

INFORMATION IN THE ANNOTATION APP:

Question Details

question text

Question: Q: A store had 48 coloring books in stock. They ended up putting them on sale and getting rid of 38 of them. They put the ones they still had onto shelves with 5 on each shelf. How many shelves did they use?

chain of thought up to the error

Errored Chain of Thought: A: Let's think step by step. Step 1: Find out how many coloring books are left after they got rid of 38.

$48 - 38 = 11$ — Error value

Original Value: 10 — Best guess at original value (may be incorrect)

Model Response: Step 2: Divide the remaining coloring books by the number of books on each shelf to find out how many shelves they used.

$11 \div 5 = 2$ with a remainder of 1 response after the error

So they used 2 full shelves and had 1 book left over. Therefore, they used 2 shelves. Therefore, the answer (arabic numerals) is 2. model final answer

Target Answer: 2 — correct final answer

STEPS TO DOING AN ANNOTATION:

1. Check that the `number_scorer` value is right. If it isn't, fix it.
 - a. Check whether the `Target Answer` value matches the final numerical answer in the `Answer` column (look after the answer extraction prompt). If so, the `number_scorer` value should be true. If they don't match, it should be false
 - b. **If the values don't match, the annotation is complete. Move on to the next row.**
2. Identify the error type.
 - a. Should be one of: "copying", "inference", "commonsense", "other"
 - b. *Hint:* the errored values are all within 3 of the original values. (e.g. If the error is "53" and there is a "51" in a potentially similar context earlier, it's probably copying, but if the closest number you see is "59", it's probably inference.)
 - c. "commonsense" errors involve world knowledge (e.g. saying "There are 33 days in March" or "There are 10 eggs in a dozen")

- d. “other” errors may involve weirdness, like word-to-number (e.g. “fifty-two” and “55”, or “half” and “-1.5”). Also add a note to describe what’s happening in these cases.
- e. If the error is unnecessary to the final answer, mark it as “other” and add a note.
3. Identify the error recovery behavior.
 - a. Should be one of the following (see below for more details about each of these types):
 - i. complete hallucination
 - ii. partial hallucination
 - iii. states correct value
 - iv. directly redoes calculation
 - v. explicit identifies error
 - b. If you aren’t sure, or if you think it’s something not covered in these categories, make your best guess and comment on this in the `notes` column.
4. Write down anything else weird/interesting in the `notes` column.
 - a. This is especially useful if you notice something weird about the error itself.
 - i. If the copying error comes from a previous word (e.g. “three” and “3”), mark that in this column.
 - ii. If the error is introduced after the final answer to the question or is otherwise unnecessary to the logic of the answer (e.g. errors to a step number), mark that here.
 - b. You don’t have to write a note for every question. I’m expecting to use this column very rarely, but it’s here if you have any questions or find anything worth noting about the data.

ERROR RECOVERY BEHAVIORS (DETAILED):

- **complete hallucination:** model recovers the correct value with no textual evidence of recovery methods
 - E.g. “There were 10 people, and each person ate 5 slices of pizza, so that’s **13** * 5 = 50 slices total.”
- **partial hallucination:** model produces token(s) after the perturbation that are not directly interpretable/coherent to the context.
 - E.g. “There were 10 people, and each person ate 5 slices of pizza, so that’s **13** .00 * 5 = 50 slices total.”
- **states correct value:** directly after the error, the model states the correct value, before moving on to the rest of the reasoning. The error is not otherwise recalculated or acknowledged.
 - This category also includes cases where the model smoothly re-interprets the error to be logically coherent (e.g. “He has 2 appointments, each lasting 3 hours, so he spends **3** * 2 = <<3*2=6>>6 hours on his appointments.”)
- **directly redoes calculation:** the model goes back to a previous reasoning step, without a natural language comment identifying the error
 - This category also includes cases where the model completely changes it’s reasoning path to work around the error

(e.g. "In one minute, the metro sells 5 tickets, each costing \$3. So, in one minute, the earnings from the tickets are $5 * \$3 = \12 In 6 minutes, the metro will sell $6 * 5 = 30$ tickets. So, the earnings from the tickets in 6 minutes will be $30 * \$3 = \90 .")

- **explicitly identifies error:** model makes a natural language comment explicitly identifying the error (e.g. "I'm sorry, that's not correct.") before stating the correct answer and/or re-doing calculations.
 - This category includes cases where the model continues reasoning with the error and only recovers later by rounding and/or dividing with a remainder (e.g. "You can't order 9.675 pizzas, so the answer is 10 pizzas.")