Appendix I: More details on mechanical turk interface for the various steps of the question and answer-options creation process
Instructions

You need to read a passage and construct questions about the information presented in the passage. For each question, you also need to provide the correct answer. The question should be answerable from the passage. Most importantly, only those questions are allowed whose answer cannot be determined from a single sentence.

Below are a few example passages and sample questions (and answers).

**Paragraph:**
Sent 1: A Republican bid to keep the government open past Friday includes no new money for the construction of a border wall along the U.S.-Mexico border, according to several congressional aides familiar with ongoing talks.
Sent 2: GOP leaders submitted the new offer Tuesday afternoon in an effort to appease Democrats, whose votes are needed to avert a shutdown of federal agencies, several House and Senate aides said.
Sent 3: However, Republicans also insisted on increases in border security and defense spending, including money to repair existing fencing and new surveillance technology to patrol the nearly 2,000-mile border.
Sent 4: Democrats have indicated that they would support such appropriations so long as no money goes toward an actual wall.

Here are a few example questions with our criteria:

**Question:** Who has to be appeased to keep the government open?
**Correct Answer:** "Democrats"
**Sentences needed to answer this question:**
1 2 3 4
**Explanation:** This is a multi-sentence question because, "appease Democrats" is mentioned in Sent 2, while "keep the government open" is part of Sent 1.

**Question:** When was the republican bid submitted?
**Correct Answer:** "Tuesday afternoon"
**Sentences needed to answer this question:**
1 2 3 4
**Explanation:** Republican bid is mentioned in Sent 1 while Tuesday afternoon occurs in Sent 2.

**Question:** Who does not want a border wall?
**Correct Answer:** "Democrats"
**Sentences needed to answer this question:**
1 2 3 4
**Explanation:** Democrats will be appeased by the new offer (Sent 2). The new offer includes no money for the wall (Sent 1). This implies that Democrats don't want the wall.

Here is another example paragraph:

**Paragraph:**
Sent 1: Obama was born on August 4, 1961, at Kapiolani Maternity & Gynecological Hospital in Honolulu, Hawaii.
Sent 2: He is the only President to have been born in Hawaii.
Sent 3: He was born to a white mother and a black father.
Sent 4: His mother, Ann Dunham (1942-1995), was born in Wichita, Kansas, of mostly English descent, with some German, Irish, Scottish, Swiss, and Welsh ancestry.

**Question:** How old was Obama's mother when he was born?
**Correct Answer:** "19"
**Sentences needed to answer this question:**
1 2 3 4
**Explanation:** Obama was born in 1961. His mother was born in 1942. This question reasoning about time across Sentences 1 and 4 (1961 - 1942 = 19).

**Question:** Where was Ann living in August, 1961?
**Correct Answer:** "Hawaii"
**Sentences needed to answer this question:**
1 2 3 4
**Explanation:** Ann is Obama's mother (Sent 4). She must have been there when Obama was born (Sent 1). Obama was born in Hawaii (Sent 2).

**Question:** Was one of Ann's children a future president?
**Correct Answer:** "Yes"
**Sentences needed to answer this question:**
1 2 3 4
**Explanation:** Ann is Obama's mother (Sent 4). Obama was a future president (Sent 2).

**Example of a bad question:**

**Paragraph:**
Sent 1: Saffi succumbs to Edwina’s charms and falls hopelessly in love with her.
Sent 2: Lord Eustach becomes aware of this but Saffi saves him from a man-eating tiger during a safari.

**Question:** What is the name of the man who succumbs to Edwina’s charms and saves her husband from a man-eating tiger during a safari?
**Correct Answer:** "Saffi"
**Explanation:** This is a BAD question because even though it is constructed using phrases from more than one sentence ("succumbs to Edwina’s charms" and "a man-eating tiger during a safari") but the answer can be guessed by reading ONLY ONE of the two sentences.

- For the paragraph below, create at least 3 questions that require multiple sentences (examples above).
- If you create more than 3 questions, you'd get bonus rewards.
- For each generated questions you write the answer in the space followed by the question, as well as denoting what sentences are required for answering it.
- Items with * are required.
Figure 6: Instructions and Interface for the Step 1 of the dataset creation process. In this step, we request taskers to write a question and its correct answer for a given paragraph. We emphasize the requirement for multi-sentence questions, and additionally ask them to indicate which sentence ids would be needed for answering the composed question. We also provide two example paragraphs with 3 multi-sentence questions each and the corresponding answers. We also provide an explanation indicating why these questions fit our criteria. Lastly, we provide an example of bad question before asking them to compose questions. We ask them to write 3-5 questions per paragraph but the figure only shows the form for one question.
Figure 7: Instructions and Interface for the Step 2 of the dataset creation process. This step was included to ensure that the questions collected from the previous step are indeed multi-sentence. In this step we show a sentence and a question to the turkers and ask them if they can answer the question using only the information present in the shown sentence. In the instructions, we discourage them from using background knowledge since some of our paragraphs are from non-fictional sources. We also provide examples along with an explanation. The bottom part of the figure shows the form for one sentence-question pair.
Instructions

Writing Answer Options for a Reading Comprehension Task

You are designing a reading comprehension test for your students. Students will read a paragraph, a question based on the paragraph; and a few options for its answer. They have to choose the correct answer from the options. For the test to make sense, you want the correct answer to be part of the options. However, you don’t want the test to be too easy! To make the test difficult for the students, you want to add incorrect options that might be confused with the correct answer to the list of options. A student who tries to answer a question without reading the paragraph carefully might choose the incorrect option! In this task, your goal is to write the list of options that will be shown to your students.

The questions showed to you in this HIT were written by people and might sometimes contain typing mistakes or other kinds of errors. In such cases, you could additionally check the box underneath the question indicating the “question is bad,” and also correct it if possible. The answer-options you write would be in response to the corrected questions.

Below we show some good and bad examples of correct and incorrect options. We also show examples of bad questions.

Correct Options: A correct option is one that correctly and completely answers the question. Keep in mind that for any question, there can be multiple correct answers. A bad correct option addresses the question only partially or incorrectly. We want to include only good correct options in the list of options. Here are some good and bad examples of correct options:

Paragraph:
Sent 1: Obama was born on August 4, 1961, at Kapiolani Maternity & Gynecological Hospital in Honolulu, Hawaii.
Sent 2: He is the only President to have been born in Hawaii.
Sent 3: He was born to a white mother and a black father.
Sent 4: His mother, Ann Dunham (1942-1995), was born in Wichita, Kansas, of mostly English descent, with some German, Irish, Scottish, Swiss, and Welsh ancestry.

Question: How old was Obama’s mother when he was born?

Correct options (separated via “/”): 19, almost twenty
Bad options: 20
Explanation: Obama was born in 1942 and 1961–1942 = 19. The good options (19 and ‘almost twenty’) answer the question correctly.

Incorrect Options: An incorrect option will be closely related to the content of the paragraph and/or the question but will not truthfully answer the question. They should not be trivial, in the sense that they should not be easily eliminated as the answer without reading the paragraph. They are presented as options to confuse your students and to make the reading comprehension task challenging for them. Here are some good and bad examples of incorrect options:

Paragraph:
Sent 1: It was hot that day.
Sent 2: The temperature on the wall of the backyard was showing something well over 100 F.
Sent 3: Meanwhile Tom, at home, was trying to finish the remainder of carrots from last night, and packing for his trip to Chicago tomorrow.
Sent 4: As employees of Art Museum, Tom and his older cousin often had to travel to Chicago.

Question: What was the temperature outside, when Tom was eating carrots?

Correct options (separated via “/”): Not very hot // Far below 100 F
Bad options: Hot
Explanation: The good options appear in the paragraph (at least partially). They are also related to the content of the question (temperature), but do not answer the question correctly. Your student will have to carefully read the paragraph to identify that these don’t answer the question correctly. The bad options appear in the paragraph (e.g. ‘Hot’) and also sometimes in the question (e.g. carrots) but they are trivially incorrect. A student can easily determine that they don’t answer the question correctly without even reading the paragraph.

Incorrect Options: A bad question could have grammatical errors, typing mistakes, etc. or might not make sense in the context of the paragraph (for instance, it might not be related to the content of the paragraph or not be answerable at all). In such cases, please mark the question as a bad question. If the question contains minor errors please correct it in the boxed space provided underneath the question, otherwise leave the boxed space empty. Here are some examples of bad questions.

Paragraph:
Sent 1: It was hot that day.
Sent 2: The temperature on the wall of the backyard was showing something well over 100 F.
Sent 3: Meanwhile Tom, at home, was trying to finish the remainder of carrots from last night, and packing for his trip to Chicago tomorrow.
Sent 4: As employees of the Art Museum, Tom and his older cousin often had to travel to Chicago.

Question: What was the temperature outside, when Pem was eating carrots?

Corrected question: If bad: What was the temperature outside, when Tom was eating carrots?
Explanation: It is easy to identify that there is a typing mistake, and Pem should be Tom.

Question: When did Tom walk out of the party?

Corrected question: If bad: What did Tom walk out of the party?
Explanation: The question doesn’t make sense for this paragraph. The paragraph does not discuss Tom attending any party. So, this is a bad question and cannot be corrected.

For each of the following questions, write good correct and incorrect options that would be shown to your students. Additionally, if the question is bad please indicate that and correct it if possible.
Figure 8: Instructions and Interface for the Step 3 of the dataset creation process. This step was included to collect correct and incorrect answer options for the questions in our dataset. In the instructions we define what we mean by correct and incorrect answer options—correct options answer the given question correctly and completely, whereas incorrect options do not answer the question but are non-trivial and cannot be easily eliminated as an incorrect answer without reading the passage. We also show examples of good and bad inputs for both along with explanations. The bottom part of the figure shows the form for entering correct and incorrect answer-options for one paragraph-question pair.
Instructions

Answering a Reading Comprehension Task

You are answering a reading comprehension test. You will read a paragraph and some questions, and a few options for its answer. You will have to choose the correct answer(s) from the options.

Note that sometimes, more than one option can answer the question correctly and completely. In those cases, please select all such options.

Here is an example:

**Paragraph:**

Obama was born on August 4, 1961, at Kapi’olani Maternity & Gynecological Hospital in Honolulu, Hawaii.

He is the only President to have been born in Hawaii. He was born to a white mother and a black father. His mother, Ann Dunham (1942-1995), was born in Wichita, Kansas, of mostly English descent, with some German, Irish, Scottish, Swiss, and Welsh ancestry.

**Question:** How old was Obama’s mother when he was born?

- teenager
- in his 40s
- mid 20s
- almost twenty

Explanation: Obama was born in 1961. His mother was born in 1942 and 1961-1942=19. The good option (almost twenty) answers the question correctly. The bad answers give an incorrect answer.

**Question:** Where was Ann living in August, 1961?

- Kapi’olani Maternity
- Honolulu, Hawaii
- Wichita
- England
- Honolulu

Explanation: Ann is Obama’s mother (Sent 4). She must have been there when Obama was born (Sent 1). Obama was born in Hawaii (Sent 2). Hawaii is also in Honolulu. So both Hawaii and Honolulu answer the question correctly. Kansas and Scottland don’t answer the question correctly.

For the following paragraph, select all the correct answer(s):

**Paragraph:**

CNN -- A Christian evangelical group said Thursday that a Bible school -- backed by American evangelist Franklin Graham -- was destroyed in the latest bombing raid to hit South Kordofan, an oil-rich Sudanese province that borders the newly created independent country of South Sudan.

Sent 1: At least eight bombs were dropped in the area Wednesday during the school’s first day of classes, according to a statement by Samaritan’s Purse, Graham’s Christian humanitarian group, which supports the school.

Sent 2: Two bombs landed inside the compound located in the region’s Nuba Mountains -- destroying two Heiban Bible College buildings and igniting grass fires across the area, the group said in a statement No injuries were reported.

Sent 4: “It was a miracle that no one was injured,” the statement added.

Sent 5: Graham, who has called on the international community to take out Sudan’s air assets and establish a no-fly zone in the region, said in a statement Thursday that he blamed Sudan’s air force for the strike.

Sent 6: At least four churches have been destroyed since August, the group said.

Sent 7: “We are deeply concerned for the welfare and lives of the people of South Kordofan and we condemn the bombing of churches and Christian facilities,” added Graham, son of the famed Rev Billy Graham.

Sent 8: More than 78,000 people have fled South Kordofan and Blue Nile states since August of last year after an armed rebellion took root, the United Nations reported.

Sent 9: The Sudanese government is thought to have responded to the rebellion by conducting sustained air raids with the use of Russian-made Antonov bombers, which have raised concerns over civilian casualties.

Sent 10: Decades of civil war between the north and south, costing as many as 2 million lives, formally ended with a U.S.-brokered peace treaty in 2005.

**Question:** What did the bombs destroy?

☐ school
☐ Command Center
☐ Various churches
☐ dhimmi
☐ people
☐ Arab’s personalities
☐ aircraft
☐ car

*Is this a bad question?* e.g. typo, or not making sense?

Figure 9: Instructions and Interface for the Step 4 of the dataset creation process. In this step we ask the annotators to read the questions and answer-options, and indicate if they find any errors in them. We also ask them to select all answer-options that correctly answer the question (illustrated through examples). Any possibly discrepancies or errors were later reviewed and corrected manually. The bottom part of the figure shows a sample form for one question.
### Appendix II: Examples from the corpus

| Sent 1: Dirk Diggler was born as Steven Samuel Adams on April 15, 1961 outside of Saint Paul, Minnesota.  |
| Sent 2: His parents were a construction worker and a boutique shop owner who attended church every Sunday and believed in God.  |
| Sent 3: Looking for a career as a male model, Diggler dropped out of school at age 16 and left home.  |
| Sent 4: He was discovered at a falafel stand by Jack Horner.  |
| Sent 5: Diggler met his friend, Reed Rothchild, through Horner in 1979 while working on a film.  |

**Question:** How old was Dirk when he met his friend Reed?

| A) 18 | B) 16 | C) 22 |

**Reasoning needed:** Temporal and mathematical reasoning

One needs to identify that Dirk was born in 1961 (Sent 1) and he met Reed in 1979 (Sent 5). At that time he must have been 18 (=1979-1961) years old.

| Sent 1: The hijackers attacked at 9:28.  |
| Sent 2: While traveling 35,000 feet above eastern Ohio, United 93 suddenly dropped 700 feet.  |
| Sent 3: Eleven seconds into the descent, the FAA's air traffic control center in Cleveland received the first of two radio transmissions from the aircraft.  |
| Sent 4: During the first broadcast, the captain or first officer could be heard declaring “Mayday” amid the sounds of a physical struggle in the cockpit.  |
| Sent 5: The second radio transmission, 35 seconds later, indicated that the fight was continuing.  |
| Sent 6: The captain or first officer could be heard shouting: “Hey get out of here-get out of here-get out of here.”  |
| Sent 7: On the morning of 9/11, there were only 37 passengers on United 93-33 in addition to the 4 hijackers.  |
| Sent 8: This was below the norm for Tuesday mornings during the summer of 2001.  |
| Sent 9: But there is no evidence that the hijackers manipulated passenger levels or purchased additional seats to facilitate their operation.  |
| Sent 10: The terrorists who hijacked three other commercial flights on 9/11 operated in five-man teams.  |
| Sent 11: They initiated their cockpit takeover within 30 minutes of takeoff.  |
| Sent 12: On Flight 93, however, the takeover took place 46 minutes after takeoff and there were only four hijackers.  |

**Question:** Which two factors were different between the three other hijacked planes and United 93?

| A) The amount of time that passed before the takeover started  |
| B) United 93 took longer and had less hijackers  |
| C) The airline operating the planes  |
| D) The weather and fuel used by the airplane  |
| E) The navigation system used by the planes  |

**Reasoning needed:** Discourse relation (contrast)

One needs to identify that the discourse marker *however* in Sent 12 indicates a contrast relation between Flight 93 and the flights mentioned in Sent 10. Also, *only* in Sent 12 indicates that the number of hijackers were fewer than in the contrasted other flights.

| Sent 10: The terrorists who hijacked three other commercial flights on 9/11 operated in five-man teams.  |
| Sent 11: They initiated their cockpit takeover within 30 minutes of takeoff.  |
| Sent 12: On Flight 93, however, the takeover took place 46 minutes after takeoff and there were only four hijackers.  |

**Question:** What was below average for this particular day?

| A) the number of passengers in the first class.  |
| B) the number of passengers on board.  |
| C) the number of hijackers  |
| D) the amount of air traffic in the skies  |
| E) the temperature  |

**Reasoning needed:** Event coreference

One needs to identify that *This* in Sent 8 co-refers to (event of) number of passengers in Sent 7. Note that Sent 12 contains *only four hijackers* and understanding that *only* indicates a smaller number of entities than expected (as in previous question), might mislead a system into believing that (C) is the correct answer.

---

Figure 10: Examples from our MultiRCCorpus. Each example shows relevant excerpts from a paragraph; multi-sentence question; and corresponding answer-options. The correct answer(s) is indicated by a *. Each example also contains a list of reasoning phenomenon needed to answer that question.
Sent 1: Suddenly, Amy screamed.
Sent 2: I whirled around and threw up my arm just in time to knock Dvorov away.
Sent 3: His eyes were glowing red, and he had bared a set of fangs that could probably take my hand off.

...  

Sent 8: On the floor was Dvorov’s head, separated from his body by several feet and a growing puddle of dark blood.
Sent 9: Nepthys stood over him with a blue sword in his hand.
Sent 10: It flashed in the stroke where it wasn’t streaked with Dvorov’s blood.

<table>
<thead>
<tr>
<th>Question: Who cut off Dvorov’s head?</th>
</tr>
</thead>
<tbody>
<tr>
<td>A) Nepthys</td>
</tr>
</tbody>
</table>

**Reasoning needed:** Implicit semantic role (agent) and causation
From sentences 8–10, indicate that Nepthys stood over Dvorov’s (dead) body with a sword which was streaked with Dvorov’s blood. Therefore, one can infer that the blood on the sword must have been caused by the action of killing of Dvorov and that Nepthys is the implicit agent of this action.

<table>
<thead>
<tr>
<th>Question: What was the weapon used to separate Dvorov’s head from his body?</th>
</tr>
</thead>
<tbody>
<tr>
<td>A) a hook</td>
</tr>
</tbody>
</table>

**Reasoning needed:** Implicit semantic role (instrument) and causation
From sentences 8–10, indicate that Nepthys stood over Dvorov’s (dead) body with a sword which was streaked with Dvorov’s blood. Therefore, one can infer that the blood on the sword must have been caused by the separation of Dvorov’s head from his body and that the sword must have been the implicit instrument of this action.

---

Figure 11: More example questions and corresponding reasoning phenomena from our MultiRC corpus.