arXiv:2603.00993v1 [cs.AI] 1 Mar 2026

CollabEval：通过多智能体协作增强大语言模型评审

Yiyue Qian [1*], **Shinan Zhang** [1*], **Yun Zhou** [2], **Haibo Ding** [2], **Diego Socolinsky** [1], **Yi Zhang** [2]

{iamyiyue, shinanz, yunzzhou, hbding, sclinsky,yizhngn}@amazon.com[1]
Amazon AWS 生成式人工智能创新中心[2]Amazon AWS Bedrock

摘要

大语言模型（LLM）彻底改变了人工智能生成内容的评估方式，大语言模型评审范式也日益普及。然而，当前的单一大语言模型评估方法面临诸多挑战，包括判定不一致以及预训练数据带来的固有偏见。为解决这些局限性，我们提出**CollabEval**，一种新颖的多智能体评估框架，通过三个阶段实现协作式**评估**流程：初始评估、多轮讨论和最终判定。与依赖竞争式辩论或单模型评估的现有方法不同，CollabEval 强调多个智能体之间的协作，并通过策略性共识检查提升效率。大量实验表明，CollabEval 在多个维度上始终优于单一大语言模型方法，即使个别模型表现不佳，也能保持稳健性能。该框架全面支持多种评估标准，并通过协作式设计确保效率。

引言

大语言模型（LLM）的迅速发展彻底改变了人工智能生成内容评估，使大语言模型评审范式日益普及（Chiang 等人，2024；Wang 等，2024b；Raina、Liusie 和 Gales，2024；Chan 等，2024；Li 等，2025；Ma 等，2025b；Song 等，2025；Ma 等，2025a；Qian 等，2025）。近期研究展示了使用单个大语言模型作为评估者的潜力，其中一项重要工作（Bai 等人，2024）推出了 MT-bench，用于系统性的大语言模型评估；另一项工作（Chiang 等人，2024）则开发了 Chatbot Arena，将其打造为一个通过人类偏好对齐来评估大语言模型的开放平台。这些方法在自动化评估流程方面取得了令人鼓舞的成果，涵盖连贯性、相关性和流畅性等多个维度。然而，近期研究也发现，当前评估方法存在显著局限。一项近期研究（Raina、Liusie 和 Gales，2024）指出，基于大语言模型的评估容易受到通用对抗攻击，引发了人们对其可靠性的担忧。此外，Wang 等人（Wang 等，2024b）发现，自训练评估器往往难以做到评估一致且客观，这凸显了构建更稳健评估框架的迫切需求。

总的来说，当前的评估方法面临几项关键挑战：(i) 单一大语言模型评估由于其预训练数据和知识带来的固有偏见，缺乏稳健性（Huang 等人，2024年）。近期研究（Bai 等人，2024年；Huang 等人，2024年）发现，不同大语言模型评估器的判断质量存在显著差异；Chat-Eval（Chan 等，2024）进一步指出，单个大语言模型可能在某些维度表现出色，却在其他维度表现欠佳。(ii) 尽管近期研究（Chan 等，2024；Chen、Saha 和 Bansal，2023年；Chern 等人，2024年）开发了基于智能体的框架来应对这些局限，其中 ChatEval（Chan 等，2024）尤其采用了多个辩论智能体进行评估，但这些方法往往缺乏适应多样化评估场景所需的灵活性和效率。这些挑战凸显了构建更稳健、更灵活的评估框架的必要性。

为应对这些局限，我们提出了**CollabEval**，一种新颖的多智能体评估框架，它实现了结构化（即三阶段）的协作式评估过程。与以往方法（Chan 等，2024；Chen、Saha 和 Bansal，2023）不同，我们的框架强调协作而非竞争式辩论，从而回应了（Verga 等人，2024年）所指出的评估中需要多样化模型视角的需求。具体而言，CollabEval 采用精密的三阶段设计：（1）初始评估，不同智能体独立评估内容；（2）多轮协作讨论，智能体通过结构化对话分享并完善各自的评估，其中包括置信度分数、一致意见、不一致意见和推理过程；（3）最终判定，基于先前讨论做出最终评估决策。值得注意的是，CollabEval 在每个阶段执行共识检查，从而在达成一致时允许提前终止，因此相较现有的基于智能体的 LLM 评审方法更能确保效率。我们工作的主要贡献包括：

- **新颖之处**：我们提出了一种三阶段评估框架，独特地将独立评估与智能体之间的协作式完善相结合。
- **全面**：CollabEval 支持基于标准的评估和跨多个维度的成对比较，通过广泛的实验验证，证明其性能优于单个大语言模型评估。

*这些作者对本文的贡献相同。版权所有 ©2025年，美国人工智能促进协会（www.aaai.org）。保留所有权利。

• **稳健且高效**：即使个别大语言模型表现不佳，我们的框架仍能保持出色的性能，同时通过策略性地检查共识并提前终止，确保效率。

相关工作

大语言模型评审。近年来，大语言模型取得了长足进步，大语言模型评审范式也因此日益广泛地应用于人工智能生成内容的评估。Bai 等人提出了 MT-bench，将其作为大语言模型评估的系统性框架，并建立了用于从多个维度评估模型表现的基准测试（Bai 等人，2024）。Chiang 等人开发了 Chatbot Arena，将其打造为一个开放平台，利用人类偏好对齐来评估大语言模型，展现了结构化评估框架的潜力（Chiang 等人，2024）。然而，现有的单一大语言模型方法存在显著局限。Raina 等人揭示了基于大语言模型的评估易受通用对抗攻击的严重漏洞（Raina、Liusie 和 Gales，2024）；Wang 等人则表明，自训练评估器难以做到一致和客观（Wang 等，2024b）。Huang 等人进一步指出，由于预训练数据和知识带来的固有偏差，单一大语言模型评估往往缺乏稳健性（Huang 等人，2024），不同大语言模型评估器的判断质量也存在显著差异。

大语言模型中的多智能体。近期研究探索了多智能体方法（Chen、Saha 和 Bansal，2023；Hong 等人，2024；Shah 等人，2024；Wang 等人，2024a；Wu 等人，2024；Zhang 等人，2024；Han 等人，2024），以提升大语言模型在各种任务中的能力。例如，ReConcile（Chen、Saha 和 Bansal，2023）是一种通过不同大语言模型之间的圆桌会议来改进推理过程的框架。该方法通过多轮讨论，让大语言模型智能体协作推理，并采用置信度加权投票机制以达成更好的共识。

在大语言模型评审这一背景下，一些研究（Zhuge 等，2024年；Chan 等，2024；Chern 等人，2024年；Rasheed 等人，2024年）探索了多智能体评估框架。ChatEval（Chan 等，2024）通过部署多个辩论智能体，自主讨论并评估人工智能生成内容。研究表明，通过辩论进行协作评估能够带来更可靠的评价。此外，Chern 等人还探讨了智能体辩论在元评估中的潜力（Chern 等人，2024年）。这些方法表明，多智能体评估系统能够有效克服单一大语言模型评委的局限，尤其是在稳健性和一致性方面。然而，许多现有方法过度依赖竞争式辩论，而非协作，这可能导致评估过程效率低下。这一局限促使我们开展 CollabEval 研究，该方法强调协作而非竞争式辩论，以实现更可靠、高效的评估。

拟议框架

本节将详细介绍 CollabEval 的三个主要阶段：初始评估、智能体之间的多轮协作讨论，以及最终判断。

初始评估

单个大语言模型评估器往往存在固有偏见，这些偏见源于其预训练数据和知识库。不同大语言模型的预训练数据集和知识各不相同，这些差异与偏见相结合，在评估人工智能生成内容时可能导致判断不一致。为了解决这些局限性并利用不同大语言模型的多样性，我们提出了一种多智能体协作评估框架。

在阶段 1 中，如图1所示，CollabEval 会使用多个独立评估者进行初始评估。每位评估者都会独立分析内容，并提供三个关键要素：评估结果、置信度分数，以及对其评估的详细理由。这种独立评估确保能够捕捉每个智能体的独特视角和能力，不受其他评估者影响。所有评估者完成初始评估后，CollabEval 会进行共识检查，以确定评估者的判断是否达成一致。如果达成共识，系统便会返回最终评估结果，体现出提前终止的效率。然而，如果评估者之间存在重大分歧，流程将进入阶段 2，评估者通过多轮讨论来解决分歧并完善评估。

多轮讨论

智能体协作。 在阶段 2 中，评估者会彼此分享初始评估、置信度分数和理由。协作重点在于找出评估中的一致意见和不一致意见。每位评估者都会审阅其他人的评估，并根据集体见解提供更新后的评估。通过这一过程，评估者能够结合多种视角，并解决初始评估中可能存在的偏见或疏漏，从而完善自己的判断。

迭代过程。 讨论以迭代方式进行，评估者利用初始评估和持续讨论中的所有可用数据来完善评估。每位评估者都会考虑：

- 所有智能体的初始评估结果
- 此前轮次的置信度分数
- 其他评估者意见一致和不一致之处

- 其他评估者提供的理由

例如，如图1所示，在第1轮讨论中，评估者1首先分析所有初始评估结果，并给出更新后的评估，明确指出一致意见和不一致意见。随后，评估者2综合考虑初始评估和评估者1更新后的评估，再给出完善后的评估。最后，评估者3审阅此前的所有评估——包括

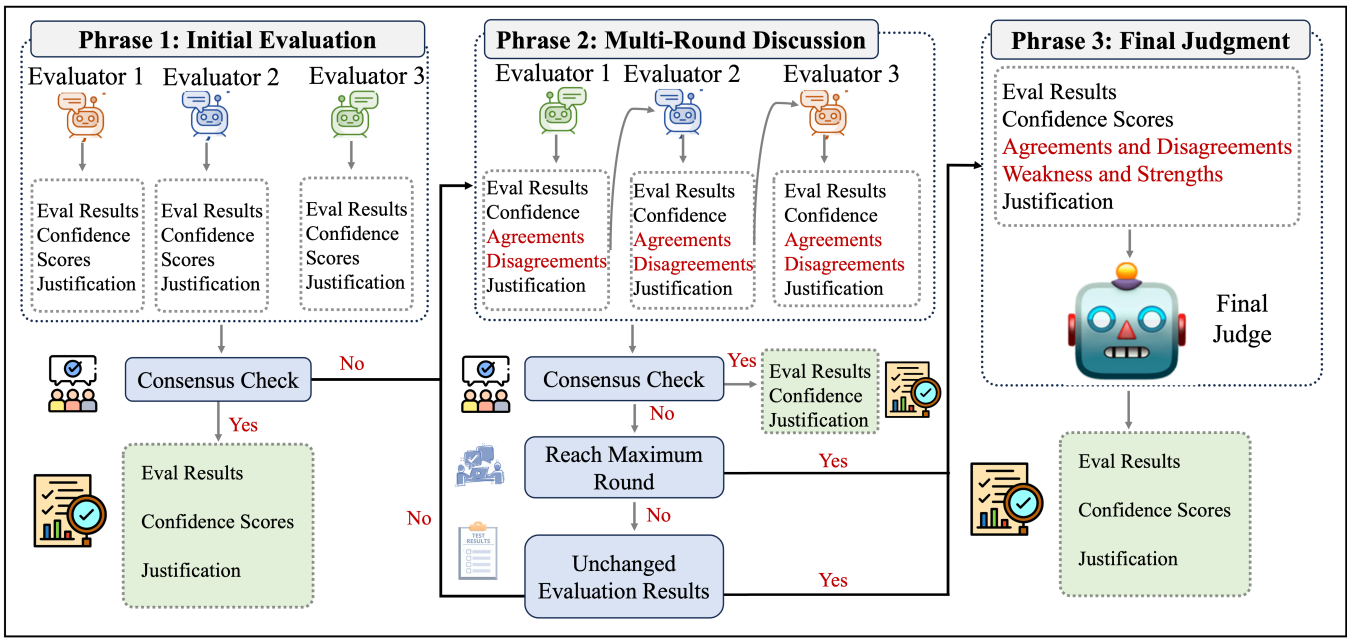


图1：CollabEval 框架由三个主要阶段组成：(a) 阶段 1：初始评估——三位评估者独立评估内容，并提供评估结果、置信度分数和理由。随后进行共识检查；如果达成共识，则返回最终结果，否则进入阶段 2。(b) 阶段 2：多轮讨论——评估者通过协作讨论分享一致意见、不一致意见和理由。每轮讨论结束后都会进行共识检查。如果达成共识，则返回结果；如果未达成共识，系统会在继续之前检查是否已达到最大轮数或结果是否未变。(c) 阶段 3：最终判定——如果讨论仍无法达成共识，最终裁判会分析此前的所有评估结果、置信度分数、一致意见／不一致意见和理由，作出最终评估决定。

初始评估和更新后的评估——然后给出自己的评估。为减轻模型能力可能造成的偏差，我们会在讨论轮次中随机打乱评估者的顺序。

共识检查。每轮讨论结束后，CollabEval 会进行三项关键检查，以确定评估流程的后续步骤。首先，系统会检查所有评估者是否已在当前轮讨论中就评估达成共识。如果达成共识，系统将返回最终结果。否则，CollabEval 会进一步核查另外两个条件：是否已达到最大讨论轮数，以及评估结果是否与上一轮相比保持不变。如果满足这两个条件中的任意一个——即达到最大轮数，或评估结果保持不变——流程便会进入最终评估阶段。然而，如果这两个条件均未满足，CollabEval 将启动新一轮讨论，进一步完善评估结果。

最终裁判评估

当多轮讨论未能达成共识或评估结果保持不变时，CollabEval 会启用一个强大的模型作为最终裁判。最终裁判会分析此前各轮的所有评估结果、置信度分数和理由，以及各方评估者之间的共识与分歧，以及评估在讨论轮次中的变化过程。最终裁判会综合考虑这些信息，作出果断的评估，兼顾评估过程中提出的所有观点和推理过程。

实验

在本节中，我们全面评估 CollabEval 在两种不同评估模式下的表现：基于标准的评估和成对比较。我们使用三个基准数据集开展实验，以评估该框架的性能。最后，我们讨论从实验中得出的重要发现和见解。

实验设置

评估模式。为了全面评估 CollabEval 的能力，我们在两种不同的评估模式下开展实验：

基于标准的评估：此模式从多个预先定义的维度评估内容，例如连贯性、一致性、流畅性和相关性。每个维度均按照特定量表评分，从而能够对生成内容的不同方面进行细粒度评估。

成对比较：在此模式下，评估者通过直接比较两个回答，判断哪个更好。这种方法特别适用于相对质量评估，

并有助于确定不同模型输出之间的偏好排名。

数据集。我们使用了三个基准数据集，其中一个是基于标准的数据集（即 SummEval 数据集（Fabbri 等人，2021）），另外两个是成对比较数据集（即 chatbot_arena_conversationdataset（LMSYSOrg 2023）和 lmsys_arena_human_preference_55k dataset（Chiang 等人，2024））。接下来，我们将进一步介绍这些基准数据集。

基于标准的数据集：我们首先使用 SummEval（Fabbri 等人，2021），这是一个全面的基准数据集，包含从 100 篇源新闻文章生成的 1600 个示例。这些摘要由 16 个不同的模型生成，涵盖了丰富多样的生成质量与风格。每篇摘要都由 8 位专家标注人员从四个关键维度进行严格评估：连贯性、一致性、流畅性和相关性。评分系统采用 1 到 5 的五分制，可对质量进行细粒度评估。数据集采用详细的格式组织，包括 (id, machine_summary, source_news, coherence_score, consistency_score, fluency_score, relevance_score)，从而能够分别对每个维度进行全面分析。

成对比较数据集：在成对比较评估中，我们采用两个不同的数据集。具体而言，对于 chatbot_arena_conversations（LMSYSOrg 2023）数据集，我们并未使用全部数据，而是随机选取了 1,000 个样本。该数据集侧重于在对话场景中对不同模型响应进行直接比较。此外，对于 lmsys_arena_human_preference_55k 数据集（Chiang 等人，2024年），我们同样使用了 1,000 个随机样本。该数据集尤为重要，因为它纳入了人类偏好判断，为评估提供了可靠的真实标注。两个数据集均遵循 (id, query, response_a, response_b, winner) 的标准化格式，从而能够对两种备选回答进行直接比较，并清晰识别更优选项。

基线。在本研究中，我们将模型 CollabEval 与两组基线方法进行比较，包括单一大语言模型评审和基于智能体的 LLM 评审。

B1：单一大语言模型评审：该基线代表传统方法，即由单个大语言模型独立评估内容。我们实现了多种先进模型，包括 Mistral Large（AI 2024b）、Claude Haiku（Anthropic 2024b）、Claude Sonnet 3（Anthropic 2024b）和 Llama 3 70b（AI 2024a），作为独立评估者。每个模型都作为独立裁判，以展示依赖单个模型判断的能力，并建立用于比较的性能基准。

B2：基于智能体的 *LLM* 评审：对于基于智能体的方法，我们还探索了另一种圆桌讨论机制，称为表 1 中的 Round-Table Agents Eval。我们没有采用三阶段机制，而是在本研究中采用圆桌机制（Chen、Saha 和 Bansal，2023），并实现了一种顺序圆桌讨论，让智能体依次评估内容。具体而言，对于每项评估任务，我们随机选择一名智能体提供初步评估。下一名智能体随后审阅这项评估，提出自己的评估，并选择同意或修订之前的评估。这个过程会依次经过所有智能体。为减少智能体顺序可能造成的偏差，我们会针对每项新的评估任务随机打乱智能体的顺序。讨论会持续进行，直到所有智能体达成共识或达到最大轮数。如果达到最大轮数后仍未达成共识，则采用多数投票机制确定最终评估结果。这种更复杂的基线采用圆桌讨论方式，让多个大语言模型参与协作评估。该方法是单智能体方法与我们提出的 CollabEval 方法之间的一个过渡方案。

实验设置。CollabEval 采用多个先进的大语言模型智能体（Mistral Large、Claude Haiku、Claude Sonnet 和 Llama 3 70b）进行评估。在阶段 1 中，每个智能体独立提供初始评估。为减轻模型顺序可能带来的偏差，我们在初始评估和多轮讨论阶段均随机打乱评估者的顺序。在阶段 2 中，若最初未达成共识，评估过程将通过多轮讨论继续进行，最多 3 轮。若讨论阶段结束后仍未达成共识，我们在阶段 3 中采用 Claude Sonnet 3.5（Anthropic 2024a）作为最终裁判，利用其强大的推理能力分析完整的评估历史并作出最终决定。

性能讨论

关于基于标准的评估的讨论。 表 1 展示了各方法在 SummEval 数据集上进行基于标准的评估的比较结果。该表采用了若干关键指标来评估性能。**准确率** 衡量评估结果与真实标签相比正确的比例。**平均轮数** 表示评估者达成共识所需的讨论迭代次数。**差距比率**(1-4)衡量所有误评样本中，大语言模型评分与真实标签之间绝对差值为相应数值的样本比例，其中差距 1 表示相差 1 分，差距 2 表示相差 2 分，以此类推。评估偏差通过**过度评估比率**和**低估比例**来衡量：前者表示所有误评样本中，大语言模型评分高于真实标签的样本比例；后者则表示评分低于真实标签的样本比例。

我们的实验结果表明，CollabEval 在所有评估维度上的表现均优异。在相关性评估中，CollabEval 的准确率达到 49.5%，平均轮数为 2.073，差距 1 比率最高（87.8%），且严重误判极少（差距 3 为 0.5%，差距 4 为 0%）。连贯性评估显示，CollabEval 表现稳健，准确率为 40.4%，错误分布均衡（差距 1 为 77.8%，差距 2 为 20.8%，Gap 3 为 1.5%），显著优于单一大语言模型方法。在流畅性评估中，CollabEval 的准确率具有竞争力（46.9%），错误分布（差距 1 为 77.8%，差距 2 为 18.0%，差距 3 为 4.5%）也优于分布更分散的单一大语言模型 Sonnet。在一致性评估中，

表 1：CollabEval 与单个大语言模型评审在 SummEval 数据集上进行基于标准的评估的比较结果。每个维度的最佳准确率以粗体显示。

模型 设置	准确率 (%)	Avg 轮次	差距 1 比例 (%)	差距 2 比例（%）	Gap 3 比例（%）	差距 4 比例（%）	过高- 评估比例（%）	不足- 评估比例（%）
相关性								
CollabEval	**49.5**	2.073	87.8	12.0	0.5	0	31.9	68.1
单一大语言模型 Sonnet	47.7	1	85.5	13.7	1.6	0	29.7	70.3
单一大语言模型 Haiku	47.6	1	84.9	14.7	1.1	0	30.2	69.8
单一大语言模型 Llama3	22.8	1	76.7	23.3	0.0	0	100.0	0.0
连贯性								
CollabEval	**40.4**	2.343	77.8	20.8	1.5	0	63.3	36.7
单一大语言模型 Sonnet	37.4	1	71.4	23.9	4.9	0	66.4	33.6
单一大语言模型 Haiku	38.9	1	76.9	22.4	0.8	0	63.4	36.6
单一大语言模型 Llama3	29.5	1	77.0	22.0	2.2	0	25.4	74.6
流畅性								
CollabEval	**46.9**	2.103	77.8	18.0	4.5	0	21.9	78.1
单一大语言模型 Sonnet	46.8	1	65.9	24.0	21.4	5	29.7	70.3
单一大语言模型 Haiku	13.8	1	75.9	22.3	6.2	0	30.2	69.8
单一大语言模型 Mistral	45.8	1	86.7	13.3	0.0	0	25.0	75.0
一致性								
CollabEval	**48.2**	2.181	79.6	18.2	7.0	0	10.2	89.8
单一大语言模型 Sonnet	46.9	1	65.8	25.2	19.8	0	10.4	89.6
单一大语言模型 Haiku	12.6	1	77.7	20.9	5.3	0	4.7	95.3
单一大语言模型 Mistral	55.9	1	93.8	5.4	2.8	0	24.4	75.6

CollabEval 的准确率达到 48.2%，错误分布最为均衡（差距 1 为 79.6%，差距 2 为 18.2%，差距 3 为 7% ）。

值得注意的是，尽管需要更多计算轮次（平均为 2.073-2.343），CollabEval 在所有维度上的高估/低估比率始终更为均衡，尤其是在相关性（31.9%/68.1%）和连贯性（63.3%/36.7%）方面。这表明，与单一大语言模型方法相比，CollabEval 有了显著改进；例如，Llama3 在相关性评估中的高估/低估比例极端地达到 100%/0%。对高估和低估比率的分析揭示了不同大语言模型在评估倾向方面各自鲜明的行为模式。最显著的是，Llama3 表现出极端的评估模式，在相关性维度上的高估率达到 100%，说明其持续倾向于给出更高分数。与此形成鲜明对比的是，在连贯性评估中，它表现出 74.6% 的低估倾向。Sonnet 和 Haiku 等其他大语言模型则表现得更为温和，高估与低估比率较为均衡。CollabEval 的评估模式最为均衡，展现了其通过协作式评估缓解极端评估偏差的能力。

关于成对比较评估的讨论。表 2 展示了在两个 Arena 数据集上，多智能体与单个大语言模型评审进行成对评估的比较结果。该表采用四项关键指标来评估模型表现。**准确率**衡量预测正确的比例，即与真实标注一致的预测占比。**平均轮数**表示达成共识所需的讨论迭代次数。**真实标注_胜出_预测_平局比例** 表示在所有误评样本中，真实标注显示存在明确胜者、但模型预测所有选项打平的样本比例，而**真实标注_平局_预测_胜出比例** 表示在所有误评样本中，真实标注为平局、但模型预测有一方胜出的样本比例。

表2中的实验结果表明，CollabEval在两个数据集上均表现出更优的性能。在Chatbot Arena数据上，CollabEval以1.542的平均轮数取得了60.2%的最高准确率，显著优于Round-Table Eval（57.7%）和单一大语言模型方法。CollabEval展现出均衡的评估能力，GT_Win_Pred_Tie比例为50.00%，而GT_Tie_Pred_Win比例仅为2.63%，表明其在模糊案例中具有判别能力。对于评估难度更高的Arena Human Preference数据，CollabEval以51.5%的准确率和1.517的平均轮数保持性能优势，而Round-Table Eval为48.7%，单一大语言模型方法则介于48.4%至50.5%之间。尽管单一大语言模型Llama3 70b的准确率具有竞争力，但其极端比例（Arena数据为53.85%/0.00%，Preference数据为55.47%/0.39%）表明其决策可能存在偏差。单一大语言模型Sonnet的表现更为均衡，但准确率较低（48.4%），且GT_Tie_Pred_Win比例更高（13.95%），表明其倾向于在模糊案例中做出确定性判断。这些结果一致表明，尽管CollabEval的多智能体方法需要额外的计算轮次，但与圆桌评估和单一大语言模型评估方法相比，它能提供更可靠、更均衡的评估。

研究结果与分析

讨论轮次。讨论轮次对 CollabEval 性能的影响揭示了若干关键模式和

表2：在两个 Arena 数据集上，针对成对比较评估，多智能体与单个大语言模型评审的比较结果。每个数据集上准确率最高的结果以粗体显示。

模型设置	准确率（%）	平均轮数	GT_胜 _预测_平局比例（%）	GT_平局_预测_胜比例（%）
Chatbot Arena 数据				
CollabEval（我们的方法）	**60.2**	1.542	50.00	2.63
圆桌智能体评估	57.7	1.214	15.84	43.97
单一大语言模型 Mistral Large	58.2	1	45.54	4.22
单一大语言模型 Haiku	57.2	1	46.30	3.38
单一大语言模型 Llama3 70b	59.7	1	53.85	0.00
Arena 人类偏好数据				
CollabEval（我们的方法）	**51.5**	1.517	53.20	9.07
圆桌智能体评估	48.7	1.258	12.70	47.37
单一大语言模型 Sonnet	48.4	1	48.06	13.95
单一大语言模型 Mistral Large	50.5	1	54.95	5.25
单一大语言模型 Llama3 70b	48.8	1	55.47	0.39

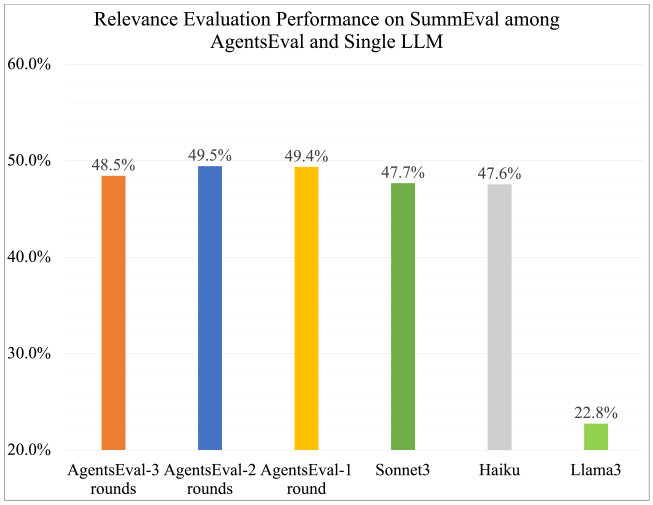


图 2：相关性评估的准确率表现分析。

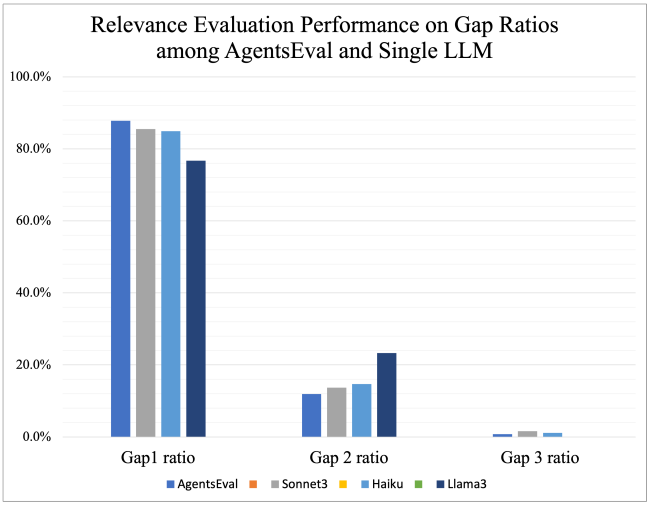


图 3：相关性评估的 Gap 比率表现分析。

潜在因素。在相关性评估中，如图 2 所示，我们观察到准确率从一轮到三轮呈现逐步提升后回落的趋势：CollabEval 进行 1 轮时准确率为 49.4%，进行 2 轮时升至 49.5%，进行 3 轮时则略微降至 48.5%。这一模式体现了效率与准确率之间的权衡：初期协作带来显著提升，但超过两轮后收益递减。

收益递减现象可归因于几个关键机制。首先，随着评估者在早期轮次中交流了大部分关键见解，信息饱和随之出现，后续轮次带来的新视角十分有限。表 1 中的差距 1 比率分析印证了这一点：CollabEval 达到 87.8%，高于单模型表现（Sonnet：85.5%，Haiku：84.9%），说明评估结果的主要改进发生在早期。其次，与单一大语言模型表现（Sonnet：47.7%，Haiku：47.6%）相比，即使 CollabEval 只进行一轮讨论（49.4%），表现也优于这些基线，表明多智能体框架的主要优势来自初始评估和第一轮讨论。

这些发现对实际部署具有重要意义：尽管增加讨论轮次可以完善评估，但最佳配置应在模型表现与合理的计算开销之间取得平衡。这一认识与 CollabEval 的设计原则一致：在保持各维度全面评估能力的同时，做到经济高效。

Gap 比率分析。 图 3 中对相关性评估的 Gap 比率分析揭示了不同模型在评估精度方面的重要规律。CollabEval 表现最佳，Gap 1 比率最高，其次是 Sonnet、Haiku 和 Llama3。这种分布模式反映了评估行为的若干关键发现。首先，CollabEval、Sonnet 和 Haiku 的 Gap1 比率非常接近，表明先进模型的精度水平较为一致；而 Llama3 的表现较低，则说明其在相关性评估方面存在差距。

错误严重程度的变化趋势进一步揭示了模型可靠性。CollabEval 从差距 1 到差距 2 再到 Gap 3 呈现出急剧下降的趋势，这表明错误一旦出现，

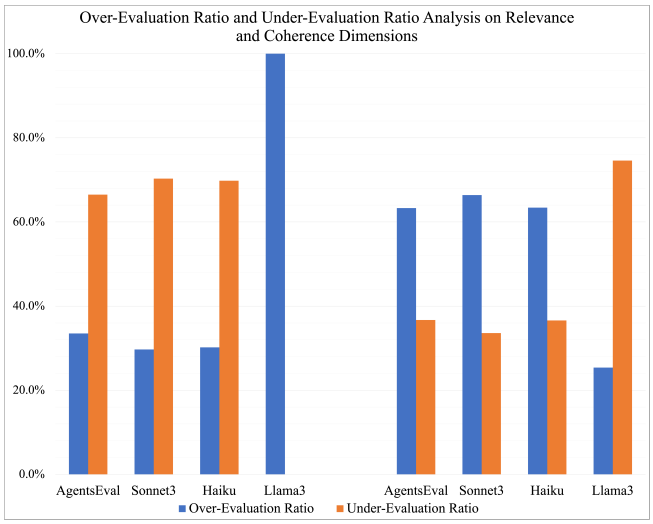


图4：相关性（左）和连贯性（右）评估的趋势分析。

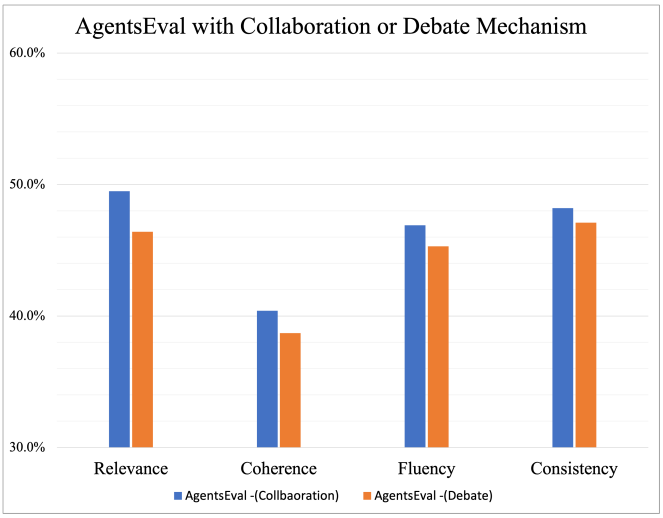


图5：CollabEval 在 SummEval 数据集上的准确率分析（采用协作机制或辩论机制）。

通常都比较轻微。这与 Llama3 较为平缓的分布形成对比，说明其对错误程度的区分能力较弱。在所有模型中，Gap 3错误和 Gap 4错误都极少出现，这表明严重误判并不常见，而 CollabEval 此类错误的发生率最低。

这些发现表明，尽管所有模型通常都能避免严重误判，但 CollabEval 的协作方法能够带来更精细的评估，使误差更多集中在极小范围内，展现了多智能体评估在保持精确性的同时最大限度减少严重评估错误的有效性。

评估模式。图4中的评估模式分析揭示了不同模型在各个维度上表现出的不同特征。在相关性维度上，CollabEval、Sonnet 和 Haiku 均表现出一致的低估倾向，表 1 中的低估比例分别约为 68.1%、70.3% 和 69.8%。这种保守的评估方法表明，这些模型在评估相关性时更加严格。相反，在连贯性维度上，这些模型均明显转向高估，CollabEval 的过度评估比率为 63.3%，Sonnet 为 66.4%，Haiku 为 63.4%，这表明它们对连贯性质量的评估更加宽松。

Llama3 的案例尤为有趣，其评估模式极端，与其他模型显著不同。在相关性评估中，它的高估比例高达 100%，表明其一贯倾向于给出较高分数。与之形成鲜明对比的是，在连贯性评估中，它表现出 74.6% 的低估倾向。这些极端模式揭示了两个重要观点：首先，单个模型可能会在特定维度上形成强烈偏差；其次，采用平衡的多智能体方法对于缓解此类极端倾向至关重要。

不同维度和模型之间截然不同的评估模式凸显了 CollabEval 协作方法的价值。该方法通过多智能体共识帮助平衡这些固有偏见，从而在不同维度上得出更细致、更可靠的评估结果。

稳健性与一致性。CollabEval 在不同评估场景中进一步展现出卓越的稳健性，尤其体现在即使单个模型存在局限，也能保持稳定的性能。在相关性评估中，尽管 Llama3 的性能显著下降（22.8%）（见图 2 和表 1），CollabEval 即使仅经过一轮讨论，仍能保持 49.4% 的高准确率。这种性能稳定性也体现在其他维度：CollabEval 在连贯性评估中的准确率达到 40.2%，而 Llama3 为 29.5%；其他维度也呈现出类似趋势。

组装机制通过多种渠道发挥作用。首先，当某个模型表现出极端的评估模式时（例如 Llama3 在相关性评估中的过度评估倾向达到 100%），CollabEval 的协作框架能够通过整合其他评估者的意见有效平衡这种倾向，从而得出更为适度且准确的评估结果（过度评估比例为 31.9%）。其次，多智能体设置支持对评估结果进行交叉验证，较强的模型能够帮助纠正较弱模型的偏差。这一点在差距比率分析中尤为明显：尽管纳入了个体表现各异的模型所提供的意见，CollabEval 仍保持了最高的差距 1 比率（87.8%）。

这些发现表明，CollabEval 的稳健表现并非仅仅源于平均效应，而是得益于一种编排机制：它在通过协作评估发挥各模型优势的同时，减轻了它们各自的弱点。

协作优势。我们上文探讨了 CollabEval 的机制——协作机制或辩论机制，如图 5 所示。实验结果表明，CollabEval 的协作方法

在所有评估维度上始终优于辩论机制。在相关性评估中，协作机制在这一关键维度上展现出明显优势。在连贯性、流畅性和一致性评估中也呈现出类似趋势。

这种稳定的差距表明，协作中智能体专注于分享见解并在彼此评估的基础上进一步推进，其效果优于竞争性的辩论机制。协作方法之所以表现更出色，是因为它强调建设性的信息共享与共识构建，而非对抗性讨论。这与我们基于标准的评估结果一致：CollabEval 的协作框架在不同维度上均展现出稳健的表现，同时通过设置合理的轮次限制，保持了较高的计算效率。

结论

本文提出了 CollabEval，这是一种用于评估人工智能生成内容的新型多智能体框架。通过大量实验，我们证明，CollabEval 在多个维度上始终优于单一大语言模型方法；经过数轮讨论后，其性能达到最佳，并展现出卓越的能力。即使单个模型表现不佳，该框架仍能保持稳健的性能，这验证了我们所采用的协作评估方法的有效性。未来的工作可以探索将该框架扩展到更复杂的评估场景，并研究不同模型组合对评估结果的影响。

参考文献

AI, M. 2024a. 隆重推出 Meta Llama 3：功能最强大的开放可用大语言模型。AI, M. 2024b. Mistral Large：模型概述。Anthropic. 2024a. Claude 3.5 Sonnet：它是什么、如何运作、使用场景与成果。Anthropic. 2024b. 隆重推出新一代 Claude。Bai, G.; Liu, J.; Bu, X.; He, Y.; Liu, J.; Zhou, Z.; Lin, Z.; Su, W.; Ge, T.; Zheng, B.; 等. 2024. Mt-bench-101：用于评估大语言模型多轮对话能力的细粒度基准。*arXiv* 预印本 *arXiv:2402.14762*. Chan, C.-M.; Chen, W.; Su, Y.; Yu, J.; Xue, W.; Zhang, S.; Fu, J.; 和 Liu, Z. 2024. ChatEval：通过多智能体辩论实现更好的基于大语言模型的评估者。发表于 *ICLR*。Chen, J. C.-Y.; Saha, S.; 和 Bansal, M. 2023. Reconcile：圆桌会议通过多样化大语言模型之间的共识提升推理能力。*arXiv* 预印本 *arXiv:2309.13007*。Chern, S.; Chern, E.; Neubig, G.; 和 Liu, P. 2024. 大语言模型能否胜任评估工作？通过智能体辩论对大语言模型评估者进行可扩展的元评估。*arXiv* 预印本 *arXiv:2401.16788*。Chiang, W.-L.; Zheng, L.; Sheng, Y.; Angelopoulos, A. N.; Li, T.; Li, D.; Zhang, H.; Zhu, B.; Jordan, M.; Gonzalez, J. E.; 和 Stoica, I. 2024. Chatbot Arena：一个根据人类偏好评估大语言模型的开放平台。arXiv:2403.04132。Fabbri, A. R.; Kry´sci´nski, W.; McCann, B.; Xiong, C.; Socher, R.; 和 Radev, D. 2021. SummEval：重新审视摘要评估。计算语言学协会会刊，9：391–409。Han, S.; Zhang, Q.; Yao, Y.; Jin, W.; Xu, Z.; 和 He, C. 2024. 大语言模型多智能体系统：挑战与开放问题。*arXiv* 预印本 *arXiv:2402.03578*。Hong, S.; Zhuge, M.; Chen, J.; Zheng, X.; Cheng, Y.; Wang, J.; Zhang, C.; Wang, Z.; Yau, S. K. S.; Lin, Z.; 等. 2024. MetaGPT：用于多智能体协作的元编程框架。发表于 *ICLR*。Huang, Y.; Bai, Y.; Zhu, Z.; Zhang, J.; Zhang, J.; Su, T.; Liu, J.; Lv, C.; Zhang, Y.; Fu, Y.; 等. 2024. C-eval：面向基础模型的多层级、多学科中文评估套件。发表于*NeurIPS*，第 36 卷。Li, H.; You, S.; Di Palo, F.; Qian, Y.; 和 Jain, A. 2025. 基于分层大语言模型编辑器的工具调用验证引导上下文优化。*arXiv* 预印本 *arXiv:2512.13860*。LMSYSOrg. 2023. Chatbot Arena 对话数据集发布。https://lmsys.org/blog/2023-07-20-dataset/。Ma, T.; Qian, Y.; Wang, Z.; Zhang, Z.; Zhang, C.; 和 Ye, Y. 2025a. 利用大语言模型增强的类别不平衡图提示学习，实现在线毒品贩运检测。发表于 计算语言学协会：*ACL 2025* 论文集，14095–14114。Ma, T.; Qian, Y.; Zhang, Z.; Wang, Z.; Qian, X.; Bai, F.; Ding, Y.; Luo, X.; Zhang, S.; Murugesan, K.; 等. 2025b. AutoData：用于开放网络数据收集的多智能体系统。发表于第三十九届神经信息处理系统年会。

Qian, Y.; Zhang, S.; Chen, L.; Socolinsky, D.; Sokhandan, N.; Cui, S.; Chen, D.; and Sathyanarayana, S. 2025年。通过超图对比学习和可解释的大语言模型驱动分析增强电子商务表示学习。载于*ACM 万维网大会 2025*年附属会议论文集，2512–2520。Raina, V.; Liusie, A.; and Gales, M. 2024年。大语言模型评审是否稳健？探究针对零样本大语言模型评估的通用对抗攻击。*arXiv* 预印本 *arXiv:2402.14016*。Rasheed, Z.; Waseem, M.; Systä, K.; and Abrahamsson, P. 2024年。通过多智能体进行大语言模型评估：初步结果。*arXiv* 预印本 *arXiv:2404.01023*。Shah, M. I. A.; Wahid, A.; Barrett, E.; and Mason, K. 2024年。点对点能源交易中的多智能体系统：全面综述。人工智能工程应用，107847。Song, H.; Razdan, D.; Qian, Y.; Chowdhury, A. G.; Patwa, P.; Chadha, A.; Zhang, S.; Keshava, S.; and Marlowe, H. 2025年。从泛化模式中学习：一种以评估为驱动的方法，用于增强小语言模型微调的数据增强。*arXiv* 预印本 *arXiv:2510.18143*。Verga, P.; Hofstatter, S.; Althammer, S.; Su, Y.; Piktus, A.; Arkhangorodsky, A.; Xu, M.; White, N.; and Lewis, P. 2024年。以陪审团取代裁判：使用多样化模型小组评估大语言模型生成内容。*arXiv* 预印本 *arXiv:2404.18796*。Wang, Q.; Wang, Z.; Su, Y.; Tong, H.; and Song, Y. 2024a。重新思考大语言模型推理的界限：多智能体讨论是关键吗？*arXiv* 预印本 *arXiv:2402.18272*。Wang, T.; Kulikov, I.; Golovneva, O.; Yu, P.; Yuan, W.; Dwivedi-Yu, J.; Pang, R. Y.; Fazel-Zarandi, M.; Weston, J.; and Li, X. 2024b。自训练评估器。*CoRR*。Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Li, B.; Zhu, E.; Jiang, L.; Zhang, X.; Zhang, S.; Liu, J.; et al. 2024年。AutoGen：通过多智能体对话赋能新一代大语言模型应用。载于*ICLR 2024*年大语言模型（*LLM*）智能体研讨会。Zhang, Y.; Yang, S.; Bai, C.; Wu, F.; Li, X.; Wang, Z.; and Li, X. 2024年。迈向高效的大语言模型接地，以实现具身多智能体协作。*arXiv* 预印本 *arXiv:2405.14314*。Zhuge, M.; Zhao, C.; Ashley, D.; Wang, W.; Khizbullin, D.; Xiong, Y.; Liu, Z.; Chang, E.; Krishnamoorthi, R.; Tian, Y.; et al. 2024年。智能体即裁判：用智能体评估智能体。*arXiv* 预印本 *arXiv:2410.10934*。