大规模多模态基准的评判模型

Min-Han Shih、Yu-Hsin Wu、Yu-Wei Chen南加州大学维特比工程学院电气与计算机工程系，美国 {minhansh, yuhsinwu, ychen543}@usc.edu

arXiv:2601.06106v1 [cs.LG] 3 Jan 2026

摘要

我们提出了一种专用的多模态评判模型，旨在针对各类任务提供可靠且可解释的评估。我们的基准涵盖文本、音频、图像和视频等模态，采用经过精心抽样的公共数据集，并使用固定随机种子，以确保可复现性并尽量减少训练-测试泄漏。我们的框架不只是进行简单评分，而是整合多模态评判，分析模型输出的质量和推理一致性，并生成诊断反馈。我们在 280 个多模态样本上评估了最先进的多模态大语言模型，包括 Gemini-2.5-flash、Phi-4 和 Qwen-2.5-omni，并将评判模型的评估结果与人工标注员的评估结果进行比较。结果显示，评判模型与人工评分高度一致，展现了其作为可扩展、可解释的评估流水线的潜力，可服务于未来的多模态人工智能研究。

1 引言

大语言模型（LLMs）和多模态大语言模型（MLLMs）的最新进展，展现了其在文本、图像、音频和视频等多种模态上的理解与推理能力取得了显著进步。GPT-5、Gemini 2.5 和 Claude 4 Sonnet 等模型如今能够解读复杂的多模态输入，在长上下文上执行推理，并生成结构化且具有上下文感知能力的输出。然而，随着 MLLMs 的能力和多样性不断扩展，如何客观且高效地评估其性能，已成为日益严峻的挑战。

传统评估流水线依赖人工标注或静态基准评分，这些方式往往耗时、不一致，且局限于特定任务的性能评估。与此同时，近期的一些方法（如大语言模型评审）利用大型模型充当评估器。（Calderon 等人，2025年；Dorner 等人，2025年；D'Souza 等人，2025年）尽管这些方法减轻了人工负担，但也带来了一些问题：（1）缺乏多模态理解能力，因为许多评判系统纯粹基于文本；（2）在复杂推理或跨模态对齐任务中，与人工判断相比，可解释性和可靠性不足。

为应对这些局限，我们的项目提出了一种**评判模型**，专门用于大规模多模态基准。我们的框架不进行单一评分，而是整合并校准来自多个视角的评判结果。该评判器将评估各种 MLLMs 的输出，例如 Gemini 2.5、Phi-4（Abdin 等人，2024年）和通义千问-2.5（通义千问等人，2025年），分析响应的质量、模态之间的一致性以及推理轨迹的一致性。除了评分之外，它还将生成全面的反馈和指导，旨在为多模态人工智能系统建立可解释且可靠的评估流程。

2 相关工作

2.1 推理大语言模型

大语言模型知道何时需要逐步思考，并且只在必要时才生成逐步推理。Wei 等人，2023年）先前关于思维链（CoT）的研究大多聚焦于思维链推理步骤的正确性，或这些推理步骤是否忠实于问题并能支撑最终答案。Ye 和 Durrett，2022；Golovneva 等人，2023年）

2.2 大语言模型作为评判者

基于大语言模型的评估器是指通过提示来根据特定标准评判某些样本质量的大语言模型。基于大语言模型的评估器会使用李克特分数等评分方式（李克特，1932年），按照 1 到 5 的量表评估单个样本的质量。（Chiang 等人，2024年；Chiang 和 Lee，2023b、a、2024年；Zheng 等人，2023年）基于大语言模型的评估器还可以比较一对样本的质量，并判断哪一个更好。（Zheng 等人，2023年）

2.3 多模态基准

视觉基准从图像基准扩展到视频基准，后者引入了时间结构和更丰富的场景上下文（Nguyen et al., 2025），例如视觉问答（Agrawal et al., 2016; Goyal et al., 2017）、图表理解（Masry et al., 2022; Wang et al., 2024b; Mathew et al., 2021）以及通用能力套件（Fu et al., 2025; Li et al., 2023; Liu et al., 2024a）。不同的视频数据集侧重于不同能力：时间排序（Cai et al., 2024; Liu et al., 2024b）、程序性理解（Tang et al., 2019; Xiao et al., 2021）、自我中心感知（Grauman et al., 2022; Damen et al., 2018）和世界建模（Hong et al., 2025）。AudioSet（Kim et al., 2019）和 VGGSound（Chen et al., 2020）等音频基准以事件级声学分类为目标，而语音数据集则涵盖交流的特定方面：VoxCeleb（Nagrani et al., 2017; Chung et al., 2018）用于说话人身份识别，AudioCaps / SpeechCaps（Kim et al., 2019; Huang et al., 2025b）用于音频/语音描述，LRS（Afouras et al., 2018）用于转录，AVA-ActiveSpeaker（Kim et al., 2021）用于帧级说话活动和可听性标签。Dynamic-SUPERB（Huang et al., 2024a, 2025a）提出了一个面向语音与音频任务的动态协作基准。

3 数据集

3.1 文本到文本数据集

为了构建一个紧凑且可复现的基准，我们从 HuggingFace Hub 上广泛使用的公共数据集中抽取固定大小的小样本。对于每个数据集，我们都会明确注明所用的数据划分，以便读者评估未来模型训练中发生训练-测试泄漏的风险。我们使用固定随机种子进行无放回抽样。数据源及各项任务的数量列于表1。

任务类别	数据集（HFID）	数据划分	#项目数	答案类型
推理-代码	nvidia/OpenCodeReasoning	数据划分_0	20	代码/最终字符串
推理—数学	nvidia/OpenMathReasoning 思维链		20	生成式(最终值)
专家级多项选择题	Idavidrein/gpqa（旧项：fingertap/GPQA-Diamond）	测试集	10	多项选择
阅读理解	ucinlp/drop	验证集	20	抽取式片段
常识推理	Rowan/hellaswag	验证集	15	多项选择
常识推理	jet-ai/social_i_qa	验证集	15	多项选择
指令遵循	google/IFEval	训练集	15	生成式（受约束）
指令遵循	YuxinJiang/FollowBench	训练集（*fallback*: 验证集）	15	生成式（受约束）
总计			130	

表1：我们文本基准测试中使用的数据集、划分、数量和答案类型。

3.2 文本到文本任务

1 .**推理-代码：**需要结构化输出或最终值的程序合成问题求解。（Ahmad等人，2025年）

2. **推理—数学：**旨在得出最终标准答案的多步数学推理。（Moshkov等人，2025年）

3. **专家级选择题：**研究生水平的科学、技术、工程和数学多项选择题（Rein 等人，2023年；Yee 等人，2021；Huang 等人，2024b）。

4. **阅读理解：**基于文章、答案为抽取式的问答（Dua 等人，2019年）

5. **常识推理：**以选择题形式进行日常物理／社会推理。（Zellers 等人，2019年；Sap 等人，2019年）

6. **指令遵循：**遵守明确的约束和格式要求。（Jiang 等人，2023年；Zhou 等人，2023年）

3.3 音频语言数据集

我们的音频语言数据集由托管在 Hugging Face 上的四个开源数据集构建而成。我们经过仔细筛选，确保涵盖多种音频模态和任务类型，包括环境声音理解、鸟鸣分类、音乐生成和语音问答。数据来源和任务类型汇总于表2。

任务类别	数据集（HF ID）	数据划分	#项目
音频描述	kuanhuggingface/audiocaps_hallucination	测试集	10
鸟鸣检测	tglcourse/5s_birdcall_samples_top20	验证集	10
音乐风格检测	nyuuzyou/suno	训练集	10
语音问答	AudioLLMs/public_sg_speech_qa_test	test	20
总计			50

表2：音频基准测试中使用的数据集、数据划分、数量和答案类型。

3.4 音频-语言任务

1. **问答：** 涉及识别并描述特定声音是否存在，或根据音频内容回答问题。（Shih et al., 2024年；Kim et al., 2019年；Wang etal., 2025a）

2. **指令遵循：** 评估模型理解并执行针对听觉输入的文本指令的能力。（Wang etal., 2025a）

3.5 视频数据集

我们的视频数据集来自 Hugging Face 公共数据集，由 Sora、Veo2 和 Kling 等生成式视频模型生成的人工智能生成的短片组成。每段短片都配有人工整理的问题，旨在测试模型对常识和物理一致性的推理能力。为确保可复现性，我们使用固定随机种子进行抽样，从非训练集中总共选取了 50 段视频片段，这些片段

降低了与大语言模型在预训练期间可能已见过的材料重叠的概率。该子集提供了一个受控且多样化的评估集，用于测试多模态大语言模型的视频内容理解能力和幻觉鲁棒性。来源和任务类型汇总于表3。

数据集（HF ID）	数据划分	#项
IntelligenceLab/VideoHallu	验证集	25
IntelligenceLab/VideoHallu	test	25
总计		**50**

表 3：我们的视频基准所用的数据集、数据划分、数量和答案类型。

3.6 视频任务

由于使用了人工智能生成的合成视频，我们的任务侧重于常识和物理推理，而非传统的视觉识别。模型需要分析每个短视频片段，解读给定的问题，并判断所描绘的场景是否合理或异常。通过这种方式来界定问题，我们将重点从简单的视觉感知转向批判性推理，从而评估多模态大语言模型能否有效地将视觉信息与其对现实世界知识的内部理解相结合。

3.7 图像数据集

为了构建一个多样且易于解读的图像-语言理解基准，我们从 Hugging Face 上的四个公共数据集中筛选样本，每个数据集都代表视觉推理的一个独立子领域。对于所有数据集，我们使用一致的随机种子抽取规模较小且固定的子集，以确保可复现性。我们仅使用评估集或测试集划分，以尽量减少预训练语料库可能造成的数据泄漏。数据来源和任务类型总结见表4。

Task	数据集（HF ID）	数据划分	#项目数
图像描述	nyu-visionx/CV-Bench	test	10
图像描述	nirajandhakal/realworldqa	test	10
图表理解	ChartFoundation/ECDBench	test	15
数学推理	数学大语言模型/数学视觉	test	15
总计			**50**

表4：图像基准测试中使用的数据集、划分、数量和答案类型。

3.8 图像任务

1. **图像描述：** 评估模型进行常识推理和整体场景理解的能力。该评测考察模型是否能够生成符合物理合理性、空间与语境一致性以及真实视觉解读的描述。图像涵盖日常现实场景和复杂情境，后者需要物体识别、语境推断及更高层次的推理。（Zhu 等人，2025年）

2. **图表理解：** 衡量对条形图等结构化视觉图像进行定量推理和关系推理的能力，这需要比较趋势和数值关系。（Yang 等人，2025年）

3. **数学推理：** 评估运用数学图示和几何图形进行符号推理与空间推理的能力，这些图示和图形需要经过解析并进行定量推断。（Wang 等人，2024a，2025b）

4 方法

4.1 评估流程概述

我们设计了一套可扩展的自动化评估流程，用于评估多模态大语言模型在文本、音频、图像和视频任务中的能力。如图 1 所示，该流程包含三个主要组成部分：多模态数据、受测多模态大语言模型和评判模型。对于每个实例，系统都会检索一个包含查询及其相应上下文的样本，并将这些输入提供给受测多模态大语言模型，由其生成纯文本回答和理由。

尽管我们的框架在架构设计上支持反馈循环——将错误返回模型以重新生成内容——但本项目侧重于验证评判模型的可靠性。因此，在我们的实验中，流水线采用全面的单向评估方式。也就是说，评判模型会根据真实参考答案和多模态证据评估生成结果，并给出自己的反馈。

4.2 受测多模态大语言模型

为确保我们的基准涵盖多种能力，我们评估了三个代表不同模型系列和规模的多模态大语言模型。

- **Gemini-2.5**：一款高效的多模态模型，针对速度和推理能力进行了优化。
- **Phi-4**：由微软开发的强大开放权重模型，能够处理复杂任务。
- **Qwen-2.5**：一款全模态模型，旨在处理各种模态输入并进行推理。

4.3 评判模型

我们方法的核心是由**Gemini-3-pro**驱动的评判模型。与传统的仅处理文本的评判器不同，我们会明确提示该智能体

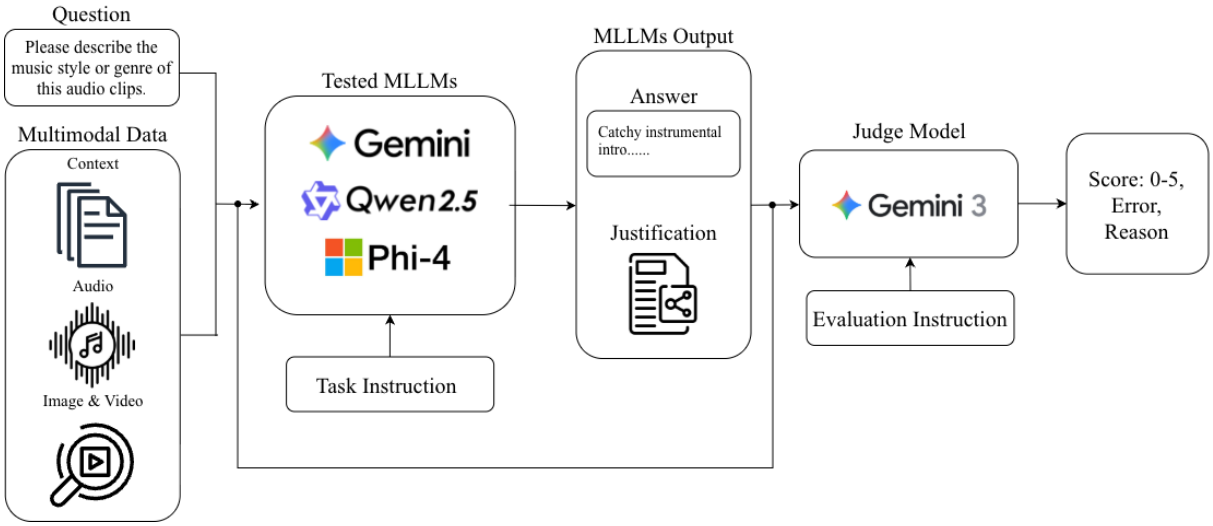

图 1：多模态大语言模型评估框架的系统架构。该框架接收多模态输入和任务指令，以查询多个受测多模态大语言模型。每个模型都会给出回答及其理由。评判模型依据评估指令评估这些回答，并输出分数（0–5）和错误分析。

执行多模态推理。它会接收原始任务指令、多模态上下文、真实答案以及被测模型的输出。评判模型有三个主要目标：

1. **回答质量分析：** 用于判断最终答案是否与真实答案或可通过物理方式观察到的现实相符。

2. **推理验证：** 分析受测模型提供的“理由”，以检查其逻辑一致性和是否存在幻觉。

3. **诊断反馈：** 该智能体不再仅输出简单的二元通过/未通过结果，而是给出结构化评估，其中包含数值分数（0-5）、具体错误类型以及详细的评分理由。

4.4 评估指标

为量化模型性能，我们采用了严格的李克特量表评分标准。评判模型根据表中定义的以下标准，对模型进行 0 到 5 分的评分 5。除了数值评分外，该方案还包括定性错误分析。如案例研究所示，评判器能够识别具体的失效模式，例如视觉幻觉或听觉遗漏。

分数	描述
5	答案正确，理由充分。
4	答案基本正确，理由基本充分，但存在一些小错误。
3	答案部分正确，理由基本合理，但存在错误。
2	答案部分正确，但理由不合理。
1	答案大部分错误，理由不合理。
0	答案完全错误或未作答。

表5：用于量化受测多模态大语言模型性能的评判模型和人工评判者评估指标。

4.5 实验结果

表6 总结了三位人工标注员与我们的评判模型在四种模态和三个多模态大语言模型上的平均评分。总体而言，对所有模型的评分取平均后，我们发现，涉及音频和视频的任务得分往往高于涉及纯文本或图像输入的任务。这表明，在我们的基准测试设置下，当前的多模态大语言模型可能已经具备相对成熟的时序理解和声学理解能力。相比之下，基于图像的任务，尤其是纯文本任务，揭示了模型间更大的性能差距：最强的系统得分接近 4 分，而较弱的模型得分通常接近或低于 2 分，这表明细粒度阅读理解和视觉依据推理仍有很大的提升空间。

从模型角度来看，Gemini-2.5 展现出最稳健的跨模态性能。在大多数情况下，其分数在全部四种模态上始终保持较高水平，人工评判者评分和评判模型评分大致集中在 3.8 至 4.5 之间，且没有在任何特定模态上出现灾难性失效。这一模式表明，Gemini 的内部表征能够在文本、音频、图像和视频输入之间良好迁移，其回答依据通常也被评为正确且有充分根据。换言之，在我们的评估流程下，Gemini 表现得像一个真正全面的多模态模型，而非针对某一种主导模态调优的专用模型。

相比之下，Phi-4 和通义千问-

2.5 展现出更为明显的模态特定弱点。Phi-4 在音频和视频上的表现处于中等水平，但在图像任务上的分数大幅下降；在这些任务中，幻觉性视觉细节和对图表元素的误读经常导致分数接近 2。通义千问-2.5 则呈现出互补的模式：它在音频和视频任务中具有竞争力，但其纯文本得分在所有模型中最低，反映出其在直接提取和推理方面频繁失败，包括在问题已明确给出答案时出现错误拒答。总体而言，这些结果表明，当前多模态大语言模型之间的跨模态鲁棒性远非均衡：一些模型能够在不同模态之间平稳泛化，而另一些模型则仍然高度依赖主导模态以及所需的推理类型。

		人类 1	人类 2	人类 3	评判器
Text	*Gemini-2.5*	3.98	**4.44**	4.12	4.28
	Phi-4	2.24	2.12	2.98	2.45
	Qwen-2.5	1.37	1.28	1.5	1.15
音频	*Gemini-2.5*	3.58	3.62	**4.16**	4.04
	Phi-4	2.66	3.26	3.30	2.82
	Qwen-2.5	3.26	3.78	3.56	3.38
视频	*Gemini-2.5*	3.70	4.10	4.20	**4.54**
	Phi-4	2.78	3.60	3.50	3.12
	Qwen-2.5	2.82	3.50	3.02	3.06
图像	*Gemini-2.5*	3.46	3.82	**3.98**	3.94
	Phi-4	2.20	2.66	1.80	2.24
	Qwen-2.5	2.52	3.28	2.28	2.32

表6：三位人工标注员与评判模型 Gemini-3-pro 对 Gemini-2.5、Phi-4 和 Qwen-2.5 在多模态任务中的评估平均分。

5 分析

5.1 分数分布

图2展示了各模型三位人工标注员与评判模型的分数分布，与表 6中报告的平均分形成补充。对于 Gemini-2.5、Phi-4 和 Qwen-2.5，我们发现，评判器的曲线在全部四种模态中都与人工评估的趋势高度一致：在人类评估中平均分较高的系统，在评判器评估中得分也较高，而且各模型在每种模态中的相对排序均保持不变。与此同时，评判器在绝对分数校准上略显保守——其分数通常比三位标注员的平均分低约 0.1–0.3 分。这种系统性偏差反映了评判器会更严格地惩罚最终答案与所给理由之间的不一致，例如答案部分正确，但其依据包含幻觉或逻辑不完整的推理。尽管存在这种轻微的负偏差，排序几乎完全一致，表明我们的自动评估器可用于对多模态大语言模型进行可靠的比较评估，即使在解读其李克特分数时可能仍需进行小幅修正，尤其是在孤立解读时。

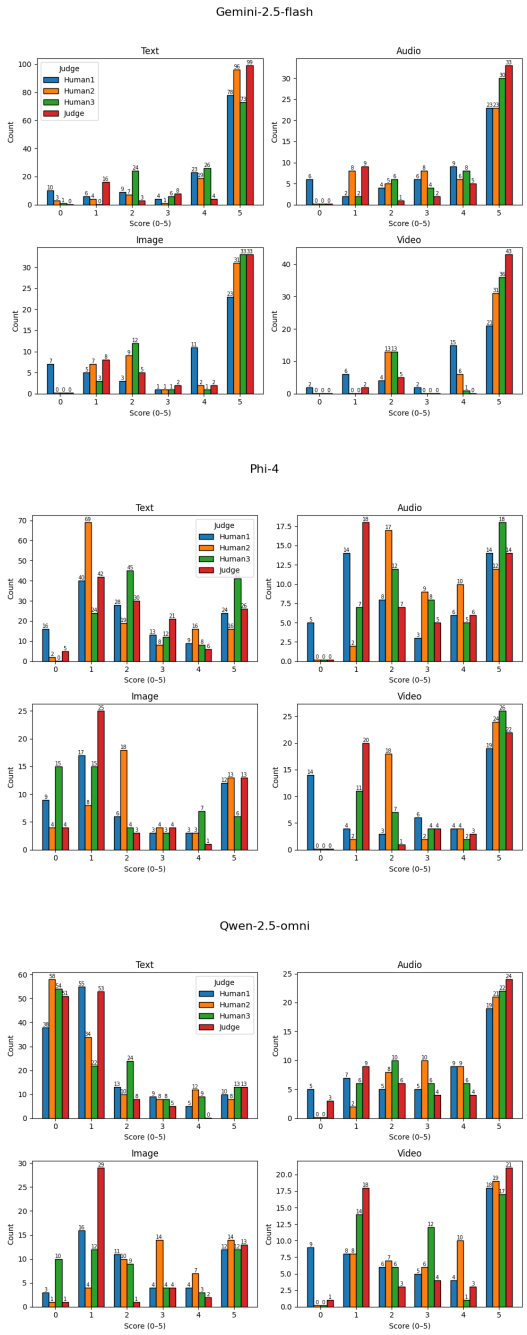

图2：人工与裁判的一致性在多模态大语言模型之间的对比。

5.2 案例研究

5.2.1 文本

我们展示一个典型的文本问答示例：模型需要根据一段简短的美式橄榄球比赛回顾，识别卡罗来纳的四分卫。段落明确写道：“四分卫 Jake Delhomme 完成了一次 3 码的达阵传球，目标是外接手穆辛·穆罕默德。”问题是：“卡罗来纳的

四分卫是谁？”如表7所示，Gemini-2.5 正确提取出 *Jake Delhomme* ，并获得评判器的满分。相比之下，通义千问-2.5 错误地声称无法回答该问题，尽管上下文中已明确给出答案，因此评判器将此情况标记为错误拒答并给出较低分数。这个示例表明，我们的评判器不仅会检查最终答案的准确性，还会在输入中可以找到真实答案时，对遗漏相关证据的情况予以扣分。

模型	预测	得分	错误类型
Gemini-2.5	Jake Delhomme	5	None
Qwen-2.5	不，这个问题无法回答。	1	错误拒答

表7：文本问答案例研究：根据比赛回顾段落识别卡罗来纳队的四分卫。

5.2.2 图像

在图像案例研究中，我们考察一道图表理解题，要求模型读取图中合成时间线图的横轴。3 并回答：“横轴上标注了多少个时代，它们分别叫什么？”真实图表包含四个时代：古代算法、中世纪方法、文艺复兴时期技术和现代模拟。如表8所示，Gemini-2.5 正确识别了时代数量及全部四个标签，因此评判器给予其最高分。相比之下，Phi-4 虽然也预测有四个时代，却虚构了完全不同的一组名称（例如，早期采用者、历史时期的采用者），这些名称在图表中完全没有出现。因此，评判器将其视为幻觉案例，仅给予部分分数。这个例子表明，我们的评估器不仅能识别粗略的数值正确性，也能判断答案是否在细粒度语义层面以视觉证据为依据。

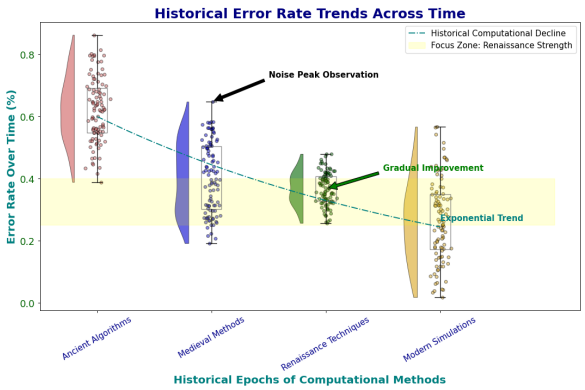

图3：图像案例研究的输入图表示例。

模型	预测	分数	错误类型
Gemini-2.5	共有 4 个标注的时代：古代算法——算法，中世纪方法，文艺复兴技术和现代模拟。	5	None
Phi-4	横轴标出了 4 个时代：早期采用者、历史时期的采用者、普遍采用和晚期采用者。	3	幻觉

表8：图像案例研究：读取图表横轴上标注的时代。

6 未来工作

未来工作的一个有前景方向，是将我们的评判器输出重新用作监督信号。每个实例都已包含一个标量分数、一种错误类型和一段自然语言解释，这些内容结合起来构成了丰富的信号，可用于构建从大语言模型反馈中进行强化学习的奖励模型，或作为宪法式 AI 框架中的一组软约束。我们不必止步于离线评估，还可以通过微调多模态大语言模型，使其最大化评判器分数，并减少幻觉和错误拒答等特定失效模式。将这一思路扩展到单轮问答之外，未来工作还可以把同样的反馈机制应用于多轮智能体场景，在此类场景中，评判器评估的是完整的交互轨迹，而不仅仅是最终答案。这样扩展后的流水线将使我们的评判器从静态转变为主动，从而实现动态地“从大语言模型中学习”。

7 结论

我们推出了一项多模态基准[1] ，并配备评判器，用于评估多模态大语言模型在文本、音频、图像和视频任务中的表现；该基准结合了李克特量表评分、结构化错误类型和自然语言解释。我们在 Gemini-2.5、Phi-4 和 Qwen-2.5 上进行的实验表明，评判器给出的评分与人工标注员的评分高度吻合，但略显保守，并且能够可靠地保留不同模态下各系统的相对排名。案例研究进一步表明，评判器能够以精细的粒度区分正确答案、幻觉答案和错误拒答。

[1]https://github.com/oscar-shih/multimodal_评判器_基准 https://huggingface.co/multi-judge

参考文献

Marah Abdin、Jyoti Aneja、Harkirat Behl、Sébastien Bubeck、Ronen Eldan、Suriya Gunasekar、Michael Harrison、Russell J Hewett、Mojan Javaheripi、Piero Kauffmann等。2024年。Phi-4 技术报告。*arXiv* 预印本 *arXiv:2412.08905*。

Triantafyllos Afouras、Joon Son Chung 和 Andrew Zisserman。2018。LRS3-TED：用于视觉语音识别的大规模数据集。

Aishwarya Agrawal、Jiasen Lu、Stanislaw Antol、Margaret Mitchell、C. Lawrence Zitnick、Dhruv Batra 和 Devi Parikh。2016。视觉问答：视觉问题回答。

Wasi Uddin Ahmad、Sean Narenthiran、Somshubra Majumdar、Aleksander Ficek、Siddhartha Jain、J ocelyn Huang、Vahid Noroozi 和 Boris Ginsburg。2025年。OpenCodeReasoning：推进竞赛编程的数据蒸馏。

Mu Cai、Reuben Tan、Jianrui Zhang、Bocheng Zou、Kai Zhang、Feng Yao、Fangrui Zhu、Jing Gu、Yiwu Zhong、Yuzhang Shang、Yao Dou、Jaden Park、Jianfeng Gao、Yong Jae Lee 和 Jianwei Yang。2024年。TemporalBench：为多模态视频模型的细粒度时间理解进行基准测试。

Nitay Calderon、Roi Reichart 和 Rotem Dror。2025年。大语言模型评审的替代标注者测试：如何从统计学上论证以大语言模型取代人工标注者。载于计算语言学协会第*63*届年会论文集（第*1*卷：长篇论文），第16051–16081页，奥地利维也纳。计算语言学协会。

Honglie Chen、Weidi Xie、Andrea Vedaldi 和 Andrew Zisserman。2020年。VGGSound：大规模音视频数据集。

Cheng-Han Chiang、Wei-Chih Chen、Chun-Yi Kuan、Chienchou Yang 和 Hung-yi Lee。2024年。将大语言模型用作作业评估器：对 1000+ 学生课程中的见解、反馈与挑战。载于*2024*年自然语言处理实证方法会议论文集，第2489–2513页，美国佛罗里达州迈阿密。计算语言学协会。

Cheng-Han Chiang 和 Hung-yi Lee。2023a。大语言模型能否成为 HumanEval 的替代方案？载于计算语言学协会第 *61* 届年会论文集（第 *1* 卷：长篇论文），第 15607–15631 页，加拿大多伦多。计算语言学协会。

Cheng-Han Chiang 和 Hung-yi Lee。2023b。深入探讨使用大语言模型进行自动评估。载于 计算语言学协会论文集：*EMNLP 2023*，第8928–8942页，新加坡。计算语言学协会。

Cheng-Han Chiang 和 Hung-yi Lee。2024年。大语言模型的过度推理与冗余计算。载于 第*18*届计算语言学协会欧洲分会会议论文集（第*2*卷：短篇论文），第161–169页，马耳他圣朱利安。计算语言学协会。

Joon Son Chung、Arsha Nagrani 和 Andrew Zisserman。2018。VoxCeleb2：深度说话人识别。载于 *Interspeech 2018*，第 1086–1090 页。

Dima Damen、Hazel Doughty、Giovanni Maria Farinella、Sanja Fidler、Antonino Furnari、Evangelos Kazakos、Davide Moltisanti、Jonathan Munro、Toby Perrett、Will Price 和 Michael Wray。2018。扩展第一人称视觉：EPIC-KITCHENS 数据集。

Florian E. Dorner、Vivian Y. Nastl 和 Moritz Hardt。2025年。 前沿领域可扩展评估的局限：大语言模型评判器无法胜过数据量翻倍。

Jennifer D'Souza、Hamed Babaei Giglou 和 Quentin Münch。2025年。 YESciEval：面向科学问答的稳健大语言模型评审。收录于 第*63*届计算语言学协会年会论文集（第*1*卷：长篇论文），第13749–13783页，奥地利维也纳。计算语言学协会。

Dheeru Dua、Yizhong Wang、Pradeep Dasigi、Gabriel Stanovsky、Sameer Singh 和 Matt Gardner。2019年。DROP：一种需要对段落进行离散推理的阅读理解基准。收录于*NAACL*论文集。

Chaoyou Fu、Peixian Chen、Yunhang Shen、Yulei Qin、Mengdan Zhang、Xu Lin、Jinrui Yang、Xiawu Zheng、Ke Li、XingSun、Yunsheng Wu、Rongrong Ji、Caifeng Shan 和 Ran He。2025年。MME：多模态大语言模型的综合评估基准。

Olga Golovneva、Moya Chen、Spencer Poff、Martin Corredor、Luke Zettlemoyer、Maryam Fazel-Zarandi 和 Asli Celikyilmaz。2023年。Roscoe：用于评估逐步推理的一套指标。

Yash Goyal、TejasKhot、Douglas Summers-Stay、Dhruv Batra 和 Devi Parikh。2017年。让 VQA 中的 V 发挥作用：提升图像理解在视觉问答中的作用。

Kristen Grauman、Andrew Westbury、Eugene Byrne、Zachary Chavis、Antonino Furnari、Rohit Girdhar、Jackson Hamburger、Hao Jiang、Miao Liu、Xingyu Liu、Miguel Martin、Tushar Nagarajan、Ilija Ra-dosavovic、Santhosh Kumar Ramakrishnan、Fiona Ryan、Jayant Sharma、Michael Wray、Mengmeng Xu、Eric Zhongcong Xu、Chen Zhao、Siddhant Bansal、Dhruv Batra、Vincent Cartillier、Sean Crane、Tien Do、Morrie Doulaty、Akshay Erapalli、Christoph Fe-ichtenhofer、Adriano Fragomeni、Qichen Fu、Abrham Gebreselasie、CristinaGonzalez、James Hillis、Xuhua Huang、Yifei Huang、Wenqi Jia、Weslie Khoo、

Jachym Kolar、Satwik Kottur、Anurag Kumar、Fed-erico Landini、Chao Li、Yanghao Li、Zhenqiang Li、Karttikeya Mangalam、Raghava Modhugu、Jonathan Munro、Tullie Murrell、Takumi Nishiyasu、Will Price、PaolaRuizPuentes、MereyRamazanova、Leda Sari、Kiran Somasundaram、Audrey Southerland、Yusuke Sugano、Ruijie Tao、Minh Vo、Yuchen Wang、Xindi Wu、Takuma Yagi、Ziwei Zhao、Yunyi Zhu、Pablo Arbelaez、David Crandall、Dima Damen、Gio-vanni Maria Farinella、Christian Fuegen、Bernard Ghanem、Vamsi Krishna Ithapu、C. V. Jawahar、Han-byul Joo、Kris Kitani、Haizhou Li、Richard New-combe、Aude Oliva、Hyun Soo Park、James M. Rehg、Yoichi Sato、Jianbo Shi、Mike Zheng Shou、Antonio Torralba、Lorenzo Torresani、Mingfei Yan，以及 Jiten-dra Malik。2022。Ego4D：3000 小时自我中心视频，纵览世界。

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang,Yao Hu, and Weidi Xie. 2025年。 Worldsense：评估多模态大语言模型对真实世界全模态的理解能力。

Chien-Yu Huang, Wei-Chih Chen, Shu wen Yang, Andy T. Liu, Chen-An Li, Yu-Xiang Lin, Wei- Cheng Tseng, Anuj Diwan, Yi-Jen Shih, Jiatong Shi, William Chen, Chih-Kai Yang, Wenze Ren, Xu- anjun Chen, Chi-Yuan Hsiao, Puyuan Peng, Shih- Heng Wang, Chun-Yi Kuan, Ke-Han Lu, Kai-Wei Chang, Fabian Ritter-Gutierrez, Kuan-Po Huang, Siddhant Arora, You-Kuan Lin, Ming To Chuang, Eunjung Yeo, Kalvin Chang, Chung-Ming Chien, Kwanghee Choi, Jun-You Wang, Cheng-Hsiu Hsieh, Yi-Cheng Lin, Chee-En Yu, I-Hsiang Chiu, Heitor R. Guimarães, Jionghao Han, Tzu-Quan Lin, Tzu-Yuan Lin, Homu Chang, Ting-Wu Chang, Chun Wei Chen, Shou-Jen Chen, Yu-Hua Chen, Hsi-Chun Cheng, Kunal Dhawan, Jia-Lin Fang, Shi-Xin Fang, Kuan-Yu Fang Chiang, Chi An Fu, Hsien-Fu Hsiao, Ching Yu Hsu, Shao-Syuan Huang, Lee Chen Wei, Hsi-Che Lin, Hsuan-Hao Lin, Hsuan-Ting Lin, Jian- Ren Lin, Ting-Chun Liu, Li-Chun Lu, Tsung-Min Pai, Ankita Pasad, Shih-Yun Shan Kuan, Suwon Shon, Yuxun Tang, Yun-Shao Tsai, Jui-Chiang Wei, Tzu-Chieh Wei, Chengxi Wu, Dien-Ruei Wu, Chao-Han Huck Yang, Chieh-Chi Yang, Jia Qi Yip, Shao- Xiang Yuan, Vahid Noroozi, Zhehuai Chen, Haibin Wu, Karen Livescu, DavidHarwath, Shinji Watanabe, and Hung yi Lee. 2025a。 Dynamic-SUPERB Phase-2：一个用于衡量语音语言模型能力的协作式扩展基准，包含180项任务。

Chien-Yu Huang、Ke-Han Lu、Shih-Heng Wang、Chi- Yuan Hsiao、Chun-Yi Kuan、Haibin Wu、Siddhant Arora、Kai-Wei Chang、Jiatong Shi、Yifan Peng、Roshan Sharma、ShinjiWatanabe、Bhiksha Ramakr- ishnan、Shady Shehata 和 Hung yi Lee。2024a。Dynamic-SUPERB：迈向动态、协作且全面的语音指令微调基准。

Chien-Yu Huang、Min-Han Shih、Ke-Han Lu、Chi-Yuan Hsiao 和 Hung-Yi Lee。2025b。SpeechCaps：通过多说话人说话风格描述推进基于指令的通用语音模型。载于*ICASSP 2025 - 2025*年 *IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第 1–5 页。

Shih-Cheng Huang、Shih-Heng Wang、Min-Han Shih、Saurav Sahay 和 Hung-yi Lee。2024b。预训练语言模型预热在参数高效微调中的系统性分析。载于*2024*年北美计算语言学协会分会会议论文集：人类语言技术（第*4*卷：学生研究研讨会），第1–7页，墨西哥城，墨西哥。计算语言学协会。

Yuxin Jiang、Yufei Wang、Xingshan Zeng、Wanjun Zhong、Liangyou Li、Fei Mi、Lifeng Shang、Xin Jiang、Qun Liu 和 Wei Wang。2023年。Follow-bench：面向大语言模型的多层次细粒度约束遵循基准。

Chris Dongjoo Kim、Byeongchang Kim、Hyunmin Lee 和 Gunhee Kim。2019年。AudioCaps：为真实场景中的音频生成描述。载于计算语言学协会北美分会*2019*年会议论文集：人类语言技术，第*1*卷（长短论文），第119–132页，明尼阿波利斯，明尼苏达州。计算语言学协会。

You Jin Kim、Hee-Soo Heo、Soyeon Choe、Soo-Whan Chung、Yoohwan Kwon、Bong-Jin Lee、Youngki K won 和 Joon Son Chung。2021。看看是谁在说话：真实场景中的主动说话人检测。载于*Interspeech 2021*，第3675–3679页。

BohaoLi, RuiWang, GuangzhiWang, Yuying Ge, Yix- iao Ge, and Ying Shan. 2023年 Seed-Bench：使用生成式理解能力对多模态大语言模型进行基准测试。

Rensis Likert. 1932年。一种测量态度的技术。《心理学档案》。

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, ZiweiLiu, Kai Chen, and Dahua Lin. 2024a Mmbench：你的多模态模型是全能选手吗?

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and LuHou. 2024b Tempcompass：视频大语言模型真的理解视频吗?

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022 ChartQA：通过视觉和逻辑推理回答有关图表的问题的基准。载于 计算语言学协会论文集：*ACL 2022*，第 2263–2279 页，都柏林，爱尔兰。计算语言学协会。

Minesh Mathew、Dimosthenis Karatzas 和 C. V. J awa- har。2021。Docvqa：用于文档图像视觉问答的数据集。

Ivan Moshkov、Darragh Hanley、Ivan Sorokin、Shubham Toshniwal、Christof Henkel、Benedikt Schifferer、Wei Du 和 Igor Gitman。2025年。Aimo-2 获胜方案：利用 OpenMathReasoning 数据集构建最先进的数学推理模型。*arXiv* 预印本 *arXiv:2504.16891*。

Arsha Nagrani、Joon Son Chung 和 Andrew Zisser- man。2017年。Voxceleb：一个大规模说话人识别数据集。载于*Interspeech 2017*，第 2616–2620 页。

Le Thien Phuc Nguyen、Zhuoran Yu、Samuel Low Yu Hang、Subin An、Jeongik Lee、Yohan Ban、SeungEun Chung、Thanh-Huy Nguyen、JuWan Maeng、Soochahn Lee 和 Yong Jae Lee。2025年。看、听与理解：对多模态大语言模型中的视听人类语音理解进行基准测试。

通义千问：An Yang、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chengyuan Li、Dayiheng Liu、Fei Huang、Haoran Wei、Huan Lin、Jian Yang、Jianhong Tu、JianweiZhang、Jianxin Yang、Jiaxi Yang、Jingren Zhou、Junyang Lin、Kai Dang、Keming Lu、Keqin Bao、Kexin Yang、Le Yu、Mei Li、Mingfeng Xue、Pei Zhang、Qin Zhu、Rui Men、Runji Lin、Tianhao Li、Tianyi Tang、Tingyu Xia、Xingzhang Ren、Xuancheng Ren、Yang Fan、Yang Su、Yichang Zhang、Yu Wan、Yuqiong Liu、Zeyu Cui、Zhenru Zhang 和 Zihan Qiu。2025年。Qwen2.5 技术报告。

David Rein、Betty Li Hou、Asa Cooper Stickland、Jackson Petty、Richard Yuanzhe Pang、Julien Dirani、Julian Michael 和 Samuel R. Bowman。2023年。Gpqa：研究生级谷歌检索不可解问答基准。

Maarten Sap、Hannah Rashkin、Derek Chen、Ronan LeBras 和 Yejin Choi。2019年。Social i qa：关于社会互动的常识推理。收录于*EMNLP*。

Min-Han Shih、Ho-Lam Chung、Yu-Chi Pai、Ming-Hao Hsu、Guan-Ting Lin、Shang-Wen Li 和 Hung yi Lee。2024年。GSQA：用于生成式口语问答的端到端模型。收录于*Interspeech 2024*，第2970–2974页。

Yansong Tang、Dajun Ding、Yongming Rao、Yu Zheng、Danyang Zhang、Lili Zhao、Jiwen Lu 和 Jie Zhou。2019年。Coin：用于综合性教学视频分析的大规模数据集。收录于 *2019 IEEE/CVF* 计算机视觉与模式识别会议（*CVPR*），第1207–1216页。

Bin Wang、Xunlong Zou、Geyu Lin、Shuo Sun、Zhuohan Liu、Wenyu Zhang、Zhengyuan Liu、AiTi A w 和 Nancy F Chen。2025a。Audiobench：音频大语言模型的通用基准。*NAACL*。

KeWang、JuntingPan、Weikang Shi、ZimuLu、Houxing Ren、Aojun Zhou、Mingjie Zhan 和 Hongsheng Li。2024a。使用 Math-Vision 数据集衡量多模态数学推理。载于第三十八届神经信息处理系统会议数据集与基准赛道。

Ke Wang、Junting Pan、Linda Wei、Aojun Zhou、Weikang Shi、Zimu Lu、Han Xiao、Yunqiao Yang、Houxing Ren、Mingjie Zhan 和 Hongsheng Li。2025b。Mathcoder-VL：连接视觉与代码，提升多模态数学推理能力。载于第*63*届计算语言学协会年会。

Zirui Wang、Mengzhou Xia、Luxi He、Howard Chen、Yitao Liu、Richard Zhu、Kaiqu Liang、XindiWu、Haotian Liu、Sadhika Malladi、Alexis Chevalier、Sanjeev Arora 和 Danqi Chen。2024b。Charxiv：探索多模态大语言模型在真实图表理解方面的不足。

Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Brian Ichter、Fei Xia、Ed Chi、Quoc Le 和 DennyZhou。2023年。思维链提示能够激发大语言模型的推理能力。

Junbin Xiao、Xindi Shang、Angela Yao 和 Tat- Seng Chua。2021。Next-qa：从问答迈向时间动作解释的新阶段。

Yuwei Yang、Zeyu Zhang、Yunzhong Hou、Zhuowan Li、GaowenLiu、AliPayani、Yuan-Sen Ting 和 Liang Zheng。2025年。用于提升多模态大语言模型图表理解能力的有效训练数据合成。载于 *IEEE/CVF* 国际计算机视觉大会（*ICCV*）论文集。

Qinyuan Ye、Bill Yuchen Lin 和 Xiang Ren。2021。Crossfit：一项旨在提升自然语言处理中跨任务泛化能力的少样本学习挑战。

Xi Ye 和 Greg Durrett。2022。用于文本推理的少样本提示中解释的不可靠性。

Rowan Zellers、Ari Holtzman、Yonatan Bisk、Ali Farhadi 和 Yejin Choi。2019年。Hellaswag：机器真的能续写你的句子吗？载于第*57*届计算语言学协会年会论文集。

Lianmin Zheng、Wei-LinChiang、Ying Sheng、Siyuan Zhuang、Zhanghao Wu、Yonghao Zhuang、Zi Lin、Zhuohan Li、Dacheng Li、Eric P. Xing、Hao Zhang、Joseph E. Gonzalez 和 Ion Stoica。2023年。使用 MT-Bench 和 Chatbot Arena 评判大语言模型评审。

Jeffrey Zhou、Tianjian Lu、Swaroop Mishra、Siddhartha Brahma、Sujoy Basu、Yi Luan、Denny Zhou 和 Le Hou。2023年。 大语言模型的指令遵循评估。

Nannan Zhu、Yonghao Dong、Teng Wang、Xueqian Li、Shengjun Deng、Yijia Wang、Zheng Hong、Tiantian Geng、Guo Niu、Hanyan Huang、Xiongfei Yao 和 Shuaiwei Jiao。2025年。 Cvbench：评估复杂多模态理解和推理中的跨视频协同。