哪些文体特征会误导 ChatGPT研究评价?

Kayvan Kousha
统计网络计量学与研究评价小组，伍尔弗汉普顿大学商学院，英国。
https://orcid.org/0000-0003-4827-971X；k.kousha@wlv.ac.uk

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

大语言模型（LLM）有望用于支持研究评价，并具备一定能力，可根据期刊论文的标题和摘要估计其研究质量。本文评估了是否存在与研究质量无关、却会影响ChatGPT评分的语言相关因素。我们使用了提交至英国卓越研究框架（REF）2021评审的99,277篇期刊文章构成的数据集，从摘要中计算了若干可读性指标，并将其与ChatGPT评分和院系卓越研究框架评分进行相关性分析。结果显示，在许多学科领域，语言复杂度和长度与ChatGPT研究质量评分的关联，比与REF专家评分的关联更强。虽然本文并未检验因果关系，但这些结果表明，与人类专家相比，ChatGPT可能更倾向于奖励语言复杂度，从而在许多领域可能偏向较长且可读性较低的摘要。大语言模型对复杂语言表现出的明显偏好，是大语言模型在研究质量评价实际应用中的一个不理想特征，除非能够找到解决方案。

关键词：大语言模型；ChatGPT；人工智能偏见；研究评价；摘要可读性；卓越研究框架（REF）。

引言

已发表学术研究的质量经常需要在任职聘用、晋升以及国家或地方层面的院系研究评价中接受评估，例如英国的卓越研究框架（REF）。这些流程耗时很长，得出的结果也可能不准确，尤其是在评估者没有足够时间或专业知识来妥善评估分配给他们的研究成果时。过去，人们一直使用基于引文的指标来辅助或取代专家评估，但不同的大语言模型（LLM）似乎给出的结果与专家评分的排序更为接近（Thelwall，2025c；Thelwall & Yaghi，2025；Thelwall & Mohammadi，2026）。然而，基于大语言模型的质量预测为何能与人工评分达到中等程度的一致，目前尚不清楚；尤其值得注意的是，与输入全文相比，仅向大语言模型提供文章标题和摘要时，它们似乎表现得更好（Thelwall，2025ab）。卓越研究框架影响案例也呈现出类似模式：根据标题和摘要得出的 ChatGPT 评分，与专家评分的相关性高于使用全文时的相关性（Kousha & Thelwall，2025）。这不禁令人怀疑，大语言模型是否在通过利用与研究质量无关的文档属性来“作弊”。因此，识别这些属性有助于理解大语言模型的这一关键特性，也有助于评估作者是否可能采取简单措施来操纵或蒙骗基于大语言模型的研究评价系统。

以往关于大语言模型用于研究质量评价的分析，往往着重于尽可能提高其与专家评分的一致性。然而，少数研究尝试

识别 或检验潜在的偏见来源。一项针对所有 REF2021 评估单元中 ChatGPT 评分的回归分析发现，ChatGPT 对来自某些国家的研究、发表于高影响力期刊的文章以及摘要较长的文章评分更高。其中，最后一点最令人担忧，因为它似乎与论文质量无关。相比之下，有些国家在研究方面投入更多，而引用率较高的期刊平均而言可能会发表质量更好的研究。不同领域的平均评分也存在显著差异，这表明存在主题偏差（Thelwall & Kurt, 2025）。一项对旅游学标题和摘要中高分与低分文本所特别常见的词语进行比较的研究发现，ChatGPT 往往会给采用复杂统计方法的研究较高评分，而给调查研究较低评分（Nunkoo & Thelwall, 2025）。对于图书馆与信息科学，大语言模型似乎会给使用更多技术术语、理论、统计学、实验和算法的论文较高评分；而对于抽象性较强的论文，以及涉及图书馆、学生或调查的研究，评分则相反（Zhu et al., 2026; Thelwall, 2025d）。目前尚不清楚这些现象中有些或全部是否反映了潜在的质量差异（由人类专家评判），但至少表明可能存在人工智能偏见。关于大语言模型用于其他任务的更广泛研究表明，如果告知大语言模型信息来源的可信度或国籍，它们可能会受到影响（Germani & Spitale, 2025; Vasu et al., 2025），但语言风格对其影响可能不大（Kundu & Barbosa, 2024）。

尽管已有上述研究发现，且下文将回顾相关研究，但摘要的文本特征与大语言模型评分之间的关系仍鲜为人知。为此，本文探讨以下研究问题。

- RQ1：摘要的哪些文本特征（如可读性、语言复杂度和长度）与大语言模型评分相关?
- RQ2：与 REF 专家评分相比，这些关联与大语言模型评分的关联是否更强?
- RQ3：上述问题的答案是否因学科（评估单元）而异?

背景

尽管本文聚焦于研究质量，但关于研究质量与摘要语言特征之间关系的相关研究很少。因此，本节转而回顾利用引文数据开展的相关研究。

摘要长度与引文影响力

摘要概述学术研究的目标、方法和研究发现。通常，摘要和标题是读者阅读论文时首先接触的部分，读者会据此判断一项研究是否相关。因此，鉴于摘要具有这一重要特点，许多科学计量学研究将摘要用作文本数据的主要来源，以探究其与研究质量或引文影响力之间的关联。摘要的某些特征，如长度、语言复杂度和可读性，可能会影响论文被引用的频率（Kousha & Thelwall, 2024）。

多项研究发现，摘要较长与更高的引文影响力相关，这或许是因为较长的摘要包含更多信息，有助于读者理解研究的关键方面，进而使论文更容易被引用。例如，对《柳叶刀》1997–1999 年发表的文章进行的一项研究发现，高被引论文的摘要长得多，词数是低被引论文的 2.5 至 7 倍（Kostoff, 2007）。一项涵盖八个科学领域、一百万篇摘要的大规模研究发现，较短的摘要（词数和句子数较少）与较少的引文相关，不过在数学和物理学领域，较短的句子则与更多引文相关，

这表明存在学科差异（Weinberger 等，2015）。另一项大规模研究分析了来自 1500 多种期刊的 430 万篇论文，也发现几乎所有期刊中，摘要长度都与被引次数呈正相关（Sienkiewicz 与 Altmann，2016）。其他针对不同学科的研究也得出了类似的发现（Didegah 与 Thelwall，2013；Hafeez 等，2019；Sohrabi 与 Iraj，2017；Van Wesel 等，2014）。然而，实验性证据表明，较长的摘要未必能提高读者关注度（Helbach 等，2025）。因此，在研究质量评估的背景下，尚不清楚专家或其他评估者在判断研究质量时是否会受到摘要长度的影响。

摘要可读性与引文影响力

许多研究探讨了摘要可读性较高（例如弗莱施阅读易读性分数较高）是否与被引次数较多相关，结果发现，复杂度较高（可读性较低）的摘要获得的引文更多，但不同学科之间存在差异。例如，高被引信息科学文章的摘要可读性较低（以弗莱施阅读易读性分数衡量）（Gazni，2011）。一项规模庞大的研究分析了 430 万篇期刊文章，发现摘要复杂度（以冈宁雾指数衡量）与被引次数呈正相关（Sienkiewicz 与 Altmann，2016）。另一项研究发现，在生物学与生物化学领域，摘要可读性（以弗莱施阅读易读性分数衡量）与被引次数呈负相关，而在化学或社会科学领域则未发现显著关联（Didegah 与 Thelwall，2013）。类似地，一项比较多个科学领域中 Web of Science 高被引论文与未被引论文的研究发现，高被引文章的摘要更长、语言复杂度更高且可读性更低（Hu、Wang 与 Deng，2021）。对于新兴技术领域的文章，使用多项可读性指标（弗莱施阅读易读度、弗莱施-金凯德年级水平和冈宁雾指数）测得的摘要可读性越低，其引文影响力越高（Ante，2022）。一项针对语言与语言学期刊 71,628 篇摘要（ 1991—2020 年）的研究也发现，摘要可读性较低与被引次数较多相关，但这种关系较弱（Wang、Liu 与 Zhou，2022）。一项规模较小的研究分析了发表于《科学》的 550 篇文章，也报告称摘要可读性与文章获得的在线关注度（Altmetric 分数）总体上相关（Jin 等，2021）。

方法

本研究的研究设计是获取大量不同学科领域学术期刊论文的大语言模型研究质量评分及其摘要的文体特征（如可读性、语言复杂度和长度），并识别两者之间是否存在关联。研究设计的第二部分是重复这一过程，但使用专家评分而非大语言模型评分，以评估专家与大语言模型在所发现的模式上是否存在差异。这类差异可能表明存在人工智能偏见，而非与研究质量相关的间接因素。研究设计的第三部分（不涉及研究问题，仅为提供背景）是用引文影响力指标替代大语言模型评分，以评估摘要的文体特征与引文影响力之间的关联是否相同或有所不同。

数据集

近期唯一一个涵盖所有领域已发表期刊文章并附有质量指标的大规模数据集，是英国 REF2021 成果集。它收录了研究

成果提交评估的项目。每项成果都由相应评估单元（大致按领域划分）至少两位领域专家按照 1 至 4 的整数等级评定研究质量评分（https://2021.ref.ac.uk/publications-and-reports/panel-criteria-and-working-methods-201902/）。虽然单项成果评分已被销毁，但每个“提交单位”中获得四个评分等级的成果数量是公开的。“提交单位”可以近似看作一所大学中的院系或学院，但并非一定如此。为方便起见，本文将使用“院系”一词。每项成果均被赋予提交该成果的院系的平均评分，作为代理质量评级（如果一篇文章由多位作者提交，则取所有提交院系的平均评分）。由于 REF2021 中每个院系都有多个不同的评分，这种做法在所有情况下都不准确。尽管如此，由于单项成果评分与院系评分相关，在不存在偏见的情况下，这种做法会减弱相关性，但不会使任何正相关失效。

此前一篇论文（Thelwall & Mohammadi，2026）为UoA 1至6的每个评估单元各500篇文章生成了个体分数，本研究采用了这些分数。这些分数由本文第二作者得出，代表他对这些文章REF质量评分的估计。

我们从REF2021网站（https://results2021.ref.ac.uk/all-submission-data）获取了受评成果的电子表格，并通过机构和提交单位（院系）两列，将其与院系评分电子表格（https://results2021.ref.ac.uk/profiles/export-all）交叉参照。评分的一个不理想之处在于，它们完全公开，可能已被纳入大语言模型训练数据。不过，这种可能性似乎不大，因为数据采用Excel格式，并非大语言模型常用的输入格式。此外，要将文章与评分对应起来，还需要交叉参照这两个电子表格，这也不太可能。最后，此前的测试表明，ChatGPT 的 API 版本很少能仅凭文章标题和摘要识别出其所属院系，甚至连期刊也无法识别（Thelwall，2026），因此不太可能查找到与REF相关的文章质量信息。

对于这个数据集，每个评估单元中，非文章类成果（例如书籍、章节）以及摘要长度最短的 10% 的文章均被剔除。后者中，许多短摘要来自短篇文章（例如读者来信、通信），或是不完整的。文章数量过多，无法逐篇核查，尤其是很难判断某篇文章是否属于短篇贡献，因此采用 10% 这一启发式方法，粗略剔除不合适的文章。最终数据集包含 99,277 篇不含短摘要的独特期刊文章（计入重复项则为 106,843 篇）。

引用率

为每篇文章计算了按学科和年份标准化的引文率，作为一项额外指标。采用标准化对数转换引文得分（NLCS），该指标会对发表学科和年份进行调整，同时也考虑到引文数据高度偏斜的特性（Thelwall，2017）。计算时，先使用 ln(1+x) 转换所有被引次数，以降低偏斜程度；然后将每篇文章的（对数转换后）被引次数除以同一学科和年份所有文章的平均对数转换被引次数（涉及多个学科时，对各学科的数值取平均）。相关数据取自 2024 年的 Scopus，因此每篇论文至少有 3 年的引文数据。标准化所用的学科分类依据 Scopus 使用的全科学期刊分类（ASJC）。这一过程可减少引文数据中的年代偏见和学科偏见。

评分来自多个大语言模型，均沿用自先前的一篇论文（Thelwall，2025c）。详情见该论文，简而言之，所用模型包括 ChatGPT-4o、其较小版本 ChatGPT-4o mini，以及 ChatGPT-5 mini。对于每个模型，研究者都通过 ChatGPT 的应用程序编程接口（API）提交文章。

这些评分依据系统提示生成。该提示描述了任务，并包含最初提供给 REF2021 人类评审员的指示，要求他们根据文章的标题和摘要进行评分。评分采用 REF2021 的等级标准，从 1*（具有国家层面的相关性）到 4*（世界领先）（确切提示内容见：Thelwall，2025c）。除一个案例外（下文将对此进行说明），所有评分均为五次独立请求的平均值，因为研究表明，取平均值可提高评分与专家评分之间的相关性（Thelwall，2025abc）。

最后一个未进行五次重复的案例采用了简单的结构化提示，要求分别对原创性、严谨性和重要性这三个质量要素进行评论，随后为每项评分，最后再给出总体评论和总体评分（确切提示内容见：Thelwall，2025c）。这与可能要求专家采用的流程类似，不同之处在于，用户提示说明允许使用小数评分（例如 3.5*），而人类专家必须给出整数评分：1*、2*、3* 或 4*。

在主要分析中，对所有评分取平均值，得出一个综合估值，以合并 Chat GPT-4o、ChatGPT-4o mini 和 ChatGPT-5 mini 的评分。先前的研究发现，按与专家评分的相关强度衡量，这种组合的整体质量排名最佳（Thelwall, 2025c）。对新旧模型的评分取平均后，效力有所提升，这可能是由于增加了迭代次数，或结合了不同模型的视角。

可读性特征

使用基于 *textstat* 库（https://pypi.org/project/textstat/）的 Python 脚本，计算数据集中摘要的可读性和文本特征。对于每篇摘要，提取了若干标准可读性指标：弗莱士阅读易读度、Flesch–Kincaid 年级水平、冈宁雾指数、SMOG指数和自动可读性指数。由于先前研究表明摘要长度与大语言模型评分相关（Thelwall & Kurt, 2025），因此也计算了摘要长度。主要分析聚焦于弗莱施阅读易读性分数和 Flesch–Kincaid 年级水平这两项指标，因为它们是科学计量学研究中广泛使用的可读性指标。弗莱施阅读易读性分数根据平均句长和每词音节数估算文本的阅读难度；分数越高，文本越容易阅读（Flesch, 1948）。Flesch–Kincaid 年级水平采用相同的两项语言学要素，但将结果表示为理解文本所需的美国学校年级水平（Kincaid et al., 1975）。主要分析聚焦于 Flesch–Kincaid 年级水平，因为该指标应用广泛。其他可读性指标（弗莱士阅读易读度和冈宁雾指数）见附录（分别为图 S1 和 S2）。

文本特征

可计算的文本指标种类繁多。本研究仅采用以往引文分析研究中考察过的指标，以便控制计算工作量，并避免推测性地加入过多指标。在缺乏理论依据的情况下测试过多指标，可能会因测试数量过多而发现假阳性结果（即邦费罗尼家族错误率问题）。基于此，选取了以下指标

用于主要分析：摘要长度（词数）、每句词数和每词音节数。这些指标体现了摘要风格的关键方面，符合标准可读性指标的用法，也体现了长度。为避免冗余，不报告许多彼此密切相关的指标（例如每词字符数、每句字符数或字符总数）。

分析

此前关于大语言模型对学术成果评分的研究往往发现，个体分数通常不准确，且明显偏向3*；但这些评分的排名包含有用信息，通常与专家评分具有中等程度的一致性（与质量代理指标的一致性稍弱）（Thelwall，2025c）。因此，衡量其有用性的唯一相关指标是秩相关。秩相关有多种，本文选择斯皮尔曼相关系数而非肯德尔秩相关系数，因为后者不会对误差大小进行惩罚。为报告统计显著性，我们为所有斯皮尔曼相关系数计算了自助法95%置信区间。图表中未显示这些区间，因为加入误差线会使图表难以阅读。

结果

主评审组*A*：健康与生命科学（文章层面比较）

对于每个评估单元有 500 个个体评分的健康与生命科学（UoA 1至6）（n=3,000，包含在多个评估单元中重复出现的文章），可以直接将 ChatGPT评分（而非院系平均分）和单篇文章评分与可读性指标进行比较。总体而言，最一致的斯皮尔曼相关系数出现在每词音节数、每句词数，以及由这两者综合得出的弗莱施-金凯德年级水平上。在所有情况下，ChatGPT的相关性都比 REF 专家评分的相关性更为正向，这表明 ChatGPT 可能更容易受到这些因素的影响，但因果关系尚未得到证实。例如，在临床医学中，弗莱施-金凯德年级水平与 ChatGPT评分之间的斯皮尔曼相关系数为 0.329，而与 REF 专家评分之间的相关系数为 0.124。公共卫生（0.218 对 0.037）、辅助卫生专业（0.291 对 0.072）和心理学（0.294 对 0.137）也存在类似差异。ChatGPT评分与构成这一指标的基础语言特征之间的相关性也更强，例如每句词数（临床医学中为 0.306 对 0.150）和每词音节数（0.220 对 0.039）。其他词汇复杂度指标也呈现类似模式，包括每词字母数（0.234 对 0.023）。

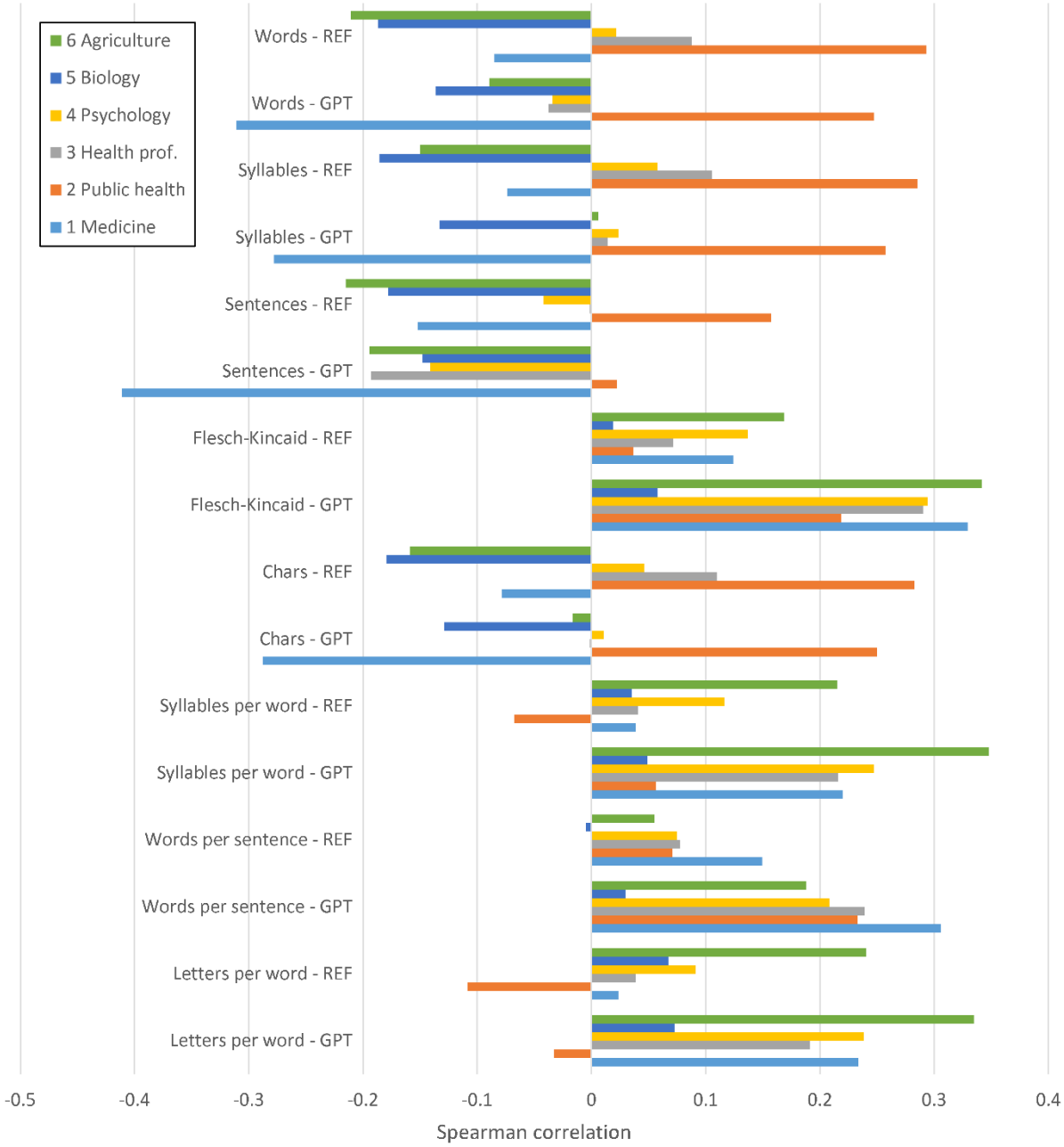


图1. ChatGPT或专家的REF评分与REF2021健康与生命科学评估单元文本复杂度指标的相关性（每个评估单元的n=500）。

*UoA 1*至*6*较小与更大的数据集分析

由于无法获得主评审组A中更大数据集（评估单元 1至6；n=39,832 篇文章）的个体 REF 评分，因此采用院系平均 REF 分数作为单篇论文研究质量的代理指标。为评估每个评估单元 500 篇文章的分析（采用个体评分）中观察到的模式，能否在使用代理指标时重现，研究人员对六个健康与生命科学评估单元的更大数据集进行了分析，并采用院系平均 REF 分数。表 1 至 3 比较了 500 篇论文数据集（采用个体评分）与更大数据集（采用院系平均分）中，摘要语言复杂度指标与评估分数之间的相关性。在全部六个健康与生命科学评估单元中，两种数据集里的 ChatGPT评分相关性都始终高于 REF 评分。即使将 REF 评分表示为

更大数据集中的院系平均值代理指标。这表明，基于院系平均 REF 分数的分析可以用于得出推论，其效果与基于个体分数的分析相当。

表1. 500篇论文样本和更大的数据集中，弗莱施-金凯德年级水平与研究质量评分之间的相关性。

UoA	ChatGPT (500 论文）	REF 个体分数	N	ChatGPT （更大的数据集）	REF 部门的评分	N
临床医学	0.220	0.039	500	0.273	0.062	9,602
公共卫生	0.056	-0.067	500	0.045	0.009	3,821
辅助卫生专业	0.216	0.041	500	0.201	0.055	9,337
心理学	0.247	0.116	500	0.195	0.085	7,881
生物科学	0.049	0.035	500	0.135	0.052	6,123
农业	0.348	0.215	500	0.326	0.208	3,068

表2. 500 篇论文样本和更大的数据集中，每句词数与研究质量评分之间的相关性。

UoA	ChatGPT （500 篇论文）	REF 个体分数	N	ChatGPT （更大的数据集）	REF 部门的评分	N
临床医学	0.306	0.150	500	0.266	0.039	9,602
公共卫生	0.233	0.071	500	0.191	-0.013	3,821
辅助卫生专业	0.239	0.077	500	0.225	0.049	9,337
心理学	0.208	0.075	500	0.160	0.000	7,881
生物科学 生物科学	0.030	-0.005	500	0.021	-0.032	6,123
农业	0.188	0.055	500	0.174	0.088	3,068

表3. 500篇论文样本和更大的数据集中每词音节数与评估分数之间的相关性。

UoA	ChatGPT (500 论文）	REF 个体分数	N	ChatGPT （更大的数据集）	REF 部门的评分	N
临床医学	0.329	0.124	500	0.353	0.069	9,602

公共卫生	0.218	0.037	500	0.185	0.003	3,821
辅助卫生专业	0.291	0.072	500	0.285	0.071	9,337
心理学	0.294	0.137	500	0.244	0.057	7,881
生物科学	0.058	0.019	500	0.113	0.013	6,123
农业	0.342	0.168	500	0.322	0.177	3,068

对全部*34*个*REF*评估单元的分析

根据主评审组A中 ChatGPT 与 REF 专家评分之间相关性的比较（图1），我们选取了三个指标，对全部34个 REF评估单元进行进一步分析：弗莱施-金凯德年级水平、每句词数和词数。选择这些指标是因为它们反映了摘要写作风格的不同方面。弗莱施-金凯德年级水平反映整体语言复杂度，每句词数反映句法复杂度，而词数衡量摘要的整体长度，这一指标在科学计量学研究中得到广泛应用。附录报告了采用具有不同语言特征的风格指标进行的补充分析，包括冈宁雾指数、SMOG指数、平均每词字母数、句子数和音节数。

弗莱施-金凯德年级水平

弗莱施-金凯德年级水平是每词音节数与每句词数的加权组合，用于估算理解文本所需的教育年级水平。评分越高，表示摘要的语言复杂度越高，因此阅读难度也越大（Kincaid 等，1975）。在全部34个评估单元中，23个单元的ChatGPT评分与弗莱施-金凯德年级水平呈统计显著正相关；1个单元呈统计显著负相关，其余评估单元则未显示统计显著的相关关系，这可能是由于样本量较小和/或相关关系较弱。这表明，语言复杂度较高的摘要往往获得更高的ChatGPT评分。然而，在34个评估单元中，REF 专家评分的院系平均分仅在9个单元中与年级水平呈统计显著正相关，方向与ChatGPT相同，但这些相关性通常要弱得多（图2）。另有6个评估单元的年级水平与REF 专家评分呈统计显著负相关，其中包括政治学与国际研究、教育学以及若干人文学科。

在若干生物科学和物理科学领域中，与摘要复杂度特征的正相关相比，ChatGPT评分的相关性强于 REF 专家评分。例如，在工程学（ChatGPT：0.215；REF：0.073）、化学（0.212；0.083）和生物科学（0.113；0.013）中均是如此。这些结果表明，ChatGPT评分与摘要中的语言复杂度关联更强，而 REF 专家评分通常与这一可读性指标仅有较弱关联，或毫无关联。

在许多社会科学学科中，随着年级水平的提高，ChatGPT评分更有可能上升；相比之下，REF 专家评分与这一指标之间通常仅存在微弱相关，或不具有显著相关性。在若干人文学科中，REF 专家评分与弗莱施-金凯德年级水平之间的关联也较弱且不一致，其中现代语言与语言学出现了一些显著的负相关（–0.163），英语语言文学

（-0.109），以及哲学（-0.112），这表明，在这些学科中，语言复杂度较高的摘要并未受到专家评估的青睐。

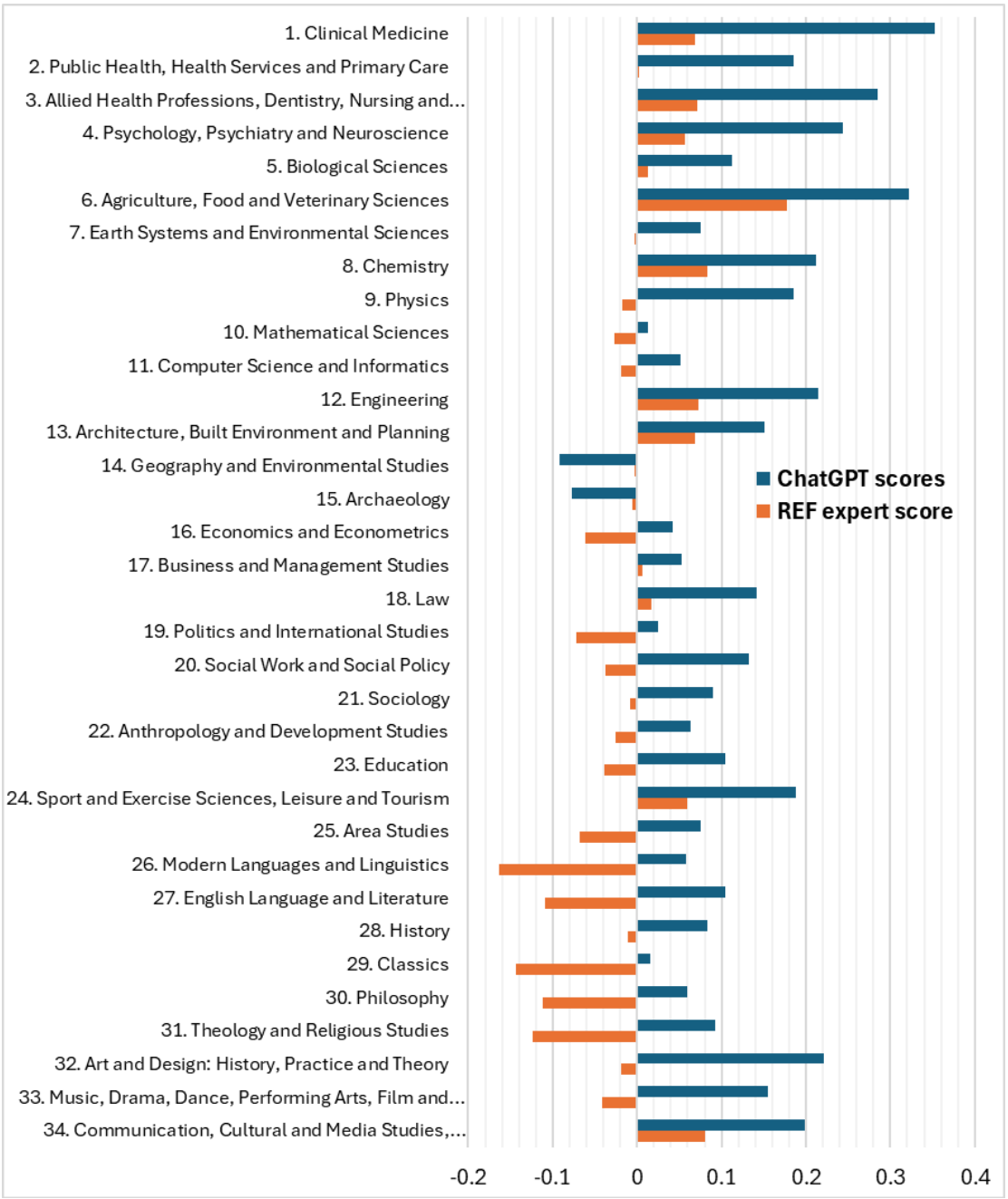


图2. 在34个REF评估单元中，弗莱施-金凯德年级水平与ChatGPT及REF院系平均专家评分之间的斯皮尔曼相关系数。

摘要长度

在图3中，34个评估单元中有29个显示ChatGPT评分与摘要长度（词数）之间存在显著相关性，其中包括22个正相关和7个负相关。正相关表明，较长的摘要往往会获得更高的ChatGPT评分，而负相关则意味着相反的情况。负相关出现在多个生物医学和科学、技术、工程和数学领域，例如临床医学（-0.311）和生物科学（-0.188）；而正相关则出现在许多社会科学和人文学科

学科，包括历史学（0.218）、英语语言文学（0.194）和哲学（0.214）。在部分生物医学和STEM学科中，摘要长度与ChatGPT评分呈负相关，尤其是在临床医学领域；这可能与这些领域摘要的写作风格有关。医学期刊通常要求使用结构式摘要（例如目标、参与者、方法、结果），并严格限制词数，以便清晰地概述关键研究发现。不过，摘要较短并不一定意味着更容易阅读。如图2所示，临床医学领域的语言复杂度与ChatGPT评分之间的正相关也最强。换言之，ChatGPT可能更倾向于给较短但语言复杂度较高的摘要打高分。

相比之下，院系平均 REF 专家评分仅在 15 个评估单元中与摘要长度存在统计学显著关联，且其中大多为负相关（12 个负相关，3 个正相关），这表明在这些学科中，较短的摘要往往与较高的平均专家评分相关。

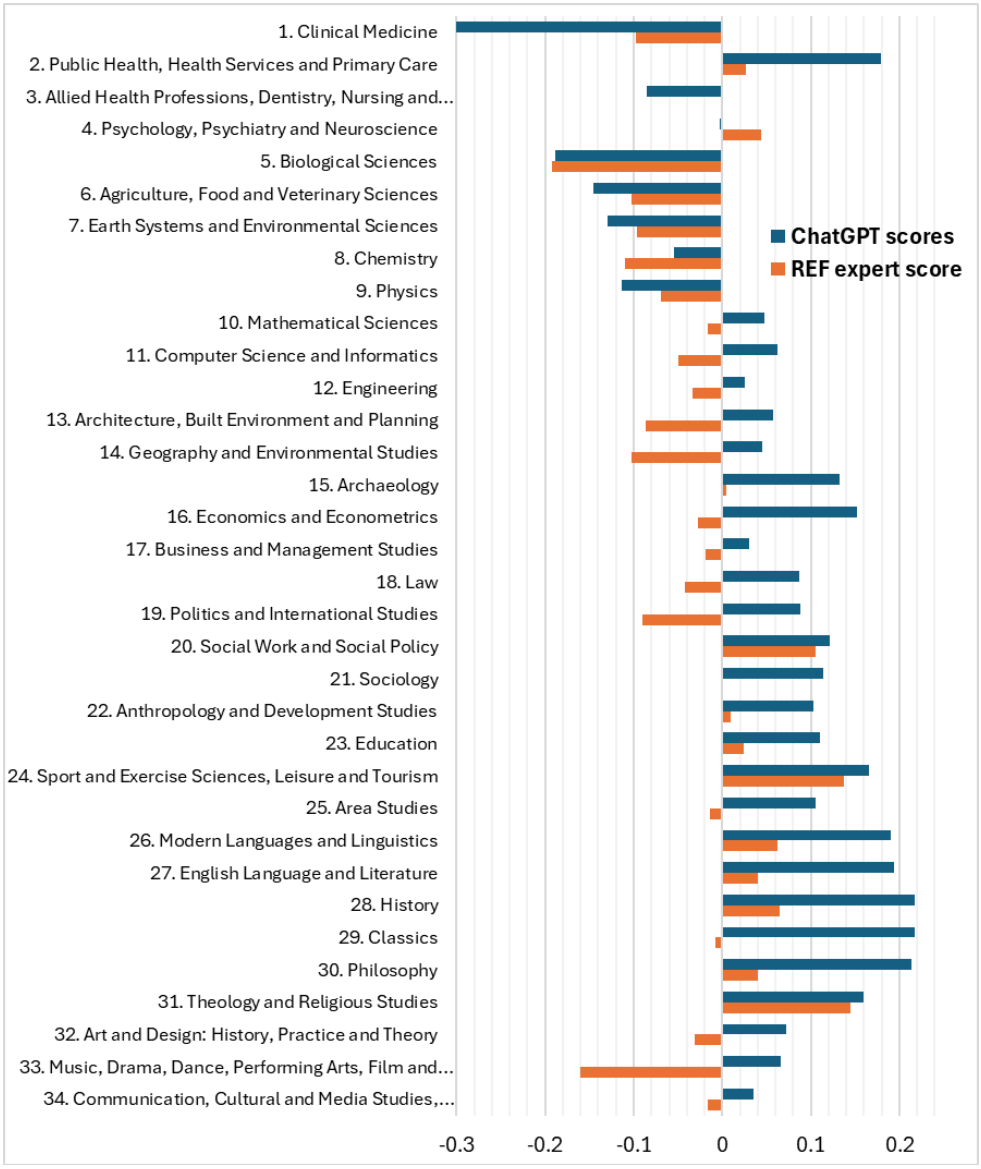


图3. 34个REF评估单元中，摘要长度（词数）与ChatGPT评分及REF院系平均专家评分之间的斯皮尔曼相关系数。

每句词数

每句词数在一定程度上反映了文本的句法复杂度，体现为平均句长。许多医学和科学、技术、工程和数学（STEM）评估单元中，ChatGPT评分与每句词数之间存在统计学显著的正相关；与 REF 院系平均专家评分相比，这些相关性通常更强、出现得也更频繁（图4）。然而，在若干艺术与人文学科中，REF 专家评分与句长呈负相关，而ChatGPT评分与句长的关系很弱或不存在，例如现代语言与语言学（–0.198）以及英语语言文学（–0.135）。这表明，部分 REF 专家评估可能更青睐句子较短、表达更清晰的写作。

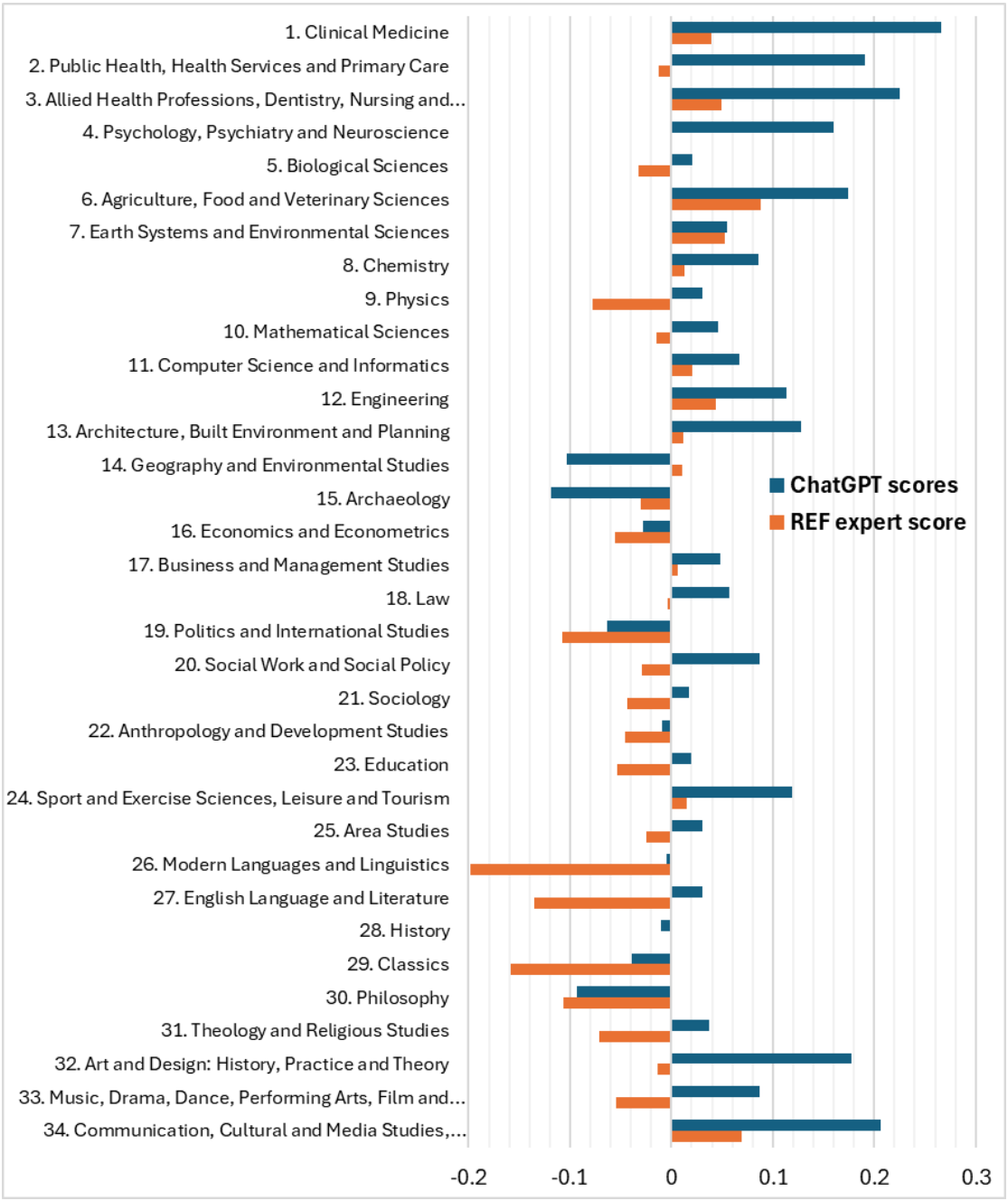


图4. 每句词数与ChatGPT和REF专家评分在34个REF评估单元中的斯皮尔曼相关系数。

每词音节数

每词音节数是一个简单的词汇复杂度指标。数值越高，通常表示文本中的词语越长或越专业。在大多数学科中，每词音节数较高的摘要往往比 REF 专家院系平均评分获得更高的 ChatGPT评分。在 34 个 REF评估单元中，有 28 个单元的 ChatGPT评分与每词音节数呈统计显著的正相关，而专家评分仅在 11 个单元中呈统计显著的相关性（图 5）。例如，在临床医学（0.273）、物理学（0.262）、化学（0.236）和工程学（0.210）中，每词音节数与 ChatGPT评分均呈正相关，而 REF 专家评分在这些学科中与每词音节数的相关性要弱得多。若干社会科学和人文学科也呈现类似模式，包括哲学（0.211）、历史学（0.190）、教育学（0.148）以及英语语言文学（0.145）；在这些学科中，REF 相关性普遍较弱、接近于零或为负。这表明，与专家评估相比，ChatGPT评分可能对摘要中的词汇复杂度更为敏感。

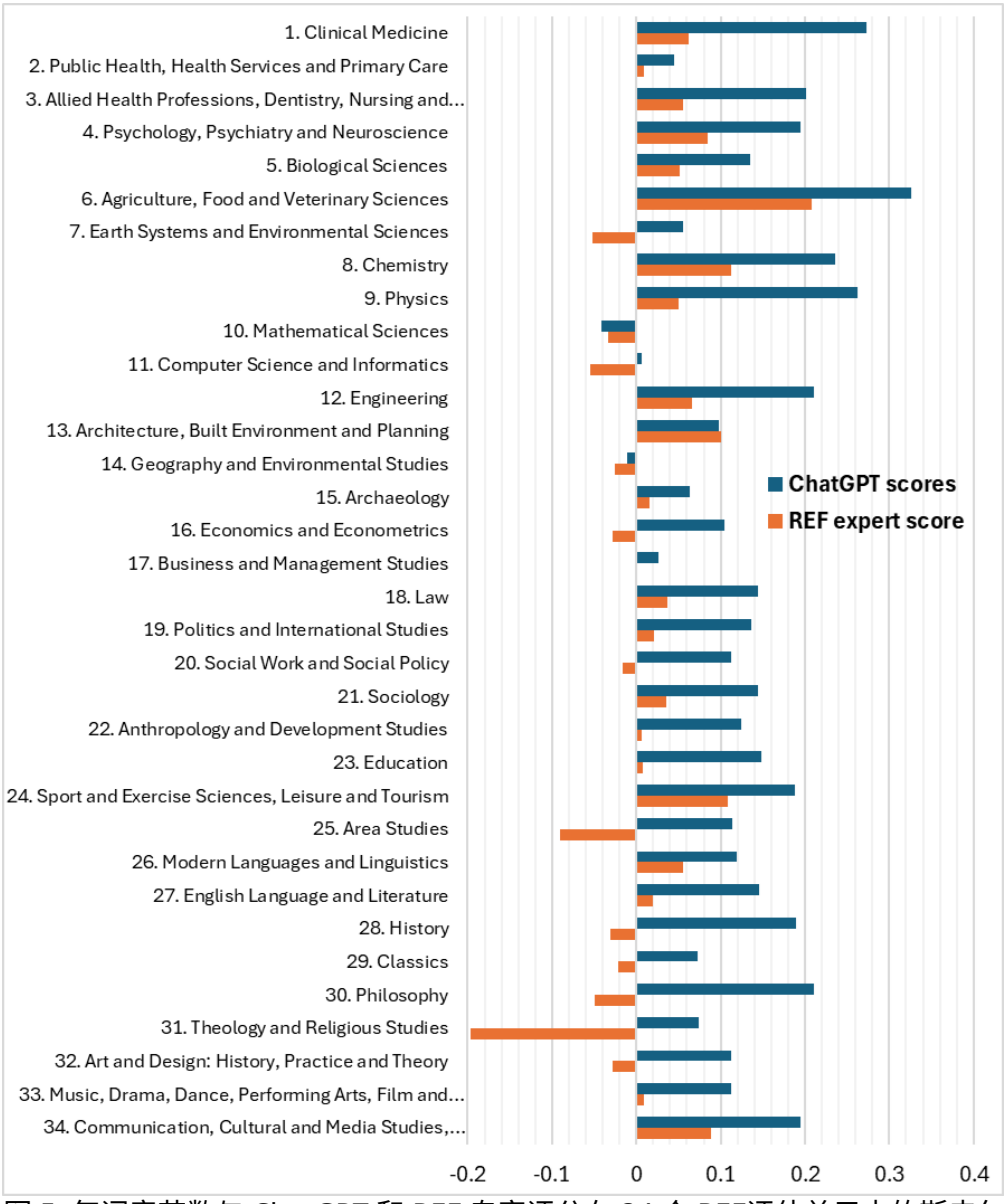


图 5. 每词音节数与 ChatGPT 和 REF 专家评分在 34 个 REF评估单元中的斯皮尔曼相关系数。

讨论

*ChatGPT*评分与引文影响力的比较

作为进一步的分析，我们考察了 ChatGPT评分和标准化对数变换引文得分（NLCS 2024）与摘要可读性（弗莱施-金凯德年级水平）之间的相关性。ChatGPT评分和 NLCS 均在单篇文章层面收集，因此可以直接比较这两个指标。

图6比较了34个REF评估单元中，ChatGPT评分和引文影响力评分与弗莱施-金凯德年级水平之间的相关性。在若干生物医学和STEM学科中，ChatGPT评分显然与较高的语言复杂度相关，但引文影响力并非如此：其相关性通常较弱，而且多为负相关。这表明，基于大语言模型的评估通常比基于引文的研究影响力指标对摘要复杂度更敏感。

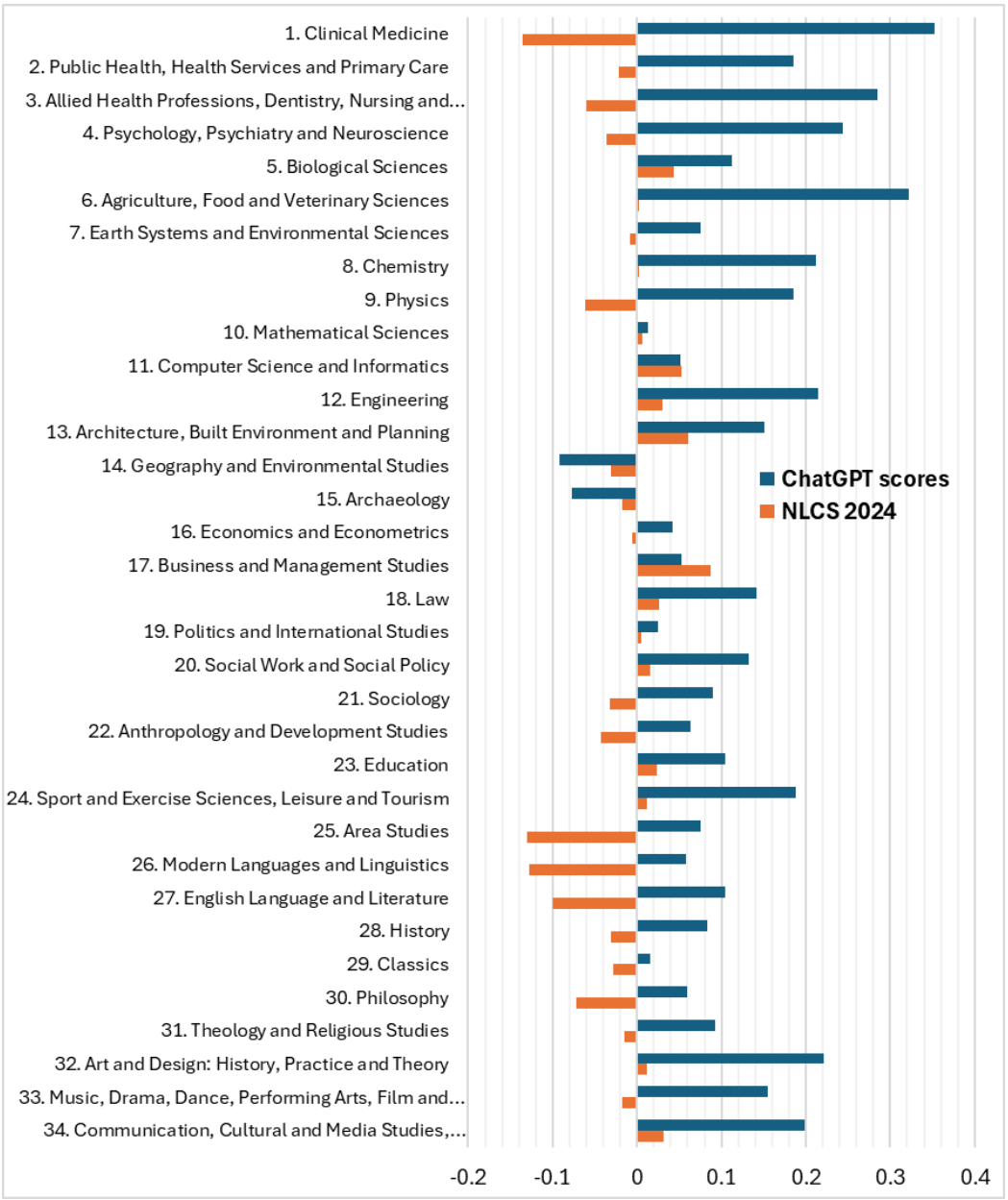


图6. Flesch–Kincaid 年级水平与 34 个 REF评估单元中 ChatGPT评分和引文影响力（NLCS 2024）之间的斯皮尔曼相关系数。

局限性

这项研究在数据、方法和解读方面存在若干局限。期刊文章均来自同一个国家，其他国家可能呈现不同的模式，尤其是在英语（科学领域的主导语言）并不通行的国家。由于各院系都会尽力提交最好的成果参加评估，这些文章也经过了质量筛选。这可能削弱了相关性的强度。另一项局限是，采用其他大语言模型、提示策略或文本指标可能会得到不同的结果。

遗憾的是，图2–4所用的更大数据集中没有单篇论文的 REF 专家评分。因此，我们使用院系平均 REF 分数作为单篇论文评分的替代指标。这种做法降低了分配给文章的专家评分的变异，因为同一院系的多篇论文会获得相同的平均分。因此，摘要特征与 REF 评分之间的相关性可能弱于使用单篇论文专家评分时的相关性，不过，对 UoA 1至6 的直接比较表明，这种差异不会很大。尽管如此，对 ChatGPT 与 REF 相关性的比较仍应谨慎解读，因为 ChatGPT评分是在文章层面获得的，而 REF 评分代表院系平均分。

另一项局限是，分析聚焦于关键文本特征和可读性特征，例如摘要长度和语言复杂度，而未考虑其他可能相关的特征，例如摘要结构（如背景、目标、方法和结论）或方法学细节。这些额外的文本特征也可能影响大语言模型的评分。

最后，本研究采用了相关性分析，无法揭示因果关系。例如，较长或语言复杂度较高的摘要未必会直接导致更高的 ChatGPT评分，而可能只是与其他会影响模型评估的学术写作特征相关。

结论

尽管因果关系尚未得到证实，但结果表明，相较于专家评估者，复杂风格的摘要更能给 ChatGPT 留下深刻印象，也就是说，ChatGPT 往往会给它们更高的评分。这一发现符合大语言模型进行文本模式匹配的特点。例如，ChatGPT 可能曾摄取过许多满怀希望的作者所写、尚未发表且质量低劣的稿件，因此“学会”将简单语法与薄弱的研究联系起来。也有可能是其历史数据中对刻意使用复杂语言的高度推崇，产生了同样的效果。更普遍地说，ChatGPT 可能已经发现，在学术界之外，人们容易被复杂语言打动。

研究发现还提出了一种可能性：作者或许会故意修改摘要的某些文本特征，以获得更高的大语言模型评估分数。如果大语言模型系统倾向于给篇幅更长或语言更复杂的摘要打出更高分，作者就可能调整这些特征来提高评分。如果使用大语言模型来支持同行评审或 REF 等国家评估中的研究评估流程，这可能会促使人们转而修改写作风格，而非提升研究本身的质量。这类动机带来的行为后果目前尚不明确。因此，确定哪些文本特征会影响大语言模型的评价，对于

理解操纵可能带来的风险，并推动大语言模型在研究评价中的负责任使用。

鉴于这些结果，似乎有必要降低语言复杂度对大语言模型评分的表面影响，从而（a）避免作者因语言表达而非研究成果获得奖励；（b）避免作者为了给大语言模型留下深刻印象而刻意采用更复杂的文风。这或许可以通过在提示指令中要求忽略语言风格，或在可能的情况下使用仅以相关学术文本训练的大语言模型来实现。理论上，给复杂文本施加评分惩罚以降低其大语言模型评分排名，或许可以根除这一问题；但这可能会不适当地干预评价过程，而且或许难以达成共识。

致谢

本研究部分由英国经济与社会研究理事会（ESRC）资助（UKRI2101）。

参考文献

Ante, L.（2022）。可读性与科学影响力之间的关系：来自新兴技术话语的证据。《信息计量学杂志》，16(1)，101252。Didegah, F., & Thelwall, M.（2013）。哪些因素有助于作者产出影响力最高的研究？合作、期刊与文献特征。《信息计量学杂志》，7(4)，861- 873。Flesch, R.（1948）。一种新的可读性衡量标准。《应用心理学杂志》，32(3)，221。https://comp311.wordpress.com/wp-content/uploads/2010/11/flesch_rudolph.pdfGazni, A.（2011）。高影响力文章的摘要是否更易读？考察来自全球顶尖研究机构的证据。《信息科学杂志》，37(3)，273-281。Germani, F., & Spitale, G.（2025）。来源框架会触发大语言模型中的系统性偏差。《科学进展》，*11*（45），eadz2924。Hafeez, D. M., Jalal, S., & Khosa, F.（2019）。影响引文的稿件特征的文献计量分析：对六种主要精神病学期刊的比较。《精神病学研究杂志》，108，90-94。Helbach, J., Wandscher, K., Pieper, D., & Hoffmann, F.（2025）。读者对较短与较长系统综述摘要的关注：一项随机对照试验。《英国医学杂志循证医学》。Hu, H., Wang, D., & Deng, S.（2021）。科学文献摘要写作风格与引文的分析。《在线信息评论》，45(7)，1290-1305。Jin, T., Duan, H., Lu, X., Ni, J., & Guo, K.（2021）。摘要可读性更高的研究论文是否会获得更多在线关注？来自《科学》的证据。《科学计量学》，126(10)，8471-8490。Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S.（1975）。为海军士兵推导新的可读性公式（自动化可读性指数、雾指数和Flesch阅读易读性公式）（编号 RBR875）。https://stars.library.ucf.edu/cgi/viewcontent.cgi?article=1055&context=istlibraryKostoff, R.（2007）。《柳叶刀》中高被引与低被引医学文章之间的差异。《科学计量学》，72，513-520。

Kousha, K., & Thelwall, M.（2024）。与被引次数更多或质量更高的期刊论文相关或可用于预测这些论文的因素：一篇《信息科学与技术年度评论》（ARIST）论文。《信息科学与技术协会会刊》，75(3)，215-244。Kousha, K., & Thelwall, M.（2025）。利用ChatGPT评估学术研究的社会影响：影响案例研究评价。《信息科学与技术协会会刊》，76(10)，1357-1373。Kundu, A., & Barbosa, D.（2024）。大语言模型是优秀的论文评分器吗？*arXiv*预印本 *arXiv:2409.13120*。Nunkoo, R. & Thelwall, M.（2025）。使用ChatGPT评估旅游研究质量。《旅游领域当前问题》。https://doi.org/10.1080/13683500.2025.2596265Sienkiewicz, J., & Altmann, E. G.（2016）。词汇与情感因素对科学论文受欢迎程度的影响。《英国皇家学会开放科学》，3(6)，160140。Sohrabi, B., & Iraj, H.（2017）。摘要中的关键词重复和期刊关键词频率对预测被引次数的影响。《科学计量学》，110(1)，243-251。Thelwall, M. & Yaghi, A.（2025）。在哪些领域，ChatGPT能够检测期刊文章质量？对REF2021结果的评估。《信息管理趋势》，13(1)，1-29。Thelwall, M.（2017）。用于研究评价的三种实用领域标准化替代指标公式。《信息计量学杂志》，*11*（1），128-151。Thelwall, M.（2025a）。使用大语言模型评价研究质量：分析ChatGPT在不同设置和输入条件下的有效性。《数据与信息科学杂志》，10(1)，7-25。https://doi.org/10.2478/jdis-2025-0011Thelwall, M.（2025b）。Google Gemini在评价研究质量方面是否胜过ChatGPT？《数据与信息科学杂志》，10(2)，1–5。https://doi.org/10.2478/jdis-2025- 0014，扩展版见：https://doi.org/10.6084/m9.figshare.28089206.v1Thelwall, M.（2025c）。在哪些领域，ChatGPT评分比引文更能体现研究质量？arXiv预印本 arXiv:2504.04464。Thelwall, M.（2025d）。ChatGPT能否取代引文来评价学术文章和期刊的质量？来自图书馆与信息科学领域的实证证据。《文献学杂志》，81(4)，1078-1094。Thelwall, M.（2026）。大语言模型了解期刊文章的基本事实吗？《文献学杂志》。https://doi.org/10.1108/JD-11-2025-0330Thelwall, M., & Kurt, Z.（2025）。使用ChatGPT进行研究评价：是否存在年龄、国家、长度或领域偏见？《科学计量学》，130(10)，5323-5343。https://doi.org/10.1007/s11192-025-05393-0Thelwall, M., & Mohammadi, E.（2026）。小型及推理型大语言模型能否对期刊论文的研究质量进行评分？平均评分和少样本提示是否有帮助？《科学计量学》。arXiv预印本 arXiv:2510.22389。Van Wesel, M., Wyatt, S., & ten Haaf, J.（2014）。一个冒号带来的差异：表层因素如何影响后续引文。《科学计量学》，98(3)，1601-1615。Vasu, S. S. M., Sheth, I., Wang, H. P., Binkyte, R., & Fritz, M.（2025）。评判中的公正：揭示大语言模型辅助同行评审中的（隐性）偏见。arXiv预印本 arXiv:2509.13400。Wang, S., Liu, X., & Zhou, J.（2022）。语言与语言学领域的可读性正在下降。《科学计量学》，127(8)，4697-4729。Zhu, Y., Haunschild, R., Ye, J., & Xie, H.（2026）。生成式人工智能能否有效地对社会科学领域进行质量评价？一项图书馆与信息科学领域的案例研究。《科学计量学》，1-21。

附录：附加的可读性分析

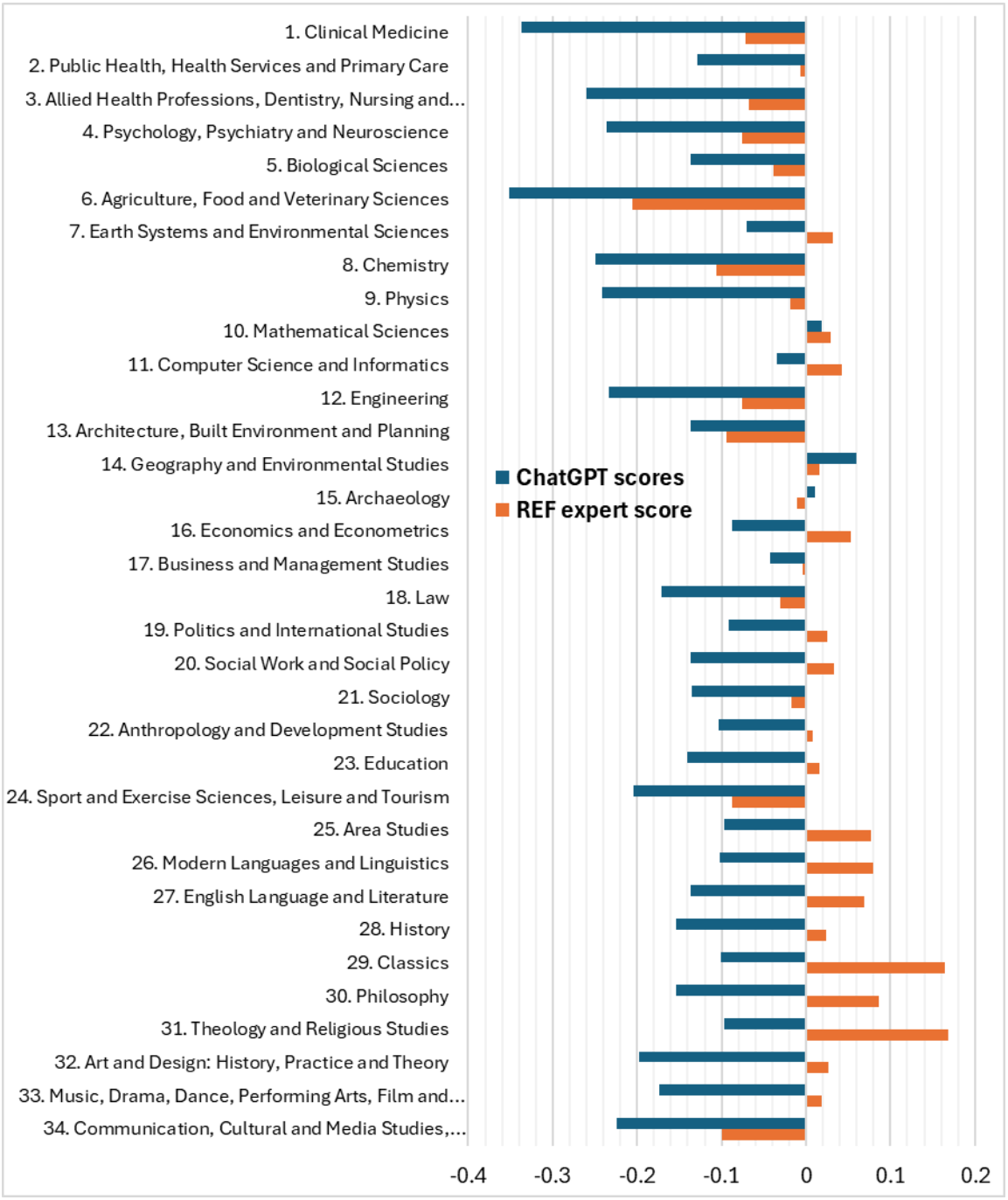


图 S1。Flesch 阅读易读度与 ChatGPT 及 REF 专家评分在 34 个 REF评估单元中的斯皮尔曼相关系数。Flesch 阅读易读度评分越高，表示摘要越易于阅读（Flesch，1948）。正相关表示可读性更高的摘要往往获得更高的评分，而负相关表示可读性较低（语言复杂度更高）的摘要往往获得更高的评分。

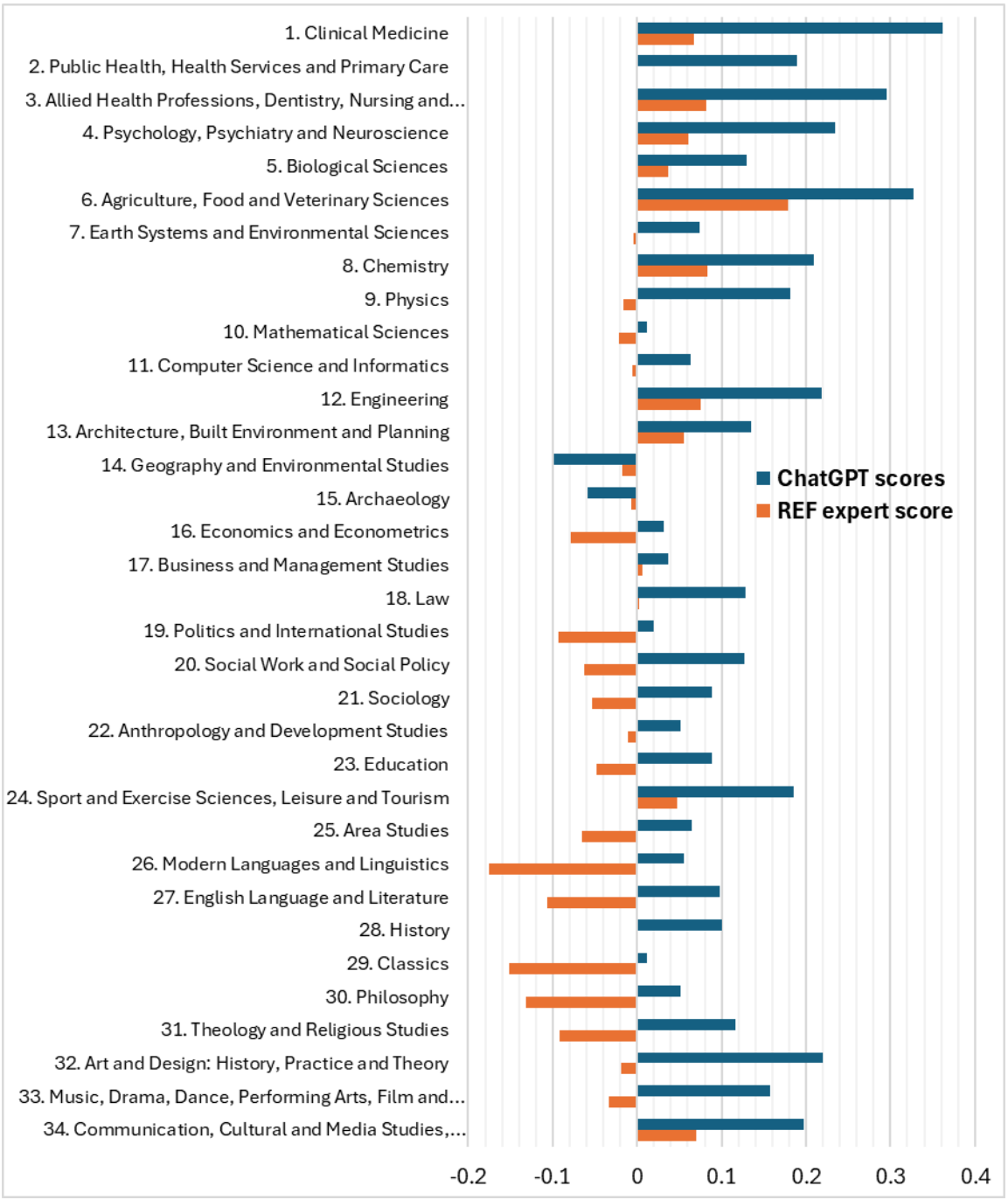


图 S2. 34 个 REF评估单元中，**冈宁雾指数**可读性评分与 ChatGPT 和 REF 专家评分之间的斯皮尔曼相关系数。冈宁雾指数评分越高，表示摘要的语言复杂度越高。正相关表示语言较复杂的摘要往往获得较高评分，而负相关则表示较简单的摘要往往获得较高评分。