一种多对一访谈范式用于高效 MLLM 评估

Ye Shen[1,2], Junying Wang[2,3], Farong Wen[1,2], Yijin Guo[1,2], Qi Jia[2], Zicheng Zhang[†2], Guangtao Zhai[†1,2]

[1]Shanghai Jiao Tong University, [2]Shanghai AI Laboratory, [3]Fudan University
[†]Corresponding author.

ABSTRACT

多模态大型语言模型（MLLMs）的快速发展促进了众多基准测试的创建。然而，传统的全范围问题回答评估存在高冗余和低效率的问题。受到人类面试过程的启发，我们提出了一种**多对一访谈范式**用于高效地评估 MLLM。我们的框架包括（i）一个包含预面试和正式面试阶段的两阶段面试策略，（ii）动态调整面试官权重以确保公平性，以及（iii）一种自适应机制来选择问题难度级别。在不同基准测试上的实验表明，所提出的范式与全范围结果的相关性比随机抽样高出显著提高，PLCC 提高了 17.6%，SRCC 提高了 16.7%，同时减少了所需的问题数量。这些发现证明了所提出的范式为大规模 MLLM 基准测试提供了一种可靠且高效的替代方案。

Index Terms— MLLM 评估，多对一面试

1. 介绍

多模态大型语言模型（MLLMs）在涉及图像、视频、音频和 3D 内容的各种任务中取得了显著的性能 [1] 。随着这些模型的迅速发展，**可靠且高效的评估**已经成为一个中心研究挑战，由此产生了各种各样的基准测试 [2] 。然而，传统的全范围问题回答（Q&A）评估存在**严重冗余**：许多实例高度相似，对模型评估贡献的新信息很少 [3] 。也就是说，可靠排名所需的实例比全面评估所需更少，从而推动了**更高效的** MLLM **评估范式**的发展。

受现实世界招聘实践的启发，其中**多对一访谈**使评估者能够通过少量精心选择的问题快速衡量候选人的能力，我们引入了一种新颖的面试范式来解决全面覆盖问答测试效率低下的问题。我们的方法包括三个关键组成部分：(i) 一个**两阶段面试策略**，包括一个轻量级预面试进行初步难度校准和正式面试进行全面能力评估；(ii) **访谈者权重的动态调整**，允许不同模型更公平、全面地评价被面试者；以及 (iii) 一个**自适应难度机制**，根据当前轮次的难度和被面试者的表现在更新后续问题，确保广泛覆盖各个能力水平。这些组成部分共同实现了对 MLLMs 的**全面的**、**准确的**、**公平的和高效的**评估。

我们的主要贡献如下：

- 我们提出了一种用于 MLLM 评估的**多对一访谈范式**，包括（一）两阶段面试策略，（二）动态面试官权重调整，以及（三）自适应难度机制。
- 我们证明了这一范式提供了**可靠**、**公平和高效的反映**的多语言模型能力，涵盖了整体性能和难度感知分布。
- 在 MMT-Bench、ScienceQA 和 SEED-Bench 上进行的大量实验表明，我们的方法始终优于随机采样，并且在使用较少问题的情况下最多可达到 17.6% **的** PLCC **和** 16.7% **的** SRCC **改进**超过全范围问答测试。

2. 相关工作

2.1. 多模型大型语言模型

早期的大语言模型（LLMs）表现出卓越的文本处理能力 [4]。GPT-2 [5] 展示了令人惊讶的人机对话能力。在 LLMs 的基础上，MLLMs 进一步具备处理多模

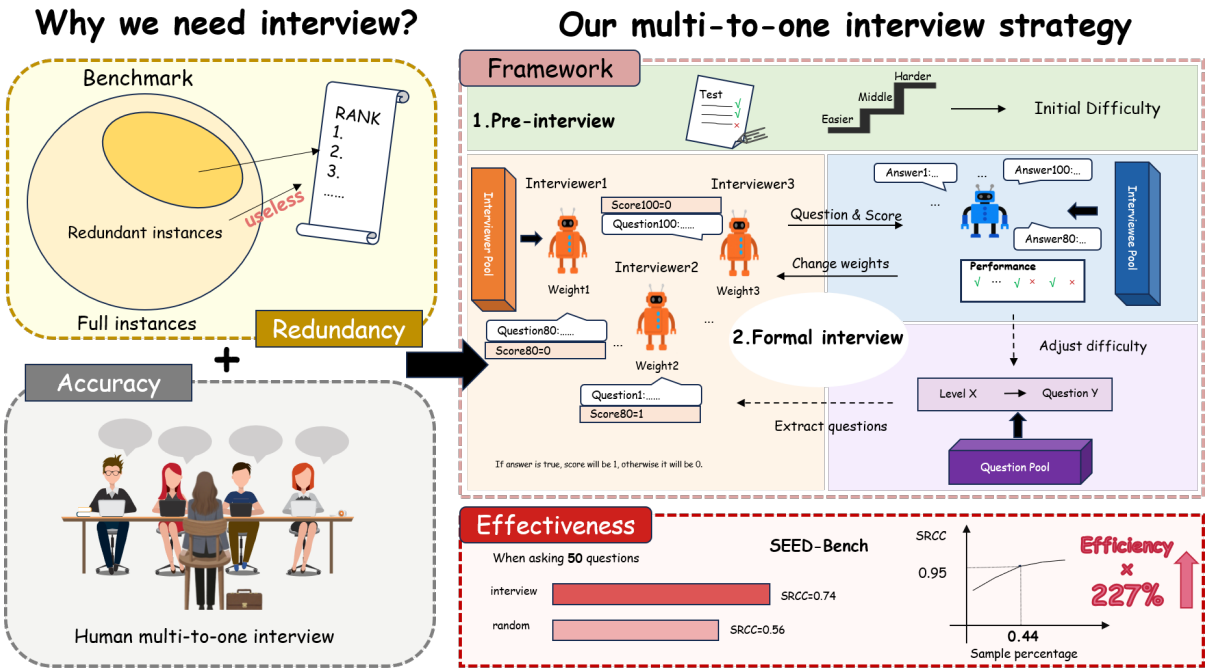


Fig. 1. 使用完整实例进行基准测试的评估效率较低。考虑到人类多对一访谈的效率，我们提出了一种多对一访谈范式来评估模型的能力，并展示了其有效性、准确性和公平性。

态信息的能力 [4]。CLIP-ViT [6] 为视觉文本对齐奠定了基础。随后，诸如 Claude-4-Sonnet [7], GPT-4o [8], Grok-4 [9] 和 Gemini-2.5-Pro [10] 等模型不断改进能力，在众多多模态任务中表现出色，甚至在现实世界中的复杂任务中也是如此。然而，如何评估模型**更准确、全面、公平和现实地**的能力已成为一个重要问题。

2.2. MLLM 基准测试

MLLMs 的快速发展推动了基准测试进入多维度能力综合评估 [11] 。早期的基准测试如 VQA [12] 评估视觉识别能力，但未能捕捉到高级推理能力。因此，研究人员开始设计针对 MLLMs 特性的全面评测基准。SEED-Bench [13] 涵盖的维度显著增加，而 MMT-Bench [14] 进一步扩大了范围，涵盖了实际应用场景中的复杂任务。尽管基准测试不断进步，但大多数仍然采用全覆盖问答测试，**这通常会导致冗余和缺乏适应性，因此迫切需要更有效的评估策略**。

3. 面试框架

总体而言，精心设计的访谈框架结合了（一）两阶段访谈策略，（二）面试官权重的动态调整，以及（三）自适应难度机制。

3.1. 两阶段面试策略

面试过程包括一个有效的预面试，初步划分难度等级，并进行正式面试以全面评估模型的能力。首先，在预面试阶段，通过书面测试来初步评估模型的能力。具体来说，从中等难度的问题库中随机抽取问题来测试模型。然后计算相应的准确率以确定正式面试的初始难度。

$$\text{Initial difficulty} = \begin{cases} middle + 1 & \text{if } acc > \beta, \\ middle & \text{if } acc = \beta, \\ middle - 1 & \text{otherwise,} \end{cases} \quad (1)$$

其中 acc 表示应试者的测试结果，β 表示可调节的准确率阈值。其次，在每轮正式面试中，选定的面试官负责（i）从与当前难度匹配的类别堆栈中抽取多个经

典问题；(ii) 判断应试者答案是否正确。然后根据应试者的表现在每一轮调整问题的难度。一旦达到设定的问题数量，整个面试结束。

3.2. 访谈者权重的动态调整

为了提高访谈范式的可靠性，我们引入了一种动态调整面试官权重的策略。具体来说，从面试官池中选择多名面试官，最初它们的权重相等。在问答环节之后，根据受访者的先前回答来计算面试官的权重。所选的提问面试官模型由这些权重决定。

$$w_i^{t+1} = \begin{cases} (1-\alpha)\cdot w_i^t & \text{if } acc_i = 0 \text{ or } acc_i = 1, \\ (1+\alpha)\cdot w_i^t & \text{otherwise,} \end{cases} \tag{2}$$

w 表示约束在 0.5 和 2 之间的权重，以避免极端偏差，i 表示第 i 位面试官，t 表示第 t 轮问答，acc_i 表示在第 i 位面试官审查下的被面试者的表现，α 是一个可调的权重阈值。

3.3. 自适应难度机制

为了提高访谈范式的准确性，我们设计了一种策略来自适应地调整每轮的难度。具体来说，在一轮之后，通过公式（3）计算准确性以确定下一轮的难度。

$$\mathrm{L}^{r+1} = \begin{cases} \mathrm{L}^r + 1 & \text{if } acc^r > \beta, \\ \mathrm{L}^r & \text{if } acc^r = \beta, \\ \mathrm{L}^r - 1 & \text{if } acc^r < \beta, \end{cases} \tag{3}$$

其中 L 表示难度级别，r 表示第 r 轮，acc^r 表示受试者在第 r 轮难度级别的整体表现，β 表示一个可调的准确性阈值。为了防止面试陷入局部难度冲击，如果难度变化在 $Level^{r-1}$ 和 $Level^r$ 之间重复三次，则使用公式 (4) 适当调整难度级别。

$$\mathrm{L}^{r+1} = \begin{cases} \mathrm{L}^{r-1} + 3 & \text{if } \mathrm{L}^{r-1} > n, \\ \mathrm{L}^{r-1} - 3 & \text{if } otherwise, \end{cases} \tag{4}$$

其中 n 表示可调级阈值。当在 $Level^r$ 处没有可用问题时，我们根据公式 (5) 调整难度级别。

$$\mathrm{L}^{r+1} = \begin{cases} \mathrm{L}^r + 1 & \text{if } acc^r > \beta, \\ \mathrm{L}^r - 1 & \text{if } otherwise. \end{cases} \tag{5}$$

4. 实验

4.1. 准备工作

4.1.1. 相关的基准测试

为了构建具有难度属性的具体基准，选择了 10 个典型模型来考察问题，其中包括 GPT-4o [8], Deepseek-VL [15], Qwen-2.5-VL [16], Gemini-1.5-pro [17], Grok-2 [18], Kimi-VL [19], Phi-3 [20], Claude3-7 [21], HunYuan [22] 和 InternVL3 [23]。从正确回答的模型数量中减去 11 以获得每个问题的难度。如果计算出的级别为 11，则视为 10。最终，MMT-Bench [14], ScienceQA [24] 和 SEED-Bench [13] 中的问题成功地被划分为 10 个难度等级。

4.1.2. 相关模型

实验中使用的访谈者和受访者模型如表 1 所示。

调用方法	模型
API Call	GPT-4.1-Nano [25], GPT-4o-Mini [8], GPT-4o [8], Grok-2 [18], Grok-4 [9], Claude-3.5-Sonnet [21], Claude-4-Sonnet [7], Qwen-VL-Plus [26], Gemini-2.0-Flash [10], Gemini-2.5-Pro [10], Gemini-2.5-Flash [10]
Local Call	Phi-4-mini-Instruct [20], Qwen2.5-VL-7b-Instruct [16], Qwen2.5-VL-72B-Instruct [16], Qwen2.5-VL-32B-Instruct [16], Internlm3-8B-Instruct [27], Internlm2.5-20B-Chat [27], Gemma-3-27B-It [28], Llama-3.2-11B-Vision-Instruct [29]

Table 1. MLLM 访谈对象和 MLLM 访谈者的示例。使用了 11 个先进的闭源 MLLM 和 8 个先进的开源 MLLM。

4.2. 设置

表 1 中的每个模型都被指定为面试者，其中 3 个通过 API 随机选定为面试官。预面试会随机提出 3 个 5 级问题。正式面试每轮都包含 3 个问题。面试官初始权重为 1.0。参数设置包括 $\alpha = 0.2$ 以确保权重调整的合理性，$\beta = 0.5$ 准确衡量模型级别，以及 $n = 7$ 覆盖全部难度。设定数量为 200。

不同难度下的准确率被计算出来以验证受访者的能力分布，并建立一个对照组，我们在其中随机选择基准问题。MLLMs 在全范围问答测试中的表现作为真实值。分析中采用了 3 个广泛采用的指标，包括斯皮尔曼等级相关系数（SRCC）[30] 、皮尔逊线性相关系数（PLCC）[31] 和肯德尔等级相关系数（KRCC）[32] 。

4.3. 分析

该范式的有效性通过三个基准进行了展示：MMT-Bench、ScienceQA 和 SEED-Bench，与随机策略相比。

4.3.1. 我们范式的准确性

根据表 2 中揭示的数值，多对一访谈范式在大多数问题设置上实现了比随机采样更高的 SRCC、PLCC 和 KRCC 值。例如，在询问 50 个问题时，提议的范式在 SEED-Bench 上的 SRCC 为 0.7377，而随机策略仅为 0.5577。这种现象与 SEED-Bench 中的大量问题有关，证实了访谈范式能够更准确地评估可用问题有限的模型的能力。

4.3.2. 我们范式的效率

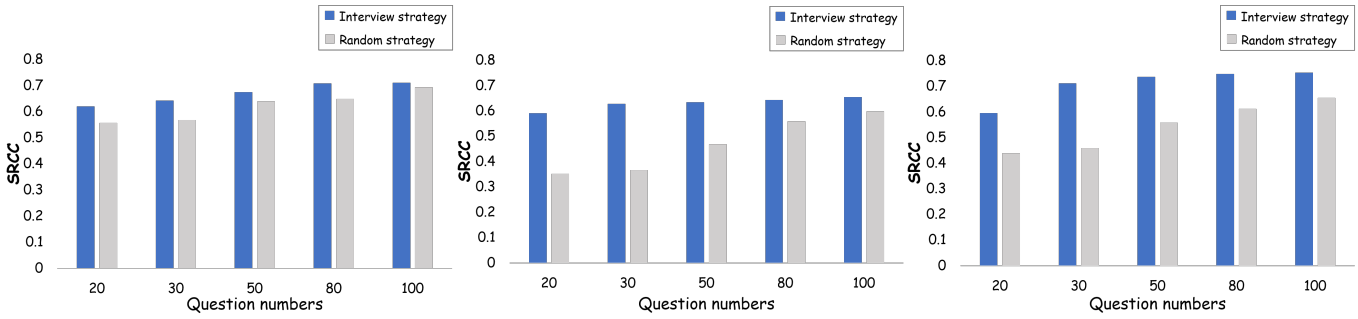
两种策略在图 2 中 SRCC 值的明显差距验证了我们方法的优越效率。当选择 30 个问题时，该范式在 ScienceQA 上实现了 0.6289 的 SRCC，而随机策略仅达到 0.3668，这表示有 71.46%的提升。这种性能差异可以归因于 ScienceQA 的问题相对较难，但访谈范式可以在评估阶段早期就针对最具有信息量的问题。

5. 结论

总之，多对一访谈范式提高了 MLLM 评估效率并改革了现有的评估方法。然而，仍然存在一些限制，包括问答评估的局限性和相对简单的面试者选择。希望一方面可以将其扩展到自动化基准构建；另一方面，在跨语言评估中具有重要意义。

Table 2. 多对一访谈范式与随机策略的比较，其中 Avg.improvement 表示我们的范式在每个指标上相对于随机采样的平均绝对增加（百分点）。

基准测试	MMT-**基准测试**			**科学问答**			**种子基准测试平台**		
数字	**信源相关系数**	PLCC	KRCC	**信度系数**	PLCC	KRCC	**信度系数**	PLCC	KRCC
随机策略：随机选取问题									
20	0.5573	0.6654	0.4279	0.3515	0.4205	0.2284	0.4277	0.4391	0.3090
30	0.5687	0.6705	0.4377	0.3668	0.4316	0.2433	0.4590	0.4876	0.3283
50	0.6397	0.7021	0.4955	0.4688	0.6183	0.3512	0.5577	0.5894	0.4085
80	0.6496	0.6985	0.5373	0.5592	0.6376	0.4345	0.6129	0.6925	0.5091
100	0.6932	0.6470	0.5426	0.5984	0.6402	0.4556	0.6554	0.6907	0.4820
多对一访谈范式（提议）：在访谈中挑选的问题									
20	0.6202	0.7275	0.4273	0.5911	0.7255	0.4366	0.5958	0.6780	0.4372
30	0.6424	0.7608	0.4557	0.6289	0.7225	0.4956	0.7122	0.6579	0.5494
50	0.6749	0.7957	0.5148	0.6348	0.7262	0.5118	0.7377	0.7110	0.5733
80	0.7091	0.8303	0.5385	0.6435	0.7213	0.4882	0.7491	0.7239	0.5852
100	0.7109	0.8308	0.5503	0.6547	0.7328	0.5013	0.7532	0.7251	0.5962
平均改进	4.98%	11.23%	0.91%	16.17%	17.60%	14.41%	16.71%	11.93%	14.09%



(a) MMT-基准　(b) 科学问答　(c) SEED-基准测试平台

Fig. 2. SRCC 性能概述对应于 3 个基准测试中的问题编号，清楚地区分了两种方法之间的差距。

6. REFERENCES

[1] Zicheng Zhang et al., "Large multimodal models evaluation: A survey," https://github.com/aiben-ch/LMM-Evaluation-Survey, 2025, Project Page: AIBench, available online.

[2] Junying Wang et al., "The ever-evolving science exam," arXiv preprint arXiv:2507.16514, 2025.

[3] Zicheng Zhang et al., "Redundancy principles for mllms benchmarks," arXiv preprint arXiv:2501.13953, 2025.

[4] Chaoyou Fu et al., "Mme-survey: A comprehensive survey on evaluation of multimodal llms," 2024.

[5] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al., "Language models are unsupervised multitask learners," OpenAI blog, vol. 1, no. 8, pp. 9, 2019.

[6] 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. PmLR, 2021, pp. 8748–8763.

[7] Anthropic, "Introducing claude 4," https://www.anthropic.com/news/claude-4, May 2025.

[8] OpenAI, "Hello GPT-4o," https://openai.com/index/hello-gpt-4o/, 2024.

[9] xAI, "Grok 4," https://x.ai/news/grok-4, July 2025.

[10] Gemini Team et al., "Gemini: a family of highly capable multimodal models," arXiv preprint arXiv:2312.11805, 2023.

[11] Zicheng Zhang et al., "Aibench: Towards trustworthy evalua-tion under the 45 law," .

[12] Stanislaw Antol et al., "Vqa: Visual question answering," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 2425–2433.

[13] Bohao Li et al., "Seed-bench: Benchmarking multimodal llms with generative comprehension," arXiv preprint arXiv:2307.16125, 2023.

[14] Kaining Ying et al., "Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi," arXiv preprint arXiv:2404.16006, 2024.

[15] Zhiyu Wu et al., "Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding," arXiv preprint arXiv:2412.10302, 2024.

[16] Shuai Bai et al., "Qwen2. 5-vl technical report," arXiv preprint arXiv:2502.13923, 2025.

[17] Gemini Team et al., "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," arXiv preprint arXiv:2403.05530, 2024.

[18] xAI, "Grok-2 beta release," https://x.ai/news/grok-2, 2024.

[19] Kimi Team et al., "Kimi-vl technical report," arXiv preprint arXiv:2504.07491, 2025.

[20] Marah Abdin et al., "Phi-4 technical report," arXiv preprint arXiv:2412.08905, 2024.

[21] AI Anthropic, "The claude 3 model family: Opus, sonnet, haiku," Claude-3 Model Card, vol. 1, no. 1, pp. 4, 2024.

[22] Xingwu Sun et al., "Hunyuan-large: An open-source moe model with 52 billion activated parameters by tencent," arXiv preprint arXiv:2411.02265, 2024.

[23] Zhe Chen et al., "Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 24185–24198.

[24] Saikh et al., "Scienceqa: A novel resource for question answering on scholarly articles," International Journal on Digital Libraries, vol. 23, no. 3, pp. 289–301, 2022.

[25] Josh Achiam et al., "Gpt-4 technical report," arXiv preprint arXiv:2303.08774, 2023.

[26] Jinze Bai et al., "Qwen technical report," arXiv preprint arXiv:2309.16609, 2023.

[27] Zheng Cai et al., "Internlm2 technical report," arXiv preprint arXiv:2403.17297, 2024.

[28] Gemma Team et al., "Gemma 3 technical report," arXiv preprint arXiv:2503.19786, 2025.

[29] Aaron Grattafiori et al., "The llama 3 herd of models," arXiv preprint arXiv:2407.21783, 2024.

[30] Charles Spearman, "The proof and measurement of association between two things.," 1961.

[31] Karl Pearson, "Vii. note on regression and inheritance in the case of two parents," proceedings of the royal society of London, vol. 58, no. 347-352, pp. 240–242, 1895.

[32] Maurice G Kendall, "A new measure of rank correlation," Biometrika, vol. 30, no. 1-2, pp. 81–93, 1938.