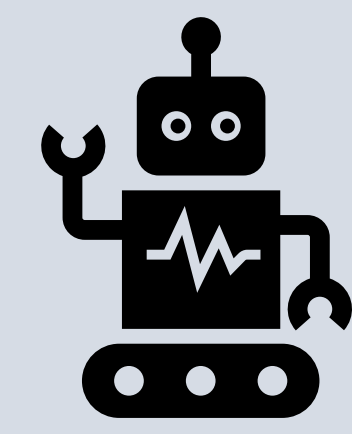


If you were able to find accounting records from the Middle East that slightly predated Stevin's publication, would you likely see decimals being used for transactions, even if they were in an Arabic script?



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



No

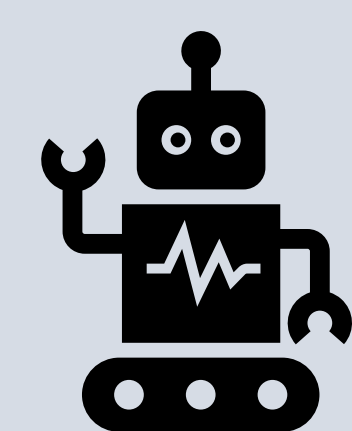


+ **Affirmative Interpretation**

(paraphrase without negation)



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



Yes



## Generating Affirmative Interpretation



- T5-HB** - Off-the-shelf generator
- Trained on 153k (sent w/ neg, affirmative interpretation) pairs
  - Pairs created using back-translation
  - Affirmative Interpretations:  $A_{HB}$



- T5-CG** - 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
  - Simple Paraphrases:  $S_{CG}$
  - Affirmative interpretations:  $A_{CG}$

## Experimental Results (CondaQA)

| Model                     | Input               | Acc.        |
|---------------------------|---------------------|-------------|
| UnifiedQA-V2-Large (770M) | P+Q                 | <b>66.7</b> |
| RoBERTa-Large (335M)      | P+Q                 | <b>64.9</b> |
|                           | P+Q+S <sub>CG</sub> | <b>65.7</b> |
|                           | P+Q+A <sub>CG</sub> | <b>66.4</b> |
|                           | P+Q+A <sub>HB</sub> | <b>67.1</b> |

**Table 1:** Results on CondaQA test set (7240 instances), the largest question-answering corpus that requires reasoning over negation. P and Q stand for passage and question. *Incorporating affirmative interpretations is useful and outperforms the UnifiedQA model with twice parameters and trained on ~1M pairs of question answers. Affirmative interpretations are better than simple paraphrases.*

## Qualitative Analysis

Affirmative interpretations should:

- (a) Not contain negation (b) Preserve the meaning

Example:

**Negated Sentence**

It is *not rare* to find pearls that measure as much as 14mm across.

**Correct Affirmative Interpretation**

It is *common* to find pearls that measure as much as 14mm across.

Manual analysis:

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

## Experimental Results (NLU Tasks)

|  | CmmnsnsQA | QNLI  | WiC   | WSC   |
|--|-----------|-------|-------|-------|
| RoBERTa with affirmative interpretations obtained using T5-HB ( $A_{HB}$ ) |           |       |       |       |
| All  | +2.9%     | +1.1% | -1.4% | -1.4% |
| w/ neg.  | +1.4%     | +0.0% | +6.1% | +4.2% |
| w/o neg.   | +4.3%     | +1.1% | +0.0% | -7.5% |
| RoBERTa with affirmative interpretations obtained using T5-CG ( $A_{CG}$ ) |           |       |       |       |
| All  | +1.4%     | +1.1% | +6.1% | +4.2% |
| w/ neg.  | +1.4%     | +0.0% | +2.8% | +1.5% |
| w/o neg.   | +2.9%     | +1.1% | +2.8% | +2.9% |

**Table 2:** Results on additional NLU tasks. We present improvements on macro F1 over RoBERTa-Large trained on the original input without coupling affirmative interpretations. Results on STS-B are available in the paper.