**Ai rebuttal数据集调研**

**1.1** [**Exploring Jiu-Jitsu Argumentation for Writing Peer Review Rebuttals**](https://arxiv.org/abs/2311.03998) **Accepted at EMNLP Main Conference 2023**

**基础数据来源**

* DISAPERE数据集（ICLR 2019-2020）：
* 包含9,946条审稿句子（review sentences）和11,103条反驳句子（rebuttal sentences）。
* 复用三层标注：
* 审稿方面（Review Aspect） → 作为态度根源（Attitude Root）（如Clarity、Substance等）。
* 审稿-反驳链接（Review-Rebuttal Links）→ 用于关联审稿与对应反驳。
* 反驳行为（Rebuttal Actions）（如Task Done、Reject Criticism等）。
* Peer-Review-Analyze数据集（ICLR 2018）：
* 提供论文章节标签（Paper Sections） → 作为态度主题（Attitude Theme）（如Methodology、Experiments等）。

**数据集统计与结构**



**点击图片可查看完整电子表格**

**总结**

|  |  |  |  |
| --- | --- | --- | --- |
| |  | | --- | | **优点** | | |  | | --- | | **不足** | |

|  |  |
| --- | --- |
| 1. 概念转化： 将心理学中的"态度根源"映射到学术审稿场景，用审稿方面和论文章节分别代理态度根源和主题。 2. 领域自适应： 通过MLM任务在审稿数据上继续预训练语言模型（如SciBERT\_ds\_neg），提升主题分类效果。 3. 混合筛选策略： 结合自动分类器粗筛与人工偏好标注细筛，最终用图排序算法提取典型反驳。 | * *数据仅来自ICLR，未覆盖自然科学或人文学科* * *稀疏映射：并非所有审稿句都有对应典型反驳，原始数据有缺失* * 论文"Limitations"部分提到，典型反驳作为模板需人工细化，需要避免直接替代人类写作 |

**1.2 :**[**A Consistency-ensured Dataset for Full-stage Peer Review and Multi-turn Rebuttal Discussions**](https://arxiv.org/abs/2505.07920)

原文摘要：同行评审是AI等领域科学进步的关键环节，但提交量的激增使得评审系统不堪重负，这不可避免地导致了评审人员短缺和评审质量下降。除了研究热度持续攀升外，造成这种超负荷的另一个关键因素是不合格稿件的重复提交，这很大程度上源于作者在投稿前缺乏有效的工具进行自我评估。大语言模型 (LLMs) 在协助作者和评审人员方面展现出巨大潜力，但其性能从根本上受限于同行评审数据的质量。然而，现有的同行评审数据集面临三大局限：

1.数据多样性有限；2.由于使用的是修订后版本而非初始投稿，导致数据不一致且质量较低；3.对涉及反驳及评审人-作者互动的任务支持不足。

为应对这些挑战，我们推出了最大规模且确保一致性的同行评审与反驳数据集——Re²。该数据集包含来自OpenReview平台上24个会议和21个研讨会的 19,926份初始投稿、70,668条评审意见以及53,818份反驳。此外，我们将反驳和讨论阶段构建为多轮对话范式，旨在支持传统的静态评审任务以及动态交互式LLM助手，从而为作者优化稿件提供更具实践性的指导，并有助于缓解日益增长的评审负担。

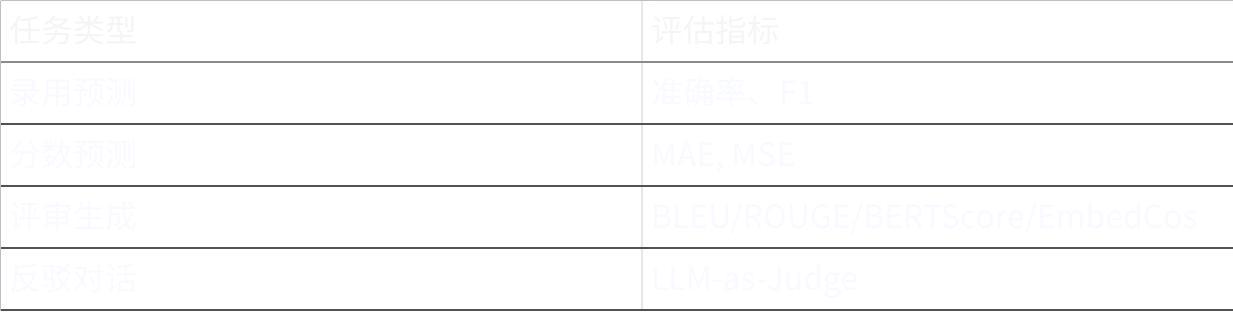
[A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications](https://aclanthology.org/N18-1149/)这是peer read v1

他们根据这篇文章提出了peer read v2

* 从 PeerRead v1 的约 14,000 篇增长到 ~100,000 篇。
* 从 PeerRead v1 的约 4,000 条增长到 ~170,000 条。
* 覆盖范围45个学术活动，远超 v1 仅覆盖 3 个会议的范围。

**数据来源**

* 从OpenReview爬取 24个会议与21个研讨会，远超现有数据集比如如PeerRead仅14,700篇论文，无反驳数据
* 初始投稿获取方式
* （1）通过会议截稿日期确定初始提交时间
* （2）从论文修订历史中提取截稿前最新版本
* 使用商用工具Doc2X保障公式识别精度，PDF转纯文本
* 关键技术挑战：
* 作者/审稿人的多次连续回复合并为单轮对话
* 针对多位审稿人的相似问题，将全局回复插入对话流
* 混合人工与自动化方法剔除催促类回复，比如请审稿人回复
* 结构化多轮对话，支持动态交互式LLM训练。



**点击图片可查看完整电子表格**

一些insight：

* 反驳对话任务中，模型在准确性和建设性表现最佳
* 零样本LLM倾向讨好作者
* 专用模型过于严格，评分均过低

**局限**

仅处理文本内容，未整合图表等多模态信息

**rebuttal\_test.json中以8xC5NNej-l\_ 为例**

**系统初始化 → 2. 论文提交 → 3. 初始评审 → 4. 作者回应 → 5. 争议解决 → 6. 最终裁决**

**1. 顶层结构**

JSON 文件是一个 **对象数组**​，每个对象代表 **一篇论文的完整评审记录**​，包含以下核心字段：

|  |
| --- |
| Plain Text [  {  "paper\_id": "8xC5NNej-l\_", // 论文唯一ID  "conference\_year\_track": "CoRL 2021 Conference", // 会议名称+年份  "reviewer\_id": "Reviewer\_bi2U", // 评审员ID  "messages": [ ... ] // 评审对话记录  },  { ... } // 其他论文记录 ] |

**2. messages 对话结构**

messages 是一个 **有序对话列表**​，按时间顺序记录评审员与作者的交互。每条记录包含：

|  |
| --- |
| Plain Text {  "role": "system" | "user" | "assistant", // 角色类型  "content": "..." // 对话内容 } |

**角色说明**



**点击图片可查看完整电子表格**

**3. 典型对话流程**

**阶段 1：初始化**

|  |
| --- |
| Plain Text {  "role": "system",  "content": "You are an experienced academic paper reviewer..." // 系统设定评审员角色 } |

**阶段 2：论文提交**

|  |
| --- |
| Plain Text {  "role": "user",  "content": "Now your reviewer ID is assigned..." // 用户提交论文 } |

**阶段 3：初始评审**

|  |
| --- |
| Plain Text {  "role": "assistant",  "content": "Review Content: summary: ... strengths\_and\_weaknesses: ..." // 评审员输出完整评审报告 } |

**阶段 4：作者回应**

|  |
| --- |
| Plain Text {  "role": "user",  "content": "Title: Response to Reviewer..." // 作者反驳/澄清 } |

**阶段 5：最终决议**

|  |
| --- |
| Plain Text {  "role": "assistant",  "content": "I am glad to have had the above discussion..." // 评审员总结并确认评分 } |

**4. 评审内容的关键字段**

评审员回复 (assistant) 包含结构化字段（不同评审员格式略有差异）：

|  |
| --- |
| Plain Text {  "role": "assistant",  "content": "Review Content:   summary: ... // 论文摘要总结  strengths\_and\_weaknesses: ... // 优缺点分析  summary\_of\_recommendation: ... // 最终推荐意见  issues: ... // 具体问题清单  Rating: ... // 评分（如 7/10）  Confidence: ... // 评审员信心指数 }" |