**ai rebuttal数据集**

**开放学术评审平台**

1. OpenReview

* 内容：公开ICLR等顶会的论文评审全过程，包括审稿意见、作者rebuttal及分数变动记录。
* 访问：[OpenReview官网](https://openreview.net/)

1. PubPeer

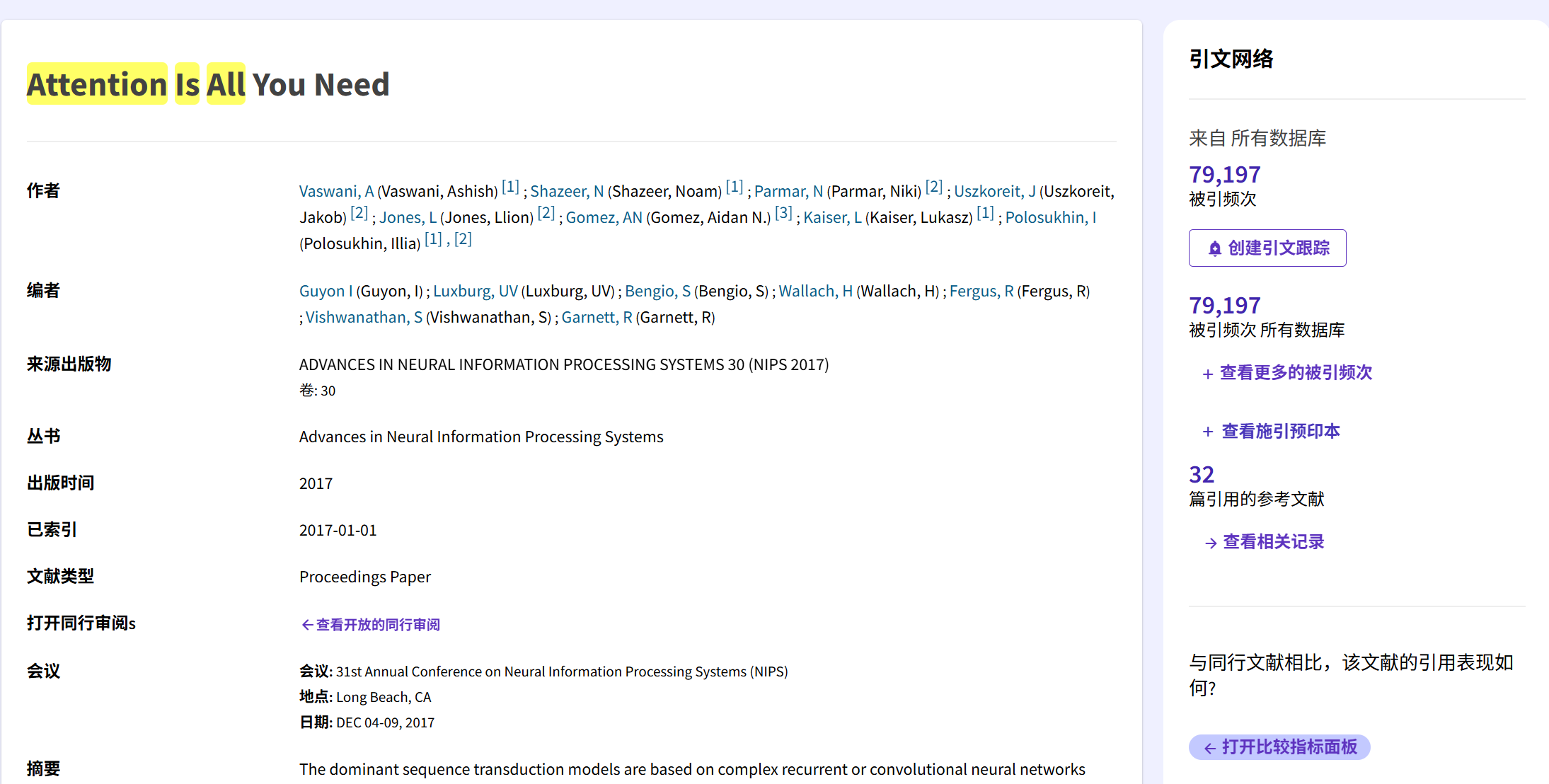
* 内容：学者对已发表论文的公开评论平台，涵盖数据质疑、方法争议等。
* 作用：分析争议性rebuttal的公共舆论场。
* 访问：[PubPeer](https://pubpeer.com/)

1. Web of Science

* 内容：传统权威数据库，包含大量论文审稿记录（需订阅），但评价体系（影响因子）被批易操纵。
* 访问：[Web of Science官网](https://www.webofscience.com/)，点击查看开放的同行评审，似乎没有rebuttal只有review？而且不是每个文献都**支持开放同行评审，这似乎跟作者选择开不开放有关**

**路径一：从具体文献入手**

1. 在Web of Science中检索并点击**感兴趣的文献**；
2. 若该文献**支持开放同行评审**，将在详情页看到**“View open peer reviews” 链接**, **点击进入**；
3. 在新页面中点击 “Open and View All” 或选择具体环节，即可查看该论文的审稿意见、编辑决策、作者回复等评审信息。





这与下面的publishing的案例给的是同一篇论文

**路径二：从主题筛选入手**

1. 在Web of Science中检索**感兴趣的研究主题**（如“hypertension”）；
2. 在页面左侧的 **“Quick Filters”**（快速过滤） 栏中，勾选 **“Open publisher-invited reviews”（公开出版商-受邀审阅）**，点击**“Refine”（精炼）**；
3. 系统将展示符合条件的文献，点击任意一篇，即可查看其开放评审记录（操作方式同路径一）。

但是似乎wos关掉了这个功能？

可能解决方式如下：

如您的界面截图所示，左侧的highly cited paper和hot paper的图标，您可以登陆Essential Science Indicators的数据库后，再打开Web of Science平台的检索界面进行检索。如无法登陆或仍不能出现相关图标，您可以与咱们学校图书馆老师沟通，确认是否当前已经有[Essential Science Indicators](https://incites.clarivate.com/)数据库的访问权限。

似乎需要购买wos的查找资源？

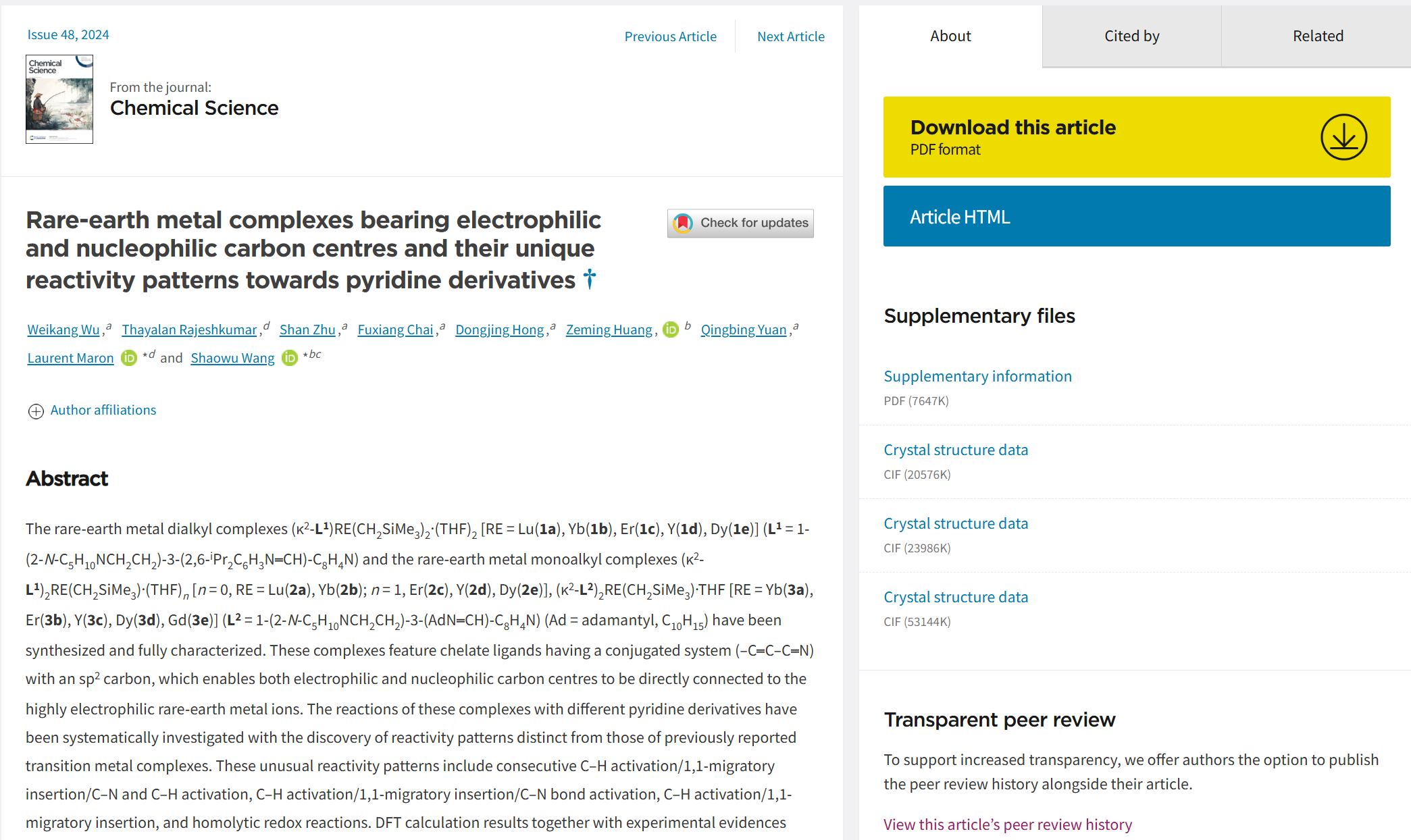
**[图片下载失败]**

**好处**

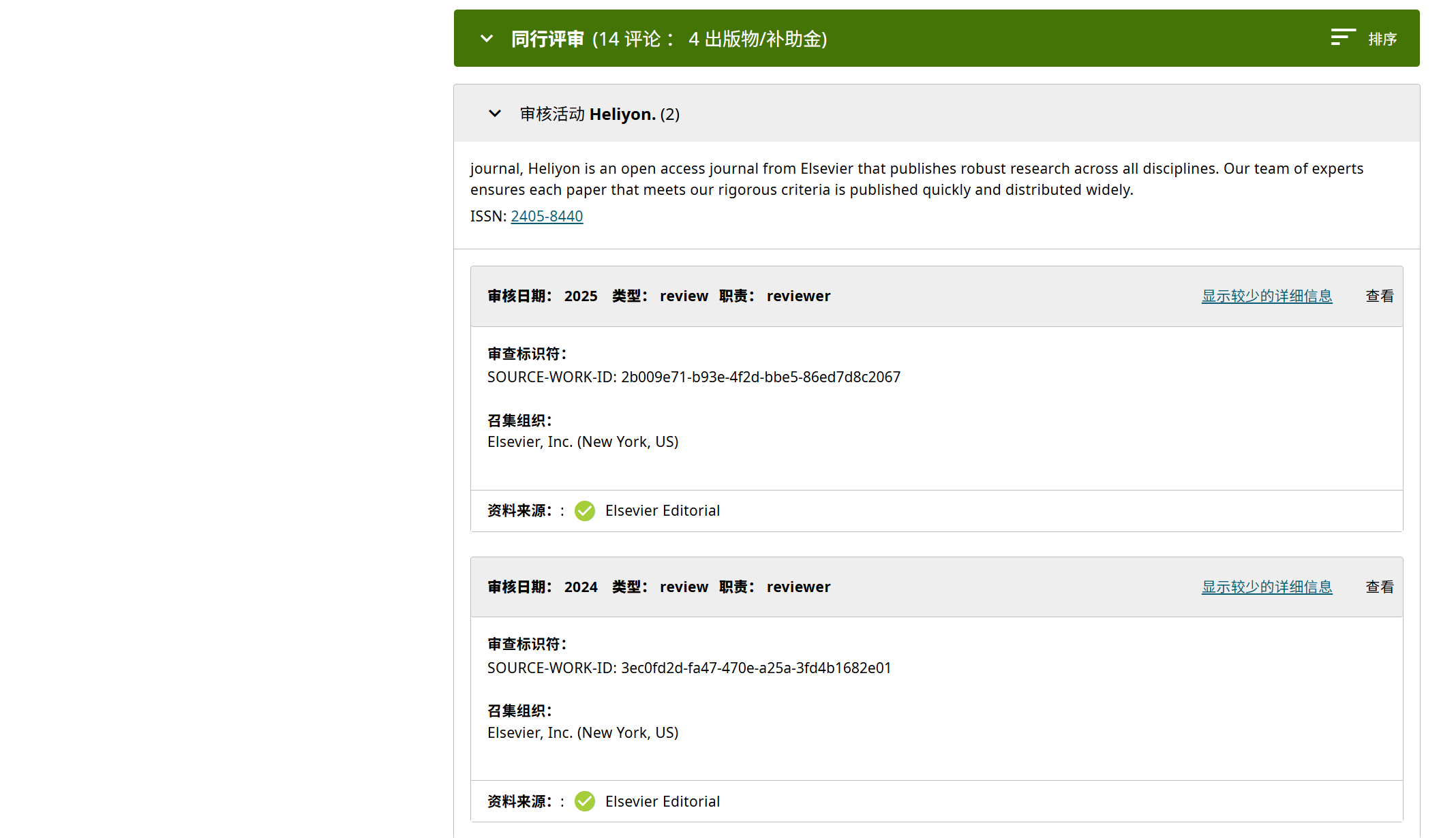
[可以直接在网站内导出文献全记录信息](https://webofscience.help.clarivate.com/zh-cn/Content/full-record.htm)

1. [publishing](https://pubs.rsc.org/)

* 案例来源https://pubs.rsc.org/en/content/articlelanding/2024/sc/d4sc04197f
* 在期刊论文主页右侧Transparent peer review（透明的同行评审）下有一个连接“View this article’s peer review history”（查看此文章同行评审历史）如下图，点击后直接跳转到上面的网站Web of Science了，这3和4是一个平台啊（）



1. orcid



举例：https://orcid.org/0000-0002-6844-1289

其实似乎它跟wos是一起的？

**专业Rebuttal研究数据集**

1. JITSUPEER数据集

* 内容：结合心理学理论分析审稿人的态度根源（如潜在信仰）与论文主题，提供“态度-主题-反驳行动”关联的标注数据。
* 基础数据源：基于ICLR 2019-2020的评审数据（DISAPERE数据集）和2018年ICLR数据（PEER-REVIEW-ANALYZE数据集），包含审稿句子分解、极性标注、反驳行动分类等3。
* 应用：支持生成针对性反驳模板，提升辩论有效性。
* 访问：[JITSUPEER论文与数据](https://arxiv.org/pdf/2311.03998.pdf)

1. DISAPERE数据集

* 内容：收录ICLR会议的审稿与作者反驳内容，细分为单句级注释，涵盖审稿方面（如创新性、实验严谨性）、反驳行动（如认同、部分修正）及二者关联。
* 访问：[DISAPERE项目页面](https://github.com/tanfiona/DISAPERE)

1. PEER-REVIEW-ANALYZE数据集

* 内容：标注审稿意见针对的论文具体部分（如方法、结论），辅助识别态度主题。
* 访问：[PEER-REVIEW-ANALYZE数据](https://github.com/allenai/Peer-Review-Analyze)

**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 |

下面是一些有启发性的工作，似乎都是在反驳帖子

**1.2** [**CounterGeDi: A controllable approach to generate polite, detoxified and emotional counterspeech**](https://arxiv.org/abs/2205.04304)

Accepted at IJCAI-ECAI 2022

**仇恨言论-反言论原始数据集来源**

**1. CONAN数据集**

* 来源：人工合成的仇恨言论 + 专家撰写的反言论
* 规模：408条仇恨实例 → 3,864条仇恨-反言论对
* 特点：针对伊斯兰恐惧症（Islamophobic speech），反言论由NGO专家编写

**2. Reddit数据集**

* 来源：Reddit平台真实仇恨言论 + 众包（Amazon Mechanical Turk工人）撰写的反言论
* 规模：5,257条仇恨实例 → 14,223条仇恨-反言论对

**3. Gab数据集**

* 来源：Gab平台真实仇恨言论 + 众包（AMT工人）撰写的反言论
* 规模：14,614条仇恨实例 → 41,580条仇恨-反言论对

|  |
| --- |
| 关键点：   * Reddit/Gab数据来自公开学术数据集，非作者主动爬取。 * 数据配对方式：每条仇恨言论关联至少一条反言论，构成对话结构（hate speech + counterspeech pair）。 |

反言论策略多样性：  
每条反言论标注了反驳策略类型，例如：

* 事实澄清（Fact-based Refutation）
* 情感呼吁（Emotional Appeal）
* 道德质疑（Moral Questioning）
* 幽默化解（Humor Defusion）

**属性数据集构建与分类**

为控制反言论的礼貌性、去毒性和情感属性，作者使用三类公开标注数据集训练GeDi分类器：

**1. 礼貌性**

* 数据来源：139万条社交媒体帖子
* 分类标注：
* 9级标签（P1-P9），P9为"礼貌"，其余为"非礼貌"
* 二值化处理：P9 vs. P1-P8

**2. 去毒性**

* 数据来源：Kaggle毒性评论分类挑战数据集
* 分类标注：
* 二分类：toxic（有毒） vs. non-toxic（无毒）
* 训练集/验证集/测试集按90%-10%分层划分

**3. 情感**

* 数据来源：416,809条多情感标注帖子
* 分类标注：
* 7种情感：悲伤（sadness）、喜悦（joy）、恐惧（fear）、愤怒（anger）、惊讶（surprise）、爱（love）
* 筛选：剔除低频标签（surprise, love <10%），保留4类（sadness, joy, fear, anger）
* 训练方式：为每类情感训练独立GeDi模型，目标情感为正向标签，其余情感为负向标签

**1.3** [**Auditing Counterfire: Evaluating Advanced Counterargument Generation**](https://arxiv.org/abs/2402.08498)

Accepted at AAAI ICWSM 2025

* Candela数据集（Hua et al., 2019）：
* 包含 70,000条 英文帖子和对应的反驳论点（Counter-arguments），来源于Reddit的r/ChangeMyView社区。
* 每条数据包含：
* 原始帖子（Original Post）
* 人工撰写的反驳论点
* 检索到的证据：从包含1,200万篇文章的数据库中提取（来源包括Wikipedia、《华盛顿邮报》《纽约时报》、Reuters等），经过去重和排序。

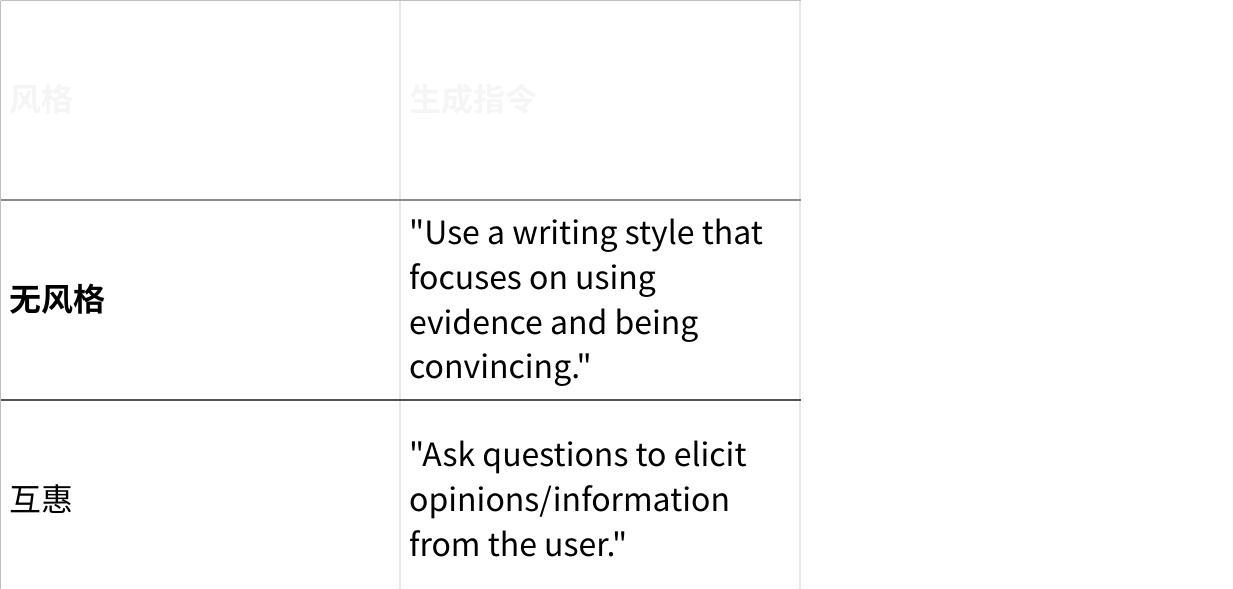
**构建Counterfire数据集**

**(1) 输入样本选择**

* 从Candela数据集中随机抽取 2,000条 原始帖子及对应证据作为输入样本。

**(2) 多模型、多风格生成**

* 模型选择：使用6种LLM生成反驳论点：
* GPT-3.5 Turbo、Koala-13B、GPT-4o、LLaMA-3.1、PaLM-2、Claude Haiku。
* 风格控制：为每条输入生成三种风格的反驳论点：



**点击图片可查看完整电子表格**

* 生成规模：  
  2,000输入 × 3风格 × 6模型 = 36,000条生成论点  
  最终整合为 38,000条 的Counterfire数据集（含部分人工标注数据）。

**数据集特点**

* 三元组结构：每个实例包含 [原始帖子, 证据, 生成的反驳论点]

**三、其他**

[从 OpenReview 获取顶会接收论文集并保存至本地数据库](https://zhuanlan.zhihu.com/p/628467095)

[能不能用Chatgpt回复审稿人意见](https://www.zhihu.com/question/582718387?write) （有趣的回答）

[这个项目是一个用于从同行评审和反驳中提取论点对的自然语言处理工具](https://github.com/LiyingCheng95/ArgumentPairExtraction)