用于院系专家评审质量评分的大语言模型

Liv Langfeldt、Dag W. Aksnes、Henrik Karlstrøm，北欧创新、研究与教育研究所（NIFU），挪威奥斯陆 N-0608。Mike Thelwall，谢菲尔德大学信息、新闻与传播学院，英国。

摘要
可以推想，同行评审人和大语言模型（LLM）在受邀评估研究时所做的事情截然不同。不过，近期证据表明，大语言模型能够在一定程度上预测已发表学术期刊论文的质量评分。大语言模型一项尚未经过检验的潜在应用是院系内部评审，这种评审可用于支持聘任与晋升决策，或筛选国家评估所需的成果。本研究首次评估：（1）大语言模型给出的质量评分与院系内部质量评分的一致程度；（2）大语言模型报告与专家报告之间的差异。研究使用一组私人数据集，其中包含谢菲尔德大学信息学院发表的58篇期刊文章，以及院系内部质量评分和报告。无论输入仅为标题/摘要还是全文，ChatGPT-4o、ChatGPT-4o mini 和 Gemini 2.0 Flash 给出的评分均与院系内部评分呈中度正相关。院系评审意见往往更具体，也体现出领域层面的知识；而 ChatGPT 报告则往往较为模式化、笼统且重复，还会主动提出改进建议。因此，研究结果：（a）证实大语言模型能够较为准确地猜测已发表学术研究的质量评分；（b）证实这种能力属于猜测而非评估（因为仅凭标题/摘要即可做到）；（c）将这种能力拓展至院系内部专家评审；（d）表明对于已发表的学术期刊论文，大语言模型报告不如人类专家报告具有洞见。

1 引言

学术晋升与聘任流程以及研究评估体系都高度依赖专家判断。在涉及研究人员的决策中，院系专家或外部专家可能会评审候选人的工作质量。最正式的例子或许是英国的研究卓越框架（REF）：该框架主要依据原创性、重要性和严谨性评估研究成果，并采用由定性描述符界定的四分制量表（REF2021，2019）。在这一流程中，所有文章至少都会经过两次正式评估：一次由院系进行，以决定提交哪些成果；另一次由外部评审人进行，以给出最终分数。长期以来，关于同行评审成本、负担及其差异性的讨论不断（Aczel 等人，2021），尤其是考虑到 REF 评估的规模（Wilsdon 等人，2015）。与此同时，大语言模型的迅速普及也引发了人们对自动化文本评估能否支持或部分取代人工评估的新兴趣。本文主要关注部门内部评估，但也借鉴了其他类型的发表后质量评估，并探讨了本文对这些评估的启示。

人工智能在研究评价中的早期应用往往侧重于提升效率，而非取代人工。例如，一家资助机构试行了人工智能辅助预筛选，

以识别不太可能获得资助的提案，同时保留由人类作出决策的权力，这引发了实践、伦理和法律方面的考量（Carbonell Cortés 等，2024年）。在出版领域，人们已探索将大语言模型用作生成手稿反馈的工具；有证据表明，在某些情境下，模型生成的评论与人类审稿人的观点重合程度，可与两位人类审稿人之间的重合程度相当（Liang 等，2024年）。这些进展促使我们进一步开展实证研究，考察大语言模型在正式评估量规下生成的“评审意见”究竟涵盖哪些内容。

一个关键的不确定性在于，大语言模型是否在有意义的层面上预测质量，还是只是在复现通用的评价性语言以及人们对“优秀科学”的感知所依据的表面相关因素，例如指出是否包含常规的方法论部分或研究局限性讨论。既有研究表明，使用适当提示词和足够强大的大语言模型进行分数预测，在所有领域都与人工评估呈正相关，但相关性因领域和输入类型而有很大差异；对多个模型的输出取平均值往往能提高相关性（Thelwall & Yaghi, 2025b；Thelwall & Yang, 2025）。然而，最强的正相关出现在向大语言模型提供文章标题和摘要、而非全文时，因此这些分数并非传统意义上的“评价”，而只是基于模式的猜测（Thelwall & Yang, 2025）。

结合上述讨论，仍有三个关键问题尚未得到解决，也是本文的研究目标。第一，所有此前对大语言模型的评估，使用的都是发表前同行评审分数或评分（Thelwall & Yang, 2025）、外部专家评审分数（Thelwall & Yang, 2025；Wu et al., 2025），或由研究团队得出的分数（Thelwall, 2025a）。尽管英国及采用类似REF的国家科研评估体系的国家普遍使用院系内部评审分数，但迄今尚无研究使用这类分数。第二，尽管已有大量研究比较大语言模型评语与发表前同行评审意见（例如 Zhou 等，2024），但尚无研究将大语言模型评语与发表后专家评审意见进行比较。这是一项重要的缺漏，因为发表前评论侧重于论文可以如何修改以及是否适合发表，而发表后专家评审意见则可能更侧重于对质量的整体考量。了解大语言模型与专家之间的差异，或许有助于揭示大语言模型如何在不评估文章的情况下猜测分数。第三，所有此前的发表后研究都使用了研究团队生成的质量评分（Thelwall, 2025a）或质量代理指标（Thelwall & Yang, 2025；Wu et al., 2025），因此，采用直接的外部质量评分进行评估，将有助于确认大语言模型总体上的质量评分能力。

本研究通过考察在使用类似REF的标准进行提示时，大语言模型如何对期刊文章进行评分和评审，来回应上述三个问题。研究将模型生成的评分和评审文本，与院系评估流程中由该领域专家针对同一院系的信息科学期刊文章所作的、与REF相关联的评估进行比较；这一评估流程来自谢菲尔德大学信息学院。本研究的总体目标不仅是量化大语言模型在这一新情境下的预测能力，还旨在更好地理解大语言模型在生成类似REF的判断时，似乎会采用何种评估信号。具体而言，研究将回答以下问题。

- 研究问题1：大语言模型生成的同行评审分数是否与院系专家评审分数呈正相关?

- 研究问题2：上述问题的答案强度在全文输入与标题/摘要输入之间，以及不同大语言模型之间有何差异?
- 研究问题3：专家评审人与大语言模型评审人撰写的发表后质量评估报告有何不同?

2 背景

2.1 使用大语言模型进行研究质量评分

目前，已有大量研究探讨大语言模型在发表前和发表后的研究评估中的潜在用途。在发表前情境中，大语言模型被用作评审人，撰写叙述性评审、结构化评议或录用建议，有时还会将其输出与编辑决定或评审人评分进行比较（Chauhan & Currie，2024）。研究表明，大语言模型在预测评分或决定（Zhou 等，2024）、识别出与人工评审者相同的优缺点（Du 等人，2024），以及为作者或编辑提供有用反馈（Lu 等，2024; Saad 等，2024）方面具有中等准确度。一项一致的发现是，大语言模型通常能够生成看似可信的评审意见，并识别常见的方法学问题，但在评估新颖性、重要性或领域特定贡献方面可能不够可靠（例如，Du 等人，2024）。

大规模证据表明，大语言模型的反馈在不少方面可以与人工评审反馈相似，同时仍存在系统性局限。对多个投稿场合数千篇论文的分析发现，GPT-4 生成的反馈与人工评审者意见之间的重合度，可能与这些场合中两位人工评审者意见之间的重合度相当；该研究也记录了不同情境下的局限性与差异（Liang et al., 2024）。这些结果支持这样一种观点：大语言模型能够提供有用的批评意见，尤其有助于发现仅凭稿件文本就能看出的常见问题。但与此同时，这些结果并不能证明模型输出可以复现选择性同行评审或国家评估项目中实际开展的专家评估。

另一相关研究方向更直接地关注预测发表后专家质量评分，而非生成反馈。对于类似 REF 的评估，大语言模型 ChatGPT、Gemini 和 Gemma3 在某些情境下能够与研究质量的外部指标或代理指标呈现弱至中等的相关性（Thelwall, 2025ade）。对多次运行的评分取平均值，始终能改善结果；不过，无论输入给大语言模型的是全文，还是仅有标题和摘要，所得效果都相近（Thelwall & Yaghi, 2025b; Thelwall, 2024; Thelwall, 2025ae）。目前有限的证据表明，Gemini 处理全文时效果更好，而 ChatGPT 仅处理文章标题和摘要时效果更好。一个持续存在的问题是评分分布较为集中（通常聚集在中高分段），而且通常高于人工评分。因此，单个分数预测通常没有意义，但用于对一组文章进行排序时便能发挥作用。例如，大语言模型给出的 3.05 分预测几乎毫无用处，除非将其转换为百分位排名（例如，在所分析文章的大语言模型评分中位列前 10%）。另一个问题是大语言模型可能存在偏见，例如偏袒或排斥某些方法论或主题（Nunkoo & Thelwall, 2025），以及对较早期研究和摘要较短的论文存在偏见（Thelwall & Kurt, 2025）。

从更广阔的视角来看，大语言模型也具备一定能力，可支持更广泛的研究评估任务，例如评估有关……影响的叙事性主张

学术研究（Kousha & Thelwall, 2025），或院系研究环境的稳健性（Kousha et al., 2025）。在这些设置中，另一个值得关注的问题是，模型可能会纳入与感知声望相关、而非与内在质量相关的声誉线索、领域惯例或训练数据伪影。对于文章评估而言，这种情况似乎不太可能发生，因为大语言模型似乎并未嵌入关于具体学术文章的知识，例如文章发表在哪份期刊、作者是谁以及发表时间（Thelwall, 2025e）。

2.2 人类与大语言模型评审者

期刊文章和会议论文等学术成果的（发表后）专家评审和（发表前）同行评审，通常以领域专门知识为基础。评审者通常依据其在论文所属研究领域的专长来遴选，既可能是较为狭窄的专长（例如期刊评审中的主题专长或方法专长），也可能是一般性的领域知识，院系评审或国家专家评审便可能属于后一种情况。评审者的专长似乎只有在编辑／管理者拥有大量潜在评审者，或可以自由选择评审者（例如期刊评审）而非只能从有限人选中挑选（例如院系评审、英国研究卓越框架）时，才最有可能与被评估文章相匹配。无论是哪种情况，评审者也可能运用一些跨领域专长来判断一般科学标准、严格程度或清晰度；而他们对论文具体研究主题和方法的专长水平则可能差异很大。尽管如此，“学术同行评审”的核心含义仍然是，研究由同行进行评估，也就是说，评审者与被评审者拥有相同或相近的能力。

鉴于现代科学高度专业化，能够胜任特定学术领域内研究评审工作的人寥寥无几（Chubin & Hackett，1990）。所需的专业水平可能会因所评估的研究方面而略有不同。例如，对学术价值／重要性和原创性的评估，可能要求评审者深入了解特定领域的前沿水平；而对方法论严谨性／严谨性的评估，则可能只需具备更广泛的评审者群体所拥有的方法论专长，除非所用方法是专业性较强或不常见的。因此，同行／专家评审意味着评审者（预期）会在其所属学科和研究领域的知识库中寻找答案，并关注其知识共同体为推动科学知识发展所关心的问题（Langfeldt 等人，2020）。

与专家／同行评审人不同，大语言模型是在海量文本语料上训练的，而非主要针对特定研究领域或知识共同体进行训练。生成式大语言模型（本文分析的类型）还能够生成语法可靠、意义连贯的文本（Radford 等人，2019），但也存在产生幻觉的风险（Sriramanan 等人，2024）。尽管训练语料库极其庞大，大语言模型也能够从其当前的“知识”中进行选择（Vaswani 等人，2017），从而恰当地、流畅地回应有关蛋糕食谱、高能物理评审意见或艺术评估的请求。因此，当它们收到一篇学术作品并被要求对其进行评估时，可能会借鉴受评文本，以及此前接触过并认为相关的评审意见和相关信息。由此生成的内容很可能反映出从表面上相关的文本中提取的模式以某种未知方式组合而成的结果。由于大语言模型采用概率方法进行学习，并且在知识重复出现时记忆得更牢，它们似乎更可能回想并复现较为笼统的评审要点（例如“文献综述不充分。”），而不是更具体的批评意见（例如“关于古诺尔斯语的讨论应提及中动态动词中有歧义的施事性。”），因为后者单独出现的频率会低得多。

不过，要确定两者之间的差异，仍需逐一进行比较，因为大语言模型非常复杂，其中包括使其能够恰当回应人类请求的各种过程（Ouyang 等人，2022年）。

3 方法

研究设计的第一部分是收集一组期刊文章及其相关的院系评审和评分。第二部分是将这些文章以不同的输入形式（即全文、标题/摘要）提交给一系列大语言模型，并将结果与针对相同期刊文章的人类专家评估进行比较。鉴于先前研究已证实，对多次试验的评分取平均值能使结果更有参考价值，因此每篇文章在每种配置下均提交 30 次，并对评分取平均值。

第三部分包括计算评审者间与评审者内一致率，并评估文本相似度。最后，两位作者阅读并定性比较了部分人类和人工智能评审报告。

3.1 数据

本论文的数据包括来自英国谢菲尔德大学（原）信息学院的 58 篇期刊文章，以及该学院匿名评审者给出的内部保密质量排名；这些排名由该学院的一组高级研究人员取平均值并进行裁定。每篇文章还附有一份摘要评审意见，由该学院的 REF 负责人根据原始个人评审意见撰写。

作为 REF 筹备工作的一部分，信息学院的员工会定期被要求提交其最优秀的作品，以供院系内部评审。内部给出的评分属于保密信息，通常只有评审员、文章作者和学院的卓越研究框架管理团队知晓。这些评分用于协助学院遴选最终提交给 REF 的作品，不用于晋升或考核。每项成果由学院卓越研究框架管理团队从学院内选出两位评审员进行单盲评审，并建议一个 REF 评分。随后，学院卓越研究框架管理团队商定最终评级（而非评分），并将两位评审员的反馈整合成一份报告，连同商定的评级一并发送给提交作品的作者。这些评级与外部 REF 评分的主要区别在于：(a) 平均而言，学院员工的资历可能不及卓越研究框架评估员；(b) 员工不会收到直接的分数预测，而是会得到关于可能分数的非正式说明，这些说明可转换为四个不同的分数等级。

由于评审报告和分数等级属于私人信息，因此需要获得学院的卓越研究框架管理团队及相关学者的许可才能查阅。经伦理审查后，我们通过发放许可申请表及一份说明数据用途的网络表单，获得了许可。填写表单的教职员工需批准将其提交的成果用于本研究。学院的卓越研究框架管理团队提供了已授权教职员工的成果名称、评审报告和分数等级。本研究以期刊文章作为原始数据。由于灰色文献成果的文档类型与期刊文章有很大差异：篇幅长得多，且面向不同的受众（从业者），因此未予采用。书籍章节本可纳入研究，但其中没有摘要；鉴于摘要在先前研究中至关重要，为本次研究专门撰写摘要似乎并不合理。

相比之下，部分会议论文与期刊文章相当。这些论文均为计算领域的全文；在该领域，会议论文往往比期刊文章更重要，因此两者属于同一类文档。然而，与期刊文章不同，这些会议论文的全文包含大量公式和数学符号，无法用标准纯文本文件加以描述。这些 PDF 文件的源文件是 LaTeX 文档描述文件。原本可以向作者索取这些文件，但这就意味着要将纯文本文件与 LaTeX 文档描述文件进行比较，造成比较不平等。因此，为避免由此导致结果偏差，我们排除了所有会议论文，仅保留期刊文章。有一篇文章是一种特殊的两部分短篇文档：期刊发表了一份两页的摘要，并附上作者自行出版的扩展版本。由于它属于不同类型的文档，因此被排除在外。所有这些抽样决策均在向大语言模型提交任何材料之前作出，以避免结果影响决策，造成偏差。

为方便起见，学院评分等级采用四点评分类似 REF 的量表，标记为 1*、2*、3* 或 4*；但出于政策原因，实际使用的标签已隐去。

我们在线获取了所提交论文的摘要和 PDF 文件，以及每篇论文的版权状态。我们编辑了 PDF 文件，删除所有可识别期刊和作者的信息，以降低大语言模型利用出版商或作者信息进行评分的可能性。此过程包括删除页码和页眉，也可能有助于大语言模型将论文正文与出版相关信息区分开来。我们对每篇论文进行了版权检查，以确定所有权和再利用条件。所有者可能是作者，也可能是出版商。前一种情况下，我们获得了再利用许可。后一种情况下，版权声明始终为知识共享署名许可（CC BY）的某种版本，允许在注明出处的情况下再利用。但该许可并未明确允许在公共大语言模型中使用内容，因为这类模型会从输入中学习，却生成不注明出处的材料。因此，我们只使用了保证不会从输入中学习的大语言模型接口。由此得出了以下数据集。

- 58 期刊文章标题和摘要。根据英国法律，这些内容可以自由转载，因此可用于机器学习。
- 58 期刊文章 PDF 文件，并按上述说明对元数据进行删减。这些文件包括作者拥有版权的 PDF 文件，以及期刊拥有版权且版权条款允许在不注明出处的情况下转载的 PDF 文件。在英国，研究版权豁免适用于文本处理，其中包括合法获取的受版权保护内容。
- 58 包含 PDF 文件中除图表外所有文本的全文文档。版权情况如上所述。
- 58 截短版文档，包含 PDF 文件中除图表和参考文献外的所有文本。版权情况如上所述。

请记住，先前使用大语言模型评估已发表期刊文章的研究，要么采用带有公开间接评分信息的已发表文章（Thelwall, 2025b），要么采用单一作者的私人评分（Thelwall, 2024）。因此，本研究是首次使用这些提示词，同时采用私人评分和不同作者撰写的文章（其中一位评分者与一位文章作者是同一人，但这些文章均未在此前的任何测试中出现）。

3.2 提示

大多数生成式大语言模型都可以通过在用户提示之前提交系统提示（指令），针对特定任务进行配置。系统指令定义任务，也可以为输出提供风格指南。本研究沿用了先前类似研究使用的同一系统提示词。具体而言，该提示词是 REF 主评审组 D（艺术、人文学科及部分社会科学）评审人员所用的 REF2021 指南（REF2021, 2019），仅对措辞稍作调整，以适应 ChatGPT 的指令风格。提示词的确切文本见在线附录（Thelwall & Yaghi, 2025a）；指令围绕严谨性、原创性和重要性展开，这些通常被认为是研究质量的核心组成部分（Langfeldt 等人, 2020）。

每次请求的用户提示均为“为这篇文章评分”，后接文章正文。此前的一项研究测试了 ChatGPT 中这一提示及参数设置的不同变体，但均未改善结果（Thelwall, 2025a），因此仅使用标准 REF 提示和 ChatGPT 的默认设置（温度=1、top_p=1）。Gemini 也采用了默认设置。

3.3 大语言模型

此前使用大语言模型预测已发表期刊文章质量评分的尝试，主要仅使用 ChatGPT，但有一项研究使用了 Google Gemini Flash 1.5。有些研究使用了下载的小型量化版本的大语言模型（精度较低的精简版本），这些模型也被用于支持发表前投稿评估，并取得了积极结果（Du 等人，2024）。本文使用了完整版本的 ChatGPT-4o（gpt-4o-2024-11-20）、成本更低的 ChatGPT-4o mini（gpt-4o-mini-2024-07-18），以及 Gemini 2.0 Flash；在数据处理时，后者是 Google 用于与 ChatGPT-4o mini 竞争的模型。由于 ChatGPT API 不允许自动提交文档，因此仅使用 Gemini 对 PDF 文件进行了测试。所有查询均于2025年3月提交。

3.4 计量指标

人工智能生成的评分与人工评分之间的关联性采用皮尔逊积差相关系数（r）和斯皮尔曼等级相关系数（ρ）进行评估。纳入皮尔逊积差相关系数是为了评估两种评分体系之间线性关系的程度，尽管其中一种评分体系采用等级；同时采用斯皮尔曼ρ，以考虑人工评分量表的顺序性质，以及可能存在的非线性但单调的关联。斯皮尔曼相关是将大语言模型评分预测与人工评分进行比较的主要指标，因为大语言模型评分的可能使用场景是对一组出版物进行排序。

先前研究发现，大语言模型评分往往集中在 3* 左右，有时集中程度很高（例如 Thelwall, 2024）。虽然可以创建查找表，将这些数值转换为 1*、2*、3* 或 4* 的完整等级，但这种做法相对武断，也会造成信息损失。相比之下，排名不受这种聚集现象的影响。此外，排名也与 REF2021 中某些评估单元使用引文数据的方式十分接近：通过报告文章的引文百分位数来提供支持。排名也是 REF 院系评估中最重要的因素，因为每个院系都会提交其排名前 n 的出版物，供国家评估使用；其中 n 根据统计日期从事研究的工作人员人数，按照特定公式确定。

为衡量模型输出之间的文本相似度，我们将大语言模型报告表示为向量空间文本嵌入；研究表明，与杰卡德相似度或余弦相似度等传统文本相似度度量相比，这种表示方式能更好地捕捉语义含义（Khosla 等人，2024）。对于每条评审意见，我们使用专用嵌入模型获取其向量嵌入，并计算每对评审意见向量之间的余弦相似度，包括模型内部和模型之间的比较。随后，对每组三十个分数取平均值，从而为每篇文章得出三个模型内相似度分数，并计算各模型之间的成对相似度分数。

3.5 文本分析

为了解大语言模型如何预测同行评审分数，本研究采用的策略是考察人工智能与人工评分相关性低的论文的大语言模型评审文本和人工评审文本，并将其与人工智能与人工评分相关性高的论文评审意见进行比较。本部分仅使用了三个数据集：基于全文和标题/摘要生成的 ChatGPT-4o 报告，以及基于全文生成的 Gemini 2.0 Flash 报告。先前研究中，ChatGPT-4o 与专家评审分数的相关性最强，因此选择了它。先前研究还发现，标题/摘要输入效果最佳，因此选择它合乎逻辑。采用全文输入也合乎逻辑，因为这与评审者所看到的内容更接近。尽管评审者会阅读 PDF 文件，但这里使用的 ChatGPT-4o API 无法处理这些文件。纳入基于全文生成的 Gemini 2.0 Flash 报告，是为了与 ChatGPT-4o 全文报告相对应，从而有机会比较全文的三个分析来源（人工、ChatGPT、Gemini）。虽然本可以为此使用 Gemini 的 PDF输入，但采用与 ChatGPT 相同的输入应有助于凸显不同模型之间的潜在差异。

定性分析所用的文章采用两步程序进行选择。首先，根据每篇论文的人工评分和人工智能评分者间一致性，我们选取人工与人工智能评分匹配度最高和最低的文章。评审者间一致性采用克里彭多夫 α 系数计算。对于每个人工评分类别（REF 评分 1、2、3 和 4），我们选取人工智能评分者间一致性最低和最高的论文。最终样本包含 15 篇文章（定量分析中的 58 篇文章总数中的一部分），其分数分布如附录中的表 A1 所示。第二步，我们在每个模型中选取第一篇（共 30 篇）整体分数与该模型平均评分相近的评审文本。最终用于文本分析的样本包含 60 份评审意见（15 份人工评审和 45 份人工智能评审）。

两位作者阅读了人工和人工智能评审报告以及摘要/论文。我们首先进行了一轮测试，分别对相同的 10 份人工评审进行编码，然后比较结果。在这一初始编码方案中，我们针对每项评审标准（原创性、重要性、严谨性），将评审意见编码为正面、负面、正面和负面兼有，或犹豫不决/未发表评论；同时记录这些评论是详细的还是笼统的，以及是否评论了三个 REF 评审标准之外的其他标准。在审阅这一初步练习的结果后，我们认为该编码方式无助于理解不同评审意见之间的差异，因此需要进行更具探索性的分析。

本文的分析基于对每篇文章的一份人工评审意见和三份人工智能评审意见进行逐一阅读（包括文章摘要，并在相关时核查文章内容），同时记录这些评审意见之间的异同。

随后，我们比较了大语言模型能够预测人工评分和无法预测人工评分的案例，也比较了模型评分低于人工评分和高于人工评分的案例。分析旨在找出有助于理解差异与相似之处的模式。

4 结果

首先报告定量证据，随后介绍定性发现 s.

4.1 相关大语言模型与院系专家评分之间的关系（研究问题*1*，*RQ 2)*

总体而言，大语言模型给出的平均分较高，ChatGPT 的所有平均分均高于 3，Gemini 的所有平均分均至少为 2.73（图 1）。对于每个模型，最低评分均出现在标题/摘要输入中。如上所述，这可能是因为文章因未包含全文中的信息而被扣分。请注意，人工评分等级属于名义尺度，不能直接比较。

就单个模型和输入下不同文章之间评分的变异性而言（即每篇论文 30 个评分的平均值），ChatGPT 的评分范围从低于 3 到 4 或接近 4 不等，但标题/摘要输入的评分要低得多（图 1）。对于全文或截断文本输入，ChatGPT 对任何文章给出的最低分为 2.89；而对于标题/摘要输入，这一分数低得多，为 2.13。

Gemini 的变异模式截然不同。除标题/摘要输入外，它几乎一味倾向于给出分数 3，很少给出更低的分数，几乎从不给出更高的分数。例如，对于 PDF输入，共有 58*30=1740 个查询，其中 1657 个（95.2%）返回了 3。因此，Gemini 非常确信这 58 篇文章几乎都值得评为 3（即在单篇报告中几乎总是预测为 3）；相比之下，ChatGPT 更愿意考虑其他分数，从 2 到 4 不等（即在针对单篇文章的报告中更频繁地提及这些分数）。

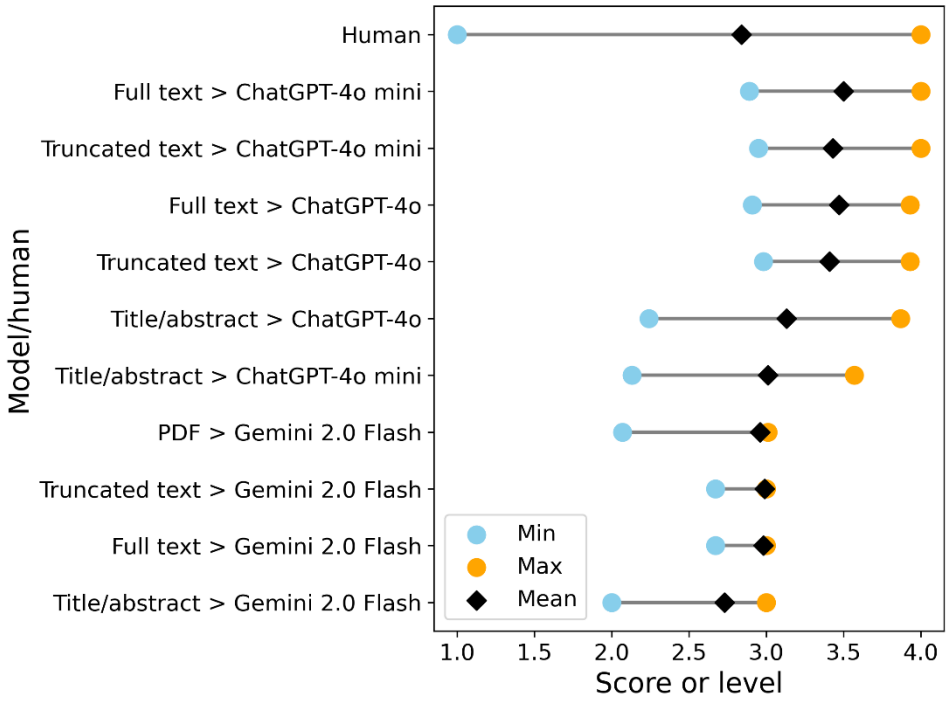

图 1. 不同大语言模型和输入的最低分、最高分和平均分，以及人类（即部门内部）名义评分等级的相应数据（n=58 篇文章）。大语言模型数据是在对 30 次迭代求平均后计算得出的。

无论使用哪种大语言模型，在所有情况下，人工评分水平较高的文章往往也会获得较高的大语言模型评分（图 2）。然而，不同水平的大语言模型评分之间存在大量重叠，没有任何一个大语言模型评分范围与单一的人工评分水平相对应。因此，无法在任何情况下仅凭大语言模型评分确定人工评分。

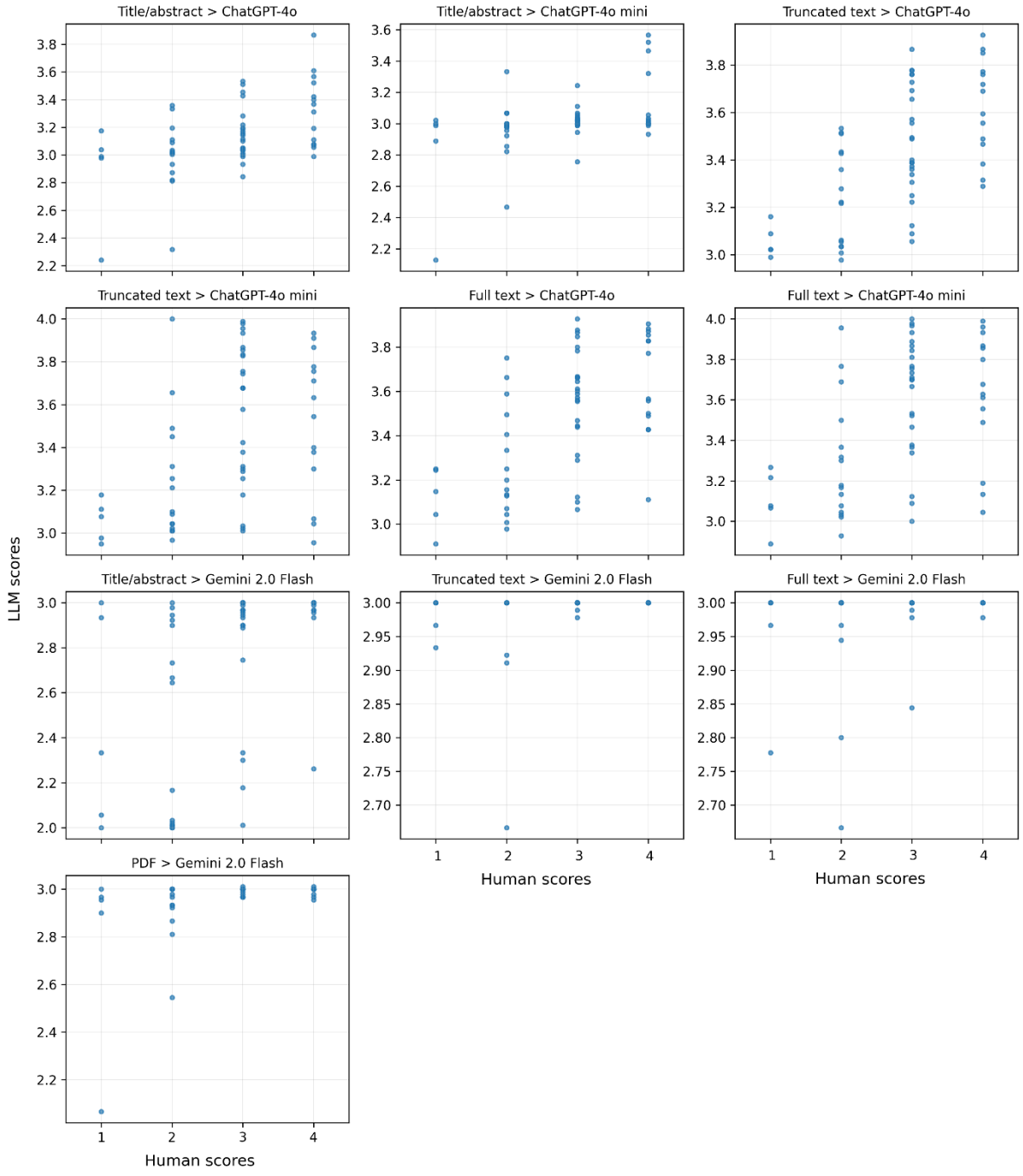

图 2：58 篇信息科学期刊文章的院系评分等级与大语言模型评分预测。

在所有情况下，58 篇信息科学文章的大语言模型评分与人工评分等级之间的斯皮尔曼相关均为正值，且除一个边缘情况外，该值在统计上均显著异于 0（图 3）。遗憾的是，95% 置信区间过宽，无法从统计学证据判断哪个大语言模型或哪种输入效果最佳；但总体而言，ChatGPT-4o 的表现最佳（相关性最高的三项），而四种输入似乎都没有持续表现为最好或最差。图 3 中，包含评估所需全部信息（文本、图表和参考文献）的唯一文档格式——PDF格式——相关性却排在倒数第三，这倾向于证实大语言模型评分只是猜测，而非判断。皮尔逊相关系数的结果也类似（图 3）。

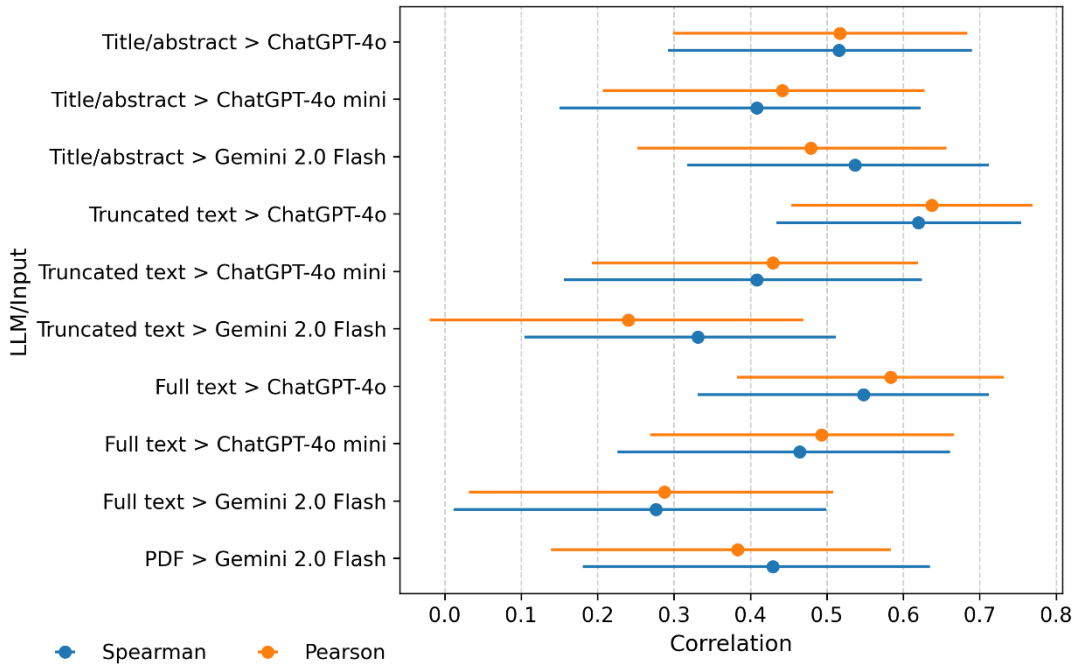

图 3. 按大语言模型及输入分类的大语言模型评分预测（对 30 次迭代取平均后）与人工评分等级之间的斯皮尔曼相关和皮尔逊相关系数（n=58 篇文章）。图中显示自助法估计的 95% 置信区间。

我们绘制了热力图，以展示大语言模型之间评分的相似程度（图4）。对于 Chat GPT，各版本之间以及不同输入之间的相关性往往高于其与人工评分的相关性；但 Gemini 的相关性普遍较低，标题/摘要输入除外。

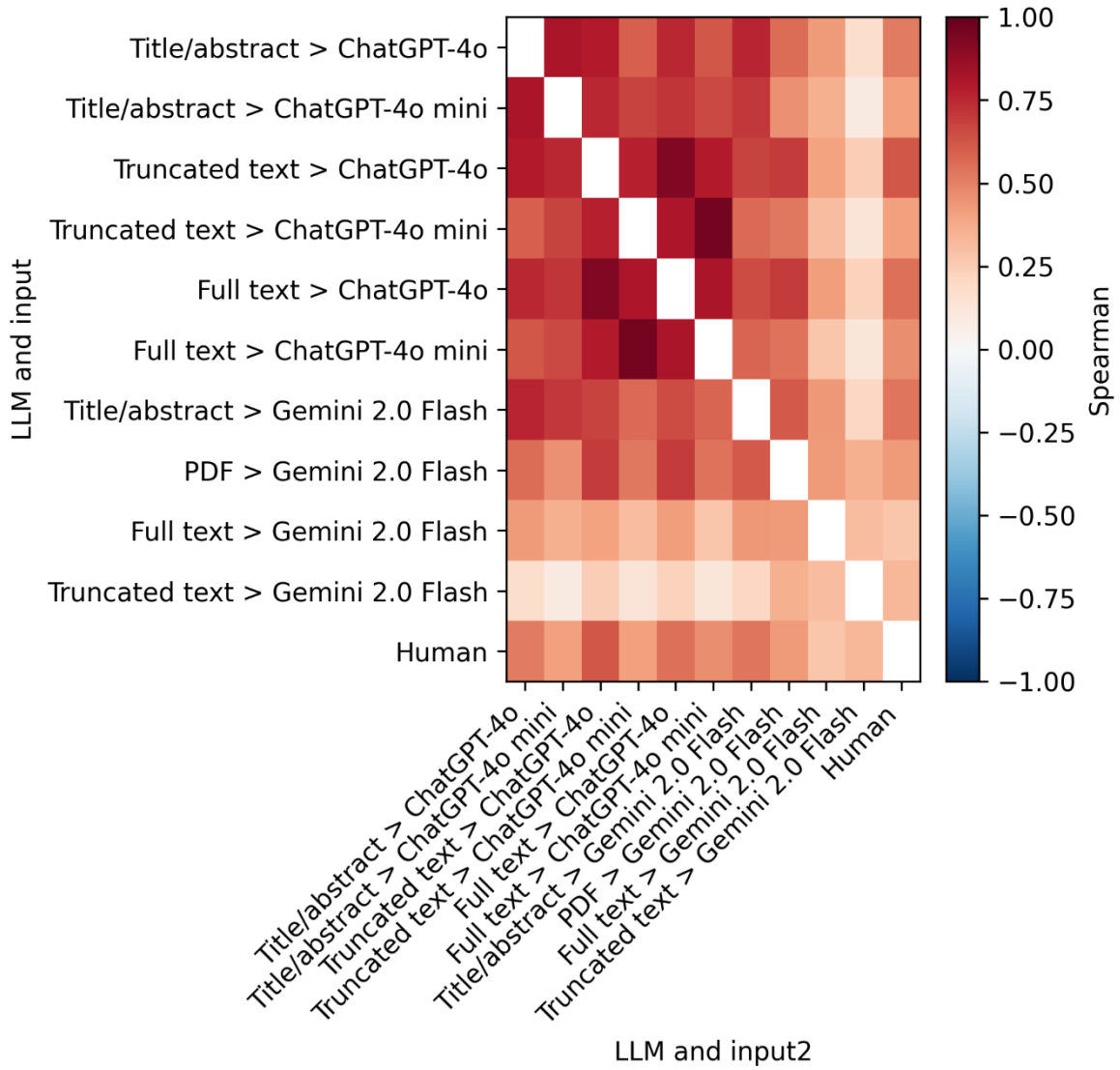

图4. 大语言模型评分预测（对30次迭代取平均）与人工评分等级之间的成对斯皮尔曼相关性热力图，按大语言模型和输入分类（n=58 篇文章）。

4.2 大语言模型评审报告中文本相似度的定量证据（研究问题***3*** ***）***

针对每篇论文和每个模型生成的 30 份评审报告，其平均文本相似度非常高。在全部 58 篇论文中，模型内相似度的变化范围都很窄。例如，ChatGPT-4o 使用全文输入时，相似度分数介于 0.92 至 0.97；使用标题/摘要输入时，介于 0.91 至 0.97；Gemini 2.0 Flash 使用全文输入时，介于 0.91 至 0.96。因此，在所有情况下，当向大语言模型反复提供相同的输出时，其报告往往非常相似，尽管从未完全相同，有时给出的评分也不尽相同。

跨模型文本相似度同样很高，尤其是在两个 ChatGPT变体之间；即使对于同一篇论文的不同输入，情况也是如此。ChatGPT-4o 使用标题/摘要输入与 ChatGPT-4o 使用全文输入之间的平均相似度介于 0.89 至 0.97 之间。当任一 ChatGPT变体与 Gemini 进行比较时，相似度分数略低：ChatGPT-Abstract 与 Gemini 之间为 0.85–0.94，ChatGPT-Full 与 Gemini 之间为 0.85–0.93。在所有模型配对中，相似度中位数均接近各自范围的中点。总体而言，这些结果表明

评审文本中的模型内差异和模型间差异都很小。较高的相似度分数表明，这些模型倾向于生成高度重复且标准化的评审报告，在措辞和内容方面的多样性有限。

对大语言模型评审报告的检查揭示了几种有趣的模式。总体而言，这些报告比人工撰写的院系评审摘要长得多。这些文本的特点是冗长：用大量文字解释一个观点，而且经常换种说法反复阐述。例如，一篇180词的摘要由ChatGPT评审后，得到了一份800词的评审报告。

4.3 大语言模型与人工报告的定性比较（研究问题***3***）

对评审报告进行定性考察，有助于解释上文所发现模式的某些方面。大语言模型生成的评审意见往往比人工评审摘要语气更为积极。虽然难以确切判断它们与人工评分存在差异的原因，但一个重要区别似乎在于模型如何使用评分量表。基于标题/摘要生成的ChatGPT-4o评审意见所依据的信息有限，因此评分往往更为严格。相比之下，基于全文的ChatGPT-4o评估给出的分数更高。如上所述，Gemini则不同，其评分高度集中在量表中点附近：几乎所有文章的评分都接近3分。Gemini的评审文本通常也更短、更少重复，语气不那么积极，在风格上与人工评审意见更为相似。

总体而言，AI生成文本显得连贯且有意义。这些模型似乎接受过大量类似评估文本的训练，尤其涉及研究质量、标准评审标准以及不同科学方法的优缺点。虽然系统提示词篇幅较长（有一页半），但其中并未包含有关科学方法的信息，因此大语言模型必须从其已有知识中提取这类信息。在评估中，大语言模型主要关注研究质量的宏观特征，例如可推广性、方法论创新、理论贡献和可重复性。因此，报告中常见的局限性也相应包括回复率较低、可推广性有限等问题。在评审报告中，常常会使用系统指令中的评价性术语（例如“世界领先”），却很少为这些术语提供依据，也不太解释它们在实践中意味着什么。

出乎意料的是，仅根据摘要进行评估时，ChatGPT似乎没有针对可用信息有限这一情况作出调整；评审报告有时会指出一些并不准确的局限性，因为论文全文中已经处理了相关问题。例如，一份评审报告指出：“不过，若能更详细地讨论样本量、参与者特征、数据编码方法和研究结果验证，将有助于增强严谨性”（案例25）。这条意见具有误导性，因为全文已对这些内容作出说明。

另一个重要区别在于，大语言模型主要强调研究质量的一般特征——例如清晰度、可推广性和方法论严谨性——而人类审稿人往往更加具体，经常指出人工智能生成的评审意见中未明确提及的具体问题、背景细节或领域特定关切。这表明，尽管大语言模型能够近似

总体评价模式，但可能缺乏人工同行评审所具有的细致程度和情境敏感性。

此外，AI生成文本更加聚焦于原创性、重要性和严谨性这三项评审标准；而人类审稿人有时也会评论他们认为相关的其他方面，例如指出论文结构清晰、文笔流畅、表达方式通俗易懂，或肯定论文丰富的参考文献或补充材料。换言之，AI模型遵循指示，按照相同标准评审所有论文；而人工评审摘要有时会在评审中纳入针对具体文章的背景信息。

除了将评论限定在正式评审标准的范围内，AI生成的评论往往还包含评审标准和评分定义中的术语，因此给人一种严格遵循指南的印象。例如，人类评审意见中没有出现“世界领先”（用于定义4*分）和“国际卓越”（用于定义3*分）这两个术语，但它们却频繁出现在AI生成文本中。尤其是，AI生成文本经常指出相关研究具有国际卓越水平，和/或尚未达到世界领先水平：

> “这项工作的原创性、重要性和严谨性有助于其为该领域作出贡献，但某些局限性使其未能达到世界领先的最高水平”（*Gemini*，案例*22*）。

> “这篇论文在原创性、重要性和严谨性方面展现出国际卓越的质量，但尚未达到卓越的最高标准”（*Gemini*，案例*54*）。

> “这篇论文展现出国际卓越的严谨性，其方法结构严谨，文献综述内容广泛”（*ChatGPT-4o* 全文，案例*24*）。

> “由于未能充分证明其对当前思想、实践或政策具有直接而深远的影响或变革潜力，这篇文章未能达到世界领先的重要性”（*ChatGPT-4o* 全文，案例*15*）。

另一个观察结果是，大语言模型经常会评论研究方法的优势与局限，包括就如何减轻局限性以及提高稳健性或普遍性提出建议。人工评审也会评论此类问题，但阐述得不那么详细，指导性也较弱。以下是一些人工智能生成文本的示例：

> “该调查数据收集在方法上严谨，并清楚地列出了确保匿名性、效度和信度的步骤；然而，调查的样本量有限，且存在潜在的无应答偏差（作者已承认这一点），这削弱了研究结果更广泛的可推广性”（*ChatGPT-4o* 全文，第*15*个案例）。

> “便利抽样策略、缺乏对人口统计学多样性的控制以及调查响应率未知等局限性，降低了方法论稳健性”（*ChatGPT-4o* 全文，案例*45*）。“数据分析方法适合研究问题，但应承认该研究存在的局限性，例如

便利抽样策略以及参与者样本中可能存在的偏差”（*Gemini 2.0 Flash* 全文，案例*45*）。[1]

“方法部分描述得极为详尽，确保了全文具有可重复性。本文结合理论推理、实证研究结果和统计方差控制，展现了卓越的科学诚信。*[...]*进一步加强研究的建议 *[...]* 尽管当前研究全面探讨了中等规模数据集 *[...],* 但若能在规模大得多的数据集上开展 *[...]* 更多分析，将有助于深入了解极端规模下的计算效率。*[*该策略*]* 在类别不平衡数据集上的效果各不相同。未来若能进一步完善将 *[...]* 事后阈值技术（例如校准曲线）的组合纳入模型开发工作流程的方式，或许能更好地说明其效用”（*ChatGPT-4o* 全文，案例*22*）。

在最后一个案例中，即使给出了最高分数，评语仍然提出了提高严谨性的建议，这表明模型并不总能使所给分数与（全部）评审意见之间保持一致。大语言模型的表达方式之所以更具说教意味，可能是因为向用户提出改进建议是它们最常见的用途，因此可以合理推测，即使用户没有要求它们提供帮助，它们也倾向于默认这样做。

考察人工评分水平较低而大语言模型评分并不低的文章时，可以发现多个案例：人工审查指出研究意义有限以及方法存在不足，而人工智能生成的评审意见也指出研究意义和方法方面的局限性，但语气更为积极，且没有给出低分。例如，在一个案例中，人工审查评论称，研究结果可能在研究开展地当地具有实用价值，但不太可能在当地之外产生显著影响（案例51）。相比之下，大语言模型提到：“尽管样本量和可推广性方面存在一些局限性，但该研究在应对一个关键问题方面展现了国际卓越水平”（Gemini）；此外，“尽管该研究具有相当大的影响力和实际影响，但其重要性仍略逊于世界领先水平”（ChatGPT-Full）。换言之，大语言模型推断该研究并非世界领先，但难以预测某项研究未被认为在研究地点之外具有实用价值。

在评估研究贡献时，相关性范围和潜在影响力是至关重要的方面，但学者对这些方面可能持有不同看法，例如，他们对研究中的不同价值，或科学与社会中的不同需求有不同的理解或侧重。因此，同行评审分数通常很难预测。但为什么大语言模型不愿给出低分呢?

部分原因或许在于，AI模型对研究主题持中立态度——也就是说，它们将所有主题都呈现为重要且有趣，而在学术领域内，人们对于哪些主题有趣且重要，通常存在共识或不同看法。另一种可能的解释是，大语言模型可能严重依赖论文中的信息，来评论其相关性或重要性范围方面的局限性；而同行评审人也会依据其学术

[1] 人工审查没有评论便利抽样，只是简单地指出样本量不错。

领域的价值观和兴趣。若要做出诸如“某项研究在其研究地点之外毫无用处”这类严厉判断，可能很难仅从论文中提取依据——除非论文中确实明确写出了这一点，而这种情况并不常见。综合来看，这意味着低分数可能难以预测，因为AI模型对研究主题持中立态度；而当学术论文从积极的角度（而非批判性评审人的角度）阐述主题及其研究结果的重要性时，AI模型预测低分数比预测高分数更加困难。

5 讨论

研究结果仅适用于一个国家、一个院系和一个领域；此外，样本量过小，无法识别大语言模型之间或不同输入长度之间具有统计显著性的差异。而且，所考察的文章既不是随机样本，也并非通过中立的筛选流程产生，而是由学者预先选定的最佳成果，符合 REF 式的提交实践。研究结果还仅涉及三个大语言模型（一个 Gemini 和两个 ChatGPT），且只采用一种提示方式，即默认模型设置。因此，研究结果无法推广到其他国家、领域和模型。采用少样本提示或微调也可能改善结果。

5.1 与先前结果的比较

本文发现的相关性强度，大体上与以下研究的结果一致：一项针对单一作者群体的信息科学论文、采用作者评分的研究（Thelwall, 2024）；以及一项针对 REF2021 期刊文章的数据集研究，该数据集涵盖更广泛的传播、文化与媒体研究、图书馆与信息科学类别，并采用代理质量评分（Thelwall & Yaghi, 2025a）。这些结果对先前的一项主张提出了质疑：对于 ChatGPT-4o，输入仅为标题/摘要时的评分比输入全文时更有参考价值（Thelwall, 2025a）；而 Gemini 的情况恰好相反（Thelwall, 2025c）。因此，当前结果以一种新型数据（院系内部评审）巩固了先前关于大语言模型能够预测该领域研究质量的证据，并表明标题/摘要与全文或 PDF输入之间的差异可能很小。对此可能有多种解释，包括以下几点：

- 大语言模型可能难以从长篇文档中找到相关信息（Liu et al., 2024），因此可能会重点关注开头（摘要）和结尾（结论——与摘要类似——或参考文献，而参考文献并不相关）。
- 大语言模型并不理解文档，可能会根据全文进行错误的模式匹配并得出结论，例如将文献综述中的批评意见作为依据。
- 摘要可能概括了论文的核心价值，而该价值已通过同行评审得到验证。

此前已有研究指出，人们不愿给出低分（Thelwall, 2024; Thelwall & Yaghi, 2025a），因此当前结果也倾向于证实这一点。上文讨论了可能的解释，另一种可能是，大语言模型会尝试给出高分，以此向用户提供正向反馈；这也符合其更常见的用途，即帮助作者改进作品。

与 Gemini 的评分相比，ChatGPT 评分的范围要宽得多，这是一个新的观察结果（例如，Thelwall, 2025c 中未曾提及）。结合单份 Gemini 报告中 3* 评分占比很高（95%）这一点来看，这表明需要进行更大程度的平均，才能有效利用 Gemini 对

一篇文章可能获得较高还是较低分数的不确定性；正是这种不确定性，才是区分文章并进行排名所必需的。

5.2 关于大语言模型报告及其与人工评审报告差异的新发现

尽管有些研究提到了大语言模型评审报告，但此前没有研究对其进行详细分析。此外，也没有研究将大语言模型报告与任何类型的发表后报告进行比较，尽管已有研究比较过大语言模型与人工发表前评论。大语言模型评语往往偏正面的发现，与有关发表前同行评审的类似观察相符（Zhou 等，2024）；然而，当前研究并未发现大语言模型对严谨性给予更多关注，而 Shin 等，2025 和 Zhu 等，2025 曾在会议论文同行评审报告中发现这一点。对文本相似性（即向量空间中的嵌入）的分析发现，大语言模型评审文本在模型内和模型间的差异都极小，这似乎是此前研究中未曾出现过的发现。

另一项新发现是，大语言模型评审似乎更注重评审标准，而人工评审则更加具体（这与同行评审中的情况相呼应：Zhou 等，2024），且更多样化。此外，大语言模型评审通过指出论文应获得某一特定等级的分数（通常以“国际卓越”来论证 3* 分），将评分与评论联系起来，并以概括的方式对此作出解释（通常认为论文因原创性、重要性和/或严谨性方面存在局限性，不足以达到“世界领先”水平／获得 4* 分）。评分与评论之间的联系仍不完全一致，因为即使是评分最高的论文，也被指出存在局限性。

一项新的观察是，仅向 ChatGPT 提供摘要时，它有时会给出有缺陷的评论（指出论文全文中并未出现的局限性），而提供全文时则不会出现这种情况。尽管如此，相关性与人工评分大致相同。因此，在审阅摘要和论文全文时，大语言模型预测人工评分的能力可能存在不同的局限性。

5.3 院系内部评审的大语言模型评分

这项分析引出了一个自然的问题：在编制 REF 申报材料时，耗时的院系内部同行评审能否借助大语言模型来完成？英国大多数院系可能每年都会开展此类评审，为下一轮 REF（REF2029）做准备。显然，大语言模型的结果还不足以取代人类专家（此前大多数大语言模型研究也都提出了类似的否定建议），但它们能否为专家提供支持?

在支持专家判断方面，这些模型无法可靠地区分真正杰出的出版物与仅仅达到合格水平的出版物。不过，ChatGPT 4o 使用截断文本作为输入时的图表（图 2）显示，评分可以通过某些方式发挥作用。在信息学院，所有评分为 4 级的文章都被暂定选入 REF，而评分为 3 级的文章中也需要选出一部分，以满足名额要求。一种选出这些文章的方法，是选取 ChatGPT 4o 评分最高的文章，因为人工评估者无法区分这些文章。实际上，学院几乎肯定会更倾向于由院系 REF 团队来选定文章，但大语言模型评分或许至少能为这一艰难决定提供一个起点。

大语言模型评分可能有助于院系评审流程的第二种方式，是指出哪些情况需要进一步人工审查。评审意见中的局限性

可能是因为信息科学具有高度的多学科性质，而学院的潜在评审人选较少。这可能包括对人工评定为 4*、但大语言模型评分较低的文章进行额外人工审查，也可以检查大语言模型评分最高的 2* 文章是否值得评为 3* 或 4*。

上述两项建议存在冲突：第一项可以节省时间，第二项则恰恰相反，但可能改善结果。要判断其中任一项是否合适，需要考虑院系对其人类审稿人和大语言模型预测的相对信任程度。

6 结论

本研究比较了 58 篇信息科学期刊文章的部门内部（人工）质量评级和评审摘要，以及由十种大语言模型配置生成的评分和评审文本。定量结果显示，无论向 ChatGPT 提供全文还是仅提供摘要，其评分与人工评分之间都存在统计显著的中度正相关；而 Gemini 的相关性较弱。不同模型之间以及不同输入文本之间的差异均无统计显著性。结合先前采用相关质量分析的研究来看，向大语言模型提供标题和摘要还是全文，其预测能力似乎并无太大差异。

文本分析有助于解读这些模式。不同模型生成的评审报告内容连贯，但措辞和内容高度标准化，多次运行之间的差异极小。模型经常复现指南中的措辞，并给出大体合理的优势与局限，但表述往往较为笼统，有时与指定的评分等级关联不强。仅依据摘要作答时，ChatGPT 有时会推断出一些实际上已在全文中讨论过的局限，表明该模型无法始终准确体现因信息缺失而产生的不确定性。与此同时，提供全文往往会提高 ChatGPT 的评分，却不会增强其评分与人工评分之间的相关性，这进一步说明，作者以自身视角呈现的额外细节可能会增加正面倾向，却未能提升与专家判断的一致性。

就支持院系内部评审而言，尽管相关性尚未达到足以表明可以通过用大语言模型替代专家评审来节省时间的程度，但预测评分仍可能有助于决策。两个可能的作用是对并列评分进行仲裁，以及识别存在异常人工评分的文章。

7 参考文献

Aczel, B., Szaszi, B., & Holcombe, A. O. (2021). 一项十亿美元的捐赠：估算研究人员花费在同行评审上的时间成本。《研究诚信与同行评审》，6(1), 1-8。Carbonell Cortés, C. Parra-Rojas, C., Pérez-Lozano, A., Arcara, F., Vargas-Sánchez, S., Fernández-Montenegro, R., & López-Verdeguer, I. (2024). 人工智能辅助预筛选生物医学研究提案：伦理考量与“laCaixa”基金会的试点案例。Data & Policy, 6, e49。Chauhan, C., & Currie, G. (2024). 生成式人工智能对科学手稿外部评审与编辑同行评审流程的影响。*American Journal of Pathology*。
https://doi.org/10.1016/j.ajpath.2024.08.002

Chubin, D. E., & Hackett, E. J. (1990). 无与伦比的科学：同行评审与美国科学政策。奥尔巴尼：纽约州立大学出版社。Du, J., Wang, Y., Zhao, W., Deng, Z., Liu, S., Lou, R., & Yin, W. (2024). 大语言模型协助自然语言处理研究人员：论文批评与（元）评审。《2024年自然语言处理实证方法会议论文集》。https://doi.org/10.18653/v1/2024.emnlp-main.292Khosla, S., Jain, S., Anupama, M. A., & Thavva, R. S. R. (2024, November). 基于文本的文档相似性多种嵌入模型的比较分析。载于国际人工智能与语音技术会议（第169—180页）。查姆：施普林格·自然瑞士。Kousha, K., & Thelwall, M. (2025). 使用ChatGPT评估学术研究的社会影响力：影响案例评估。《信息科学与技术协会期刊》，76(10), 1357-1373。https://doi.org/10.1002/asi.25021 Kousha, K., Thelwall, M., & Gadd, E. (2025). ChatGPT能否评估研究环境？来自REF2021的证据。 *arXiv*预印本 *arXiv:2512.05202*。Langfeldt, L., Nedeva, M., Sörlin, S., & Thomas, D. A. (2020). 共存的研究质量概念：研究特定语境下对优质研究理解的框架。《密涅瓦》，58(1), 115-137。Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D. Y., Yang, X., & Zou, J. (2024). 大语言模型能否为研究论文提供有用反馈？一项大规模实证分析。*NEJM AI*，AIoa2400196。https://doi.org/10.1056/AIoa2400196 Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). 迷失于中间：语言模型如何利用长上下文。《计算语言学协会汇刊》，*12*，157-173。Lu, Y., Xu, S., Zhang, Y., Kong, Y., & Schoenebeck, G. (2024). 利用大语言模型获取信息丰富的文本评估。《第25届ACM经济学与计算会议论文集》（第582—612页）。Nunkoo, R. & Thelwall, M. (2025). 使用ChatGPT评估旅游研究质量。《旅游业当前问题》。https://doi.org/10.1080/13683500.2025.2596265 Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., & Lowe, R. (2022). 训练语言模型遵循人类反馈指令。《神经信息处理系统进展》，*35*，27730-27744。Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). 语言模型是无监督多任务学习器。*OpenAI*博客。https://storage.prod.researchhub.com/uploads/papers/2020/06/01/language-models.pdf REF2021（2019）。评审小组标准与工作方法（2019/02）。https://2021.ref.ac.uk/publications-and-reports/panel-criteria-and-working-methods-201902/index.html Saad, A., Jenko, N., Ariyaratne, S., Birch, N., Iyengar, K. P., Davies, A. M., & Botchu, R. (2024). 探索ChatGPT在同行评审流程中的潜力：一项观察性研究。《糖尿病与代谢综合征：临床研究与综述》，18(2), 102946。

Shin, H., Tang, J., Lee, Y., Kim, N., Lim, H., Cho, J. Y., & Kim, J. (2025年)。《警惕盲点：用于评估大语言模型评审的焦点层级评估框架》。arXiv 预印本 arXiv:2502.17086。Sriramanan, G., Bharti, S., Sadasivan, V. S., Saha, S., Kattakinda, P., & Feizi, S. (2024年)。Llm-check：探究大语言模型中的幻觉检测。《神经信息处理系统进展》，37，34188-34216。Thelwall, M. (2024年)。ChatGPT 能否评估研究质量？《数据与信息科学杂志》，9(2)，1-21。Thelwall, M. (2025a)。使用大语言模型评估研究质量：分析 ChatGPT 在不同设置和输入下的有效性。《数据与信息科学杂志》，10(1)，7-25。https://doi.org/10.2478/jdis-2025-0011 Thelwall, M. (2025b)。在哪些领域，ChatGPT 4o 的评分比引文更能与研究质量保持一致？arXiv 预印本 arXiv:2504.04464。Thelwall, M. (2025c)。Google Gemini 在评估研究质量方面是否优于 ChatGPT？《数据与信息科学杂志》，10(2)，1–5。https://doi.org/10.2478/jdis-2025-0014 扩展版本见：https://doi.org/10.6084/m9.figshare.28089206.v1 Thelwall, M. (2025d)。较小的大语言模型能否评估研究质量？《马来西亚图书馆与信息科学杂志》，30(2)，66-81。https://doi.org/10.22452/mjlis.vol30no2.4 Thelwall, M. (2025e)。大语言模型是否了解有关期刊文章的基本事实？《文献学杂志》。Thelwall, M., Jiang, X., & Bath, P. (2025年)。使用 ChatGPT 估计已发表的医学研究质量。《信息处理与管理》，62(4)，104123。https://doi.org/10.1016/j.ipm.2025.104123 Thelwall, M., & Kurt, Z. (2025年)。使用 ChatGPT 进行研究评价：它是否存在年龄、国家、篇幅或领域偏差？《科学计量学》，130(10)，5323-5343。https://doi.org/10.1007/s11192-025-05393-0 Thelwall, M., & Yaghi, A. (2025a)。ChatGPT 能在哪些领域检测期刊文章质量？对 REF2021 结果的评估。《信息管理趋势》，13(1)，1-29。arXiv 预印本 arXiv:2409.16695。Thelwall, M., & Yaghi, A. (2025b)。评估 ChatGPT 在多个平台上预测学术同行评审结果的能力。《科学计量学》，1-23。Thelwall, M. & Yang, Y. (2025年)。ChatGPT 产生的隐式与显式研究质量评分概率。《定量科学研究》。6，1271–1293。https://doi.org/10.1162/QSS.a.393 Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017年)。《注意力机制即所需的一切》。《神经信息处理系统进展》，30。Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., & Johnson, B. (2015年)。《计量学潮汐》。关于计量指标在研究评估与管理中作用的独立审查报告。https://sussex.figshare.com/articles/report/The_metric_tide_report_of_the_independent_review_of_the_role_of_metrics_in_research_assessment_and_managem ent/23418680 Wu, W., Zhang, Y., Haunschild, R., & Bornmann, L. (2025年)。利用大语言模型进行发表后同行评审：潜力与局限性。载于：编辑：

Shushanik Sargsyan, Wolfgang Glänzel, Giovanni Abramo：第20届国际科学计量学与信息计量学会议，2025年6月23日至27日，埃里温，亚美尼亚，（第1176-1195页）。Zhu, C., Xiong, J., Ma, R., Lu, Z., Liu, Y., & Li, L. (2025年)。当你的评审者是大语言模型：同行评审中的偏差、分歧与提示注入风险。*arXiv* 预印本 *arXiv:2509.09912*。Zhou, R., Chen, L., & Yu, K. (2024年)。大语言模型是可靠的评审者吗？对大语言模型执行自动论文评审任务的全面评估。载于《*2024*年计算语言学、语言资源与评估联合国际会议论文集（*LREC-COLING 2024*）》（第9340-9351页）。

附录

表 A1：研究案例概览，包括人工评分、人工智能评分和人工智能评分者间一致性。按人工评分和人工智能评分者间一致性排序。定性分析样本以粗体显示。

人工评分者	GPT 全文		GPT 摘要		Gemini 全文 文章		人工与人工智能平均分差			人工智能评分者间一致性
分数	分数	Std. 偏差	分数	Std. 偏差	分数	Std. 偏差	H GPT-full	H GTP-abst	H Gemini	克里彭多夫 α
1	**3.03**	**0.18**	**2.93**	**0.25**	**2.90**	**0.31**	**-2.03**	**-1.93**	**-1.90**	**0.02**
1	**3.10**	**0.31**	**2.97**	**0.18**	**2.97**	**0.18**	**-2.10**	**-1.97**	**-1.97**	**0.04**
1	**3.17**	**0.38**	**3.00**	**0.00**	**2.97**	**0.18**	**-2.17**	**-2.00**	**-1.97**	**0.09**
2	**3.00**	**0.26**	**2.90**	**0.31**	**2.93**	**0.25**	**-1.00**	**-0.90**	**-0.93**	**-0.01**
2	**3.03**	**0.18**	**2.87**	**0.35**	**2.93**	**0.25**	**-1.03**	**-0.87**	**-0.93**	**0.03**
2	3.17	0.38	3.17	0.38	3.00	0.00	-1.17	-1.17	-1.00	0.03
2	3.00	0.00	2.87	0.35	2.80	0.41	-1.00	-0.87	-0.80	0.04
2	3.27	0.45	3.03	0.49	3.00	0.00	-1.27	-1.03	-1.00	0.06
2	3.10	0.31	2.97	0.18	2.87	0.35	-1.10	-0.97	-0.87	0.07
2	3.13	0.35	2.97	0.18	2.93	0.25	-1.13	-0.97	-0.93	0.07
2	3.03	0.18	2.80	0.41	2.97	0.18	-1.03	-0.80	-0.97	0.08
2	3.17	0.38	3.00	0.00	3.00	0.00	-1.17	-1.00	-1.00	0.09
2	3.20	0.41	3.00	0.00	3.00	0.00	-1.20	-1.00	-1.00	0.11
2	3.43	0.50	3.07	0.25	2.97	0.18	-1.43	-1.07	-0.97	0.24
2	3.03	0.18	2.43	0.50	2.53	0.51	-1.03	-0.43	-0.53	0.26
2	3.43	0.50	3.00	0.00	3.00	0.00	-1.43	-1.00	-1.00	0.32
2	3.70	0.47	3.37	0.49	3.00	0.00	-1.70	-1.37	-1.00	0.33
2	3.77	0.43	3.13	0.57	3.00	0.00	-1.77	-1.13	-1.00	0.41
2	**3.67**	**0.48**	**3.03**	**0.18**	**3.00**	**0.00**	**-1.67**	**-1.03**	**-1.00**	**0.51**
2	**3.00**	**0.00**	**2.30**	**0.47**	**2.03**	**0.18**	**-1.00**	**-0.30**	**-0.03**	**0.66**
3	**3.07**	**0.25**	**3.03**	**0.18**	**2.97**	**0.18**	**-0.07**	**-0.03**	**0.03**	**0.01**
3	**3.10**	**0.31**	**3.03**	**0.18**	**2.97**	**0.18**	**-0.10**	**-0.03**	**0.03**	**0.02**
3	3.03	0.18	2.83	0.38	3.00	0.00	-0.03	0.17	0.00	0.09
3	3.27	0.45	3.03	0.18	3.00	0.00	-0.27	-0.03	0.00	0.13
3	3.23	0.43	3.00	0.00	3.00	0.00	-0.23	0.00	0.00	0.14
3	3.70	0.47	3.50	0.51	3.00	0.00	-0.70	-0.50	0.00	0.34
3	3.47	0.51	3.00	0.00	2.97	0.18	-0.47	0.00	0.03	0.34
3	3.47	0.51	3.00	0.00	2.97	0.18	-0.47	0.00	0.03	0.34
3	3.63	0.49	3.17	0.38	3.00	0.00	-0.63	-0.17	0.00	0.35
3	3.63	0.49	3.13	0.35	3.00	0.00	-0.63	-0.13	0.00	0.37
3	3.60	0.50	2.97	0.41	3.00	0.00	-0.60	0.03	0.00	0.38
3	3.73	0.45	3.23	0.43	3.00	0.00	-0.73	-0.23	0.00	0.41
3	3.57	0.50	2.97	0.18	2.97	0.18	-0.57	0.03	0.03	0.43
3	3.57	0.50	3.00	0.00	3.00	0.00	-0.57	0.00	0.00	0.45
3	3.63	0.49	3.00	0.26	3.00	0.00	-0.63	0.00	0.00	0.46
3	3.63	0.49	3.03	0.18	3.00	0.00	-0.63	-0.03	0.00	0.47
3	3.73	0.45	3.13	0.35	2.97	0.18	-0.73	-0.13	0.03	0.48
3	3.90	0.31	3.43	0.50	3.00	0.00	-0.90	-0.43	0.00	0.53
3	3.73	0.45	3.07	0.25	3.00	0.00	-0.73	-0.07	0.00	0.55
3	3.93	0.25	3.43	0.50	3.00	0.00	-0.93	-0.43	0.00	0.57
3	3.97	0.18	3.50	0.51	3.00	0.00	-0.97	-0.50	0.00	0.61
3	3.87	0.35	3.17	0.38	3.00	0.00	-0.87	-0.17	0.00	0.61
3	**3.90**	**0.31**	**3.17**	**0.38**	**3.00**	**0.00**	**-0.90**	**-0.17**	**0.00**	**0.66**
3	**3.90**	**0.31**	**3.13**	**0.35**	**3.00**	**0.00**	**-0.90**	**-0.13**	**0.00**	**0.69**
4	**3.17**	**0.38**	**3.10**	**0.31**	**2.97**	**0.18**	**0.83**	**0.90**	**1.03**	**0.04**
4	**3.47**	**0.51**	**3.27**	**0.45**	**2.97**	**0.18**	**0.53**	**0.73**	**1.03**	**0.18**
4	3.50	0.51	3.20	0.41	3.00	0.00	0.50	0.80	1.00	0.21
4	3.47	0.51	3.00	0.26	3.00	0.00	0.53	1.00	1.00	0.30
4	3.57	0.50	3.13	0.35	3.00	0.00	0.43	0.87	1.00	0.30
4	3.60	0.50	3.17	0.38	3.00	0.00	0.40	0.83	1.00	0.31
4	3.50	0.51	3.03	0.18	3.00	0.00	0.50	0.97	1.00	0.33
4	3.83	0.38	3.47	0.51	3.00	0.00	0.17	0.53	1.00	0.46
4	3.90	0.31	3.53	0.51	3.00	0.00	0.10	0.47	1.00	0.53
4	3.90	0.31	3.30	0.47	3.00	0.00	0.10	0.70	1.00	0.57
4	3.93	0.25	3.40	0.50	3.00	0.00	0.07	0.60	1.00	0.58
4	4.00	0.00	3.60	0.50	3.00	0.00	0.00	0.40	1.00	0.67
4	**3.93**	**0.25**	**3.80**	**0.48**	**3.00**	**0.00**	**0.07**	**0.20**	**1.00**	**0.68**
4	**3.93**	**0.25**	**3.03**	**0.18**	**3.00**	**0.00**	**0.07**	**0.97**	**1.00**	**0.85**