Visual Cognition in Vision-Language Models

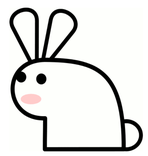
Krista A. Ehinger
School of Computing and Information Systems, The University of Melbourne

Introduction

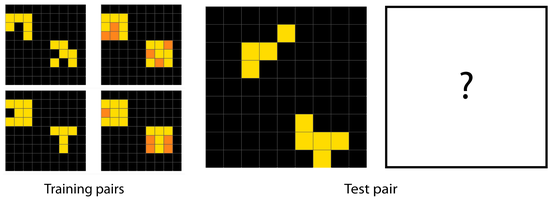
Multimodal Large Language Models (MLLMs) are state of the art for visual interpretation tasks such as describing images and answering questions based on photos or diagrams. MLLMs even show emergent behavior such as interpreting the humor in memes [1].

Despite strong performance on these tasks, MLLMs struggle with more abstract visual reasoning tasks [2]. The failures often suggest problems with basic perception: lack of shape understanding, inability to locate objects or understand their spatial relationships.

How well can MLLMs perform visual cognition tasks?



Easy task for MLLM: Describe this image. (GPT-4o: The image shows a stylized drawing of a bunny rabbit.)



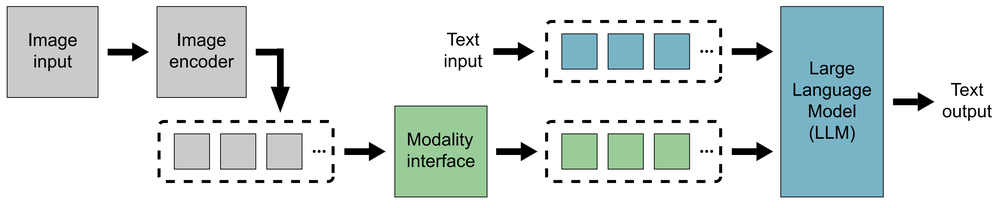
Hard task for MLLM: sample task from the Abstract Reasoning Corpus (ARC) [3]. The task is to complete the test pair following the pattern shown in the training pairs.

Multimodal Large Language Models

Accept text and image as input, produce text as output

Typical architecture consists of:
- Pretrained large language model
- Pretrained image encoder (e.g., CLIP [4]), trained on image-text alignment), which converts image pixels into a high-level abstract representation
- Interface model which converts the image abstraction into language-like tokens

Models are trained on large image-text datasets:
- Pretraining on image captioning (input image, output caption)
- Instruction tuning on a variety of visual question answering tasks (input image and query, output answer)

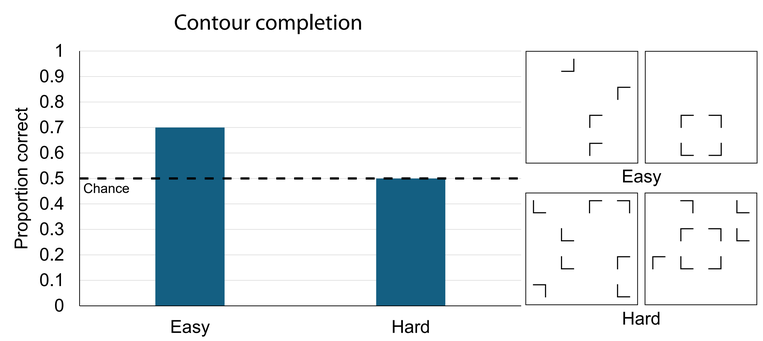


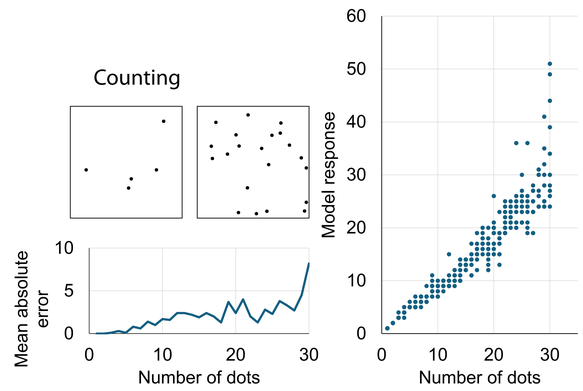
Typical architecture of an Multimodal Large Language Model for vision tasks

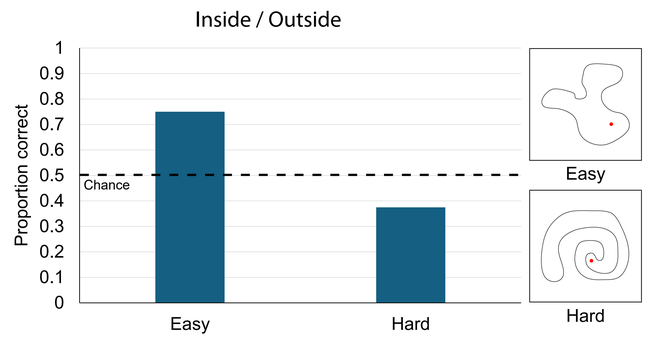
[1] Yin, S., et al. (2024). A survey on Multimodal Large Language Models. *IEEE TPAMI*.
[2] Ahrabian, K., et al. (2024). The Curious Case of Nonverbal Abstract Reasoning with Multi-Modal Large Language Models. *COLM*.
[3] Chollet, F. (2019). On the Measure of Intelligence. Arxiv.org/abs/1911.01547.
[4] Radford, A., et al. (2021) Learning transferable visual models from natural language supervision, *ICML*.
[5] Ullman, S. (1984). Visual routines. *Cognition*, 18(1-3), 97–159.

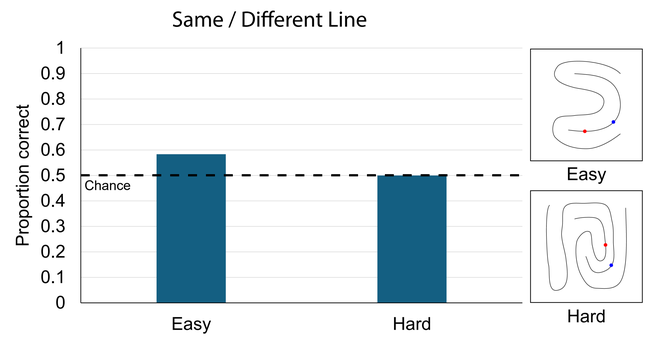
Performance on Visual Cognition Tasks

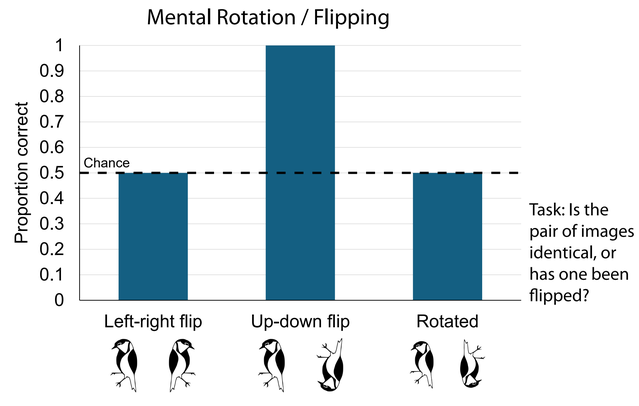
Model: GPT-4o, "zero shot" performance (no additional training)

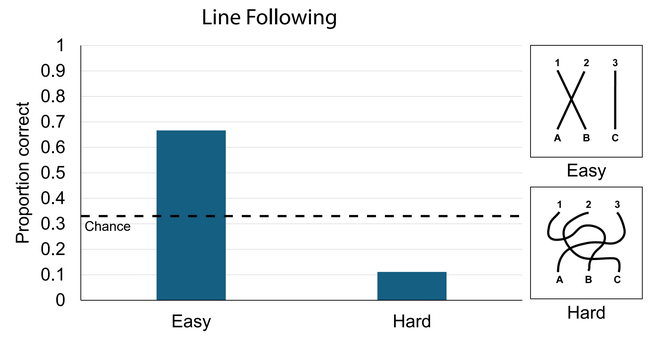












Conclusion

Despite excellent performance on recognition tasks, Multimodal Large Language Models (MLLMs) perform poorly on a range of visual cognition tasks that are easy for humans. They also show odd idiosyncrasies, like being able to recognize when objects are flipped upside down but not left-right.

MLLMs learn to translate images into a language-like representation, which may not be well-suited for visuospatial tasks or tasks which require "visual routines" [5]. At the same time, these results suggest that fundamental visual abilities, like contour completion, may not be strictly necessary for high-level object recognition and understanding, since MLLMs can recognize objects well despite lacking these fundamental abilities.