大语言模型能否评估资助申请书质量？重新审视 Wennerås 和 Wold 同行评审数据

Ulf Sandström
产业经济与管理，瑞典皇家理工学院，瑞典斯德哥尔摩。
https://orcid.org/0000-0003-1292-8239 ulf.sandstrom@indek.kth.se, forskningspolitik@gmail.com

Mike Thelwall
信息、新闻与传播学院，谢菲尔德大学，英国。
https://orcid.org/0000-0001-6065-205X m.a.thelwall@sheffield.ac.uk

目的：尽管同行评审对科研资助决策至关重要，学者往往不愿参与评审。这可能导致从提交申请到最终决策之间出现长时间延误，也可能使忙碌或非相关领域的学者给出不合格的评审意见。如今，至少有一家资助机构使用大语言模型（LLM）来减轻评审负担，但仍需评估大语言模型为资助申请书评分的准确性。

设计/方法/途径：本文将多种中等规模开放权重大语言模型的评分，与一个经过充分研究的数据集中的同行评审评分进行比较。该数据集为瑞典医学研究委员会 1994 年的博士后奖学金申请。

研究发现：大语言模型评分之间呈中度相关（平均斯皮尔曼相关：0.34）；与专家平均评分相比，它们的相关性较弱但为正，且大多具有统计显著性（平均斯皮尔曼相关：0.22）。专家评分与大语言模型之间最高的秩相关为 0.33，来自 Gemma 3 27b 对提案标题和摘要（不含正文）的评分；这一数值约为评审人之间相关性的 56%。

研究局限性：样本量较小、资助项目年代久远，以及评估标准各不相同，这些因素都削弱了分析的稳健性。

实践意义：尽管至少在这一特定案例中，大语言模型为资助申请书评分的量化表现弱于专家，但它们或可用于申请初筛或平局裁决。

原创性/价值：这是对大语言模型评分在资助申请书中的价值进行的首次评估。**关键词**：资助项目同行评审；大语言模型；开放权重大语言模型；《科学计量学》；研究评估

1 引言

竞争性科研资助是大多数科研体系的核心组成部分。这一过程旨在将资金投向能够得到有效利用的领域，通常用于支持资助方目标宽泛或明确的项目。决策过程通常涉及同行评审：资助方一般会遴选具备相关专业知识的学者，邀请他们提交评估意见并打分，有时还需给出多个分数。随后，评审组可能会根据这些报告和自身判断进行审议，作出最终决定。决策过程可能耗时较长，因为评审人往往不愿承担评审工作（例如，英国研究与创新署收到的评审邀请中，10 个有 9 个遭到拒绝：个人通信），而且需要给他们留出时间，以便在可能十分繁忙的日程中完成评审。主要问题在于：（a）该体系会占用专家评审人的时间，而这些时间往往

被浪费在最终遭拒的提案上，且有经验的评审人从中能学到的有限；（b）延误可能会降低研究的价值，并使紧急或时机性强的想法难以获得资助。

资助机构已经使用人工智能（AI）来协助选择评审人（Newman- Griffis 等，2025），以提高效率；如今，其中一家机构还将其用于评审。La Caixa 使用大语言模型（LLM）对提案进行初筛，将得分最低的提案交由人类专家进行快速评审。之后，所有三方（大语言模型和两名评审人）均判定为质量较弱的提案将不经完整评审（三次详细的专家评审）直接被拒绝。这使所需的评审数量略有减少（Carbonell Cortés 等，2024）。

在上述背景下，鉴于大语言模型已展现出较强的能力，能够估计已发表学术期刊文章（Thelwall，2025；Thelwall 与 Yang，2025）以及其他学术文件的研究质量，例如研究环境陈述（Kousha 等，2025）、会议论文（Thelwall 和 Yaghi，2025a）和书籍（Thelwall 和 Cox，2025），因此有必要探究它们能否评估资助申请的质量。尽管此前已有一些研究使用人工智能评估资助申请（例如 Sikimić 和 Radovanović，2022），但似乎没有任何研究取得足够理想的结果，因而被资助机构采用，所以 La Caixa 似乎是先驱。对研究人员而言，有两个实际问题：申请材料通常属于机密，不像开放获取的期刊文章那样可以公开查阅，因此需要采取额外步骤和保障措施才能获取；此外，资助计划和同行评审流程的类型很多。本研究通过聚焦于一类资助提案——1994年瑞典医学研究委员会博士后奖学金申请——（部分）解决了这些问题；由于瑞典的官方文件公开原则，这些材料可以获取。[1]

本研究评估大语言模型对资助申请书进行评分的能力，考察其评分是否与专家评分一致。出于实际考虑，研究聚焦于1994年的瑞典医学研究奖学金申请，因为这些材料可以获取，而针对这一案例获得的证据也将有助于推测大语言模型在其他案例中的表现。研究还聚焦于中等规模的“开放权重”大语言模型：这类模型可以下载并在本地运行；鉴于信息的敏感性，这似乎很可能是资助方评估资助申请书时的必要条件。先前研究表明，在学术评估方面，这些大语言模型的表现几乎与 ChatGPT 等更大型的云端模型一样好（Thelwall 和 Mohammadi，2026年）。以下研究问题均围绕1994年的瑞典医学研究奖学金申请展开，并构成本研究的核心。

1. 大语言模型评分是否与专家评审评分呈正相关？这似乎是大语言模型能够在提案评估中发挥任何作用的必要条件。2. 不同大语言模型给出的评分是否彼此相关？这将揭示它们试图共同衡量的“大语言模型质量”在多大程度上是共通的。3. 大语言模型与专家评审人的一致程度，与不同专家评审人之间的一致程度相比如何？这对于理解 RQ1 的答案至关重要。

[1] 这些申请材料由 Agnes Wold 提供；根据瑞典官方文件公开原则，提交给 MRC 的资助申请书可以公开获取。该原则赋予公众一项宪法权利，即查阅公共机关持有的官方记录（《出版自由法》，第2章）。

2 背景

本研究的实证依据是1994年提交给瑞典医学研究委员会（MRC）的博士后奖学金申请。当时，这些研究员奖学金是瑞典生物医学领域最具声望、竞争最激烈的早期职业资助机会之一。它们面向刚获得博士学位的人，旨在支持其向独立研究过渡，通常包括在新的研究环境中接受一段时间的培训。申请材料由学科评审小组通过结构化的同行评审流程进行评估，评审综合考量多个标准，通常包括申请人以往的成就、拟议项目的科学质量与可行性，以及接收单位环境是否合适。评审人给出数值评分，随后由评审小组汇总，并将其作为资助决策的依据。本研究的一项关键条件是，由于瑞典长期奉行官方记录公开查阅原则，申请材料及相关评审文件（包括评分）均可获取。

这些申请材料后来构成了 Wennerås 和 Wold（1997）颇具影响力的研究基础。该研究考察了瑞典医学研究委员会同行评审中的性别偏见。研究分析揭示，评审评分存在系统性差异，而仅凭传统的生产力指标无法解释这些差异；这表明，女性申请人必须比男性申请人具有显著更高的生产力，才能获得相当的评分。除其实质性发现外，Wennerås 和 Wold 的论文之所以在方法论上具有重要意义，还在于它将量化的同行评审评分与独立的文献计量指标所衡量的申请人优劣联系起来。因此，该研究已成为研究评估文献中被引用和讨论最频繁的研究之一（Sandström & Riis，2025年）。

随后对 Wennerås 和 Wold 相关材料的重新分析与讨论，进一步关注了 MRC 评审过程的其他特征，包括评估标准的多维性质，以及评审小组在形成和调节单个评审人判断方面所发挥的作用。总体而言，这一系列研究已将 1994年 MRC 博士后 fellowship 竞赛确立为考察资助项目同行评审运作方式、优势与局限性的基准案例。尤其是，该案例难得地揭示了在现实资助条件下，数值评分如何从专家判断中产生，而不是在实验或模拟环境中形成。

在此背景下，本研究重新利用同一批历史材料，探讨一个不同但互为补充的问题：将当代大语言模型应用于真实资助申请书时，它们能否近似专家同行评审评分。这个案例的价值不在于其时效性，而在于我们对评审过程及其结果拥有异常丰富的背景知识。因此，我们可以不孤立地解读大语言模型与评审人之间的一致性，而是将其置于已知的评审人之间一致性水平，以及此前记录的专家评分变异来源和偏见之中加以考察。

2.1 资助项目同行评审的评审人间信度

研究评估中一个长期存在的问题是，不同评审人在评估同一份资助申请时，其意见一致到何种程度。在资助项目同行评审中，无论学科、资助计划还是国家如何，评审人间信度通常都不高。早期的定量研究已表明，评审人之间的一致性往往较低，独立评分之间的相关性通常从较弱到中等不等（例如，Cole 等人，1981年；Cicchetti，1991年）。此后的大规模分析也在生物医学、社会科学和跨学科资助等领域证实了这一模式。

较新的证据表明，评审者间一致性较低，主要并非源于评审人能力不足，而是反映了资助评审本身具有多维度、依赖判断的性质。即使遵循共同准则，评审人对原创性、可行性、研究者履历和潜在影响等标准的权衡也各不相同（Langfeldt，2006年；Fogelholm 等人，2012年）。因此，评审人之间的差异往往是系统性的，而非随机的，因为不同评审人会着重考量提案质量的不同方面。

对主要资助机构开展的实证研究一致发现，资助评分的评审人间相关性远低于期刊同行评审或文献计量指标中通常观察到的水平。例如，对美国国立卫生研究院、美国国家科学基金会和欧洲资助计划的分析显示，评审人之间的相关性通常介于0.2至0.4之间，具体取决于学科领域和研究设计（Bornmann 等，2010；Li 和 Agha，2015）。评审小组讨论可以降低部分差异，但很少能将其完全消除；最终的资助决策往往取决于协商形成的判断，而非单纯汇总评分。

这方面的文献对比较同行评审与其他评估方法的研究具有重要启示。评审人间信度较低，这为任何外部系统——包括文献计量指标或大语言模型——所能达到的合理一致程度设定了上限。因此，解读大语言模型生成的评分与专家评审评分之间的弱至中等相关性时，应将专家之间本身存在的基础性分歧纳入考量，而不应将其视为系统存在根本缺陷的证据。

2.2 用于研究评估的大语言模型

许多研究探讨了大语言模型能否评估已发表期刊论文的研究质量，或能否在计算机会议同行评审过程中为作者提供有用的反馈建议。这两项任务与本文相关，但不同于资助申请评分。

多种云端大语言模型和开放权重大语言模型能够评估已发表期刊论文的研究质量，具体而言，其评分与专家评审评分呈正相关，且具有统计显著性（Thelwall，2025年；Thelwall 和 Mohammadi，2026年；Thelwall 与 Yang，2025年）。这种相关性的强度因领域而异，从弱到中等不等；在健康科学、生命科学和物理科学领域，相关性更强（Thelwall 和 Yaghi，2025b）。尽管输入给大语言模型的只有论文标题和摘要，而这些信息不足以评估一篇论文，但仍会出现这种相关性。因此，尽管大语言模型评分与专家评分呈正相关，它们也只是猜测。尚不清楚这种一致性从何而来，但大语言模型可能会识别出某个领域的专家评审人通常认为高于或低于平均水平的话题或方法选择（例如，Nunkoo 和 Thelwall，2025年）。

除了估算发表后质量评分，大语言模型还可以在同行评审流程中根据未发表稿件的全文提出建议（Liang et al., 2024），并汇总评审意见以支持编辑或作者（Du et al., 2024）。目前，只有少数计算机科学会议和一家出版商支持正式使用大语言模型进行同行评审，但即使明确告知评审人不得使用，大语言模型辅助仍常被他们采用（Naddaf, 2026），这也推动了检测由大语言模型撰写的评审意见的方法的发展（Liu et al., 2025; Ryu et al., 2025）。

国际学习表征会议（ICLR）会在评审人完成初次评审后，为他们提供大语言模型的建议。在接受这些

反馈的人中，27% 随后修改了评审意见（Thakkar 等，2025），这表明这些建议在此情境下似乎很有用。对于不确定该任务中哪些做法合适的新评审人而言，大语言模型的建议或许尤其有价值（Yun 等，2026）。较小的开放权重大语言模型似乎已足以胜任这项任务（Sadallah 等，2025），与使用云端版本相比，能提供更高的安全性。

基于智能体的大语言模型评审支持或许是未来的发展方向，例如安排不同的智能体分别负责撰写、整理和筛选评审意见（Bharti 等，2026a；Farooq 和 Iqbal，2026；Goyal 等，2026）。完全禁止在同行评审中使用大语言模型似乎行不通，因此，确保有效使用大语言模型似乎是更合理的做法（例如 Parra 等，2026），尤其是在同行评审负担日益加重的背景下。

3 方法

本研究的设计是将 1994 年的瑞典医学研究奖学金申请转换为英文纯文本文件，然后提交给一系列大语言模型进行评分，再将这些评分与专家评分进行比较。由于专家同行评审评分由多个依据不同项目任务要求开展工作的评审小组给出，因此在汇总和解读这些评分时需要格外谨慎；下文将明确讨论这一问题。

3.1 数据

3.1.1 申请材料

这些瑞典医学研究奖学金申请是按照标准模板填写的文字处理文档打印件。各部分包括申请人的个人信息、摘要以及对拟议工作计划的描述。其中一份提案只有摘要，是一份“虚设”提案，因为该申请在上一资助轮次中已作为一份双重提案获批。该提案已被剔除。所有提案均以英语或瑞典语撰写。为便于本地大语言模型处理，这些提案按以下方式转换成英文纯文本文件。

142份提案被扫描为PDF文件，随后提取了摘要和提案正文部分。这两个部分仍为PDF格式，之后使用文本识别软件Tesseract（Smith，2007）转换为纯文本文件，并设置为同时识别瑞典语和英语。该软件部分针对质量较差的扫描文档以及非英语的欧洲语言而设计，因此似乎适用于这项任务。总体而言，这一处理过程效果良好，但也出现了一些错误，尤其是化学式方面的错误较多。Tesseract还尝试识别图像中的文字，例如图表坐标轴图例。所有提案均由母语人士逐一校正了这些异常（瑞典语由Ulf校正，英语由Mike校正）。

瑞典语提案和摘要由大语言模型 Gemma 3 27b 翻译成英语（见下文），并要求其严格遵循原提案的结构。如果没有这项指示，它可能会重新组织提案，使其更连贯或更符合英语表达习惯。这种重组可能会因误解而引入错误，也可能改善申请人的表达能力，因此在此情境下并不可取。翻译是大语言模型的一项核心任务，而 Gemma 3 27b 似乎出色地完成了这项任务。不过，所有译文都由一位瑞典语双语母语人士（Ulf）检查，以确保翻译准确。所使用的提示词如下。

> 将以下瑞典语学术文本翻译成正式的英语学术文本。

> 重要规则：- 将换行符和空行视为不可更改的格式标记。- 不要合并、拆分、重排或删除段落。- 不要添加任何评论。- 仅输出翻译后的文本。文本：

这一流程最终得到142份英文提案标题和摘要文件，以及142份标题、摘要和提案文件。尽管评估完整提案比仅评估其摘要更为合理，但已有研究表明，对于学术文本，大语言模型处理摘要的效果与处理全文相当，因此本研究采用了这两种方法。在两种情况下，均只使用提案文本，并对个人信息进行脱敏处理。由于离线大语言模型尚无法处理文档中的图表，因此排除了图表。与出版物相比，这一遗漏的影响可能较小，因为图表通常是对文本的可视化说明，而非研究结果。此外，申请人的个人信息（例如封面上的电话号码和地址）也未进行处理。

3.1.2 同行评审得分

所有提案均由独立评审人进行评估和评分，以下各项的评分范围均为0至4分：

- 研究问题
- 方法学
- 申请人能力

评审组的审议似乎也将主办机构的研究环境纳入了考量（Sandström, & Riis, 2025），但这并非明确的评审标准，因此未纳入分析。尽管所有申请表的格式都相同，申请人可以将申请材料提交给一个或多个学科评审小组。因此，在一定程度上，提案文本针对不同的目标，有时甚至针对多个目标。尽管如此，从提出相关且具有创新性的医学研究计划这一广泛目标来看，它们的目的相同。因此，将这些提案作为一个整体进行分析，并分别在各评审组内进行分析，都是合理的；但由于样本量较小，后者无法实现。

除评审组之间的差异外，每份申请还可在最多三个不同的资助计划中接受评估——一般项目（n=119）、优先领域（n=19）和流动项目（n=24）。在每个项目中，申请材料由最多两个专家评审小组独立评审（在数据中分别标记为 PK1 和 PK2），每个小组都会针对三个维度提供数值评分：研究问题、方法学和申请人能力。因此，同一位申请人可能会在同一维度上获得多个评分，这些评分来自不同的评审小组和项目背景。如果提案被认为具有跨学科性质，且与两个评审组都相关，也可能会被送交两个评审组。

由于评审小组评估嵌套在项目之中，且不同项目的任务和评估标准各不相同，因此没有唯一正确的方法可以将这些评分汇总为单一的申请人层面指标。我们将每项评审组评分视为一项独立的专家判断，并保留所有原始评分。随后，我们分两个明确阶段汇总评分。第一阶段，将各项目内的评审组评分合并（在 PK1 和 PK2 均有数据时取其均值）。我们对所有评分进行了整体分析，也按项目分别进行了分析。

理想情况下，我们本应采用不同的汇总规则重复分析，以避免在异质的评估情境中依赖单一构造分数，但

样本量不足以使结果在统计学上具有可区分性，因此我们只报告最简单可行的方法。遗憾的是，对不同项目的分数取平均，实际上预设了不同任务要求下产生的评估具有可比性，而这未必成立，也是本研究的一项局限。

3.2 大语言模型处理

3.2.1 提示词

生成提示词以说明任务。理想的提示词应当仿照提供给人工评审者的说明。在本例中，我们并不知道这些说明，因此改用量表描述。由于申请的时效性可能至关重要，提示词中反复加入提交年份1994年，以降低提案因过时而被降级的可能性。在使用英国资助申请进行实验、完善措辞并确保提示词有效后，我们采用了以下文本。

> 分析以下于1994年提交给瑞典医学研究委员会的研究提案的优势。评估时请考虑提交年份为1994年。请在报告末尾按照以下量表给出0-4分的评分，必要时可使用小数：研究问题：___其中0 = 不正确；1 = 不清楚：研究问题不符合当前研究前沿。该研究已有人开展过，或正在以不同方向进行；2 = 清楚且有趣；3 = 清楚且非常有趣；4 = 清楚，且在1994年具有原创性。方法学：___其中 0 = 不正确； 1 = 材料或方法不完全恰当； 2 = 材料恰当，方法常规但适用于该研究目的；3 = 材料恰当，方法恰当且部分具有创新性；4 = 材料具有原创性或难以获取，和/或在1994年具有技术原创性，且由申请人开发。申请人能力：___其中0 = 缺失；1 = 存疑；2 = 良好；3 = 非常好；4 = 在1994年表现极为出色。

3.2.2 大语言模型

选取了一系列现代中型大语言模型进行评估。科研资助机构似乎不太可能拥有运行大型模型（如完整版本的 DeepSeek）所需的资源，但中型模型既可以在专用设备上快速运行，也可以在标准个人计算机上缓慢运行。因此，它们似乎处于典型资助机构实际可行能力的边界。大语言模型种类繁多，包括较早的模型以及学术机构基于较新模型定制的版本，但为获得最具一般性的结果，仅选取了近期广为人知的大语言模型。推理模型和非推理模型均被纳入，以便比较它们的能力。因此选取了以下大语言模型；在实验进行时，这些模型均为各自系列的最新版本。模型名称中“b”前的数字表示模型包含的参数数量（以十亿计），是衡量模型规模大小的近似指标。

- Gemma 3 27b：来自 Google 的非推理模型。
- Magistral Small（24b）。Mistral AI 推理模型。
- Llama Scout（17b）。Meta 的推理模型。

- Qwen3（32b）。阿里巴巴的推理模型。
- DeepSeek-R1（32b）。来自 DeepSeek 的推理模型。
- DeepSeek-R1（72b）。来自 DeepSeek 的大型推理模型。作为额外检验，用于考察更大的模型是否可能表现更好。

3.2.3 请求处理

每份提案都提交给每个大语言模型五次，并使用该大语言模型的平均分。先前研究表明，对多次迭代的结果取平均值能提高其与人类判断之间的相关性（例如，Thelwall 与 Yang，2025）；这可能是因为大语言模型在运用部分随机机制理解任务并生成报告时，借此调动了更多内部知识。

报告遵循了要求的结构，并编写了一个简单的 Python 程序，用于提取四项评分，并计算每项评分在五次迭代中的提案平均分。最终，每份文档和每个大语言模型都得到一组四项评分（每份提案各两项：一项针对标题/摘要，另一项针对完整提案）。

3.3 分析

从信息效用的角度来看，大语言模型评分最重要的方面是其排名，而非具体数值。这是因为大语言模型通常会给任务打高分，而不是严格遵循量表中各分值的描述。尽管这些评分基本上大多不准确，排名往往仍然有用（Thelwall 和 Mohammadi，2026 年）。因此，评估大语言模型评分效用的主要检验标准，是其对提案的排名与专家排名的一致程度。这也符合资助工作的核心任务：实质上是选出排名靠前的提案予以资助，直到资金用完为止。通常，委员会会对决策做一些小幅调整，而不会严格按照排名执行。但瑞典医学研究委员会并非如此。每位评审者最初给某项提案打出的分数在整个过程中都没有更改。因此，研究采用斯皮尔曼秩相关来评估大语言模型预测的实用性，并通过自助法计算95% 置信区间，以判断统计学显著性。

另一个需要采用不同统计量的目标，是评估任务评分中潜在一致性的程度。对此有许多不同的衡量方法。要评估各个评审人与各个大语言模型之间的一致性，最合适的指标是组内相关系数（ICC）。具体而言，当两组评审评分来自多位评审人时，ICC（1,1）适用。它类似于皮尔逊相关，但不假设评审评分具有相同的均值和方差（Koo 与 Li，2016；Shrout 与 Fleiss，1979）。然而，这一指标在此处并不适用，因为它要求评审人使用相同的评分范围，而大语言模型并非如此——它们往往倾向于给出一个较高的分数。因此，唯一有用的衡量方式是考察排名顺序，此时适合采用斯皮尔曼相关。

4 结果

大语言模型对所有维度及总体的平均评审评分，均呈现较弱但通常为正且具有统计显著性的相关性；无论输入的是标题和摘要，还是标题、摘要和正文，结果均如此（图 2–4）。尽管样本量过小，无法进行稳健比较，但大语言模型的预测在流动性评审组似乎最差。在这些模型中，Gemma 3 27b 的相关性最强，不过模型之间的差异大多

不具有统计显著性。无论总体而言，标题和摘要还是全文都没有明显优势：向任何大语言模型输入哪种文本，似乎都没有影响。此外，对不同模型的结果取平均似乎能提高相关性，不过提升幅度很小，而且在任何情况下都不具有统计显著性。

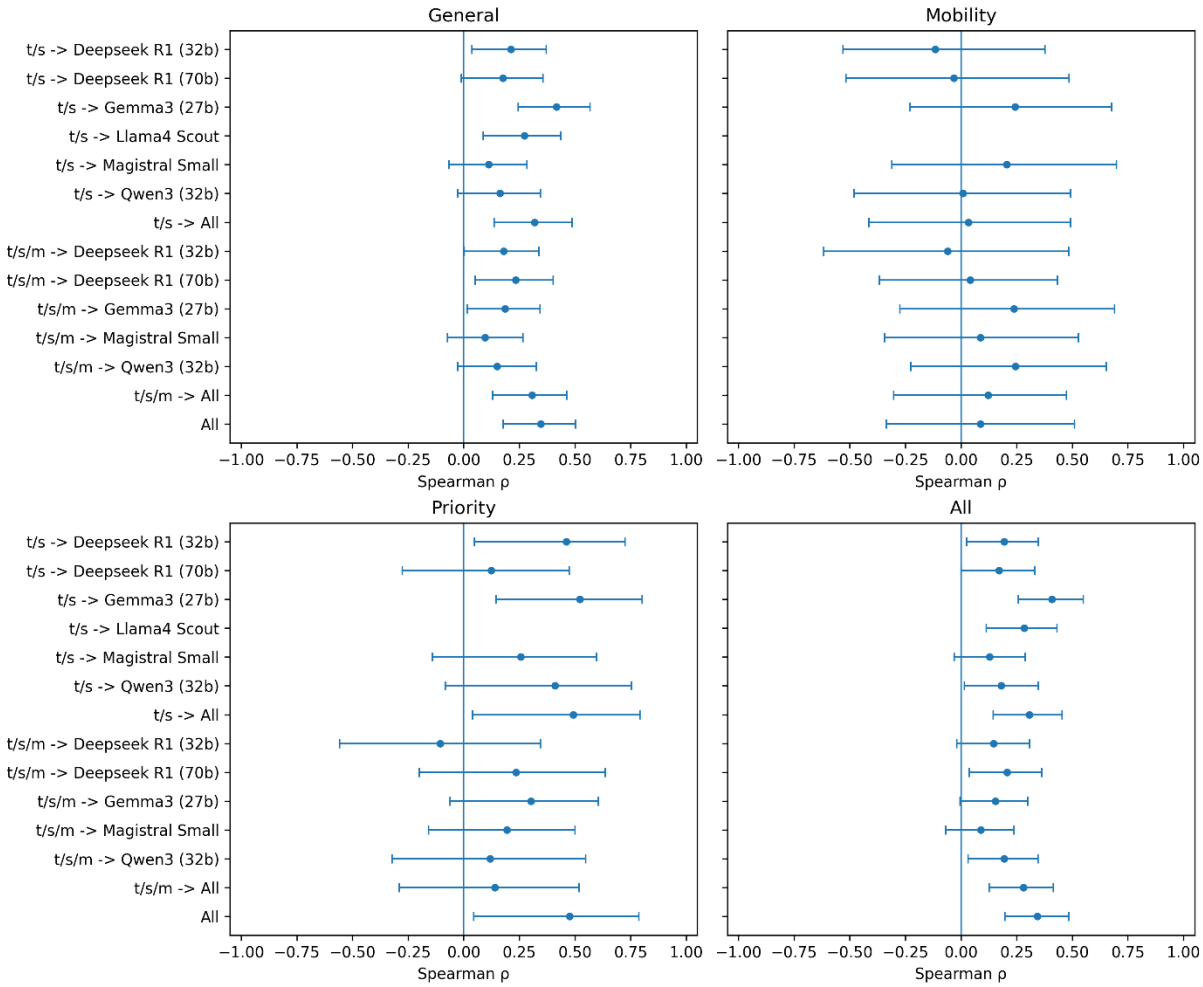


图 1. 基于提案标题和摘要（t/s）或标题、摘要和正文（t/s/m）得出的各评审组及总体研究问题的大语言模型评分（五次评分的平均值）与平均评审评分之间的斯皮尔曼相关系数。“全部”表示所有大语言模型评分的平均值，包括基于提案标题和摘要得出的评分。误差线表示 95% 置信区间。由于有效评分数量较少，标题、摘要和正文数据集中未纳入 Llama Scout。

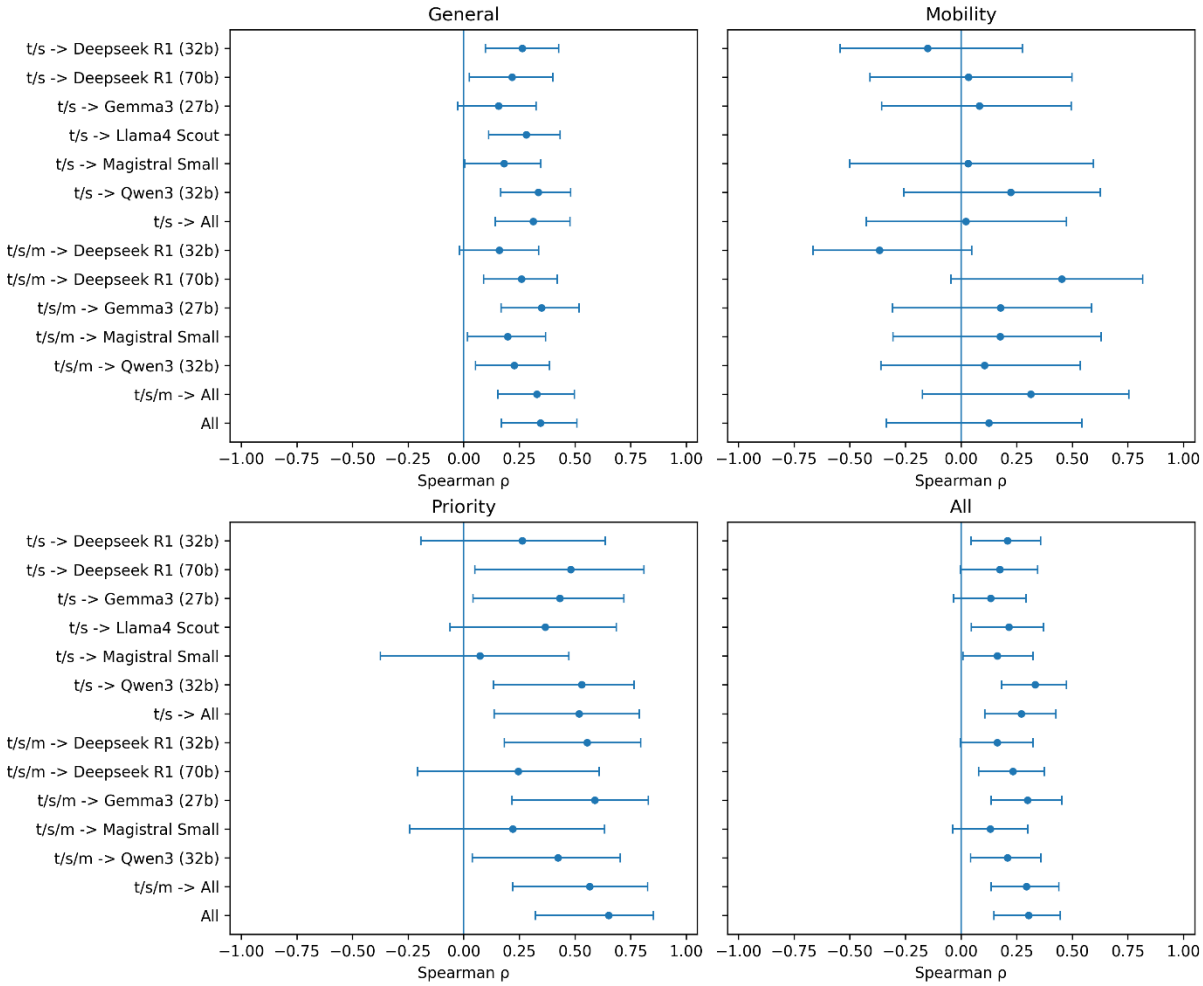


图2. 按评审组及总体划分的大语言模型评分（五次评分的平均值）与方法评审评分均值之间的斯皮尔曼相关系数，评分依据为提案标题和摘要（t/s），或标题、摘要和正文（t/s/m）。“All”表示所有大语言模型评分的平均值，包括依据提案标题和摘要得出的评分。误差线表示95%置信区间。由于 Llama Scout 产生的有效评分很少，因此未纳入标题、摘要和正文数据集。

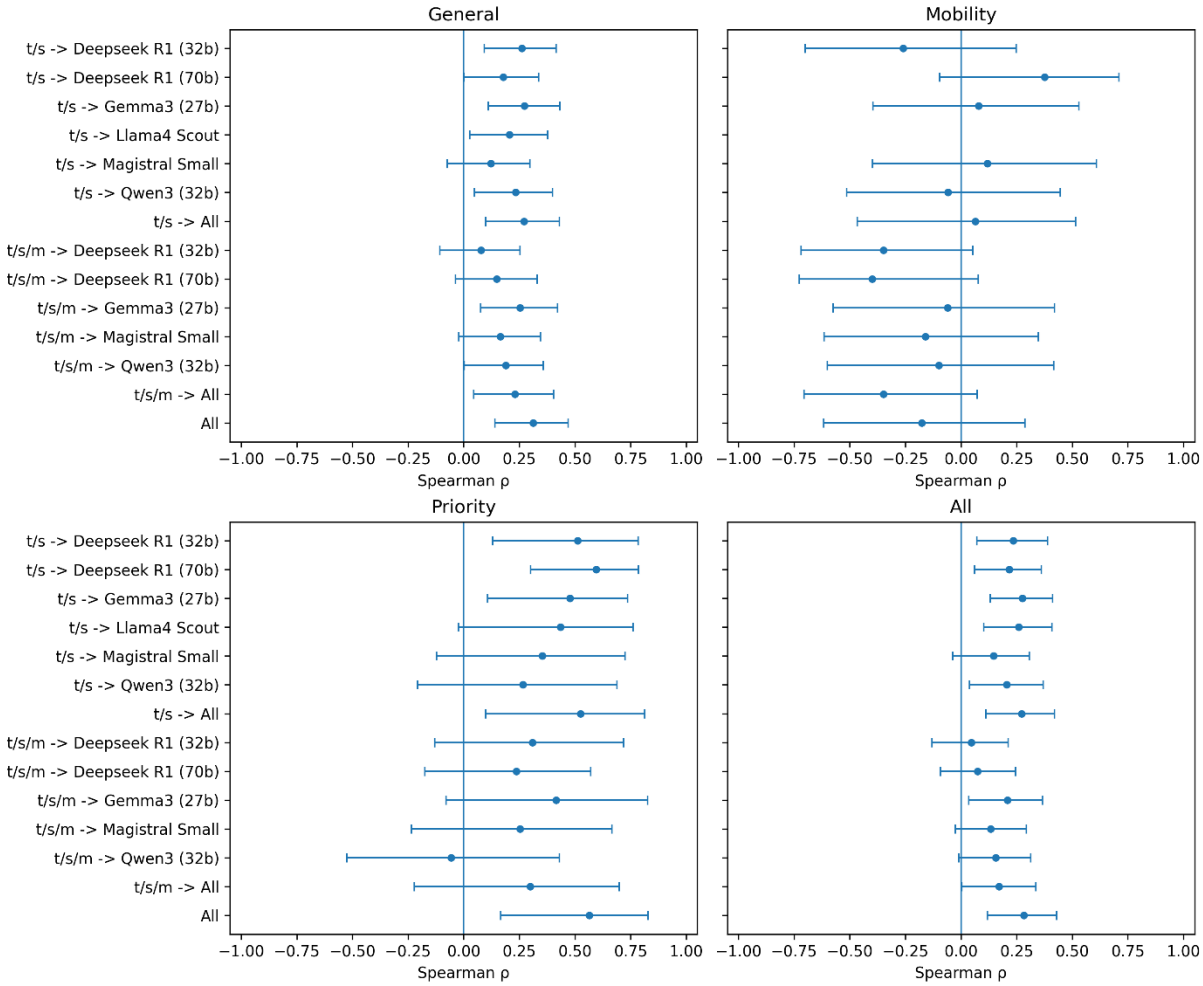


图 3. 基于提案标题和摘要（t/s），或基于标题、摘要和正文（t/s/m），按评审组及总体计算的人工智能评分（五次评分的平均值）与申请人能力平均评审评分之间的斯皮尔曼相关系数。“全部”表示所有人工智能评分的平均值，包括基于提案标题和摘要得出的评分。误差线表示 95% 置信区间。由于产生的有效评分较少，Llama Scout 未包含在标题、摘要和正文数据集中。

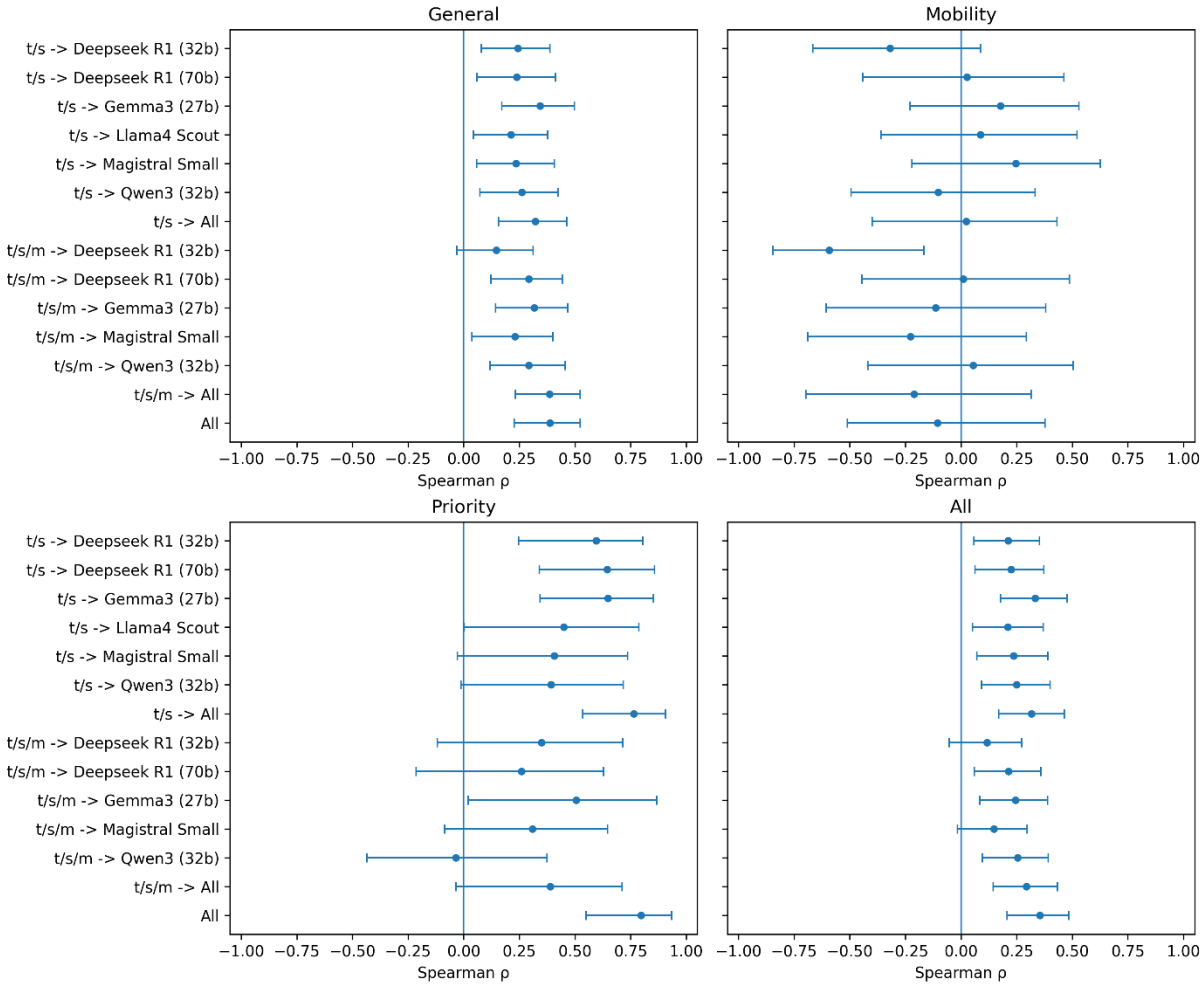


图 4. 大语言模型评分（五次评分的平均值）与评审评分均值之间按评审组及总体划分的斯皮尔曼相关**overall**，基于提案标题和摘要（t/s）或标题、摘要和正文（t/s/m）计算。“全部”表示所有大语言模型评分的平均值，包括基于提案标题和摘要得出的评分。误差线表示 95% 置信区间。由于 Llama Scout 产生的有效评分很少，因此在标题、摘要和正文数据集中缺少该模型。

在 26 个案例中，同一份提案有两个评审组评分。针对这些案例，计算了评审组评分之间的斯皮尔曼相关，结果如下。

- 一般项目：0.63（0.26，0.83），n=21。
- 流动项目：0.83（-1.00，1.00），n=4。
- 优先项目：0.35（-0.54，0.94），n=9。
- 全部：0.62（0.32，0.81），n=26。

因此，在样本量足以进行合理检验的类别（一般项目和全部项目）中，评审组彼此之间的意见一致程度，约为大语言模型与其平均分数一致程度的两倍。当然，个别评审人之间的意见一致率可能低于评审组之间的一致率，因此大语言模型的意见一致率可能更接近个别专家之间的一致率。

总体而言，大语言模型彼此之间的一致性高于它们与人工评审者之间的一致性。例如，对于总体评分，大语言模型之间的平均相关系数为0.34，而它们与评审评分之间的平均相关系数为0.22。大语言模型之间的排名一致性在输入为标题和摘要时最高（图5）。因此，它们似乎以大体相似的方式处理文本。通过对大语言模型进行平均而获得的更高相关性表明

它们可能会从提案中提取不同的相关模式，而平均处理则利用了这一点。较长文本中的较低一致性可能反映出，随着文本信息增多，算法变体有更大的空间以不同方式利用这些信息。

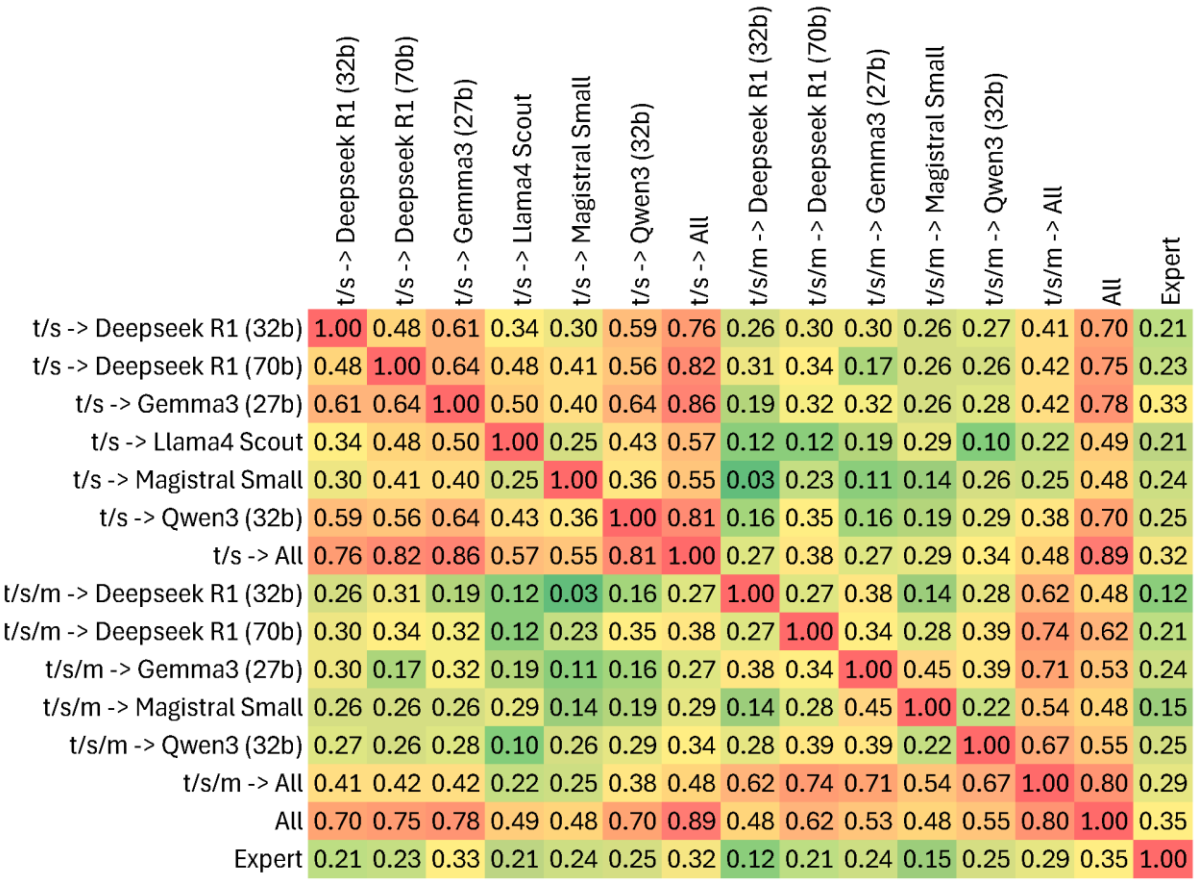


图5. 大语言模型评分（五次评分的平均值）与评审评分总体均值之间的斯皮尔曼相关系数。由于 Llama Scout 生成的有效评分较少，因此未纳入标题、摘要和正文集合。

5 讨论

本研究存在许多局限，其中大多数可能导致大语言模型与专家判断之间的相关性低于理想测试中的相关性。主要局限在于数据年代久远，且样本量太小，无法分别分析各评审小组，而这将是最稳健的做法。此外，大语言模型可能没有考虑接收单位环境质量和声誉，而人工评审者很可能考虑了这些因素。另一项可能提高相关性的重要局限是，资助流程的结果部分已为人所知，并且可能包含在训练数据中，因为获资助研究人员可能会在简历和发表的论文中注明其获得的资助。这就带来一种风险：大语言模型可能通过获知结果来作弊；但这种情况似乎不太可能，因为它们似乎并不了解个别学术文献的具体内容。例如，即使是被引次数很高的已发表期刊论文，它们也很少能回答有关这些论文的基本问题（Thelwall，2026）。此外，1994年的评审标准和申请流程如今是否仍然适用，也并不明确。而且，瑞典申请书的翻译步骤可能影响了评分。

与以往研究相比，尽管此前尚无研究使用大语言模型预测资助评审分数，但本研究的发现为不断增加的证据添砖加瓦：针对多种文档类型，大语言模型能够生成与专家评分呈正相关的分数（Kousha 等，2025；Thelwall，2025；Thelwall 和 Cox，2025；Thelwall 和 Yaghi，2025a）。当前发现的相关性弱于以往大多数研究结果，这表明这项任务对大语言模型而言可能相对困难。一种可能的解释是，资助评审本身就很困难；不过，人际相关性要高得多（尽管置信区间较宽），因此这种解释似乎不太可能。另一种可能是，提案年代久远增加了任务难度。还有一种解释是，对于研究员奖学金，申请书评估高度以个人为中心；对于瑞典医学研究人员这样规模相对较小的群体，人工评审者在评估个人时可能考虑了更多不言明的背景知识。

6 结论

研究结果首次表明，尽管规模有限，大语言模型在为资助申请书评分时，虽然评分幅度与人工评分不一致，但其排名与人工评分相符的能力虽弱，却具有统计显著性。这一结论适用于中等规模的大语言模型；截至2026年初，Gemma 3 27b 可能是表现最佳的可用模型，但并非所有模型都经过测试，而且模型间的差异不具有统计显著性。仅依据申请书标题和摘要、不看正文，也几乎能同样准确地做出预测，因此这些预测显然是基于模式的猜测，而非任何有实质意义的“评估”。此外，将多个模型对两种输入类型（即有正文和无正文的标题及摘要）给出的评分取平均，似乎能取得最佳结果，其相关性达到专家评审小组之间相关性的56%。

当前研究的结果总体上支持 La Caixa 现有的做法，即使用大语言模型进行申请初筛，并由人工做出最终决定。另一种可能的应用场景是在决胜阶段：当多个评分相近的申请中只有一部分能够获得资助时，目前委员会可能会在阅读提案后做出决定；让大语言模型评分来裁决平局，似乎至少同样公平，而且速度更快。如果认为这种做法不可接受，那么至少可以让大语言模型先进行排名，为最终决定提供参考。当然，大语言模型也可以作为人工评分的补充，用于交叉核查偏见，或发现那些因多位评审人给出严苛评价而“不走运”的申请，以及因多位评审人给出宽松评价而“走运”的申请。

7 致谢

本研究部分由 ESRC 元科学基金 UKRI1486 资助。

8 参考文献

Bharti, P. K., Dalal, V. 和 Panchal, M.（2026年）。Co-Reviewer：人工智能能像人类一样进行评审吗？用于同行评审中大语言模型与人工评审一致性的智能体框架。《科学计量学》，1-42。https://doi.org/10.1007/s11192-026-05557-6 Bornmann, L., Mutz, R. 和 Daniel, H.-D.（2010年）。期刊同行评审的信度泛化研究：评审者间信度及其决定因素的多层次荟萃分析。《公共科学图书馆·综合》 2010年12月14日，https://doi.org/10.1371/journal.pone.0014331。Carbonell Cortés, C., Parra-Rojas, C., Pérez-Lozano, A., Arcara, F., Vargas-Sánchez, S., Fernández-Montenegro, R. 和 López-Verdeguer, I.（2024年）。人工智能辅助预筛选

生物医学研究提案：伦理考量与“la Caixa”基金会的试点案例。《数据与政策》，6，e49。Cicchetti DV。手稿和基金申请提交的同行评审信度：一项跨学科调查。《行为与脑科学》1991年；14：119–135。Cole S、Cole JR、Simon GA。同行评审中的偶然性与共识。《科学》1981年；214：881–886。Du, J., Wang, Y., Zhao, W., Deng, Z., Liu, S., Lou, R. 和 Yin, W.（2024年）。大语言模型协助自然语言处理研究人员：批评论文（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.292 Farooq, A. 和 Iqbal, K.（2026年）。通过多智能体机制设计重构同行评审流程。arXiv预印本 arXiv:2601.19778。Goyal, P., Parmar, M., Song, Y., Palangi, H., Pfister, T. 和 Yoon, J.（2026年）。ScholarPeer：用于自动化同行评审的上下文感知多智能体框架。arXiv预印本 arXiv:2601.22638。Koo, T.K. 和 Li, M.Y.（2016）《选择和报告组内相关系数的指南》，《脊椎按摩医学杂志》，15/2：155-163。Kousha, K., Thelwall, M. 和 Gadd, E.（2025年）。ChatGPT能评估研究环境吗？来自REF2021的证据。arXiv预印本 arXiv:2512.05202。Langfeldt, L.（2006年）。同行评审的政策挑战：管理偏见、利益冲突和跨学科评估。《研究评估》，15(1)，31–41。Li, D. 和 Agha, L.（2015年）。大名气还是大创意：同行评审小组能否选出最佳科学研究提案？《科学》，348(6233)，434–438。https://doi.org/10.1126/science.aaa0185。Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D. Y., Yang, X. 和 Zou, J.（2024年）。大语言模型能否对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, Y., Zhao, X., Kruegel, C., Song, D. 和 Bu, Y.（2025年）。大语言模型的上下文内水印。arXiv预印本 arXiv:2505.16934。Naddaf, M.（2026年）。人工智能常用于同行评审——且经常违反指导原则。《自然》，649，273。Newman-Griffis, D., Woods, H. B., Wu, Y., Thelwall, M. 和 Holm, J.（2025年）。算法资助——科研资助机构负责任使用人工智能和机器学习指南。https://doi.org/10.6084/m9.figshare.29041715 Nunkoo, R. 和 Thelwall, M.（2025 年）。使用 ChatGPT 评估旅游研究质量。《旅游当前问题》。https://doi.org/10.1080/13683500.2025.2596265 Parra, E., Haiduc, S., Chatterjee, P., Ehsani, R. 和 Iaremchuk, P.（2026年）。迈向软件工程同行评审的可持续未来。arXiv预印本 arXiv:2601.21761。Ryu, H., Jang, D., Lee, H. S., Jeong, J., Kim, G., Cho, D. 和 Yang, E.（2025年）。ReviewScore：利用大语言模型进行错误信息同行评审检测。arXiv预印本 arXiv:2509.21679。Sadallah, A., Baumgärtner, T., Gurevych, I. 和 Briscoe, T.（2025年11月）。好的、坏的与建设性的：自动测量同行评审对作者的效用。载于2025年自然语言处理实证方法会议论文集（第28979-29009页）。Sandström, U. 和 Riis, U.（2025年）。一项高影响力同行评审研究的信度与效度：探究Wennerås 和 Wold在《自然》中的数据。《定量科学研究》，6（2）：1315–1335。https://doi.org/10.1162/QSS.a.395

Shrout, P.E. 和 Fleiss, J.L.（1979）《组内相关系数：在评估评审者信度中的应用》，《心理学通报》，86/2：420。Sikimić, V. 和 Radovanović, S.（2022）。科研资助评审中的机器学习：对高能物理项目效率进行算法预测。《欧洲科学哲学杂志》，*12*（3），50。Smith, R.（2007年9月）。Tesseract OCR 引擎概述。载于第九届文档分析与识别国际会议（*ICDAR 2007*）（第2卷，第629–633页）。IEEE。Thakkar, N.、Yuksekgonul, M.、Silberg, J.、Garg, A.、Peng, N.、Sha, F. 和 Zou, J.（2025年）。大语言模型反馈能否提升评审质量？ICLR 2025上2万份评审的随机研究。*arXiv* 预印本 *arXiv:2504.09737*。Thelwall, M.（2025年）。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.（2026年）。大语言模型了解有关期刊文章的基本事实吗？《文献学杂志》。https://doi.org/10.1108/JD-11-2025-0330 Thelwall, M. 和 Cox, A.（2025年）。使用 ChatGPT 根据学术图书的描述估计其质量。《学术图书馆学杂志》，51（2），文章103023。https://doi.org/10.1016/j.acalib.2025.103023Thelwall, M. 和 Mohammadi, E.（2026年）。小型大语言模型和推理型大语言模型能否为期刊文章的研究质量评分？平均化和少样本方法是否有帮助？《科学计量学》。https://doi.org/10.1007/s11192-026-05585-2 Thelwall, M. 和 Yaghi, A.（2025a）。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》，130（10），5285–5301。https://doi.org/10.1007/s11192-025-05287-1 Thelwall, M. 和 Yaghi, A.（2025b）。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。https://doi.org/10.48550/arXiv.2409.16695Thelwall, M. 和 Yang, Y.（2025年）。ChatGPT 给出的隐性和显性研究质量评分概率。《定量科学研究》。6，1271–1293。https://doi.org/10.1162/QSS.a.393Wennerås, C. 和 Wold, A.（1997年）。同行评审中的裙带关系与性别歧视。《自然》，387，341–343。https://doi.org/10.1038/387341a0 Yun, J.、Kwon, J.、Kim, M. 和 Kim, Y.（2026年）。关于大语言模型辅助同行评审的立场：通过指导和反馈应对评审者差距。*arXiv* 预印本 *arXiv:2601.09182*。