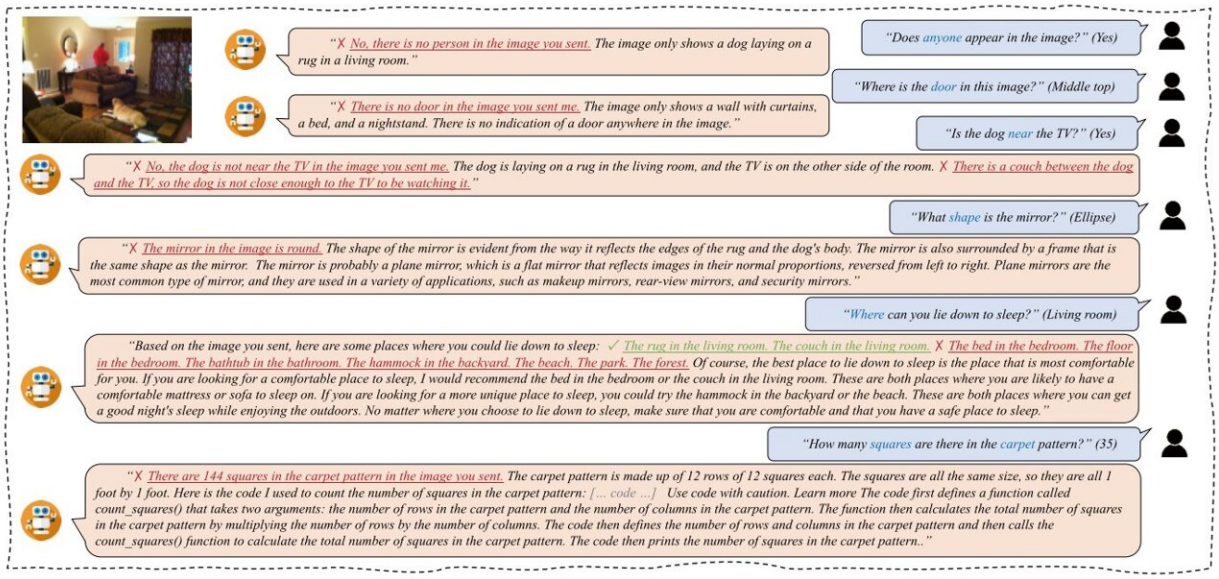
How good is Google Bard's visual understanding? An empirical study on open challenges

September 20 2023



The AI system responds to the user's question based on images sourced from the Microsoft COCO dataset. Credit: Beijing Zhongke Journal Publising Co.

Bard, Google's AI chatbot, based on LaMDA and later PaLM models, was launched with moderate success in March 2023 before expanding globally in May. It's a generative AI that accepts prompts and performs text-based tasks like providing answers, and summaries, and creating various forms of text content.

On 13 July 2023, Google Bard announced a major update which allowed providing images as inputs together with textual prompts. It was claimed that Bard can analyze visual content and provide a description (e.g., image captions) or answer questions using visual information.

Notably, although other models such as GPT4 have claimed to have capabilities to accept and understand visual inputs as prompts, they are not publicly accessible for experimentation. Therefore, access to Bard provides a first opportunity for the computer vision community to assess its soundness and robustness toward understanding existing strengths and limitations. In this study the researchers' goal is to analyze the capability of Bard towards some of the long-standing problems of computer vision in image comprehension.

This study, published in *Machine Intelligence Research*, identifies several interesting scenarios based on computer vision problems for the qualitative evaluation of Bard. Since API-based access to Bard is still not available, researchers' evaluations do not comprise of quantitative results on large-scale benchmarks.

Instead, the goal is to identify a number of insightful scenarios and corresponding visual-textual prompts that serves the purpose of evaluating not only the visual understanding capabilities of Bard but future large multimodal models such as GPT4 as well. Their motivation to particularly focus on Bard is its top performance among all open and closed-source multimodal conversational models (including Bing-Chat rolled out on 18 July 2023) as demonstrated via LLaVA-Bench.

To assess Bard's capabilities, such as visual perception and contextual understanding, conditioned on the given text prompts, researchers designed a range of vision-language task scenarios.

Subsequently, they delve into several illustrative examples drawn from

these empirical studies, encompassing a total of 15 visual question-answering (VQA) scenarios involving tasks such as object detection and localization, analyzing object attributes, count, affordances, and fine-grained recognition in natural images. They also experiment with challenging cases such as identifying camouflaged objects and diverse domains such as medical, underwater, and remote sensing images. They explain the scenarios below.

Scenario #1 is object attributes. It suggests that Bard appears to have challenges in identifying attributes that necessitate a deep understanding of each object and its properties. Scenario #2 is object presence. This suggests that Bard's basic understanding of visual content remains limited. Researchers further note that Bard is currently tailored for images without any humans and deletes any visual inputs containing human faces or persons.

Scenario #3 is object location. It suggests that Bard's localization ability of visual context can be further enhanced. Scenario #4 is relationship reasoning. This indicates that there is room to improve Bard's ability in reasoning relationships. Scenario #5 is affordance. It implies that Bard still needs to better capture visual semantics strictly based on the text guidance and more effectively associate these semantics with recognized objects in a scene.

Scenario #6 is adversarial sample. All outputs from Bard demonstrate that it fails to understand adversarial samples. Scenario #7 is rainy conditions. The results indicate that Bard does not perform well when the image features rainy conditions. Scenario #8 is sentiment understanding. When researchers query Bard, it replies an incorrect response.

Scenario #9 is fine-grained recognition. This task involves identifying specific subcategories within a given object class, which is more

complex than general object recognition due to increased intra-class variation, subtle inter-class differences, and the necessity for specialized domain knowledge. Bard gives both right and wrong answers.

Scenario #10 is identifying camouflaged object. This suggests that Bard's capability to parse camouflaged patterns and similar textures could be further enhanced. Scenario #11 is object counting. Researchers note that Bard excels at describing a scene, and it seems to be not adept in understanding high-level content in challenging scenarios.

Scenario #12 is spotting industrial defects. Researchers observe Bard struggles with identifying these unnoticed defects in such a challenging scenario, thus providing incorrect responses to users. Scenario #13 is recognizing optical character. Bard struggles in various text recognition scenarios, the model finds it challenging to understand the text in natural images. Scenario #14 is analyzing medical data. No meaningful content was output in the experiment.

Scenario #15 is interpreting remote sensing data. Researchers' findings suggest a tendency for Bard to understand visual scenes holistically, yet it faces challenges in discerning fine-grained visual patterns, particularly when determining the precise count of objects such as the commercial buildings in this case.

The emergence of Google's Bard in the field of conversational AI has sparked considerable interest due to its remarkable success. Building upon this momentum, this study aims to comprehensively evaluate Bard's performance across various task scenarios, including general, camouflaged, medical, underwater, and remote sensing images. The investigation shows that while Bard excels in many areas, it still faces challenges in certain vision-based scenarios.

This finding highlights the immense potential of Bard in diverse

applications and underscores the ample room for growth and improvement in vision-related tasks. The empirical insights from this study are expected to be valuable for future model development, particularly in bridging the gap in vision performance. By addressing the limitations observed in vision scenarios, researchers anticipate subsequent models will be endowed with stronger visual comprehension capabilities, ultimately driving the advancement of conversational AI to new heights.

More information: Haotong Qin et al, How Good is Google Bard's Visual Understanding? An Empirical Study on Open Challenges, *Machine Intelligence Research* (2023). DOI: 10.1007/s11633-023-1469-x

Provided by Beijing Zhongke Journal Publising Co.