Utrecht University

An Evaluation of Temporal Order and Causal Understanding in LLMs and LVLMs

Soheila Amiri

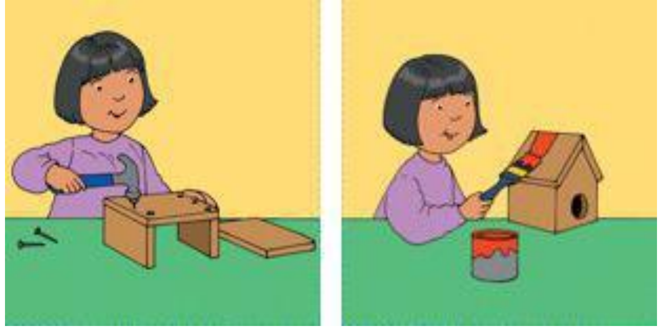
20/11/2025

Introduction

Mia built a little wooden birdhouse, painted it bright red, and hung it on a tree branch. Soon, a tiny bird moved in, and Mia smiled, proud of her new feathery neighbor. (GPT 5-1)

Wait a minute…

Introduction



+

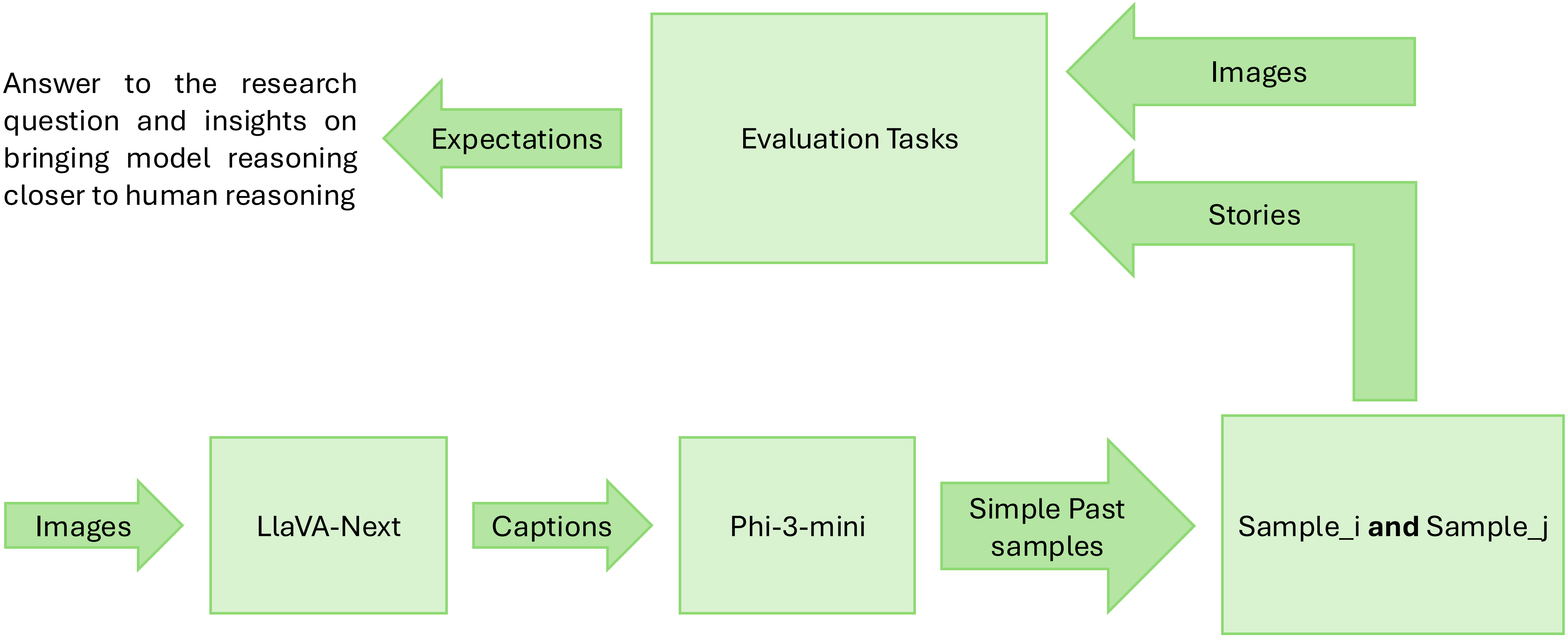
the girl nailed the wooden birdhouse and the girl painted the wooden birdhouse.

- How dependent the occurrence of the second event is on the first?
- What is the right temporal/causal order of events?
- Which event happened after the other?

....

?

Research Overview



Answer to the research question and insights on bringing model reasoning closer to human reasoning

Research Question:

How do different models handle temporal reasoning compared to causal reasoning in storytelling tasks?

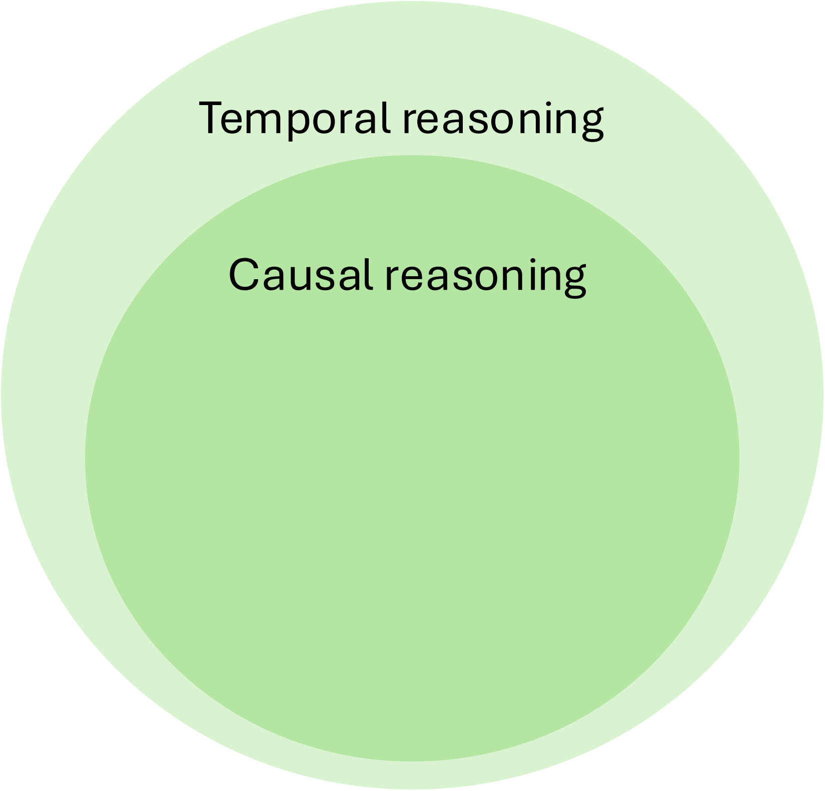
- **Temporal Reasoning: When?**

  A happened before/after B.

- **Causal Reasoning: Why?**

  A caused B to happen.

  B happened because A happened before B.



Temporal reasoning

Causal reasoning

Temporal Reasoning vs Causal Reasoning

The girl watched TV and the girl slept.

- ➢ Did she fall asleep because she watched TV?

  OR
- ➢ She fell asleep simply because it was her usual bedtime.

Data - Motivation

Event A happened AND Event B happened.

- **Temporal relationships between events**
  - **Symmetric** : The girl read and the girl played.
  - **Asymmetric**
    - The dog got dirty and the dog took a bath.
    - The boy drank the poison and the boy died.

Data - Details

- 342 single events distributed across 80 stories.
- Stories consisting of 2 or 3 images/events, capturing both symmetric and asymmetric relationships between events.

	Symmetric	Asymmetric	total
2-event stories	142	271	413
3-event stories	114	125	239

Data - Examples

The woman made the bed and the woman did the laundry.

The dog was dirty and the dog took a bath.

The children stacked the body part of the snowman and the children added the head of the snowman on top.

Evaluation Tasks

Considering both unimodal and multimodal tasks:

- Rating Task
- Multiple Choice Question

Models Under Evaluation:

- GPT5-mini
- Qwen3-VL
- Llama 3
- llava-onevision-qwen2-7b-ov

Evaluation Taks - Rating Task

Temporal Reasoning:

Temporal Order Ratings:

- **1** = not correct
- **2** = can be correct
- **3** = completely correct

Causal Reasoning:

Dependency Ratings:

- **1** = not dependent
- **2** = can be dependent
- **3** = completely dependent.

Rating Task- Prompting

Causality:

I will show you a sentence describing two events, in the form "A and B". I want you to think about the causal link between A and B. Rate how dependent the occurrence of B is on A, considering the given order in the text. Use these values: 1 = not dependent, 2 = can be dependent, 3 = completely dependent. Assume events are happening on the same day. Provide your answer in the following form: A: <event A> B: <event B> Final rating: [value] Explanation: <your explanation> Sentence: ...

Utrecht
University

Rating Task- Prompting

Time:

I will show you a sentence describing two events, in the form "A before B". I want you to think about the temporal link between A and B. Rate how correct the order of events is in the sentence. Use these values: 1 = not correct, 2 = can be correct, 3 = completely correct. Assume events are happening on the same day. Provide your answer in the following form: A: <event A> B: <event B> Final rating: [value] Explanation: <your explanation> Sentence: ...

Rating Task - Challenges

To be continued…

- **Ground Truth:**

  Humans Judgments + Model-Generated Answers

- **Prompting:**

  Words matter…

Challenges - Prompting

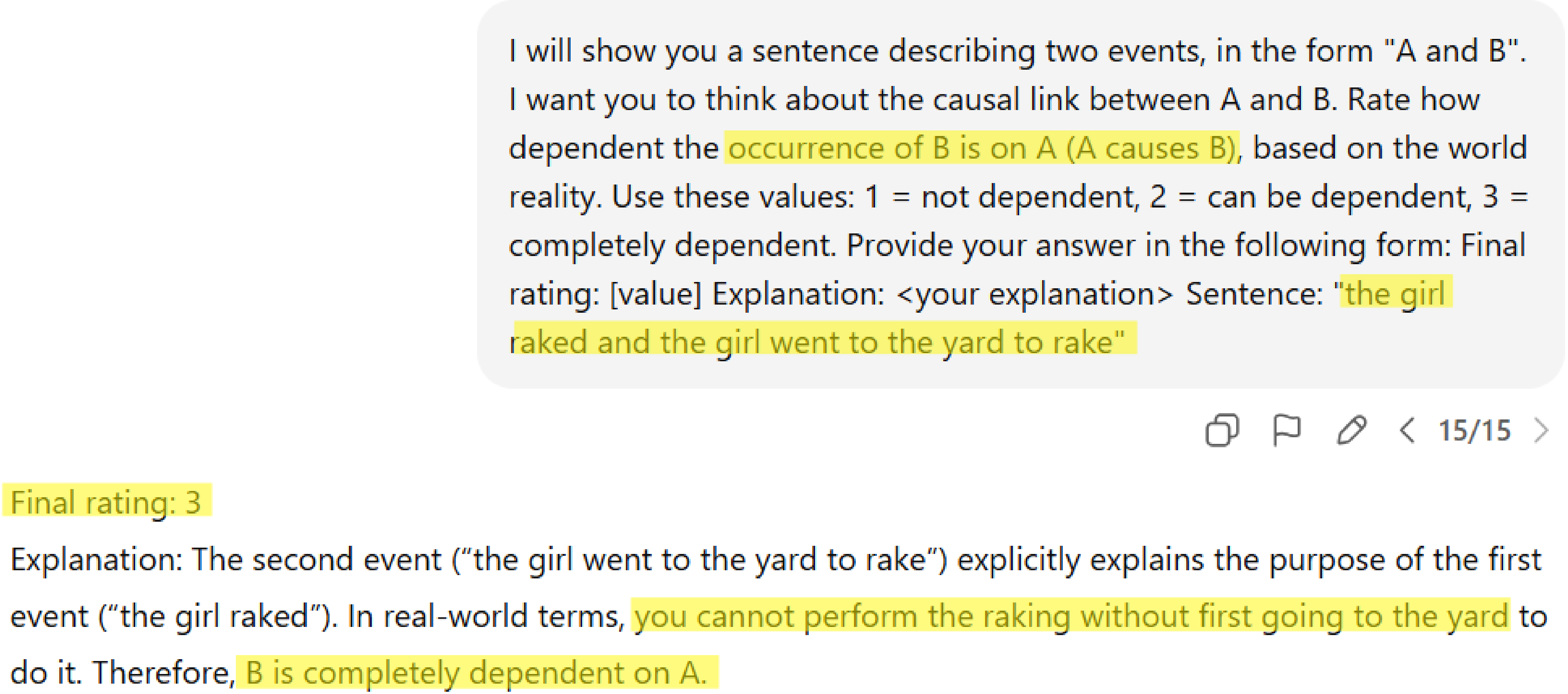
You are performing a task that involves reasoning about time and causality based on reality. Given the following text, based only on the order of events in the text, rate how correct the order of events is. Use: 1 = not correct, 2 = absolutely or possibly correct. Provide also the numerical value as: Final Rating: value. Generate 5 different answers. "The girl got poisoned and the girl drank water."

Rating Task - Findings

To be continued...

- **Detail Neglect in Prompts**
- **Confusion in Reasoning**
- **Bias in the Generated Response**
- **Model Uncertainty in Generated Responses**
- **Temporal Gaps between events**

Findings – Confusion in reasoning



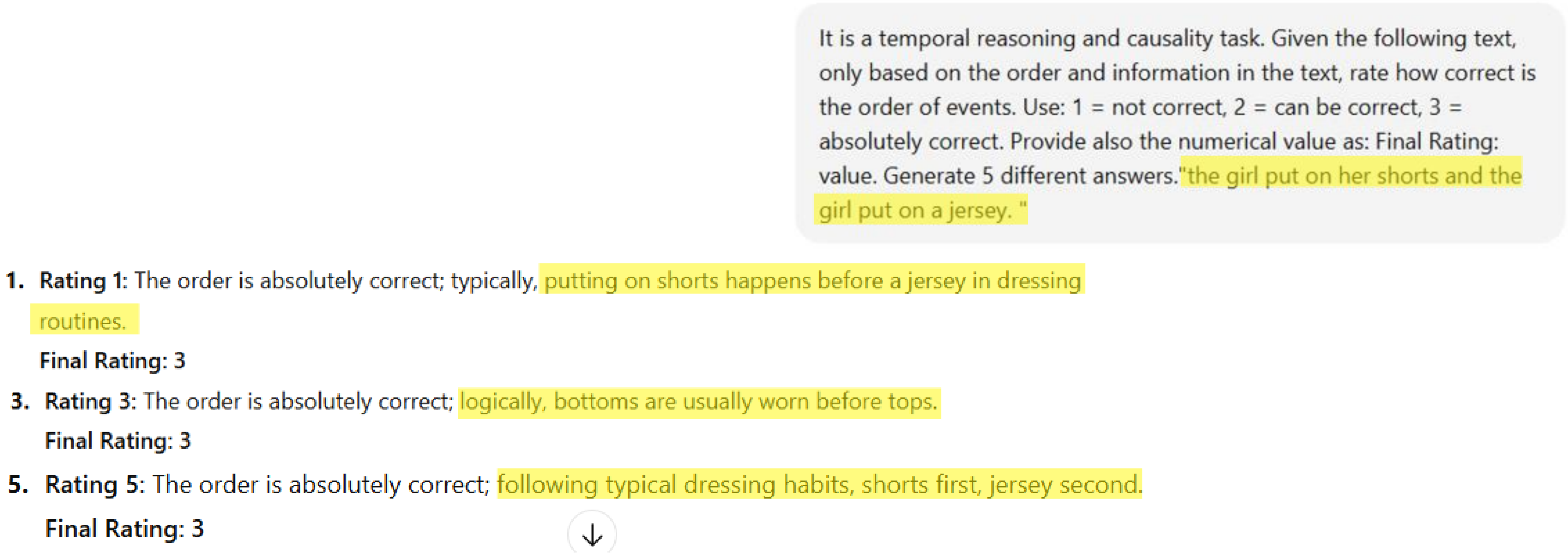
I will show you a sentence describing two events, in the form "A and B". I want you to think about the causal link between A and B. Rate how dependent the occurrence of B is on A (A causes B), based on the world reality. Use these values: 1 = not dependent, 2 = can be dependent, 3 = completely dependent. Provide your answer in the following form: Final rating: [value] Explanation: <your explanation> Sentence: "the girl raked and the girl went to the yard to rake"

Final rating: 3

Explanation: The second event (“the girl went to the yard to rake”) explicitly explains the purpose of the first event (“the girl raked”). In real-world terms, you cannot perform the raking without first going to the yard to do it. Therefore, B is completely dependent on A.

Utrecht University

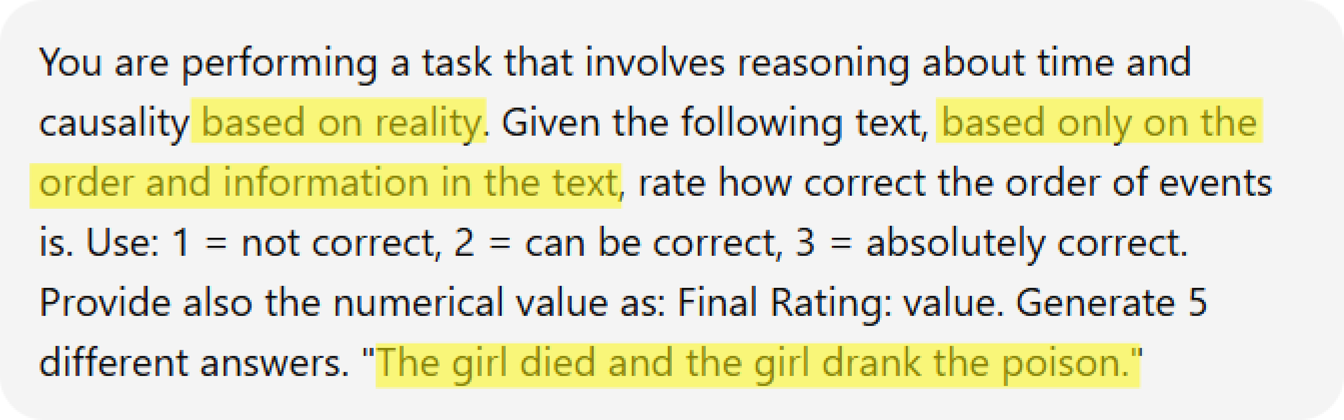
Findings - Bias



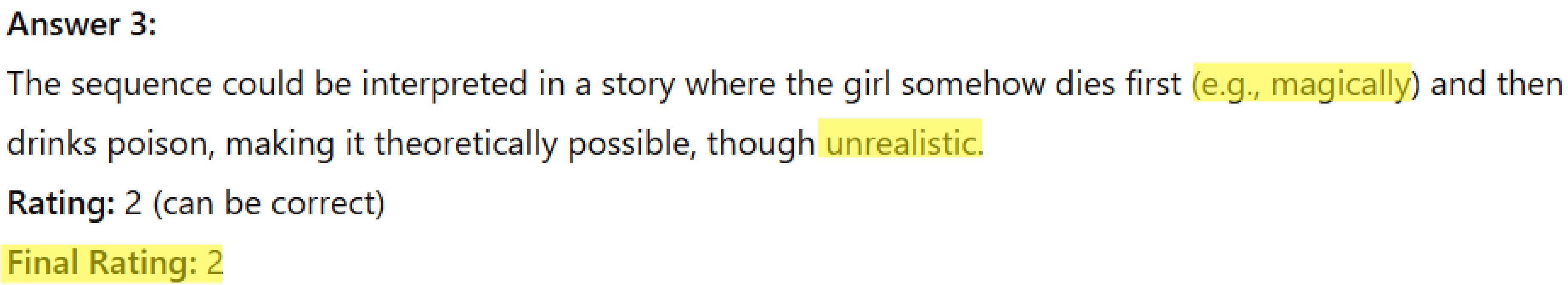
It is a temporal reasoning and causality task. Given the following text, only based on the order and information in the text, rate how correct is the order of events. Use: 1 = not correct, 2 = can be correct, 3 = absolutely correct. Provide also the numerical value as: Final Rating: value. Generate 5 different answers."the girl put on her shorts and the girl put on a jersey. "

1. **Rating 1:** The order is absolutely correct; typically, putting on shorts happens before a jersey in dressing routines.
   Final Rating: 3
3. **Rating 3:** The order is absolutely correct; logically, bottoms are usually worn before tops.
   Final Rating: 3
5. **Rating 5:** The order is absolutely correct; following typical dressing habits, shorts first, jersey second.
   Final Rating: 3

Findings – Detail Neglect



You are performing a task that involves reasoning about time and causality based on reality. Given the following text, based only on the order and information in the text, rate how correct the order of events is. Use: 1 = not correct, 2 = can be correct, 3 = absolutely correct. Provide also the numerical value as: Final Rating: value. Generate 5 different answers. "The girl died and the girl drank the poison."



Answer 3:

The sequence could be interpreted in a story where the girl somehow dies first (e.g., magically) and then drinks poison, making it theoretically possible, though unrealistic.

Rating: 2 (can be correct)

Final Rating: 2

Thank you!