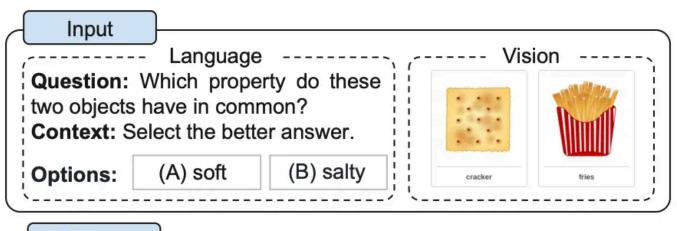
Multimodal CoT Prompting

Zhang et al. (2023) recently proposed a multimodal chain-of-thought prompting approach. Traditional CoT focuses on the language modality. In contrast, Multimodal CoT incorporates text and vision into a two-stage framework. The first step involves rationale generation based on multimodal information. This is followed by the second phase, answer inference, which leverages the informative generated rationales.

The multimodal CoT model (1B) outperforms GPT-3.5 on the ScienceQA benchmark.



Output

Rationale: Look at each object. For each object, decide if it has that property. Potato chips have a salty taste. Both objects are salty. A soft object changes shape when you squeeze it. The fries are soft, but the cracker is not. The property that both objects have in common is salty.

Answer: The answer is (B).

Figure 1. Example of the multimodal CoT task.

Image Source: Zhang et al. (2023).

Further reading:

<u>Language Is Not All You Need: Aligning Perception with Language Models</u> (Feb 2023)

Copyright © 2024 DAIR.AI