

AI 和人类顺序性协同的管理决策过程 ——来自期刊论文初审筛选的启示

史会斌¹, 李成汉²

(1. 西安电子科技大学经济与管理学院, 陕西西安 710126; 2. 同济大学经济与管理学院, 上海 200092)

摘要:本文深入期刊论文初审筛选决策的具体过程,使用大语言模型,针对“先 AI 后人类”的顺序性管理决策,提出了提高“人类-AI”系统协同性的决策方法。本文指出“先 AI 后人类”的顺序性管理决策的关键是在对 AI 决策方案筛选效能评估基础上,设计 AI 的决策方案筛选阈值。在引入和设计了 spearman 相关系数、优序比较系数、灵敏性、特异性、准确性及 AI 最小保留率指标后,使用聚合多个人类主体的评价基准,完成了确立决策方案筛选阈值的过程。本文对一个期刊三组论文的分析表明,利用现有大语言模型的论文初审筛选决策中,通过保留 75% 的论文,就能够覆盖相关评估基准中的高质量论文,从而使编辑的初审工作量降低约 25%,编辑总体工作效率提高约 15%,编审全流程效率提高约 8%。本文提出的决策方法能够适应不同业务场景以及大语言模型能力的动态演进,具有较强的适用性。

关键词:人机协同;管理决策;期刊论文评审;筛选效能;最小保留率

中图分类号:F272.3

文献标识码:A

人工智能(Artificial Intelligence, AI),特别是大语言模型(Large Language Model, LLM),以其强大的信息搜集能力及分析处理能力,在组织管理决策中发挥着越来越重要的作用^[1],现有研究一方面表明 AI 技术能够在人力资源管理、运营管理、市场营销、会计与金融、战略管理等领域的决策当中发挥重要作用^[2-4],另一方面也对 AI 应用于管理决策的模式进行了提炼,Raisch 和 Krakowski^[5]指出,对 AI 在管理决策中的自动化及增强化功能不能简单的强调某一个方面,而需要使用悖论视角从时间和空间上对自动化及增强化进行综合管理;Raich 和 Fomina^[6]进一步把“人类-AI”的混合型解决问题过程分为三个阶段,解决方案搜索前的目标设计及数据提供阶段和搜索后的选择阶段是由人类完成的,解决方案的搜索过程是一个“人类-AI”的协同过程,可以分为自主搜索、顺序搜索(先 AI 后人类)和交互搜索三种协同方式;与 Raich 和 Fomina 的观点类似,Shrestha 等^[7]认为人类和 AI 协同的管理决策结构可划分为三种类型:完全授权 AI 决策、AI 与人类顺序决策(包括先 AI 后人类和先人类后 AI 两种顺

序)以及 AI 与人类聚合决策。虽然这些研究为开展“人类-AI”协同的管理决策提供了落地场景的扩展及理论指引,但是 Vaccaro 和 Malone^[8]开展的一个跨场景元分析发现,现有的“人类-AI”系统和管理决策上并没有产生很好的协同效应,其决策绩效低于人类或 AI 中的最佳者,这就提出了如何增强“人类-AI”系统的协同性,以提高管理决策绩效的问题。

关于如何提高“人类-AI”系统在管理决策中的协同性问题,Doshi 等人^[4]指出通过聚合多个 LLM 的结果,可以抵消单个 LLM 的偏差,提升 AI 整体判断的准确性。Vaccaro 和 Malone^[8]指出可以把决策任务分成不同的子任务,并依据人类和 AI 的优势进行子任务的分配;也可以避免单一化的绩效指标设计,开发更稳健的“人类-AI”系统评估指标;还可以设定人类、AI 及“人类-AI”系统的质量阈值,对决策的质量进行约束。这些改进“人类-AI”系统协同性的建议需要结合具体场景验证其可行性。

本文选择期刊论文初审这一业务场景,深入

“先 AI 后人类”顺序性协同的具体过程,探讨管理决策中提高“人类 - AI”系统协同性的路径。首先明确提出,“先 AI 后人类”顺序性协同的管理决策当中,关键是要了解 AI 的能力范围,并据此设置 AI 对决策方案的筛选比例,然后在期刊论文初审的场景下,对 AI 决策能力的评估指标、评估基准、决策方案筛选的阈值进行了设计、分析和应用,为开展“先 AI 后人类”顺序性协同的管理决策提供了可操作的程序及方法,同时也为寻求提升“人类 - AI”系统的协同性提供了可行路径。

1 研究设计

期刊论文评审是保持学术届活力的重要基础,但是面对论文评审的繁重任务,学者们一直在探讨如何使用人工智能降低评审工作量,提升评审的效率和质量。虽然在 LLM 出现之前,学者们就开始探讨自行设计的人工智能算法在论文评审中的应用范围及可行性^[9],但是由于 LLM 在阅读理解,知识整合,甚至逻辑推理方面表现出很强的能力,以及其天然的使用方便性,这就使得利用 LLM 改进甚至替代人类进行论文评审的研究大量出现^[10]。现有研究已经发现 LLM 在论文质量的总体排序方面与人类具有一定的一致性^[11,12],但是在当前的一致性水平下,如何将 LLM 应用于论文评审的具体过程还未很好地解决。另外现有研究也发现 LLM 有扩大评分的倾向^[13],而且单一 LLM 生成的结果往往存在多样性不足^[14]以及偏见性^[4]的问题,因此完全依赖单一 LLM 评分进行论文筛选决策可能会存在评分准确性及淘汰率过低的问题。

期刊论文的评审主要包括初审、外审和复审三个阶段,初审由编辑做出拒稿或进行外审的决策,外审则邀请外审专家对论文进行评价,并给出评审意见,然后编辑依据外审结果进行复审,并做出录用决策。在论文的初审阶段,需要编辑全面评估每一篇论文的质量,在投稿量较大的情形下,将形成很大的工作量。本文认为如果我们能够利用 AI 的评分结果,在编辑开始初审工作之前就淘汰一些质量较低的投稿论文,把剩余评分较高的论文让编辑进行初审筛选,就能够发挥 AI 对论文质量的初步把关作用,并降低编辑评审的论文数量,从而利用“先 AI 后人类”的顺序协同模式,从整体上提高论文评审的工作效率。在这个过程中对 AI 的决策方案筛选

能力评估成为开展协同决策的关键。

因此本文基于现有 LLM 的能力,综合多个 LLM 的评分结果,设计相关方法,评估 LLM 的评分结果能够在多大程度上用于准确筛选出质量较低的论文,以降低需要编辑初审的论文数量。具体来说,本文针对某一个管理类期刊,搜集了其在三个不同时期,编辑对三组提交论文的最终排序及录用结果。针对这些提交的论文,本文同时让两位专家独立重新进行了质量排序,也使用主流的 LLM(包括 Deepseek - R1、通义千问、智谱清言、豆包、文心一言)对每篇论文进行了评分,并形成 AI 排序。基于第一和第二组数据,本文首先使用 spearman 相关系数和优序比较系数对 AI 综合排序的效能进行评价;然后综合编辑部和两位人类专家排序中的高质量论文(各自排序的前 1/3),形成 4 个高质量论文集合,作为对 AI 筛选效能评估的基准,并使用灵敏性(sensitivity)、特异性(specificity)、准确性(accuracy)、AI 最小保留率 4 个指标对 AI 的筛选效能进行评估。评估的结果表明 AI 最小保留率约为 75% 时,筛选出的结果能够覆盖编辑部和人类专家相关基准中的全部高质量论文,这也是应用 AI 进行论文筛选时一个阈值。按照这一筛选效能,把 AI 应用于论文的初审,大约能提高初审效率 25%,提高编辑总体工作效率 15%,提高编审全流程效率 8%。最后,利用第三组论文的数据,对本文提出的筛选阈值进行了验证,基本确认了本文在 AI 筛选效能评估和 AI 筛选阈值确立方面所设计的方法及其指标的可行性。

2 数据来源

2.1 编辑部评审数据

本研究的样本基于某管理类期刊对投稿论文的评审记录,包括 A、B、C 三组不同时期论文的评审及录用结果。该期刊的论文评审主要包括三个阶段:初审、外审和复审,编辑依据外审结果进行复审时,会使用评价量表给出复审评价,并做出录用决策。在复审过程中还会存在论文的退修及再评审过程,由于这一过程与本文研究关系不大,因此本文在复审环节只考虑最终录用决策结果,而对复审过程中的其他环节不做考虑。

如表 1 所示,本研究获得的 A 组论文数量为 13 篇,编辑部的评审结果为:4 篇初审拒稿(31%),9 篇进入外审,最终 1 篇录用。B 组论文数量为 12

篇,评审结果为:4 篇初审拒稿(33%),8 篇进入外审,最终 1 篇录用;C 组论文数量为 13 篇,评审结果为:5 篇初审拒稿(38%),8 篇进入外审,最终 1 篇录用。初审拒稿的论文没有评分,进入复审的论文编

辑会按照一个评价量表对各维度进行等级性评估。本研究将使用 A 组和 B 组论文对 AI 的论文初审筛选效能进行评估,并确立使用 AI 进行论文初审筛选的阈值,然后应用于对 C 组论文进行初审筛选。

表 1 本研究样本概况
Table 1 Sample overview of this study

组别	论文总数	编辑初审拒稿数	编辑初审拒稿率	编辑初审保留率	外审/复审论文数	最终录用数
A 组	13	4	31%	69%	9	1
B 组	12	4	33%	67%	8	1
C 组	13	5	38%	62%	8	1

2.2 人类专家评审数据

对于 A 组和 B 组的论文,本文分别邀请了两位专家对每组论文的质量从高到低进行排序。考虑到编辑部的结果只是综合了编辑和两位外审专家意见形成的结论,决策人数较少,只有 3 个,而管理决策的准确性会随着决策主体规模的增加而增加^[4],因此本文将编辑部的评审结果与两位人类专家的结果进行综合后,形成对 AI 筛选效能评估的基准。

2.3 AI 评审数据

我国有多家综合性能领先的 LLM,综合考虑国内主流 LLM 平台对输入输出内容的审查制度和 API 接口的便捷性,本文选择市场评价较高的五个 LLM:Deepseek-R1、通义千问、智谱清言、豆包、文心一言。

为了确保学术论文的质量,并筛选出符合期刊要求的稿件,本文尽量使 LLM 以统一的标准对论文进行评分。按照本文所选样本期刊的评审原则,最终将论文质量评价的标准汇总为问题前沿性,理论创新性以及研究严谨性 3 个主要维度。Doshi 等^[4]的研究结果表明,对 LLM 进行明确的角色设置以及思维链条的引导能够获得更准确的结果,因此本文设计了提示词,要求 LLM 作为论文评审专家从问题前沿性、理论创新性以及研究严谨性这 3 个维度对论文进行打分,并对 3 个维度的得分取平均值得到每篇论文最终的质量评价分数(0-100)。

本文使用的提示词如下:
“你是一位论文期刊的评审专家,请你阅读全文(不需要阅读末尾的参考文献部分),从问题前沿性(1 到 25 分代表研究问题很多人讨论过,相关成果很多;26 到 50 分代表研究问题有人关注过,但相关成果不多;51 到 75 分代表文章的研究问题之前

较少人关注,相关成果很少;76 分到 100 分则代表文章的研究问题十分前沿,几乎没有人提出过相关研究成果)、理论创新性(1 到 25 分代表文章理论部分几乎没有创新;26 到 50 分代表文章的理论部分验证了前人的理论成果,但创新仍然不足;51 分到 75 分代表文章扩展了现有理论的边界,提出了一些建设性的意见;76 分到 100 分代表文章填补了现有理论的空白,做出了突破性的理论贡献)和研究严谨性(1 到 25 分代表文章的研究方法有严重漏洞;26 到 50 分代表文章的研究方法有部分漏洞,需要完善;51 到 75 分代表文章的研究方法较为严谨,但仍有可以改进的地方;76 分到 100 分代表文章的研究方法十分完备,结论可靠)这 3 个维度分别对论文的质量进行评分,每个维度满分为 100 分,分数越高代表文章的质量越高,最后将 3 个维度的分数取平均得到论文的综合得分,不需要列出打分依据。”

由于 LLM 运行过程的“黑箱”问题,导致同一个提示词多次运行可能会得到不同的评分,因此,本文使用每一个 LLM 对每一篇论文评分 5 次,在去除最高分和最低分后,对其余三次打分结果取平均值作为该 LLM 对特定论文的最终评分。为了避免单一 LLM 评分的偏见性^[4],我们将 5 个模型对特定论文的单独评分结果去除最高分和最低分之后,剩余的 3 个分数取平均值,代表 5 个 LLM 对该论文的综合评分,这样就得到了每一篇论文的 AI 质量评分,再按照这个评分对 A、B、C 三组论文分别排序,最终得到每组论文的 AI 质量排名。

3 评估 AI 初审筛选效能的方法

3.1 AI 总体评审效能评估方法

本文通过比较 AI 评分产生的排序与人类主体(包括编辑部、两位独立人类专家)评价形成的排序,对 AI 的总体评审效能进行评估, AI 评审的排序结果与人类主体实际排序结果的一致性程度越高,表明 AI 的评审效能越高。

编辑部对论文的排序是依据编辑在复审过程中形成的评价量表产生的。本文把复审中评价量表结果转化为具体的分数,将论文质量从高到低进行排序。如果两篇论文的排名相同,例如都处于第三名的位置,那么这两篇论文的排名均为 3.5。初审被淘汰的论文由于编辑未进行打分,不参与排名,默认排在复审论文之后。

spearman 相关系数可以用于评估两个等级变量之间的相关性大小,有了各主体(编辑部、两位独立人类专家、AI)对论文的评分,就可以分析他们的 spearman 相关系数,从中可以初步判断 AI 评审结果的准确性。另外针对特定主体的排序,通过将每组内的论文进行两两配对比较,可以形成优序比较集合,例如对于 A 组中的论文 A_i , AI 的排名若高于该组中的论文 A_j 的排名,则 $R_{ij} = R(A_i, A_j) = 1$, 反之为 0, 如果两篇论文排名相同,则 $R(A_i, A_j) = 0.5$ 。

由于只有进入复审的论文才有编辑部排序结果,所以 A、B、C 组分别有 9 篇、8 篇、8 篇论文参与一致性指标分析。针对每一个主体, A、B、C 各组内的论文可以分别形成由 72 对、56 对以及 56 对论文的比较结果构成的优序比较集合。最后,通过比较 AI 与人类优序比较集合的一致性,可以评估 AI 评审结果与人类结果在论文优劣比较上的差异或者共识程度,计算优序比较系数(consistency)的公式如公式(1)所示:

$$consistency = 1 - \frac{\sum_{i=1}^n \sum_{j=1}^{n-1} |R_{ij} - R'_{ij}|}{n(n-1)} \quad (1)$$

其中, n 为组内的论文数量(在这里三组论文数分别为 9, 8, 8), R_{ij} 和 R'_{ij} 分别代表不同主体针对同一组论文评审结果形成的优序比较集合中的一个元素(不同主体的优序比较集合的排序要一致), consistency 的值介于 0 和 1 之间, AI 评审与人类主体对任意两篇论文优劣的共识程度越高,则 consistency 值越趋近于 1。

3.2 评估 AI 初审筛选效能的方法

本文的出发点是利用 AI 降低编辑初审的工作量。如果认为 AI 按照某一保留率(排名前 $x\%$) 识

别出来的论文为高质量论文,那么通过比较 AI 保留下来的高质量论文与编辑部识别的高质量论文的重合率,就能够判断 AI 的筛选效能,重合率越高就说明 AI 在保留率为 $x\%$ 情形下的筛选效能越高。

在比较 AI 筛选结果与人类主体结果的重合率时,参考 Khraisha 等^[10]的指标设计,本文使用到了 4 个筛选状态:真阳性(TP True Positives)、真阴性(TN True Negative)、假阳性(FP False Positives)、假阴性(TN True Negative)。这里的阳性指的是按照人类的基准应该保留的样本(高质量论文),阴性指的是按照人类的基准应该删除的样本(其余论文)。

真阳性 TP,若 AI 按照前 $x\%$ 的保留率保留的某一篇文章与人类基准保留的某一篇文章重合,则说明 AI 对该论文的筛选结果是真阳性,也就是说 AI 与人类同时判断该论文应该保留。

真阴性 TN,若 AI 按照 $(1 - x\%)$ 的淘汰率淘汰的某一篇文章与人类基准淘汰的某一篇文章重合,则说明 AI 对该论文的筛选结果是真阴性,也就是说 AI 与人类同时判断该论文应该淘汰(退稿)。

假阳性 FP,若 AI 按照 $x\%$ 的保留率保留的某一篇文章与人类基准保留的任何一篇文章都不重合,则说明 AI 对该论文的筛选结果是假阳性,也就是说 AI 与人类基准对该论文的判断存在分歧, AI 认为应该保留该论文以进入下一个评审环节,人类基准则认为应该淘汰该论文,该论文不应该进入下一个评审环节。

假阴性 FN,若 AI 按照 $(1 - x\%)$ 的淘汰率淘汰的某一篇文章与人类基准淘汰的任何一篇文章都不重合,则说明 AI 对该论文的筛选结果是假阴性,也就是说 AI 与人类基准对该论文的判断存在分歧, AI 认为该论文应该淘汰,不应该进入下一个评审环节,而人类基准则认为该论文应该保留,并进入下一个评审环节。

基于这些状态,本文进一步使用三个结果性的指标评估 AI 的筛选效能,以判断 AI 筛选结果的有用性。这三个评估指标分别是灵敏性(sensitivity)、特异性(specificity)、准确性(accuracy)。

灵敏性(sensitivity)指 AI 正确识别阳性样本的能力,其计算如公式(2)所示。其中 TP 数量为 AI 识别出来的真实阳性的数量, FN 数量为 AI 把真实阳性识别为阴性的数量, $(TP \text{ 数量} + FN \text{ 数量})$ 表示真实的阳性样本的总数量。sensitivity 数值越大,表

明 AI 识别出真实阳性样本的比例越大,当 sensitivity = 1 时, FN 数量 = 0,表明 AI 能够识别出全部的真实阳性样本。

$$sensitivity = \frac{TP \text{ 数量}}{(TP \text{ 数量} + FN \text{ 数量})} \tag{2}$$

特异性 (specificity) 指 AI 正确识别阴性样本的能力,其计算如公式 (3) 所示。其中 TN 数量为 AI 识别出来的真实阴性的数量,FP 数量为 AI 把真实阴性识别为阳性的数量,(TN 数量 + FP 数量) 表示真实的阴性样本的总数量。specificity 数值越大,表明 AI 识别出真实阴性样本的比例越大,当 specificity = 1 时,FP 数量 = 0,表明 AI 能够识别出全部的真实阴性样本。

$$specificity = \frac{TN \text{ 数量}}{(TN \text{ 数量} + FP \text{ 数量})} \tag{3}$$

准确性 (accuracy) 指 AI 判断结果与真实情况一致的程度,其计算如公式 (4) 所示。(TP 数量 + TN 数量) 为 AI 识别出来的真实阳性和真实阴性样本的总数量,分母为总的样本数量。

$$accuracy = \frac{(TP \text{ 数量} + TN \text{ 数量})}{(TP \text{ 数量} + FN \text{ 数量} + TN \text{ 数量} + FP \text{ 数量})} \tag{4}$$

表 2 AI 筛选论文比例的确定

Table 2 Determination of the proportion of papers screened by AI

序号	比例名称	比例解释
1	AI 高质量论文比例	AI 排序的前 1/3 论文
2	AI 最小保留率比例	AI 排序的前 x% 论文,如果刚好能够覆盖某一个人类基准所包含的全部论文,就说 x% 是这个基准的 AI 最小保留率

另外,虽然编辑部的排序结果是综合编辑和外审意见后得到的,但是对论文的质量还可能存在误判,特别是在编辑部初审淘汰的论文中可能存在高质量论文,因此我们进一步综合两个人类专家的判断结果,形成高质量论文的集合,作为判断 AI 筛选效能的基准。在综合两个人类专家判断结果时,使用了不同的投票决策原则,如表 3 所示,我们一共确立了 4 个由多个人类主体的判断聚合而成的用于评估 AI 论文筛选效能的基准:(1) 编辑部独立基准,包含编辑部排序的前 1/3 论文;(2) 编辑部与专家合并基准,包含编辑部排序的前 1/3 的论文、专家 1 排序的前 1/3 的论文、专家 2 排序的前 1/3 的论文所组成的集合,也就是编辑部、两个专家,只要有一个主体认为是高质量的论文都应该纳入基准当中。(3) 专家增强编辑部的基准,包含编辑部排序的前

3.3 AI 初审筛选效能的评估基准设定

根据本文所选期刊的原则,编辑初审时基本是按照 1/3 的比例进行淘汰的,被淘汰的论文可以被认为是低质量的论文,而在经过外审后,编辑部综合外审意见从剩余论文中筛选出高质量论文并择优刊发。从编辑部评审结果来看,每组复审论文中排名前 50% 的论文可以保证涵盖该组所有最终录用论文,因此我们将这部分论文定义为高质量论文(论文总数的前 1/3),其余论文定义为低质量论文,需要被淘汰。

在使用 AI 进行论文初审筛选时,我们的目标是使 AI 初审筛选的结果在包含高质量论文的情况下,尽可能降低 AI 初审筛选保留比率,这样就能够提高初审的效率。例如以编辑部的高质量论文(编辑部排序前 1/3)作为基准,分析 AI 保留的高质量论文(AI 排序前 1/3)与编辑部高质量论文的一致性程度,一致性程度越高,说明 AI 的筛选效能越高。另外,通过分析 AI 需要保留多大比例的论文(AI 最小保留率),才能使得编辑部保留的高质量论文都被 AI 保留下来,就可以得到使用 AI 进行论文筛选的阈值。相关定义如表 2 所示。

1/3 的论文,以及专家 1 和专家 2 各自排序前 1/3 论文中同时出现的论文。(4) 编辑部与专家共同决策的基准,包含被编辑部、专家 1 和专家 2 中至少两个主体同时排序前 1/3 的论文。其中编辑部与专家共同决策基准对论文质量的判断最为严格,筛选出的高质量论文数量也最少。

4 主要结果

4.1 利用 A、B 组数据评估 AI 总体评审效能

本文通过比较 A 组与 B 组论文的 AI 排序结果与人类专家及编辑部排序结果,进行 AI 总体评审效能的判断。由于只有进入复审的论文编辑部才有排序结果,所以在比较时,A、B 组分别包括 9 篇和 8 篇论文。

表 3 评价 AI 论文筛选效能的基准

Table 3 Benchmarks for evaluating the effectiveness of AI paper screening

序号	基准名称	基准解释
1	编辑部独立基准	包含编辑部排序的前 1/3 论文
2	编辑部与专家合并基准	编辑部、专家 1、专家 2 各自排序前 1/3 论文的并集
3	专家增强编辑部的基准	专家 1、专家 2 各自排序前 1/3 论文的交集,再与编辑部排序的前 1/3 论文合并
4	编辑部与专家共同决策的基准	包含被编辑部、专家 1 和专家 2 三个主体中至少两个主体排序前 1/3 的论文。

首先我们通过 spearman 相关系数对各主体的判断一致性进行比较,如表 4 和表 5 所示,从表中可以看出,各主体排序在 A 组和 B 组上的相关系数都不高,而且都不显著,这与期刊论文评审这一决策的复杂性有关,现有研究也发现人类专家在论文评审意见上的重合率只有 28.58%^[15],这就导致人类专家之间对论文的排序会出现不一致。另外 2024 年的一项研究也指出^[16],现有 AI 的能力在长论文的处理方面还比较有限,其在论文筛选决策中的准确率约为 20%。虽然 AI 的能力在不断地快速演进,但是其提高程度可能还有限,导致其与人类主体判断的相关性也不会太高。虽然表 4 中显示 A 组中各主体相关系数不显著,但也给出一些提示:(1)各人类主体排序的相关系数不高,说明人类之间的排序一致性不强;(2)AI 与人类主体的相关性系数大于 0,排序结果与人类表现出一定的一致性。同样,表 5 中的相关系数也给出一些提示:(1)人类主体排序结果的相关系数大多为负,排序结果的一致性较差,对论文质量判断的分歧较大;(2)AI 与人类主体的相关性出现负相关,说明在人类有较大分歧的论文组,AI 也形成了有别于人类的独立判断。

表 4 A 组论文各主体排序的相关性

Table 4 Spearman correlation of various subjects on group A papers

评价主体	编辑部	专家 1	专家 2	AI
编辑部	1.0			
专家 1	0.00	1.0		
专家 2	0.25	-0.13	1.0	
AI	0.15	0.15	0.23	1.0

其次,本文通过优序比较系数,对各主体的判断一致性进行比较,如表 6 和表 7 所示。从表 6 中可以看出:(1)人类主体对 A 组中任何两篇论文优劣的认识上,具有较高共识,最低的共识率也接近

50%;(2)在 A 组论文中,AI 与人类主体之间能够达成的共识大于 50%,甚至接近 60%,这就为使用 AI 进行 A 组论文筛选提供了基础。从表 7 中可以看出:(1)B 组论文中人类主体间关于论文优劣比较的共识总体低于 50%,特别是专家 2 与编辑部的优序比较系数较低;(2)B 组中 AI 与人类主体关于论文优劣比较具有一定的共识,与编辑部和专家 1 的优序比较系数大于 0.4,与专家 2 的优序比较系数更是达到了 0.536。这就为使用 AI 进行 B 组论文筛选提供了基础。

表 5 B 组论文各主体排序的相关性

Table 5 Spearman correlation of various subjects on group B papers

评价主体	编辑部	专家 1	专家 2	AI
编辑部	1.0			
专家 1	0.02	1.0		
专家 2	-0.49	-0.17	1.0	
AI	-0.25	-0.12	0.14	1.0

表 6 A 组论文各主体排序的优序比较系数

Table 6 Priority comparison coefficient of various subjects on group A papers

	编辑部	专家 1	专家 2	AI
编辑部	1.0			
专家 1	0.500	1.0		
专家 2	0.583	0.472	1.0	
AI	0.500	0.556	0.583	1.0

4.2 利用 A、B 组数据评估 AI 的初审筛选效能

(1)A 组论文初审筛选效能评估

A 组论文一共 13 篇,编辑部独立基准包括 4 篇高质量论文,编辑部与专家合并基准包括 9 篇高质量论文,专家增强编辑部的基准包括 5 篇高质量论文,编辑部与专家共同决策的基准包括 3 篇高质量

论文。本文把 AI 对 A 组论文排序结果的前 1/3 与四个基准分别进行比较,可以确定出每个基准下 TP、FP、TN、FN 的数量,从而计算出灵敏性、特异性、准确性三个指标。

表 7 B 组论文各主体排序的优序比较系数

Table 7 Priority comparison coefficient of various subjects on group B papers

	编辑部	专家 1	专家 2	AI
编辑部	1.0			
专家 1	0.500	1.0		
专家 2	0.286	0.429	1.0	
AI	0.429	0.464	0.536	1.0

如表 8 所示,从这些指标中我们可以看出保留 AI 排序前 1/3 的论文实现的筛选效果。在编辑部独立基准下的灵敏性和特异性指标均为最低,其中灵敏性只有 0.25,表明 AI 保留前 1/3 论文的情况下,能够覆盖编辑部独立基准中 25% 的高质量论文。特异性指标为 0.67,表明大概 67% 的低质量论文都被 AI 识别了出来。准确性指标为 0.54,说明 AI 保留前 1/3 论文的情况下,正确保留和淘汰的论文总比率为 54%。在考虑编辑部与专家的聚合基准时,灵敏性提高到了 0.4,表明 AI 保留前 1/3 论文的情况下最多能够识别出 40% 的高质量论文。特异性提高到了 0.75,表明 AI 能够识别出 75% 的低质量论文。准确性也提高到了 0.62,意味着 AI 对 62% 的论文判断都与人类基准的筛选结果一致。

以上的数据表明,相比于编辑部独立基准,AI 与编辑部和专家的聚合基准一致性更高,但 AI 保留 1/3 论文能够覆盖的高质量论文的比例并不是很高,要想达到对各基准中高质量论文的全覆盖,还需要进一步计算各基准对应的 AI 最小保留率。

AI 的最小保留率,是指当 AI 的排序结果按照

这个保留率保留时,保留下来的论文刚好能够覆盖基准中的全部高质量论文。对于编辑部独立基准来说,AI 保留 77% 的论文(10 篇)就可以覆盖基准中 4 篇高质量。对于编辑部与专家合并基准来说,AI 保留 92% 的论文(11 篇)就能够覆盖基准中的 9 篇高质量论文。对于专家增强编辑部的基准来说,AI 保留 77% 的论文(10 篇)就能够覆盖基准中的 5 篇高质量论文。对于编辑部与专家共同决策的基准来说,AI 保留 77% 的论文(10 篇)就能够覆盖基准中的 3 篇高质量论文。

如果我们以编辑部独立基准来看,当我们按照 AI 保留 77% 的比例进行论文筛选时,能够提高编审工作多大的效率呢?

在未使用 AI 之前,编辑需要先审核全部的论文(n 篇次),然后淘汰其中的 1/3 送外审,外审专家处理的论文数为 $4n/3$ 篇次(一个论文邀请两个外审专家进行评审)。对外审回来的论文,编辑复审要处理 $2n/3$ 篇次。整个编审系统处理的论文篇次是 $(n + 4n/3 + 2n/3) = 3n$ 篇次,编辑的初审工作量是 n 篇次,编辑的总工作量是 $5n/3$ 篇次。

在使用了 AI 之后,由 AI 先依据最小保留率 $x\%$ 进行初审,筛选掉 $n(1 - x\%)$ 的论文,接着由编辑对 $n * x\%$ 篇的论文进行初审,工作量为 $n * x\%$ 篇次,并保留总论文数的 $2n/3$ 篇论文送外审,外审专家处理论文 $4n/3$ 篇次,对外审回来的论文,编辑复审要处理 $2n/3$ 篇次,整个编审系统人工处理的论文篇次是 $(n * x\% + 2n/3 + 4n/3) = (2 + x\%)n$ 篇次,编辑初审工作量是 $n * x\%$ 篇次,编辑的总工作量是 $n * x\% + 2n/3 = (2/3 + x\%)n$ 篇次。

由此对编辑初审工作量可以降低 23% (计算公式为 $1 - x\%$),对编辑总体工作量可以降低 14% (计算公式为 $(3/5) * (1 - x\%)$),对编辑部总体工作量可以降低约 8% (计算公式为 $(1/3) * (1 - x\%)$)。

表 8 A 组论文在不同基准下的 AI 筛选效能

Table 8 The AI screening efficiency of group A papers under different benchmarks

比较基准	保留 AI 排序的前 1/3 实现的筛选效能			AI 最小保留率
	灵敏性	特异性	准确性	
编辑部独立基准	0.25	0.67	0.54	0.77
编辑部与专家合并基准	0.33	0.75	0.46	0.92
专家增强编辑部的基准	0.4	0.75	0.62	0.77
编辑部与专家共同决策基准	0.33	0.7	0.62	0.77

(2) B 组论文初审筛选效能评估

B 组论文一共 12 篇,编辑部独立基准一共包括 4 篇高质量论文,编辑部与专家合并基准包括 9 篇高质量论文,专家增强编辑部的基准包括 6 篇高质量论文,编辑部与专家共同决策的基准包括 3 篇高质量论文。本文把 AI 对 B 组论文排序结果的前 1/3 与四个基准分别进行比较,进行相关指标计算,如表 9 所示。通过这些指标的数值,我们发现 AI 保留

前 1/3 论文实现的筛选效果,在 B 组与 A 组的特征基本相同。在编辑部独立基准下的灵敏性和特异性指标均为最低,其中灵敏性只有 0.25,特异性只有 0.63,准确性只有 0.5。而在虑编辑部与专家共同决策基准下,灵敏性提高到了 0.67,特异性提高到了 0.78,准确性也提高到了 0.75,意味着 AI 最高能够对 75% 的论文做出与人类基准筛选结果一致的判断。

表 9 B 组论文在不同基准下的 AI 筛选效能

Table 9 The AI screening efficiency of group A papers under different benchmarks

比较基准	保留 AI 排序的前 1/3 实现的筛选效能			AI 最小保留率
	灵敏性	特异性	准确性	
编辑部独立基准	0.25	0.63	0.5	1
编辑部与专家合并基准	0.44	1	0.58	1
专家增强编辑部的基准	0.5	0.83	0.67	1
编辑部与专家共同决策基准	0.67	0.78	0.75	0.75

但是在 AI 最小保留率上,对于前三个基准,AI 保留下来的论文想要覆盖基准中的全部论文,就必须一篇论文也不淘汰,也就是基于 AI 的排序结果,在前三个基准下,不能实现筛选功能。这是由于 AI 把人类判别的某些高质量论文排在了最后一位,因此只有全部保留才能覆盖基准中的论文。

但是在最后一个编辑部与专家共同决策的基准中,AI 的最小保留率为 75%,也就是说按照 AI 的排序保留 75% 的论文就能够覆盖 B 组该基准下的全部高质量论文,此时对编辑初审工作量可以降低 25%,对编辑总体工作量可以降低 15%,对编辑部总体工作量可以降低约 8%。

(3) A、B 组论文筛选效能评估总结

综合 A、B 组的结果来看,AI 在编辑部与专家综合判断的三个基准下的灵敏性、特异性和准确性指标都高于编辑部独立基准下的对应指标,在编辑部与专家共同决策基准下的指标最优(准确性最高,最小保留率最低),这说明相比于编辑部的独立判断,AI 更能反应多个专家的综合意见。通过把人类主体确定的高质量论文作为 AI 筛选论文的目标,在最好的情况下 AI 必须保留 75% 的论文才能全部覆盖高质量的论文,能够发挥一定的初审筛选效能。在最差的情况下 AI 必须保留 100% 的论文才能全部覆盖高质量的论文,不能发挥初审筛选效能。

4.3 C 组数据应用结果

以上 AI 筛选的结论表明,AI 在编辑部与专家共同决策的基准下,通过保留 75% 左右的论文(A 组 10 篇,B 组 9 篇),就能够覆盖质量较高的论文,从而发挥初审筛选论文的效果。因此我们把这一结论应用于 C 组论文,验证是否能够达到较好的筛选效果。C 组论文共 13 篇,其编辑部独立基准的高质量论文包含 4 篇论文,当 AI 的保留率为 75% 时(10 篇论文),AI 选择出来的论文可以覆盖编辑部确定的全部高质量论文,此保留率也为 C 组的最小保留率。

通过 A、B、C 三组数据分析,可以初步得出结论,AI 的恰当运用在保留高质量论文的前提下,可以提高初审论文筛选的效率,但是目前提高的程度还比较有限,大约提高论文初审效率约 25%,提高编辑总体工作效率约 15%,提高编审全流程效率约 8%。

5 总结与结论

本文以期刊论文初审为场景,深入管理决策的具体过程,探索实现管理决策中“人类 - AI”系统协同性的方法。针对“先 AI 后人类”的顺序性管理决策,需要在对 AI 的决策方案筛选效能评估基础上,设计 AI 的决策方案筛选阈值,使 AI 的筛选结果能够尽可能多地淘汰质量较低的方案,然后交由人类

进行后续决策方案的选择。本文的主要贡献集中于对AI的决策方案筛选效能评估上:首先,通过引入spearman相关系数及创新性的设计优序比较系数,实现了对AI总体决策效能的评估;其次通过定义筛选结果的阴性和阳性状态,引入了灵敏性、特异性、准确性指标,实现了对AI决策方案筛选效能的评估;然后,通过聚合人类集体判断的方式进行基准设计,并提出多个人类专家之间的判断可以通过合并、增强、共决三种方式形成不同的评价基准,这就避免了使用单一评价基准不能准确反映AI决策能力的局限^[8],使对AI决策方案筛选效能的评估更加全面;最后,通过设计最小保留率这一指标,明确了AI在决策方案筛选中的阈值设定方法,这一指标不仅能够反映AI的方案筛选效能,而且也使得对AI的应用不只是停留在AI与人类决策结果相关性的讨论层面,而是实现了把AI直接应用于管理决策的可能,也就是按照最小保留率这一阈值首先让AI进行决策方案的筛选,然后在此基础上让人类进一步筛选决策方案,就可以提高整体的决策效率。

同时本文的研究不拘泥于AI(或者说LLM)的能力水平,而是关注如何在现有AI能力水平的基础上,把AI切实引入到管理决策当中,这样就可以加快AI落地应用的步伐。本文提出的AI决策能力的相关评估指标及其使用逻辑,不仅能够应用于其他场景中的决策方案筛选,而且能够适应LLM能力的动态演进,在AI能力有显著变革后,重新分析计算AI决策方案筛选的阈值,实现“人类-AI”系统协同性水平的不断提升。

对于期刊论文初审场景来说,本文的分析表明,现有AI的能力在期刊论文的初审筛选决策中能够实现一定的筛选效应,从而实现“先AI后人类”的顺序性协同决策。本文的分析结论显示,以现有通用LLM的能力,通过保留75%的论文,就能够覆盖被评审论文中质量较高的论文,能够使编辑的初审工作量降低大约25%,提高编辑总体工作效率约15%,提高编审全流程效率约8%。如果随着未来LLM能力的提升,其在论文筛选中的最小保留率能够降低到67%,那对于本文研究的样本期刊来说(该期刊初审淘汰约1/3论文),AI将能够完全替代编辑的初审工作,实现期刊评审效率的大幅度提高。

虽然本研究从管理决策的过程角度,提出了提高“人类-AI”系统协同性的方法,但是还存在以下局限,需要未来的研究进行拓展。第一,由于受到样

本的可获得性限制,本文分析中使用的样本量较小,使相关指标在对应的子样本量发生一个较小变化时,指标数值会发生较大的跳跃,从而影响了本文对AI提升方案决策效率的准确判断。但是这个不足并不影响本文提出的方法的可行性。未来的研究可以把本文的方法应用于决策方案数量较大的场景,例如一些具有较大论文投稿量的大型会议,以增加评估指标数值的稳定性,帮助我们更加准确地认识AI在决策方案筛选中的效能。第二,在对本文决策方法有效性进行验证时,使用的C组论文与A组和B组来自同一个期刊的不同时间,这虽然能够为本研究基于A、B组论文得出的结论提供一定的效度检验,但是与使用来自其他期刊的论文进行效度检验相比,对本文研究结论有效性的验证还是略显不足,未来的研究可以使用不同来源的样本,对本文的结论进行验证。第三,使用AI进行决策方案筛选,相当于改变了决策中“守门人”规则的实现方法,Korena^[17]指出初步筛选的“守门人”规则会导致整个管理决策过程产生偏见,那么AI作为“守门人”会对管理决策结果产生哪些偏见及如何进行矫正,将是未来可以进一步研究的内容。第四,现有研究发现AI的引入会改变对人类的技能要求^[18,19],在本文研究的期刊论文评审场景中,把AI应用于论文的初审筛选对编辑的能力要求会带来哪些变化,也是未来研究可以进一步关注的内容。

参考文献:

- [1] Cao G, Duan Y, Edwards J S, et al. Understanding managers' attitudes and behavioral intentions towards using artificial intelligence for organizational decision-making[J]. *Technovation*, 2021, 106: 102312.
- [2] 张亚莉,李辽辽,丁振斌. 组织管理中的人工智能决策:述评与展望[J]. *外国经济与管理*, 2024, 46(10): 18-38. Zhang Y, Li L, Ding Z. AI decision-making in organizational management: A review and prospects[J]. *Foreign Economics & Management*, 2024, 46(10): 18-38.
- [3] 谷炜,刘亚金, Lu F S,等. 人工智能驱动管理决策:应用、感知与偏见[J]. *中国管理科学*, 2025, 33(5): 99-112. Gu W, Liu Y, Lu F S, et al. AI-driven decision sciences: Application, perception and bias[J]. *Chinese Journal of Management Science*, 2025, 33(5): 99-112.
- [4] Doshi A R, Bell J J, Mirzayev E, et al. Generative arti-

- ificial intelligence and evaluating strategic decisions[J]. Strategic Management Journal, 2025, 46 (3): 583 – 610.
- [5] Raisch S, Krakowski S. Artificial intelligence and management: The automation – augmentation paradox[J]. Academy of Management Review, 2021, 46 (1): 192 – 210.
- [6] Raisch S, Kateryna F. Combining human and artificial intelligence: Hybrid problem – solving in organizations [J]. Academy of Management Review, 2025, 50 (2): 441 – 464.
- [7] Shrestha Y R, Ben – Menahem S M, Von Krogh G. Organizational decision – making structures in the age of artificial intelligence [J]. California Management Review, 2019, 61 (4): 66 – 83.
- [8] Vaccaro M, Almaatouq A, Malone T. When combinations of humans and AI are useful: A systematic review and meta – analysis [J]. Nature Human Behaviour, 2024, 8 (12): 2293 – 2303.
- [9] Kousha K, Thelwall M. Artificial intelligence to support publishing and peer review: A summary and review[J]. Learned Publishing, 2024, 37 (1): 4 – 12.
- [10] Khraisha Q, Put S, Kappenberg J, et al. Can large language models replace humans in systematic reviews? Evaluating GPT – 4's efficacy in screening and extracting data from peer – reviewed and grey literature in multiple languages[J]. Research Synthesis Methods, 2024, 15 (4): 616 – 626.
- [11] Thelwall M, Abdallah Y. Evaluating the predictive capacity of ChatGPT for academic peer review outcomes across multiple platforms[J]. Scientometrics, 2025, 25 (3): 1 – 23.
- [12] Li Y, Datta S, Rastegar – Mojarad M, et al. Enhancing systematic literature reviews with generative artificial intelligence: Development, applications, and performance evaluation[J]. Journal of the American Medical Informatics Association, 2025, 32 (4): 616 – 625.
- [13] Latona G R, Ribeiro M H, Davidson T R, et al. The AI review lottery: Widespread AI – assisted peer reviews boost paper scores and acceptance rates[J]. arXiv preprint arXiv, 2024, 2405.02150.
- [14] Boussieux L, Lane J N, Zhang M, et al. The crowdless future? Generative AI and creative problem – solving [J]. Organization Science, 2024, 35 (5): 1589 – 1607.
- [15] Liang W, Zhang Y, Cao H, et al. Can Large Language Models provide useful feedback on research papers? A large – scale empirical analysis[J]. arXiv preprint arXiv, 2023, 2310.01783.
- [16] Zhou R, Chen L, Yu K. Is LLM a Reliable Reviewer? A Comprehensive Evaluation of LLM on Automatic Paper Reviewing Tasks[C]. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC – COLING 2024), 2024; 9340 – 9351.
- [17] Koren M. The gatekeeper effect: The implications of pre – screening, self – selection, and bias for hiring processes[J]. Management Science, 2024, Published Online.
- [18] Jia N, Luo X, Fang Z, et al. When and how artificial intelligence augments employee creativity[J]. Academy of Management Journal, 2024, 67 (1): 5 – 32.
- [19] Law K K, Shen M. How does artificial intelligence shape audit firms[J]. Management Science, 2025, 71 (5): 3641 – 3666.

The process of sequential AI – Human collaborative decision – making ——Insights from initial manuscript screening in journal peer review

SHI Hui – bin¹, LI Cheng – han²

(1. School of Economics and Management, Xidian University, Xi'an 710126, China;

2. School of Economics and Management, Tongji University, Shanghai 200092, China)

Abstract: The widespread application of the "AI – human" system in management decision – making has sparked discussions on enhancing the synergy of the "AI – human" system to improve management decision – making performance. Although existing research has proposed some methods to improve the synergy of the "AI – human" system in management decision – making, there is still a lack of validation in specific business scenarios. This article chooses the initial screening of journal articles as a scenario and delve into the

(下转第 215 页)

receiving fewer citations, suggesting a greater propensity to challenge existing paradigms. In contrast, Distinguished Young Scientists generate research with higher content novelty and team diversity, resulting in more citations. This pattern demonstrates the complementary roles that researchers at different career stages play in scientific advancement. The distribution analysis of our indicators shows that while most research concentrates in low – novelty regions, there exists a significant peak in high – novelty areas, reflecting researchers’ strategic choices between incremental advances and radical innovation. Team background diversity shows a right – skewed distribution, indicating that while most research teams maintain relatively homogeneous backgrounds, there is a growing trend toward more diverse team compositions, particularly among senior researchers. These findings have important implications for science funding policies and talent development strategies. First, evaluation criteria should be redesigned to better support potentially disruptive research, particularly from early – career scientists, whose innovative contributions might be undervalued by traditional metrics. Second, differentiated support mechanisms should be developed for researchers at various career stages, recognizing their distinct approaches to innovation. Third, funding programs should actively promote interdisciplinary collaboration while maintaining domain expertise. Our results suggest that current evaluation metrics focusing primarily on citation counts might systematically undervalue young researchers’ contributions to scientific disruption. This indicates a need to develop more sophisticated assessment systems that can capture different dimensions of scientific impact. These insights contribute to our understanding of how research novelty and team diversity influence scientific outcomes in China’s research funding system, offering practical guidance for optimizing funding mechanisms to foster both steady knowledge advancement and breakthrough discoveries.

Key words: research novelty; team diversity; research impact; disruptive innovation; research funding policy

(上接第 50 页)

specific process of sequential AI – Human collaboration to exploring the path to improving the synergy of the "AI – human" system in management decision – making.

This article proposes that we can use the scoring results of AI to eliminate some low – quality submitted papers before the editor starts the initial screening work. Then, the editor screens the remaining papers with higher scores. In this process, AI’s filtering effect will reduce the number of papers reviewed by editors. Therefore, we can use the sequential AI – Human collaborative mode to improve the overall efficiency of paper review work. That is to say, the key to this type decision – making lies in designing AI screening thresholds based on an evaluation of AI’s screening effectiveness.

Firstly, this article introduces Spearman correlation and innovatively designs priority comparison coefficient to evaluate the overall decision – making efficiency of AI. Secondly, by defining the negative and positive states of the screening results, sensitivity, specificity, and accuracy indicators were introduced to evaluate the AI’s decision screening effectiveness; Then, the design of evaluating benchmarks is carried out by aggregating human collective judgments, and it is proposed that the judgments of multiple human experts can be combined, enhanced, and jointly determined to form different evaluation benchmarks. These benchmarks can avoid the limitation of using a single evaluation benchmark that cannot accurately reflect AI’s decision – making ability; Finally, by designing the minimum retention rate as an indicator, the threshold setting method for AI in screening decision was clarified. This indicator not only reflects AI’s screening effectiveness, but also enables the rapid implementation of AI applications in the specific scenario. The analysis of three groups of papers in a journal demonstrates that by retaining 75% of the papers during the initial screening phase using existing large language models, high – quality papers identified in relevant evaluation benchmarks can be effectively covered. This reduces editorial workload by approximately 25%, increases overall editorial efficiency by about 15%, and improves the efficiency of the entire editorial process by roughly 8%.

The research is not limited to the level of AI (or LLM) capabilities, but focuses on how to effectively introduce AI into management decision – making on the basis of existing AI capabilities, so as to accelerate the pace of AI implementation and application. The relevant evaluation indicators of AI decision – making capability and their usage logic proposed in this article can not only be applied to decision screening in other scenarios, but also adapt to the dynamic evolution of LLM capability. After significant changes in AI capability, the threshold for AI decision screening can be reanalyzed and calculated, achieving continuous improvement in the synergy level of the "AI – human" systems.

Key words: AI – Human Collaboration; management decision – making; peer review; screening effectiveness; minimum retention rate