Unsupervised Commonsense Question Answering with Self-Talk

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula and Yejin Choi
EMNLP 2020
Unsupervised Commonsense Question Answering with Self-Talk

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula and Yejin Choi
EMNLP 2020
Unsupervised Commonsense Question Answering with Self-Talk

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula and Yejin Choi
EMNLP 2020
Unsupervised Commonsense Question Answering with Self-Talk

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula and Yejin Choi
EMNLP 2020
Commonsense Question Answering
COPA: Choice of Plausible Alternatives

Context: The man broke his toe.
Question: What was the cause?
Choices:
1) He got a hole in his sock.
2) He dropped a hammer on his foot.
### COPA: Choice of Plausible Alternatives

**Context:** The man broke his toe.

**Question:** What was the cause?

**Choices:**
1) He got a hole in his sock.
2) He dropped a hammer on his foot.

### CommonsenseQA

**Question:** Where on a river can you hold a cup upright to catch water on a sunny day?

**Choices:**
1) waterfall, 2) bridge, 3) valley, 4) pebble, 5) mountain
## COPA: Choice of Plausible Alternatives

**Context:** The man broke his toe.

**Question:** What was the cause?

**Choices:**
1) He got a hole in his sock.
2) **He dropped a hammer on his foot.**

## CommonsenseQA

**Question:** Where on a river can you hold a cup upright to catch water on a sunny day?

**Choices:**
1) waterfall, 2) bridge, 3) valley, 4) pebble, 5) mountain

## WinoGrande

**Context:** Katrina had the financial means to afford a new car while Monica did not, since ____ had a high paying job.

**Choices:** 1) Katrina 2) Monica
<table>
<thead>
<tr>
<th>Commonsense Question Answering</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>COPA: Choice of Plausible Alternatives</strong></td>
</tr>
</tbody>
</table>
| **Context:** The man broke his toe.  
**Question:** What was the cause?  
**Choices:**  
1) He got a hole in his sock.  
2) He dropped a hammer on his foot. |
| **CommonsenseQA** |
| **Question:** Where on a river can you hold a cup upright to catch water on a sunny day?  
**Choices:**  
1) waterfall, 2) bridge, 3) valley, 4) pebble, 5) mountain |
| **WinoGrande** |
| **Context:** Katrina had the financial means to afford a new car while Monica did not, since ____ had a high paying job.  
**Choices:** 1) Katrina, 2) Monica |
| **MC-TACO:**  
**Multiple Choice Temporal Commonsense** |
| **Context:** [...] dream of becoming a judge.  
**Question:** How many years did it take for Mark to become a judge?  
**Choices:** 1) 63 years, 2) 7 weeks, 3) 7 years, 4) 7 seconds, 5) 7 hours |
<table>
<thead>
<tr>
<th>Commonsense Question Answering</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>COPA: Choice of Plausible Alternatives</strong></td>
</tr>
</tbody>
</table>
| **Context:** The man broke his toe.  
**Question:** What was the cause?  
**Choices:**  
1) He got a hole in his sock.  
2) He dropped a hammer on his foot. |
| **CommonsenseQA** |
| **Question:** Where on a river can you hold a cup upright to catch water on a sunny day?  
**Choices:**  
1) waterfall, 2) bridge, 3) valley, 4) pebble, 5) mountain |
| **WinoGrande** |
| **Context:** Katrina had the financial means to afford a new car while Monica did not, since ____ had a high paying job.  
**Choices:** 1) Katrina 2) Monica |
| **MC-TACO: Multiple Choice Temporal Commonsense** |
| **Context:** [...] dream of becoming a judge.  
**Question:** How many years did it take for Mark to become a judge?  
**Choices:** 1) 63 years 2) 7 weeks 3) 7 years 4) 7 seconds 5) 7 hours |
| **SocialIQa: Social Interaction QA** |
| **Context:** In the school play, Robin played a hero in the struggle [...] angry villain.  
**Question:** How would others feel as a result?  
**Choices:** 1) sorry for the villain. 2) hopeful that Robin will succeed. 3) like Robin should lose the fight. |
## Commonsense Question Answering

### COPA: Choice of Plausible Alternatives

**Context:** The man broke his toe.

**Question:** What was the cause?

**Choices:**
1) He got a hole in his sock.
2) He dropped a hammer on his foot.

### CommonsenseQA

**Question:** Where on a river can you hold a cup upright to catch water on a sunny day?

**Choices:**
1) waterfall, 2) bridge, 3) valley, 4) pebble, 5) mountain

### WinoGrande

**Context:** Katrina had the financial means to afford a new car while Monica did not, since ____ had a high paying job.

**Choices:** 1) Katrina, 2) Monica

### MC-TACO: Multiple Choice Temporal Commonsense

**Context:** [...] dream of becoming a judge.

**Question:** How many years did it take for Mark to become a judge?

**Choices:**
1) 63 years, 2) 7 weeks, 3) 7 years, 4) 7 seconds, 5) 7 hours

### SocialIQa: Social Interaction QA

**Context:** In the school play, Robin played a hero in the struggle [...], an angry villain.

**Question:** How would others feel as a result?

**Choices:**
1) sorry for the villain, 2) hopeful that Robin will succeed, 3) like Robin should lose the fight

### PIQA: Physical Interaction QA

**Question:** To separate egg whites from the yolk using a water bottle, you should

**Choices:**
1) [...] Release, which creates suction and lifts the yolk.
2) [...] Keep pushing, which creates suction and lifts the yolk.
Unsupervised Multiple Choice Question Answering
Unsupervised Multiple Choice Question Answering

$P_L M (\text{The answer is answer}_1)$

$P_L M (\text{The answer is answer}_2)$

$\ldots$

$P_L M (\text{The answer is answer}_k)$
Unsupervised Multiple Choice Question Answering

\[
\text{Predict most “plausible” answer choice}
\]

\[
\begin{align*}
\text{P}_{\text{LM}}(\text{The answer is } & \text{answer}_1) \\
\text{P}_{\text{LM}}(\text{The answer is } & \text{answer}_2) \\
\vdots \\
\text{P}_{\text{LM}}(\text{The answer is } & \text{answer}_k)
\end{align*}
\]
Unsupervised Multiple Choice Question Answering

\[ P_{LM}(\text{The answer is } \text{answer}_1) \]
\[ P_{LM}(\text{The answer is } \text{answer}_2) \]
\[ \ldots \]
\[ P_{LM}(\text{The answer is } \text{answer}_k) \]

\{ Predict most “plausible” answer choice \}

Because Brett found an internship while in college but Ian was unable to, **Brett** found a job less quickly after graduation.

Because Brett found an internship while in college but Ian was unable to, **Ian** found a job less quickly after graduation.
Why unsupervised?

1. Supervised models often learn to solve a dataset, rather than the task.
Why unsupervised?

1. Supervised models often learn to solve a *dataset*, rather than the *task*.
2. Making a good use of large and expensive-to-train pre-trained language models.
Why unsupervised?

1. Supervised models often learn to solve a dataset, rather than the task.
2. Making a good use of large and expensive-to-train pre-trained language models.
Why unsupervised?

1. Supervised models often learn to solve a *dataset*, rather than the *task*.
2. Making a good use of large and expensive-to-train pre-trained language models.
Can we use pre-trained language models to generate otherwise implicit knowledge?*
Can we use pre-trained language models to generate otherwise implicit knowledge?*

* Acknowledging limited coverage (Gordon and Van Durme, 2015) and limited precision (e.g. Kassner and Suchutze, 2019).
Self-Talk
Self-Talk
Self-Talk
Self-Talk

instance $\text{LM}_1$ clarification question clarification
Self-Talk
Self-Talk

instance → clarification question → clarification → answer

LM₁ → LM₂
1. Generating Clarifications

instance \[\text{LM}_1\] clarification question clarification
Example: WinoGrande
Example: WinoGrande

Question Generation:
Because Brett found an internship while in college but Ian was unable to, ___ found a job less quickly after graduation.
Example: WinoGrande

Question Generation:

Because Brett found an internship while in college but Ian was unable to, (Ian) found a job less quickly after graduation.
Example: WinoGrande

Question Generation:

Because Brett found an internship while in college but Ian was unable to, (Ian) found a job less quickly after graduation.

What is the purpose of

Question & Answer Prefixes

What is the purpose of
The purpose of ____ is
Example: WinoGrande

Because Brett found an internship while in college but Ian was unable to, (Ian) found a job less quickly after graduation.

Question Generation:

What is the purpose of the internship?

What is the purpose of ___ is

Question & Answer Prefixes
Example: WinoGrande

Question Generation:
Because Brett found an internship while in college but Ian was unable to, \textit{(Ian)} found a job less quickly after graduation.

What is the purpose of

Answer Generation:
Because Brett found an internship while in college but Ian was unable to, \textit{___} found a job less quickly after graduation.

What is the purpose of \textit{the internship}?
Because Brett found an internship while in college but Ian was unable to, (Ian) found a job less quickly after graduation.

What is the purpose of the internship?
Because Brett found an internship while in college but Ian was unable to, (Ian) found a job less quickly after graduation.

What is the purpose of the internship?

The purpose of the internship is to help people find jobs.
## Question & Answer Prefixes

### WinoGrande

<table>
<thead>
<tr>
<th>Question Prefix</th>
<th>Answer Prefix</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is the definition of</td>
<td>The definition of ___ is</td>
</tr>
<tr>
<td>What are the properties of a ___</td>
<td>The properties of a ___ are</td>
</tr>
<tr>
<td>What is ___</td>
<td>___ is</td>
</tr>
<tr>
<td>What is the purpose of ___</td>
<td>The purpose of ___ is</td>
</tr>
<tr>
<td>What is the main function of ___</td>
<td>The main function of ___ is</td>
</tr>
<tr>
<td>What does mean to ___</td>
<td>___ Means</td>
</tr>
</tbody>
</table>
## Question & Answer Prefixes

### WinoGrande

<table>
<thead>
<tr>
<th>Question Prefix</th>
<th>Answer Prefix</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is the definition of</td>
<td>The definition of __ is</td>
</tr>
<tr>
<td>What are the properties of a</td>
<td>The properties of a __ are</td>
</tr>
<tr>
<td>What is</td>
<td>A __ is</td>
</tr>
<tr>
<td>What is the purpose of</td>
<td>The purpose of __ is</td>
</tr>
<tr>
<td>What is the main function of</td>
<td>The main function of __ is</td>
</tr>
<tr>
<td>What does mean to</td>
<td>__ Means</td>
</tr>
</tbody>
</table>

### PIQA

<table>
<thead>
<tr>
<th>Question Prefix</th>
<th>Answer Prefix</th>
</tr>
</thead>
<tbody>
<tr>
<td>How to</td>
<td>The way to do __ is</td>
</tr>
<tr>
<td>How do you</td>
<td>The way you do __ is</td>
</tr>
<tr>
<td>How can one</td>
<td>One can __ by</td>
</tr>
<tr>
<td>What can be used for __</td>
<td>__ can be used for __</td>
</tr>
<tr>
<td>What can one do in order to</td>
<td>In order to __, one can</td>
</tr>
<tr>
<td>What should you use for</td>
<td>For __, you should use</td>
</tr>
<tr>
<td>What is the definition of</td>
<td>The definition of __ is</td>
</tr>
<tr>
<td>What are the properties of a</td>
<td>The properties of a __ are</td>
</tr>
<tr>
<td>What is</td>
<td>A __ is</td>
</tr>
</tbody>
</table>
2. Predicting the Correct Answer Choice

instance + clarification \[\rightarrow\] \[\rightarrow\] answer

\[\text{LM}_2\]
Zero-shot Model with Self-Talk

Because Brett found an internship while in college but Ian was unable to, **Brett** found a job less quickly after graduation. The purpose of the internship is to help people find jobs.

Because Brett found an internship while in college but Ian was unable to, **Ian** found a job less quickly after graduation. The purpose of the internship is to help people find jobs.

Because Brett found an internship while in college but Ian was unable to, **Brett** found a job less quickly after graduation. The definition of “job” is to be employed by someone.

Because Brett found an internship while in college but Ian was unable to, **Ian** found a job less quickly after graduation. The definition of “job” is to be employed by someone.
Baselines
Because Brett found an internship while in college but Ian was unable to, **Brett** found a job less quickly after graduation.

Because Brett found an internship while in college but Ian was unable to, **Ian** found a job less quickly after graduation.
Knowledge-informed Baselines

Example: Social IQa

Taylor was doing her job so she put the money in the drawer. What will Taylor do next?
Knowledge-informed Baselines

Example: Social IQa

Taylor was doing her **job** so she put the **money** in the drawer.

What will Taylor do next?
Knowledge-informed Baselines

Example: Social IQa

Taylor was doing her job so she put the money in the drawer.

What will Taylor do next?
Taylor was doing her job so she put the money in the drawer.

What will Taylor do next?

Job is a type of work. You would work because you want money.
Taylor was doing her job so she put the money in the drawer.

What will Taylor do next?

Job is a type of work. You would work because you want money.
Taylor was doing her job so she put the money in the drawer.

What will Taylor do next?

Job is a type of work. You would work because you want money.

Job to earn money.
Knowledge-informed Baselines
Example: Social IQa

Taylor was doing her job so she put the money in the drawer.

What will Taylor do next?

Job is a type of work. You would work because you want money.
Job to earn money.
Taylor was doing her job so she put the money in the drawer.

Job is a type of work. You would work because you want money. Job to earn money. As a result, Taylor wants to keep the money in the drawer.
Results
Results
Accuracy on the validation set
Results

Accuracy on the validation set

COPA  MC-TACO  Social IQa  PIQA  Winogrande  CommonsenseQA

Majority  Baseline  Knowledge-informed  Self-talk  Best supervised  Human
Results

Accuracy on the validation set

<table>
<thead>
<tr>
<th>Test Set</th>
<th>Majority</th>
<th>Baseline</th>
<th>Knowledge-informed</th>
<th>Self-talk</th>
<th>Best supervised</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>COPA</td>
<td>55</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC-TACO</td>
<td>40.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social IQa</td>
<td>33.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PIQA</td>
<td>50.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Winogradne</td>
<td>50.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CommonsenseQA</td>
<td>20.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Results

Accuracy on the validation set

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Majority</th>
<th>Baseline</th>
<th>Knowledge-informed</th>
<th>Self-talk</th>
<th>Best supervised</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>COPA</td>
<td>55</td>
<td>53</td>
<td>53</td>
<td>53</td>
<td>55</td>
<td>37.2</td>
</tr>
<tr>
<td>MC-TACO</td>
<td>40.3</td>
<td>53.1</td>
<td>41.1</td>
<td>40.3</td>
<td>53.1</td>
<td>37.2</td>
</tr>
<tr>
<td>Social IQa</td>
<td>33.6</td>
<td>41.1</td>
<td>50.5</td>
<td>41.1</td>
<td>50.5</td>
<td>41.1</td>
</tr>
<tr>
<td>PIQA</td>
<td>62.6</td>
<td>54.8</td>
<td>50.4</td>
<td>54.8</td>
<td>50.4</td>
<td>50.4</td>
</tr>
<tr>
<td>Winogrande</td>
<td>20.9</td>
<td>37.2</td>
<td>50.5</td>
<td>50.5</td>
<td>50.5</td>
<td>50.5</td>
</tr>
<tr>
<td>CommonsenseQA</td>
<td>33.6</td>
<td>40.3</td>
<td>55</td>
<td>55</td>
<td>55</td>
<td>55</td>
</tr>
</tbody>
</table>
# Results

## Accuracy on the validation set

<table>
<thead>
<tr>
<th></th>
<th>COPA</th>
<th>MC-TACO</th>
<th>Social IQa</th>
<th>PIQA</th>
<th>Winogrande</th>
<th>CommonsenseQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority</td>
<td>55</td>
<td>40.3</td>
<td>33.6</td>
<td>50.5</td>
<td>50.4</td>
<td>20.9</td>
</tr>
<tr>
<td>Baseline</td>
<td>53</td>
<td>53.1</td>
<td>41.1</td>
<td>62.6</td>
<td>54.8</td>
<td>37.2</td>
</tr>
<tr>
<td>Knowledge-informed</td>
<td>69</td>
<td>58.8</td>
<td>47.5</td>
<td>69.6</td>
<td>55.4</td>
<td>39.7</td>
</tr>
<tr>
<td>Self-talk</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Best supervised</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Accuracy values are shown for each dataset, with the best performance highlighted in green.
### Results

**Accuracy on the validation set**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Majority</th>
<th>Baseline</th>
<th>Knowledge-informed</th>
<th>Self-talk</th>
<th>Best supervised</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>COPA</td>
<td>55</td>
<td>53</td>
<td>69</td>
<td>66</td>
<td>58.8</td>
<td>59.9</td>
</tr>
<tr>
<td>MC-TACO</td>
<td>40.3</td>
<td>53.1</td>
<td>58.8</td>
<td>59.9</td>
<td>56.2</td>
<td>59.9</td>
</tr>
<tr>
<td>Social IQa</td>
<td>33.6</td>
<td>47.5</td>
<td>46.2</td>
<td>50.5</td>
<td>60.6</td>
<td>69.6</td>
</tr>
<tr>
<td>PIQA</td>
<td>50.5</td>
<td>62.6</td>
<td>69.6</td>
<td>70.2</td>
<td>54.8</td>
<td>55.4</td>
</tr>
<tr>
<td>Winogrande</td>
<td>50.4</td>
<td>54.8</td>
<td>55.4</td>
<td>54.7</td>
<td>62.6</td>
<td>69.6</td>
</tr>
<tr>
<td>CommonsenseQA</td>
<td>20.9</td>
<td>37.2</td>
<td>39.7</td>
<td>32.4</td>
<td>50.4</td>
<td>50.5</td>
</tr>
</tbody>
</table>
## Results

### Accuracy on the validation set

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Majority</th>
<th>Baseline</th>
<th>Knowledge-informed</th>
<th>Self-talk</th>
<th>Best supervised</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>COPA</td>
<td>55</td>
<td>53</td>
<td>40.3</td>
<td>33.6</td>
<td>46.2</td>
<td>37.2</td>
</tr>
<tr>
<td>MC-TACO</td>
<td>69</td>
<td>66</td>
<td>58.8</td>
<td>47.5</td>
<td>54.7</td>
<td>55.4</td>
</tr>
<tr>
<td>Social IQa</td>
<td>66</td>
<td>59.9</td>
<td>59.9</td>
<td>46.2</td>
<td>54.7</td>
<td>55.4</td>
</tr>
<tr>
<td>PIQA</td>
<td>76.6</td>
<td>79.2</td>
<td>79.2</td>
<td>70.2</td>
<td>54.7</td>
<td>54.7</td>
</tr>
<tr>
<td>Winogrande</td>
<td>79.3</td>
<td>76.6</td>
<td>76.6</td>
<td>76.6</td>
<td>54.7</td>
<td>54.7</td>
</tr>
<tr>
<td>CommonsenseQA</td>
<td>83.7</td>
<td>79.3</td>
<td>83.7</td>
<td>79.3</td>
<td>54.7</td>
<td>54.7</td>
</tr>
</tbody>
</table>
Results

Accuracy on the validation set

<table>
<thead>
<tr>
<th>Test Set</th>
<th>Majority</th>
<th>Baseline</th>
<th>Knowledge-informed</th>
<th>Self-talk</th>
<th>Best supervised</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>COPA</td>
<td>55</td>
<td>53</td>
<td>55</td>
<td>55</td>
<td>55</td>
<td>55</td>
</tr>
<tr>
<td>MC-TACO</td>
<td>40.3</td>
<td>53.1</td>
<td>40.3</td>
<td>40.3</td>
<td>40.3</td>
<td>40.3</td>
</tr>
<tr>
<td>Social IQa</td>
<td>33.6</td>
<td>41.1</td>
<td>33.6</td>
<td>33.6</td>
<td>33.6</td>
<td>33.6</td>
</tr>
<tr>
<td>PIQA</td>
<td>50.5</td>
<td>50.5</td>
<td>50.5</td>
<td>50.5</td>
<td>50.5</td>
<td>50.5</td>
</tr>
<tr>
<td>Winogrande</td>
<td>20.9</td>
<td>37.2</td>
<td>20.9</td>
<td>20.9</td>
<td>20.9</td>
<td>20.9</td>
</tr>
<tr>
<td>CommonsenseQA</td>
<td>83.7</td>
<td>32.4</td>
<td>83.7</td>
<td>83.7</td>
<td>83.7</td>
<td>83.7</td>
</tr>
</tbody>
</table>
What do people think of useful clarifications?

Passage: Kendall was being careless and accidentally set fire to Sydney. Sydney had to go to the hospital because of the burns.

Main question: How would Kendall feel afterwards?

Evaluate the following clarification question asked about the passage/main question and its answer.

Question: What is the relationship between 'careless' and 'reckless'?
- □ The question is completely gibberish, I can't understand it at all.
- □ The question is not perfectly grammatical, but I think I can understand it.
- ○ The question is grammatical.
- □ The question is on topic with respect to the given passage.

Answer: Careless is similar to reckless
- □ The answer is completely gibberish, I can't understand it at all.
- □ The answer is not perfectly grammatical, but I think I can understand it.
- ○ The answer is grammatical.
- □ The answer is on topic with respect to the given passage and/or main question.
- □ The answer is factually correct or likely true.
- □ The answer provides useful information for answering the main question.

* Clarifications that change baseline prediction from incorrect to correct
What do people think of **useful clarifications**?

Most clarifications are grammatical or at least understandable:

- **Grammatical**: 87.2%
- **Understandable**: 10.1%
- **Gibberish**: 2.7%
What do people think of useful clarifications?

Most clarifications are grammatical or at least understandable:
What do people think of useful clarifications?

Clarifications judged as often relevant and factually correct but less frequently as helpful
What are the errors in Harmful clarifications?

* Clarifications that change baseline prediction from **correct** to **incorrect**
What are the errors in Harmful clarifications?

The children were not vaccinated, which was fine with Betty but annoyed Mary. _____ believed they made kids autistic.

What does it mean to be “autistic”?

Be “autistic” means to have problems in social interaction and communication skills.

* Clarifications that change baseline prediction from correct to incorrect
Takeaways
Takeaways

✓ Generating knowledge with LMs improves upon the baseline and performs similarly to knowledge-informed models.
Takeaways

- Generating knowledge with LMs improves upon the baseline and performs similarly to knowledge-informed models.
- Generated clarifications don’t align with what humans consider helpful → different “reasoning process”?
Takeaways

- Generating knowledge with LMs improves upon the baseline and performs similarly to knowledge-informed models.
- Generated clarifications don’t align with what humans consider helpful → different “reasoning process”?

Future directions:
- Multiple hops
- Introspection
Takeaways

- Generating knowledge with LMs improves upon the baseline and performs similarly to knowledge-informed models.
- Generated clarifications don’t align with what humans consider helpful → different “reasoning process”?

Future directions:
- Multiple hops
- Introspection

Thank you!

@VeredShwartz vereds@allenai.org
References


