

UPME: An Unsupervised Peer Review Framework for Multimodal Large Language Model Evaluation

Qihui Zhang^{1,3*} Munan Ning^{1*} Zheyuan Liu^{1,3} Yue Huang⁴ Shuo Yang¹
 Yanbo Wang¹ Jiayi Ye¹ Xiao Chen⁵ Yibing Song^{2,3} Li Yuan^{1†}

¹School of Electrical and Computer Engineering, Peking University ²Hupan Lab

³DAMO Academy, Alibaba Group ⁴University of Notre Dame ⁵Tsinghua University

Abstract

Multimodal Large Language Models (MLLMs) have emerged to tackle the challenges of Visual Question Answering (VQA), sparking a new research focus on conducting objective evaluations of these models. Existing evaluation methods face limitations due to the significant human workload required to design Q&A pairs for visual images, which inherently restricts the scale and scope of evaluations. Although automated MLLM-as-judge approaches attempt to reduce the human workload through automatic evaluations, they often introduce biases. To address these problems, we propose an **Unsupervised Peer review MLLM Evaluation** framework. It utilizes only image data, allowing models to automatically generate questions and conduct peer review assessments of answers from other models, effectively alleviating the reliance on human workload. Additionally, we introduce the vision-language scoring system to mitigate the bias issues, which focuses on three aspects: (i) response correctness; (ii) visual understanding and reasoning; and (iii) image-text correlation. Experimental results demonstrate that UPME achieves a Pearson correlation of 0.944 with human evaluations on the MMstar dataset and 0.814 on the ScienceQA dataset, indicating that our framework closely aligns with human-designed benchmarks and inherent human preferences.

1. Introduction

Recently, Large Language Models (LLMs) have made significant strides in reasoning and application capabilities [3, 32, 47], enabling diverse applications such as text and code generation [1, 19, 40, 48]. Notably, due to the exceptional natural language understanding capabilities of LLMs, they are increasingly employed for automated evaluations, a pro-

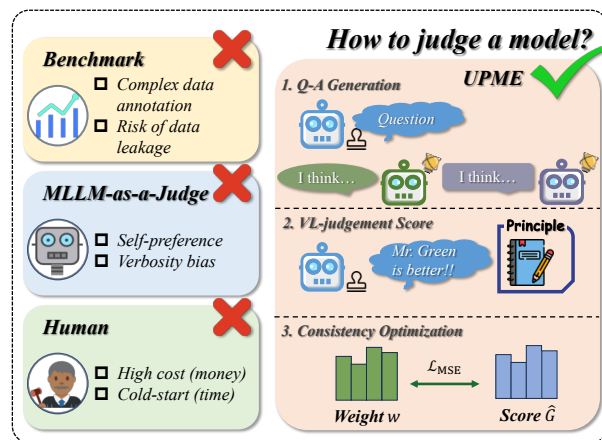


Figure 1. Existing methods for evaluating MLLMs face various challenges. Our proposed UPME framework addresses these limitations by leveraging a peer review mechanism, reducing annotation costs, and aligning closely with human judgment.

cess known as LLM-as-a-Judge [55]. With the integration of vision encoders [36], Multimodal Large Language Models (MLLMs) [24, 32] can incorporate multiple modalities (e.g., text, image, and audio) [23, 56] and showcase remarkable performance in downstream applications, including visual conversation [7] and embodied intelligence [16]. Furthermore, they can act as agents to interact with GUI [9, 18].

Subsequently, the challenge of conducting objective and comprehensive evaluations of MLLMs while providing users with authoritative guidelines has emerged as a key research focus. A common approach involves constructing benchmarks based on Visual Question Answering (VQA), which assess the performance of MLLMs across various dimensions by setting image-based questions and answers [10, 28, 52, 53]. However, most of these works heavily depend on substantial human involvement in creating Q&A pairs for images [21, 44], leading to limited evaluation volume and making it insufficient for a comprehen-

* Equal contribution † Corresponding author
 {maskhui1003, munanning}@gmail.com
 yuanli-ece@pku.edu.cn

sive assessment of visual content [6]. Another category of methods is MLLM-as-a-Judge [17, 20, 26, 55], which aims to reduce reliance on human annotations by allowing models to evaluate each other directly. However, these methods often still depend on human-designed questions and cannot entirely eliminate the need for human input. Additionally, a significant challenge is the presence of verbosity and self-preference biases [8, 51], where MLLMs tend to favor longer responses or prefer their own outputs. These biases can lead mutual evaluation mechanisms to deviate from truly understanding visual content.

To address these challenges, we propose UPME, the Unsupervised Peer review MLLM Evaluation framework, which is designed to evaluate MLLMs without requiring human QA annotations, as shown in Figure 1. During each iteration, UPME selects two candidate models and a review model from the MLLM pool. The review model generates questions for a given image and evaluates the responses provided by the candidate models. This evaluation is conducted using a vision-language scoring system that assesses textual response correctness, visual understanding and reasoning, as well as image-text consistency through CLIP scores. Each model is initialized with a confidence weight and alternates roles as candidate and review models. Their estimated scores are calculated based on weighted evaluations provided by the review models. The framework employs dynamic weight optimization, iteratively updating model scores and refining weights through consistency optimization. Ultimately, this process generates a comprehensive and less-biased estimated score list.

We selected the MMStar [10] and ScienceQA [28] datasets as benchmarks because they encompass visual content reasoning and multimodal content across various disciplines. Experimental results show that under unsupervised settings, UPME achieved Pearson correlations of 0.944 and 0.814 with human evaluation results on these two datasets, while Spearman similarities reached 0.972 and 0.886, indicating a high degree of similarity between our approach and human-annotated QA benchmarks. Further experiments demonstrated that UPME achieved higher consistency with human assessments compared to purely peer review-based methods, effectively reducing verbosity and self-preference issues in the MLLM-as-a-Judge framework.

Overall, our contributions can be summarized as follows:

- We propose the first Unsupervised Peer review MLLM Evaluation (UPME) framework, addressing the challenge of reliance on human annotations in MLLM evaluation
- Our framework incorporates a Vision-Language Scoring System that targets visual performance and image-text association, providing a visually-focused evaluation.
- We conduct extensive experiments demonstrating that UPME achieves high consistency with human QA assessments and strong alignment with human preferences.

2. Related Work

2.1. MLLM Benchmark

MLLM benchmarks rely on human-annotated QA pairs to measure the accuracy of MLLMs in answering questions. They focus on the different capabilities of MLLMs. ScienceQA [28] evaluate the multi-turn reasoning ability for a broad range of scientific subjects by introducing Chain of Thought (CoT) [47]; MMVet [52] presents an LLM-based open-ended output evaluation method, enabling unified scoring across various question types and answer formats; MMMU [53], inspired by MMLU, primarily assesses model performance in complex reasoning tasks requiring college-level knowledge, covering six major disciplines and 30 fields with multimodal questions; MMStar [10] focuses on creating a benchmark requiring visual content for reasoning, rigorously selecting visual-dependent samples and detecting data leakage to ensure genuine performance gains from multimodal training; RealWorldQA [49] is designed to assess the ability of MLLMs to understand real-world space. Although these benchmarks propose various evaluation methods, their evaluation scope is limited by the reliance on substantial human annotation.

2.2. MLLM-as-a-Judge

To reduce the human annotation workload, LLM-as-a-Judge [55] is proposed to enable models to perform mutual evaluation. This approach primarily involves using LLMs to assess responses in two ways: comparing pairs of answers to determine superiority [20, 25, 55], or directly scoring individual answers based on specific criteria [17, 25]. This strategy is extended to MLLM-as-a-Judge [8] to assess the performance of MLLMs across multiple evaluation tasks. MJ-Bench [11] evaluates the effectiveness of multimodal reward models in text-to-image generation. However, these methods face two main biases [38, 51]: the verbose bias and self-preference bias, which means MLLMs prefer the answers from themselves or with longer text length. Human preferences have also not been effectively incorporated into these automated evaluation systems [31].

2.3. Human Preference-based Evaluation

Predefined image and QA-pair benchmarks for MLLMs may encounter issues like data leakage and targeted optimization, which can prevent an accurate reflection of the true capabilities of these models [15, 50]. Therefore, evaluation methods based on human preferences offer a better assessment of whether MLLMs meet human needs [30]. Chatbot Arena [12] provides a public platform specifically for evaluating LLMs based on human preference. This platform uses a pairwise comparison approach, where users interact with two anonymous models and select the one that performs better, generating a dynamic leaderboard via the

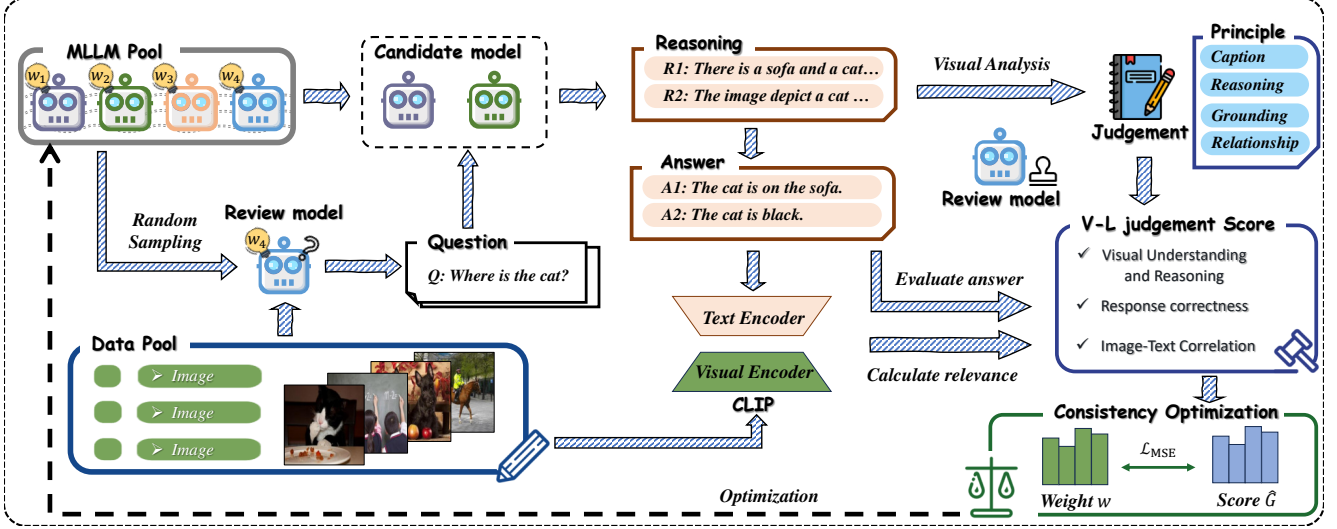


Figure 2. The UPME framework consists of three main components: (i) Peer Review Mechanism, where two candidate models and one review model are randomly selected from the MLLM pool. The review model generates questions based on a selected image, and candidate models provide responses. (ii) Vision-Language Judgment Scoring System, which evaluates answers based on textual correctness, visual understanding and reasoning, and image-text correlation. (iii) Dynamic Weight Optimization, ensuring consistency between confidence weights and estimated scores through iterative optimization cycles.

Elo rating system. Although open-ended user evaluations can more realistically reflect LLM performance in real-world scenarios [46], the high economic cost of collecting large volumes of high-quality user feedback, combined with the long wait time for obtaining such feedback, limits the widespread adoption of this approach.

3. Methodology

Assigning higher weights to better-performing models leads to a significant improvement in the accuracy of the entire evaluation framework. This conclusion is supported by preliminary experiments detailed in subsection 8.1 and confirmed by recent research [13, 31, 41, 43]. Therefore, we focus on characterizing the relationship between model performance and its judging capability through consistency optimization.

3.1. Peer review formulation

Problem definition. The UPME framework facilitates mutual evaluation among MLLMs via a peer review mechanism, aiming to produce evaluation scores that closely reflect human preferences. Given an unsupervised image dataset $\mathcal{I} = \{I_i\}_{i=1}^n$ of n images (with no human annotated questions and answers) and a model pool $\mathcal{M} = \{M_j\}_{j=1}^m$ consisting of m MLLMs (both open-source and closed-source), we can obtain an UPME-learned score list $\hat{G}$:

$$\hat{G} := [\hat{G}_{M_1}, \hat{G}_{M_2}, \hat{G}_{M_3}, \dots, \hat{G}_{M_m}]. \quad (1)$$

On the other hand, given G_{M_1} represents the score of

model M_1 on the human-annotated benchmark, we define the human-annotated score list G as:

$$G := [G_{M_1}, G_{M_2}, G_{M_3}, \dots, G_{M_m}]. \quad (2)$$

Our objective is to optimize the UPME-learned scores $\hat{G}$ by maximizing their similarity to the human-annotated score list G , i.e., $\text{Sim}(G, \hat{G}) \rightarrow 1$, ensuring that UPME-learned scores align closely with human benchmark standards.

Overview of UPME. Figure 2 shows an overview of our UPME framework. During each review iteration, we randomly select two different candidate models M_j and M_k from the pool $\mathcal{M}$ for comparison, along with another model M_r to serve as the reviewer. M_r generates a question $Q_i^r = M_r(I_i)$ for each image I_i and then prompts the candidate models M_j and M_k to provide responses $A_i^{j,r}$ and $A_i^{k,r}$, respectively. The answers can be written as:

$$A_i^{j,r} = M_j(I_i, Q_i^r), \quad A_i^{k,r} = M_k(I_i, Q_i^r). \quad (3)$$

Based on the vision-language judgment scoring System S_{VL} (detailed in subsection 3.2), M_r then evaluates these responses, and compares the responses of M_j and M_k to produce a review score $\text{Review}_i^{j,k,r}$ as:

$$\text{Review}_i^{j,k,r} = S_{VL}(A_i^{j,r}, A_i^{k,r}, Q_i^r, I_i | M_r). \quad (4)$$

By assigning each model an initiated confidence weight w , we can calculate the learned score $\hat{G}_{M_j}$ of each MLLM:

$$\hat{G}_{M_j} = \sum_i \sum_{k \neq j} \sum_{r \neq k, r \neq j} \text{Review}_i^{j,k,r} \times w_r, \quad (5)$$

then we can subsequently obtain the UPME-learned score list $\hat{G}$, as illustrated in Equation 1.

3.2. Vision-Language Judgment Scoring System

In this subsection, we present our vision-language judgment scoring system, denoted as S_{VL} in Equation 9. Our evaluation is structured around three core criteria: Response Correctness, Visual Understanding and Reasoning, and Image-Text Correlation. The first two aspects are evaluated by employing M_r , which reviews responses based on prompts outlined in Section 10. The Image-Text Correlation is measured using CLIP scores.

Response Correctness. The primary criterion is the correctness of the candidate models' responses to questions posed by the review model. Existing studies indicate that paired comparative evaluation yields better accuracy than isolated assessments [26, 27], so we adopt a pairwise scoring approach. Based on the review model M_r 's evaluation Judge_i^r for candidates M_j and M_k , we calculate the model's score for Response Accuracy $S_{\text{Correct}}(j)$ as follows:

$$S_{\text{Correct}}(j) = \begin{cases} 1 & \text{if } \text{Judge}_{\text{Correct},j} > \text{Judge}_{\text{Correct},k} \\ 0.5 & \text{if } \text{Judge}_{\text{Correct},j} = \text{Judge}_{\text{Correct},k} \\ 0 & \text{if } \text{Judge}_{\text{Correct},j} < \text{Judge}_{\text{Correct},k} \end{cases} \quad (6)$$

where $\text{Judge}_{\text{Correct}}$ represents the review model's assessment of response correctness. If $\text{Judge}_{\text{Correct},j} > \text{Judge}_{\text{Correct},k}$, it indicates that model M_j is judged superior in response accuracy to M_k , scoring 1 point; a tie yields 0.5 points, and a lower score results in 0 points.

Visual Understanding and Reasoning. Inspired by the significant CoT ability demonstrated in the o1 model [34] through adversarial training, we leverage M_r to evaluate the interpretative and analytical capabilities of the model with respect to visual content.

$S_{\text{Visual}}(j)$ is computed using the function Γ , which evaluates the response across four dimensions: caption, reasoning, grounding, and relationship, as follows:

$$S_{\text{Visual}}(j) = \Gamma(\text{cap., rea., gro., obj.} \mid A_j, M_r), \quad (7)$$

where captioning evaluates the ability to generate precise descriptions, reasoning measures logical consistency, grounding assesses accurate object localization, and relationship captures the connections between subjects. Within this structure, the review model M_r provides a combined

score via Γ , yielding $S_{\text{Visual}}(j)$. $S_{\text{Visual}}(j)$ reflects a model's capability to go beyond basic VQA, integrating both description and reasoning to demonstrate a deeper level of visual comprehension.

Image-Text Correlation. To capture the alignment between image content and textual responses, we introduce the CLIP [37] score:

$$S_{\text{Clip}}(j) = \text{CLIP}(I_i, A_i^{j,r}), \quad (8)$$

where $S_{\text{Clip}}(j)$ measures the correlation between image I_i and response $A_i^{j,r}$ from model M_j . The details are described in subsection 9.2.

Combining the above components, we compute the final vision-language judgment score as a weighted sum of the three criteria: Response Accuracy, Visual Understanding and Reasoning, and Image-Text Correlation:

$$S_{VL} = \gamma_1 S_{\text{Correct}} + \gamma_2 S_{\text{Visual}} + \gamma_3 S_{\text{Clip}}, \quad (9)$$

where γ_1 , γ_2 , and γ_3 are weights reflecting the relative importance of each component in the overall evaluation.

3.3. Dynamic Weight Optimization

In each iteration, we use Mean Squared Error (MSE) as the loss function to update the weights w , enhancing the consistency between the weights w and the estimated scores $\hat{G}$:

$$\mathcal{L}_{\text{MSE}}(\hat{G}, w) = \frac{1}{m} \sum_{j=1}^m \left(\hat{G}_{M_j} - w_{M_j} \right)^2. \quad (10)$$

The entire optimization process is dynamic. After each round of weight updates, the weights and scores continue to change. The loss function computes new scores based on the updated weights, which are then iteratively used to optimize the weights further, forming a continuous iterative optimization loop, as detailed in Algorithm 1.

4. Experiments

4.1. Experimental Setup

Datasets. To validate the effectiveness of our framework, we conduct experiments on two distinctly styled datasets: ScienceQA [28] and MMStar [10]. ScienceQA underscores the importance of CoT in scientific reasoning tasks, comprising approximately 21,208 multimodal multiple-choice questions covering various science topics. MMStar focuses on evaluating the multimodal capabilities of MLLMs, and consists of 1,500 meticulously curated samples including 6 core capabilities and 18 detailed axes. Each sample exhibits visual dependency and minimal data leakage.

Candidate MLLMs. Our MLLM pool includes 5 closed-source models and 1 open-source model. The open-source

Models	MMstar			ScienceQA		
	Pearson ($\uparrow$)	Spearman ($\uparrow$)	Per.Ent. ($\downarrow$)	Pearson ($\uparrow$)	Spearman ($\uparrow$)	Per.Ent. ($\downarrow$)
LLama-3.2-11b-V	0.314 ± 0.0757	0.550 ± 0.0577	0.983 ± 0.2310	0.160 ± 0.0430	0.225 ± 0.0957	1.099 ± 0.0000
Claude-3-haiku	0.095 ± 0.1301	0.225 ± 0.1500	1.099 ± 0.0000	-0.145 ± 0.0308	-0.525 ± 0.2061	1.099 ± 0.0000
Claude-3.5-sonnet	0.780 ± 0.0146	0.825 ± 0.0500	0.752 ± 0.2310	0.437 ± 0.0452	0.450 ± 0.1732	0.752 ± 0.2310
Gemini-1.5-pro	0.864 ± 0.0068	0.850 ± 0.0577	0.637 ± 0.0000	0.4147 ± 0.0464	0.725 ± 0.2062	0.434 ± 0.5352
GPT-4o-mini	0.668 ± 0.0073	0.725 ± 0.1258	0.868 ± 0.2886	0.3354 ± 0.0635	0.600 ± 0.0000	1.099 ± 0.0000
GPT-4o	0.878 ± 0.0038	0.875 ± 0.0050	0.159 ± 0.3183	0.617 ± 0.0071	0.625 ± 0.1258	0.637 ± 0.0000
Methods						
Peer Review	0.725 ± 0.0044	0.771 ± 0.1616	1.040 ± 0.2830	0.463 ± 0.0193	0.686 ± 0.1777	1.040 ± 0.2830
Majority Vote [42]	0.757 ± 0.0013	0.757 ± 0.0857	1.299 ± 0.1733	0.509 ± 0.0181	0.524 ± 0.0660	1.040 ± 0.0000
Rating Vote [2]	0.795 ± 0.0013	0.743 ± 0.2309	0.628 ± 0.0755	0.623 ± 0.0084	0.629 ± 0.1895	0.920 ± 0.2387
PRD [22]	0.838 ± 0.0027	0.864 ± 0.0317	0.427 ± 0.0087	0.636 ± 0.0042	0.694 ± 0.0734	0.746 ± 0.0016
UPME	0.944 ± 0.0011	0.972 ± 0.0330	0.141 ± 0.2812	0.814 ± 0.0024	0.886 ± 0.0286	0.422 ± 0.2812

Table 1. Main results of different models and methods on MMstar and ScienceQA datasets, evaluated by Pearson, Spearman, and Permutation Entropy metrics. Since the model cannot judge itself, scores from the other five models involved in the computations are used for evaluation. Optimal results are highlighted in bold and underlined. Standard deviation (std) is calculated based on four runs.

model is Llama-3.2-11b-vision-instruct [29], which adapts the Llama-3 architecture to the multimodal domain through visual instruction tuning [24]. The closed-source models consist of GPT-4o [35], GPT-4o-mini [33], Claude-3.5-Sonnet [3], Claude-3-Haiku [4], and Gemini-1.5-Pro [45]. These commercial models employ enhanced human alignment strategies to optimize performance and user experience across various applications.

Metrics. For all experiments, we employ three popular metrics to evaluate the setups mentioned above and our UPME method: the Pearson Correlation Coefficient [14] to evaluate the similarity of score lists, Spearman’s Rank Correlation Coefficient [54] to measure the correlation of ranking lists, and Permutation Entropy [5] to assess the complexity and unpredictability of the scoring distributions. All experiments are performed for 4 runs with different seeds ($seed = 1, 2, 3, 4$) and record the average results.

Comparison Methods. Since we are the first to investigate unsupervised peer review in MLLM evaluations, we choose the original peer review mechanism and two popular classical methods, Majority Vote [42] and Rating Vote [2], for our comparative analysis. We also adopt a semi-supervised method, PRD [22], and conduct experiments using individual models as judges to further bolster our comparisons.

4.2. Experimental Results

4.2.1. Comparison Results

The experimental results in Table 1 indicate that among the candidate MLLMs, GPT-4o exhibits the highest review capability on the MMstar dataset, achieving Pearson and Spearman correlation coefficients of 0.878 and 0.875, respectively. In contrast, the open-source LLama-3.2 model demonstrates weaker performance. Although Claude-3.5-

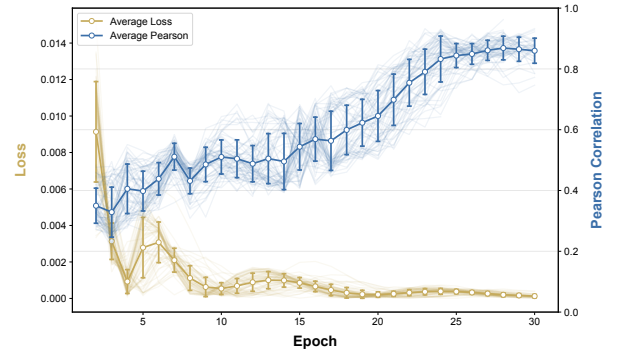


Figure 3. Convergence experiments.

Sonnet and Gemini-1.5-Pro claim to offer performance comparable to GPT-4o, they still show limitations when evaluating other models, falling noticeably short of human-level performance. The original peer review method yields Pearson and Spearman coefficients lower than that of GPT-4o, indicating that the biases of weaker models influence the Peer Review approach. Our UPME method achieves a Pearson correlation of 0.944 and a Spearman correlation of 0.972 on the MMstar dataset, demonstrating a high degree of similarity with human annotations and validating the effectiveness of our optimizations in visual QA comprehension and consistency evaluation.

On the ScienceQA dataset, even the best-performing GPT-4o achieves Pearson and Spearman correlations of 0.617 and 0.625, reflecting the difficulty of the task. The original peer review method shows more modest results, with Pearson and Spearman of 0.463 and 0.686, respectively. Following a series of optimizations and alignments, UPME achieves Pearson and Spearman correlations

Method	Cor.	Vis.	Cli.	MMstar		ScienceQA	
				Pearson	Spearman	Pearson	Spearman
Peer Review				0.727	0.714	0.457	0.600
Score Optimization	✓			0.854	0.771	0.713	0.657
Visual Alignment		✓		0.873	0.886	0.701	0.771
		✓	✓	0.785	0.829	0.548	0.771
		✓	✓	0.903	0.943	0.775	0.886
UPME	✓	✓	✓	0.944	0.972	0.814	0.886

Table 2. Ablation study with Pearson and Spearman similarity metrics for each dataset. Optimal results are highlighted in bold and underlined. ‘Cor.’ stands for correctness, ‘Vis.’ stands for visual understanding, and ‘Cli.’ stands for Clip.

of 0.814 and 0.886, showcasing its robustness and superior performance in complex datasets. While the performance on ScienceQA may not be as high, it is important to note that traditional datasets like ScienceQA might not fully align with the evaluation needs of current MLLMs because they may lack visual dependency or sufficient discrimination [10, 52]. UPME achieves higher Pearson and Spearman correlations, demonstrating its robustness and superior performance.

Moreover, as shown in Figure 3, the loss consistently decreases, and the Pearson similarity metric increases over epochs in 64 different initial settings, which demonstrates the convergence of w and $\hat{G}$. Under the current experimental setup (25 images and 1500 judgments), the framework reliably converges within the 30-epoch limit.

4.2.2. Ablation Study

In the ablation study presented in Table 2, we evaluated the impact of progressively introducing different components on model performance. The Score Optimization phase showed consistent improvements in Pearson and Spearman correlation coefficients as modules were added. For instance, using only the correctness module resulted in Pearson scores of 0.854 for MMstar and 0.713 for ScienceQA. Introducing the visual understanding module alone yielded scores of 0.873 and 0.701, respectively. Adding the Clip component alone gave scores of 0.785 for MMstar and 0.548 for ScienceQA. Combining visual understanding and Clip modules further improved performance to 0.903 and 0.775. Ultimately, integrating all components within the UPME framework resulted in the highest performance, with Pearson and Spearman correlations of 0.944 and 0.972 on MMstar and 0.814 and 0.886 on ScienceQA. These results confirm that optimizing the scoring mechanism and visual alignment significantly boosts evaluation accuracy and robustness across different datasets.

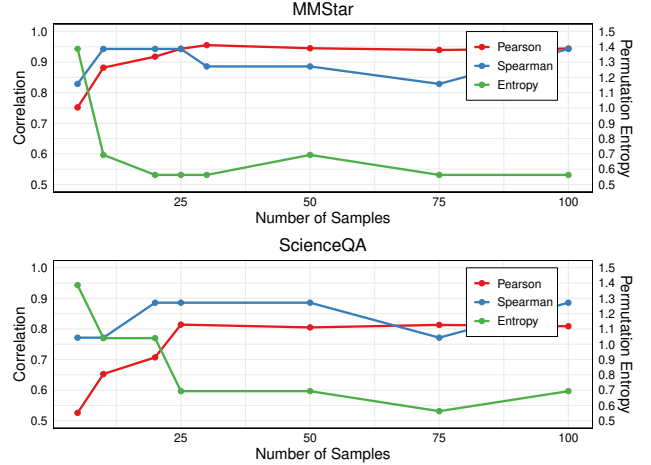


Figure 4. The performance of UPME in different sample size.

4.2.3. Impact of Sample Size

In Figure 4, we analyze the impact of sample size on the alignment with human-annotated QA benchmarks, measured through Pearson and Spearman correlation coefficients, as well as Permutation Entropy. As the sample size increases from 5 to 25, there is a clear improvement in both correlation metrics and a decline in Permutation Entropy, indicating that the generated score lists and rank lists show greater alignment with human evaluations. Beyond a sample size of 25, the metrics stabilize, showing minimal changes with further increases from 25 to 100 in sample size. This stabilization suggests that the consistency of the generated outputs relative to human judgments has reached a satisfactory level, making 25 samples a reliable point for evaluation. Therefore, we primarily selected a sample size of 25 for our experiments, as it provides an effective balance between alignment accuracy and data efficiency.

5. Human Preference Alignment Analysis

The aforementioned experiments primarily reflect the consistency between the UPME framework and the predefined QA-based benchmarks. To further explore the alignment between UPME and human preferences, as well as to examine the two types of biases inherent in MLLM-as-Judge approaches, we selected a subset of responses for human evaluation and analysis of UPME scoring results. The detailed experimental design is provided in subsection 9.4.

5.1. Alignment with Human Preference

To validate whether our proposed UPME method can align with human preferences without manual labeling, we conducted experiments on the MMstar and ScienceQA datasets. The experimental results on Table 3 indicate that the baseline method exhibited relatively low human agreement and model consistency rates, suggesting that the Peer

Dataset	Method	Agreement (%)	Consistency (%)
MMstar	Peer Review	71.1	67.5
	UPME	95.9	89.8
ScienceQA	Peer Review	68.2	61.8
	UPME	87.4	82.6

Table 3. Human Agreement (%) and Human Consistency (%) Rates on Various Datasets.

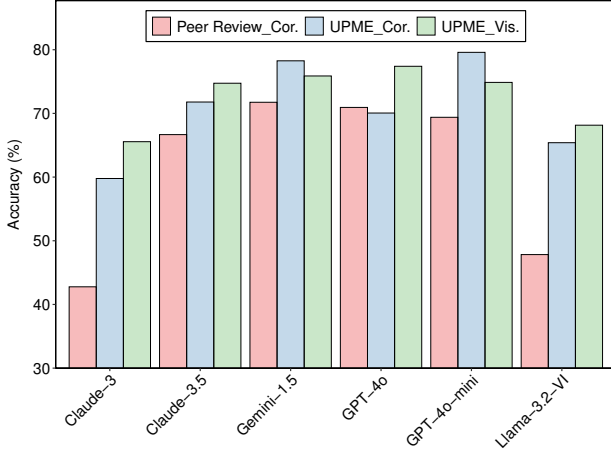
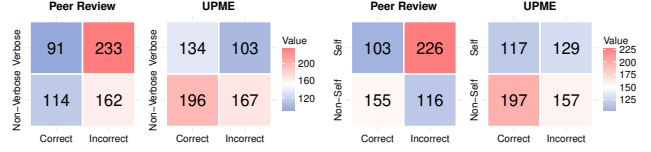


Figure 5. Model accuracy comparison in peer review framework w/ and w/o UPME, where *Peer Review_Cor.* represents the correctness of the original peer review, and *UPME_Cor.* and *UPME_Vis.* correspond to the two judgment dimensions of response correctness and visual understanding, introduced in subsection 3.2.

Review mechanism under an unsupervised setting without weight optimization struggles to align with human preferences. In contrast, UPME demonstrated significant improvements by incorporating Correctness, Visual Understanding, and Clip Correlation. On the MMStar, UPME achieved an agreement rate of 95.9% and a consistency rate of 89.8%, showing that the optimized scoring criteria significantly enhance the accuracy of evaluation outputs and alignment with human preferences. By capturing key multimodal understanding metrics without relying on manual labeling, UPME effectively achieved high consistency with human annotations, highlighting its substantial advantages in improving response consistency and accuracy under an unsupervised framework.

5.2. Performance w/ and w/o UPME

To assess the improvement in model judgment accuracy within the Peer Review mechanism and the UPME framework, we conducted an experiment, the results are displayed in Figure 5. The calculation of accuracy is based on $ACC = \frac{\text{judge}_{\text{correct}}}{\text{judge}_{\text{all}}}$. Experimental results show that the UPME significantly improves the accuracy of models in



(a) Verbosity Bias

(b) Self Preference

Bias Type	Method	Chi-Square	p-value	Phi Coefficient
Verbosity Bias	Peer review	10.996	0.00091	0.135
	UPME	0.280	0.59696	0.022
Self Preference	Peer review	39.584	3.142e-10	0.257
	UPME	3.489	0.0618	0.076

Figure 6. Heatmap and Table for Peer Review and UPME on Self Preference and Verbosity Bias.

both correctness and visual understanding compared to the original correctness judgment. Specifically, the average accuracy in correctness evaluation increased from 61.56% in the baseline to 74.48% with UPME, an improvement of 12.92%. This indicates a significant advantage of UPME in assessing the correctness of model responses. Furthermore, in the evaluation of visual understanding, UPME achieved an average accuracy of 73.93%, which is also notably higher than the baseline. This improvement is primarily attributed to UPME’s unsupervised evaluation framework, which employs a multi-model peer review mechanism that combines correctness and visual understanding scores to achieve a more efficient and accurate assessment.

5.3. Mitigating the Bias of MLLM-as-a-Judge

Mitigation of Verbosity bias. As shown in Figure 6, the verbosity-preference bias is significantly alleviated by our method. The original Peer Review approach had a p-value of 0.00091 (< 0.05), with a Chi-square value of 10.996 and a Phi coefficient of 0.135, suggesting a significant preference for verbose responses during evaluation. In contrast, the UPME framework yielded a p-value of 0.59696 (> 0.05), with the Chi-square value reduced to 0.280 and the Phi coefficient decreased to 0.022, indicating that it does not exhibit a significant bias toward selecting verbose responses. Additionally, the number of incorrect selections of verbose responses dropped from 233 to 103 in the UPME framework. Thus, the UPME approach effectively mitigates the bias toward verbose responses in models, enhancing evaluation diversity and alignment with human preferences in an unsupervised setting.

Mitigation of Self-preference. Self-preference bias is also significantly alleviated, as shown in Figure 6. In the original Peer Review method, the Chi-square value was 39.584, the Phi coefficient was 0.257, and the p-value was



Figure 7. Case study illustration of UPME. We provide the original human-designed questions and the UPME-generated questions, along with the answer analysis. The upper case presents the *Disability of review model*, where the review model can not answer the original question itself. The middle case demonstrates cases exhibiting *verbosity bias*. The bottom case shows *self-preference bias*.

3.142e-10, indicating a highly significant bias. In contrast, the UPME framework reduced the Chi-square value to 3.489, the Phi coefficient to 0.076, and the p-value to 0.0618, approaching insignificance. The incorrect selections due to self-preference dropped from 226 to 129, further demonstrating our effectiveness in mitigating this bias.

5.4. Case Study

We selected three cases to demonstrate how the UPME effectively addresses the core issues in traditional MLLM-as-a-judge as shown in Figure 7. In the first case, the review model initially failed to identify the error in Candidate B's response because it was unable to answer the question itself. After applying UPME, the review model was able to generate a question within its capabilities and accurately determine that Candidate A's response showed a deeper understanding of the image, which aligned with human preferences, thus avoiding the misjudgment issue. In the second case, although Candidate B's response was longer, it did not

provide additional useful information. Traditional evaluations might wrongly consider the longer answer as more informative, but with UPME, the review model, through precise text evaluation, clearly identified Candidate A's concise and clear answer as more valuable, thus avoiding verbosity bias. Finally, in the third case, UPME effectively avoided self-preference bias, enhancing the fairness and objectivity.

6. Conclusion

This study proposes an unsupervised peer review-based framework, UPME, effectively reducing the annotation workload in traditional human-designed QA benchmarks. Our vision-language scoring system emphasizes a vision-oriented judgment, addressing issues such as verbosity and self-preference biases inherent in MLLM-as-judge evaluation methods. Further experiments validate that our evaluation results achieve higher alignment with human preferences, providing a promising and reliable approach.

Acknowledgement

This work was supported by Alibaba DAMO Academy through the Alibaba Innovative Research Program.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1
- [2] Mohammad Allahbakhsh and Aleksandar Ignjatovic. Rating through voting: An iterative method for robust rating. *arXiv preprint arXiv:1211.0390*, 2012. 5
- [3] Anthropic. Claude 3.5: A sonnet. <https://www.anthropic.com/news/claude-3-5-sonnet>, 2024. 1, 5
- [4] Anthropic. Claude 3 haiku. <https://www.anthropic.com/news/claude-3-haiku>, 2024. 5, 1
- [5] Christoph Bandt and Bernd Pompe. Permutation entropy: a natural complexity measure for time series. *Physical review letters*, 88(17):174102, 2002. 5
- [6] Han Bao, Yue Huang, Yanbo Wang, Jiayi Ye, Xiangqi Wang, Xiuying Chen, Mohamed Elhoseiny, and Xiangliang Zhang. Autobench-v: Can large vision-language models benchmark themselves? *arXiv preprint arXiv:2410.21259*, 2024. 2
- [7] Yuzhe Cai, Shaoguang Mao, Wenshan Wu, Zehua Wang, Yaobo Liang, Tao Ge, Chenfei Wu, Wang You, Ting Song, Yan Xia, et al. Low-code llm: Visual programming over llms. *arXiv preprint arXiv:2304.08103*, 2, 2023. 1
- [8] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Pan Zhou, Yao Wan, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. *arXiv preprint arXiv:2402.04788*, 2024. 2
- [9] Dongping Chen, Yue Huang, Siyuan Wu, Jingyu Tang, Liuyi Chen, Yilin Bai, Zhigang He, Chenlong Wang, Huichi Zhou, Yiqiang Li, et al. Gui-world: A dataset for gui-oriented multimodal llm-based agents. *arXiv preprint arXiv:2406.10819*, 2024. 1
- [10] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 1, 2, 4, 6
- [11] Zhaorun Chen, Yichao Du, Zichen Wen, Yiyang Zhou, Chenhang Cui, Zhenzhen Weng, Haoqin Tu, Chaoqi Wang, Zhengwei Tong, Leria HUANG, et al. Mj-bench: Is your multimodal reward model really a good judge? In *ICML 2024 Workshop on Foundation Models in the Wild*, 2024. 2
- [12] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024. 2
- [13] Zhumin Chu, Qingyao Ai, Yiteng Tu, Haitao Li, and Yiqun Liu. Pre: A peer review based large language model evaluator. *arXiv preprint arXiv:2401.15641*, 2024. 3
- [14] Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4, 2009. 5
- [15] Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. Benchmark probing: Investigating data leakage in large language models. In *NeurIPS 2023 Workshop on Backdoors in Deep Learning-The Good, the Bad, and the Ugly*, 2023. 2
- [16] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palme: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 1
- [17] Chujie Gao, Qihui Zhang, Dongping Chen, Yue Huang, Siyuan Wu, Zhengyan Fu, Yao Wan, Xiangliang Zhang, and Lichao Sun. The best of both worlds: Toward an honest and helpful large language model. *arXiv preprint arXiv:2406.00380*, 2024. 2
- [18] Wenyi Hong, Weihang Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Yuxiao Dong, et al. Cogagent: A visual language model for gui agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14281–14290, 2024. 1
- [19] Zdeněk Kasner and Ondřej Dušek. Beyond traditional benchmarks: Analyzing behaviors of open llms on data-to-text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12045–12072, 2024. 1
- [20] Zdeněk Kasner and Ondřej Dušek. Beyond traditional benchmarks: Analyzing behaviors of open llms on data-to-text generation, 2024. 2
- [21] Jian Li and Weiheng Lu. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*, 2024. 1
- [22] Ruosen Li, Teerth Patel, and Xinya Du. Prd: Peer rank and discussion improve large language model based evaluations. *arXiv preprint arXiv:2307.02762*, 2023. 5, 2
- [23] Bin Lin, Yang Ye, Bin Zhu, Jiayi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 1
- [24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 5
- [25] Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Zhuoer Feng, Bosi Wen, Jiale Cheng, Pei Ke, et al. Alignbench: Benchmarking chinese alignment of large language models. *arXiv preprint arXiv:2311.18743*, 2023. 2
- [26] Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. Aligning with human judgement: The role of pairwise preference in large language model evaluators. *arXiv preprint arXiv:2403.16950*, 2024. 2, 4
- [27] Adian Liusie, Potsawee Manakul, and Mark JF Gales. Zero-shot nlg evaluation through pairwise comparisons with llms. *arXiv preprint arXiv:2307.07889*, 2023. 4

- [28] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 1, 2, 4
- [29] Meta. Llama 3.2 11b-vision-instruct. <https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct>, 2024. 5, 1
- [30] Birger Moell. Evaluating large language models with human feedback: Establishing a swedish benchmark. *arXiv preprint arXiv:2405.14006*, 2024. 2
- [31] Kun-Peng Ning, Shuo Yang, Yu-Yang Liu, Jia-Yu Yao, Zhen-Hui Liu, Yu Wang, Ming Pang, and Li Yuan. Peer-review-in-llms: Automatic evaluation method for llms in open-environment. *arXiv preprint arXiv:2402.01830*, 2024. 2, 3
- [32] OpenAI. Gpt-4 technical report. <https://openai.com/research/gpt-4>, 2023. 1
- [33] OpenAI. Gpt-4o mini: Advancing cost-efficient intelligence. <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>, 2024. 5, 1
- [34] OpenAI. Introducing openai o1: Advancing model reasoning with visual and logical insights. <https://openai.com/index/introducing-openai-o1-preview/>, 2024. 4
- [35] OpenAI. Hello gpt-4o, 2024. 5, 1
- [36] Renjie Pi, Lewei Yao, Jiahui Gao, Jipeng Zhang, and Tong Zhang. Perceptionpt: Effectively fusing visual perception into llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27124–27133, 2024. 1
- [37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4, 3
- [38] Vyas Raina, Adian Liusie, and Mark Gales. Is llm-as-a-judge robust? investigating universal adversarial attacks on zero-shot llm assessment. *arXiv preprint arXiv:2402.14016*, 2024. 2
- [39] Refuel.ai. Llm labeling: Technical report, 2025. 2
- [40] Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023. 1
- [41] Andreas Stephan, Dawei Zhu, Matthias Aßenmacher, Xiaoyu Shen, and Benjamin Roth. From calculation to adjudication: Examining llm judges on mathematical reasoning tasks. *arXiv preprint arXiv:2409.04168*, 2024. 3
- [42] James Surowiecki. The wisdom of crowds/james surowiecki. NY.: Anchor, 2005. 5
- [43] Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. Judgebench: A benchmark for evaluating llm-based judges. *arXiv preprint arXiv:2410.12784*, 2024. 3
- [44] Ruixue Tang, Chao Ma, Wei Emma Zhang, Qi Wu, and Xiaokang Yang. Semantic equivalent adversarial data augmentation for visual question answering. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pages 437–453. Springer, 2020. 1
- [45] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 5, 1
- [46] Li Wang, Hui Zhang, and Rajesh Patel. A user-centric benchmark for evaluating large language models. *arXiv preprint arXiv:2404.13940*, 2024. 3
- [47] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 1, 2
- [48] Siyuan Wu, Yue Huang, Chujie Gao, Dongping Chen, Qihui Zhang, Yao Wan, Tianyi Zhou, Xiangliang Zhang, Jianfeng Gao, Chaowei Xiao, et al. Unigen: A unified framework for textual dataset generation using large language models. *arXiv preprint arXiv:2406.18966*, 2024. 1
- [49] xAI. Realworldqa: A benchmark for evaluating real-world spatial understanding in multimodal ai models, 2024. 2
- [50] Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*, 2024. 2
- [51] Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, et al. Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*, 2024. 2
- [52] 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. *arXiv preprint arXiv:2308.02490*, 2023. 1, 2, 6
- [53] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 1, 2
- [54] Jerrold H Zar. Spearman rank correlation. *Encyclopedia of biostatistics*, 7, 2005. 5
- [55] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023. 1, 2
- [56] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*, 2023. 1