

# Automated Generation of Challenging Multiple-Choice Questions for Vision Language Model Evaluation

Anonymous CVPR submission

Paper ID 4379

## Abstract

The rapid development of vision language models (VLMs) demands rigorous and reliable evaluation. However, current visual question answering (VQA) benchmarks often depend on open-ended questions, making accurate evaluation difficult due to the variability in natural language responses. To address this, we introduce AutoConverter, an agentic framework that automatically converts these open-ended questions into multiple-choice format, enabling objective evaluation while reducing the costly question creation process. Our experiments demonstrate that AutoConverter can generate correct and challenging multiple-choice questions, with VLMs demonstrating consistently similar or lower accuracy on these questions compared to human-created ones. Using AutoConverter, we construct VMCBench, a benchmark created by transforming 20 existing VQA datasets into a unified multiple-choice format, totaling 9,018 questions. We comprehensively evaluate 28 state-of-the-art VLMs on VMCBench, setting a new standard for scalable, consistent, and reproducible VLM evaluation.

## 1. Introduction

With the rapid advancements in vision language models (VLMs) [2, 26, 39, 50], rigorous and reliable evaluation methods are critical to gauge model performance and guide further research. Visual question answering (VQA) has emerged as a standard evaluation protocol, typically structured with open-ended or multiple-choice questions. Open-ended questions [12, 34, 57] allow models to generate free-form, natural language responses, while multiple-choice questions [30, 46, 58] ask models to choose the answer from predefined options. Most existing VQA benchmarks are in open-ended format [9, 22] because of the complexity and extensive resources required to design high-quality multiple-choice questions.

We revisit current evaluation strategies for open-ended

questions, focusing on the challenge of measuring semantic similarity between model-generated and ground-truth answers (§3). Two main evaluation methods exist: rule-based, which evaluates word or phrase level overlap, and model-based, which uses large language models (e.g., GPT-4o) for semantic matching. While rule-based evaluation is computationally efficient, our experiments on the VQAv2 dataset [12] demonstrate that it fails to capture semantic nuances and formatting differences, yielding a poor correlation (0.09) with true VLM performance. In contrast, model-based metrics, though more semantically accurate, are costly and unstable — updates in model versions (e.g., GPT-4o from 0513 to 0806) constantly increase scores by 6% on the MMVet dataset [57], making results incomparable and raising future reproducibility issues.

To mitigate these issues, we propose an alternative for VLM evaluation by converting open-ended VQA questions into multiple-choice format. This shift allows for more objective and reproducible assessment by providing defined options and simplifying answer validation. Nonetheless, creating multiple-choice questions is inherently complex, particularly in generating distractor options that are plausible yet challenging given the correct answer — a process traditionally requiring extensive human expertise and effort to simulate different errors students could make.

We introduce *AutoConverter* (§4), a novel multi-agent system powered by GPT-4o, designed to automatically convert open-ended questions into challenging yet correct multiple-choice questions. To enhance difficulty, specialized distractor proposer agents collaborate with a reviewer agent in an iterative manner to generate a large pool of challenging distractors focused on common error types, including conceptual misunderstandings, visual misinterpretations, and reasoning errors. A selector agent then chooses the most challenging distractors. To ensure correctness, a question correctness evaluator agent assesses the similarity between distractors and the correct answer, providing feedback that allows a refiner agent to adjust distractors to improve question correctness. This agentic pipeline ensures that the generated multiple-choice questions are both cor-

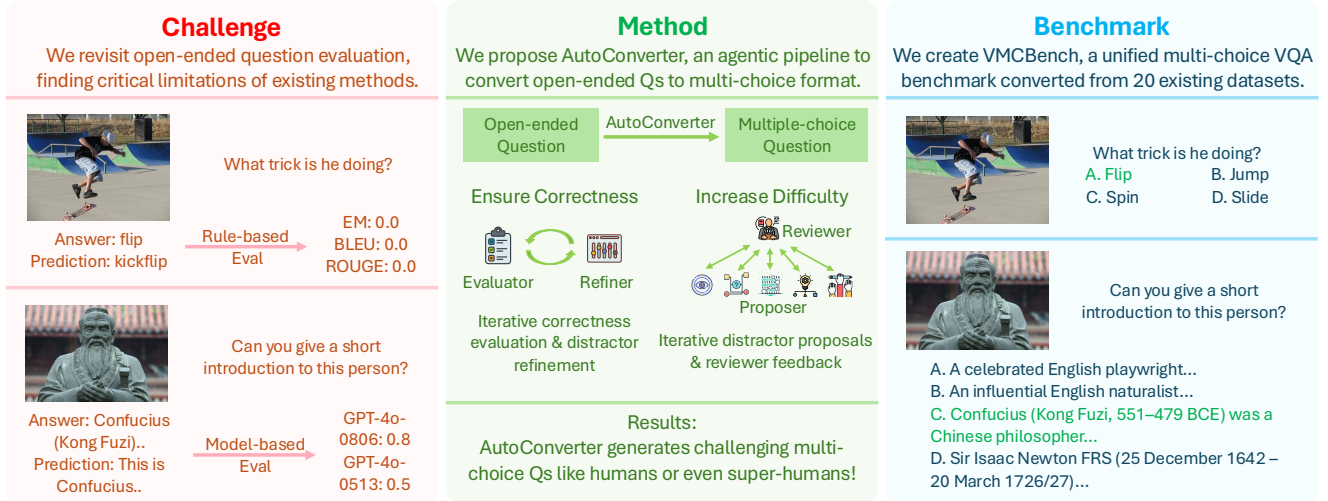


Figure 1. **Overview.** (Left) We analyze existing open-ended VQA evaluation metrics, underscoring their limitations in providing accurate and reproducible assessments. (Middle) We introduce *AutoConverter*, a multi-agent system that automatically converts open-ended questions into multiple-choice format, enabling objective assessment while reducing the costly question creation process. (Right) Using *AutoConverter*, we convert and refine 20 existing VQA datasets into a unified multiple-choice benchmark to support future VLM research.

rect and challenging, delivering strong discriminative power for VLM evaluation.

Our experiments validate *AutoConverter*’s ability to produce correct and challenging multiple-choice questions. When applied to established multiple-choice VQA datasets such as MMMU [58], MathVista [31], and AI2D [18], *AutoConverter* generates distractors that match or even surpass the difficulty of human-crafted alternatives, as evidenced by comparable or lower accuracy scores across various VLMs, with only 3% of questions with the highest correctness score marked incorrect by human annotators. A detailed ablation study further demonstrates that each component of *AutoConverter* significantly enhances the correctness and difficulty of questions compared to naive conversion methods. These results indicate *AutoConverter*’s broad applicability beyond converting open-ended questions to multiple-choice format, such as refining existing multiple-choice datasets to increase difficulty or generating challenging questions for students in an educational context.

Building on *AutoConverter*, we introduce *VMCBench*, a benchmark comprising 9,018 multiple-choice VQA questions compiled from 20 existing datasets (§5). These original datasets, which contain either open-ended or multiple-choice questions, have been converted or refined using *AutoConverter*. Any questions flagged as uncertain by the correctness evaluator were manually verified to ensure quality and accuracy. *VMCBench* provides a standardized benchmark to evaluate diverse model capabilities across various question types. We evaluated 28 state-of-the-art VLMs on *VMCBench*, establishing a new standard for scalable, consistent, and reproducible vision language model evaluation. Our contributions are summarized in Figure 1.

## 2. Related Works

**Vision language model evaluation.** With the rapid development of vision language models (VLMs) [2, 26, 39, 50], numerous benchmarks have emerged to assess various capabilities of VLMs, including general understanding [6, 12, 13, 23, 34, 46, 57], reasoning [16, 30, 31, 52, 55, 58], OCR [38, 48], and document and chart comprehension [18, 19, 35–37]. These datasets typically use either open-ended or multiple-choice formats. Open-ended questions are easier to create but challenging to evaluate accurately, while multiple-choice questions simplify evaluation but demand greater effort in their creation. Previous studies have found limitations of open-ended evaluation for VLMs [5, 20, 33], often focusing on case studies and comparisons with human performance. Their proposed solutions are model-based evaluation methods [10, 63]. In this work, we systematically analyze various VLMs on two widely-used open-ended VQA datasets in zero-shot evaluation settings, quantitatively revealing significant challenges in VLM evaluation for open-ended questions. We demonstrate how rule-based evaluation poorly correlates with actual model rankings in zero-shot scenarios. Additionally, we highlight the substantial limitations of model-based evaluation: model-based evaluations are inconsistent across versions, which leads to serious reproducibility issues. These issues seem intrinsic and difficult to resolve. Consequently, we propose to convert open-ended questions to multiple-choice format in order to streamline reliable VLM evaluation.

**Converting open-ended to multiple-choice.** Converting open-ended questions to multiple-choice requires generating distractors that are both challenging and accurate—a task that traditionally demands extensive human exper-

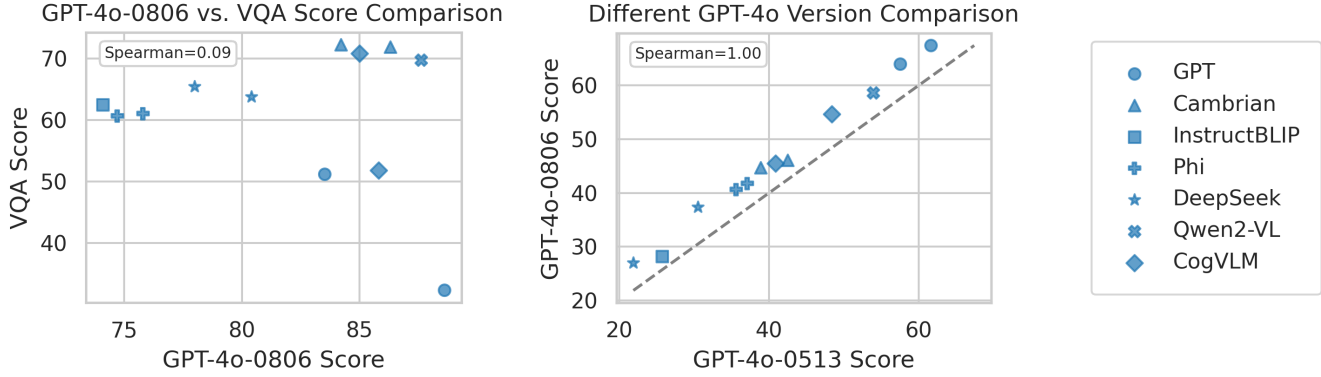


Figure 2. **Challenges in evaluating open-ended questions.** (Left) Rule-based metrics significantly underestimate model performance and penalize models that do not strictly follow the expected format. (Right) Model-based evaluations using two different versions of GPT yield substantially different scores, making comparisons inconsistent and raising reproducibility issues.

tise and has limited the construction of multiple-choice benchmarks. Some approaches attempt to use knowledge graphs [56], retrieval [32], reinforcement learning [29], and supervised learning [8] to automatically produce distractors that resemble correct answers. However, these methods require external resources and often result in relatively simple distractors. Recent advances in large language models (LLMs) have inspired attempts to automate distractor generation to support human creators, yet these approaches still necessitate significant human input to ensure question quality [54, 59]. In our work, we systematically define two essential criteria—correctness and difficulty—for effective multiple-choice questions, and develop an agentic pipeline [27, 42, 47] to automate their creation. Our approach generates multiple-choice questions with distractors that are both challenging and correct, achieving quality comparable to or even surpassing human-crafted questions while significantly reducing the need for human effort. This tool has broad utility: it can convert open-ended questions to multiple-choice for VLM evaluation, refine poorly constructed questions, and generate questions beyond VLM evaluation, such as educational assessments.

### 3. Open-Ended Question Evaluation Challenge

In this section, we discuss the challenges of evaluating vision language models (VLMs) using open-ended questions. The primary issue lies in accurately and robustly measuring the semantic similarity between model-generated answers and ground-truth answers, a long-standing challenge in natural language processing [11, 61, 63].

#### 3.1. Rule-based Evaluation

**Background.** When ground-truth answers are short, rule-based evaluation metrics such as exact match or quasi-exact match (e.g., BLEU [41], ROUGE [24]) are typically used to measure word-level overlap between the ground truth and model predictions. However, these methods are limited in

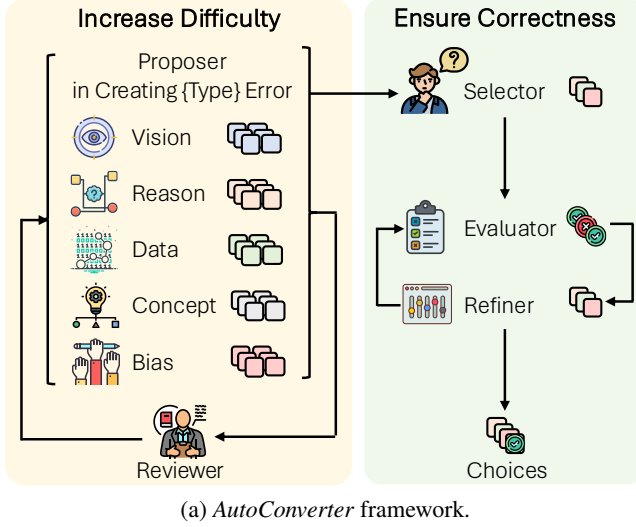
accounting for semantic variations, such as synonyms, paraphrasing, or formatting differences. This limitation is even more pronounced when evaluating current VLMs in zero-shot settings, where models are instruction-tuned to generate slightly more detailed and lengthier responses for clarity.

**Experiment.** To examine the limitations of these evaluation metrics, we tested 12 state-of-the-art VLMs on the widely-used VQAv2 dataset [12]. In a zero-shot setting, we prompted the models with, “Please try to answer the question with short words or phrases if possible.” and used VQAv2’s official rule-based evaluation code. We then conducted a detailed analysis of the cases where GPT-4o did not receive full marks, using human judgment to correct scores where the rule-based evaluation failed. Based on these human annotations, we developed a model-based evaluation method using GPT-4o with a well-crafted prompt to assess the semantic match between model outputs and ground-truth answers. Our model-based metric achieved a 0.95 agreement with human evaluations on VQAv2, making it a reliable proxy for human judgment.

**Findings.** Analysis of GPT-4o’s errors revealed that approximately 66% of the misclassifications were due to evaluation failures rather than the model’s inability to answer the questions. With human corrections, GPT-4o achieved an accuracy of 88% on VQAv2, significantly higher than the 32% accuracy reported by rule-based metrics. A comparison between rule-based and model-based evaluations (i.e., human-proxy evaluation) showed a Spearman correlation of only 0.09, as seen in Figure 2 (left), demonstrating that rule-based metrics provide nearly random and unreliable scores, failing to distinguish between state-of-the-art VLMs.

#### 3.2. Model-based Evaluation

**Background.** When ground-truth answers are longer, reflecting more realistic use cases for VLMs, rule-based metrics become even less effective at capturing semantic equivalence between predictions and answers. Consequently, the community increasingly relies on model-based evalu-



Method	$C$ ( $\uparrow$ )	$M_1$	$M_2$	$M_3$	$M_4$	Avg ( $\downarrow$ )
Human	4.59	33.4	35.2	46.0	52.4	41.8
Naive	4.59	34.2	40.6	50.2	59.0	46.0
w/o Concept	4.66	33.6	37.8	46.4	57.4	43.8
w/o Reason	4.69	30.0	40.2	47.8	56.2	43.5
w/o Vision	4.68	29.0	39.2	45.8	57.2	42.8
w/o Data	4.66	29.4	36.6	47.0	57.8	42.7
w/o Bias	4.65	30.6	37.6	48.6	60.2	44.2
w/o Reviewer	4.64	30.2	38.2	45.4	57.2	42.7
w/o Refiner	4.28	25.0	34.2	42.0	50.0	37.8
<i>AutoConverter</i>	4.69	27.8	37.2	44.2	53.6	40.7

(b) Ablation of *AutoConverter* on MMMU.  $C$  is correctness (higher is better), Avg is average model performance (lower is better).  $M_1$  is PaliGemma-3B,  $M_2$  is LLaVA-1.5-7B,  $M_3$  is Phi-3.5-Vision,  $M_4$  is Qwen2-VL-7B.

Figure 3. ***AutoConverter* framework and results.** (Left) *AutoConverter* is a multi-agent framework with two key steps: increasing difficulty and ensuring the correctness of the converted question. (Right) We perform an ablation study on *AutoConverter* and find that each component is crucial for enhancing question correctness and achieving the desired level of difficulty.

ation [10, 63], which leverages advances in language models like GPT-4o with carefully designed prompts to generate similarity scores between predictions and answers. Although these similarity scores correlate well with human evaluations, changes in model versions can lead to significant shifts in evaluation results.

**Experiment.** To evaluate the stability and robustness of model-based metrics, we tested 12 state-of-the-art VLMs on the MMVet dataset [57], where answer length is 63 words on average. Using two versions of GPT-4o (GPT-4o-0806 and GPT-4o-0513) as evaluators, we kept all other variables constant, such as model responses and evaluation prompts.

**Findings.** While model-based evaluation is reliable in ranking models comparatively, it is costly and introduces notable shifts in absolute scores. As shown in Figure 2 (right), we observed a perfect 1.0 correlation in VLM rankings between the two GPT-4o versions, underscoring the model-based evaluation’s strong alignment with human judgment. However, the absolute scores from GPT-4o-0806 were consistently 6% higher than those from GPT-4o-0513. Further analysis indicated that GPT-4o-0806 tended to assign a score of 1.0 instead of 0.9 for nearly-correct responses, while GPT-4o-0513 showed the opposite tendency. These differences in absolute values hinder the comparability of results across different model versions, significantly impacting the reproducibility of research findings, especially when older versions are deprecated.

#### 4. *AutoConverter*: An Agentic Pipeline Generating Challenging Multi-Choice Questions

Given the challenge of evaluating open-ended questions for vision language models (VLMs) detailed in §3, how can

we mitigate these issues? We propose to convert open-ended questions into a multiple-choice format, capitalizing on the simplicity and objectivity of evaluating multiple-choice questions. However, traditionally, creating multiple-choice questions, especially reasonable yet challenging distractor options, requires substantial human expertise and effort. In this section, we introduce *AutoConverter*, an agentic pipeline that automatically generates high-quality multiple-choice questions from open-ended ones.

##### 4.1. Problem Formulation and Key Desiderata

We define the multiple-choice question conversion process as follows: given an image, a question, and the correct answer  $(v, q, a) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}$ , where  $\mathcal{X}$  and  $\mathcal{Y}$  represent sets of images and texts respectively, we aim to generate a set of distractor choices  $\mathcal{C}_d = \{c_i \in \mathcal{Y} \mid i \in [N]\}$ , where  $N$  is the number of distractors. We define the candidate choice set as  $\mathcal{C} = \mathcal{C}_d \cup \{a\}$ , forming the final multiple-choice question as  $(v, q, \mathcal{C})$ . We choose  $N = 3$  in our work because 4-choice is the most common configuration for multiple-choice questions and minimizes the risk of option selection bias for language models [62].

Two key desiderata are crucial for the multiple-choice conversion process. The first is **correctness**, ensuring that the multiple-choice question is valid with only one correct answer. The second is **difficulty**, ensuring that the question cannot be answered trivially and possesses sufficient discriminative power for test takers. We leverage recent advances in language model agent research, particularly in role-playing [42] and self-reflection techniques [47], to accomplish these goals.



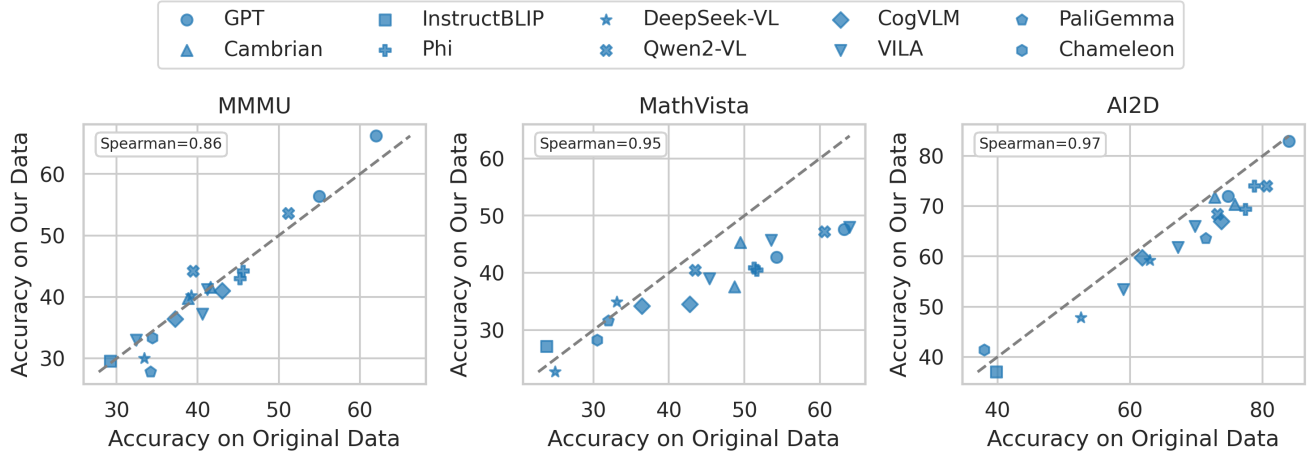


Figure 4. ***AutoConverter* generates challenging multiple-choice questions.** Using *AutoConverter*, we generated distractors for questions and answers from three existing multiple-choice datasets: MMMU, MathVista, and AI2D, and compared them with original human-created distractors. We evaluated various VLMs on both the *AutoConverter*-generated and the original questions, finding that VLMs consistently achieved similar or even lower accuracy on the *AutoConverter*-generated questions compared to the original ones.

## 4.2. Ensuring Correctness

Correctness is essential when converting open-ended questions to a multiple-choice format, defined as having exactly one correct answer. Since our conversion process retains the original open-ended question stem and its answer, we only need to ensure that the other distractors are incorrect given the question and its correct answer. To achieve this, we employ a multi-agent framework to rigorously evaluate distractor options, further enhancing the correctness of the generated multiple-choice questions through iterative refinement.

**Evaluating correctness.** We use GPT-4o as a correctness checker, which evaluates each distractor’s similarity to the correct answer and assigns a 5-point Likert score for the overall correctness of the question. A score of 5 indicates strong confidence in having a single correct answer, while a score of 1 suggests ambiguity or the presence of multiple correct answers. The evaluator was developed based on a small-scale study of correct and incorrect questions. To validate the effectiveness of this evaluator, we conducted large-scale human annotations on 2,400 questions to assess their correctness (details in §5). Our results indicate that 51%, 51%, 63%, 84%, and 95% of questions are deemed correct when assigned correctness scores of 1, 2, 3, 4, and 5, respectively, by our evaluator. Thus, we can use this correctness evaluator as a judge to assess question correctness and refine questions with low correctness scores.

**Refining questions to enhance correctness.** Since the correctness evaluator can accurately flag problematic questions, we refine questions with low correctness scores in an iterative manner to enhance their accuracy. Specifically, when the correctness score of a generated question falls below a threshold of 4, we use GPT-4o as a refiner to adjust distractors based on feedback from the correctness evalua-

tor. This refinement continues until the evaluator’s correctness score meets the required threshold or until a maximum of three refinement rounds is reached.

The entire process is shown in Figure 3a (right). Our ablation studies demonstrate that this process plays a significant role in enhancing the correctness of the questions.

## 4.3. Increase Difficulty

Difficulty is another critical factor in designing effective multiple-choice questions. Our goal is to prevent questions from being trivially answered due to obviously incorrect options or shortcuts, ensuring that each question retains the discriminative power needed to rigorously test VLM capabilities. Inspired by human strategies for creating challenging distractors, we introduce a multi-agent framework powered by GPT-4o. This framework iteratively generates and refines a large set of difficult distractors from multiple perspectives and ultimately selects those that pose the highest level of difficulty.

**Generating a diverse set of distractors.** We first create a large pool of candidate distractors based on common error types to capture a range of potential challenges. We categorize these error types as follows: concept misunderstanding, visual misinterpretation, reasoning error, data processing error, and question bias. Detailed definitions for each error type are provided in the Appendix. For each error type, GPT-4o proposes a set of distractors based on the image, question, and correct answer, along with accompanying rationales. This pool of candidates forms a comprehensive foundation for our selection process.

**Iterative refinement of distractors.** After generating initial distractors, we employ another GPT-4o agent as a reviewer to provide targeted feedback for each distractor. This reviewer assesses the plausibility and challenge level


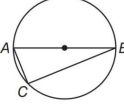

Dataset	Image	Question	Answer	Original Distractors	Naive Distractors	AutoConverter Distractors
MMMU		The painting focuses on the...	idea of Manifest Destiny	<ul style="list-style-type: none"> <li>idea of Pastoral Idealism</li> <li>concept of the sublime in nature</li> <li>concept of untamed wilderness</li> </ul>	<ul style="list-style-type: none"> <li>concept of Romanticism</li> <li>idea of Urbanization</li> <li>theme of Industrial Revolution</li> </ul>	<ul style="list-style-type: none"> <li>representation of untouched landscapes (Similar concept)</li> <li>focus on the sublime in nature (Visually correct)</li> <li>theme of natural beauty and wilderness (Similar concept)</li> </ul>
MathVista		AB is a diameter, AC = 8 inches, and BC = 15 inches. Find the radius of the circle.	8.5	<ul style="list-style-type: none"> <li>7.5</li> <li>8</li> <li>17</li> </ul>	<ul style="list-style-type: none"> <li>7.5</li> <li>12</li> <li>10</li> </ul>	<ul style="list-style-type: none"> <li>23 (Misinterpret <math>AB=AC+BC</math>)</li> <li>17 (Confuse radius and diameter)</li> <li>9.5 (Calculate <math>17/2</math> incorrectly)</li> </ul>
AI2D		From the above food web diagram, which species get directly affected if all rabbit dies	lion	<ul style="list-style-type: none"> <li>grass</li> <li>cricket</li> <li>mouse</li> </ul>	<ul style="list-style-type: none"> <li>deer</li> <li>cricket</li> <li>frog</li> </ul>	<ul style="list-style-type: none"> <li>owl (Not directly affected)</li> <li>snake (Follow the wrong arrow)</li> <li>deer (Visually close to rabbit)</li> </ul>

Figure 5. **Qualitative comparison of the original questions, naive baseline-generated questions, and *AutoConverter*-generated questions.** *AutoConverter* simulates errors from different perspectives and produces correct and challenging multiple-choice questions.

of each distractor given the context of the question and answer, offering suggestions for improvement. The specialized proposer agent then refines the distractors based on this feedback, ensuring that each distractor is as challenging as possible and removing those considered too close to the correct answer to maintain correctness.

**Selecting high-difficulty distractors.** From the refined pool of distractors with their creation rationales, we finally use GPT-4o as a selector to evaluate each candidate’s difficulty and correctness, selecting the most challenging distractors along with justifications for their selection. This process results in a final set of high-quality distractors that rigorously test VLM performance.

This process is illustrated in Figure 3a (left). Our detailed ablation studies demonstrate that each of these agents contributes to improving the difficulty of the questions.

#### 4.4. AutoConverter Results

To evaluate the effectiveness of *AutoConverter* in generating high-quality multiple-choice questions, we leverage three existing multiple-choice VQA datasets: MMMU [58], MathVista [31], and AI2D [18]. We retain all questions with four choices and regenerate the distractors based on the image, question, and correct answer. We then compare *AutoConverter*-generated distractors with the original human-crafted ones by evaluating various VLMs on both the original and converted datasets.

***AutoConverter* generates correct multiple-choice questions.** Using a correctness evaluator and refiner pipeline, we ensure the accuracy of the generated questions. Our analysis shows that the generated questions maintain a correctness level comparable to that of the original human-crafted questions. After filtering to keep only questions with a correctness score of 5, large-scale human annotation indicates that only 3% of questions are marked as incorrect for MMMU, MathVista, and AI2D. Among these, 52% of errors are due to incorrect

original answers, while only 48% are introduced by the *AutoConverter* process. These results demonstrate that most *AutoConverter*-generated questions are accurate and suitable for reliable VLM evaluation.

***AutoConverter* generates challenging multiple-choice questions.** As shown in Figure 4, *AutoConverter* produces highly challenging questions, with a broad range of VLMs achieving similar or even lower accuracy on the generated distractors compared to the original human-crafted ones across MMMU, MathVista, and AI2D. Notably, MMMU is regarded as one of the most challenging VQA datasets, as it is sourced from exams and textbooks. These results highlight *AutoConverter*’s capability not only in converting open-ended questions to multiple-choice but also in refining existing multiple-choice questions to enhance their difficulty. Furthermore, *AutoConverter* could have applications beyond VLM evaluation, such as generating challenging questions for educational purposes.

**Each agent in *AutoConverter* contributes to jointly improving correctness and difficulty.** To analyze each agent’s role in *AutoConverter*, we conduct an ablation study by removing each component, as shown in Table 3b. First, removing specialized error proposers results in a decrease in difficulty, with an average increase of 1.6% in relative accuracy. Next, omitting the reviewer for iterative distractor refinement leads to a 4.9% relative increase in average VLM performance, indicating lower difficulty. Finally, removing the question evaluator and refiner, which serve as correctness safeguards, causes an 8.7% relative decrease in correctness score. We also compare *AutoConverter* with a naive baseline that uses the prompt “Please generate 3 distractors given the question, answer, and image.” As shown in Table 3b, *AutoConverter* significantly outperforms this baseline in both correctness (2.2% relative increase) and difficulty (11.5% relative decrease in average VLM performance), demonstrating the effectiveness of our approach.

We provide qualitative comparisons of original and *Auto-*

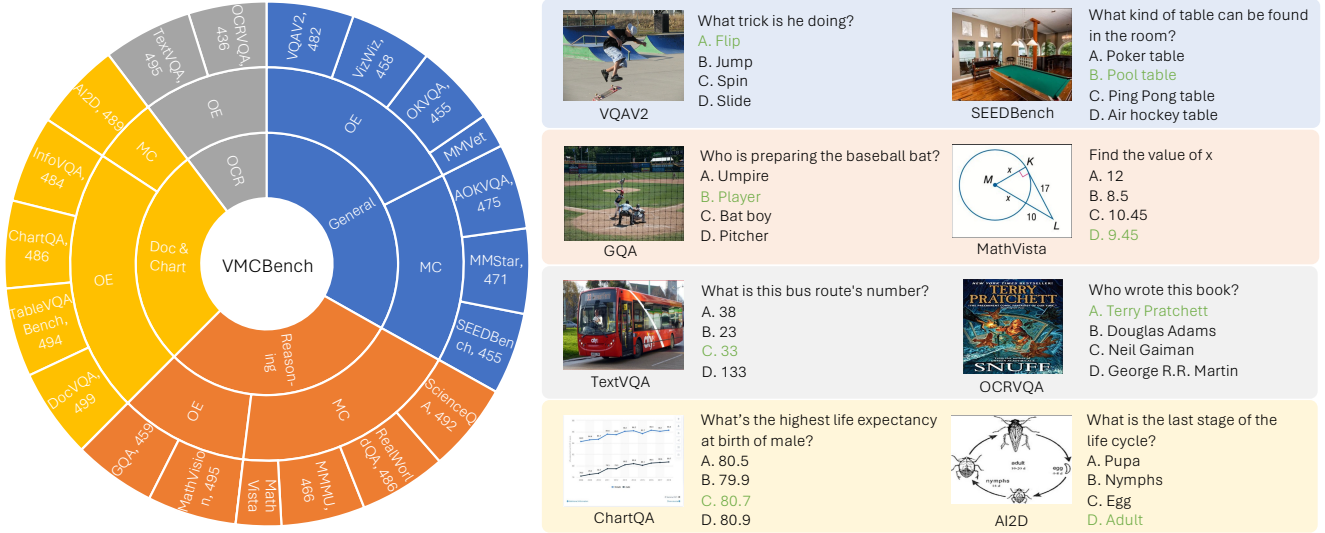


Figure 6. **VMCBench overview.** (Left) *VMCBench* is constructed by converting 12 open-ended (OE) and refining 8 multiple-choice (MC) VQA datasets into a unified multiple-choice format, with human validation ensuring correctness. The number of questions per dataset is listed. (Right) Example questions from *VMCBench*, showcasing diverse question types across multiple domains.

toConverter-generated questions, as well as the naive baseline, in Figure 5, which demonstrate the effectiveness of *AutoConverter* in creating hard yet correct questions.

## 5. VMCBench: A Unified Multiple-Choice Visual Question Answering Benchmark

Using our *AutoConverter* method, detailed in §4, we introduce *VMCBench* — a benchmark that unifies 20 existing visual question answering (VQA) datasets into a consistent multiple-choice format. *VMCBench* spans a diverse array of visual and linguistic contexts, rigorously testing various model capabilities. By transforming open-ended questions into multiple-choice format, *VMCBench* enhances vision language model evaluation by mitigating ambiguities while preserving the complexity of the tasks. This benchmark provides a reliable and reusable resource for the evaluation of future vision language models (VLMs).

### 5.1. Benchmark Overview

*VMCBench* transforms 20 widely-used VQA datasets into a unified multiple-choice benchmark. These datasets can be broadly categorized to assess general capabilities of VLMs (VQAV2 [12], OKVQA [34], MMVet [57], VizWiz [13], A-OKVQA [46], MMStar [6], SEEDBench [23]), reasoning capabilities (MathVision [52], GQA [16], MMMU [58], RealWorldQA [55], MathVista [31], ScienceQA [30]), OCR tasks (OCRQA [38], TextVQA [48]), and document and chart understanding (DocVQA [36], InfoVQA [37], ChartQA [35], TableVQABench [19], AI2D [18]).

Among these, 12 datasets consist of open-ended questions, which we convert into the multiple-choice format to facilitate evaluation. The remaining 8 datasets are already in

multiple-choice format; for these, we apply *AutoConverter* to refine distractors, increasing their difficulty. These refined datasets can further test model robustness and help identify dataset contamination [14, 40, 45, 60].

For each dataset, we randomly sampled up to 500 questions, as recent studies suggest this sample size may suffice for evaluating model performance [43, 44], resulting in a total of 9,450 questions across the 20 datasets. We apply *AutoConverter* to these 9,450 questions.

To ensure the correctness of the converted questions, *AutoConverter* includes an internal correctness evaluator that assigns a correctness score to each converted question, ensuring only one correct answer exists. Out of the 9,450 converted questions, 103, 162, 399, 635, and 8,151 questions received correctness scores of 1, 2, 3, 4, and 5, respectively. Six PhD student annotators reviewed all questions with correctness scores below 5 and randomly sampled 1,101 questions with a score of 5, resulting in a total of 2,400 annotated questions. The correctness rates for questions with scores of 1, 2, 3, 4, and 5 were 51%, 51%, 63%, 84%, and 95%, respectively. This high accuracy of the correctness evaluation supports the reliability of questions with a confidence score of 5 for future dataset construction. Additionally, we annotated the source of errors, distinguishing between errors from the ground-truth answers (original dataset errors) and those introduced by *AutoConverter*. Notably, 59% of errors were due to issues in the ground-truth answers, indicating that if all original open-ended questions were correct, *AutoConverter* would introduce only 2% errors in total for questions with a correctness score of 5. We removed all questions that were identified as incorrect, resulting in a final total of 9,018 questions.



Model	General Reasoning OCR Doc&Chart Avg				
GPT-4o	85.2	68.1	96.6	83.3	80.8
Qwen2-VL-7B	84.7	64.8	96.5	80.2	78.8
Claude-3.5-Sonnet	81.0	62.6	93.3	85.7	78.1
Cambrian-34B	84.3	65.2	95.1	73.6	76.9
Gemini-1.5-Pro	79.8	65.3	92.2	72.2	74.8
GPT-4o-Mini	81.1	61.1	94.2	75.0	74.9
VILA1.5-40B	83.4	64.7	92.9	67.9	74.7
CogVLM2-19B	79.2	56.8	92.1	73.6	72.4
Qwen2-VL-2B	78.5	57.1	93.0	73.0	72.2
Phi-3-Vision	74.8	57.2	90.5	74.3	71.1
Cambrian-13B	80.0	56.0	92.5	67.0	70.7
Cambrian-8B	78.9	57.5	89.7	65.1	70.0
Idefics2-8B	78.6	58.4	92.6	62.0	69.6
Phi-3.5-Vision	72.3	57.0	86.3	68.7	68.3
VILA1.5-13B	75.8	55.3	85.1	50.8	63.9
DeepSeek-VL-7B	73.9	54.2	85.7	53.1	63.7
CogVLM-17B	73.7	49.5	77.9	55.1	62.0
VILA1.5-8B	73.4	51.8	81.4	47.0	60.7
Gemini-1.5-Flash	66.8	53.6	77.7	52.1	60.1
PaliGemma-3B	73.3	53.2	55.0	52.1	59.7
VILA1.5-3B	71.6	48.3	78.4	42.5	57.5
DeepSeek-VL-1.3B	70.0	44.9	79.4	44.1	56.6
LLaVA1.5-13B	67.6	47.2	74.8	37.7	54.2
LLaVA1.5-7B	65.3	46.9	73.3	35.2	52.6
Chameleon-30B	53.5	40.7	52.4	35.2	44.6
InstructBLIP-7B	56.4	36.4	48.2	30.1	42.5
InstructBLIP-13B	55.4	35.7	49.5	26.4	41.1
Chameleon-7B	42.6	36.2	39.3	31.3	37.3

Table 1. Performance of 28 VLMs on *VMCBench*.

*VMCBench* spans a diverse range of visual and linguistic contexts and includes a broad array of question types, offering a comprehensive framework to evaluate VLMs across multiple vision language tasks. It provides a unified interface for efficient and accurate assessment of VLM capabilities. The question distribution and examples of converted questions are shown in Figure 6.

## 5.2. Evaluation Results

We evaluated 28 state-of-the-art vision language models (VLMs) on the *VMCBench* to assess their performance, including GPT-4 [39], Gemini-1.5 [50], Claude-3.5 [2], Qwen2-VL [53], Cambrian [51], VILA [25], CogVLM [15], Phi [1], Idefics2 [21], DeepSeek-VL [28], PaliGemma [3], LLaVA1.5 [26], Chameleon [49], and InstructBLIP [7]. The evaluation is conducted in the zero-shot setting with prompt “Question: {question} Options: A. {A} B. {B} C. {C} D. {D} Answer with the option’s letter from the given choices directly.” The results are summarized in Table 1. Detailed results are in Appendix.

Our findings reveal several key insights:

**Private models remain the top performers, but the gap is narrowing.** As shown in Table 1, the best-performing model on *VMCBench* is GPT-4o, achieving 80.6% overall accuracy. However, the gap between GPT-4o and the second-best model, Qwen2-VL-7B, is only 2.0%. This open-source model demonstrates that the performance difference between private and public models is decreasing.

**Rapid advances in VLM development.** Another notable trend is the significant improvement from InstructBLIP to Qwen2-VL-7B, nearly doubling in performance (40.9% vs. 78.6%). These models were released in 2023 and 2024, respectively, illustrating the rapid pace of progress in VLMs.

**Scaling up models generally improves performance.** As expected, scaling up model sizes typically leads to performance gains. For instance, across different model families (such as VILA, Cambrian, Chameleon, DeepSeek, and LLaVA), we observe a clear improvement with larger models, similar to scaling trends in language models [4, 17].

**Evaluating across diverse datasets is essential to uncovering model limitations.** A comparison between Phi-3 and Phi-3.5 reveals that while Phi-3.5 achieves better performance on some reasoning datasets such as MMMU and MathVista, it performs significantly worse on OCR and document/chart understanding tasks, which reduces its overall performance. This underscores the importance of a holistic evaluation of models across a broad range of datasets. Previously, this was challenging due to variations in dataset formats and evaluation protocols. *VMCBench* addresses this by unifying the task into a multiple-choice format, thus reducing evaluation costs and standardizing comparisons.

**Continuous feature inputs outperform discrete tokens.** Chameleon is the only model in our experiments that utilizes VQGAN discrete tokens, yet it performs significantly worse than smaller models trained with continuous features using less computational resources. This observation raises an intriguing question: why might VQGAN tokens be less effective for image understanding tasks?

We believe that *VMCBench* is a valuable benchmark that will play a significant role in advancing VLM development by facilitating simple, fast, and accurate model evaluation.

## 6. Conclusion

We address limitations in open-ended visual question answering benchmarks for evaluating vision language models (VLMs) by introducing *AutoConverter*, a multi-agent system that transforms questions into multiple-choice format with challenging distractors. *AutoConverter* effectively creates questions that match or exceed the difficulty of human-written distractors, with VLMs often achieving similar or lower accuracy. Building on this, *VMCBench* provides a benchmark of 9,018 multiple-choice questions, setting a new standard for consistent and rigorous VLM evaluation.



## References

- [1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 8
- [2] Anthropic. Introducing the next generation of claude, 2024. 1, 2, 8
- [3] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024. 8
- [4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020. 8
- [5] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *ICML*, 2024. 2
- [6] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? In *NeurIPS*, 2024. 2, 7
- [7] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023. 8
- [8] Wenjian Ding, Yao Zhang, Jun Wang, Adam Jatowt, and Zhenglu Yang. Can we learn question, answer, and distractors all from an image? a new task for multiple-choice visual question answering. In *COLING*, 2024. 3
- [9] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *MM*, 2024. 1
- [10] Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. AlpacaFarm: A simulation framework for methods that learn from human feedback. *arXiv preprint arXiv:2305.14387*, 2023. 2, 4
- [11] Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. Summeval: Re-evaluating summarization evaluation. *TACL*, 2021. 3
- [12] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. 1, 2, 3, 7
- [13] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, 2018. 2, 7
- [14] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadam, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 2021. 7
- [15] Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. CogVLM2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024. 8
- [16] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 2, 7
- [17] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 8
- [18] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016. 2, 6, 7
- [19] Yoonsik Kim, Moonbin Yim, and Ka Yeon Song. TableVQA-bench: A visual question answering benchmark on multiple table domains. *arXiv preprint arXiv:2404.19205*, 2024. 2, 7
- [20] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024. 2
- [21] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024. 8
- [22] Tony Lee, Haoqin Tu, Chi Heem Wong, Wenhao Zheng, Yiyang Zhou, Yifan Mai, Josselin Somerville Roberts, Michihiro Yasunaga, Huaxiu Yao, Cihang Xie, et al. Vhelm: A holistic evaluation of vision language models. *arXiv preprint arXiv:2410.07112*, 2024. 1
- [23] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *CVPR*, 2024. 2, 7
- [24] CY LIN. Rouge: A package for automatic evaluation of summaries. In *ACL Workshop*, 2004. 3
- [25] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *CVPR*, 2024. 8
- [26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1, 2, 8
- [27] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. In *ICLR*, 2024. 3
- [28] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 8

- [29] Jiaying Lu, Xin Ye, Yi Ren, and Yezhou Yang. Good, better, best: Textual distractors generation for multiple-choice visual question answering via reinforcement learning. In *CVPR*, 2022. 3
- [30] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafford, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *NeurIPS*, 2022. 1, 2, 7
- [31] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2024. 2, 6, 7
- [32] Haohao Luo, Yang Deng, Ying Shen, See-Kiong Ng, and Tat-Seng Chua. Chain-of-exemplar: enhancing distractor generation for multimodal educational question generation. In *ACL*, 2024. 3
- [33] Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. Improving automatic vqa evaluation using large language models. In *AAAI*, 2024. 2
- [34] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019. 1, 2, 7
- [35] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *ACL Findings*, 2022. 2, 7
- [36] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, 2021. 7
- [37] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *WACV*, 2022. 2, 7
- [38] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, 2019. 2, 7
- [39] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 2, 8
- [40] Yonatan Oren, Nicole Meister, Niladri S Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. Proving test set contamination in black-box language models. In *ICLR*, 2024. 7
- [41] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 3
- [42] Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *UIST*, 2023. 3, 4
- [43] Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein Dor, Eyal Shnarch, Noam Slonim, Michal Shmueli-Scheuer, and Leshem Choshen. Efficient benchmarking (of language models). In *NAACL*, 2024. 7
- [44] Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples. In *ICML*, 2024. 7
- [45] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019. 7
- [46] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *ECCV*, 2022. 1, 2, 7
- [47] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *NeurIPS*, 2024. 3, 4
- [48] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, 2019. 2, 7
- [49] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 8
- [50] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1, 2, 8
- [51] Shengbang Tong, Ellis L Brown II, Penghao Wu, Sanghyun Woo, ADITHYA JAIRAM IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *NeurIPS*, 2024. 8
- [52] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv preprint arXiv:2402.14804*, 2024. 2, 7
- [53] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 8
- [54] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024. 3
- [55] xAI. Realworldqa dataset, 2024. 2, 7
- [56] Han-Cheng Yu, Yu-An Shih, Kin-Man Law, Kai-Yu Hsieh, Yu-Chen Cheng, Hsin-Chih Ho, Zih-An Lin, Wen-Chuan Hsu, and Yao-Chung Fan. Enhancing distractor generation for multiple-choice questions with retrieval augmented pre-training and knowledge graph integration. *arXiv preprint arXiv:2406.13578*, 2024. 3
- [57] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *ICML*, 2024. 1, 2, 4, 7
- [58] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024. 1, 2, 6, 7
- [59] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Ming Yin, Botao

- 774 Yu, Ge Zhang, et al. Mmmu-pro: A more robust multi-  
775 discipline multimodal understanding benchmark. *arXiv*  
776 *preprint arXiv:2409.02813*, 2024. 3
- 777 [60] Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine  
778 Wu, Will Song, Tiffany Zhao, Pranav Raja, Dylan  
779 Slack, Qin Lyu, et al. A careful examination of large lan-  
780 guage model performance on grade school arithmetic. *arXiv*  
781 *preprint arXiv:2405.00332*, 2024. 7
- 782 [61] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang,  
783 Kathleen McKeown, and Tatsunori B Hashimoto. Bench-  
784 marking large language models for news summarization.  
785 *TACL*, 2024. 3
- 786 [62] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer  
787 Singh. Calibrate before use: Improving few-shot perfor-  
788 mance of language models. In *ICML*, 2021. 4
- 789 [63] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan  
790 Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan  
791 Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with  
792 mt-bench and chatbot arena. *NeurIPS*, 2023. 2, 3, 4