Common Sense Physical Reasoning in Dutch

Sander Land
sander@writer.com

Judith Land

1 Dataset Creation

One hundred samples were created manually and with LLM assistance over several weeks. Manually created samples were typically based on brainstorming about culturally relevant topics, including chocolate sprinkles on bread, ice skating, dikes, local sports, and specific dishes.

Language models, including GPT-5, Gemini 2.5 Pro, and Claude Sonnet 4, were used in drafting samples, suggesting topics, and proofreading. Overall, we found their performance severely lacking in understanding the task and generating suitable samples. Models tended to use overly obvious wrong answers, go in the direction of factual knowledge questions, or both. Table 1 (b) shows such an example falling below the required standards. Including the full project guidance and fifty examples increased the rate of useful output, but it remained relatively low at approximately one in four samples being either directly acceptable or requiring minor edits. More effective strategies included finding basic tutorials for various actions on the internet and rewriting them as short samples.

One specific sample shown in Table 1 (a) revealed reasoning errors across leading AI models, including Claude Sonnet 4, GPT-5, and Gemini 2.5 Pro. When asked about entering a canal lock to descend, the models unanimously chose an answer suggesting one should wait for the water to fall. This demonstrates that even the strongest models can still consistently fail at physical reasoning, and become distracted by simple word and phrase similarity.

(a) Difficult Sample

Question

You are sailing your boat into a lock that goes to a lower water level. How do you get to the other side?

Correct Solution

If necessary, wait for the water in the lock to **rise**. Sail into the lock. Tie your boat to the floating bollard. Wait until the water in the lock is equal to the water level you are going to, untie your boat and sail out.

Incorrect Solution

If necessary, wait for the water in the lock to **lower** [...]

(b) Poor LLM-generated sample

Situation

You want to reheat old rice without it becoming dry.

Correct Solution

Add a little water and heat covered in the microwave.

Incorrect Solution

Heat the rice on high heat in a dry pan.

Table 1: **Notable samples:** (a) is a sample even strong current day models struggle with. (b) is a typical example of an LLM-generated sample when asked. Both samples were translated to English.

2 Dataset Checking

Samples were proofread by both authors, one of whom is a specialist in developing large language models, and one of whom is a professional scientific writer, for correctness, grammar, and spelling.