**AI review工作总结**

**一、项目概述**

|  |
| --- |
| 近年来，随着人工智能技术在学术研究、内容生成和科研评估领域的深度融合，一系列创新的AI评审系统应运而生，旨在解决传统评价方法中效率低下、主观性强及维度单一等核心问题。这些系统通过结合大型语言模型、多模态分析、动态维度提取等技术，不仅提升了评审的自动化水平，更推动了评价标准向可解释性、多维度化和专业化方向演进。 |

**1.1 项目一**[**REVIEWER2: Optimizing Review Generation Through Prompt Generation**](https://arxiv.org/pdf/2402.10886.pdf)

**问题背景**

|  |
| --- |
| * 现有AI评审生成方法易产生笼统的评审意见​（如“方法合理但实验不足”），缺乏具体指导性。 * 模型倾向于生成“平均化”评审，无法覆盖人类评审员可能关注的不同方面（如理论缺陷、实验设置等）。 * 论文作者无法指定评审聚焦的具体方面（如“请重点评审数学推导”） |

**方法**

1. 提示生成模型（Mₚ）​​：分析论文全文，生成多个方面提示​，例如：

* *评估方法部分的理论严谨性*
* *检查实验的可复现性*

1. 评审生成模型（Mᵣ）​​：根据论文内容+方面提示生成针对性评审。

* 关键技术​：
* 基于LongLoRA支持32K长上下文，避免传统方法依赖摘要导致的信息损失。
* 提出PGE流程​（Prompt Generation with Evaluation）自动生成高质量提示：

**成果作用**

|  |
| --- |
| 作者端​：提供针对性修改建议  评审端​：生成多样评审视角辅助人类评审员 |

**经验总结**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| |  | | --- | | **优点** | | |  | | --- | | **不足** | | |  | | --- | | **改进方向** | |

|  |  |  |
| --- | --- | --- |
| * *生成评审包含具体指向​，而非泛泛而谈* * PGE流程有效性​，通过LLM生成+自评筛选，93.6%提示在3轮内达到质量标准 * 作者可通过选择不同提示获取多元视角​ | * *PGE提示生成与REVIEWER2评审生成独立进行​，导致提示可能偏离评审生成需求* * *PGE仅基于评审文本生成提示，而REVIEWER2需输入全文，两者信息源不统一。因为*Llama-2的4K上下文限制无法处理长论文 * 模型依赖预训练知识，对高度专业化论文（如理论物理）可能生成不准确评审 * BERTScore无法捕捉事实准确性​，此外人工评估依赖GPT-4作为裁判可能引入模型固有偏见 | * *将提示生成（Mₚ）与评审生成（Mᵣ）统一为多任务学习框架* * 对于文本规模过大，使用主干捕捉全局结构，分支保留细节[ring attention](https://arxiv.org/abs/2405.02822) * 在1M+学术文献上继续预训练或者使用RAG * 为BERTScore设计新的评估指标，比如计算评审与论文事实对齐度，同一论文不同提示的评审差异度 |

**1.2 项目二**[**AI-Driven Review Systems: Evaluating LLMs in Scalable and Bias-Aware Academic Reviews**](https://arxiv.org/pdf/2408.10365)

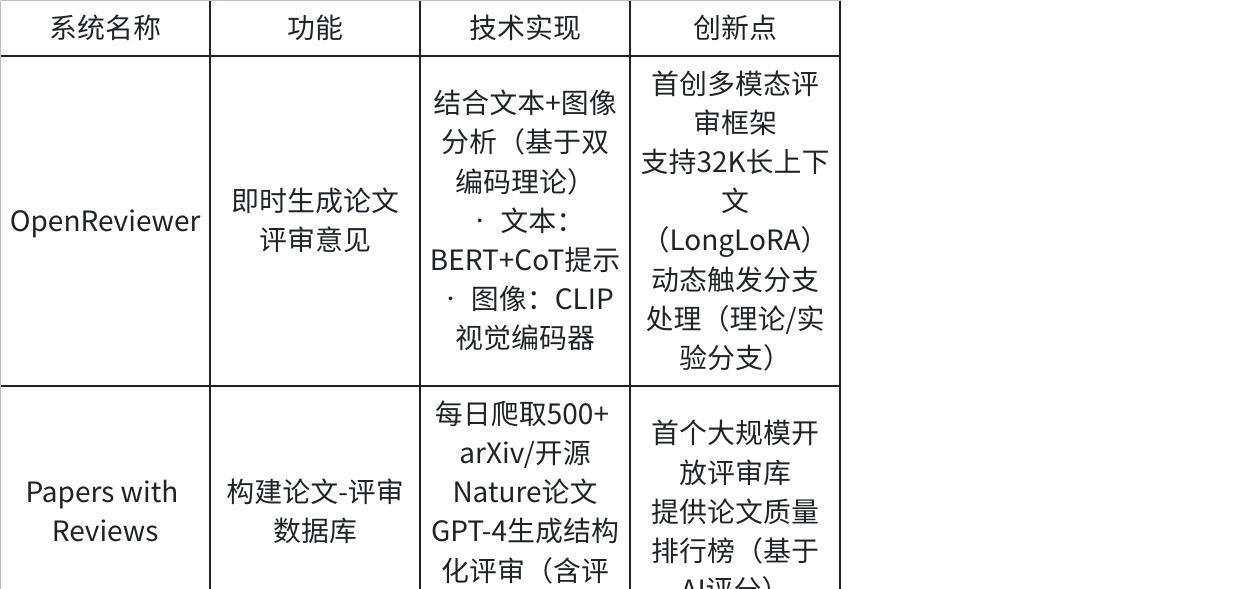
**具体运用方向**

• 帮助作者改进他们的论文，充分引用相关工作、清晰度、合理性等。

• 让审稿人帮助查找和完善分配给审稿人的论文的审稿要点。

• 协助会议项目主席或期刊编辑在人工监督下快速识别低质量的作品，以便拒稿。

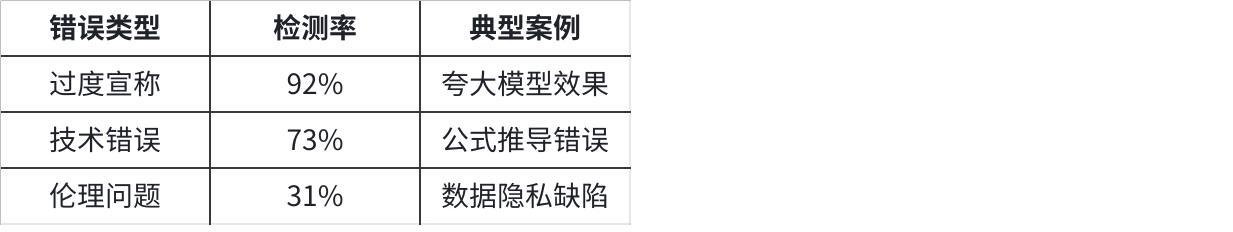
**审查系统**



**点击图片可查看完整电子表格**

**错误分析**

检测能力分级​



**点击图片可查看完整电子表格**

**四维评估体系**

1. 人类偏好评估​：专家投票对比AI/人类评审
2. LLM自动评估​：GPT-4作为裁判生成偏好理由
3. 人类偏好预测​：微调LLM预测人类选择
4. 错误注入分析​：人为植入论文缺陷，检测模型盲点

* *缺陷类型*：理论错误/实验缺陷/过度宣称/伦理问题
* *检测工具*：GPT-4修改原文+对比评审差异

**经验总结**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| |  | | --- | | **优点** | | |  | | --- | | **不足** | | |  | | --- | | **改进方向** | |

|  |  |  |
| --- | --- | --- |
| * *动态评审生成，LLM基于内容生成定制问题* * 多模态理解   图像处理：CLIP提取视觉特征  文本处理：BERT提取关键主张  融合机制：交叉注意力加权   * 量化评估   胜率矩阵​计算人类/GPT-4对不同评审的偏好胜率；Bradley-Terry系数 | * *提示生成（Mₚ）与评审生成（Mᵣ）分离 → 提示可能偏离评审需求；Mₚ仅用摘要 → 丢失全文细节* | * *端到端联合训练，共享编码器减少信息损失* * 层次化长文本处理​（图11）   主干网络​：RingAttention捕捉全局结构  分支网络​：动态摘要器提取关键句理论/实验章节中的句子  门控融合​：加权整合特征   * 伦理风险控制，添加LLM水印[AI-Generated]标识 |

**成果总结与未来展望**

**人类-AI评审差异**

* 评分分布对比​
* GPT-4评分分布更集中（标准差↓0.3）
* 人类对“创新性”评分更严格（均值低0.8分）
* 优化策略  
  引入动态校准机制​：  
  最终分数 = LLM原始分 × 人类评分分布 / LLM评分分布

下一步将探索自演进评审系统​——通过人类反馈强化学习（RLHF）使模型持续迭代

**1.3 项目三**[**MARG: Multi-Agent Review Generation for Scientific Papers**](https://arxiv.org/abs/2401.04259)

**架构设计**

三层智能体分工：

领导者​​：协调任务分配与通信

* 动态规划评审流程（高维任务拆解→步骤执行）
* 消息广播机制**SEND MESSAGE协议**

工作者​​：处理论文分块（每块≤4K tokens）

* 按段落切分，保留位置标记（e.g. "Section 3.1, para 2"）

专家​​：聚焦特定评审维度（实验/清晰度/影响力）

* 定制化提示模板，实验专家需"设计假设实验对比原文"

跨智能体通信协议：

错误纠正机制：

* 循环对话检测（重复消息自动终止）
* 协议遵守强化（遗忘协议时追加系统提示）

上下文修剪策略：

* 工作者：仅保留最近3条消息 + 初始论文块
* 领导者：剪枝历史消息，保留发出指令

三阶段评审生成​：

1. 问题发现​：专家智能体生成针对性问题

* *实验专家* → "方法未说明随机种子设置"

1. 内部辩论​：智能体交叉质疑

* 辩论记录：Agent2质疑Agent1：样本量不足未论证

1. 评论精炼​：领导者汇总争议点，生成最终评审

* 过滤冗余评论（重复率>80%则合并）

**结果描述**

* 长度-效果拐点​：论文>8K tokens时收益显著
* 特异性提升瓶颈​：专家提示模板质量比智能体数量更重要（5专家vs.3专家仅+7%）
* 灾难性错误溯源​：73%严重错误源于领导者规划失误

**经验总结**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| |  | | --- | | **优点** | | |  | | --- | | **不足** | | |  | | --- | | **改进方向** | |

|  |  |  |
| --- | --- | --- |
| * 突破长度限制，全文本覆盖​，将万token级论文分块处理 * *专家智能体定制化提示模板，会提示哪一模块需要修改* * 动态辩论机制​，智能体相互挑战减少盲点 * 检测重复消息自动终止（如无限致谢循环） * 实时校验SEND MESSAGE格式 | * 计算成本过高，单篇评审需1.2M tokens，平均耗时18分钟/篇 * 实验专家无法处理伦理问题，需预设角色类型；工作者无法访问其他分块，致全局推理缺失 * 各个模块存在依赖，领导者规划错误→全体智能体执行偏差 * 精炼阶段遗漏​：13%无效评论未被过滤，如"建议多写结论" | * ①技术性回复→向量编码（Sentence-BERT提取关键语义）   ②辩论内容→争议点摘要（仅保留异议维度+数据支撑）  ③低负载时合并工作者/专家职能  ④分层模型部署​：  工作者→轻量模型（Phi-3）  专家/领导者→GPT-4 Turbo   * 工作者决定核心主张，领导者构建全局知识图谱 |

**未来方向**

* 实时人类介入​：争议分数>阈值时请求人工仲裁
* 领域自适应专家​：预训练科学领域专家（SciBERT微调）
* 伦理约束模块​：添加审查智能体（检测偏见/歧视表述）

**1.4** [**MLR-Bench: Evaluating AI Agents on Open-Ended Machine Learning Research**](https://arxiv.org/pdf/2505.19955)

**项目描述**



**点击图片可查看完整电子表格**

**结果描述**

**经验总结**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| |  | | --- | | **优点** | | |  | | --- | | **不足** | | |  | | --- | | **改进方向** | |

|  |  |  |
| --- | --- | --- |
| * *多维度量化指标dimensions= ["Hallucination","Consistency", "Soundness", "Significance"]* * *人类对齐验证，双评委机制与10位领域专家评分对比* * 检索最新10篇相关论文（2023-2025年）解决模型知识滞后问题 * 编码代理遇错时，自动降级到简化实验方案 * 分层计算 * 策略构思/提案：轻量模型（Ministral-8B） * 实验/写作：重量模型（Claude-3.7） | * 80%实验输出含未验证数据 * 实验仅覆盖ML领域，未验证跨学科泛化性 * 新颖性评分依赖文本相似度，无法识别“旧方法新组合”式伪创新 * 计算成本瓶颈 | * 实时监控实验异常，连续3次运行失败触发阈值 * *创新性评分指标 = 1 - 最大相似度* |

未来需建立科研伦理防火墙​，对高风险领域设置结果双人复核机制。

**三、未来可以做的工作**

目前ai rebuttal的论文似乎还没有人做，但是这篇论文的想法很好[ReasonFlux-PRM: Trajectory-Aware Long Chain-of-Thought Reasoning in LLMs](https://arxiv.org/abs/2506.18896)，但是似乎只是设置了奖励机制，没有运用到rebuttal上。

* 工作流程​：

1. 生成初始评审意见（类比作者提交论文）
2. 接收反馈信号（类比评审意见）
3. 动态调整响应（类比作者rebuttal）

* 案例佐证​：  
  在实验分析中，模型根据PRM标记的错误步骤重新生成推理路径，实现类似rebuttal的自我修正。

而另外一篇论文[Exploring Jiu-Jitsu Argumentation for Writing Peer Review Rebuttals](https://arxiv.org/pdf/2311.03998.pdf) 是关于rebuttal的数据集，该数据集包含了2019年和2020年ICLR会议的审稿和相应反驳。这些审稿和反驳被细致地分解成单个句子，并被三层注释标记，包括审稿方面和极性、审稿与反驳之间的链接，以及反驳行动的直接注释。

**四、总结**

目前所做的ai review/rebuttal的工作所面临的问题有如下几点：

* 无法处理开放式质询，就类似[*From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge*](https://arxiv.org/abs/2411.16594)*处理动态与复杂任务能力有限*
* 伦理盲区​，对利益冲突等非技术性反馈无响应机制
* 长程依赖断裂​，步数过多时的推理链修正成功率骤降
* 偏差与脆弱性，如位置偏差、长度偏差等，会影响模型评判的公平性
* 当模型既是生成者又是评判者时，可能出现自我偏向的问题。
* 计算成本瓶颈
* 精炼阶段遗漏​
* 各个模块存在依赖，单个模块对最后结果的产生影响很大
* 事实准确性不足，另外可能预训练专业知识不够

**未来研究方向**

* 通过引入人工校正，进一步提高评估的可信度。
* 动态评估框架：为复杂任务设计更灵活的评估机制。
* 零样本评估技术：通过更强大的提示设计，让模型无需额外训练即可准确评估。

**商业化的ai review**

完全为这个量身定做的产品几乎没有，但是有一些可借鉴的产品

1. [53AI知识库](https://www.53ai.com/news/LargeLanguageModel/2025042610369.html)

* 传统软件测试局限性及AI增量代码自动Review的必要性
* 基于AI大模型搭建增量代码自动Review工具的目标与方案
* 模型选择与技术方案详解，以及实际应用效果评估

1. [网易有道QAnything引擎](https://qanything.ai/)

* 功能：开源RAG系统进化成“企业AI大脑”，支持私有化部署与智能体生成，已服务上百家企业。
* 适配场景：可构建领域知识库驱动的论文质量评估系统，自动比对文献创新点是否有重复