超越“不够新颖”：通过 LLM 辅助反馈丰富学术批评

Osama Mohammed Afzal[1], Preslav Nakov[2], Tom Hope[3], Iryna Gurevych[1]

[1] UKP Lab, TU Darmstadt and Hessian Center for AI (hessian.AI)
[2]MBZUAI, [3]The Allen Institute for AI (AI2),
www.ukp.tu-darmstadt.de

Abstract

创新性评估是同行评审中一个核心但尚未深入研究的方面，特别是在 NLP 等高负荷领域，评审者的能力日益紧张。我们提出了一种结构化方法，用于自动化创新性评估，通过三个阶段模拟专家评审者的行为：从投稿中提取内容，检索和综合相关工作，以及结构化比较以进行基于证据的评估。我们的方法是通过对大量人类撰写的创新性评审进行分析而得出的，并捕捉了关键模式，如独立的主张验证和上下文推理。在对 182 篇 ICLR 2025 投稿以及人工注释的评审创新性评估进行的评估中，该方法实现了与人类推理 86.5 个百分点的匹配率和 75.3 个百分点的创新性结论一致性——显著优于现有的基于大型语言模型的基线。该方法生成详细的、关注文献的分析，并提高了临时评审判断的一致性。这些结果强调了结构化语言模型辅助方法在不取代人工专业知识的情况下，支持更严格和透明的同行评审的潜力。数据和代码已提供。[1]

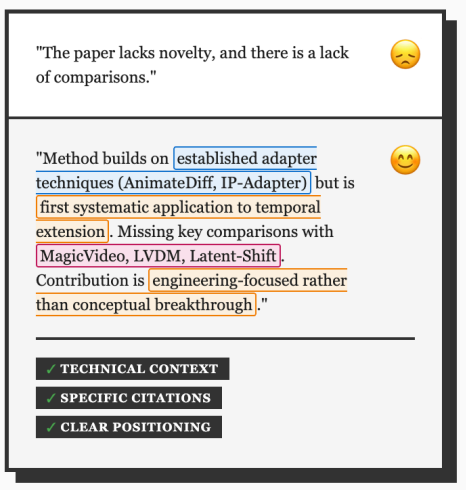

Figure 1: 表面水平的人类（顶部）与 LLM 书写的新颖性评估（底部）比较。

1 介绍

同行评审系统在其自身的成功下崩溃。在 NeurIPS 2021 上，两组独立委员会对 23 % 存在相同意见的论文 (Beygelzimer et al., 2023) 表示不同意——这种一致性问题的破裂表明比仅仅是容量限制更深层次的问题。随着稿件提交数量每 15 年翻倍 (Larsen and Ins, 2010)，而评审员目前每年处理 14 份评审 (Díaz et al., 2024)，该系统每年 1500 万小时的评审时间 (Aczel et al., 2021) 正在产生越来越不可靠的结果。

在同行评审任务中，创新性评估尤为突出并成为问题之一。创新性评估要求评审者通过识别提交内容相较于现有研究的具体进展来确定是否具有足够的原创贡献，评估这些进展是否足够显著以至于值得发表，并验证作者相较于之前研究是否准确地描述了他们的贡献。这个知识密集的过程要求评审者对相关领域的相关工作有全面的了解，并能够精确地区分有意义的创新和渐进式修改——随着出版率加快和研究领域的专业化，这项工作变得越来越困难。不堪重负的评审者常常诉诸浅层分析，产生诸如“缺乏足够创新性”之类的模糊反馈而没有明确的理由。当评审者遇到超出其特定专业领域的论文时，挑战将进一步加剧，导致过于保守的拒绝或无法捕捉渐进研究的不足评估。

近年来，大型语言模型的进步为大规模解决这些新颖性评估挑战提供了前所未有的机会。这些突破性技术革新了文本处理，并在知识密集型任务中展示了卓越的性能 (Raiaan et al., 2024)，最近的技术进步更是扩展了其在专业推理和高效推断上的能力 (Li et al., 2024a; Zhang et al., 2025)。

虽然最近的 LLM 进展创造了这个机会，但现有的工作中并没有专门作为同行评审过程中的独立任务来处理新颖性评估。之前的研究将新颖性评估纳入了想法生成的流程中 (Radensky et al., 2025; Lu et al., 2024; Li et al., 2024b)，生成的同行评审由于其存在于来自训练数据的同行评审中而进行了新颖性评估 (Idahl and Ahmadi, 2025; D'Arcy et al., 2024)，或者为

[1]https://ukplab.github.io/arxiv2025-assessing-paper-novelty

改善而在评审合成流程中增加新颖性评估步骤 (Zhu et al., 2025) 。然而，这些方法要么在合成想法上操作而不是实际的研究贡献上，要么未能独立评估新颖性评估能力。这代表着一个关键的缺口，需要专门的方法来进行同行评审新颖性评估。

为了填补这一空白，我们提出了一种用于同行评审提交的端到端创新性评估流程。我们的方法包括三个阶段：文档处理和内容提取、相关工作检索和排序，以及结构化创新性评估。最后阶段实施四个连续步骤：从提交的 pdf 中选择与创新相关的内容，从检索的论文中建立对相关工作的全面理解，将声明的创新性与前一步的综合分析进行比较，并生成包含比较中引用证据的摘要。该流程适用于真实研究论文，直接评估创新性评估能力，从而解决现有方法的局限性。重要的是，我们首次使用实际人类数据（包括标注的创新性评估声明）对大型语言模型进行创新性评估，并提供跨多个维度的全面评估。该流程适用于真实研究论文，直接评估创新性评估能力，从而解决现有方法的局限性。

研究问题与贡献 这项工作旨在解决以下研究问题：

1. 我们的具有人类信息的创新性评估流程与现有方法相比如何？
2. 我们的评估在关键评价维度上与人工审稿人偏好的一致性如何？
3. 自动化评估能否可靠地替代人工判断用于评估新颖性评估质量？

我们的贡献有三个方面：

- 人类分析数据集和见解：一个系统整理的包含 182 篇论文的数据集，这些论文来自 ICLR 2025，并带有人类新颖性评估的注释，以及对专家评审者的推理模式、评估标准和可为 AI 系统的新颖性评估设计提供信息的论点结构的经验见解。
- 人类知情的流程：一个以文献为基础的流程，结合了人类新颖性评估实践中的见解，具有结构化的提示策略和根据观察到的专家审稿人行为进行的针对性内容提取。
- 综合评估与分析：系统比较我们的人类信息方法与现有基线和人类评审人员，并在多个维度上进行细致评估，以及对自动评估方法进行验证。

2 相关研究

基于人工智能辅助的同行评审系统 我们的工作定位在科学研究的同行评审阶段，即在稿件提交评估时我们的系统开始运作。虽然以前的研究 (D'Arcy et al., 2024) (Idahl and Ahmadi, 2025) (Zhu et al., 2025) (Chitale et al., 2025) (Chang et al., 2025) (Nemecek et al., 2025) 开发了可能隐含包含新颖性评估步骤的端到端同行评审生成管道，但我们是第一个专注于构建专门的新颖性评估管道，并且是第一个系统地评估 LLM 在此任务中的表现的。一个相关的工作方向是在研究的创意阶段 (Radensky et al., 2025) (Shahid et al., 2025) (Li et al., 2024b) (Lu et al., 2024) 进行操作，开发用于研究创意生成的管道，旨在通过来自新颖性评估者的反馈循环来提高新颖性。相比之下，我们在一个更成熟的阶段进行操作，在该阶段想法已经得到了充分的执行，比较分析已经很好地制定。创意阶段工作的评估集中在通常是抽象且定义不清的合成想法上，而我们评估的是具体、精炼的研究贡献，这些贡献已经经过执行和稿件准备的精细化过程。

我们的工作采用了一个广泛的相关工作发现流程，该流程收集提交论文中引用的文献，并通过 GPT-4.1 生成的提示进行查询，另外检索相关论文。然后使用基于嵌入的方法对论文进行排名，并使用 RankGPT 重新排名。我们从现有工作 (Radensky et al., 2025) (Shahid et al., 2025) (Li et al., 2024b) 中调整这个通用方法，对排名和过滤进行了修改以适应我们特定的任务。类似的检索-排名-重新排名的流程已用于相关工作的生成 (Agarwal et al., 2025) 。另一种检索方法是 OpenScholar (Asai et al., 2024) ，其使用基于 LLM-RAG 的方法通过识别 4500 万开放获取论文中的相关段落来回答科学查询。像 DeepReviewer (Zhu et al., 2025) 这样的工作结合了 OpenScholar 进行新颖性验证。然而，我们对 OpenScholar 在新颖性评估中的主要批评在于，它仅提供一般性的比较，而不是我们任务所需的横跨方法论、问题表述、评估方法和新颖性声明的详细分析。

大语言模型生成文本的评估 以前的工作在评估生成的同行评审时采用了两种方法：一种是量化评估，比较 LLM 分配的评分（如总体评分、严谨性等）与人类在评审表上分配的评分；另一种是使用传统指标进行定性评估，如 BERTScore (Zhang et al., 2020) 、ROUGE (Lin, 2004) 和 BLEU (Papineni et al., 2002) ，或者更近期的方法如 LLM-as-Judge (Zheng et al., 2023) 。我们采用了 LLM-as-Judge 的方法进行评估。值得注意的是，之前的工作中没有专门评估 LLM

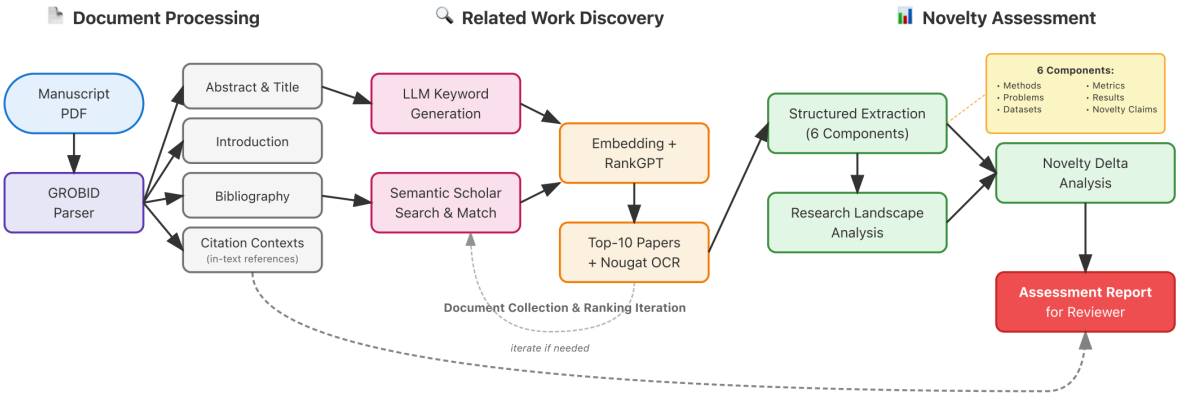

Figure 2: 自动化新颖性评估流程。系统通过三个阶段处理稿件: (1) 文档处理使用 GROBID 提取内容, (2) 相关工作发现通过嵌入相似性和 LLM 重新排序识别并排序相关论文, (3) 新颖性评估执行结构化分析以生成基于证据的新颖性评估。

在新颖性评估上的表现作为独立任务，使得我们的评估框架成为首创。

3 方法论

3.1 人类分析用于提示设计

为了理解人类如何进行新颖性评估，我们分析了来自 ICLR 2025 的评论，该会议明确要求在专门的评论部分进行新颖性评估，使得关于新颖性的讨论比在其他会议场合更为频繁。

我们从 OpenReview 获取投稿，并使用基于关键字的搜索，关键字包括 “novel”、“original”、“research gap”、“innovation”、“incremental”、“prior work” 和 “existing work”。论文的排名使用了一种评分函数，其优先级如下: (1) 包含超过 4 个新颖性关键字的评论，(2) 在评论中有一致的新颖性讨论模式，以及 (3) 评论总数。我们选择了排名前 200 篇论文进行分析。

为了加速标注过程，我们使用多个 GPT-4o mini 实例来进行句子级别的分类，以判断单个句子是否讨论了创新性。这种分类帮助人工标注者识别出创新性讨论发生的一般区域，之后人工选择所有包含实际创新性评估的句子。这个过程揭示了在 200 篇抽样论文中，有 18 篇 (9 %) 包含有限的真正创新性评估，通常是通过关键词匹配触发，这些关键词指向论文的组成部分而非创新性评估。其余的 182 篇论文形成了我们分析的最终数据集。我们系统地分析了选定的评估，以识别评论者推理、评估标准和论证结构中反复出现的模式。该分析集中于评论者如何构建他们的创新性论点、他们优先考虑哪些证据以及他们如何将投稿与先前的工作进行比较。

该分析揭示了专家审稿人在评估新颖性时的几个关键模式:

验证胜于接受: 审稿人不是直接接受作者的主张，而是独立验证与先前工作的关系，并批判性地审查作者如何描述相关研究，通常在作者的论述和实际技术关系之间进行区分。我们的提示明确要求模型 “独立验证关系” 和 “区分作者声称的区别与独立观察到的区别”，与这种关键验证方法相呼应，如图 10 和 11 所示。

多样粒度: 审稿人以不同的细节评估贡献——有些提供整体的新颖性评估，而另一些则分别检查每一个贡献与相关的先前工作进行比较。(我们通过 “贡献增量分析” 部分解决这个问题，该部分系统地检查每个声称的贡献与最相似的先前工作进行比较，确保无论作者的展示风格如何，都能全面覆盖，如图 11 所述。)

不同的分析视角: 一些评审者专注于方法学创新，而另一些评估系统时则采取整体性视角，根据领域的成熟度调整期望。我们的提示通过分别的章节——研究定位、方法关系以及领域背景考量——整合了多种分析视角，有助于根据领域成熟度调整新颖性的期望，这在图 10 和 11 中有所展示。

差距识别: 审稿人系统地识别相关工作的讨论中的差距，并区分实现层面的改进和真正的概念进步。(“相关工作考虑” 部分特别指导模型识别缺失的比较，并评估是否改进源于 “实现细节而非概念进步”，在图 11 中直接解决了这一审稿人行为。) 这些见解为我们的提示任务设计和 LLM 的输入提供了信息。

3.2 我们的方法

我们的流程通过三个阶段处理提交的 PDF 并生成结构化的新颖性评估 (图 2): (i) 文档处理从提交中提取关键内容，(ii) 相关工作发现识别并排名相关的先前工作，(iii) 新颖性评估进行比较分析以生成基于证据的新颖性评估。

我们使用 GROBID 从提交的 PDF 中提取结

构化内容，以获得后续阶段所需的标题、摘要、参考文献和引文上下文。

3.3 阶段 2：相关工作的发现

这个阶段通过一个多步骤的检索流程来识别和排名相关工作，该流程旨在捕获明确引用的作品和潜在相关但未被引用的研究。

引用文献处理 参考文献条目与 Semantic Scholar 匹配，以获取标准化的元数据（标题、摘要、作者、发布日期、发布平台），以便进行一致的后续处理。

为了识别未被作者引用的相关工作，我们使用 GPT-4.1 生成 5 个关键词查询，并在 Semantic Scholar 上进行搜索。结果经过过滤，以去除与提交内容完全相同标题的文献（避免潜在的预印本）以及提交日期之后发表的论文。

我们使用 SPECTER v2 (Singh et al., 2022) 在拼接后的标题和摘要上生成所有收集到的论文的嵌入。通过与提交物嵌入的余弦相似性对论文进行排名，以识别语义上相似的工作。

为了优先选择概念上而非纯语义上相似的论文，我们采用基于 LLM 的重新排序 (Sun et al., 2023b,a)，通过提示强调方法论方法、新颖性声明和问题陈述。我们选择排名前 K（k=20）的论文进行新颖性评估。

对于选择的论文，我们通过在 Semantic Scholar、ACL Anthology 和 arXiv 上的层次搜索来获取 PDF。检索到的论文使用 MinerU (Wang et al., 2024; He et al., 2024) 处理以提取引言部分，对于处理失败的情况使用 Nougat OCR (Blecher et al., 2023) 作为后备。我们在此使用这些工具进行 OCR，因为它们输出更准确的 OCR，而且我们将在下一阶段使用这些论文内容。

3.4 阶段 3：新颖性评估

我们使用 GPT-4.1 (OpenAI, 2024)，因为它在遵循指令方面有更好的能力。这个阶段包括四个连续的步骤。

将检索到的论文作为原始文本进行处理会导致上下文优化挑战，从而降低 LLM 性能。最近的研究表明，即使任务复杂性保持不变，模型性能也会随着输入长度的增加而持续下降 (Hong et al., 2025)。这发生的原因是，模型被无关信息所淹没会降低准确性 (Zhu et al., 2025; Idahl and Ahmadi, 2025)，或者通过大量截断导致的上下文不足限制了理解 (Radensky et al., 2025)。

我们从每篇论文的标题、摘要、引言中提取符合创新性评估要求的六个结构化组成部分：（i）方法，（ii）解决的问题，（iii）数据集，（iv）结果，（v）评估方法，和（vi）创新性声明。通过此方式可以保留重要信息，同时减少上下文长度，以缓解长而无结构化输入引起的性能下降（图 8）。

Decision	Papers	Reviews	Words/rev	Rev/paper
No Decision / Withdrawn	51	110	1002	2.16
Reject	81	195	919	2.41
Accept (Poster)	45	102	962	2.27
Accept (Spotlight)	4	10	997	2.50
Accept (Oral)	1	2	1182	2.00
Total	182	419	959	2.30

Table 1: ICLR 2025 决策结果中关于新颖性讨论的论文和评论的分布

景观分析专家评审通常会被分配到他们专业领域的论文，这为他们提供了对已建立的基准、常用技术、评价指标和最新发展的全面领域知识。为了逼近这种背景基础，我们加入了一步景观分析，系统地组织从检索到的相关工作中先前提取的结构化组件。

使用 GPT-4.1，我们进行跨论文综合，以识别方法学集群、追踪问题随时间的演变、映射评估生态系统，并建立不同方法之间的技术关系（图 9）。这种景观分析产生了一个研究空间的层次组织，明确连接了相关方法、竞争方法和互补技术之间的关系。

这种结构化的表示形式充当后续新颖性评估的背景，模仿专家评审在评估其领域论文时自然具备的组织化领域理解。

新颖性差分分析 这一步通过三个输入进行比较分析：(1) 研究领域，(2) 提交的声明贡献，和 (3) 引用上下文——即提交稿引用相关工作的句子。引用上下文揭示了作者如何定位他们的贡献，使得能够验证声称的区别与修辞框架。

利用 GPT-4.1 并结合我们的人类分析（章节 3.1）提供的提示，系统实现了关键的审稿人模式：对作者声明的独立验证、贡献的多层次详细审查，以及识别相关工作讨论中的空白（图 10 和 11）。

评估报告生成 最后一步生成的简洁段落式总结看起来类似于实际的同行评审新颖性评估，从而能够直接与人类撰写的评估进行比较（图 12）。

4 评估

在研究人类模式时，我们使用了提示设计阶段之前标注的数据。我们用每个人工评审及其对应的标注新颖性评估语句对 GPT-4.1 进行提示，以使用图 13 中的提示生成连贯的新颖性评估。这一步骤是必要的，因为新颖性相关的评论通常分散在评审的各处，而不是作为统

System	Reasoning Alignment (%)	Conclusion Agreement (%)	Positive Shift (%)	Negative Shift (%)
OpenReviewer (Idahl and Ahmadi, 2025)	42.4 ± 0.39	46.8 ± 0.71	6.3 ± 0.27	15.3 ± 0.40
DeepReviewer (Zhu et al., 2025)	50.6 ± 0.67	51.5 ± 1.24	21.7 ± 1.89	9.1 ± 0.00
Human vs. Human	65.1 ± 1.05	62.8 ± 0.40	6.7 ± 0.79	15.0 ± 0.40
Scideator (Radensky et al., 2025)	23.7 ± 0.00	22.4 ± 0.00	0.0 ± 0.00	20.5 ± 0.00
Ours	86.5 ± 0.20	75.3 ± 0.85	16.3 ± 1.28	3.0 ± 0.43

Table 2: 推理对齐、结论一致性、正向变化和负向变化指标总结

一的评估出现。简单地将这些片段合并会在评估时引入风格偏见，因为离散的人类评论与我们系统生成的连贯评估会有显著不同。合成的 GPT-4.1 评估作为我们的评估基准。

4.1 评估方法

自动评估 评估新颖性评估系统具有显著挑战，因为此任务具有主观性且依赖于知识的密集性。" 新颖" 的定义很大程度上取决于评估者对周围研究领域的熟悉程度。即便人类审稿人得出相似的新颖性结论，他们可能通过不同的推理路径和证据基础得出这些决定。

鉴于这些挑战，我们使用一种 LLM-as-Judge 框架，以我们的风格标准化的人类新颖性评估作为基准。我们通过图 14 和 15 中的提示评价 AI 生成的评估，使用 GPT-4.1 作为我们的评判：

新颖性结论一致性：AI 评估是否与人类审稿人得出相似的新颖性结论。

新颖性推理一致性：AI 的推理过程和理由是否与人类审稿人逻辑一致。

先前工作参与：评估是否表现出了与相关文献的充分参与，而不是肤浅的分析。

分析深度：评估是否提供了实质性和详细的评价，而不是表面层次的观察。

这些维度确保 AI 评估不仅与人类判断一致，而且符合彻底的、基于证据的新颖性评估的质量标准。我们的评估采用了两阶段的过程以确保一致性。首先，我们使用 GPT-4.1 从人类评论中提取核心判断（主要新颖性优势和劣势），提示如图 14 所示。我们单独进行这一提取以建立稳定的参考判断，因为将提取与评估结合起来会导致大型语言模型在不同比较中识别出不同的观点。在第二阶段，我们使用图 15 中的提示将 AI 生成的评估与这些预先提取的判断进行对比。这一评估量化了四个方面：(1) 判断相似性，衡量 AI 是否以置信度分数识别出相同的特定新颖性方面；(2) 结论一致性，检查最终的新颖性充分性判决是否一致；(3) 对先前工作的参与，分类为无，有限（1-2 个引用），或广泛（3+）；及 (4) 分析深度，分为表面层次、适中（1-2 方面）或深入（3+ 详细比较）。表格 2 报告了在这些维度上的结果一致评分。

人工评估 为了验证我们的自动评估，我们请三位专注于科学的自然语言处理和人工智能的博士生（两位三年级，一位一年级）进行人工评估，他们都有多次会议发表经验。评估使用相同的四个维度进行成对比较。

评估者对来自不同系统（人为真实数据、我们的方法或基线）的新颖性评估进行并列比较，每对代表不同的系统类型。我们共收集了 100 个比较：每位评估者进行了 25 个重复样本（用于评估者间一致性）和 25 个独特样本的比较。

对于每个维度，评估者可以选择：候选者 A 获胜，候选者 B 获胜，平局，或不清楚。评论框用于记录具体的观察和假设，从而实现定量偏好测量和定性见解。用于人类偏好收集的网页界面可以在图 6 和 7 中看到。

评审者间一致性 表 6 显示中等的评审者间一致性（0.493-0.560），其 Kappa 值为一般（0.287-0.368），反映了在评估新颖性质量时固有的主观性。

4.2 基线方法

我们将我们的方法与三个现有系统进行比较，并为新颖性评估调整每个系统。

赛迪艾特 (Radensky et al., 2025) Scideator 包含一个新颖性分类模块，该模块使用具有少样例和任务定义的 GPT-4o 将想法分类为 “新颖” 或 “非新颖”。该模块最初设计用于想法合成管道，其中大型语言模型基于新颖性反馈迭代地完善想法。我们通过将 Scideator 的 “想法” 输入转换为 “标题和摘要” 输入来进行调整——认识到这些是科学想法的结晶形式。这保留了 Scideator 的分类方法，同时从评估初步想法转向评估完成的科学贡献。

开放审稿人 (Idahl and Ahmadi, 2025) OpenReviewer 使用 Llama-OpenReviewer-8B 生成全面的同行评审，该模型训练于来自顶级会议的 79,000 条专家评论。由于它生成的是完整的评论，而不是针对创新性的评估，我们使用相同的基于 LLM 的方法从其输出中提取与创新性相关的内容，这一方法也应用于人类评审标准化，如图 13 提示中所示。

System	Surface-Level (%)	Moderate (%)	Deep (%)
OpenReviewer	67.4	31.3	1.2
DeepReviewer	43.4	56.6	0.0
Human vs. Human	22.3	66.2	11.5
Scideator	44.9	54.5	0.6
Ours	0.0	47.9	52.1

Table 3: 推理深度分布（百分比）

System	None (%)	Limited (%)	Extensive (%)
OpenReviewer	39.9	53.1	7.0
DeepReviewer	24.7	75.3	0.0
Human vs. Human	19.6	65.2	15.2
Scideator	0.0	75.9	24.1
Ours	0.0	39.1	60.9

Table 4: 先前工作参与分布（百分比）

DeepReviewer (Zhu et al., 2025) DeepReviewer 是一个多阶段的评审框架，结合了使用 OpenScholar (Asai et al., 2024) 进行的文献检索与基于证据的论证，该论证由 DeepReviewer-14B 在结构化评审注释上训练而成。我们使用与 OpenReviewer 和人工评审相同的 LLM 基于提取方法从其输出中提取新颖性评估。值得注意的是，DeepReviewer 是在 ICLR 2025 数据上进行训练的，该数据包含了我们整个评估数据集——这是在解释其性能时需考虑的一个关键因素。

5 结果与分析

我们通过将各系统的创新性评估与人类的创新性评估作为参考进行比较来评估每个系统。对于有多个人工评审的论文，我们还进行了人与人之间的比较以建立基准。表格 2 展示了整体结果。

5.1 整体表现

我们的系统在关键指标上显著优于 AI 基线和人类对人类基线。对于推理对齐，我们的系统分别比 OpenReviewer (Idahl and Ahmadi, 2025) 和 DeepReviewer (Zhu et al., 2025) 高出 44.1 和 35.9 个百分点，并且高出人类基线 21.4 个百分点。对于结论一致性，我们的系统再次领先于所有三个基线，人类基线表现最接近，约比我们的系统低 13 个百分点。我们的系统与基线的输出对比可见于表格 7 、8 和 9 。

我们分析了从新颖性结论中得出的两个指标。正向变化衡量的是与人类参考相比，从中性/消极情绪转向积极情绪的评估，而负向变化则衡量相反的方向。较高的正向变化意味着过于乐观的评估，而较高的负向变化则表明过多的批评。人工智能系统通常表现出乐观偏见，这一点从 DeepReviewer 的高正向变化得分可以体现。我们的系统的正向变化低于

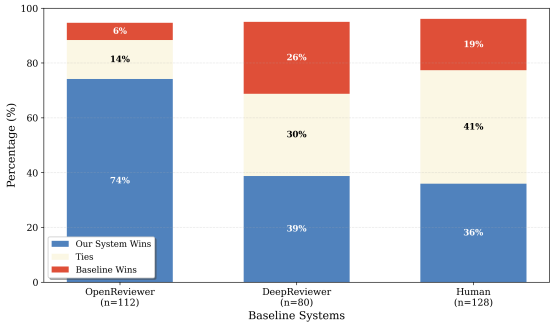

Figure 3: 我们的系统与三个基线系统基于人工评价的整体性能比较（n 值表示比较次数）

System Configuration	Reasoning (%)	Conclusion (%)
Naive Prompt	39.3	24.7
+ Our Prompt Design	80.0 (+40.7)	71.5 (+46.8)
+ Structured Extraction	83.3 (+3.3)	76.0 (+4.5)
+ Landscape Analysis (Full)	86.5 (+3.2)	75.3 (–0.7)

Table 5: 组件分析：管道组件的增量贡献

DeepReviewer，但 OpenReviewer 与人类的评分最为接近。在负向变化方面，人类往往更为批判—这种模式由 OpenReviewer 反映，随后是 DeepReviewer。我们的方法达到了最低的负向变化率。

5.2 深度和先前工作参与

表格 3 & 4 显示我们的系统在两个维度上都取得了最高分，与所有基线不同，未产生表面层次的分析。这源于我们针对新颖性评估的专业多步管道，而其他系统生成的完整同业审查中新颖性仅为一个次要组成部分。OpenReviewer 表现最差，没有检索能力并依赖于参数化知识。DeepReviewer 使用 OpenScholar 检索但未能进行比较分析。人工审稿者表现出较高的差异性，有些人参与深入而有些人仅提供最小分析。

5.3 人工评估验证

我们进行了人工评估来验证我们的 “LLM 作为评审者” 评估框架。图 3 和 4 显示了成对比较的结果。与 OpenReviewer 相比，我们的系统在 74 % 的情况下表现较好。与 DeepReviewer 和人工评审的表现更为复杂（赢率分别为 39 % 和 36 %），但较高的平局率（分别为 30 % 和 41 %）表明许多评估被认为是相当的。在所有比较中，损失率持续偏低（16-26 %）。按维度（图 4 ），“论点证明” 和 “分析质量” 表现最佳（赢率分别为 56 % 和 55 %）。创新决策显示出最多的平局（31 %），表明不同的方法通常会产生类似的结论。这些模式与我们的自动化结果一致，支持了我们评估方法的有效性。

5.4 分析：理解人类对齐模式

我们系统的高一致性评分相比于人类-人类的基线需要仔细研究。为此，我们分析了有多位人工审稿人的论文，以了解不一致的来源。

5.4.1 人工评审者变异性的来源

定性分析揭示了导致评审意见不一致的几个因素：不同的评估视角：评审者常常关注新颖性的不同方面。在提交的 Ipe4fMCBXk 中，一半的评审者强调方法学上的贡献，而另一些评审者则关注应用的新颖性，这导致从同一篇论文得出的结论相反。不同行业专家水平：评审者的背景知识会影响他们的评估结果。例如，在一篇蛋白质设计论文中，熟悉该领域历史的评审者正确地识别出了组合技术的先前工作，而其他评审者却将这些技术评估为新颖的贡献。评估的细致程度：一些评审者提供高层次的判断（“创新的方法”），而另一些评审者则聚焦于具体的技术细节。这种细致程度上的差异也导致了即使在事实基础上可能一致的情况下，评审意见的不一致。

5.4.2 系统评估的作用

我们系统的方法在应用一致的评估标准方面与人工审核不同。它对每篇论文评估多个维度（方法、应用、已有工作），在多次评估中保持分析的均匀深度，并对新颖性判断应用一致的标准。这种系统化的方法可能解释了对齐模式：当人工审核员由于关注不同方面而意见不一致时，我们系统的全面评估可以部分对齐每个观点。

5.5 组件分析

表 5 显示了每个流水线组件的增量贡献。我们基于人类认知的信息提示设计提供了最大的提升（推理增加了 40.7 %，结论增加了 46.8 %），这反映了从我们的人类分析中得出的结构化评价标准的重要性。结构化提取带来了中等的改进（推理增加了 3.3 %，结论增加了 4.5 %），但大大降低了整体计算成本和时间，而景观分析的贡献最小（推理增加了 3.2 %，结论减少了 0.7 %）。

我们提出了一种用于同行评审自动化新颖性评估的人类启发流程，解决了人工智能辅助评审系统中的一个关键缺口。我们的方法结合了系统的相关工作检索和从专家评审员模式分析中得出的结构化评估标准。实验结果表明，我们的系统在现有人工智能基准上表现优异，并在关键评估维度上比人对人的比较达成更高的一致率。该系统在深度分析和相关工作的参与方面始终表现出色，而人类评估显示，我们的评估在 41 % 的情况下与人类评审相当，各维度的损失率较低（14-21 %）。这些发现突显了系统化 AI 在新颖性评估中的潜力，以及人类新颖性判断中固有的主观性。我们的工作为科学同行评审中新颖性评估提供了更可靠、透明的基础。

6

局限性

尽管取得了强劲的表现，我们的系统还有几个重要的局限性：

评估范围：我们的评估集中在 ICLR 2025 的计算机科学论文。该系统在其他科学领域的表现尚未经过测试，并且可能需要进行特定领域的适应。

一致性与多样性：尽管我们的分析表明系统评估减少了评审者之间的分歧，但这种一致性可能会消除视角上的宝贵多样性。35-40 % 的人际分歧率可能反映了专业知识和观点上的合理差异，而不仅仅是简单的不一致。

细微的新颖性：突破性的想法常常挑战传统的评估标准。我们的系统的统一方法可能会遗漏那些人类专家通过直觉或深入领域专业知识能够识别出的范式转变的贡献。

语言范围：我们的研究仅评估了该系统在英语稿件和评论上的表现。因此，我们不能声称该方法能够推广到以其他语言编写或基于不同学术惯例的投稿；评估跨语言的表现仍是未来的工作。

References

Balazs Aczel, Barnabas Szaszi, and Alex O. Holcombe. 2021. A billion-dollar donation: estimating the cost of researchers' time spent on peer review. *Research Integrity and Peer Review*, 6(1):14.

Shubham Agarwal, Gaurav Sahu, Abhay Puri, Issam H. Laradji, Krishnamurthy DJ Dvijotham, Jason Stanley, Laurent Charlin, and Christopher Pal. 2025. Litllm: A toolkit for scientific literature review. *Preprint*, arXiv:2402.01788.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, David Wadden, Matt Latzke, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke Zettlemoyer, and 6 others. 2024. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *Preprint*, arXiv:2411.14199.

Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan. 2023. Has the machine learning review process become more arbitrary as the field has grown? the neurips 2021 consistency experiment. *Preprint*, arXiv:2306.03262.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2023. Nougat: Neural optical understanding for academic documents. *Preprint*, arXiv:2308.13418.

Yuan Chang, Ziyue Li, Hengyuan Zhang, Yuanbo Kong, Yanru Wu, Zhijiang Guo, and Ngai Wong. 2025. Treereview: A dynamic tree of questions framework for deep and efficient llm-based scientific peer review. *ArXiv*, abs/2506.07642.

Maitreya Prafulla Chitale, Ketaki Mangesh Shetye, Harshit Gupta, Manav Chaudhary, and Vasudeva Varma. 2025. Autorev: Automatic peer review system for academic research papers. *arXiv preprint arXiv:2505.14376*.

Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. 2024. Marg: Multi-agent review generation for scientific papers. *Preprint*, arXiv:2401.04259.

Oscar Díaz, Xabier Garmendia, and Juanan Pereira. 2024. Streamlining the review process: Ai-generated annotations in research manuscripts. Preprint available on arXiv.

Conghui He, Wei Li, Zhenjiang Jin, Chao Xu, Bin Wang, and Dahua Lin. 2024. Opendatalab: Empowering general artificial intelligence with open datasets. *arXiv preprint arXiv:2407.13773*.

Kelly Hong, Anton Troynikov, and Jeff Huber. 2025. Context rot: How increasing input tokens impacts llm performance. Technical report, Chroma.

Maximilian Idahl and Zahra Ahmadi. 2025. OpenReviewer: A specialized large language model for generating critical scientific paper reviews. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 550–562, Albuquerque, New Mexico. Association for Computational Linguistics.

Peder Olesen Larsen and Markus Ins. 2010. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics*, 84(3):575–603.

Baolin Li, Yankai Jiang, Vijay Gadepally, and Devesh Tiwari. 2024a. Llm inference serving: Survey of recent advances and opportunities. *2024 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–8.

Long Li, Weiwen Xu, Jiayan Guo, Ruochen Zhao, Xingxuan Li, Yuqian Yuan, Boqiang Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, Deli Zhao, Yu Rong, Tian Feng, and Lidong Bing. 2024b. Chain of ideas: Revolutionizing research via novel idea development with llm agents. *Preprint*, arXiv:2410.13185.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The AI Scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

Alexander Nemecek, Yuzhou Jiang, and Erman Ayday. 2025. The feasibility of topic-based watermarking on academic peer reviews. *Preprint*, arXiv:2505.21636.

OpenAI. 2024. Gpt-4.1 (june 2024 version). `https://platform.openai.com/`. Large language model.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Marissa Radensky, Simra Shahid, Raymond Fok, Pao Siangliulue, Tom Hope, and Daniel S. Weld. 2025. Scideator: Human-llm scientific idea generation grounded in research-paper facet recombination. *Preprint*, arXiv:2409.14634.

Mohaimenul Azam Khan Raiaan, Md. Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Jashim Sakib, Most. Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, 12:26839–26874.

Simra Shahid, Marissa Radensky, Raymond Fok, Pao Siangliulue, Daniel S Weld, and Tom Hope. 2025. Literature-grounded novelty assessment of scientific ideas. *arXiv preprint arXiv:2506.22026*.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2022. Scirepeval: A multi-format benchmark for scientific document representations. In *Conference on Empirical Methods in Natural Language Processing*.

Weiwei Sun, Zheng Chen, Xinyu Ma, Lingyong Yan, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023a. Instruction distillation makes large language models efficient zero-shot rankers. *ArXiv*, abs/2311.01555.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023b. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937, Singapore. Association for Computational Linguistics.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and Conghui He. 2024. Mineru: An open-source solution for precise document content extraction. *Preprint*, arXiv:2409.18839.

Dalong Zhang, Jun Xu, Jun Zhou, Lei Liang, Lin Yuan, Ling Zhong, Mengshu Sun, Peilong Zhao, QiWei Wang, Xiaorui Wang, Xinkai Du, Yangyang Hou, Yu Ao, ZhaoYang Wang, Zhengke Gui, ZhiYing Yi, Zhongpu Bo, Haofen Wang, and Huajun Chen. 2025. Kag-thinker: Interactive thinking and deep reasoning in llms via knowledge-augmented generation. *ArXiv*, abs/2506.17728.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2025. DeepReview: Improving LLM-based paper review with human-like deep thinking process. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29330–29355, Vienna, Austria. Association for Computational Linguistics.

A 人工评估协议：新颖性评估比较

A.1 任务设计

我们进行了一项比较评估，评估者由人类评估员判断 AI 生成的创新性评估与专家撰写的参考评估的质量。每位评估员将成对的 AI 生成评估（标记为 A 和 B）与人类专家对同一研究论文的黄金标准创新评审进行比较。

A.2 评估框架

在每次评估中，评估者会收到：(1) 一份由专家撰写的金标准新颖性评论作为参考，(2) 两份新颖性评估 (A 和 B)，系统身份被隐藏。

A.2.1 评估维度

评估人员从四个维度评估每一对。

1. 推理一致性：哪个评价更好地捕捉到了参考文献中的关键创新推理？评估者考虑了创新主张的相似性、逻辑论证和关注领域。
2. 决策一致性：哪个评估在新颖性判断上与参考最一致？这包括对整体判断（新颖性/增量/混合）的意见一致以及对新颖性因素的相似权重。
3. 论点证明：哪种评估更有效地通过证据支持其创新性主张？评估者查找具体引用、文章中的具体例子，以及是否存在无支撑的概括。
4. 分析质量：哪个评估提供了更深入的技术分析以揭示新颖性？这包括对技术讨论的深度、分析的具体性，以及对优势和局限性的平衡考虑。

对于每个维度，评估者从四个选项中选择一个：A 胜，B 胜，平局（两者同样好/差），或不清楚（无法确定）。

A.3 评估指南

A.3.1 评估员说明

评估者被指示在评估 A 和 B 之前，先彻底阅读参考评估，独立地评估每个维度，并基于实质内容而非风格差异做出判断。他们为每个例子分配 4 到 7 分钟以确保彻底的评估，并在必要时用解释性评论标记模棱两可的案例。

A.3.2 评估重点

评估者被指示要优先考虑新颖性推理的实质性和准确性，与参考判断（特别是维度 1 至 2）的对齐，技术分析的质量和深度（特别是维度 3 至 4），以及支持论点的具体证据和引用。他们被指示忽略写作风格、语法或格式的差异；与新颖性无关的论文改进建议；具有同等意义的轻微措辞变化；以及如果内容质量相当，则忽略长度差异。

A.4 实现细节

A.4.1 评估平台

评估是通过一个自定义的网页界面进行的，该界面以标准化格式呈现材料（见图 6 和 7）。每位评估者都会收到一个唯一的评估者 ID、50 对随机分配的论文评估对，以及保存进度和标记不明确案例的功能。

我们计算了评估者之间的一致性，并在表 6 中报告了 Cohen's kappa。

A.4.2 数据收集

完成的评估被提交为结构化的 JSON 文件，其中包含维度上的选择（A/B/平局/不清楚）、每次评估花费的时间，以及对标记案例的评论。

Category	Agreement	Kappa	Comparisons
Novelty Reasoning Alignment	0.520	0.341	75
Novelty Decision Alignment	0.533	0.346	75
Claim Substantiation	0.493	0.287	75
Analytical Quality	0.560	0.368	75

Table 6: 跨类别的评估者间一致性指标

B 输出示例

我们的流程的输出可以在表格 7 、8 和 9 中看到。很明显，在所有四个维度中，我们的系统与人类的对齐性优于基准。

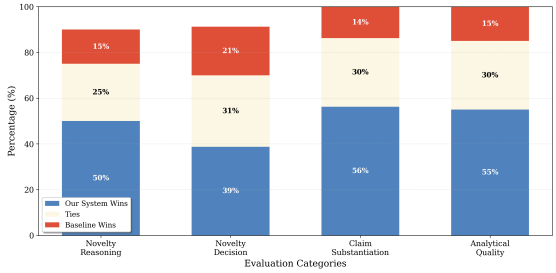

Figure 4: 在所有基线比较中汇总的各评估类别的性能分析。

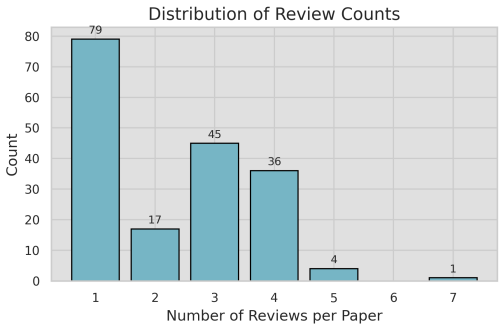

Figure 5: 每篇论文收到的评论数量分布。大多数论文收到了 1 到 4 篇评论。

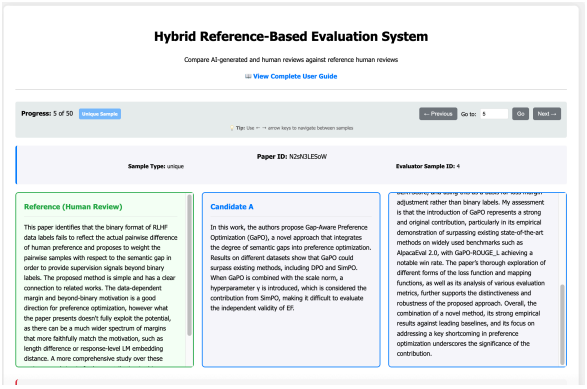

Figure 6: 用于人工评估的自定义界面的截图。注释者在多个维度上比较了 AI 生成和人工撰写的新颖性评估，包括推理深度、对先前工作的参与度以及结论的一致性。

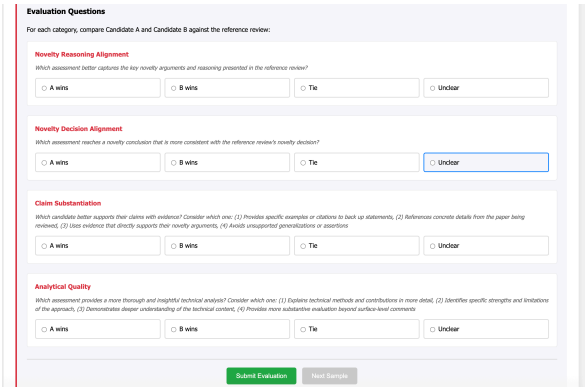

Figure 7: 用于人工评估的自定义界面的截图 (2)。标注者在多个维度上比较 AI 生成和人类撰写的新颖性评估，包括推理深度、涉及先前工作和结论一致性。

Human (Reference)	Scideator (Baseline)	Ours (Proposed)
The proposed approach aims to significantly reduce memory consumption during long-sequence inference while maintaining model performance and requiring minimal modifications to existing LLM frameworks. While the integration of CGE and RGL is presented as a novel solution to the KV cache memory challenge, [redacted]. Sparse attention mechanisms have already been [redacted], which diminishes the originality of the proposed methods. Furthermore, the CGE component [redacted]. The feedback mechanism using sliding windows has also been well-studied in H20 [3]. As such, [redacted]. The paper would benefit from a clearer articulation of how its methods differ from these established approaches to better establish its unique contribution.	[redacted] The idea is novel because IntelLLM introduces [redacted] by using strategies like center of gravity eviction (CGE) and remote gap localization (RGL), which are [redacted] such as RazorAttention[o] and LeanKV[1]. These methods prioritize retaining essential tokens and leveraging positional features to enhance compression efficiency, offering [redacted] for balancing compression rate and performance.	IntelLLM introduces new heuristics—Center of Gravity Eviction (CGE) and Remote Gap Localization (RGL)—for token selection and long-range dependency preservation in KV cache compression for LLM inference, but [redacted]. The submission [redacted], as [redacted], and similar methods (e.g., RazorAttention, PyramidKV, L $_2$ Norm) achieve comparable goals without model changes or fine-tuning. [redacted], and [redacted]. [redacted]. Reviewers should view IntelLLM as [redacted] and may wish to request more comprehensive comparisons and a more accurate positioning within the current research landscape.

Table 7: 来自人工审核者（参考）、Scideator 基线以及我们提出的系统的完整新颖性评估。关键短语被高亮显示以展示判决的一致性：[redacted]、[redacted]、[redacted] 和 [redacted]。

Research Paper Information Extraction Prompt

```
You are tasked with extracting key information from a research paper
for building a knowledge representation.
Paper title: {title}
Based on the paper content provided below, extract the following information:
- "methods": [List of methods/approaches proposed in the paper],
- "problems": [List of problems the paper addresses],
- "datasets": [List of datasets used for evaluation],
- "metrics": [List of evaluation metrics used],
- "results": [List of objects with 'metric' and 'value' fields
  representing key quantitative results],
- "novelty_claims": [Claims about what is novel in this work]
Be precise and specific.
Paper content:
{abstract}
{introduction}
```

Figure 8: 研究论文信息提取提示

Human (Reference)	DeepReviewer (Baseline)	Ours (Proposed)
The paper "DuRND: Rewarding from Novelty to Contribution for Reinforcement Learning via Dual Random Networks Distillation" proposes [redacted] by introducing two distinct random network modules—one for states deemed "successful" and another for states associated with "failure." This innovation allows for the derivation of both a "novelty" and a "contribution" reward signal, [redacted]. While [redacted], I still believe [redacted].	This paper proposes [redacted], to address the exploration-exploitation dilemma in reinforcement learning by combining novelty-based exploration and value-based exploitation in a unified framework. The novelty and contribution rewards in DuRND are based on prediction errors from random network modules, which may not capture complex state relationships in certain environments, and the reliance on prediction error as a proxy for novelty might be insufficient where the state space has a complex structure or the prediction error does not align with actual information gain. The approach introduces [redacted] but [redacted]. While the framework demonstrates effectiveness and low computational overhead, [redacted]. Overall, [redacted], but there remain [redacted] due to dependency on specific proxy measures and heuristic schedules.	DuRND is [redacted], introducing a dual random network architecture to compute both novelty (exploration) and "contribution" (exploitation) rewards, with a dynamic scheduling mechanism that shifts emphasis over time. While the use of random networks for both signals is [redacted], [redacted] (e.g., SORS, ROSA, Sibling Rivalry, SERENE). The authors accurately position DuRND as a low-overhead, scalable method, but [redacted]. Empirical results show improved performance on standard sparse-reward benchmarks, though [redacted]. Overall, DuRND offers [redacted], with its main contribution being [redacted].

Table 8: 来自人工审稿人（参考）、DeepReviewer 基线和我们提出的系统对 DuRND 论文的完整创新性评估。关键短语已加亮以显示判决的一致性：[redacted]、[redacted]、[redacted] 和 [redacted]。

Human (Reference)	OpenReviewer (Baseline)	Ours (Proposed)
This paper introduces a method for generating image inputs to Vision Language Models (VLMs) that embed "meta-instructions," enabling the model to respond to any user query appended to the image with a specified "spin," such as a particular sentiment or language. While the concept of providing meta-instructions through image inputs is [redacted] this paper offers the most thorough and comprehensive treatment of the subject to date. The general algorithm employed, which involves creating a dataset of input-output pairs and training an image via projected gradient descent to maximize the likelihood over this dataset, [redacted]. As such, [redacted]. However, the paper distinguishes itself through [redacted], particularly the study of semantic changes in images resulting from various attacks, with a focus on how meta-instruction attacks can preserve the original image meaning. The transferability experiments and the breadth of evaluation across multiple VLMs further strengthen the contribution. Overall, [redacted], as well as the new insights it provides into the vulnerabilities of VLMs to image-based prompt injection. The paper would benefit from more explicitly acknowledging the overlap with prior methods and more clearly articulating how its results advance the understanding of this threat model.	This paper introduces [redacted] as a way to attack visual language models, which is [redacted]. While the approach is new, the paper could provide more insight into how the proposed attacks compare to existing forms of adversarial attacks on VLMs, highlighting the unique aspects of meta-instruction attacks.	This assessment finds that the submission addresses indirect, cross-modal prompt injection in Visual Language Models (VLMs) by embedding hidden meta-instructions in images, aiming to steer model outputs while preserving image semantics. The work is [redacted], but distinguishes itself through [redacted]. The main substantive contributions are [redacted], and empirical evidence that image-based meta-instructions can be more effective than explicit text prompts. However, the assessment notes that [redacted], as [redacted], and [redacted]. Overall, [redacted].

Table 9: 对 VLMs 论文中元指令的人类审稿人（参考）、OpenReviewer 基线以及我们提出的系统进行的完整新颖性评估。关键短语已突出显示以展示判决的一致性：[redacted]，[redacted]，[redacted] 和 [redacted]。

Research Landscape Analysis

```
# Research Landscape Analysis

## Task
Analyze the collection of research papers provided below to create a
comprehensive map of the research landscape they represent. The submission
paper is the focus of our analysis, and the related papers provide context.

## Input Format
You will be provided with structured information extracted from multiple
research papers including:
- A submission paper that is the focus of our analysis
- Multiple related papers that form the research context

Each paper contains:
- methods: List of methods/approaches proposed
- problems: List of problems addressed
- datasets: List of datasets used
- metrics: List of evaluation metrics
- results: Key quantitative results
- novelty_claims: Claims about what is novel in the work

## Output Format
Provide a comprehensive analysis with the following sections:

1. METHODOLOGICAL LANDSCAPE
   - Identify and describe the main methodological approaches across the papers
   - Group similar or related methods into clusters
   - Highlight methodological trends or patterns
   - Describe relationships between different methodological approaches

2. PROBLEM SPACE MAPPING
   - Identify the key problems being addressed across the papers
   - Analyze how different papers approach similar problems
   - Highlight patterns in problem formulation

3. EVALUATION LANDSCAPE
   - Analyze the common datasets and evaluation methods
   - Identify patterns in how performance is measured
   - Compare evaluation approaches across papers

4. RESEARCH CLUSTERS
   - Identify groups of papers that appear closely related
   - Describe the key characteristics of each cluster
   - Analyze relationships between clusters

5. TECHNICAL EVOLUTION
   - Identify any visible progression or evolution of ideas
   - Highlight building blocks and their extensions
   - Note any competing or complementary approaches

## Example Output Format
METHODOLOGICAL LANDSCAPE
- Cluster 1: [Description of similar methods across papers]
  - Papers X, Y, Z employ transformer-based approaches with variations in...
  - These methods share characteristics such as...
  - They differ primarily in...

PROBLEM SPACE MAPPING
- Problem Area 1: [Description of a common problem addressed]
  - Papers A, B, C all address this problem but differ in...
  - The problem is formulated differently in Paper D which focuses on...

... [additional sections] ...

Ensure your analysis is comprehensive, identifying significant patterns
and relationships across the collection of papers.

## Papers:
{papers}
```

Figure 9: 研究领域分析提示

Novelty Delta Analysis for Reviewer Support - Part 1

```
# Novelty Delta Analysis for Reviewer Support

## Task
Independently analyze how the submission paper's contributions relate to existing
work in the field, critically examining both author claims and actual relationships.
This analysis should help reviewers assess novelty by providing objective comparisons
with prior work.

## Input Format
You will be provided with:
1. The structured information from the submission paper
2. A comprehensive research landscape analysis
3. Citation sentences for key related papers (how authors cite and characterize these works)

## Key Analysis Principles
- Independently verify relationships between submission and prior work
- Critically examine how authors characterize and compare with prior work
- Identify discrepancies between author characterizations and actual relationships
- Present evidence-based observations without making final judgments
- Distinguish between author-claimed differences and independently observed differences
- Provide context about field maturity and related work

## Output Format
Provide a detailed analysis with the following sections:

1. RESEARCH CONTEXT POSITIONING
   - Situate the submission within the identified research landscape
   - Identify the most closely related prior works
   - Independently assess how the submission relates to existing methodological clusters
   - Analyze its place within the problem space and evaluation approaches
   - Note: Do not accept author positioning claims without verification

2. AUTHOR CITATION ANALYSIS
   - Analyze how authors characterize and compare with each cited related work
   - Identify patterns in how authors position their contributions relative to others
   - Assess whether characterizations of prior work are accurate and balanced
   - Note discrepancies between how authors describe prior work and independent assessment
   - Evaluate whether claimed improvements or differences are substantiated
   - Identify rhetoric that may overstate differences or understate similarities

3. CONTRIBUTION DELTA ANALYSIS
   For each main contribution claimed in the submission:
   - Identify the most similar prior work for this specific contribution
   - Critically examine whether claimed differences actually exist
   - Detail exactly how this contribution differs from prior work, based on evidence
   - Compare author characterizations with independently verified relationships
   - Distinguish between substantive differences and superficial variations
   - Note when author claims about novelty or extension may be overstated
   - Consider whether improvements might be due to implementation details rather than
     conceptual advances
   - Note: Present factual observations about deltas without accepting author framing

4. FIELD CONTEXT CONSIDERATIONS
   - Provide information about how active/mature this research area is
   - Identify recent survey papers or literature reviews in this space
   - Note trends in how the field has been evolving
   - Present context about typical incremental advances in this field
   - Note: Offer context that helps reviewers calibrate their expectations
```

Figure 10: 用于评审支持的新颖性差异分析 - 第 1 部分

Novelty Delta Analysis for Reviewer Support - Part 2

```
5. CRITICAL ASSESSMENT CONSIDERATIONS
   - Identify aspects where claimed novelty may be overstated
   - Analyze whether authors' characterizations of their own novelty align with evidence
   - Consider whether empirical improvements might result from factors other than claimed
     innovations
   - Assess whether terminology differences might mask conceptual similarities
   - Identify instances where "extensions" might be routine adaptations
   - Note: Frame these as considerations rather than definitive judgments

6. RELATED WORK CONSIDERATIONS
   - Identify potentially relevant work not addressed in the submission
   - Highlight areas where additional comparisons are necessary
   - Note incomplete or potentially misleading characterizations of prior work
   - Identify when claimed "limitations" of prior work may be exaggerated
   - Compare how authors cite specific works versus how they actually relate
   - Note: Present these as information that might help complete the picture

7. KEY OBSERVATION SUMMARY
   - Highlight the most significant independently verified differences from prior work
   - Summarize the main relationships to existing research
   - Identify which claimed contributions have the strongest and weakest differentiation
   - Note the most important discrepancies between author characterizations and
     independent assessment
   - Note: Frame as observations to inform the reviewer's independent judgment

## Evidence Standards
For each observation, provide:
- Specific references to prior work
- Clear distinction between author claims and independently verified differences
- Explicit identification of similarities and differences based on technical details
- Assessment of whether differences appear substantive or superficial
- Analysis of accuracy in how authors characterize related work

## Example Format for Citation Analysis
"For [Paper X], the authors characterize it as 'limited to simple datasets' and claim
their work 'extends X to complex scenarios.' The citation sentences appear in the
following contexts:
- 'Unlike X, which only works on simple datasets, our approach handles complex
  scenarios' (Introduction)
- 'X proposed the basic framework, but did not address challenge Y' (Related Work)

Independent analysis suggests that Paper X actually did address complex scenarios
in Section 3.2, though using different terminology. The authors' characterization
appears to understate X's capabilities to emphasize their contribution. The actual
primary difference appears to be [specific technical difference] rather than the
complexity of supported scenarios."

Remember that your role is to provide objective analysis that helps reviewers make
informed judgments about novelty. Carefully examine both what authors explicitly
claim and how they implicitly position their work through their characterizations
of prior research.

{structured_representation}

## Papers from related work not cited
{not_cited_paper_titles}

##Citation Context
{citation_contexts}

## Research Landscape
{research_landscape}
```

Figure 11: 新颖性增量分析以支持审稿人 - 第二部分

Reviewer Summary Prompt

```
Summarize the following assessment in 5 sentences for a reviewer reviewing at an AI conference
    .

## Delta Assessment
{novelty_assessment}
```

Figure 12: 审稿人总结提示

Novelty Assessment Normalization Prompt

```
I'll provide you with a novelty assessment extracted from an academic peer review,
along with the full review for context. Please reformat the novelty assessment into
a standardized paragraph that begins with a brief description of the paper's contribution
before analyzing its novelty.

Example of desired format:
"This paper presents a method for neural network compression using knowledge distillation
with a focus on mobile applications. The approach has limited novelty, as it largely
builds upon existing techniques in the literature. While the authors claim their technique
is the first to combine layerwise distillation with quantization-aware training, similar
combinations have been explored in prior work by Smith et al. (2022) and Jones et al. (2023).
The main contribution appears to be a specific implementation detail in how gradient flows
are managed during the distillation process, but this incremental advance does not
significantly push the boundaries of the field. The paper would benefit from more clearly
articulating the specific differences from existing approaches to better establish its
    contribution."

Full review (for context):
{full_review}

Extracted novelty assessment to be reformatted:
{novelty_statements}

Important guidelines:
1. Begin with a clear description of what the paper presents/proposes (drawn from the full
    review if needed)
2. Create a cohesive paragraph that flows from describing the contribution to analyzing its
    novelty
3. Maintain all novelty claims and critiques from the original assessment
4. Preserve references to prior work and comparisons
5. Keep the reviewer's judgment of novelty level
6. Incorporate relevant context from the full review to provide a complete picture of the
    novelty assessment
7. Follow the structure of the example paragraph: description first, then novelty analysis
8. Preserve all critical analysis regarding limitations or strengths of novelty claims

Provide the reformatted novelty assessment:
```

Figure 13: 新颖性评估规范化提示

Core Novelty Judgment Extraction Prompt

```
Extract 2-3 core novelty judgments from this assessment:

{reference_assessment}

Focus on statements that directly assess:
- How novel/original the contribution is
- How work relates to prior research
- Specific novelty limitations
- Whether advance is incremental/fundamental

Exclude general recommendations or writing suggestions.

For each judgment, explain why it's considered a core novelty assessment.
Provide rationale for your selection of these specific judgments.
```

Figure 14: 核心新颖性判断提取提示

Reviewer Novelty Evaluation Prompt

```
Compare reviewer assessment against reference using these core judgments:

Core Judgments: {extracted_core_judgments}
Reference: {reference_assessment}
Reviewer: {reviewer_assessment}

Evaluate three dimensions:

1. JUDGMENT SIMILARITY: Do they identify same novelty strengths/weaknesses?
   - For each core judgment, find corresponding judgment in reviewer assessment
   - Assess similarity and provide detailed explanation of alignment/differences
   - Include confidence score for each comparison
   - If the core judgement is referring to a very specific aspect of the methodology
     and the reviewer assessment does not mention it, then the core judgment is
     not similar to the reviewer assessment.

2. CONCLUSION ALIGNMENT: Same bottom-line about novelty sufficiency?
   - Determine overall conclusions (SUFFICIENT / INSUFFICIENT / MIXED)
   - Explain whether conclusions align and why

3. PRIOR_WORK_ENGAGEMENT:
   - How does the reviewer engage with prior work?
   - Does the reviewer mention prior work?
   - Does the reviewer compare the current work to prior work?
   - Does the reviewer provide evidence for their claims?
   - Does the reviewer use prior work to support or critique the work?
   - Evaluate number and relevance of citations to prior work
     (NONE: no citations; LIMITED: 1 to 2; EXTENSIVE: 3+ relevant citations).

4. DEPTH_OF_ANALYSIS:
   - Assesses how deeply specific novelty aspects are compared to prior work
     (SURFACE LEVEL: vague; MODERATE: 1 to 2 aspects; DEEP: 3+ or highly detailed comparisons)

Provide explanations for all assessments to support reasoning.
```

Figure 15: 审稿人新颖性评估提示