高水平论文开放同行评议意见分析及启示*
——以H1 Connect为例

刘晓娟　于姚　沈嘉宁
（北京师范大学政府管理学院，北京 100875）

摘要：开放同行评议为论文评价提供了新的视角，分析高水平论文的评议意见特征，有利于评判优质论文、优化学术评价体系。以被全球生物医学领域权威开放同行评议平台H1 Connect专家推荐的、发表在顶级期刊上有关COVID-19的论文为高水平论文的典型案例，爬取这些论文在H1 Connect中的开放同行评议意见，从指标和文本两个层面分析专家评议意见所体现的对高水平论文的关注和认可程度、情感特征以及价值特征。研究发现，高水平论文能快速吸引专家的关注，但仅有少数引发热议。开放同行评议文本中存在一定数量的情感句，明确体现了专家的观点态度：正向评价居多，主要针对论文的总体表现；负向评价较少，主要针对论文采用的研究方法。开放同行评议意见认为论文的价值体现在学术领域和实践应用的多个维度，应用贡献更为突出。通过对高水平论文开放同行评议意见的分析，建议在学术评价体系中引入开放同行评议数据源，以提升学术评价的科学性和有效性。实际应用中，仍需进一步完善同行评议指标，深入挖掘评议文本中的关键要素，加强资源建设与技术融合，利用前沿技术智能解析评议文本的深层价值。

关键词：开放同行评议；高水平论文；学术评价；H1 Connect
中图分类号：G250　DOI：10.3772/j.issn.1673-2286.2025.01.007
引文格式：刘晓娟，于姚，沈嘉宁. 高水平论文开放同行评议意见分析及启示：以H1 Connect为例[J]. 数字图书馆论坛，2025，21（1）：55-66.

论文是科研活动阶段性成果的主要载体，对论文进行分析和评价是把握学术动态和研究规律的重要手段[1]。目前论文评价主要依赖文献计量法，虽然相对客观高效，但存在忽视论文内容、时滞过长、引用动机不明确等缺点。替代计量学的发展拓宽了论文的评价维度，但其数据来源缺乏专业度，且容易受到人为操纵的影响。如何构建更加科学、规范的论文评价体系一直受到各界关注，其中一种解决思路是将定量评价与定性评价相结合，使得评价结果更加科学准确。然而，在传统同行评议模式下，这种方法始终面临同行评议意见难以获取的挑战。开放同行评议是源于开放科学运动的一种新兴的同行评议方法[2]，尽管其概念尚未形成统一定义，但其核心思想主要是开放作者和审稿人身份、开放评议流程、开放评语等[3]。通过愈加透明的学术出版过程监督评审工作，开放同行评议有效提升了同行评议结果的专业性、透明度和可信度，为论文评价提供了新视角和有力的数据支持。

在众多开放同行评议平台中，H1 Connect（原F1000Prime、Faculty Opinions）是全球生物医学领域公认的权威平台之一，汇集了近万名领域内顶尖的专家，旨在对经过传统同行评议认可后正式发表的论文进行进一步的推荐和评价，仅有少数杰出的研究成果能够获得专家的一次或多次推荐。

顶级期刊（以下简称“顶刊”）一般指由学术共

收稿日期：2024-10-29
*本研究得到国家社会科学基金一般项目“公众信任视角下科研成果社会影响的分析与评价研究”（编号：23BTQ058）资助。

同体成员一致认可的代表本领域最高水平的学术期刊[4]，在职称评定、绩效考核、学科评估、资源配置等方面发挥着重要作用。发表于顶刊的论文若再次被H1 Connect专家推荐，越发凸显其重要的学术价值和现实意义，一定程度上可作为高水平论文的典型代表。H1 Connect专家对这些高水平论文的评议意见是在论文发表后基于内容作出的权威评价，能够为进一步评判论文价值提供重要参考，有必要对其进行深入挖掘和充分利用。因此，本研究以得到H1 Connect专家推荐的顶刊论文为例，从指标和文本两个层面分析高水平论文的开放同行评议意见，深化对发表后开放同行评议意见的认识和理解，全面揭示同行评议专家视角下的论文价值，以期为评判高水平论文、完善学术评价体系提供参考。

1 相关研究

同行评议意见来自相关领域的专家，专业度和可信度较高，蕴含着丰富的价值。近年来，越来越多的期刊和会议开始采用开放同行评议机制，H1 Connect、Publons、OpenReview等开放同行评议平台也应运而生，大大降低了同行评议意见的获取难度，同行评议意见的数据类型也更加丰富。除了文本形式的评议意见，还包括评价分数、评价标签等多种形式，很多学者针对不同类型的同行评议意见进行研究。

（1）开放同行评议指标与被引频次、AAS（Altmetric Attention Score）的相关性分析。谭贝加[5]以Altmetric Top100论文为研究对象，探讨被引频次、AAS和F1000评分用于综合评价生物医学论文的可行性，发现F1000评分与被引频次弱相关，AAS与F1000评分及被引频次均不相关。许丹等[6]以Faculty Opinions中临床实践类和基础理论类学科为例，分析发现不同学科论文的多数开放同行评议指标与被引频次、AAS呈现低相关或不相关。

（2）不同标签下论文指标的差异性分析。H1 Connect的专家在推荐论文时需要从给定的标签库中挑选任意一个或多个标签，以简要概括自己的推荐理由，具体包括：New Finding（新发现）、Interesting Hypothesis（有趣的假设）、Novel Drug Target（新的药物靶点）、Technical Advance（技术进步）、Good for Teaching（有益于教学）、Changes Clinical Practice（改变临床实践）、Confirmation（证实之前的数据或假设）、Refutation（反驳之前的数据或假设）、Controversial（争议性的发现）、Negative/Null Results（结果与预期不符或无显著统计学意义）。一些学者研究该平台不同标签下论文评价指标的表现差异。Bornmann[7]研究发现具有Good for Teaching标签的论文确实比没有该标签的论文获得了更高的Altmetric计数，具有New Finding标签的论文被引频次更高。Du等[8]、姜育彦等[9]研究发现具有Interesting Hypothesis、Controversial、Novel Drug Target标签的变革性研究论文被同行专家高度评价的可能性更大，但被引频次却比具有其他标签的论文少；具有Confirmation、New Finding、Technical Advance等标签的循证性研究论文更有可能被大量引用，但被同行专家高度推荐的可能性很小。

（3）同行评议文本的情感分析。同行评议文本中往往蕴含着专家的情感倾向，能够反映专家对论文的整体态度，因此，情感分析在同行评议文本分析中得到较多应用。张明阳等[10]利用ELECTRA模型，从创新性、动机、实验设计、相关工作的完整性以及论文的写作表达5个方面自动标注同行评议文本的情感极性。Ghosal等[11]构建DeepSentiPeer模型，根据同行评议文本和作者回复中的情感信息，预测稿件分数及录用情况。有学者进一步将同行评议文本的情感极性与文献计量指标相结合。Zong等[12]探究了评议意见情感极性与被引频次之间的关系，发现：得到正向评议的论文被引频次明显高于对照组（未被评议的论文）；得到中性、负向评议和同时得到正向、负向评议的论文与对照组的被引频次没有显著差异。

（4）基于同行评议文本的特征识别。同行评议文本的语言特征、关键要素等也受到了学者们的关注。Zhang等[13]以《英国医学杂志》所刊论文的评议意见为研究对象，分析意见长度、词语分布和评语位置的差异，发现初级研究员在赞扬稿件和作者时比高级研究员更频繁地使用however。Wang等[14]对F1000Prime平台的开放同行评议意见进行分析，发现专家常用的词语包括interesting、important、first、exceptional等。秦成磊等[15]将评议意见要素划分为正向评价、负向评价、要求/建议（主要：涉及方法改进、材料补充等；次要：涉及单词拼写、数值更正等）、问题/疑问、陈述5个类别，并对比分析了传统机器学习模型、深度学习模型的识别效果。Ghosal等[16]从4个层面划分同行评议意见，即评议

意见对应的论文章节（如引言、方法、数据等）、评议的角度（如新颖性、原创性、清晰性等）、评议意图（如建议、讨论、疑问等）、评议意见的重要程度（如指出研究缺陷、格式问题、没有明显立场的一般叙述等）。梁帅等[17]对F5000平台的专家评议意见进行知识图谱和共词分析，总结出优秀论文在创新、价值、内容及写作方面的特征。

综上，开放同行评议意见的专业性和可靠性，以及数据获取难度的降低，使得越来越多的学者开始对其进行分析和利用。然而，已有研究的数据源主要集中于发表前同行评议意见，这些意见主要关注论文质量，意在指出论文的问题并决定其是否能够发表。而发表后开放同行评议则对已正式发表的论文进行进一步评估，其意见往往更侧重评估论文的价值和影响。已有针对发表后开放同行评议意见的研究主要探究其数值指标，较少关注蕴含专家深度见解的同行评议文本。为了深入挖掘和充分利用发表后开放同行评议意见，本文以H1 Connect为例，同时考虑指标和文本两个层面，并在常见情感特征的基础上进一步从价值特征角度挖掘同行评议文本，主要包含以下两个研究内容：①分析开放同行评议专家视角下的高水平论文评议意见具有哪些特征；②提出开放同行评议意见在评判高水平论文、优化学术评价体系中的应用建议。

2　研究设计

2.1　研究对象选择及数据获取

与COVID-19相关的论文直接关系到疫情防控与生命健康，在学术、政策、健康、经济和社会等多个维度产生了广泛影响。考虑到COVID-19论文能充分体现出医学研究的重要价值且与H1 Connect涉及的学科领域高度契合，选取得到H1 Connect专家推荐的有关COVID-19的项刊论文作为高水平论文的典型案例。通过分析这些论文揭示重大公共卫生事件中高水平论文的特征及价值，为多维度评判科学研究的价值和影响力提供理论支持和方法参考。

选取全球公认的三大顶刊——《细胞》《自然》《科学》，以及四大顶级医学期刊——《新英格兰医学杂志》《美国医学会杂志》《英国医学杂志》《柳叶刀》作为样本来源。首先，在Web of Science核心合集中检索发表在以上7种选定期刊上题名包含“COVID-19”“Novel coronavirus”“2019-nCoV”“SARS-CoV-2”“coronavirus 2”“Coronavirus disease 2019”“Corona virus disease 2019”的论文，时间限定为2020年1月—2023年9月，文献类型限定为“Article”或“Review”。接下来，根据DOI在Altmetric.com中获取论文的替代计量指标数据，在InCites平台获取论文的学科规范化引文影响力指标（CNCI），并利用自编Python程序爬取H1 Connect的开放同行评议意见，数据获取时间为2024年11月。最终得到1 012篇COVID-19项刊论文，其中252篇得到H1 Connect专家推荐（论文集R），760篇未被推荐（论文集N_R）。各期刊发表的COVID-19论文数量及H1 Connect推荐情况如表1所示。整体来看，《细胞》《自然》《科学》上发表的论文数量要多于四大医学期刊；近1/4（24.90%）的论文得到H1 Connect专家推荐。

表1　各期刊发表的COVID-19论文数量及H1 Connect推荐情况

期刊名称	COVID-19论文数量/篇	H1 Connect推荐论文数量/篇	H1 Connect推荐占比/%
自然	191	48	25.13
科学	173	50	28.90
细胞	166	42	25.30
英国医学杂志	134	11	8.21
新英格兰医学杂志	130	42	32.31
柳叶刀	126	34	26.98
美国医学会杂志	92	25	27.17

2.2　研究思路

基于H1 Connect的开放同行评议意见，从指标和文本两个层面进行分析。基于多种指标分析论文在不同评价方法下的关注和认可程度，并深入探究开放同行评议文本所呈现出的情感特征和价值特征，在此基础上探讨开放同行评议意见在高水平论文评判与学术评价体系优化中的应用建议。

（1）开放同行评议专家推荐论文在不同评价方法下的关注和认可程度分析。选取同行评议（指标及具体含义详见表2）、CNCI和AAS这3类评价指标表征不同评价方法下对论文的关注和认可程度。引文分析侧重专业领域内的学术影响力，替代计量学重视公众领域的社会影响力，同行评议则是从专家视角深入评价论文内

表2　同行评议指标及具体含义

指标	含义
评价星级（RStar）	专家对论文进行评级，分为Good（1星）、Very Good（2星）和Exceptional（3星），所有专家的评级加权构成最终的RStar指标
评价次数（RNumber）	论文在H1 Connect中被评价的次数
评价时滞（RDelay）	H1 Connect中首条论文评价发布的日期与论文发表日期的间隔时间（天）

注：RStar直接来源于H1 Connect，RNumber和RDelay是本研究构建的衍生指标。

容。这些评价方法各有侧重，对论文的评价结果也会有所差异。通过对比分析不同指标，探究同行评议意见在论文评价中的独特作用和价值。

（2）开放同行评议文本的情感态度分析。同行评议文本的情感特征可以反映专家对论文的观点态度，通过分析情感得分、评价对象及相应情感词，深入揭示专家的情感强度、关注重点及表达特点。

（3）开放同行评议文本所呈现的论文价值特征分析。同行评议文本中对于论文价值的洞察和评价源自权威专家，是挖掘论文价值的重要参考。围绕贡献类型和创新程度两个维度分析评议文本提及的论文价值要素，为更加科学合理地认识论文内在价值，进而对其进行识别和评价提供参考。

3　开放同行评议指标分析

3.1　专家推荐论文的引用与社会关注

鉴于专家的专业性和权威性，其推荐论文更可能获得大量学者引用和社会关注。为探究这一推测是否成立，本研究选取H1 Connect专家推荐论文（论文集R，252篇）和未被推荐论文（论文集N_R，760篇）进行比较，分析这两个论文集的CNCI和AAS。

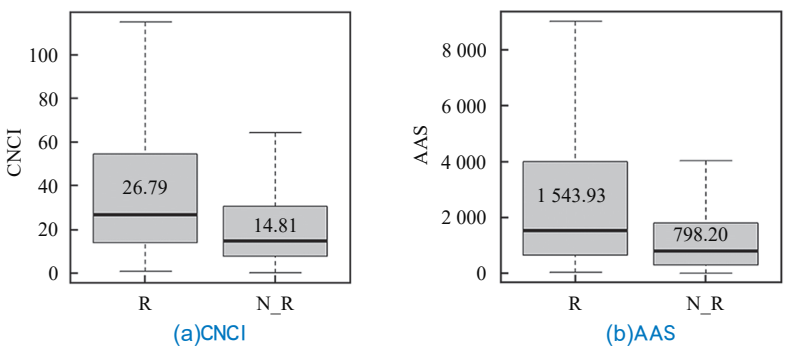
（1）CNCI、AAS分布。CNCI是将论文的被引频次除以同出版年、同学科领域、同文献类型论文的平均被引频次得到的，能够消除论文所属学科、出版年份以及文献类型对被引频次的影响。论文集R中，CNCI的分布范围是0.61~590.67，平均值是51.59。84.52%的论文CNCI在10以上，11.51%的论文CNCI在100以上。AAS的分布范围是34.35~31 549.12，平均值是3 308.21。绝大多数（98.41%）论文AAS在100以上，8.33%的论文AAS极高，在10 000以上。

（2）论文集R与N_R对比。进一步检验发现两个论文集中，CNCI、AAS均存在显著性差异（p值均小于0.001）。为了更直观地分析这种差异，分别绘制两个论文集CNCI箱线图和AAS箱线图（见图1），因极差过大，未显示异常值。数据总体呈右偏分布；从中位数看，论文集R的CNCI和AAS约为论文集N_R的2倍；论文集R的CNCI和AAS分布较为离散，而论文集N_R的分布相对集中。

综上可见，专家推荐论文在学者和社会公众中受到青睐，因此，在论文发表早期计量指标尚不明确时，可基于专家推荐遴选潜在的高影响力研究成果。

3.2　专家推荐论文的同行评议表现

（1）RNumber可以反映专家对论文的关注度。如表3所示，超过7成的论文仅获得了1次评价，近1/10的论文获得2次以上的评价。Waltman等[18]指出生物医学领域约有2%的论文能够获得H1 Connect专家的推荐，可见即使绝大多数被推荐的COVID-19项刊论文RNumber较少，专家推荐依然能反映出其杰出的学术水平。很少有论文获得很多位专家的推荐，RNumber为



(a)CNCI　(b)AAS

图1　CNCI和AAS箱线图

表3　RNumber分布

RNumber范围/次	论文数量/篇	占比/%
1	185	73.41
2	43	17.06
3	13	5.16
4	6	2.38
[5，9]	5	1.99

5次及以上的论文仅占比1.99%，主要涉及新冠发病机制、人体免疫机制以及临床疗法等主题。

（2）RDelay可以反映专家关注论文的及时性。如表4所示，近1/5的论文能够较快得到专家关注，在发表7天内获得首次评价，超过半数（56.74%）的论文能够在发表1个月内获得首次评价，极少数论文在发表1年后才获得首次评价。已有研究[9,19]指出，H1 Connect的专家通常会在论文正式发表后的两个月左右对本领域的论文进行推荐，约42%的推荐发生在论文发表后的首月内。与之相比，本研究中COVID-19项刊论文被推荐的及时性更高，表明在这一领域，专家对相关研究关注更为迅速，这可能与疫情的紧迫性密切相关。

表4　RDelay分布

RDelay范围/天	论文数量/篇	占比/%
[0，7）	49	19.44
[7，30）	94	37.30
[30，60）	45	17.86
[60，90）	31	12.30
[90，365）	31	12.30
[365，515]	2	0.80

（3）RStar可以反映专家对论文内容和价值的认可度。如表5所示，RStar普遍较低且较为集中，82.54%的论文分布在[1，5）区间内，少数论文处于[15，36]的区间内。

表5　RStar分布

RStar范围	论文数量/篇	占比/%
[1，5）	208	82.54
[5，10）	34	13.49
[10，15）	6	2.38
[15，36]	4	1.59

3.3　不同亮点的论文在3类评价方法下的表现

H1 Connect的专家赋予论文的标签在一定程度上揭示了论文的亮点。论文集R中，平均每篇论文标有2个标签，最多标有6个标签。为探究不同亮点的论文在3类评价方法下的表现是否具有一定特点，比较分析不同标签论文评价指标平均值（见表6）。

（1）Changes Clinical Practice论文的平均表现最为突出。尽管数据集中这类论文的数量比较少，但这4篇论文在开放同行评议方法下的RStar、RDelay等指标都优于论文集R的平均水平，体现出专家对这类论文的一致认可。在引文和替代计量学评价方法下的表现存在差异，其中1篇关于“地塞米松在新冠住院患者中的应用”的论文表现尤为出色，CNCI和AAS在论文集R中分别排名第2位和第39位，一定程度上拉高了4篇论文的整体均值。

（2）开放同行评议专家推荐有助于快速识别在引文和替代计量学评价方法下价值尚未凸显的论文。标有Refutation、Controversial或Negative/Null Results的论文的CNCI、AAS、RStar均值处于较低水平，但能

表6　不同标签论文指标平均值

标签	论文数量/篇	占比/%	CNCI	AAS	RStar	RDelay/天	RNumber/次
Changes Clinical Practice	4	1.59	145.21	4 259.67	4.50	10.50	1.75
Refutation	6	2.38	24.22	1 697.01	3.50	17.33	1.83
Controversial	21	8.33	48.24	2 760.38	3.33	26.95	1.67
Negative/Null Results	21	8.33	34.34	2 878.89	2.33	30.00	1.48
Technical Advance	43	17.06	59.15	3 756.75	4.91	40.63	1.79
Novel Drug Target	48	19.05	68.76	1 918.87	6.44	31.33	2.25
Confirmation	63	25.00	60.53	3 819.22	4.13	25.81	1.71
Interesting Hypothesis	65	25.79	52.74	2 729.98	5.42	42.71	2.08
Good for Teaching	117	46.43	61.43	4 119.48	4.04	35.49	1.68
New Finding	199	78.97	56.76	3 432.32	3.64	47.54	1.52
论文集R均值			51.59	3 308.21	3.31	45.39	1.46

够相对较快地得到开放同行评议专家的关注。这些论文往往与现有观点相悖，挑战了已有的认知，现阶段虽遭遇“冷落”，但随着研究的深入，有可能在未来发挥重要作用。

（3）专业性强的论文存在更高的知识壁垒，在替代计量学评价方法下表现欠佳。标有Novel Drug Target的论文的CNCI、RNumber和RStar相对较高，但AAS明显低于论文集R的平均水平。这类论文一般是较为前沿、新颖的内容，专业性更强，其价值更易被相关专家和学者识别，其社会领域传播会受到公众自身知识水平的制约。这印证了单一评价方法可能存在局限性，综合多元化的指标有助于全面、准确地评估论文价值。

4 开放同行评议文本分析

同行评议文本是对论文创新性、方法科学性、数据可靠性、探索深度和质量等方面的权威评估，往往蕴含着专家的情感倾向以及对于论文价值的重申和强调，评价角度丰富，立场相对客观，是开展论文评价的重要素材。深入挖掘开放同行评议文本，分析其所体现的情感特征以及对于论文价值特征的评判，不仅可以为情感与价值等要素的快速识别提供新的特征依据，也可以为学术论文评价体系的完善与创新提供新的思路方向。

4.1 情感特征

已有研究[20]表明在分析同行评议文本时，相比算法（TextBlob和VADER），人工分析可得出更可靠的结果。为深入挖掘隐含在同行评议文本中的专家观点态度，采用人工标注的方式提取评议文本中的情感句，将蕴含赞扬或批评含义的单词或词组都标注为广义的情感词，鉴于否定词出现次数较少且不包含多重否定等复杂结构，不再单独考虑no、not等否定词，将其整合为负向情感词组的一部分进行提取[21]，最终得到87个正向情感词和23个负向情感词。综合考虑情感词的语境及其在SentiWord词典中的分值，将情感词划分为3个等级，其中正向情感词设为1分或2分，负向情感词设为−1分，情感词提取示例如表7所示。

情感分值的量化需要考虑程度词，本研究共提取

表7 情感词提取示例

例句	情感词分数
This was an excellent study……	2
The results are interesting……	1
This was a good observational study that had a few limitations……	1；−1

注：下划线标注词为情感词。

出4个程度词，分别为very、especially、particularly、really。结合具体语义，并参考文献[22]中的程度词及其量化分值，将4个程度词的程度级别统一设为1.5。计算每条同行评议文本的正向情感得分S_P和负向情感得分S_N。S_P的计算公式如式（1）所示。

$$S_P=\sum_{i=1}^{t}W^{degree}W^{emotion} \quad (1)$$

式中：t表示评议文本包含该方向情感词的总数；W^{degree}表示程度词分值，若有程度词修饰，则W^{degree}=1.5，否则W^{degree}=1；$W^{emotion}$表示情感词分值。S_N计算方式同理。情感得分分布情况如表8所示。

表8 情感得分分布

正向情感			负向情感		
情感得分	评议文本数量/条	占比/%	情感得分	评议文本数量/条	占比/%
[1，4）	156	80.83	−1	19	51.35
[4，8）	34	17.62	−2	10	27.03
[8，14]	3	1.55	−3	8	21.62

（1）情感得分。从情感极性来看，368条同行评议文本中，共有207条（56.25%）包含情感词，其中绝大部分包含正向评价，少量包含负向评价，这与Wang等[21]分析结果一致，也契合H1 Connect作为论文推荐平台的定位。大部分评议文本的情感强度较低，包含正向情感词的评议文本中，超过80%的S_P范围为[1，4）；包含负向情感词的评议文本中，S_N为−1的最多，且最低不低于−3。仅有3条评议文本中蕴含着强烈的情感倾向，最高为14分，这条得分最高的评议文本所对应的论文共收到4次评价，在其余3位专家的评价中，情感得分的均值仅为3分。综上可见，专家评议文本情感倾向相对不明显，但也存在一定数量的情感句明确体现出专家的观点态度。专家情感表达的谨慎性进一步凸显了情感句的分析价值，在利用其评价论文时，应注意评议专家个人习惯与偏好等潜在因素的影响。

（2）评价对象及相应情感词。同行评议文本中包

含大量论文的细节信息，分别对应论文的不同方面，如方法合理、结果有趣等。为了深入了解专家的情感表达，进一步对情感句进行细粒度挖掘。参考同行评议文本与论文结构划分的相关研究[13,16,23]，将情感句的评价对象划分为研究问题、方法、结果、讨论、总体表现5个部分，划分说明及示例如表9所示。

表9　评价对象划分说明及示例

评价对象	说明	示例
研究问题	对论文的研究问题进行评价	This study raises intriguing questions about the impact of FcgRIIb on innate immune responses
方法	对论文数据收集、实验开展、数据分析的过程进行评价	The authors must be commended for carrying out a well-designed multicentered trial at the peak of the pandemic
结果	对研究得到的具体结果进行评价，如实验数据、图表或统计分析结果等	Interestingly, there are also cross-reactive CD4 cells in 40–60% of subjects not exposed to COVID-19
讨论	对研究结果的解释分析、研究主要结论及重要性进行评价	Unfortunately, the authors provided no explanation as to why these cross-reactive antibodies were non-protective, even for the ones that could bind to the spike RBD, which is a critique of this study
总体表现	不局限于论文的特定部分，而是对总体工作进行评价	This is a very important study on many levels detailing the immune response to SARS-CoV-2

注：下划线标注处为评价对象。

情感句评价对象的分布情况如图2所示，可见专家情感表达更加关注论文的总体表现、讨论和结果部分。负向情感的流露虽然较少，但更具针对性，倾向于指出论文方法（54.72%）上的不足，对于论文总体表现（1.89%）进行批评的情况较少，而正向情感表达多涉及论文的总体表现（44.68%）。这与Ghosal等[16]“总体部分收到更多的正向评价，审稿人对方法部分更为挑剔”的观点相符。

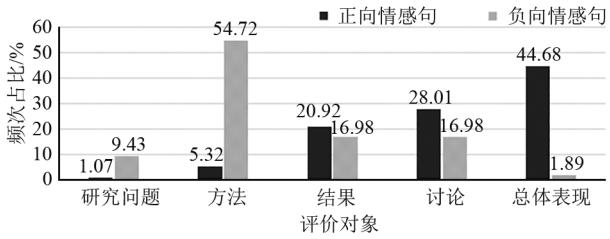


图2　情感句评价对象分布

情感词能够鲜明地展现专家对论文的具体态度和看法。为此，进一步从情感词的维度进行分析。开放同行评议文本常用情感词如表10所示。专家表达正向情感时，主要对论文进行赞扬（如good、excellent），并肯定其重要性（如importan*、essential）、新颖性（如

表10　开放同行评议文本常用情感词

评价对象	正向情感词（频次）	负向情感词（频次）
研究问题	**essential**（1）、**importan***（1）、**intriguing**（1）	not answer（2）、leave some aspects/topics untouched（1）、limit*（1）、not address（1）、not clear（1）
方法	good（2）、**importan***（2）、**impressive***（2）、**promising**（2）	limit*（23）、caveat（3）、uncertain*（3）、confounding（2）、bring into question（1）、controversial（1）、lack of（1）、not check（1）、problematic（1）、weakness（1）
结果	**importan***（17）、interest*（13）、**encouraging***（4）、informative（2）、key（2）、of note（2）、remarkabl*（2）、**significant**（2）、surprising*（2）	limit*（5）、inconclusive（2）、uncertain*（2）、lack of（1）
讨论	**importan***（18）、new（8）、interest*（7）、**critical**（4）、**novel**（4）、useful（4）、elegant*（2）、**essential**（2）、**excellent**（2）、informative（2）、invaluable（2）、notabl*（2）、promising（2）、**significant**（2）、**strong**（2）	a grain of salt（1）、critique（1）、disconcerting（1）、not discuss（1）、inconclusive（1）、no explanation（1）、not answer（1）、not clear（1）、uncertain*（1）、unfortunately（1）
总体表现	**importan***（21）、good（11）、interest*（10）、**excellent**（7）、**strong**（7）、useful（6）、elegant*（4）、**impressive***（4）、new（4）、nice*（4）、well-conducted（4）、comprehensive（3）、**essential**（3）、**great**（3）、noteworthy（3）、prospective（3）、**solid**（3）、timely（3）、**valuable**（3）、well（3）、clearly（2）、**critical**（2）、striking（2）、**wonderful**（2）	be of no use（1）

注：加粗词语为2分情感词，*表示以指定前缀开头的多个单词。

new、novel）、价值（如useful、valuable）以及发展潜力（如prospective、promising）等。专家表达负向情感时，limit*的出现频次较高，主要用于指出方法和结果的局限性。此外，还涉及各种问题缺陷和不确定性，如problematic、caveat、uncertain*等。以上结果表明不同评价对象对应情感词呈现出一定特征，这些特征对于识别专家观点态度进而深度评价论文内容具有参考意义。

4.2 价值特征

论文价值评价是衡量论文质量、推动知识创新与学科进步的必要环节。我国相继出台的《教育部办公厅关于开展清理“唯论文、唯帽子、唯职称、唯学历、唯奖项”专项行动的通知》《关于规范高等学校SCI论文相关指标使用 树立正确评价导向的若干意见》《第五轮学科评估工作方案》等政策旨在推动学术评价从单一的数量评价向质量、创新性、贡献、影响等多维度评价转变，以凸显科研成果的原创性和社会贡献。以往基于论文摘要或全文数据集分析论文价值的方法依然以作者的描述为主，存在一定的主观性。而开放同行评议文本体现的是专家对论文价值的判断，更具科学性。

论文价值具体体现在作者提出的新理论、新方法、新技术、新成果、新应用等创新贡献要素在人类社会发展与科技进步中产生的社会价值与经济效益[24]。本研究从贡献类型和创新程度两个维度分析论文的价值要素。针对贡献类型，已有研究[24-27]主要围绕理论、方法、观点和应用等方面划分贡献。同时，目前广泛使用的医学研究评价框架有回报模型（Payback Model）、研究影响框架（Research Impact Framework，RIF）和加拿大健康科学院（Canadian Academy of Health Sciences，CAHS）框架等[28]，主要围绕学术、政策、产品、健康、经济和社会效益等方面展开。本研究在上述基础上进一步结合同行评议文本，从学术贡献和应用贡献两个方面划分COVID-19项刊论文的贡献类型（见表11）。在创新程度方面，依据程度差异将其划分为开创性和增量性[29]，详见表12。

表11 贡献类型划分说明及示例

贡献类型	贡献细分	说明	示例
学术贡献	理论贡献	发现新的研究问题，针对已有的问题提出新的见解，发现新规律、提出新假设等	The paper reports a new hypothesis to explain the generation of anti-Spike antibodies
	方法贡献	提出或改进研究问题的方法、模型、途径等	The authors discovered a novel means to activate cytotoxic T cells
应用贡献	公共卫生政策制定	为相关部门制定公共卫生政策提供科学依据	Such research is particularly important to assess the effect of public health measures in conjunction with increasing vaccination coverage
	疫苗或药物研发	为疫苗、药物的研发提供指导	These are important new insights, which also bear relevance for vaccine development and monitoring of non-responders
	治疗方案或建议	为治疗新冠患者提供有效的方案或建议	These findings, which underscore the importance of targeting treatment to the stage of the disease, are practice-changing

表12 创新程度划分说明及示例

创新程度	说明	示例
开创性	研究工作具有原始性创新，可以是开创突破性研究，或是首次提出的独创性研究	The study opens up new opportunities for research on potential therapeutic targets in the management of this severe condition
增量性	在已有研究的基础上进行创新，推动领域进程或是对已有研究的扩展和优化	This manuscript greatly enhances our understanding of the pathogenesis of the Omicron BA.1 variant of the SARS-CoV-2 virus

人工提取同行评议文本中评判论文价值的语句，得到238条价值评价句。从价值要素和典型特征词的分析来看，可以得出以下结论。

（1）价值要素。贡献类型的分布（见表13）可以揭示论文主要在哪些方面产生价值。应用贡献占比略高于学术贡献，这表明专家推荐的相关论文更加注重将研究成果转化为具体的实际应用。从细分类型来看，理论贡献占比最高，可见专家推荐的COVID-19项刊论文能够为理解这一新型病毒提供坚实的理论基础；方法贡献和公共卫生政策制定占比较低，说明被推荐的

表13　贡献类型分布

贡献类型	贡献细分	频次	占比/%
学术贡献	理论贡献	99	36.00
	方法贡献	29	10.55
	总计	128	46.55
应用贡献	公共卫生政策制定	25	9.09
	疫苗或药物研发	68	24.73
	治疗方案或建议	54	19.64
	总计	147	53.45

COVID-19项刊论文很少涉及新的、突破性的研究方法，且少有研究成果能够转化为具体的公共卫生政策。总体来看，专家推荐的COVID-19项刊论文的贡献能够覆盖回报模型、RIF、CAHS框架等评价框架中的学术、政策、产品（如疫苗、药物）和健康维度，但较少涉及经济和社会效益维度。

创新程度方面，共有147条价值评价句涉及论文的创新性，其中增量性占比高达85.03%，开创性仅占14.97%，这表明专家推荐的COVID-19项刊论文更多是对已有研究的拓展和推动。为进一步探究不同创新程度论文的贡献差异，对创新程度和贡献类型进行交叉分析。不同创新程度论文的贡献类型分布（见图3）显示，不同创新程度论文在理论贡献和公共卫生政策制定方面的差异尤为显著；增量性创新的贡献类型分布更加不均衡，理论贡献占比接近50%，开创性创新中应用贡献占比高达70.84%，这表明专家推荐的COVID-19项刊论文中，增量性创新主要在已有学术理论体系的发展和完善方面发挥作用，而开创性创新则更注重实际应用领域的开辟和探索。

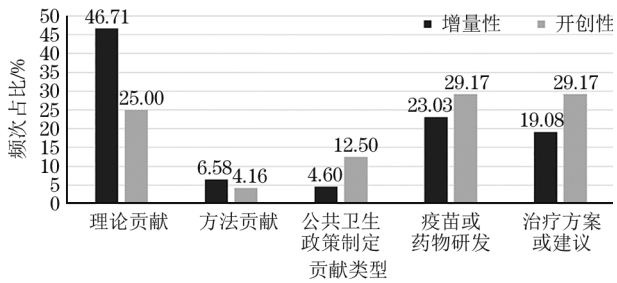


图3　不同创新程度贡献类型分布

（2）典型特征词。在分析同行评议文本价值要素时发现，不同贡献类型和创新程度的价值评价句中具有一些典型特征词，如学术贡献中的issue/question（问题）、hypothesis（假设）、concept（概念）、frame（框架）、mechanism（机制）、view/insight（观点/见解）、finding/discovery（发现）、approach/means/methodology（方法/方法论）、strategy（策略）、model（模型）、pathway（路径）等，应用贡献中的public health（公共卫生）、policy（政策）、implement（实施）、suggest/guide（建议/指导）、measure（措施）、vaccine/vaccination（疫苗/疫苗接种）、drug（药物）、dosage/dose（剂量）、therapy/treatment/therapeutic（治疗/疗法/治疗性的）、assess/evaluate（评估/评价）、effective/efficient/potency（有效的/有效率的/效力）等，创新程度中的breakthrough（突破）、open up new（开辟新的）、seminal（开创性的）、original（原创的）、promote/facilitate（促进）、help（有助于）、enhance/develop/expand（提升/发展/扩展）、provide（提供）、support（支持）、pave the way（铺平道路）、tremendous achievement（巨大成就）、milestone（里程碑）等。这些特征词的归纳总结能够为快速精准地定位评议文本中的价值要素提供有力工具，进而在论文价值自动化识别和评价中发挥作用。

5　结论与启示

开放同行评议为论文评价提供了新的选择，然而，当前对于发表后开放同行评议意见，特别是其文本内容的探索仍需加强。高水平论文的评议意见有助于了解专家评选论文时所关注的核心因素，对其进行分析能够从专家视角揭示优质论文特征。鉴于此，本研究以H1 Connect推荐的COVID-19项刊论文的评议意见为例进行分析，旨在加深对开放同行评议意见的理解，并为应用开放同行评议意见评判高水平论文、完善学术评价体系提供参考。

5.1　研究结论

（1）相较于未被推荐的论文，得到专家推荐的论文在引文和替代计量学评价方法下的表现更为出色。被推荐论文能够快速吸引同行评议专家的关注，但只有少数论文能够引发广泛讨论和获得高度认可。不同标签的论文在3类评价方法下表现各异：改变临床实践的论文在同行评议中表现较好；挑战认知的论文能较快吸引专家关注但综合表现不佳；专业性强的论文更易获得学者和同行专家认可，但社会影响相对有限。

（2）COVID-19项刊论文的开放同行评议文本整体情感倾向不明显，但部分情感句明确体现了专家的态度。正向评价居多，主要关注论文总体优势；负向评价较少，侧重论文方法上的不足。专家选用情感词具有倾向性，正向情感词强调重要性、新颖性、价值、潜力等，负向情感词主要用于指出局限性、不确定性和问题缺陷等。

（3）专家推荐的COVID-19项刊论文在学术领域和实践应用的多个维度上有所贡献，贡献类型因论文创新程度的不同而存在差异。总体而言，应用贡献更为突出；价值评价句中增量性创新占比远高于开创性创新，前者更多展现出学术贡献，后者更多展现出应用贡献。

5.2 应用启示

根据上述结论，对应用开放同行评议意见评判高水平论文、优化学术评价体系提出以下建议。

（1）引入开放同行评议数据源，增强学术评价的科学性与有效性。研究发现不同评价方法对论文的评价结果有所差异，其中同行评议指标评价时滞较短，有利于评判高水平的新发表论文，以及在引文和替代计量学评价方法下价值尚未凸显的论文。同时，同行评议文本中蕴含情感、价值等关键要素，对其进行挖掘有望促进定量与定性评价的有机结合。因此，应当在学术评价体系中引入优质开放同行评议数据源，以激发同行评议优势，增强学术评价的科学性与有效性。尤其是H1 Connect平台的同行评议团队由全球顶尖专家构成，评价数据展现出较高的权威性。此外，调研发现还有许多各具特色的平台值得关注，如：PubPeer允许匿名用户对已发表的论文进行评议，在学术打假中发挥重要作用；《大气化学与物理学》为同行专家、作者、公众和编辑搭建了一个交互式公开讨论的平台，从多个群体视角全方位呈现论文的评议意见。

（2）探索完善同行评议指标，深入挖掘评议文本关键要素。开放同行评议意见为优化学术评价体系提供了新的思路和方向，但在实际应用时仍需谨慎。一方面，应加强探索发表后开放同行评议指标。研究发现论文的RStar指标整体较低且相对集中，但H1 Connect平台缺少相关指标的明确度量标准，认识评议结果仍缺乏依据。目前已有研究探讨同行评议意见的质量，为更全面地认识和应用开放同行评议意见，建议开放同行评议平台优化评分机制，确保能充分体现文献质量的差异，减少评价结果的偶然性。同时建议科研人员加强对评议数据的质量分析，深入探究影响评价结果的各种因素，以便为未来开放同行评议数据的应用提供更加科学、准确的指导。另一方面，应深入挖掘评议文本中的关键要素。可在传统情感极性分析的基础上，更细粒度地挖掘情感句的评价对象、评价角度等，以提升情感分析的质量和深度。但开放评议模式下，同行专家透露其身份的社会压力可能会影响其反馈的情感极性[30]，同时研究发现专家情感也会受到个人习惯和偏好的影响。因此，在应用情感信息时需要注意这些潜在因素，结合论文多元特征综合考量。此外，在挖掘价值要素时需要注意论文价值的体现是多维度的，不同学科价值侧重点存在差异，如COVID-19论文侧重疫苗研发、疗法创新等应用贡献。因此，识别和运用价值要素开展学术评价时，应充分考虑学科特点，制定差异化的评价策略。

（3）加强资源建设和技术融合，智能解析评议文本深层价值。深入挖掘同行评议文本有助于揭示论文的深层次价值。本研究采用人工标注的方式以确保从微观视角准确、详尽地分析同行评议文本，然而，在对其进行大规模应用时，应探索智能化的处理技术，以提升同行评议文本的处理效率和挖掘深度，有效辅助专家的评价与决策。一方面，开放同行评议平台应尽可能提供应用程序编程接口，为科研人员分析、应用同行评议数据提供便利。科研人员可以基于相关数据进一步建设同行评议文本通用语料库，并针对不同学科、研究领域、评价类型等，逐步完善专用语料库，为智能化评价工具开发提供资源保障。另一方面，科研人员应紧跟技术前沿，借助人工智能领域新兴技术赋能同行评议文本挖掘。近年来，飞速发展的生成式人工智能具备生成、总结、提取、分类、检索与改写六大核心能力[31]，在同行评议文本贡献识别、情感分析、文本分类、评价信息集成等任务中具备极大潜力，充分利用其优势有助于推动学术评价体系的智能化、科学化发展。

本研究还存在一些不足之处：①研究对象仅包含COVID-19项刊论文，虽然这类研究具有较强的现实意义，但所覆盖的主题比较单一；②以H1 Connect平台为例进行研究，评议意见来源单一，可能具有一定倾向性和偏差。未来，应探索利用智能化技术分析大规模数据，通过纳入更多期刊、学科和平台，检验相关结论并探讨差异，从而提升研究结果的全面性和可靠性。

参考文献

[1] 索传军，盖双双，周志超. 认知计算：单篇学术论文评价的新视角[J]. 中国图书馆学报，2018，44（1）：50-61.

[2] KOVANIS M，PORCHER R，RAVAUD P，et al. The global burden of journal peer review in the biomedical literature：strong imbalance in the collective enterprise[J]. PLoS One，2016，11（11）：e0166387.

[3] 刘丽萍，刘春丽. 开放同行评议利弊分析与建议[J]. 中国科技期刊研究，2017，28（5）：389-395.

[4] 樊秀娣. 把发表“顶刊”论文视为“王道”不可取[N]. 中国科学报，2020-07-21（5）.

[5] 谭贝加. 被引频次结合Altmetrics得分、F1000评分用于生物医学论文影响力评价的可行性研究：以2014—2017年Altmetrics Top100论文为例[J]. 中国科技期刊研究，2020，31（11）：1388-1393.

[6] 许丹，韩爽，徐爽. Faculty Opinions不同评价条件下论文多元评价指标差异性及相关性分析[J]. 中国科技期刊研究，2022，33（2）：246-259.

[7] BORNMANN L. Validity of altmetrics data for measuring societal impact：a study using data from Altmetric and F1000Prime[J]. Journal of Informetrics，2014，8（4）：935-950.

[8] DU J，TANG X L，WU Y S. The effects of research level and article type on the differences between citation metrics and F1000 recommendations[J]. Journal of the Association for Information Science and Technology，2016，67（12）：3008-3021.

[9] 姜育彦，刘雪立. 绝对颠覆性指数与同行评议指标及CNCI的关系：基于病毒学论文的研究[J]. 图书情报工作，2023，67（3）：96-105.

[10] 张明阳，王刚，彭起，等. 学术论文公开评审平台数据分析[J]. 计算机科学，2021，48（6）：63-70.

[11] GHOSAL T，VERMA R，EKBAL A，et al. DeepSentiPeer：harnessing sentiment in review texts to recommend peer review decisions[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg：Association for Computational Linguistics，2019：1120-1130.

[12] ZONG Q J，FAN L L，XIE Y F，et al. The relationship of polarity of post-publication peer review to citation count[J]. Online Information Review，2020，44（3）：583-602.

[13] ZHANG G Y，WANG L C，XIE W X，et al. “This article is interesting，however”：exploring the language use in the peer review comment of articles published in the BMJ[J]. Aslib Journal of Information Management，2022，74（3）：399-416.

[14] WANG P L，WILLIAMS J，ZHANG N，et al. F1000Prime recommended articles and their citations：an exploratory study of four journals[J]. Scientometrics，2020，122（2）：933-955.

[15] 秦成磊，韩茹雪，周昊旻，等. 同行评审意见类型识别及其在不同被引频次下的分布研究[J]. 图书情报工作，2022，66（13）：102-117.

[16] GHOSAL T，KUMAR S，BHARTI P K，et al. Peer review analyze：a novel benchmark resource for computational analysis of peer reviews[J]. PLoS One，2022，17（1）：e0259238.

[17] 梁帅，高继平. 基于F5000论文评审意见的优秀论文特征识别[J]. 科学学研究，2017，35（3）：331-337.

[18] WALTMAN L，COSTAS R. F1000 recommendations as a potential new data source for research evaluation：a comparison with citations[J]. Journal of the Association for Information Science and Technology，2014，65（3）：433-445.

[19] H1 Connect. A new dawn for research evaluation[EB/OL]. [2024-05-20]. https://connect.h1.co/blog/a-new-dawn-for-research-evaluation/.

[20] LUO J W，FELICIANI T，REINHART M，et al. Analyzing sentiments in peer review reports：evidence from two science funding agencies[J]. Quantitative Science Studies，2021，2（4）：1271-1295.

[21] WANG P L，SU J. Post-publication expert recommendations in faculty opinions（F1000Prime）：recommended articles and citations[J]. Journal of Informetrics，2021，15（3）：101174.

[22] 李江波，张梁，姜春林. Altmetrics视角下的人文社会科学学术专著影响力评价研究：基于BkCI、Amazon和Goodreads的比较分析[J]. 情报学报，2020，39（9）：896-905.

[23] NWOGU K N. The medical research paper：structure and functions[J]. English for Specific Purposes，1997，16（2）：119-138.

[24] 罗卓然，蔡乐，钱佳佳，等. 学术论文创新贡献句识别研究[J]. 图书情报工作，2021，65（12）：93-100.

[25] WOBBROCK J O，KIENTZ J A. Research contributions in human-computer interaction[J]. Interactions，2016，23（3）：38-44.

[26] 章成志，李铮. 基于学术论文全文的创新研究评价句抽取研究[J]. 数据分析与知识发现，2019，3（10）：12-19.

[27] 曹树金，闫欣阳，张倩，等. 中外情报学论文创新性特征研究[J]. 图书情报工作，2020，64（1）：80-92.

[28] 胥美美. 医学研究影响评价进展与启示[J]. 科技管理研究，

2021，41（18）：80-86.

[29] 赵旸. 基于引用语境分析的科研贡献点识别方法研究[D]. 北京：中国科学院文献情报中心，2024.

[30] MATSUI A，CHEN E，WANG Y W，et al. The impact of peer review on the contribution potential of scientific papers[J]. PeerJ，2021，9：e11999.

[31] 叶继元，郭卫兵. 生成式人工智能参与学术评价的反思[J]. 中国社会科学评价，2024（1）：37-48，158.

作者简介

刘晓娟，女，博士，教授，研究方向：信息计量与科学评价，E-mail：lxj_2007@bnu.edu.cn。
于姚，女，硕士研究生，研究方向：信息计量与科学评价。
沈嘉宁，女，硕士研究生，研究方向：信息计量与科学评价。

Analysis of Open Peer Review Comments for High-Level Papers and Enlightenment: Taking H1 Connect as an Example

LIU XiaoJuan YU Yao SHEN JiaNing
(School of Government, Beijing Normal University, Beijing 100875, P. R. China)

Abstract: Open peer review provides new perspectives for evaluating papers. Analyzing the characteristics of review comments on high-level papers can aid in their assessment and optimize the academic evaluation system. Taking papers on COVID-19 that are recommended by experts of H1 Connect, a globally authoritative open peer review platform in the biomedical field, and published on top journals as typical cases of high-level papers, we collect open peer review comments of these papers on H1 Connect. We analyze the attention and recognition, emotional characteristics, and value characteristics reflected in the expert review comments of high-level papers from both indicator and textual perspectives. The findings reveal that high-level papers can quickly attract the attention of peer reviewers, but only a few spark heated discussions. The open peer review texts often contain emotional sentences reflecting the reviewers' attitudes, with more positive comments addressing the overall performance of the papers and fewer negative comments mainly addressing the research methods used. The open peer review comments indicate that the value of papers is reflected in multiple dimensions including academic fields and practical applications, with an especially strong emphasis on practical contributions. Through the analysis of open peer review comments on high-level papers, it is recommended that open peer review data sources be introduced into the academic evaluation system to enhance the scientificity and effectiveness of academic evaluation. In practical applications, it is still necessary to further improve the peer review indicators, deeply explore the key elements in the review texts, strengthen resource construction and technological integration, and utilize cutting-edge technology to intelligently analyze the deep value of review texts.

Keywords: Open Peer Review; High-Level Paper; Academic Evaluation; H1 Connect

（责任编辑：管清濛）

（上接第54页）

Triplet Extraction Method for Green and Low-Carbon Field Based on Large Language Models

WANG LiJun ZHAO ZiYan MA Li JIANG HuiChao ZHANG Ran
(State Grid Information & Telecommunication Co., Ltd., Beijing 100761, P. R. China)

Abstract: Triplet extraction aims to extract entities and their relationships from text to form structured knowledge representations, which is a key technology for building automated knowledge graphs. Although traditional deep learning-based triplet extraction methods perform well when sufficient training data is available, in vertical scenarios such as the green and low-carbon sector of the power industry, the lack of standardized supervised data, high cost of manual annotation, and the presence of many specialized terms in papers and patents limit the recognition accuracy of these methods. To address these issues, this paper proposes a triplet extraction method based on large language models. By using proprietary large models to annotate a small amount of high-quality labeled data and combining retrieval-augmented techniques to guide open-source models for extraction, high-quality and automated vertical domain extraction has been achieved. Moreover, to improve extraction efficiency and precision in few-shot scenarios, this method also includes a data streamlining and complex data segmentation module, which divides the data based on the difficulty level of extraction and further divides complex data to simplify the extraction process, thereby improving the extraction effect. To verify the performance of the model, we automatically annotate a dataset of patents and papers in the power field using GPT-4, and introduce comparisons with well-known proprietary and open-source large models such as ChatGPT and ChatGLM. The experimental results demonstrate that our method achieves better extraction performance.

Keywords: Triplet Extraction; Knowledge Graph; Large Language Model; Green and Low-Carbon

（责任编辑：王玮）