承诺清单：同行评审中作者承诺的审计 Chung-Chi Chen[1],Iryna Gurevych[2]

[1]AIST，日本[2]泛在知识处理实验室（UKP Lab），达姆施塔特工业大学计算机科学系及应用网络安全国家研究中心 ATHENE，德国

c.c.chen@acm.org，iryna.gurevych@tu-darmstadt.de

arXiv:2603.00003v1 [cs.CY] 8 Jan 2026

摘要

同行评审中的作者回复通常会承诺在最终论文中补充实验、发布代码或澄清内容。然而，目前尚无系统机制确保作者兑现这些承诺。在这篇立场论文中，我们使用大语言模型（LLM）比较答辩与最终定稿，开展大规模作者承诺审计。通过分析 ICLR-2025 和 EMNLP-2024 中的承诺，我们发现，尽管大多数承诺的修改都已落实，但仍有相当一部分（约 25%）未能兑现，其中“缺失的实验”等影响重大的事项最常未被履行。我们证明，基于大语言模型的工具能够切实可行地检测这些承诺。最后，我们提出作者承诺清单的构想，用于提醒作者和会议组织者注意尚未兑现的承诺，从而加强问责并维护同行评审流程的公正性。我们探讨这一做法的益处，并倡议未来的会议采用该做法。

1 引言

在同行评审的答辩阶段，作者通常会承诺根据审稿人反馈，对最终稿件进行具体改进。例如，作者可能会写道：“我们将针对数据集 X 进行一项额外实验，并将结果纳入最终定稿版本。”这类**作者承诺**旨在回应审稿人的顾虑，并可能影响决策过程。然而，论文一经录用，通常就没有正式机制来强制作者在终稿提交时落实这些承诺的修改。程序委员会和审稿人很少在论文录用后重新审阅，而目前“并无技术要求规定作者必须在稿件录用后充分处理所有审稿人的建议。”（Kuznetsov 等人，2024）。这种缺乏问责的情况可能会削弱人们对同行评审流程的信任，因为审稿人可能会怀疑，他们的反馈是否真正得到了采纳，还是仅仅被口头认可以换取论文录用。

学界已开始认识到这一差距。作为回应，一些会议场所尝试采用有条件录用和指导修改的做法，要求作者在监督下实施特定修改。过去，会议政策通常不鼓励作者在答辩中承诺提出全新结果，正是因为在作出决定时无法核实此类主张。然而，近年来，学界对答辩阶段更深入的讨论和补充结果变得更加开放，甚至予以鼓励。在实践中，许多作者如今会作出大量承诺，以说服审稿人。正如我们将要展示的，现代会议答辩中包含许多承诺，内容包括增加实验和证明、发布代码以及提高写作质量。

人工追踪数以万计的此类承诺是否兑现并不可行。自然语言处理和大语言模型的最新进展为此提供了契机。大语言模型日益广泛地应用于文档比较和一致性检查任务。例如，大语言模型可以核验抽象式摘要是否与源文档相符，或编辑后事实是否仍保持一致（Luo 等人，2023）。在同行评审领域，研究者提出了利用自然语言处理技术将评审意见与稿件修订内容对应起来（Kuznetsov 等人，2022），并检测最终定稿版本中的新增内容（Brierley 等人，2022）。基于这些进展，我们探究大语言模型能否有效识别作者承诺的修改是否未能出现在最终论文中。

在本文中，我们首次对作者承诺及其履行情况开展实证研究。我们汇编了一个新颖的数据集，收录来自两大2024–2025自然语言处理/机器学习会

议的作者回复承诺，即ICLR-2025与EMNLP-2024（Dycke等人，2023）。我们使用自动化流程提取并归类这些承诺，然后采用表现良好的大语言模型，对照论文的最终定稿版本逐一核验。我们的分析提供了以下方面的定量洞见：

- **承诺的普遍程度：**作者平均会做出多少项承诺？这一数量在不同会议之间有何差异？
- **履行率：**最终论文中，实际履行的承诺占全部承诺的比例是多少？
- **被违背承诺的性质：**哪些类型的承诺最常未被履行？它们主要是重大的新增工作（实验、理论），还是较小的修正？
- **自动化审计的可行性：**与人工判断相比，大语言模型能够多准确地识别已履行和未履行的承诺？

我们的结果证实，有相当一部分作者承诺并未兑现，尤其是那些涉及大量新增工作的承诺。例如，我们发现，补充缺失的实验或比较是最常见的承诺，同时也是最常未被履行的承诺之一。令人鼓舞的是，我们还发现，基于大语言模型的检查能够自动识别其中的大多数案例，这表明有望开发出一种实用的监督工具。

最后，我们提出了一种切实可行的解决方案，以改善问责：**作者承诺清单（ACC）**。ACC 可以利用大语言模型自动生成，将从答辩中提取的承诺以核查清单的形式呈现给作者，并在终稿提交时进行核验。我们概述了 ACC 如何融入现有工作流程，并讨论了其潜在益处，例如遏制不真诚的承诺，并确保发表的论文兑现评审者所期待的改进。我们认为，在未来的会议中采用 ACC，可以增强同行评审的可信度，同时不会给作者或组织者带来过重负担。

2 相关工作

近年来，同行评审正日益走向更高的透明度，包括公开评审和公开讨论论坛（例如，ICLR 使用的 OpenReview），以及发布评审流程报告（Shah 等人，2018）。然而，这些努力主要关注决策过程和审稿人的行为，而不是作者在决策后的行动。一个相关的发展是，在投稿过程中引入可复现性和伦理清单（Pineau 等人，2021）。例如，NeurIPS 和一些 ACL 会议如今要求作者填写有关代码和数据共享的核查清单。这些核查清单旨在鼓励作者在投稿时作出某些承诺（例如发布代码或详细说明评估设置）。既有研究表明，尽管引入了可复现性和报告核查清单，但对建议做法的合规仍不一致（Dodge 等人，2019）。这些核查清单鼓励作者在投稿时作出承诺，但并未明确要求作者在录用后履行所承诺的行动（例如提供代码）。另一个相关方向是由社区主导的发表后验证（Pineau 等人，2021）。如果作者未能兑现承诺（例如发布代码或数据集），读者或复现研究者可能会发现并记录这一情况。然而，这发生在发表之后，而且缺乏系统性。我们提出的 ACC 可以被视为迈向透明且负责任的研究传播这一更广泛运动的一部分。通过在发表前正式记录并核查承诺，ACC 将提高作者是否履行评审期间所陈述内容的可见度，从而为评审透明度补上缺失的一环。

3 方法论

我们的研究包含三个主要部分：（1）从两个会议收集数据；（2）承诺提取与分类流程；以及（3）基于大语言模型的验证系统，并由人工进行验证。以下将逐一详细介绍。

3.1 数据来源

我们聚焦于两个重要会议场所：ICLR 2025（国际学习表征会议）和 EMNLP 2024（自然语言处理实证方法会议）。选择这两个会议场所，是因为它们都设有作者回复（答辩）阶段，让作者能够回应审稿人意见；但二者在评审透明度和互动形式上存在显著差异。ICLR 使用 OpenReview 平台，评审意见以及随后作者与审稿人之间的讨论均可公开查阅。相比之下，EMNLP 2024 的作者回复

以及评审讨论在会议结束后不会公开发布。

对于 ICLR 2025，我们通过 OpenReview（可通过 API 公开访问）获取了所有论文及其相关审稿人意见和作者回复。对于 EMNLP 2024，我们从 NLPeer-v2 获取了最终被主会接收的论文的作者回复文档（Dycke 等人，2023）。在这两种情况下，我们还收集了论文的最终定稿版本。如果承诺的修改已落实，我们预期会在最终定稿版本中看到这些修改。表 1总结了这些数据集的基本统计信息。

3.2 承诺提取与验证

我们将**作者承诺**定义为作者在答辩或回复中明确表示将添加、修改、澄清或提供原投稿之外内容的陈述（例如，额外实验、分析、澄清或发布）。

承诺事项提取 我们首先使用大语言模型从答辩或回复文本中提取作者承诺。给定完整的作者回复及相关审稿人意见，我们提示模型识别作者作出的所有明确承诺。重要的是，大语言模型仅提取作者回复中表达此类承诺的原文片段，不会进行转述或改写。

对于每项提取出的承诺，模型还会根据主题相关性和上下文线索（例如提及的部分、图示或技术术语），识别并匹配该承诺所回应的相应审稿人评论。这一步会生成一份结构化列表，其中列出原子级承诺及其相关的审稿意见。

最终版本中的承诺验证为了评估承诺是否已兑现，我们再次在验证设置中使用大语言模型。对于每项提取出的承诺，模型会获得以下内容：(1) 一项作者承诺的原文；(2) 对应论文的最终（可供出版）版本；以及 (3) 相关审稿人评论，作为上下文信息。

模型判断承诺的更改是否在最终版本中实现，并输出二元判断（已履行或未履行），还可附上简短的理由说明，指出最终论文中的具体证据（例如，

	ICLR 2025	EMNLP 2024
收集的论文总数	7,192	1,050
提取的承诺总数	84,739	4,320
每篇论文的平均承诺数	11.78	4.11
承诺中位数	11	3
最大承诺数（单篇论文）	59	21

表1：ICLR 2025 和 EMNLP 2024 的答辩中识别出的作者承诺的汇总统计。

各部分、表格或图示）。每项承诺均单独评估，从而能够在单个作者承诺的层面进行细粒度分析。

4 结果与分析

我们现在展示研究结果，涵盖所做承诺的数量、作者履行承诺的比例，以及未履行承诺的特征。我们比较这两个会议，并考察不同类别承诺中的模式。

4.1 承诺数量：ICLR 与 EMNLP

第一个令人瞩目的发现（表1）是 ICLR 作者做出的承诺数量远高于 EMNLP。ICLR 2025 的作者平均每篇论文做出 11.8 项承诺，中位数为 11 项。相比之下，EMNLP 2024 的作者平均做出约 4.1 项承诺（中位数为 3 项）。ICLR 单篇投稿中的承诺数量最高达 59 项，而 EMNLP 的最高值为 21 项。

这些发现带来一个有趣的启示：审稿人和程序委员会主席可能心照不宣地知道，ICLR 作者会在评审后对论文做出大量修改。录用决定或许是在假设这些修改会落实的前提下做出的。这意味着，如果这些修改没有落实，最终发表的论文可能达不到评审期间承诺的水平。因此，在回复篇幅较长、承诺繁多的会议场所，审计的必要性或许更为迫切。

4.2 履行率

最终版本中实际履行的承诺占比是多少？总体而言，我们发现大多数承诺都得到了履行，但仍有相当一部分未能兑现：

- **ICLR 2025：** 提取出的承诺中，约有 75.89% 在

严重程度/难度	Type	ICLR-2025	EMNLP-2024
高/高	缺失的实验/比较	28.10%	31.54%
	缺失的理论/证明	2.80%	0.94%
高/低	未发布的代码/数据	4.76%	2.89%
	未阐明的概念/定义	8.10%	7.34%
	缺失的讨论/局限性	16.46%	17.95%
低/ 低	未修正的拼写错误/写作	7.49%	9.29%
	未调整的结构/附录	8.37%	7.65%
	缺少引文/参考文献	3.61%	5.85%
低/高	缺失图/示例	13.56%	11.63%
未知	其他	6.75%	4.92%

表 2：按严重性、难度、类型和会议划分的未履行作者承诺细目。

定稿论文中得以履行，约有 24% 未履行。

- **EMNLP 2024：** 约 72.91% 的承诺得到了履行，约 27% 未能履行。

为了解未履行承诺的性质，我们根据两个维度对其进行分类：“难度”（任务执行起来有多困难）和“严重性”（未能履行对论文科学诚信的影响有多关键）。我们提出了一个四象限框架来对这些被违背的承诺进行分类，如表 2所示。[1]

第一象限：高严重性 / 高难度此类别包含对研究科学有效性至关重要，但耗费大量资源或在技术上颇具挑战、难以落实的承诺。

- **缺失的实验/比较：** 这是两场会议中最常见的未兑现承诺类型。作者常常承诺补充基线或消融实验以满足审稿人的要求，却未能兑现，这可能是由于时间有限或计算资源不足（例如 GPU 时长用尽）。

- **缺失的理论/证明：** 虽然较少见，但缺失的理论推导会动摇论文的基础性主张。

第二象限：高严重性 / 低难度这些承诺对于可复现性和清晰度至关重要，但执行起来相对容易。未能兑现这类承诺尤其令人担忧，因为这可能表明作者疏忽大意。

- **未发布的代码/数据：** 发布代码在技术上并不复杂，但常常被推迟或遗忘。这引发了人们对可复现性以及潜在“虚假宣传”的担忧。

- **未阐明的概念/定义：** 简单的澄清对于读者理解至关重要。忽略这些问题表明作者缺乏对细节的关注。

[1]完整的分类提示词以及将类别映射到象限的映射逻辑详见附录 9。

- **缺失的讨论/局限性：** 作者常常承诺讨论局限性或更广泛的影响，以安抚审稿人，但在最终版本中却省略了这些内容，可能是为了节省空间或避免突出弱点。

第三象限：低严重性 / 低难度这些都是小问题，会影响论文的润色完善程度，但很少成为决定录用与否的因素。

- **未调整的结构/附录：** 相当一部分被违背的承诺涉及重新组织内容，例如将内容移至附录。未能做到这一点往往会导致稿件杂乱无章。

- **未修正的拼写错误/写作：** 承诺会修正但最终文本中仍然存在的轻微错误。

第四象限：低严重性 / 高难度这些承诺能改善论文的呈现效果，但并非确保科学有效性所必需。

- **缺失图/示例：** 制作高质量的可视化图示或详细示例十分耗时。尽管承诺提供这些内容以增强清晰度，但缺失这些内容往往可以接受。

这项分析表明，尽管许多被违背的承诺只是小问题，但仍有相当一部分属于高严重性类别（第一象限和第二象限），对已发表研究成果的科学诚信和可复现性构成风险。

4.3 解读未履行及透明度的作用

上述结果表明，有相当一部分作者承诺未能在最终定稿版本中得到履行。重要的是，不应简单地将这些未履行的情况解读为作者怀有恶意或未能兑现承诺。相反，这些情况反映了录用后修改过程的复杂性及其所受到的种种限制。

首先，承诺在可行性和结果方面存在很大差异是很自然的。一些承诺采取的行动——例如额外实验、消融实验，

或探索性分析——可能确实经过了尝试，但最终得出负面或不确定的结果。在这种情况下，未履行并不一定意味着该承诺毫无意义。例如，审稿人可能会建议一个研究方向，后来却发现它无法提供有用信息，但作者仍可能承诺对此进行探索，而审稿人在更新得分时也可能信赖了这一承诺。即便结果是负面的，在最终论文中明确报告这一点（例如在附录或讨论部分）仍能提升透明度，并更好地体现科学过程。现代会议的论文格式越来越允许灵活使用附录来容纳此类内容，因此更容易呈现这些结果，而不必挤占核心结果的篇幅。

其次，从录用通知到终稿提交的时间通常非常紧迫。在答辩期间作出的某些承诺可能过于乐观，尤其是在必须积极回应审稿人关切的压力下。事后看来，有些承诺可能并不现实，或提出得为时过早。这引出了两个不同的问题：一方面，应避免作者作出自己无法合理履行的承诺；另一方面，如果承诺的修改未能及时完成，那么在最终版本中明确说明原因，而不是悄然省略该修改，是合理的。

第三，有些未履行的情况涉及相对次要的问题，例如未改正的拼写错误、细微的措辞修改，或承诺进行的结构调整。这些问题不太可能影响录用决定，但仍会降低已发表论文的整体质量和完善程度。一份简明的核查清单可以提醒作者履行这些承诺，有助于发现此类疏漏并提升一致性，同时不会增加太多负担。

综上所述，这些观察表明，审计作者承诺的目标不应是惩罚或监管作者，追究他们每一项未履行的承诺。相反，首要目标应是提升同行评审过程中的透明度、共同理解和后续跟进。通过明确指出哪些承诺已履行、哪些未履行以及原因，诸如作者承诺清单这样的结构化机制可以促使答辩更加诚恳，减少过度承诺，并鼓励作者记录

指标	得分
人工参考录用率	72.86%
Claude-Sonnet-4.5	75.38%
Gemini-2.5-Pro	78.00%
GPT-5	74.05%

表 3：人工参考录用率和大语言模型评估准确率。

适当情况下的负面结果或尚未解决的问题。从这个意义上说，未履行不仅是一种需要尽量减少的失败情形，也表明哪些方面可以通过更清晰的沟通和更完善的工具来加强同行评审的公正性。

5 自动化承诺审计的实际可行性

除衡量作者承诺被履行的频率之外，一个关键问题是自动化审计机制能否在真实的会议工作流程中实际部署。在本部分中，我们从两个互补的视角评估其可行性：（1）基于大语言模型的判断相对于人类参考答案接受行为的可靠性，以及（2）大规模运行此类检查的计算与财务成本。

5.1 人类参考答案接受与模型准确率

我们抽取了由大语言模型识别出的600项承诺，这些实例由通过Prolific招募的英语母语者进行标注。为评估判断一致性，我们额外纳入了200个由多名标注者标注的重叠实例。

表 3汇总了结果。我们报告参考答案接受率，其定义为标注者提供的参考答案被评估器判定为正确的比例。人类评估器的参考答案接受率为72.86%，基于大语言模型的评估器表现出相当的接受行为。在三个模型中，接受率介于74%至78%之间，模型之间的差异相对较小。

有两点值得强调。首先，人类的参考答案接受率相对较低，这表明即使是理想系统，也不应期待其达到完美的准确率。其次，准确率达到约 75% 已足以确保大多数作者不会受到误报或过多误报的困扰。随着模型不断改进，以及更多经过标注的数据

	ICLR 2025	EMNLP 2024
平均篇幅（页数）	27.89	17.54
最大篇幅（页数）	150	62

表 4：两个会议的论文篇幅统计。

承诺识别	承诺验证
平均输入词元数	2,875.82	5,384.24
平均输出词元数	713.56	525.27
输入成本（美元）	0.0036	0.0108
输出成本（美元）	0.0071	0.0063
每篇论文的总成本：0.0278美元		

表 5：每篇论文的大语言模型词元用量与成本估算（EMNLP 2024）。

可用于校准，我们预计这些结果将进一步提升。

重要的是，这些发现也为部署时的设计选择提供了依据：审计结果应明确允许存在不确定性。具体而言，作者承诺清单应包含“不适用”、“承诺不再成立”或“可能存在检测错误”等选项，使作者能够对含糊不清的情况提出异议或作出说明，而不是将模型输出视为强制执行机制。

5.2 计算与财务成本

第二个值得关注的问题是，在现代会议的规模下，自动化承诺审计是否负担得起。我们通过测量完成以下任务所需的平均输入和输出词元数量来估算成本：(1) 识别承诺；(2) 对照定稿论文验证这些承诺。

我们首先考察论文篇幅。表4 显示，EMNLP 2024 论文的平均篇幅为 17.54 页，最长达 62 页，而 ICLR 论文则明显更长。这些篇幅决定了基于大语言模型的分析所需的词元预算。

以 EMNLP 2024 为代表性案例，表5 报告了我们流程中两个主要步骤的平均词元用量和每篇论文的估算成本。具体而言，我们使用 `gemini-2.5-Pro` 进行承诺识别（输入：每 100 万词元 1.25 美元；输出：每 100 万词元 10 美元），并使用`gemini-3-preview` 进行承诺验证（输入：每 100 万词元 2 美元；输出：每 100 万词元 12 美元）。即使将两个阶段都计算在内，每篇论文的总成本也约为 0.0278 美元。

从实践角度来看，与典型的会议预算相比，这笔成本微不足道。例如，可以通过小幅提高论文注册费来消化这笔成本，也可以直接将其视为运营开销。即使是在接收数千篇论文的大型会议场所，总成本仍然不高，不会对会议财务状况造成实质性影响。

5.3 对部署的启示

综合来看，我们的研究结果表明，对作者承诺进行自动化审计在技术和经济层面都切实可行。尽管基于大语言模型的判断并不完美，但其准确率已接近人类评审的一致程度，而且通过精心设计界面、强调透明度而非强制执行，可以减少错误的影响。同时，计算成本足够低，因此可以常规部署，而不会造成显著的财务负担。

这些结果表明，将作者承诺清单整合到现有终稿工作流程中，作为一种轻量但有影响力的干预措施，是可行的。这样的系统不会增加新的审稿人工作量，也不会实施严格监管，而主要作为提醒和记录工具，促使作者落实承诺并更清晰地沟通。由此，它为在不损害同行评审协作精神的前提下提升问责提供了一条务实路径。

6 作者承诺清单

我们的结果表明，未履行往往由以下因素驱动：(i) **非故意错误**（严重性高但难度低的遗漏，例如缺少代码发布或缺少澄清）以及 (ii)**结构性疏忽**（严重性低的整理事项，例如承诺的改写或结构调整），而非故意的不当行为。在第5部分中，我们进一步表明，基于大语言模型的验证足够准确且成本可承受，可作为轻量级的提醒与文档机制加以部署。基于这些见解，我们提出**作者承诺核查清单（ACC）**：一种结构化核查清单，用于记录答辩期间作出的承诺，并提示作者要么 (a) 提供证据，证明该承诺已在最终定稿版本中得到履行，要么 (b) 就未履行的原因提供简短而明确的理由说明。

核查清单项	作者确认
“我们将在 MovieCoref 上进行评估，并在最终版本中讨论其他建议的数据集。” 类型： 缺失的实验/比较（高严重性） 系统检查：未满足 证据： 未找到“MovieCoref”；实验仅在 LitBank/Fantasy-Coref/AFT 上进行。	☐ 已履行（指明部分/表格） ☐ 未履行（简要说明原因） 原因： 资源限制 / 超出范围 / 空间 / 其他 ☐ 系统错误（模型误判）
“我们会让幻觉句子更加简洁。” 类型： 写作修订（低严重性） 系统检查： 未检测到变化 证据： 答辩与最终版本中的句子没有变化。	☐ 已完成 ☐ 未履行（可选备注） ☐ 系统错误（模型误判）

表6：面向作者的作者承诺清单（ACC）示例。作者明确确认承诺是否履行、说明未履行的理由，或报告系统误判，体现了自动化检查不具约束力且可能出错的性质。

6.1 设计原则：透明度优先于惩罚

ACC并非旨在“监管”作者或增加额外的审稿人工作量。相反，它将§5中讨论的透明度目标加以落实：如果某项承诺未被实施（例如，因为实验结果为负、不可行或被降低了优先级），最终记录应当明确说明这一点，而不是默默地将其省略。具体而言，ACC围绕三项原则进行设计：

- **让承诺清晰可见：** 展示确切的承诺文本（原文照录），以及审稿人提出该要求的具体内容。
- **让承诺履行情况可核查：** 要求提供其在最终论文中的位置（部分/表格/图）的链接，或对未履行作简短的理由说明。
- **优先处理影响重大的项目：** 强调第一/第二象限项目（高严重性），以降低论文发表后未达到审稿人在决策时预期的风险。

6.2 工作流集成

ACC 旨在以最少的摩擦融入现有的定稿论文工作流程（例如 OpenReview、Softconf）。重要的是，该工作流程默认以信任为先：作者主要负责核实并报告自身承诺的状态。同时，此设计仍随时可供审计，使会议能够在需要加强问责时启用额外的验证步骤。

步骤 1：承诺事项提取（决定前）。在决定通知发出前，自动化系统会从作者回复中逐字提取原子级作者承诺，并可选择将其与相应的审稿人意见进行对应。提取出的清单可与作者分享，供其参考，让他们在录用决定最终敲定前审阅自己提出的承诺。

步骤 2：作者自我验证（定稿前）。在定稿论文截止日期之前，作者使用 ACC 验证工具检查自己的定稿论文草稿。该工具会为每项提取出的承诺生成一份初步履行情况报告，并附上佐证（或注明未检测到佐证）。作者审核这份报告，并将其作为自查内容与定稿版本一并提交。此步骤简便且不具约束力，旨在减少意外未履行承诺的情况，而非强制确保合规。

步骤3：可选的官方验证（终稿提交后）。对于需要更强审计保障的会议场所，会议系统可在终稿截止日期之后，对已提交的最终论文选择性地运行相同的ACC验证程序。由此生成的核查清单可返回给作者确认，使其能够（i）确认履行，（ii）简要说明未履行，或（iii）报告系统误判并指出相关部分。此步骤并非所有会议场所都必需，其主要目的是支持审计与一致性检查，而非凌驾于作者判断之上。

步骤 4：面向组织者的摘要（可选）。程序委员会主席可以选择查看汇总摘要（例如，未履行的高严重性承诺事项的数量），以支持指导修改、有条件录用或事后分析，具体取决于会议的治理模式。

6.3 作者实际看到的内容：具体示例

要使 ACC 发挥作用，关键在于以可操作、可理解且与现有终稿工作流程相一致的形式向作者呈现承诺。ACC 并非直接暴露原始模型输出或机器可读表示，而是有意设计为面向作者的检查清单。每个核查清单条目都明确一项承诺，概述系统的初步评估，并要求作者主动确认履行情况，或就未履行提供简要说明。

表 6 展示了 ACC 在作者提交终稿时的呈现方式。每一行对应一项从答辩中逐字提取的原子承诺。系统提供的状态旨在作为提醒和诊断信号，而非具有约束力的判断；作者可以覆盖该状态，但必须明确执行此操作。

这种基于表单的呈现方式有两个目的。首先，它能减少因疏忽或时间紧迫而导致的意外未履行，尤其是对于难度较低的事项，例如承诺补充说明或修改措辞。其次，当承诺无法履行时（例如，由于实验结果不理想或资源限制），它会鼓励作者明确记录这一情况，而不是不作解释地省略相应修改。重要的是，ACC 无意因未履行承诺而惩罚作者；相反，它使回应承诺的结果透明且可审计，从而加强评审期间的讨论与最终发表记录之间的一致性。

6.4 预期影响与范围

ACC 主要针对第二象限和第三象限的失误，这类遗漏既很常见，也可以通过提醒和结构化文档加以预防。通过要求提供证据或简短说明，ACC 可以减少未履行承诺却不作声的情况，提高答辩内容与发表记录之间的一致性，并遏制在答辩期间过度承诺。重要的是，ACC 仍与多样化的会议规范兼容：会议场所可以先采用纯粹面向作者的提醒（不予强制执行），之后若有需要，再采用面向组织者的可选摘要，用于指导修改或有条件录用。

7 讨论与结论

本文主张，在当前同行评审流程中补上一个简单却缺失的环节：在终稿阶段让答辩期间作出的承诺清晰可见、可供核查且易于解读。我们的实证审查表明，作者承诺十分常见，且其中相当一部分最终未履行，包括缺失的实验或未公开的制品等严重性较高的事项。同时，我们强调，未履行承诺并不自动意味着存在恶意：负面结果、可行性限制以及终稿时间安排紧迫，都可能使某些承诺难以完成。正因如此，我们将 ACC 定位为一种透明度与文档机制，而非惩罚性的强制执行工具。

重要的是，ACC 在典型情况下被设计为轻量级。对于平均约有 4–11 项承诺的论文（表 1），确认一份简短的核查清单，相较于撰写答辩和准备最终修订稿，只是很小的边际负担。当前体系中更大的负担往往来自另一来源：审稿人经常要求进行许多额外实验，而先前的研究表明，此类要求可能源于时间有限以及启发式驱动的审稿行为（Purkayastha et al.,2025）。在这些条件下，作者与审稿人的互动很容易滑向这样一种模式：作者感到不得不承诺“更多”，以获得更好的得分，即便所要求的补充内容与论文的核心主张仅有微弱关联。

从这个角度看，ACC 不仅关乎问责；它也是一种反思回应承诺应具有何种含义的工具。通过要求作者要么指出履行证据，要么明确解释未履行，ACC 鼓励社区区分：（i）真正影响论文主张的要求，以及（ii）主要增加工作量而不改变科学结论的要求。这进而可以支持更健康的规范：当讨论锚定于主张与证据，而非分数驱动的协商时，作者能够更有信心、更精确地回应审稿人的关切。

更广泛地说，我们应谨慎避免为了数值评分而优化同行评审流程，进而牺牲沟通质量。我们从事的是研究，而不是“分数洗白”。设计——

出能让答辩和修订更有意义的互动机制，是学术共同体面临的一项重要课题。ACC 是朝这个方向迈出的一小步，务实而可行：它推动系统作出更明确的承诺、更诚实地兑现承诺，并使评审期间的讨论与发表记录更加一致。

局限性

我们的研究存在若干局限，在解读研究结果和所提出的作者承诺核查清单（ACC）时应予以考虑。

覆盖范围与代表性。我们仅分析了来自两个会议场所（ICLR 2025 和 EMNLP 2024）的承诺。这些会议在透明度、答辩时长和回复格式方面有所不同，但可能无法代表其他领域、答辩政策不同的会议场所（例如不设答辩或仅进行指导修改的会议场所）或期刊。

“承诺”的定义与粒度。我们将答辩中的明确行动陈述作为承诺的操作性定义。这一选择会遗漏隐含意图（例如“我们同意这很重要”，但没有明确行动），也可能以不够准确的方式拆分或合并行动（例如一句话包含多个子承诺）。因此，所报告的数量和履行率取决于提取粒度，以及“承诺”与“讨论”之间的界限。

大语言模型提取与验证错误。承诺事项提取和履行验证都依赖大语言模型的判断，而这类判断可能出错。错误包括：(i) 假阳性，即模型将非承诺性文本判定为承诺；(ii) 假阴性，即模型遗漏了含蓄的承诺；(iii) 将承诺错误地对应到其他审稿人的评论；以及 (iv) 当证据以改述形式呈现、被移至附录、出现在图表中或以间接方式表达时，验证出错。人类一致率（表 3）表明，即使对人类而言，这项任务也存在歧义，因此模型错误并非单纯的实现问题，部分也反映了内在主观性。

履行作为二元标签。 我们将履行视为二元结果，但许多实际情况是部分履行：作者可能只进行部分而非全部承诺的实验，增加讨论却未达到所暗示的深度，或提供不完整或无法运行的代码。因此，二元评分可能会过度体现合规（将最低限度的改动计为已履行）和不合规（将部分进展计为未履行）。更忠实的评估应采用分级标签（已履行／部分履行／未履行），并将“已尝试但得到负面结果”与“未尝试”区分开来。

解读与规范性范围。 未履行不应被等同于恶意。有些承诺可能变得无法实现、超出范围，或产生负面结果。我们的分析并未估算作者的意图、投入，或每项承诺对录用的决策关键性。因此，ACC 应被视为一种透明度和文档记录机制，而非强制执行工具；在部署时应谨慎行事，避免激励流于表面的“勾选交差式”修改。

ACC 可能产生的行为影响。 虽然 ACC 可以减少无意遗漏，但也可能改变答辩行为：作者可能会减少明确承诺，使用模棱两可的措辞以避免接受审查，或将承诺改写成含糊的表述。评估这些二阶影响需要开展前瞻性研究（例如在某个会议上进行 A/B 测试），而本文并未进行此类研究。

致谢

本文基于 AIST 政策导向预算项目“面向物理领域的生成式人工智能基础模型研发”所取得的成果撰写。本研究由欧盟（欧洲研究委员会，Inter-Text，101054961）和 LOEWE 杰出讲席“无处不在的知识处理”共同资助；该讲席隶属于 LOEWE 计划，位于德国黑森州（资助编号：LOE WE/4a//519/05/00.002(0002)/81）。然而，本文所表达的观点和意见仅代表作者本人，并不一定反映欧盟或欧洲研究委员会的观点。欧盟和资助机构均不对此承担责任。

参考文献

Liam Brierley、Federico Nanni、Jessica K Polka、Gautam Dey、Máté Pálfy、Nicholas Fraser 和 Jonathon Alexis Coates。2022 年。大流行病期间预印本发布与期刊发表之间的变化追踪。《公共科学图书馆·生物学》，20(2):e3001285。

Jesse Dodge、SuchinGururangan、Dallas Card、Roy Schwartz 和 Noah A Smith。2019 年。展示你的

工作：改进实验结果的报告方式。载于《*2019年自然语言处理实证方法会议暨第九届自然语言处理国际联合会议（EMNLP-IJCNLP）论文集*》，第 2185–2194 页。

Nils Dycke、Ilia Kuznetsov和Iryna Gurevych。2023。NLPeer：用于同行评审计算研究的统一资源。载于计算语言学协会第*61*届年会论文集（第*1*卷：长论文），第5049–5073页，多伦多，加拿大。计算语言学协会。

Ilia Kuznetsov、Osama Mohammed Afzal、Koen Dercksen、Nils Dycke、Alexander Goldberg、Tom Hope、Dirk Hovy、JonathanK Kummerfeld、AnneLauscher、Kevin Leyton-Brown、Sheng Lu、Mausam、Margot Mieskes、Aurélie Névéol、Danish Pruthi、Lizhen Qu、Roy Schwartz、Noah A. Smith、Thamar Solorio及其他5人。2024。自然语言处理能为同行评审做些什么？*arXiv*预印本 *arXiv:2405.06563*。

Ilia Kuznetsov、Jan Buchmann、Max Eichler 和 Iryna Gurevych。2022。修改后重新提交：同行评审中基于文本协作的互文模型。《计算语言学》，48(4):949–986。

Zheheng Luo、Qianqian Xie 和 Sophia Ananiadou。2023。ChatGPT 作为文本摘要的事实不一致性评估器。*arXiv* 预印本 *arXiv:2303.15621*。

Joelle Pineau、Philippe Vincent-Lamarre、Koustuv Sinha、Vincent Larivière、Alina Beygelzimer、Florence d'Alché Buc、Emily Fox 和 Hugo Larochelle。2021。提升机器学习研究的可复现性（NeurIPS 2019 可复现性项目报告）。《机器学习研究杂志》，22(164):1–20。

Sukannya Purkayastha、Zhuang Li、Anne Lauscher、Lizhen Qu和Iryna Gurevych。2025。LazyReview：一个用于揭示自然语言处理同行评审中懒惰思维的数据集。载于《第*63*届计算语言学协会年会论文集（第*1*卷：长论文）》，第3280–3308页，奥地利维也纳。计算语言学协会。

Nihar B Shah、Behzad Tabibian、Krikamol Muandet、Isabelle Guyon和Ulrike Von Luxburg。2018。NIPS 2016评审流程的设计与分析。《机器学习研究杂志》，19(49):1–34。

A 提示词

表 7 给出了我们用于识别作者承诺的提示词，而表8 展示了用于核实承诺的提示词。

B 承诺分类的提示词

表 9 给出了用于将未兑现的承诺归入特定类别，并确定其严重性和难度象限的提示词。

C 象限映射逻辑

表9的分类结果被映射到四象限框架，以评估被违背的承诺的影响。我们分析中使用的映射逻辑详见表10。

部分	提示内容
任务描述	你将获得来自 OpenReview 讨论的两项输入： (1) 作者_回复： {author_response} （2）审稿人意见_（每条均用方括号括起并编号）： {joined_comments}
Task	提取作者**承诺**及其对应的**审稿人意见**。
承诺的定义	任何明确的作者承诺、澄清或行动声明都属于承诺，例如： 例如： • “我们已添加.....” • “我们将添加／修订／澄清／发布.....” • “我们已重新措辞.....” • “我们现在纳入.....” • “我们将此留待未来工作.....” • “我们添加了参考文献／图示／实验.....” 仅从作者_回复中提取表达 这些承诺。
对齐的审稿人评论定义	针对每项承诺，提取一个或多个**审稿人评论片段**（同样须为原文，摘自审稿人_评论），这些评论是该承诺所回应或涉及的内容。 • 每条审稿人摘录都应包含足够的上下文，以便理解批评意见或请求。 •如果有多条相关的审稿人意见，请全部列出。 • 如果没有明显匹配的内容，请为该承诺使用空列表。
匹配指南	• 按主题或线索词匹配（例如章节编号、图表提及、算法名称，诸如“DAS”、“OOD”、“Same-Object”等术语）。 •优先采用简洁但完整的文本片段（段落级）。 • 不要改述或重写任何内容。 • 保留原有标点和换行。

严格的输出格式	仅返回符合以下确切模式的有效 JSON，不要包含其他内容：
	{
	"承诺": ["..."] ,
	"已对齐_审稿人_评论": [["..."]]
	}

约束	• "承诺"和"已对齐_审稿人_评论"中的元素数量必须匹配。 • “aligned_reviewer_comments”中的每个元素都是一个列表（可能为空）。 •仅输出 JSON 块，不要解释，不要使用 Markdown。

表 7：用于提取作者承诺及其对齐的审稿人意见的提示词。

部分	提示内容
角色定义	你是一位资深学术审稿人。
输入	提供给你的内容包括： • 一份最终论文的 PDF 文件 • 答辩期间作出的作者承诺列表 • 审稿人意见，作为背景信息，帮助理解每项承诺的意图
Task	检查所提供的 PDF 文件中是否兑现了每项承诺。
承诺列表	承诺列表： {json.dumps(items, indent=2, ensure_ascii=False)}
评估目标	根据明确的证据，判断每项作者承诺中所声称的内容是否确实出现在最终论文 PDF 中，例如部分、表格、图示、实验或文本修改）。

输出格式	仅返回有效的 JSON 数组。
	示例：
	[
	{
	"promise_text": "我们将增加一项消融研究.....",
	"is_fulfilled\": true ,
	\"履行 –证据\": \"第 5 节中的表 4 展示了.....\"
	}
	]

输出约束	• 输出必须是有效的 JSON 数组 • 每个元素对应一项承诺 • JSON 输出之外不得包含解释、Markdown 或其他文本

表 8：用 于验证答辩期间作出的作者承诺是否在最终 al 论文。

部分	提示内容
角色定义	你是一位专业的学术审稿人和元审稿人。
任务描述	将以下作者承诺（来自学术论文修订计划）归入一个或多个类别，依据承诺文本和未履行的证据。
类别	只能从以下列表中选择： • 缺失的实验/比较 • 未发布的代码/数据 • 未修正的拼写错误/写作 • 缺少引文/参考文献 • 缺失图/示例 • 缺失的理论/证明 • 缺失的讨论/局限性 • 未调整的结构/附录 • 未阐明的概念/定义 • 其他 • 不是承诺
输入	**承诺文本：** "{promise_text}" **履行证据（失败原因）：** "{evidence}"
规则	1. 如果承诺涵盖多个不同的行动，可以使用多标签。 2. 如果文本是向审稿人提出的问题，或缺少具体承诺，请使用“不是承诺”。 3. 返回严格的 JSON 格式。

输出格式	仅返回符合以下模式的有效 JSON：
	{
	"is_promise": true ,
	"categories": ["类别名称"],
	"理由": "一句话解释"
	}

表9：用于将未履行的作者承诺归入定义的九个类别的提示词。

类别	象限	描述
缺失的实验/比较	Q1	高严重性 / 高难度
缺失的理论/证明	Q1	高严重性 / 高难度
未发布的代码/数据	Q2	高严重性 / 低难度
未阐明的概念/定义	Q2	高严重性 / 低难度
缺失的讨论/局限性	Q2	高严重性 / 低难度
未修正的拼写错误/写作	Q3	低严重性 / 低难度
缺少引文/参考文献	Q3	低严重性 / 低难度
未调整的结构/附录	Q3	低严重性 / 低难度
缺失图/示例	Q4	低严重性 / 高难度

表10：从类别到四象限框架的映射逻辑（严重性与难度）。