



# STAGER 清单: 生成式人工智能可靠性的 标准化测试和评估推荐

陈镜虹<sup>1\*</sup>, 朱凌煊<sup>1\*</sup>, 牟伟明<sup>1,2\*</sup>, 林安琪<sup>1\*</sup>, 曾东强<sup>3</sup>, 齐畅<sup>4</sup>, 刘灶渠<sup>5,6</sup>, 江爱民<sup>7</sup>, 汤步富<sup>8</sup>, 史文杰<sup>9</sup>, Ulf D Kahlert<sup>9</sup>, 周建国<sup>10,11,12</sup>, 郭世鹏<sup>13</sup>, 陆晓凡<sup>14</sup>, Xu Sun<sup>15</sup>, Trunghieu Ngo<sup>15</sup>, 蒲中机<sup>16</sup>, 贾保磊<sup>16</sup>, Che Ok Jeon<sup>17</sup>, 何勇槟<sup>18,19</sup>, 吴海洋<sup>20,21</sup>, 古书琴<sup>22</sup>, Wisit Cheungpasitporn<sup>23</sup>, 黄浩杰<sup>24,25,26</sup>, 毛卫浦<sup>27,28</sup>, 王诗翔<sup>29</sup>, 陈新<sup>30</sup>, Loïc Cabannes<sup>15</sup>, Gerald Sng Gui Ren<sup>31,32</sup>, Iain S Whitaker<sup>33,34</sup>, Stephen Ali<sup>33,34</sup>, 程全<sup>35,36#</sup>, 苗凯<sup>37,38#</sup>, 袁硕峰<sup>39,40#</sup>, 罗鹏<sup>1#</sup>

<sup>1</sup>南方医科大学珠江医院肿瘤科

<sup>2</sup>上海交通大学医学院附属第一人民医院

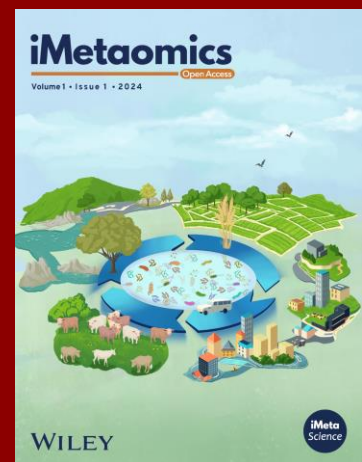
<sup>35</sup>中南大学湘雅医院

<sup>37</sup>澳门大学健康科学学院

<sup>38</sup>澳门大学MoE肿瘤学前沿科学中心

<sup>39</sup>香港大学深圳医院

<sup>40</sup>香港大学李嘉诚医学院



Jinghong Chen, Lingxuan Zhu, Weiming Mou, Anqi Lin, et al. 2024. STAGER checklist: Standardized testing and assessment guidelines for evaluating generative artificial intelligence reliability. *iMetaOmics* 1: e1.

<https://doi.org/10.1002/imo2.1>



# 简介

生成式人工智能（AI）在医疗应用方面拥有巨大的潜力。已有大量研究探讨了各种生成式人工智能模型在医疗保健领域的功效，但目前还缺乏一个全面系统的评估框架。鉴于一些评估生成式人工智能在医疗应用中的能力的研究在方法设计上存在缺陷，目前也缺乏对其进行评估的标准化指南。为此，我们的目标是为评估生成式人工智能系统在医疗领域的表现量身定制标准化评估指南。



# 简介

## 医学领域生成式AI的应用相关的文献检索

Web of Sciences  
Cochrane Library  
PubMed  
Google Scholar

01

02

## 多学科团队

由生命科学、临床医学、  
医学工程方面的专家和生  
成式人工智能用户组成

03

## STAGER清单

涵盖生成式人工智  
能在医疗应用中的  
关键评估方面的32  
个项目的清单

我们的目标是对人工智能系统进行整体评估。检查表勾勒出了从问题收集到结果评估的清晰路径，为研究人员应对潜在挑战和陷阱提供了指导。我们的框架为涉及生成式人工智能在医学中的适用性测试的研究提供了标准化和系统化的方法。它提高了研究报告的质量，有助于生成式人工智能在医学和生命科学领域的发展。



# Publication records of a PubMed search using 'ChatGPT' as the keyword

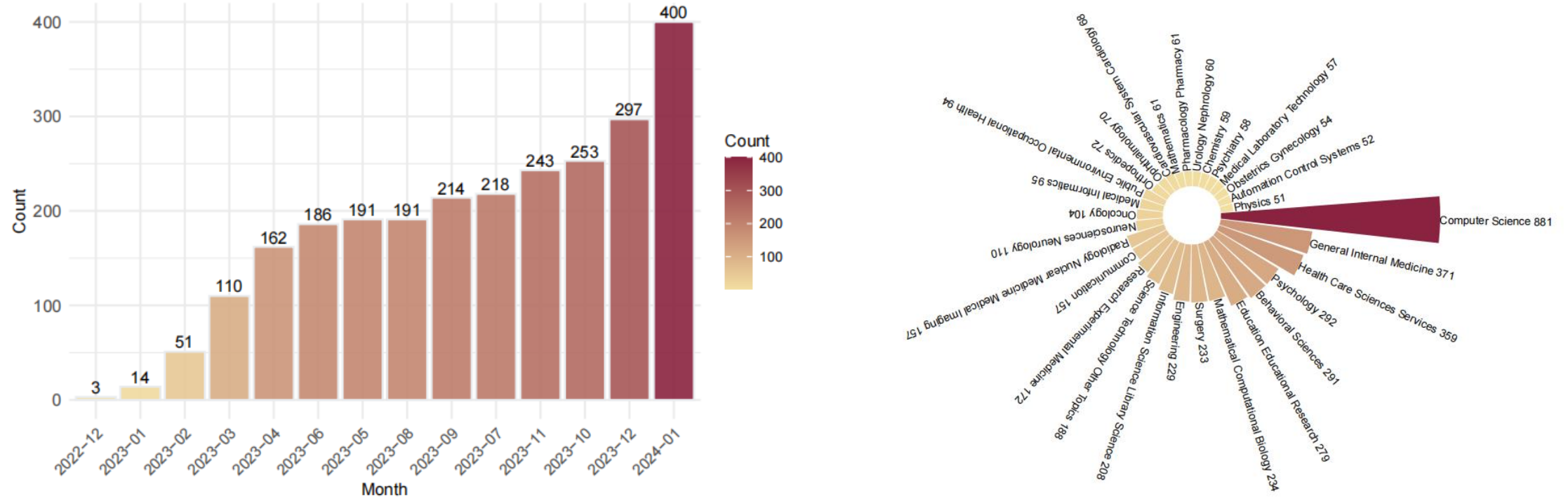
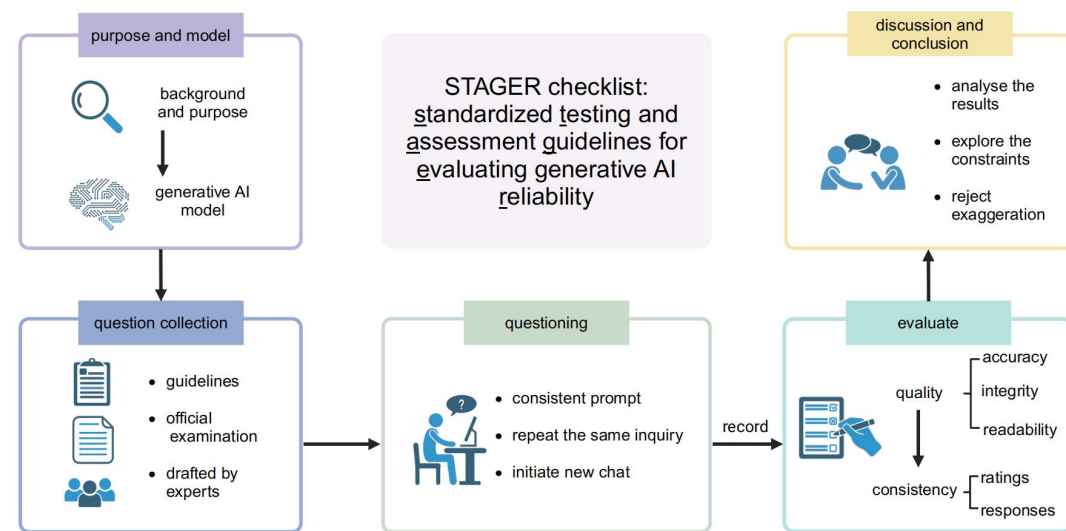


Figure 1 Publication records of a PubMed search using 'ChatGPT' as the keyword.



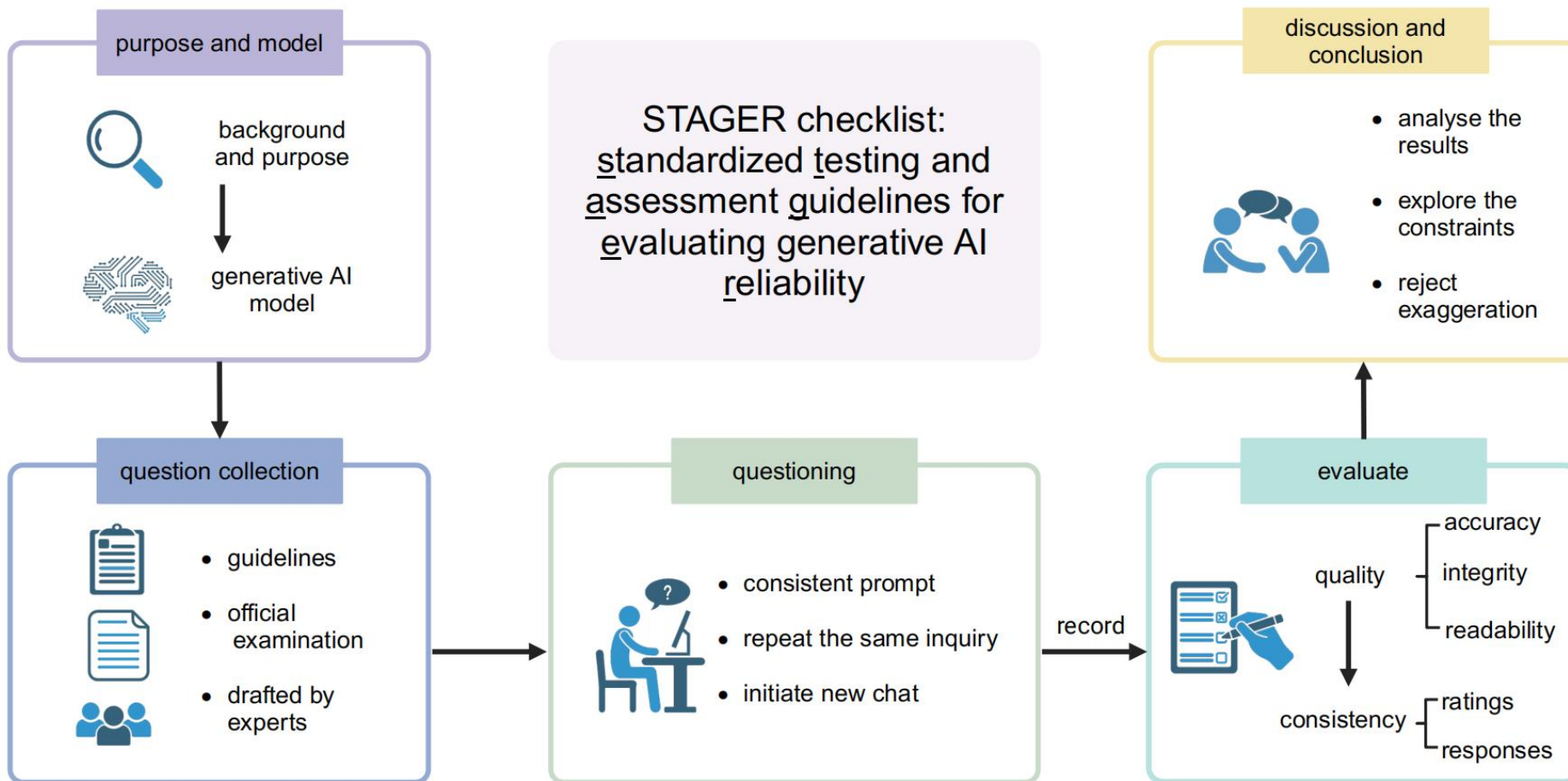
# 亮点

- STAGER清单是评估生成式人工智能（AI）可靠性的标准化测试和评估指南，这是一个有着32个项目的框架，为评估医学和生命科学背景下的生成式人工智能系统提供了量身定制的标准化评估指南。
- STAGER清单由问题收集、查询方法和评估技术等关键方面组成。
- STAGER清单提高了研究质量，促进了这一新兴领域的发展。





# STAGER清单的工作流程





# STAGER清单的详细内容

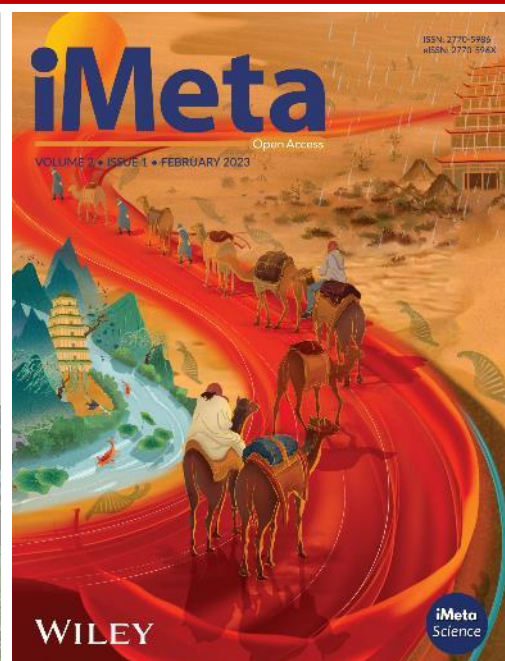
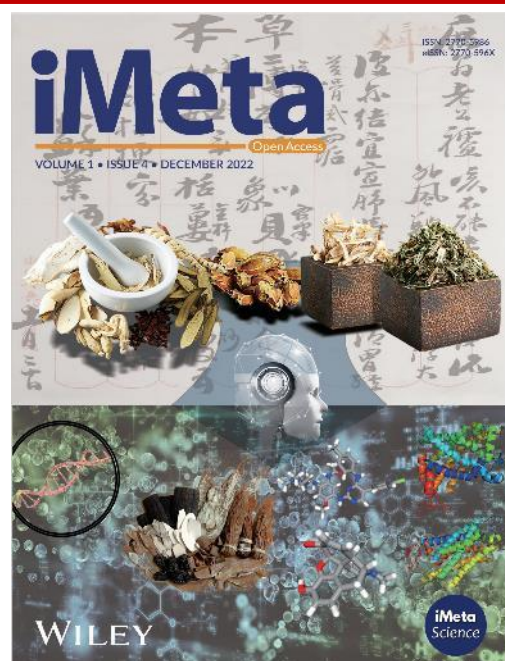
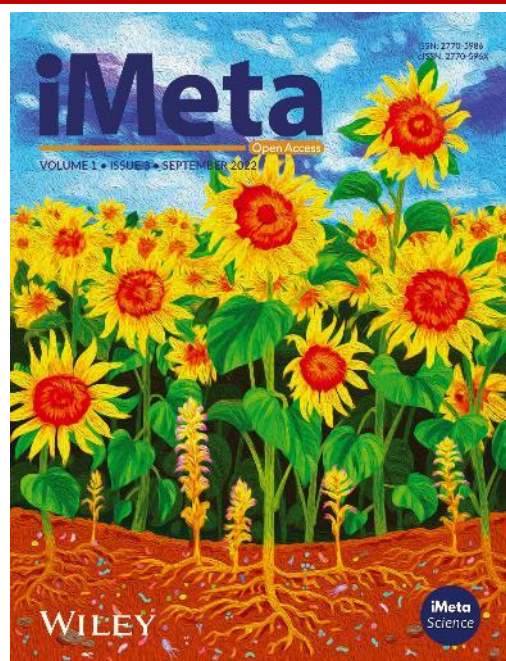
如需简明的清单，请参考文章的表S或访问[GenAIMed.org](http://GenAIMed.org)

Evaluations and explanations of generative AI for medical applications.			
主题	序号	建议	解释
标题	1	将报告确定为与评估生成式人工智能在医学中的适用性的研究相关的文章。	让读者初步了解文本的性质。
摘要	2	说明研究目的、使用的生成式人工智能模型及其版本、问题来源、方法、结果和结论。	为读者快速了解本研究奠定基础，并方便其他研究人员对本研究的设计和结果进行批判性分析。
引言			
合理解释	3	回顾现有相关信息并解释研究背景。	使读者能够把握文章的中心主题。
对象	4	说明具体目标，包括使用的生成式人工智能模型及其版本、生成式人工智能使用的训练集、问题来源	为读者理解文章提供必要的框架。
方法			
问题收集	5	从指南、官方考试题库、通过谷歌等搜索引擎找到的高频问题或专家起草的问题中选择专业问题，确保问题涵盖特定的医学分支领域。	对于指南或问题库，可以手动选择问题或使用软件提取问题，而使用应用程序接口选择问题可以减少主观错误，并对整个数据集更有意义。从搜索引擎中选择问题时，研究人员可以选择经常出现的问题。如果问题由专家增强研究的普遍性。
	6	确保问题在难度、类型和专业性方面具有代表性。	
	7	说明问题是如何收集的、问题的数量、问题是否经过预先筛选、筛选的条件、输入的方式以及相关输入模式，如文本、图像、声音、视频输入等，以及相关属性（如图像分辨率）。	
代理	8	记录所使用的模型、生成式人工智能的版本以及定制参数（如果适用的话）。说明当前使用版本的优缺点以及评估该版本的理由。	所使用的模型和人工智能生成器的版本可能会对结果产生重大影响。温度是影响文本生成随机性的一个参数。温度越高，生成的文本越多样、越新颖，但不可预测性和潜在的不准确性也会增加。温度越低，生成的文本就越稳定。
	9	如果打算将其作为函数序列报告，建议报告模型版本之间的关系（例如，是否是简单升级；如果不是，建议报告横向比较结果）	报告模型版本之间的关系可以澄清技术的演变，帮助用户了解改进或变化。此外，提供横向比较结果有助于理解每个版本的不同功能和应用。
提问	10	使用格式一致的提示语，并在文章中提供完整的提示语。	减少不同提问方式带来的客观差异以及这种差异对答案质量的影响。在文章中提供完整的提示，确保研究的透明度。
	11	多次提出同一个问题，并记录每次回答。	众所周知，生成式人工智能会对相同的询问做出不同的回答，因此有必要反复提问以衡量其一致性。
	12	请注明问题是开放式的还是多项选择式的。	主观题和客观题的评估方式不同。
	13	为每个问题发起新的聊天。	防止生成式人工智能受到上下文的影响。
	14	记录收集到的回答。	减少人工智能版本之间的性能差异和知识更新时间的影响。
准确性	15	说明在处理主观性问题时，为保证评分准确性而采用的任何方法。	准确性是指答复反映或符合现实或真相的程度。
	16	与参考答案进行比较，记录每个问题的正确答案数，并计算客观题的正确率。	生成式人工智能模型正确响应的次数越多，就越被认为是稳健的。
完整性	17	说明用于评估答复之间完整性的任何方法。	完整性是指答复是否全面、详细并涵盖相关信息。
可读性	18	说明用于评估答卷可读性的任何方法。	反映文章阅读和理解的难易程度（如语言的清晰度、结构的组织、语法和拼写的准确性）。
	19	明确评审人员的组成及其理由，建议由来自医学、人工智能和跨学科领域等不同领域的两名以上专家，以及来自伦理学、社会学和用户群体的利益相关者组成。	
评审人	20	特别注意评估答复的可实施性。	确保评估过程的公正性和有效性。
	21	评估对同一问题的不同回答的一致性，以评估生成式人工智能能否稳定地提供一致的回答。	
	22	评估评审员评分的一致性和可靠性，避免评审员之间的主观评分存在显著差异。	有效监控模型性能的方法，有助于检测模型是否出现不稳定行为。
结果			
结果选择	23	描述搜索过程的结果，从收集的问题数量到最终结果，最好使用流程图。	
研究特点	24	说明纳入分析的所有研究，并详细介绍其特点。	
独立的结果	25	介绍每项研究结果的准确性、完整性和可读性，建议使用表格或图表进行介绍。	要深入了解人工智能在医学中的应用价值和局限性，揭示其在特定子领域的表现至关重要。
总括的结果	26	介绍所进行的所有统计综合的结果，以及为探索研究结果间异质性的可能原因而进行的分析结果。	
讨论			
解释	27	根据研究目标分析结果。	全面分析生成式人工智能在准确性、完整性和可读性方面的表现。
	28	说明研究的优势。	让读者了解研究的重要性。
优势和局限	29	探讨研究的限制因素，承认可能存在的偏颇或不准确之处。	提高对研究成果的范围、准确性和适用性的认识
	30	理性讨论，拒绝夸大其词。	诚实和理性的表达是维护学术规范和推动知识进步所必需的。
结论	31	提供一个简明扼要的结论，总结研究的主要发现，重申研究的重要性，并指出未来研究的方向或建议	为今后的研究指明方向，帮助促进生成式人工智能在医疗领域的进一步发展和应用。
其他信息			
研究经费及赞助	32	请提供当前调查以及最初研究的资金支持来源和赞助商的职能。	保持研究的客观性和透明度。



# 总结

- 本清单中的评估框架引入了一种标准化和系统化的方法，用于评估医疗应用中的生成式人工智能研究，重点在于提高研究报告的质量。
- 通过提供一套明确的评估标准，该框架满足了对人工智能研究透明度和严谨性的需求，这对准确性和可靠性是至关重要的。
- 该框架有望促进学术合作和知识交流，为跨学科合作创造沃土，确保其保持前沿性、相关性，并与不断变化的医学科学领域保持一致。
- STAGER清单和更多细节请见：[GenAIMed.org](https://doi.org/10.1002/imo2.1)。



“iMeta”由威立、肠菌分会和华人科学家出版的开放获取期刊，主编由中科院微生物所刘双江和荷兰格罗宁根大学傅静远教授共同担任。目的是发表原创研究、方法和综述以促进宏基因组学、微生物组和生物信息学发展。目标是发表前10%(IF>20)的高影响力论文。期刊特色包括视频投稿、可重复分析、图片打磨、青年编委、中英双语、50万用户的社交媒体宣传等。2022年2月发行，相继被ESCI、Google Scholar、DOAJ、Scopus等数据库收录，发文161篇，被引2316次(Dimension, 2024/2/19)!



主页: <http://www.imeta.science>

出版社: <https://wileyonlinelibrary.com/journal/imeta>



投稿: <https://wiley.atyponrex.com/journal/IMT2>



[office@imeta.science](mailto:office@imeta.science)



宣传片



[iMeta](#)

