIN



# Generating Affirmative Interpretations

- T5-HB (Hossain and Blanco, 2022)
  - Off-the-shelf affirmative interpretation generator
  - Trained on Large-AFIN
    - 153k (sent w/ neg, affirmative interpretation) pairs

# Generating Affirmative Interpretations

- T5-HB (Hossain and Blanco, 2022)
  - Off-the-shelf affirmative interpretation generator
  - Trained on Large-AFIN
    - 153k (sent w/ neg, affirmative interpretation) pairs
    - Obtained using back-translation
  - Affirmative interpretations:  $A_{HB}$

# Generating Affirmative Interpretations

- T5-HB (Hossain and Blanco, 2022)
  - Off-the-shelf affirmative interpretation generator
  - Trained on Large-AFIN
    - 153k (sent w/ neg, affirmative interpretation) pairs
    - Obtained using back-translation
  - Affirmative interpretations:  $A_{HB}$
- T5-CG (Vorobev and Kuznetsov, 2023)

# Generating Affirmative Interpretations

- T5-HB (Hossain and Blanco, 2022)
  - Off-the-shelf affirmative interpretation generator
  - Trained on Large-AFIN
    - 153k (sent w/ neg, affirmative interpretation) pairs
    - Obtained using back-translation
  - Affirmative interpretations:  $A_{HB}$
- T5-CG (Vorobev and Kuznetsov, 2023)
  - Off-the-shelf paraphraser

# Generating Affirmative Interpretations

- T5-HB (Hossain and Blanco, 2022)
  - Off-the-shelf affirmative interpretation generator
  - Trained on Large-AFIN
    - 153k (sent w/ neg, affirmative interpretation) pairs
    - Obtained using back-translation
  - Affirmative interpretations:  $A_{HB}$
- T5-CG (Vorobev and Kuznetsov, 2023)
  - Off-the-shelf paraphraser
  - Trained on 419k pairs of sentences and paraphrases generated by GPT-3.5

# Generating Affirmative Interpretations

- T5-HB (Hossain and Blanco, 2022)
  - Off-the-shelf affirmative interpretation generator
  - Trained on Large-AFIN
    - 153k (sent w/ neg, affirmative interpretation) pairs
    - Obtained using back-translation
  - Affirmative interpretations:  $A_{HB}$
- T5-CG (Vorobev and Kuznetsov, 2023)
  - Off-the-shelf paraphraser
  - Trained on 419k pairs of sentences and paraphrases generated by GPT-3.5
  - Generate 5 paraphrases and choose the first one without negation, if any;

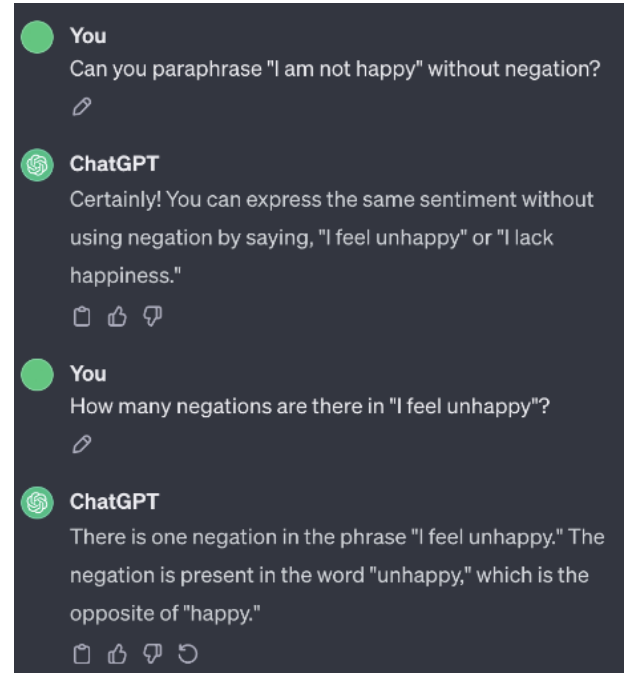
# Generating Affirmative Interpretations


- T5-HB (Hossain and Blanco, 2022)
  - Off-the-shelf affirmative interpretation generator
  - Trained on Large-AFIN
    - 153k (sent w/ neg, affirmative interpretation) pairs
    - Obtained using back-translation
  - Affirmative interpretations:  $A_{HB}$
- T5-CG (Vorobev and Kuznetsov, 2023)
  - Off-the-shelf paraphraser
  - Trained on 419k pairs of sentences and paraphrases generated by GPT-3.5
  - Generate 5 paraphrases and choose the first one without negation, if any; otherwise, choose the first one
  - Affirmative interpretations:  $A_{CG}$
  - Simple Paraphrases:  $S_{CG}$




# Generating Affirmative Interpretations








# Generating Affirmative Interpretations



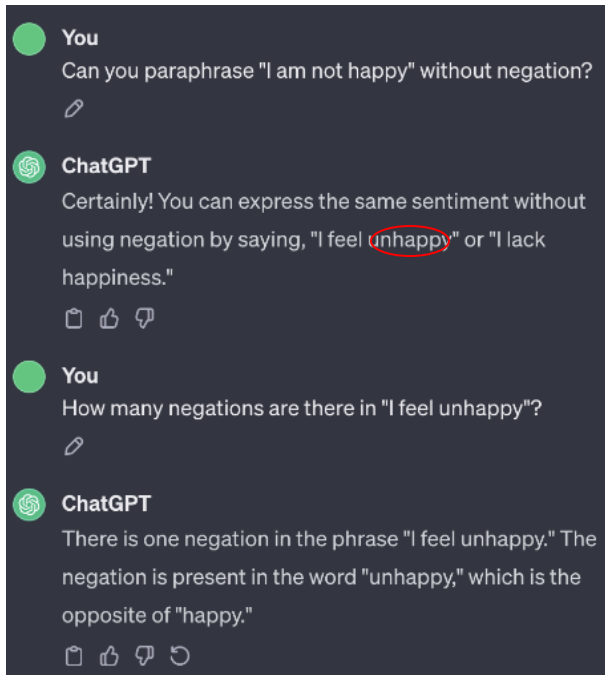
**You**  
Can you paraphrase "I am not happy" without negation?  


**ChatGPT**  
Certainly! You can express the same sentiment without using negation by saying, "I feel unhappy" or "I lack happiness."  
  


**You**  
How many negations are there in "I feel unhappy"?.  





**ChatGPT**  
There is one negation in the phrase "I feel unhappy." The negation is present in the word "unhappy," which is the opposite of "happy."  
   


# Generating Affirmative Interpretations







The image shows a chat interface with a dark background. It contains four messages in a vertical sequence. The first message is from 'You' (green circle icon) asking for a paraphrase of 'I am not happy' without negation. The second message is from 'ChatGPT' (green robot icon) providing an affirmative interpretation: 'I feel unhappy' or 'I lack happiness', with 'unhappy' circled in red. The third message is from 'You' asking how many negations are in 'I feel unhappy'. The fourth message is from 'ChatGPT' stating there is one negation in the word 'unhappy'.

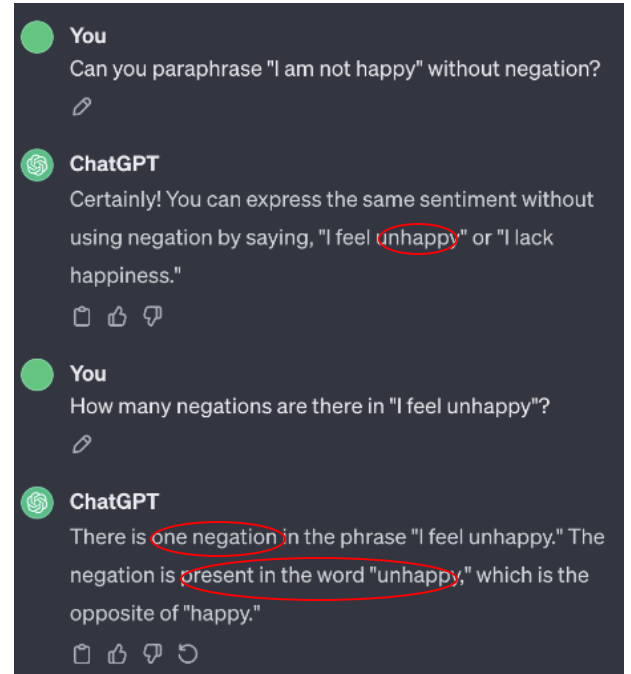
**You**  
Can you paraphrase "I am not happy" without negation?  



**ChatGPT**  
Certainly! You can express the same sentiment without using negation by saying, "I feel **unhappy**" or "I lack happiness."  
  




**You**  
How many negations are there in "I feel unhappy"?:  



**ChatGPT**  
There is one negation in the phrase "I feel unhappy." The negation is present in the word "unhappy," which is the opposite of "happy."  
   





# Generating Affirmative Interpretations



**You**  
Can you paraphrase "I am not happy" without negation?  


**ChatGPT**  
Certainly! You can express the same sentiment without using negation by saying, "I feel **unhappy**" or "I lack happiness."  
  

**You**  
How many negations are there in "I feel unhappy"??  


**ChatGPT**  
There is **one negation** in the phrase "I feel unhappy." The negation is **present in the word "unhappy,"** which is the opposite of "happy."  
   

# Evaluation

# Evaluation

- CondaQA (Ravichander et al., 2022)

# Evaluation

- CondaQA (Ravichander et al., 2022)
  - The largest question-answering corpus that requires reasoning over negation

# Evaluation

- CondaQA (Ravichander et al., 2022)
  - The largest question-answering corpus that requires reasoning over negation
- Five NLU Tasks

# Evaluation

- CondaQA (Ravichander et al., 2022)
  - The largest question-answering corpus that requires reasoning over negation
- Five NLU Tasks
  - CommonsenseQA
  - STS-B
  - QNLI
  - WiC
  - WSC



# Results

	# Pars.	Input Representation		Acc.	Group Consistency				
		Training	Testing		All	Par.	Sco.	Aff.	
From <a href="#">Ravichander et al. (2022)</a>									
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	

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  $\approx 1M$  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.

# Results

	# Pars.	Input Representation		Acc.	Group Consistency			
		Training	Testing		All	Par.	Sco.	Aff.
From <a href="#">Ravichander et al. (2022)</a>								
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

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  $\approx 1M$  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.

# Results

	# Pars.	Input Representation		Acc.	Group Consistency				
		Training	Testing		All	Par.	Sco.	Aff.	
From <a href="#">Ravichander et al. (2022)</a>									
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	

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  $\approx 1M$  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.

# Results

	# Pars.	Input Representation		Acc.	Group Consistency				
		Training	Testing		All	Par.	Sco.	Aff.	
From <a href="#">Ravichander et al. (2022)</a>									
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	

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  $\approx 1M$  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.

# Results

	# Pars.	Input Representation		Acc.	Group Consistency				
		Training	Testing		All	Par.	Sco.	Aff.	
From <a href="#">Ravichander et al. (2022)</a>									
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	

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  $\approx 1M$  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.

# Results

	# Pars.	Input Representation		Acc.	Group Consistency				
		Training	Testing		All	Par.	Sco.	Aff.	
From <a href="#">Ravichander et al. (2022)</a>									
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 <sub>CG</sub>	P+Q+S <sub>CG</sub>	65.7	28.4	60.8	42.4	48.6	
w/ Affirmative Interpretations		P+Q+A <sub>HB</sub>	P+Q	62.8	26.3	60.5	39.2	43.3	
		P+Q+A <sub>HB</sub>	P+Q+A <sub>HB</sub>	67.1*	31.4	61.9	43.8	50.7	

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  $\approx 1M$  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.

# Results

	# Pars.	Input Representation		Acc.	Group Consistency			
		Training	Testing		All	Par.	Sco.	Aff.
From <a href="#">Ravichander et al. (2022)</a>								
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 <sub>CG</sub>	P+Q+S <sub>CG</sub>	65.7	28.4	60.8	42.4	48.6
w/ Affirmative Interpretations		P+Q+A <sub>HB</sub>	P+Q	62.8	26.3	60.5	39.2	43.3
		P+Q+A <sub>HB</sub>	P+Q+A <sub>HB</sub>	67.1*	31.4	61.9	43.8	50.7

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  $\approx 1M$  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.

# Results

	# Pars.	Input Representation		Acc.	Group Consistency			
		Training	Testing		All	Par.	Sco.	Aff.
From <a href="#">Ravichander et al. (2022)</a>								
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 <sub>CG</sub>	P+Q+S <sub>CG</sub>	65.7	28.4	60.8	42.4	48.6
w/ Affirmative Interpretations		P+Q+A <sub>HB</sub>	P+Q	62.8	26.3	60.5	39.2	43.3
		P+Q+A <sub>HB</sub>	P+Q+A <sub>HB</sub>	67.1*	31.4	61.9	43.8	50.7

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  $\approx 1M$  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.



# Results

	# Pars.	Input Representation		Acc.	Group Consistency			
		Training	Testing		All	Par.	Sco.	Aff.
From <a href="#">Ravichander et al. (2022)</a>								
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  $\approx 1M$  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.

# Results

	# Pars.	Input Representation		Acc.	Group Consistency			
		Training	Testing		All	Par.	Sco.	Aff.
From <a href="#">Ravichander et al. (2022)</a>								
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 <sub>CG</sub>	P+Q+S <sub>CG</sub>	65.7	28.4	60.8	42.4	48.6
w/ Affirmative Interpretations		P+Q+A <sub>HB</sub>	P+Q	62.8	26.3	60.5	39.2	43.3
		P+Q+A <sub>HB</sub>	P+Q+A <sub>HB</sub>	67.1*	31.4	61.9	43.8	50.7
		P+Q+A <sub>CG</sub>	P+Q	61.3	23.4	59.6	37.8	37.8
		P+Q+A <sub>CG</sub>	P+Q+A <sub>CG</sub>	66.4*	31.7	62.6	44.6	49.4
		P+Q+A <sub>HB</sub> +A <sub>CG</sub>	P+Q+A <sub>HB</sub> +A <sub>CG</sub>	65.6	30.1	60.9	43.7	49.9
		P+Q+A <sub>G</sub>	P+Q	63.6	26.7	61.4	38.8	43.9
		P+Q+A <sub>G</sub>	P+Q+A <sub>HB</sub>	64.4	28.3	57.2	40.7	46.2
		P+Q+A <sub>G</sub>	P+Q+A <sub>CG</sub>	65.6	30.3	61.3	42.4	49.0
		P+Q+A <sub>G</sub> or A <sub>HB</sub>	P+Q	62.5	25.7	60.1	38.6	42.4
		P+Q+A <sub>G</sub> or A <sub>HB</sub>	P+Q+A <sub>HB</sub>	65.7	30.2	61.1	41.3	48.9
		P+Q+A <sub>G</sub> or A <sub>CG</sub>	P+Q	60.6	22.0	57.9	35.2	36.8
		P+Q+A <sub>G</sub> or A <sub>CG</sub>	P+Q+A <sub>CG</sub>	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  $\approx 1M$  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<sub>HB</sub> and A<sub>CG</sub> affirmative interpretations.

# Results

	# Pars.	Input Representation		Acc.	Group Consistency			
		Training	Testing		All	Par.	Sco.	Aff.
From <a href="#">Ravichander et al. (2022)</a>								
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 <sub>CG</sub>	P+Q+S <sub>CG</sub>	65.7	28.4	60.8	42.4	48.6
w/ Affirmative Interpretations		P+Q+A <sub>HB</sub>	P+Q	62.8	26.3	60.5	39.2	43.3
		P+Q+A <sub>HB</sub>	P+Q+A <sub>HB</sub>	67.1	31.4	61.9	43.8	50.7
		P+Q+A <sub>CG</sub>	P+Q	61.3	23.4	59.6	37.8	37.8
		P+Q+A <sub>CG</sub>	P+Q+A <sub>CG</sub>	66.4	31.7	62.6	44.6	49.4
		P+Q+A <sub>HB</sub> +A <sub>CG</sub>	P+Q+A <sub>HB</sub> +A <sub>CG</sub>	65.6	30.1	60.9	43.7	49.9
		P+Q+A <sub>G</sub>	P+Q	63.6	26.7	61.4	38.8	43.9
		P+Q+A <sub>G</sub>	P+Q+A <sub>HB</sub>	64.4	28.3	57.2	40.7	46.2
		P+Q+A <sub>G</sub>	P+Q+A <sub>CG</sub>	65.6	30.3	61.3	42.4	49.0
		P+Q+A <sub>G</sub> or A <sub>HB</sub>	P+Q	62.5	25.7	60.1	38.6	42.4
		P+Q+A <sub>G</sub> or A <sub>HB</sub>	P+Q+A <sub>HB</sub>	65.7	30.2	61.1	41.3	48.9
		P+Q+A <sub>G</sub> or A <sub>CG</sub>	P+Q	60.6	22.0	57.9	35.2	36.8
		P+Q+A <sub>G</sub> or A <sub>CG</sub>	P+Q+A <sub>CG</sub>	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  $\approx 1M$  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<sub>HB</sub> and A<sub>CG</sub> affirmative interpretations.

# Qualitative and Error Analysis

---

Negated sentence	Affirmative interpretation
------------------	----------------------------

---

---

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.

# Qualitative and Error Analysis

	Negated sentence	Affirmative interpretation
Adjective (48%)	The island became <i>completely uninhabited</i> by 1980 with the automation of the lighthouse.  They are also made to work the company <i>unpaid</i> as a form of "training".	The island became <i>vacant</i> by the 1980s because of the automation of the lighthouse.  They are made to work the company <i>free</i> as a form of "training".

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<sub>HB</sub>, 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.

# Qualitative and Error Analysis

	Negated sentence	Affirmative interpretation
Adjective (48%)	The island became <i>completely uninhabited</i> by 1980 with the automation of the lighthouse.  They are also made to work the company <i>unpaid</i> as a form of "training".	The island became <i>vacant</i> by the 1980s because of the automation of the lighthouse.  They are made to work the company <i>free</i> as a form of "training".
Verb (28%)	Early Negro leagues were able to attract top talent but <i>were unable</i> to retain them due to financial, logistical and contractual difficulties.  Although the original date is <i>not used in modern times</i> , it has become an official holiday.	Early Negro Leagues were able to attract top talent but <i>failed</i> to retain them due to financial, logistical and contractual difficulties.  Although the original date was <i>used in the ancient times</i> , it has become an official holiday.

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<sub>HB</sub>, 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.

# Qualitative and Error Analysis

	Negated sentence	Affirmative interpretation
Adjective (48%)	The island became <i>completely uninhabited</i> by 1980 with the automation of the lighthouse.	The island became <i>vacant</i> by the 1980s because of the automation of the lighthouse.
	They are also made to work the company <i>unpaid</i> as a form of "training".	They are made to work the company <i>free</i> as a form of "training".
Verb (28%)	Early Negro leagues were able to attract top talent but <i>were unable</i> to retain them due to financial, logistical and contractual difficulties.	Early Negro Leagues were able to attract top talent but <i>failed</i> to retain them due to financial, logistical and contractual difficulties.
	Although the original date is <i>not used in modern times</i> , it has become an official holiday.	Although the original date was <i>used in the ancient times</i> , it has become an official holiday.
Quantity (24%)	But <i>nobody outside of the Muslim world</i> made daily use of them before Stevin.	<i>Muslim groups were the only ones</i> to made daily use of them before Stevin.
	However, he enjoyed it but <i>not at that age</i> .	He enjoyed it at <i>another age</i> .
Drop negation without further modifications (10%)	The <i>unpopular</i> 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 <i>inorganic</i> 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.

# Qualitative and Error Analysis

- Affirmative Interpretations should:



# Qualitative and Error Analysis

- Affirmative Interpretations should:
  - (a) not contain negation

# Qualitative and Error Analysis

- Affirmative Interpretations should:
  - (a) not contain negation
  - (b) preserve the meaning

# Qualitative and Error Analysis

- Affirmative Interpretations should:
  - (a) not contain negation
  - (b) preserve the meaning

---

---

<i>% w/ negation</i>	<i>% meaning-preserving</i>
----------------------	-----------------------------

---

---

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).

# Qualitative and Error Analysis

- Affirmative Interpretations should:
  - (a) not contain negation
  - (b) preserve the meaning

	% 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).

# Conclusion

# Conclusion

- The idea is simple yet effective:

# Conclusion

- The idea is simple yet effective:
  - complement inputs that contain negation with a paraphrase that does not contain negation

# Conclusion

- The idea is simple yet effective:
  - complement inputs that contain negation with a paraphrase that does not contain negation
- Automatically obtained (noisy) affirmative interpretations yield improvements with:



# Conclusion

- The idea is simple yet effective:
  - complement inputs that contain negation with a paraphrase that does not contain negation
- Automatically obtained (noisy) affirmative interpretations yield improvements with:
  - (a) CondaQA compared with a model with twice as many parameters pre-trained with  $\approx 1\text{M}$  question-answer pairs from 20 existing corpora and

# Conclusion

- The idea is simple yet effective:
  - complement inputs that contain negation with a paraphrase that does not contain negation
- Automatically obtained (noisy) affirmative interpretations yield improvements with:
  - (a) CondaQA compared with a model with twice as many parameters pre-trained with  $\approx 1M$  question-answer pairs from 20 existing corpora and
  - (b) five NLU tasks.

# Conclusion

- The idea is simple yet effective:
  - complement inputs that contain negation with a paraphrase that does not contain negation.
- Automatically obtained (noisy) affirmative interpretations yield improvements with:
  - (a) CondaQA compared with a model with twice as many parameters pre-trained with  $\approx 1$ M question-answer pairs from 20 existing corpora and
  - (b) five NLU tasks.
- The methodology is architecture- and task-agnostic.