UNIVERSITÀ
DI TRENTO
Department of
Information Engineering and Computer Science

SiS LAB
Signals & Interactive Systems

Guidelines for Evaluating the Quality of Data Items for the SocialIQA Benchmark

Description of the annotation task

In this task you are asked to evaluate the correctness of data items, from the SocialIQA dataset, for which an LLM has given a wrong answer. SocialIQA is a question-answering benchmark (3 possible answers) for testing social commonsense intelligence. The objective of the annotation task is to evaluate the validity of the benchmark, understanding if the model had given a bad answer because of its limitations or because there are errors in the data. The task is composed of two sub-tasks.

You are given some SocialIQA items, composed of context, question, and the three possible answers (A, B, and C), the ground truth option, and the model's answer.

Sub-task 1: Evaluating Data Items

In the first sub-task, you are asked to evaluate each data item, with relative ground truth answer and model prediction, using four possible categories: *structural flaw, semantic flaw, pragmatic flaw, and model error.*

A) ***Structural flaws*** include surface level or syntactic issues, such as duplicated entries, repetitive answer options, incomplete prompts, broken syntax, and grammatical errors in the context or answer choices.

B) ***Semantic Flaws*** involve inconsistencies or ambiguities in meaning, such as mismatches in semantic roles (e.g., assigning actions to the wrong participant), violations of temporal logic, referential confusion, and under-specified or contextually vague questions that allow multiple answer options to be equally plausible.

C) ***Pragmatic Flaws*** include answer options that are implausible, socially incoherent, or misaligned with real world expectations. These include responses that fail to reflect reasonable human motivations, emotional reactions, or common social behavior, as well as answers that are awkwardly phrased or pragmatically unnatural given the context.

D) ***Model errors*** include answer options that are wrong because of the model prediction limitations, and don't contain any issues in the data itself.

Sub-task 2: Providing an explanation for the selected choice

Following the evaluation, you are asked to specify a note about your choice. The note should be between 3 to 30 words long. For example:

"depends on the personality of the Agent in the answer"

or

"slight grammatical issue in the context ('was home with the new baby my Tuesday')"

You will be asked for the note only for the three types of flaws, for the model error it is not mandatory.

Examples

Examples of Structural Flaw

<table>
<tr><th colspan="3">SocialIQA Item</th></tr>
<tr><td colspan="3"><b>Context:</b> Riley looked at Jesse.

<b>Question:</b> What will Riley want to do next?

<b>AnswerA:</b> Question: At the awards dinner, Riley looked at Jesse with admiration. Why did Riley do this? Jesse had won the top prize for the company and the team
<b>AnswerB:</b> had planned the awards dinner and it had gone very smoothly
<b>AnswerC:</b> Look at something else

<b>Ground truth:</b> C</td></tr>
<tr><th>Model Response</th><th>Evaluation</th><th>Explanation</th></tr>
<tr><td>A</td><td>Structural Flaw</td><td>AnswerA contains context of the question</td></tr>
</table>

SocialIQA Item

Context: Tracy enjoyed constructing things and made a table.

Question: How would she feel as a result?

answerA: accomplished
answerB: accomplished
answerC: proud

Ground truth: A

Model Response	Evaluation	Explanation
B	Structural Flaw	Answer A and B are identical options

Examples of Semantic Flaws

SocialIQA Item

Context: It was a pretty sunny day outside. Quinn walked instead.

Question: What does Quinn need to do before this?

AnswerA: have fun
AnswerB: notice the sun
AnswerC: move their feet

Ground truth: C

Model Response	Evaluation	Explanation
B	Semantic Flaw	Answer C is a violation of temporal logic, while answer B is more accurate

SocialIQA Item

Context: Skylar was so excited that her mother bought her a new dress.”

Question: How would you describe Skylar?

AnswerA: elated
AnswerB: very happy
AnswerC: a person who likes fashion

<table>
<tr><td colspan="3"><b>Ground truth:</b> A</td></tr>
<tr><th>Model Response</th><th>Evaluation</th><th>Explanation</th></tr>
<tr><td>B</td><td>Semantic Flaw</td><td>All three answers are equally plausible</td></tr>
</table>

Examples of Pragmatic Flaws

<table>
<tr><th colspan="3">SocialIQA Item</th></tr>
<tr><td colspan="3"><b>Context:</b> Tracy the bus driver took Jesse's students on a field trip to a remote location.

<b>Question:</b> How would Tracy feel afterwards?

<b>AnswerA:</b> lost
<b>AnswerB:</b> accurate
<b>AnswerC:</b> incompetent

<b>Ground truth:</b> B</td></tr>
<tr><th>Model Response</th><th>Evaluation</th><th>Explanation</th></tr>
<tr><td>A</td><td>Pragmatic Flaw</td><td>These are unnatural answer options, in relation to the context</td></tr>
</table>

<table>
<tr><th colspan="3">SocialIQA Item</th></tr>
<tr><td colspan="3"><b>Context:</b> Bailey was a shy kid at school. They made no friends.

<b>Question:</b> What will happen to Bailey?

<b>AnswerA:</b> get work done
<b>AnswerB:</b> go to a party
<b>AnswerC:</b> Nothing will happen to others

<b>Ground truth:</b> A</td></tr>
<tr><th>Model Response</th><th>Evaluation</th><th>Explanation</th></tr>
<tr><td>C</td><td>Pragmatic Flaw</td><td>Missing context to choose a correct answer</td></tr>
</table>

Examples of Model Errors

<table>
<tr><th colspan="3">SocialIQA Item</th></tr>
<tr><td colspan="3"><b>Context:</b> Jan used Robin's approach on how to do well in a job interview.
<b>Question:</b> What will Jan want to do next?
<b>AnswerA:</b> do a job interview
<b>AnswerB:</b> make a high income
<b>AnswerC:</b> remember
<b>Ground truth:</b> B</td></tr>
<tr><th>Model Response</th><th>Evaluation</th><th>Explanation</th></tr>
<tr><td>A</td><td>Model Error</td><td>Jan already did the job interview</td></tr>
</table>

<table>
<tr><th colspan="3">SocialIQA Item</th></tr>
<tr><td colspan="3"><b>Context:</b> Casey gave some money to Jesse so she could go to the movie.
<b>Question:</b> How would you describe Casey?
<b>AnswerA:</b> thankful
<b>AnswerB:</b> greedy
<b>AnswerC:</b> giving
<b>Ground truth:</b> C</td></tr>
<tr><th>Model Response</th><th>Evaluation</th><th>Explanation</th></tr>
<tr><td>A</td><td>Model Error</td><td></td></tr>
</table>