Evaluating Vision-Language Models on the TriangleCOPA Benchmark

Ankur Chemburkar, Andrew Feng, and Andrew S. Gordon
University of Southern California, Los Angeles, California, USA
chemburk@usc.edu, feng@ict.usc.edu, gordon@ict.usc.edu

Abstract

The TriangleCOPA benchmark consists of 100 textual questions with videos depicting the movements of simple shapes in the style of the classic social-psychology film created by Fritz Heider and Marianne Simmel in 1944. In our experiments, we investigate the performance of current vision-language models on this challenging benchmark, assessing the capability of these models for visual anthropomorphism and abstract interpretation.

The TriangleCOPA Benchmark

Benchmark evaluations have been instrumental in gauging the remarkable progress of large language models over the last decade. Early natural language benchmarks for commonsense reasoning, including the 1000-question Choice of Plausible Alternatives (Roemmele, Bejan, and Gordon 2011) and the 120-question Winograd Schema Challenge (Davis, Morgenstern, and Ortiz 2017), saw the performance of top systems advance from near random-chance to near perfect accuracy over the last ten years. As vision-language models begin to emerge from OpenAI and others, including open-weight models such as LLAVA 1.5 (Liu et al. 2023), there is a strong need for vision-language benchmark evaluations to gauge progress among models and their various designs. However, there also exists opportunities to utilize various benchmarks that were created years before the first vision-language models existed.

The TriangleCOPA benchmark (Maslan, Roemmele, and Gordon 2015) is an early commonsense reasoning benchmark evaluation that can potentially serve to assess contemporary vision-language models. Modeled after the popular Choice of Plausible Alternatives benchmark, TriangleCOPA consists of 100 English-language questions, each with an accompanying video that visualizes a short scenario. The notable feature of TriangleCOPA is that each scenario is a two-dimensional animation set in the same domain as the classic Heider-Simmel film, developed for social psychology experiments (Heider and Simmel 1944), where the characters are a large triangle, a smaller triangle, and a circle, who maneuver in and around a box with a hinged door. Figure 1 shows the text of an example TriangleCOPA question along with a frame from its accompanying video, where Choice A was annotated as the correct answer. In this paper we investigate the use of the TriangleCOPA as a benchmark for evaluating various vision-language models under multiple conditions.

Question 83: *A small triangle and big triangle are next to each other. A circle runs by and pushes the small triangle. The big triangle chases the circle. Why does the big triangle chase the circle?*
Choice A: *The big triangle is angry that the circle pushed the small triangle, so it tries to catch the circle.*
Choice B: *The big triangle and circle are friends. The big triangle wants to say hello to the circle.*

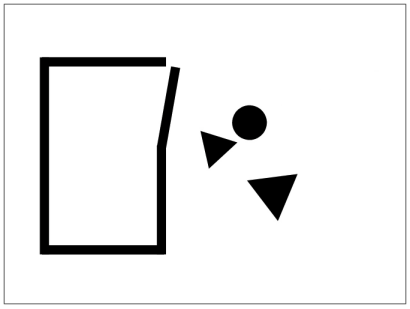

Figure 1: Text of a TriangleCOPA question with a frame from its accompanying video

Methodology

At the time of this writing, two prominent vision-language models were available for inclusion in our experiments. First, the GPT-4V model from OpenAI (OpenAI 2023) is a state-of-the-art closed-weight vision-language model, which we utilized via their commercial API. Second, Llava-1.5 (Liu et al. 2023) is an open-weight vision-language model that is competitive with GPT-4V on many tasks, which we utilized by downloading the 13B Vicuna-based checkpoint with 8-bit quantization from Huggingface and running on a local GPU cluster.

In addition, we included in our experiments several additional open-weight large language models (language-only),

evaluating their performance on the textual portion of TriangleCOPA questions. Namely, we included the Mistral 7B and Mixtral 8x7B models from Mistral AI (Mistral AI 2023a; 2023b) and Gemma 7B and Gemma Instruct 7B from Google AI (Gemma Team et al. 2024). We utilized each of these models via their corresponding Huggingface API.

To compute the accuracy of each model on the TriangleCOPA benchmark, we passed each of its 100 questions to each model with three variations of prompts (two for the language-only models), as follows:

Basic prompt: In the first prompt condition, we passed the entire TriangleCOPA question to the model, with the simple instruction *"From the given information you must answer the question."* Two more instruction sentences are appended to prevent refusals and one instruction showing the desired output format. The language-only models in our experiments are able to perform this task, as each TriangleCOPA question includes some description of the visual scene before presenting its choices.

Rich prompt: In our early tests, we found that several models struggled with answering questions about the behavior of triangles and circles, but could be coaxed into answering these questions by providing some additional context about the task. The *rich prompt* condition also provides only the textual portion of each question, but is prefaced with further instructions to anthropomorphize the shapes and answer the question as if they represented real people.

Vision prompt: In full vision-language condition, we modified the rich prompt by replacing the textual question in each TriangleCOPA item with a set of frames taken from the accompanying video that depicts the situation. To generate these image sequences for each video, a fixed number of frames were sampled at uniform intervals from the original video. Then, these frames were passed to each model along with the rich prompt. The textual question and the description was simply replaced by *"Choose the option that seems the most correct according to the video."*

When scoring models, refusal to answer a question was counted as incorrect. This criteria was most detrimental to scores of the Mixtral 8x7B, GPT-4V, and Llava-1.5 models, which refused to answer several questions when provided with the basic prompt.

Results

Table 1 lists the accuracy of each model under each prompt condition. Results show that Open AI's GPT-4V model outperforms all open-weight models across all prompt types.

The strong performance of both Mixtral 8x7B and GPT-4V on the text-only *rich prompt* condition is consistent with the excellent performance of these two systems on other text-only benchmarks. However, it should be noted that we cannot rule out the possibility that the text of entire TriangleCOPA evaluation was included in the training regimes of these models, which have not been disclosed publicly.

More notable is the mediocre performance of both vision-language models in the *vision prompt* condition. Here, Open AI's GPT-4V model fails to interpret the sequence of images as a coherent scenario for many TriangleCOPA questions, and the Llava-1.5 model barely performs above the 50% random baseline. In our subsequent querying of these two vision-language models about each question, we found that both models struggled to correctly answer even basic questions about the spatial relationships between the three characters and the box depicted in these abstract, two-dimensional scenes.

Table 1: Correct TriangleCOPA responses, out of 100

model	*basic prompt*	*rich prompt*	*vision*
random baseline	50	50	50
Mistral 7B	88	83	
Mixtral 8x7B	84	98	
Gemma 7B	67	68	
Gemma Instruct 7B	78	77	
Llava-1.5	88	87	53
GPT-4V	**90**	**100**	**64**

Conclusions

One of the qualities of a good benchmark for evaluating AI systems is that it is trivially easy for humans to complete, but extremely difficult for current AI systems. The original 1000-question Choice of Plausible Alternatives benchmark (Roemmele, Bejan, and Gordon 2011) had this quality, with a 50% random baseline and a 58.8% strong baseline when it was initially released over a decade ago. Since then, a myriad of different approaches achieved incremental improvements on this benchmark over many years, until contemporary large language models were able to achieve human-level performance.

Our experiments suggest that the TriangleCOPA benchmark also has this quality, with current vision-language models struggling to bridge the gap between a random baseline and human-level performance. Undoubtedly, much of the difficulty stems from the very abstract nature of the visual input, depicting the movements of simple shapes on a two-dimensional plane. The ease in which people anthropomorphize these shapes was the original interest of Fritz Heider and Marianne Simmel when they made the 1944 film that inspired the TriangleCOPA benchmark. As progress on vision-language models continues, the TriangleCOPA benchmark can serve as a useful gauge in assessing this human-like capacity for visual anthropomorphism and abstract interpretation.

Acknowledgments

Research was sponsored by the Army Research Office and was accomplished under Cooperative Agreement Number W911NF-20-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Davis, E.; Morgenstern, L.; and Ortiz, C. L. 2017. The first winograd schema challenge at ijcai-16. *AI Magazine* 38(3):97–98.

Gemma Team, T. M.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Sifre, L.; Rivière, M.; Kale, M. S.; Love, J.; Tafti, P.; Hussenot, L.; and et al. 2024. Gemma. Available online at: https://www.kaggle.com/m/3301. Accessed: March 16, 2024.

Heider, F., and Simmel, M. 1944. An experimental study of apparent behavior. *The American Journal of Psychology* 57(2):243–259.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. In Oh, A.; Neumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 34892–34916. Curran Associates, Inc.

Maslan, N.; Roemmele, M.; and Gordon, A. S. 2015. One hundred challenge problems for logical formalizations of commonsense psychology. In *Proceedings of the Twelfth International Symposium on Logical Formalizations of Commonsense Reasoning*, 107–113. Palo Alto, CA: AAAI Press.

Mistral AI. 2023a. Announcing mistral 7b. Accessed: March 16, 2024.

Mistral AI. 2023b. Mixtral of experts. Accessed: March 16, 2024.

OpenAI. 2023. Gpt-4v(ision) system card. Available online at: https://openai.com/research/gpt-4v-system-card. Accessed: March 16, 2024.

Roemmele, M.; Bejan, C.; and Gordon, A. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *Proceedings of the AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning, Stanford University*.