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

Description of the annotation task

In this task you are asked to evaluate the correctness of data items, from the ToMi dataset, for which an LLM has given a wrong answer. ToMi is a question-answering benchmark to test the theory of mind abilities of language models. The objective of the annotation task is to evaluate the validity of the benchmark, based on the presence of three recurrent data issues. The task is composed of two sub-tasks.

You are given some ToMi items, composed of context, question, story type, question type, and ground truth, and the model prediction.

Dataset description

The next description is taken directly from the Github page where the code for generation is available. Please, pay attention to their story types and questions classification.

Story types can be one of:

1. `true_belief` - All agents observed all actions
2. `false_belief` - An agent failed to observe an action
3. `second_order_false_belief` - An agent has a false belief about another agent's set of beliefs

Question types can be one of:

1. `first_order_(0|1)_tom` - A first order false belief question in a story where a false belief situaion has been established

2. first_order_(0|1)_no_tom - A first order false belief question in a story where the agent in question observed all actions
3. second_order_(0|1)_tom - A second order false belief question in a story where a second order false belief situation has been established
4. second_order_(0|1)_no_tom - A second order false belief question in a story where the agent in question does not have a second order false belief
5. reality - A control question (ex: "Where is object *x* now?")
6. memory - A control question (ex: "Where was object *x* at the beginning?")

Example story:

Jackson entered the hall. Chloe entered the hall. The boots is in the bathtub. Jackson exited the hall. Jackson entered the dining_room. Chloe moved the boots to the pantry.
(Memory) Where was the boots at the beginning? bathtub
(Reality) Where is the boots really? pantry
(First-order belief agent 1) Where will Jackson look for the boots? bathtub
(First-order belief agent 0) Where will Chloe look for the boots? pantry
(Second-order belief agent 1) Where does Jackson think that Chloe searches for the boots? bathtub
(Second-order belief agent 0) Where does Chloe think that Jackson searches for the boots? bathtub

As you can see in the example story,

- *first-order belief* questions test the ability to infer about other people's mental states (*What does X think?)*
- *second-order belief* questions test the ability to infer beliefs about beliefs, in other words, the beliefs of one agent about another agent's beliefs (*What does X think that Y thinks?)*

Sub-task 1: Evaluating Data Items

In the first sub-task, you are asked to evaluate each data item, with the relative model prediction, using four possible categories: *incorrect ground truth, incorrect story type, incorrect agent number, and no data issues*.

A) ***Incorrect ground truth:*** the given ground truth is referring to a wrong object location.

B) ***Incorrect story type:*** a *true_belief* story has been labeled as *false_belief* or *second_order_false_belief* and vice versa.

C) ***Incorrect agent number:*** agent number *0* is wrongly assigned number *1* or vice versa. From the authors' description, ***agent 0* is always the agent moving the object from a location to another one**.

D) ***No data issues:*** none of the above issues is present in the data item.

An item can contain more than one issue, in that case assign more labels to it.

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:

"James is agent 0 because is moving the object"

or

"the story should be true_belief because all agents observed all actions"

You will be asked for the note only for the three types of issues, for the ***no data issues*** it is not mandatory.

Examples

Examples of Incorrect Ground Truth

ToMi Item		
Context: 1 Evelyn entered the basement. 2 Owen entered the basement. 3 The shoes is in the cupboard. 4 Owen exited the basement. 5 Evelyn moved the shoes to the bucket. 6 Owen hates the suit 7 Chloe entered the basement. 8 Evelyn likes the apple **Question:** 9 Where does Evelyn think that Owen searches for the shoes? **Question type:** second_order_0_no_tom **Story type:** false_belief **Ground truth:** bucket		
Model Response	**Evaluation**	**Explanation**
Cupboard	Incorrect Ground Truth	Owen exited the room before that Evelyn moved the object to the new location

Examples of Incorrect Story Type

ToMi Item
Context: 1 Mila entered the TV_room. 2 Chloe hates the pants 3 William is in the TV_room. 4 The cap is in the drawer. 5 Chloe entered the TV_room. 6 Chloe exited the TV_room. 7 William exited the TV_room. 8 William entered the patio. 9 Mila moved the cap to the bucket. **Question:** 10 Where will William look for the cap? **Question type:** first_order_1_tom **Story type:** true_belief **Ground truth:** drawer

Model Response	Evaluation	Explanation
bucket	Incorrect Story Type	William missed Mila moving the object, so the story is a false_belief one

Examples of Incorrect Agent Number

ToMi Item
Context: 1 Jack entered the master_bedroom. 2 William entered the living_room. 3 Owen entered the master_bedroom. 4 The pajamas is in the treasure_chest. 5 Jack exited the master_bedroom. 6 William exited the living_room. 7 Owen moved the pajamas to the pantry. **Question:** 8 Where will Jack look for the pajamas? **Question type:** first_order_0_tom

Story type: false_belief

Ground truth: treasure_chest

Model Response	Evaluation	Explanation
pantry	Incorrect Agent Number	Jack didn't move the object, so he should be agent number 1

Examples of More Data Issues

ToMi Item
Context: 1 Emma entered the crawlspace. 2 Oliver entered the crawlspace. 3 Alexander likes the socks 4 Alexander entered the crawlspace. 5 The grapes is in the crate. 6 Emma exited the crawlspace. 7 Alexander hates the pineapple 8 Oliver moved the grapes to the box. 9 Alexander exited the crawlspace. 10 Emma entered the crawlspace. **Question:** 11 Where will Emma look for the grapes? **Question type:** first_order_0_no_tom **Story type:** false_belief **Ground truth:** box

Model Response	Evaluation	Explanation
crate	Incorrect Ground Truth, Incorrect Agent Number	Emma didn't move the object and exited before seeing the transfer

Examples of No Data Issues

ToMi Item		
Context: 1 Hannah dislikes the slacks 2 Isla entered the hallway. 3 Mila loves the onion 4 Hannah entered the hallway. 5 Mila entered the hallway. 6 The tie is in the treasure_chest. 7 Isla moved the tie to the drawer. 8 Hannah exited the hallway. **Question:** 9 Where does Isla think that Hannah searches for the tie? **Question type:** second_order_0_no_tom **Story type:** true_belief **Ground truth:** drawer		
Model Response	**Evaluation**	**Explanation**
treasure_chest	No Data Issues	