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.

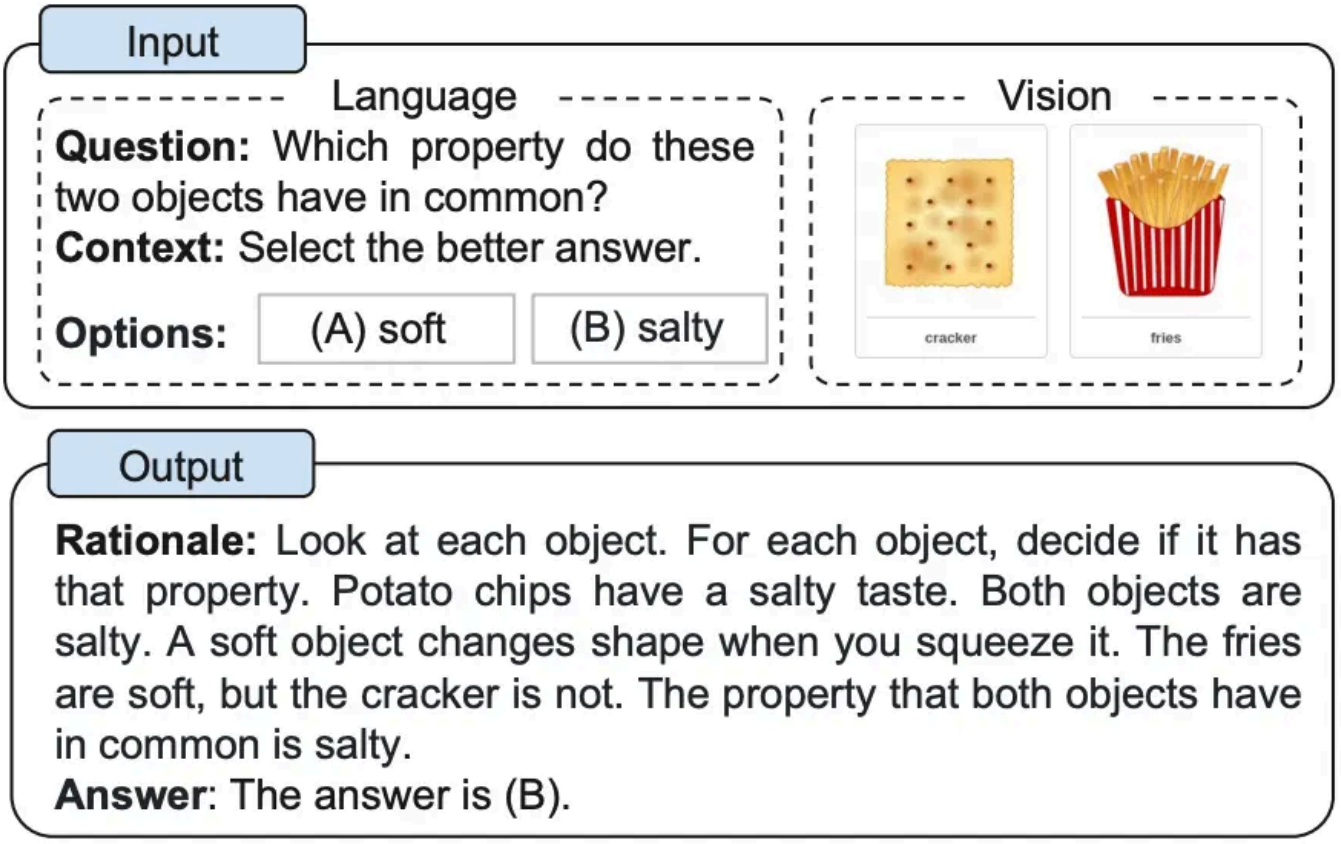

Figure 1. Example of the multimodal CoT task.

Image Source: Zhang et al. (2023)

Further reading:

- Language Is Not All You Need: Aligning Perception with Language Models (Feb 2023)