arXiv:2611.30410v1 [cs.CL] 23 Nov 2026

# 人工智能审稿人的披露偏差

**Jens Leysen**[1], **Noah Daniëls**[1],**Marco Favier**[1],**Bart Goethals**[1,2]

[1]安特卫普大学，安特卫普，比利时 [2]Froo
mle，安特卫普，比利时

**摘要**

人工智能模型日益协助评审论文，如今许多会议和期刊要求作者说明其如何使用人工智能。这份人工智能使用声明会改变人工智能审稿人对论文的评分吗？它是否会改变哪些论文被接收？在一项初步研究中，我们仅改变了150篇ICLR 2027投稿的人工智能使用声明，最多采用九个版本，这些版本改编自真实声明：从无人工智能，到写作帮助和编程，再到人工智能研究代理和人工智能作者。六个人工智能审稿人对这些版本进行了评分：决策模型Jev，以及语言模型Claude Haiku 5.5、Claude Sonnet 5.5、Gemini 3.8 Flash、Gemini 3.1 Pro和Gemini 4 Argon。在23,652次评审中，我们发现，与声明未使用人工智能相比，当声明表示人工智能完成了全部工作时，每位审稿人对论文的平均评分都会降低。人工智能研究代理声明会显著降低Jev、Sonnet和Argon的评分。对于人工智能辅助，我们没有发现披露偏差的证据：在写作帮助、编程和研究帮助方面，没有任何审稿人的评分与其在声明未使用人工智能时的评分存在显著差异。观察到的惩罚是否会改变哪些论文被接收，不仅取决于惩罚的幅度，还取决于审稿人对录用边界附近的论文进行区分时的精细程度和一致性，取决于惩罚是分散到所有论文上还是集中在少数论文上，也可能取决于有多少其他论文披露了人工智能使用情况。

## 1 引言

如今，机器学习会议要求作者说明其如何使用人工智能。从2027年起，每篇 ICLR 投稿都必须包含一个说明人工智能使用情况的部分，列出诸如实现方法、解读结果或协助证明等任务（ICLR 2027 项目主席，2026）。与此同时，人工智能模型也开始进入评审流程：人工智能会议中的相当一部分评审文本可能已经经过大语言模型的大幅修改（Liang 等人，2024a），大语言模型辅助的评审会给出不同的评分（Russo 等人，2025），ICLR 本身也曾向其审稿人发送大语言模型反馈（Thakkar 等人，2025），而 AAAI-26 则向负责作出决定的委员会成员展示了每篇完成全部评审的论文的一份人工智能生成的评审意见（Biswas 等人，2026）。因此，论文的人工智能使用声明将越来越多地被人工智能读取。

披露政策建立在这样一个假设之上：诚实的声明不会损害一篇本来没有问题的论文。有证据表明，人们对披露人工智能使用情况的人信任较少（Schilke 和 Reimann，2025）。大语言模型评审者则会反过来对有关其他类型文本作者身份的标签作出反应（Cheong 等人，2025年；Saraf 等人，2025年；Chae 等人，2026年）。这一假设是否适用于人工智能审稿人尚不得而知：据我们所知，还没有研究考察人工智能审稿人如何解读科学论文中的人工智能使用声明，其解读是否取决于人工智能参与的程度，或这种声明是否会改变哪些论文获得接收。然而，这一问题的答案决定了披露政策究竟是在奖励诚实，还是在助长少报。

我们探讨人工智能使用声明是否会改变人工智能审稿人对论文的评分，并讨论这种变化何时会改变哪些论文被接收。我们将仅由该声明引起的评分变化称为 披露偏差。对于第一个问题，我们采用了对照实验。我们选取了150篇真实的 ICLR 2027 投稿，仅替换其中人工智能使用部分的正文。在这项初步研究中，我们针对每个人工智能使用程度采用一份声明；这些声明均接近样本中的真实声明，范围从写作帮助到人工智能研究代理，并以人工智能作者作为极端情况。随后，六种不同类型的人工智能审稿人分别为这些论文版本填写 ICLR 评审表。其中一种是 Jev，这是一种以概率作答的决策模型（TypeSafe AI 2026）。另外五种是语言模型：Claude Haiku 5.5、Claude Sonnet 5.5，以及 Gemini 3.8 Flash、Gemini 3.1 Pro 和 Gemini 4 Argon。我们要求每个语言模型只给出评分，或先撰写书面审稿意见。由于每组对比都比较完全相同的论文，除重新审稿所产生的噪声之外，任何差异都源于人工智能使用声明。我们的发现如下：

- **没有证据表明人工智能辅助存在披露偏差**。 对于写作帮助、AI 编写的代码和人工智能研究帮助，没有任何审稿人的评分与其在声明未使用人工智能时的评分存在显著差异。

- **有一些证据表明，披露人工智能代理和作者使用人工智能会带来不利偏差**。 相比之下，每位审稿人在声明称人工智能完成了所有工作时，给论文的平均评分都会更低；研究智能体声明会显著拉低 Jev、Sonnet 和 Argon 的评分。

- **仅凭评分惩罚无法说明有多少决定会发生变化**。 在一项模拟中，由单个人工智能审稿人作出决定，且只有一篇论文披露了代理式使用；统计显著的惩罚会使被拒收的接收论文比例比重新审稿时高出 8 至 18 个百分点

  单独来看，在 Jev 假设的期望评分中（其粒度精细得多），同一声明会拒绝 77 个百分点

更多。

## 2 相关工作

**同行评审中的大语言模型**。大语言模型对论文的反馈与人工评审的重合程度，和人工评审彼此之间的重合程度大致相当（Liang 等人 2024b年）。据估计，在包括 ICLR 在内的四个人工智能会议上，6.5–16.9% 的评审文本可能经过大语言模型的大幅修改（Liang 等人 2024a年）；在 ICLR 2024，人工智能辅助评审给出的评分更高，而收到此类评审的边缘论文更有可能被接收（Russo 等人 2025年）。如今，会议也开始将人工智能融入评审流程：在 ICLR 2025，一项随机研究中，大语言模型为审稿人的 20,000 多份评审提供了反馈（Thakkar 等人 2025年）；ICML 2026 通过一项随机实验比较了关于使用大语言模型的评审人政策（Kim 等人 2026）；而 AAAI-26 则为其 22,977 篇完成全面评审的论文附上了人工智能生成的评审意见，这些意见对负责决定是否接收这些论文的高级程序委员会成员和领域主席可见（Biswas 等人 2026年）。

**影响大语言模型审稿人的线索**。大语言模型审稿人会对科学内容以外的因素作出反应。稿件中隐藏的指令可以提高大语言模型审稿人的评分（Ye 等，2024），而且一些作者曾在预印本中隐藏此类指令（Lin，2026）。与人类评审相比，大语言模型评审也更常提及作者指出的局限性（Ye 等，2024）。当作者来自声望较高的机构时，即使稿件完全相同，评分也会更高（Howell 等，2025；Vasu 等，2026），不过并非每项审计研究都得出这一结果（Alharbi，2026）。大语言模型审稿人给大语言模型撰写的论文打分高于人类撰写的论文（Li 等，2025）；对于相同的项目，大语言模型也更偏好大语言模型撰写的描述（Laurito 等，2025）。在实际评审中，考虑论文质量后，大语言模型辅助评审对大语言模型辅助撰写论文表现出的宽容倾向便不复存在（Sharma、Joachims 和 Dean，2026）。大语言模型评审者也偏爱自身的输出（Panickssery、Bowman 和 Feng，2024），这种偏好可能源于文本带来的熟悉感（Wataoka、Takahashi 和 Ri，2024）。

**文本不变的标签**。仅仅一个标签就能影响大语言模型评审者。与标注为由大语言模型撰写的回答相比，标注为由专家或人类撰写的回答更受青睐，而评审者在理由中并未提及该标签（Marioriyad, Rohban, and Soleymani Baghshah 2025）。一项研究发现，无论文本实际由哪个模型撰写，只要注明撰写文本的模型，评分就会随之变化：注明 Claude 时评分上升，注明 Gemini 时评分下降（Saraf 等人，2025年）。另一项研究发现，仅仅告知评审者某项研究提案是由其自身还是由另一个模型撰写，就会改变其评分（Chae 等人，2026年）。与我们的问题最为接近的是，一项研究发现，注明一篇新闻文章是在人工智能工具的帮助下创作的，会使人类读者和两个大语言模型给出的评分都略微降低（Cheong 等人，2025年）。

**人工智能使用披露**。大多数顶级期刊都对生成式人工智能提供了指导，通常还包含具体的披露要求（Ganjavi 等，2024）；ICLR 要求作者披露任何大语言模型的使用情况（ICLR 2026 项目主席，2025），并从 2027年起在强制性栏目中进行披露（ICLR 2027 项目主席，2026）。在 ICLR 2026 中，64% 的投稿包含了此类声明，其中几乎所有声明都提到了编辑或润色（Mohapatra 和 Pruthi，2026）。在其他地方，披露远远落后于实际使用情况：2025年初，在期刊论文中，每有一篇论文披露人工智能使用，就约有 40 篇论文呈现出人工智能使用的统计迹象（He 和 Bu，2026）。披露也可能带来成本。披露人工智能使用的人获得的信任较少（Schilke 和 Reimann，2025），而被知晓使用人工智能的人则被认为胜任能力较低（Reif、Larrick 和 Soll，2025）。人工智能所扮演的角色越大，惩罚就越明显：读者对人工智能生成的写作的评价低于对人工智能编辑过的写作的评价（Li 等，2024）；而“人工智能生成”标签主要在被理解为完全自动化时造成负面影响（Altay 和 Gilardi，2024）。然而，一项针对机器学习专家的研究发现，摘要并不存在这种惩罚。他们对一篇论文摘要的三个版本进行评价：由人类撰写、由大语言模型生成，或由大语言模型编辑。被告知各版本来源的读者，对三个版本的评分都高于需要自行猜测的读者，并且最常选择大语言模型编辑的版本（Akpinar 等，2026）。对于审稿人而言，证据则较为有限。在一项关于科学摘要的研究中，录用决定并不取决于文本是否由人工智能生成（Hazra 等，2026）。在真实的同行评审中，当较弱的论文包含被检测出的大语言模型文本时，人类审稿人会给出更低的评价；而较强的论文则不受影响（Sharma、Joachims 和 Dean，2026）。

据我们所知，尚无研究检验过仅一项人工智能使用声明是否会改变人类审稿人的评分。也没有研究探讨过人工智能审稿人会如何回应论文中的人工智能使用声明。与我们的研究最接近的标签研究，都是改变通用文本上的标签。大多数研究会注明撰写文本的模型或作者类型。另有一项研究添加了一句话，说明一篇新闻文章是在人工智能工具的帮助下撰写的（Cheong 等人，2025年）。针对大语言模型审稿人的审计研究改变的是作者及其所属机构，而非声明。关于人工智能参与程度的研究则关注人类读者，且研究对象主要是新闻和短文。我们在真实投稿中检验会议如今要求提供的人工智能使用声明。我们只改变声明中对人工智能所做工作的描述及其所指明的模型，并让六个人工智能审稿人填写会议审稿表。随后，我们考察仅声明本身是否会改变论文被接收的结果。

## 3 方法

我们的代码、提示词、声明和运行配置（包括随机种子）将在一个开放仓库中发布。[1]

### 论文

我们从 OpenReview 中随机抽取了 ICLR 2027 投稿，所有投稿都包含 AI 使用部分，这是 ICLR 2027 强制要求的。在每个版本中，我们保留该部分的标题，只在原位置并按照正文的样式替换其正文内容。我们的每条声明都必须适应论文自身该部分的篇幅，以确保页面的其余部分保持原位。在抽取的 160 篇论文中，有 10 篇（6%）的该部分过短，无法容纳我们最长的声明，因此我们跳过了这些论文，最终保留 150 篇。我们评审的 150 篇论文涵盖 ICLR 的 19 个主要领域，且这些投稿均可在 OpenReview 上公开获取。我们于 10月23日 至 2026年11月5日 期间开展了评审，当时这些投稿正处于评审阶段

[1]本文目前正作为研讨会论文接受评审。论文被接收后，我们将发布代码仓库。

表 1：我们的声明：针对样本中发现的每种人工智能使用级别各有一条，另有无人工智能 和人工智能作者。智能体 声明有三个变体，因此总计九条声明：智能体（*ChatGPT*）、智能体（*Claude*） 和智能体（*Gemini*）分别以相应模型名称取代“生成式人工智能工具”。

| 级别 | 声明 |
|---|---|
| 0 无人工智能 | 本作品未使用生成式人工智能工具。 |
| 1 写作 | 我们仅使用生成式人工智能工具进行语法检查和写作润色。我们未将其用于<br>研究构思、实验设计、代码开发或数据分析。作者对<br>内容负责。 |
| 2 代码 | 我们使用生成式人工智能工具来实现和重构实验代码，并编辑手稿。<br>研究构想和实验设计均由作者独立完成。所有生成的代码均经作者审核并运行。<br>作者。作者对内容负责。 |
| 3 研究帮助 | 我们使用生成式人工智能工具，就方法论和实验设计提供反馈、实施方法，<br>分析并解读结果，以及编辑手稿。作者审查并核实了所有人工智能辅助工作。<br>作者对内容负责。 |
| 4 智能体 | 这篇论文是在将生成式人工智能工具用作研究代理的长期合作中完成的：它们规划并执行了<br>实验、分析了结果，并起草了正文和图表，作者则审阅了每个步骤。<br>作者对内容负责。 |
| 5 人工智能作者 | 这篇论文完全由生成式人工智能工具撰写：它们提出了研究构想、编写了代码，并执行了<br>实验、分析了结果并撰写了正文。 |

在 ICLR。我们不公布投稿名称，也不发布修改后的版本；我们对所有人工智能审稿人所采用的设置和条款均排除了将这些论文用于模型训练的可能性。

## 声明

我们阅读了抽取的全部160篇论文中的人工智能使用声明。其中一篇仅在方法中提及人工智能；我们根据人工智能最多执行了什么，将其余159篇分为四个级别：写作、代码、研究帮助 和智能体。对于每个级别，我们的声明都紧密遵循该级别中的一条或几条真实声明，并将其缩短，同时保留相同的责任结尾句（表1）。同一级别中的真实声明在措辞和所报告的内容上差异很大。因此，每个级别都由众多可能的措辞之一来代表，审稿人对该措辞的反应可能与对级别本身的反应不相上下。

各级别的区别在于人工智能自行完成了多少研究工作。以下引号中的短语摘自论文自身的声明。写作 （61篇论文）包括文本方面的帮助，例如“轻度语言编辑、提升可读性”，或“语法校正和改写”，也包括翻译、文献检索、排版和图表。代码（37篇）还包括代码方面的帮助，例如“调试代码”。研究帮助 （37篇）还包括作者随后会据此采取行动的意见：想法、分析、结果解读、“对概念框架、实验设计的反馈”，或“协助撰写证明”。智能体（24篇）指人工智能基本独立地完成研究工作，例如运行实验或综合文献：“设置实验、检查结果”或“研究构思与执行”。没有论文报告使用了无人工智能。

我们的写作、代码和研究帮助声明读起来非常像真实声明。我们的智能体 声明在两个方面比大多数真实声明更加明确。首先，大多数真实声明将人工智能描述为帮助作者。它们会说人工智能“提供了协助”；其中一些提到实验的声明则说人工智能“帮助”完成了实验。这种措辞适用于任何比例的工作，就像一位主要独立完成实验的合作者也可以被说成帮助完成了实验一样。我们的声明则直截了当地说，人工智能规划并运行了实验。少数真实声明和我们的声明一样直白。其中一份声明称，一名人工智能助手“在作者的硬件上执行了运行”。另一份声明称，智能体团队的各个步骤与人工评审交替进行，“涵盖实验规划与执行、正文和图表生成”。

其次，它们通常只报告一两项智能体任务，而我们的声明将所有任务都列了出来：规划并运行实验、分析结果以及起草文本。我们声明的其他部分在真实声明中也很常见。大多数真实声明不注明模型，而是使用“生成式人工智能工具”这一说法。大多数声明会说作者评审或核实了人工智能的工作，也大多会说作者对内容承担责任。我们的声明也是如此。由于注明具体模型可能会引发不同理解，因此智能体 声明还分别提供了以 ChatGPT、Claude 或 Gemini 替代通用工具名称的版本。160份声明中没有无人工智能 和人工智能作者的对应项。无人工智能 是基线。人工智能作者 是一种积极对照：这是一个极端案例，但我们预计任何阅读该声明的审稿人都会对此作出反应。

有些真实声明对人工智能具体做了什么语焉不详。有些只提到一项任务，却没有说明其涵盖的内容。“协助撰写这篇论文”可能是指润色文本，但论文也是研究的最终成果，因此也可能是指协助研究本身。“数据生成”没有说明人工智能生成了哪些数据，也没有说明生成数据的目的。还有一些声明提到研究的宽泛环节，例如“研究发展”或“研究讨论”，却没有说明人工智能承担了哪些任务。我们将此类声明归入其可能含义中最宽泛的层级。

## 评审表

所有审稿人都填写相同的评审表，依据 ICLR 2026 的评审表设计。表格仅保留分数和总体评分，措辞更加简洁，并询问是否接受这篇论文（是

或否）。可靠性、表达和贡献各分为四个等级：差、一般、好和优秀（对应 ICLR 的 1 至 4 分）。总体评分分为六个等级：强拒、拒稿、边缘拒稿、弱接受、接受和强接受（对应 ICLR 的 0 至 10 分，间隔为 2 分）。

### 审稿人

- **Jev**（版本 1.13.0，TypeSafe AI）是一种决策模型：它不生成文本，而是针对一个有明确类型的问题，为每个答案返回一个概率（TypeSafe AI 2026年）。它接收论文中截至参考文献部分的文本。我们以两种方式解读 Jev 的答案。其最可能的等级，即该概率分布的众数，是表格六个等级之一，也就是审稿人会给出的评分。其预期等级，即按各等级的概率加权计算出的平均值，可能介于两个等级之间。评审表不允许给出这种评分；我们将其作为一种假设性的、更精细的解读加以报告，结果部分将用它来展示这种解读会带来哪些变化。两者来自同一批评审，因此是对同一结果的两种解读，而不是两个独立结果。

- **Claude Haiku 5.5 和 Gemini 3.8 Flash：**它们是小型通用语言模型，下文分别简称 Haiku 和 Flash。每个模型都会收到一份 PDF 格式的论文。由于模型在对论文进行推理后，判断可能会有所不同，我们在两种设置下使用每个模型，两种设置的推理力度都较低：一种是评分 设置，即要求模型直接填写表格，不撰写评审意见；另一种是评审 设置，即要求模型先撰写评审意见（包括摘要、优点、缺点以及给作者的问题），然后在同一条回复中填写表格。

- **Claude Sonnet 5.5、Gemini 3.1 Pro 和 Gemini 4 Argon：**我们还测试了各家提供商更大的模型——*Sonnet*和*Pro*，以及谷歌最新的*Argon*，以了解这些结果能否适用于不同模型。出于成本考虑，Sonnet 和 Argon 仅针对四种声明（无人工智能、写作、智能体和人工智能作者）评审论文，并分别采用评审和评分设置。

仅要求评分时，语言模型在低推理强度下几乎不进行推理：除 Pro（166 个推理令牌）外，所有模型的中位数都是 0。要求先撰写书面审稿时，Haiku 的推理量较多（中位数约为 2,100 个令牌），Pro 则约为 400 个；其他模型的中位数仍为 0。我们保留了各模型的默认温度。每位审稿人都对分配给自己的每个版本进行了两次相互独立的评审。[2]

### 评分比较

对于每位审稿人和每项声明，我们将一篇论文的两次评审取平均值，并将结果与同一论文在无人工智能 声明下的结果进行比较。我们报告这150篇论文的平均差异。我们对各项差异进行双侧配对*t*-检验，该检验还会给出其95%置信区间。每一项比较（例如智能体 与无人工智能的比较）都会进行检验涵盖每位审稿人和每种设置；我们将这些检验视为同一个检验族，并使用Holm方法（Holm 1979）进行校正，从而将该检验族中出现任何假阳性的概率控制在5%以内。只有通过该校正的结果，我们才称其具有显著性。

### 模拟决策

如果评分惩罚会改变哪些论文被接收，那么它对作者就很重要。为了了解在什么情况下会出现这种变化，我们模拟了一个简单的会议。由一位人工智能审稿人对所有投稿进行评分，会议接收评分最高的 27.4%，这也是 ICLR 2026 的录用率（ICLR 2026 程序委员会主席 2026）。因此，一篇论文能否被接收，取决于它在所有投稿中的排名，而不只取决于它的评分。现实中的会议采用不同的决策方式，每篇论文会由多位审稿人和一位领域主席评审。这项模拟旨在单独考察一位人工智能审稿人会如何做出决策。

每个论文版本都经过两次评审。首先，所有论文都被标记为无人工智能声明；随后，其中一篇论文的两份无人工智能评审之一会决定哪些论文被接收。这份评审为论文提供了基准排名。接着，我们让一篇已接收的论文逐篇披露智能体的使用情况：它获得智能体声明，而其他所有论文仍保留其基准评审。现在，这篇论文将依据其智能体版本的两份评审之一接受评判。如果该评审给出的排名低于排名最高的未接收论文，它就会被拒绝。

新的评审也可能在声明完全没有变化的情况下拒绝一篇论文，因为没有审稿人能做到完全一致。因此，我们还将每篇接收论文的另一份无人工智能评审作为对照来判断：它显示仅重新评审就会拒绝多少篇接收论文。声明造成的影响，是超出这一对照的拒绝比例。评分并列时按随机方式打破；智能体 评审和对照均采用相同规则，并对多次随机抽样的结果取平均值。

我们通过互换每个版本中两份评审的角色，分别运行两次模拟，并汇总两次运行的结果。我们报告的影响附有自助法95%置信区间，以及双侧置换检验结果；检验结果已在不同审稿人和设置之间采用Holm方法校正。

## Cost

按2026年10月的标价，运行这23,652份评审的成本约为1,380美元，即每份评审5.8美分：Jev为1美元（2,700份评审），Haiku为171美元（5,400份），Sonnet为702美元（2,400份），Flash为66美元（5,400份），Pro为190美元（5,400份），Argon则按其入门价格计为248美元（2,352份）；该价格将于2026年12月1日翻倍。批处理模式可将Claude和Gemini模型的价格减半，使总成本降至约690美元。

## 4 结果

### 六位审稿人，对同一批论文有六种看法

审稿人对相同论文的评判差异很大（图 1）。在 无人工智能 声明下，Jev（按其最可能评分）、Flash 和 Pro 的评分较为宽松：它们对论文的平均评分分别为 8.43、8.08 和 7.73，接近接受，并分别将其中 99%、96% 和 99% 评为弱接受或更高。相比之下，Haiku 更为严格，论文平均评分为 3.95，即边缘拒稿，只有 7% 的论文被评为弱接受或更高。Argon 介于两者之间，平均评分为 6.60，并将其中 80% 评为

[2]Argon 有三篇论文例外：谷歌的安全过滤器以 PDF 中含有图像为由，拦截了 Argon 对这三篇论文的所有评审，因此 Argon 评审了 147 篇论文，我们报告的评审数为 23,652 篇，而不是 23,700 篇。Flash 和 Pro 评审这三篇论文时则没有遇到问题。

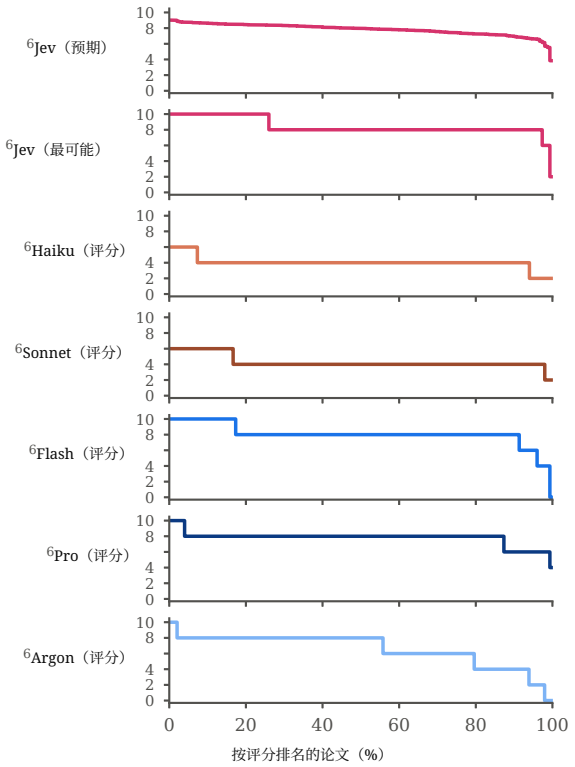



图1：每位审稿人对150篇论文的评分，按照论文评分排序；评分是在无人工智能设置下进行的首次评审。图中还显示了 Jev 的预期评分和最可能评分，以及评分设置下各语言模型的评分。

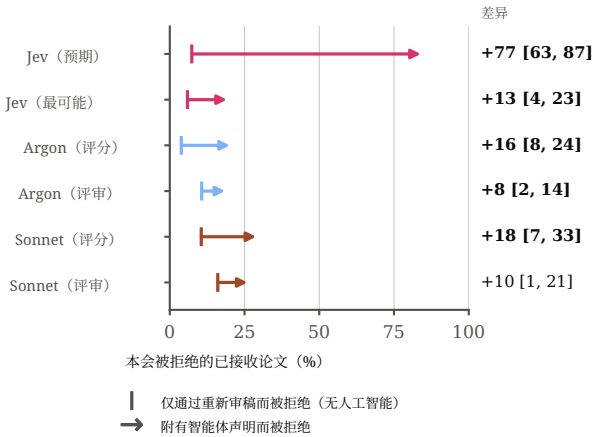



图2：在评分惩罚显著的设置中，本会被拒绝的已接收论文：仅重新审稿时采用无人工智能，以及附有智能体 声明时，若会议保留评分最高的27.4%论文（即 ICLR 2026 的录用率）。右侧为百分点差异及其95%置信区间；通过 Holm校正的结果以粗体显示。

弱接受或更高。Sonnet 与 Haiku 类似，也较为严格：它对论文的评分为 4.29，只有 17% 被评为弱接受或更高。此外，审稿人对于哪些论文更优秀也没有太多共识：在不同模型之间，其评分的斯皮尔曼秩相关系数介于 0.15 和 0.66 之间。

ICLR 2026 接收了27.4%的有效投稿（ICLR 2026 程序委员会主席 2026）。没有任何审稿人的接收率接近这一比例：Jev、Flash、Pro 和 Argon 将80%至99%的论文评为弱接受或更高，而 Haiku 和 Sonnet 仅对7%至17%的论文给出这样的评分。

每位审稿人的评分一致性同样重要。单独要求给出评分时，Jev 和 Sonnet 的一致性最高：它们对同一篇论文的两次评审中，评分不同的情况分别占5%和6%。其他设置下，这一比例为10%至38%。平均而言，Jev（按其最可能评分计算）的两次评审相差0.14分，而各语言模型的差值为0.12至0.81分。

### 未发现对辅助使用的披露偏差 ance

表2展示了主要结果。对于写作帮助、AI 编写的代码和研究帮助，没有任何审稿人的评分与其在 无人工智能 声明下的评分存在显著差异。这并不意味着这些声明没有影响，但确实限制了这种影响可能达到的程度：差异介于 $-0.07$ 和 $+0.13$ 分之间，并且没有任何 95% 置信区间低于无人工智能评分超过 0.21 分。仅智能体声明对部分审稿人以及人工智能作者声明对所有审稿人的评分产生了显著的降低作用。

### 有一些证据表明，披露会对 AI 代理和作者造成偏差

在所有设置下，每位审稿人都会明确地惩罚人工智能作者，惩罚幅度从0.22分（评审后的 Haiku）到4.12分（仅评分时的 Flash）不等。研究智能体声明的影响介于辅助使用与人工智能作者之间，在此情况下，有三位审稿人对其施加了显著惩罚。与无人工智能相比，对每篇论文的两次评审取平均后，Jev 对150篇论文中144篇的期望评分下降，平均下降0.44分，最可能评分则下降0.11分。单独要求 Argon 评分时，其评分下降0.48分：44篇论文的评分降低，8篇提高；书面审稿后，评分下降0.32分。单独要求 Sonnet 评分时，其评分下降 0.19 分：20 篇论文的评分降低，2篇提高；书面审稿后，评分下降0.13分。单独要求 Haiku 和 Flash 评分时，其评分分别下降0.13分和0.14分，Pro 则下降0.09分；这三者均未通过校正。书面审稿后，Haiku、Flash 和 Pro 也都没有表现出显著惩罚（$-0.04$、 $-0.03$和 $+0.03$）。

在注明智能体的模型后（表2，最后三行），与无人工智能相比，Jev 对所有三个注明的智能体给出的预期评分都显著更低；对于其中注明的 ChatGPT 和 Claude，其最可能评分也显著更低。要求 Haiku 仅给出评分时，它对注明为 ChatGPT 和 Gemini 的智能体评分显著更低；在进行书面审稿后，对注明为 ChatGPT 的智能体评分显著更低。要求 Flash 仅给出评分时，它对注明为 ChatGPT 的智能体评分显著更低。对于任何注明的模型，Pro 均未显示出显著的评分惩罚。

表2：两项声明之间的评分差异：前者减去后者，采用0–10评审量表，取150篇论文的平均值。粗体表示经Holm校正后达到显著性，采用双侧检验，并在行内进行校正（每项比较各为一个检验族）。评分：要求模型直接填写表格；评审：先撰写评审，再填写表格。Jev：评分分布的期望值和众数。为控制成本，Sonnet和Argon仅在四项声明下评审论文。

| | Jev | | Haiku | | Sonnet | | Flash | | Pro | | Argon | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 比较 | 预期 | mode | 评分 | 评审 | 评分 | 评审 | 评分 | 评审 | 评分 | 评审 | 评分 | 评审 |
| 写作 对比无人工智能 | +0.01 | −0.01 | +0.03 | +0.05 | −0.05 | +0.03 | +0.07 | +0.13 | −0.03 | −0.07 | −0.06 | −0.03 |
| 代码 对比无人工智能 | +0.02 | −0.01 | −0.02 | +0.06 | | | −0.03 | +0.09 | +0.02 | +0.08 | | |
| 研究帮助 对比 无人工智能 | −0.01 | −0.03 | +0.00 | −0.04 | | | −0.02 | +0.05 | −0.05 | −0.05 | | |
| 智能体 对比无人工智能 | **−0.44** | **−0.11** | −0.13 | −0.04 | **−0.19** | **−0.13** | −0.14 | −0.03 | −0.09 | +0.03 | **−0.48** | **−0.32** |
| 人工智能作者 对比无人工智能 | **−2.68** | **−2.42** | **−0.45** | **−0.22** | **−0.37** | **−0.27** | **−4.12** | **−0.51** | **−0.42** | **−0.47** | **−0.52** | **−0.43** |
| 智能体 *(ChatGPT)* 对比 无人工智能 | **−0.73** | **−0.13** | **−0.16** | **−0.15** | | | **−0.27** | −0.10 | −0.13 | −0.13 | | |
| 智能体 *(Claude)* 对比 无人工智能 | **−0.32** | **−0.09** | −0.07 | +0.03 | | | −0.19 | −0.03 | −0.04 | −0.05 | | |
| 智能体（*Gemini*）对比无人工智能 | **−0.31** | −0.07 | **−0.17** | −0.07 | | | −0.12 | −0.05 | −0.11 | −0.11 | | |

### 从评分惩罚到决策

图2显示，在存在显著评分惩罚的设置下，智能体 声明会使每位审稿人的接收论文中，有多少比例被拒绝，且这一比例超过仅重新审稿的情况。按其最可能评分计算，Jev 拒绝的接收论文比例比仅重新审稿高出13个百分点（95%置信区间为4至23）。要求 Argon 仅给出评分时，该比例高出16个百分点（8至24）；进行书面审稿后，高出8个百分点（2至14）。要求 Sonnet 仅给出评分时，该比例高出18个百分点（7至33）；进行书面审稿后，高出10个百分点（1至21）。在图2所示的效应中，只有 Sonnet 在进行书面审稿后的效应未通过校正。Haiku、Flash 和 Pro 均未显示出显著的评分惩罚。对它们而言，该声明导致的拒绝比例至多比仅重新审稿高出4个百分点，且这些效应均未通过校正。

其影响还取决于如何使用审稿人的输出。Jev 的期望评分——第3节中假设阅读后得出的评分——几乎对每篇论文都会因智能体声明而降低，平均下降 0.44 分。相比之下，最可能评分平均下降 0.11 分。期望评分有 111 个不同取值，而最可能评分最多只有六个，许多接收论文恰好略高于

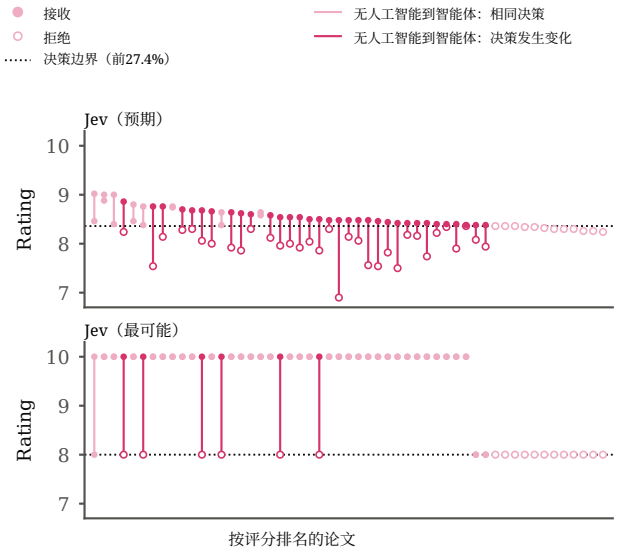



图3：Jev 在无人工智能 条件下的评分，以及对接收论文而言，在智能体声明下的评分；按无人工智能条件下的评分排序。对于这41篇接收论文中的33篇，按期望评分计算，决策发生了变化；对于其中6篇，按最可能评分计算，决策发生了变化。这里采用了两种评审顺序中的一种，以及一种平局裁决方式。图2汇总了两种顺序的结果，并扣除了单独重新审稿的影响。

边界。小幅下降就会使其中许多人跌至边界以下。因此，在期望评分下，该声明相较于单独重新审稿，多拒绝了77个百分点；而在Jev的最可能评分下，这一数值为13个百分点（图2）。

图3展示了在一种评审分配方式下对 Jev 的两种解读。按照最可能评分，大多数评分不会变动；评分下降的论文会落在边界线上，与被拒绝的论文并列。对于 Jev，该声明会将概率从最高等级转移到所有较低等级，因此期望评分会下降，但最高等级通常仍是最可能的评分。

## 5 讨论

**诚实需要付出代价吗？** 披露政策假定，诚实的声明不会损害一篇高质量论文。对于我们研究的人工智能审稿人，只要人工智能仅提供辅助，我们就没有发现披露偏差的证据。一旦声明称某个 AI 代理规划并开展了实验，这一假设就只对部分审稿人成立：Jev、Sonnet 和 Argon 给这篇论文的评分低于无人工智能时的评分。一旦声明称

人工智能完成了所有工作，每位审稿人给这篇论文的评分都会更低。因此，人工智能审稿人是否会因智能体声明而扣分，必须针对每位审稿人分别进行测量。

有人可能会质疑：审稿人有理由怀疑人工智能规划并执行的实验。然而，我们的人工智能审稿人并未核查任何内容（例如代码或数据），只是将这种怀疑用于他们刚刚读过其实验部分的论文。怀疑可以是合理的，但仅仅因为披露而降低评分，反映的是对某些人工智能用途的倾向或偏差，而不是对研究工作的公正判断。相反，怀疑应促使审稿人仔细审视研究工作本身、其证据、代码和数据。由于我们实验中的论文

除了人工智能使用声明外完全相同，因此，评分折扣只能源于审稿人读到的人工智能在研究中所发挥的作用。这引出了一个问题：任何审稿人，无论是人类还是人工智能，依据谁或什么工具参与了研究过程来评判论文，这样做是否合理？同行评审旨在评判论文的内容，而双盲评审会隐去作者身份，使其身份不会影响审稿人的判断。同样的道理也可能适用于人工智能使用声明：它们告诉审稿人哪些工具参与其中，并可能像作者身份一样左右判断。披露服务于透明度和问责制；它不是衡量论文质量的标准。审稿人是否应因为数据由硕士生收集、代码由博士生编写，或分析由博士后完成，而不是由首席研究员完成，就降低论文评分？审稿人或许应当评判数据、代码和分析，而不是评判是谁产出了它们。对于面对会因披露人工智能使用而予以惩罚的审稿人（无论是人类还是人工智能）的作者而言，激励是明确的：淡化人工智能的作用。已声明的人工智能使用量本就落后于实际使用量（He 和 Bu，2026年），而披露偏差只会进一步扩大这一差距，尤其是在我们最希望看到报告的智能体工作方面。

**什么会让评分惩罚变成拒稿**。 评分惩罚是否会导致论文被拒，取决于审稿人的输出如何被使用。Jev 的概率发生小幅变化时，其最可能评分通常保持不变，而其期望评分会随每次变化而变动。

这种影响还取决于有多少其他论文披露了人工智能使用情况。当会议录用固定比例的论文时，只有论文的相对位置才重要。如果每篇论文都披露了人工智能使用情况，并且失去相同的分数，排名就不会改变。然而，我们的审稿人对部分论文的降分幅度高于其他论文，因此即使每篇论文都进行了披露，排名也可能发生变化。我们衡量的是最简单的情况：在声明无人工智能使用的论文中，只有一位作者报告了代理式使用。只要披露智能体工作的论文很少，这种情况就具有一定现实性：在我们阅读的 160 份声明中，有 24 份进行了披露。至于随着更多作者进行披露会发生什么，我们尚未进行测试。

**启示**。 像我们这样的研究可能会被解读为人工智能审稿人存在缺陷、评审工作应当继续完全由人类负责的证明。我们对此并不完全赞同。我们没有将人工智能审稿人与人类审稿人进行比较，而人类可能同样存在偏差且缺乏一致性：在神经信息处理系统大会 2021 的一项大型实验中，882 份投稿分别由两个独立委员会进行评审；对于其中一个委员会接收的论文，另一个委员会约有一半予以拒绝（Beygelzimer 等，2023）。因此，问题不在于人工智能审稿人是否完美无缺，而在于它们可以在哪些方面改进一个本就存在缺陷的流程。对于使用人工智能审稿人对论文进行评分的会议，我们的结果提出了两点建议。第一，应根据审稿人将作出的决策，以其输出实际使用的方式对其进行测试：平均评分惩罚并不能说明它会拒绝多少篇论文，而相同的评审意见可能导致很少或很多论文被拒，具体取决于人们如何解读这些意见。第二，可以考虑不向审稿人展示人工智能使用声明；会议仍然可以要求提交该声明，但不必将其传递给审稿人。我们认为，人工智能审稿人更适合发挥作用的领域在别处，即从事那些客观、繁琐，而当今处于压力之下的系统尚无法完成的工作：核查论文主张、复现其结果以及查找参考文献中的错误。对于此类核查，作者如何使用人工智能不应成为影响因素。至于一篇论文是否有趣、新颖或重要，我们更愿意将其留给人来判断。

## 6 局限性

**模型**。 我们研究了六种审稿人：一个决策模型、来自两家提供商的各一个小型和大型语言模型，以及谷歌最新的模型。在这些语言模型中，Sonnet 和 Argon 会显著惩罚包含 智能体 声明的论文，而 Haiku、Flash 和 Pro 不会，尽管 Haiku 和 Flash 确实会惩罚一些提及智能体模型的声明。因此，这种偏差似乎并非单一提供商的特性。我们的审稿人中没有一个来自 OpenAI。

**推理力度**。 语言模型以较低的推理力度运行。我们未测试增加推理是否会减少针对智能体的披露偏差，也未测试这是否会使模型更加关注声明。

**设计**。 全部 150 篇论文均来自同一个会议，即 ICLR2027。我们使用单个人工智能审稿人模拟决策，而真实决策则综合多份评审意见和一位领域主席的意见。其中越来越多的评审意见本身是在大语言模型的帮助下撰写的（Sharma、Joachims 和 Dean，2026；Kim 等，2026）。

**声明**。 我们为每个级别使用一种措辞。与其所遵循的真实声明一样，相邻措辞之间的差异不仅在于人工智能所扮演的角色，因此我们知道哪些声明受到了惩罚，但不知道每个声明的哪一部分导致了这种惩罚。例如，人工智能作者 声明是唯一一项关于人工智能使用、但未说明作者对内容负责的声明。我们的智能体 声明也比大多数真实声明更加直白。此外，我们只替换了 AI 使用部分。论文的其他部分仍可能提及人工智能使用；对于少数论文，审稿人注意到了这一不一致。

**未来工作**。 最直接的后续研究是每次只改变 智能体 声明中的一个部分，例如由谁执行操作，或作者如何核查工作，并为每个层级测试多种措辞，以了解审稿人会对哪些内容作出反应。随后，在相同论文上测试 OpenAI 的审稿人、采用更高推理力度的审稿人以及人类审稿人，就能揭示这种偏差的普遍程度。

## 7 结论

我们没有发现人工智能辅助存在披露偏差的证据；但有些人工智能审稿人会惩罚这样一种披露：AI代理在作者监督下运行了实验；所有审稿人都会惩罚这样一种披露：人工智能完成了全部工作。这对论文造成多大影响，取决于审稿人的评分如何转化为决定，也可能取决于披露此类信息的论文数量。随着越来越多的论文由人工智能研究智能体协助产出、并由人工智能审稿，要求披露的会议应当测试在其自身的审稿程序下，修改声明是否会改变录用决定。

## 人工智能使用披露

我们通过 ClaudeCode 使用了 Anthropic 的 Claude Opus 5.5。核心理念和设计决策由作者提出。

Claude 协助设计实验、编写代码、进行分析、搜索相关文献并起草文本。作者审查了每一步，并对内容负责。

## 参考文献


Akpinar, N.-J.；Avula, S.；Lee, C.-J.；Dang, B.；Razat, K.；以及 Murdock, V. 2026。大语言模型还是人类？研究摘要中的信任与质量感知。载于*2026*年计算机系统中的人为因素会议论文集（*CHI '26*），1–25。Alharbi, E. 2026。大语言模型会审查它们所评审的内容吗？评分校准、错误检测和作者身份效应的多模态审计。arXiv:2608.28626。Altay, S.；以及 Gilardi, F. 2024。人们会怀疑被标记为人工智能生成的标题，即使其内容属实或由人类创作，因为他们假定采用了完全自动化的人工智能。*PNASNexus*，3(10)：pgae403。Beygelzimer, A.；Dauphin, Y. N.；Liang, P.；以及 Vaughan, J. W. 2023。随着该领域的发展，机器学习评审流程是否变得更加武断？NeurIPS 2021 一致性实验。arXiv:2306.03262。Biswas, J.；Schoepp, S.；Vasan, G.；Opipari, A.；Zhang, A.；Hu, Z.；Joseph, S.；Lease, M.；Li, J. J.；Stone, P.；Wagstaff, K. L.；Taylor, M. E.；以及 Jenkins, O. C. 2026。大规模人工智能辅助同行评审：AAAI-26 人工智能评审试点。arXiv:2604.13940。

Chae, S.; Kim, M.; Jung, D.; Choi, S.; and Jung, S. 2026. 自我与他者标签会在大语言模型评审者中引发双向偏差。载于计算语言学协会年会论文集：*EMNLP 2026*，33466–33482。

Cheong, I.; Guo, A.; Lee, M.; Liao, Z.; Kadoma, K.; Go, D.; Chang, J. C.; Henderson, P.; Naaman, M.; and Zhang, A. X. 2025. 惩罚透明度？人工智能披露和作者人口统计特征如何影响人类与人工智能对写作的判断。生成式人工智能披露、所有权与共创领域问责制 CHIWORK 2025 研讨会；arXiv:2507.01418。Ganjavi, C.; Eppler, M. B.; Pekcan, A.; Biedermann, B.; Abreu, A.; Collins, G. S.; Gill, I. S.; and Cacciamani, G. E. 2024. 学术与科学出版中出版商和期刊关于使用生成式人工智能的作者指南：文献计量分析。*BMJ*，384：e077192。

Hazra, S.；Lee, D.；Majumder, B. P.；以及 Kumar, S. 2026年。《经过小幅修订后接收：AI辅助科学写作的价值》。载于第*31*届国际智能用户界面会议（*IUI '26*）论文集，第1–22页。

He, Y.; and Bu, Y. 2026年. 学术期刊的人工智能政策未能遏制人工智能辅助学术写作激增。美国国家科学院院刊, 123(9): e2526734123. Holm, S. 1979. 一种简单的序贯拒绝多重检验程序。《斯堪的纳维亚统计学杂志》, 6(2): 65–70.

Howell, A.; Wang,J.; Du, L.; Melkers, J.; and Shah, V. 2025年. 声望高于能力：对同行评审中的大语言模型偏见进行改编审计。arXiv:2509.15122.

ICLR 2026 程序委员会主席。2025年。《ICLR 2026 大语言模型使用政策》。ICLR 博客，2025年8月26日，https://blog.iclr.cc/2025/08/26/policies-on-large-language-model-usage-at-iclr-2026/，访问日期：2026年10月21日。

ICLR 2026 程序委员会主席。2026年。《ICLR 2026 评审流程回顾》。htt ps://blog.iclr.cc/2026/03/31/a-retrospective-on-the-iclr-2026-review-process/，访问日期：2026年10月24日。ICLR 2027 程序委员会主席。2026年。《ICLR 2027 作者人工智能政策》。https://iclr.cc/Conferences/2027/AIPolicyForAuthors，访问日期：2026年10月21日。Kim, S. S. Y.；Deng, W. H.；Vaughan, J. W.；Su, B.；Su, W.；Agarwal, A.；Li, S.；Jaggi, M.；Goldstein, D. G.；Shah, N. B.；以及 Dudík, M. 2026年。《同行评审中大语言模型的使用及其影响：ICML 2026 的一项随机实验与调查》。arXiv:2609.19420。

Laurito, W.; Davis, B.; Grietzer, P.; Gavenčiak, T.; Böhm, A.; 和 Kulveit, J.2025年. 人工智能—人工智能偏见：大语言模型偏好大语言模型生成的通信内容。美国国家科学院院刊, 122(31): e2415697122.

Li, R.; Gu,J.-C.; Kung, P.-N.; Xia, H.; Liu, J.; Kong, X.;Sui, Z.; 和 Peng, N. 2025年. LLM-REVal：我们现在能信任大语言模型审稿人了吗？arXiv:2510.12367.

Li, Z.; Liang, C.; Peng, J.; 和 Yin, M. 2024. 披露人工智能辅助会如何影响人们对写作的看法？见*2024*年自然语言处理实证方法会议论文集, 4849–4868.

Liang, W.; Izzo, Z.; Zhang, Y.; Lepp, H.; Cao, H.; Zhao, X.; Chen, L.; Ye, H.; Liu, S.; Huang, Z.; McFarland, D. A.; and Zou, J. Y. 2024a年. 大规模监测人工智能修改的内容：ChatGPT 对人工智能会议同行评审的影响案例研究。载于第*41*届国际机器学习会议论文集，第235卷，《机器学习研究论文集》，29575–29620。

Liang, W.; Zhang, Y.; Cao, H.; Wang, B.; Ding, D. Y.; Yang, X.; Vodrahalli, K.; He, S.; Smith, D. S.; Yin, Y.; McFarland, D. A.; and Zou, J. 2024b年. 大语言模型能否为研究论文提供有用的反馈？一项大规模实证分析。*NEJM AI*，1(8)。

Lin, Z. 2026年。稿件中的隐藏提示利用人工智能辅助的同行评审。《*ACM*通讯》，69(7)：53–56。

Marioriyad, A.；Rohban, M. H.；以及 Soleymani Baghshah, M. 2025年。沉默的评审者：大语言模型评审中未被承认的捷径偏差。神经信息处理系统大会 2025不可靠数据中的可靠机器学习研讨会；arXiv:2509.26072。

Mohapatra, A.；以及 Pruthi, D. 2026年。计算机科学研究中的人工智能披露期望与实践。载于计算语言学协会年会：*EMNLP 2026*论文集，48880–48903。

Panickssery, A.；Bowman, S. R.；以及 Feng, S. 2024。大语言模型评估者识别并偏好自身生成内容。载于神经信息处理系统进展，第37卷，68772–68802。Reif, J. A.；Larrick, R. P.；以及 Soll, J. B. 2025。使用人工智能会受到社会评价惩罚的证据。美国国家科学院院刊，122(19)：e2426766122。Russo, G.；Horta Ribeiro, M.；Davidson, T. R.；Veselovsky, V.；以及 West, R. 2025。人工智能评审彩票：广泛采用人工智能辅助的同行评审提升论文评分与接收率。*ACM*人机交互汇刊，9(7)：1–28。Saraf, M.；Rezvani Boroujeni, S.；Beaudry, J.；Abedi, H.；以及 Bush, T. 2025。量化标签诱导的大语言模型自评与交叉评估偏差。arXiv:2508.21164。Schilke, O.；以及 Reimann, M. 2025。透明度困境：人工智能披露如何侵蚀信任。组织行为与人类决策过程，188：104405。Sharma, V.；Joachims, T.；以及 Dean, S. 2026年。大语言模型会偏好大语言模型吗？量化同行评审中的交互效应。载于 *2026*年*ACM*公平性、问责制与透明度会议（*FAccT '26*）论文集，3151–3189。Thakkar, N.；Yuksekgonul, M.；Silberg, J.；Garg, A.；Peng, N.；Sha, F.；Yu, R.；Vondrick, C.；以及 Zou, J. 2025。大语言模型反馈能提升评审质量吗？对 ICLR 2025 年2万份评审的随机研究。arXiv:2504.09737。TypeSafe AI。2026年。TypeSafe文档。https://docs. typesafe.ai，访问日期：2026年11月3日。Vasu, S. S. M.；Sheth, I.；Wang, H.-P.；Binkyte, R.；以及 Fritz, M. 2026年。评判中的公正：揭示大语言模型辅助同行评审中的（隐藏）偏差。载于计算语言学协会论文集：*ACL 2026*，307–330。Wataoka, K.；Takahashi, T.；以及 Ri, R. 2024。大语言模型作为评审者时的自我偏好偏差。arXiv:2410.21819。神经信息处理系统大会 2024年安全生成式人工智能研讨会。Ye, R.；Pang, X.；Chai, J.；Chen, J.；Yin, Z.；Xiang, Z.；Dong, X.；Shao, J.；以及 Chen, S. 2024。我们已经到达目标了吗？揭示在学术同行评审中使用大语言模型的风险。arXiv:2412.01708。