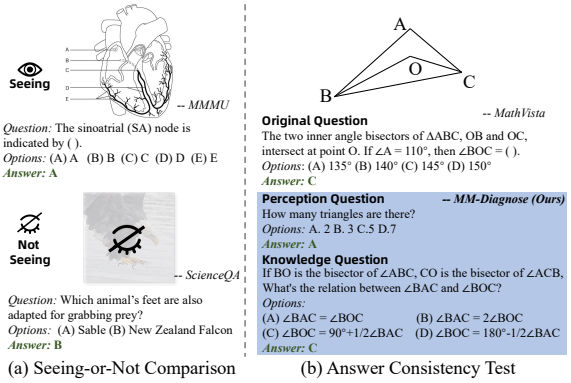
MMEVALPRO: Calibrating Multimodal Benchmarks Towards Trustworthy and Efficient Evaluation

Jinsheng Huang[1,2,3]*, Liang Chen[1,2]*, Taian Guo[1,2,3], Fu Zeng[4], Yusheng Zhao[1,2,3]
Bohan Wu[1,2,3], Ye Yuan[1,2,3], Haozhe Zhao[1], Zhihui Guo[5], Yichi Zhang[1], Jingyang Yuan[1,2,3]
Wei Ju[1,2,3], Luchen Liu[1,2,3], Tianyu Liu[6], Baobao Chang[1,2]†, Ming Zhang[1,2,3]†
[1]National Key Laboratory for Multimedia Information Processing, Peking University
[2]School of Computer Science, Peking University, [3]PKU-Anker LLM Lab
[4]Chinese Academy of Medical Sciences, [5]CUHK, [6]Alibaba Group
hjs@stu.pku.edu.cn leo.liang.chen@outlook.com {chbb,mzhang_cs}@pku.edu.cn
https://mmevalpro.github.io

Abstract

Large Multimodal Models (LMMs) exhibit impressive cross-modal understanding and reasoning abilities, often assessed through multiple-choice questions (MCQs) that include an image, a question, and several options. However, many benchmarks used for such evaluations suffer from systematic biases. Remarkably, Large Language Models (LLMs) without any visual perception capabilities achieve non-trivial performance, undermining the credibility of these evaluations. To address this issue while maintaining the efficiency of MCQ evaluations, we propose MMEVALPRO, a benchmark designed to avoid Type-I errors through a trilogy evaluation pipeline and more rigorous metrics. For each original question from existing benchmarks, human annotators augment it by creating one perception question and one knowledge anchor question through a meticulous annotation process. MMEVALPRO comprises 2,138 question triplets, totaling 6,414 distinct questions. Two-thirds of these questions are manually labeled by human experts, while the rest are sourced from existing benchmarks (MMMU, ScienceQA, and MathVista). Compared with the existing benchmarks, our experiments with the latest LLMs and LMMs demonstrate that MMEVALPRO is **more challenging** (the best LMM lags behind human performance by 31.73%, compared to an average gap of 8.03% in previous benchmarks) and **more trustworthy** (the best LLM trails the best LMM by 23.09%, whereas the gap for previous benchmarks is just 14.64%). Our in-depth analysis explains the reason for the large performance gap and justifies the trustworthiness of evaluation, underscoring its significant potential for advancing future research.



(a) Seeing-or-Not Comparison (b) Answer Consistency Test

Figure 1: Examples of the probing experiments.

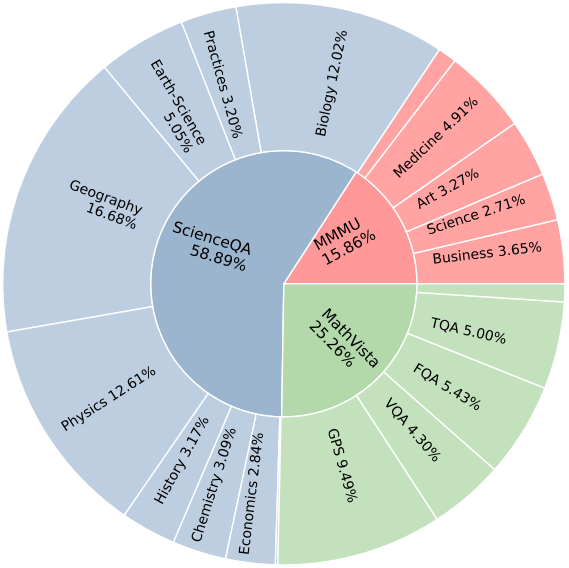


Figure 2: Topic distribution of MMEVALPRO's data.

1 Introduction

Ever since the birth of standardized testing, the credibility of its conclusions has been a significant concern. The same problem goes for the evaluation of recently popular Large Multimodal Models (LMMs) such as GPT4-o (OpenAI, 2024), Gemini-1.5 (Team et al., 2024), Qwen-VL (Bai et al., 2023b) and LLaVA (Liu et al., 2023b). One classic composition of such an evaluation is the multiple-choice question (MCQ), which includes an image, a question, possible choices, and an answer. This form of evaluation has higher usability

*Equal contribution. †Corresponding authors.