生成式人工智能科普写作能力评估

——基于“微生物”主题的科普创作

和鸿鹏　张雅欣　张力恺

（北京航空航天大学人文与社会科学高等研究院，北京 100191）

[摘　要] 生成式人工智能已经成为科普作品创作的重要工具，但与人类创作者相比，其科普写作能力仍然受到质疑。为评估生成式人工智能的科普写作能力，本研究构建了包括易读性、趣味性、科学性和传播效果的四维度评价指标，以“微生物”为主题，分别考察人类创作者、DeepSeek、ChatGPT与文心一言创作的科普作品。结果显示，ChatGPT初步提示与深度提示版本在整体得分上与人类创作版本无显著差异，ChatGPT深度提示版本在“易读性”指标上显著优于人类创作版本；DeepSeek深度提示版本在“趣味性”和“传播效果”得分上显著优于人类创作版本；评价者对所有生成式人工智能生成作品的甄别正确率均不足55%，且评价者更倾向于认为高分作品为人类所创作。研究结果表明，生成式人工智能具备替代人类科普创作者的潜力，进而提出了“人机合作科普创作”这一科普创作新模式，并呼吁学界关注“人类能力幻觉”现象。

[关键词] 生成式人工智能　科普写作　科学传播　人工智能创作能力

[中图分类号] N4；TP18　[文献标识码] A　[DOI] 10.19293/j.cnki.1673-8357.2025.02.003

1 问题提出

1.1 关于生成式人工智能科普写作能力的争论

生成式人工智能为科学家和科普工作者提供了便捷工具，可快速生成科普内容，也吸引了研究者的关注。马科维茨（David M Markowitz）倡导将生成式人工智能融入科学传播（Science Communication）[1]；阿尔瓦雷斯（Amanda Alvarez）等认为其可能彻底改变科学传播的方式[2]；麦克（Mike S. Schäfer）在承认生成式人工智能科学传播能力的同时，也指出人工智能可能导致准确性挑战和“大规模错误”，并对科学传播本身和更大的科学传播生态系统产生影响[3]。事实上，生成式人工智能凭借高效的信息处理和语言生成能力，参与科普写作已成为现实，如亚马逊网站上架了超百本ChatGPT参与创作的科普相关图书[4]。但不同研究者对于生成式人工智能的科普写作能力尚存争议，如雪莉（Shirley S. Ho）认为，“生成式人工智能工具通过可能比以前更简单、更快速的方式生成内容，为科学传播引入了新的协同作用”，但是阿尔瓦雷斯却指出，“鉴于生成式人工智能工具产生无意义

收稿日期：2025-03-15
基金项目：北京市社会科学基金项目“智能时代人机合作科研的学术诚信问题及其治理”（24ZXC011）；中央高校基本科研业务费资助项目“生成式人工智能的学术伦理风险与化解路径”（501XYGG2024132002）。
作者简介：和鸿鹏，北京航空航天大学人文与社会科学高等研究院助理教授，研究方向：科技与社会、科技伦理等，E-mail：he-2004mail@163.com。

的倾向，科学传播者应该考虑生成式人工智能是否实际与他们的工作目的完全对立”[2]。可见，研究者对生成式人工智能的科普写作能力是否达到（甚至超过）人类作者的水平，在观点层面尚存争议，特别是由于不同版本生成式人工智能（如 GPT3.5 与 4.5 版本）的科普写作能力存在较大差异[5]，以及科普写作能力涉及多个维度的评价等原因，使得这一争议变得愈发复杂，有待实证研究作出回答。

1.2 生成式人工智能写作能力评估

生成式人工智能在不同领域的广泛应用，推动对其能力的评估成为学界关注的焦点，如论文评价能力、数据分析能力等[6-8]，其中对写作能力的评价是研究重点之一。对生成式人工智能写作能力评估的相关研究主要从不同写作主题切入，如文学写作[9]、医学写作[10]；还有学者关注生成式人工智能的学术写作能力，如摘要和引言撰写[11]、论文衔接[12]、材料准确性[13]、写作技能[14]等，并提示了潜在的学术伦理风险[15]。但目前鲜有研究对生成式人工智能的科普写作能力作出评估，又因科普写作兼顾通俗性与专业性，现有研究的结论不易直接类推至科普写作。因此，本文的研究问题为：不同类型的生成式人工智能在科普写作的总体表现和关键评价指标上，是否可以达到（甚至超过）人类科普创作者的水平？回答这一问题有助于了解生成式人工智能在科普写作方面的优势、效果及不足，深化对人工智能时代科普工作变革的理解。

2 实验设计与实施

2.1 实验设计

2.1.1 科普写作能力的评价指标构建

《芝加哥科学传播指南》指出，“科普写作”是运用日常的、非专业的语言写科学，而判断一篇科普作品好坏的标准有三。第一，是否有趣，第一句话或者第一段话是否有可能会让大部分读者对这个主题感兴趣，文章的主体部分是否能够满足这种兴趣。第二，是否准确，文中的事实和数据是否无误，概念界定是否清楚，使用是否得当。第三，是否易读，文中的语言是否简洁且不依赖于术语，读者是否能够非常顺畅地从头读到尾，且不会出现卡壳或困扰的地方。第四，是否有良好的过渡衔接，内容之间是否存在良好的组织逻辑[16]。综上，科普作品的评价标准至少包括趣味性、科学性、易读性和组织逻辑 4 个维度。

然而，当前大型语言模型在组织逻辑的呈现形式方面有明显的可辨识特征。为避免因此造成的实验误差，本研究通过适当提示，使人工智能生成的科普作品在呈现形式上趋近于人类创作的作品。由此，“组织逻辑”将不被纳入评价标准。为更好地评价科普作品质量，将“传播效果”纳入指标体系，从而形成表 1 所示的科普作品评价指标。

表 1　科普作品评价指标及测量

评判标准		指标测量
一级指标	二级指标	
易读性	通俗性	我认为这篇文章的语言日常、简洁，不依赖于专业术语
	通顺性	我能够非常顺畅地将这篇文章从头读到尾，没有出现卡壳或困扰的地方
趣味性	有吸引力	这篇文章的题目、第一句话或者第一段话引起了我对这一话题的兴趣
	满足期待	这篇文章的后续内容满足了我此前产生的兴趣和期待
科学性	可靠性	这篇科普文章的事实、信息等内容是真实、准确、可靠的
	概念界定与使用	这篇科普文章的概念界定清晰、使用得当
传播效果	启发性	这篇文章对我有所启发（认知、生活等方面）
	分享意愿	我愿意将这篇文章转发、推荐给身边可能感兴趣的朋友

2.1.2 实验变量与评价者

本实验考察不同类型科普作品创作者的科普写作能力，具体包括 3 个变量：（1）科普作品的创作主体，包括人类与生成式人工智能；（2）大型语言模型类型，包括 DeepSeek R1 版本、ChatGPT 4.0 版本与文心一言 3.5 版本；（3）对大型语言模型的提示程度，分为“初步提示”与“深度提示”。为确保实验的科学性和严谨性，以下变量被严格控制：第一，科普作品主题被限定为“微生物”，以减少主题差异对评价结果的影响。选择该主题的主要原因是微生物学既是现代科学（特别是生命科学）中发展最快的领域之一，同时也与公众生活密切相关，是理解“科学知识”和“科学与社会”这两个科普创作维度的重要切口。第二，实验评价者均采用盲评方式，从而尽可能减少评价者预期对实验结果的影响。

在微生物主题下，实验选定两个子主题——“微生物与社会”和“微生物知识”，每个子主题包括 5 篇科普作品，分别是人类创作版本（A）、文心一言初步提示版本（B）、ChatGPT 初步提示版本（C）、ChatGPT 深度提示版本（D）和 DeepSeek 深度提示版本（E），共计 10 篇科普文章（见表 2）。每篇文章均同时由一般评价者与专家评价者在不同维度进行评价，其中一般评价者 40 位（评价除科学性之外的其他指标）、专家评价者 3 位（评价科学性指标并进行作品甄别）。其中，专家评价者均已取得生物学博士学位且研究方向为微生物；一般评价者均为在校本科生，并被分为两组，每组 20 人，分别评价“微生物与社会”和“微生物知识”主题的科普文章。

表 2 实验中科普作品的类型、题目及编号

实验文章编号		科普文章题目	
		微生物与社会	微生物知识
科普文章创作者	A 人类创作版本	《我们的口腔，竟然有一座“微生物城”》（A1）	《寻找外星生命时，为何要先找微生物》（A2）
	B 文心一言初步提示版本	《口腔微生物：健康与疾病的微妙平衡》（B1）	《探索宇宙奇迹：为何寻找外星生命先从微生物着手》（B2）
	C ChatGPT 初步提示版本	《口腔中的微生物：看不见的守护者与潜在威胁》（C1）	《从微生物入手：寻找外星生命的科学起点》（C2）
	D ChatGPT 深度提示版本	《口腔里的隐秘世界：微生物的生存与战争》（D1）	《为何寻找外星生命，先从微生物开始》（D2）
	E DeepSeek 深度提示版本	《看不见的微型租客：探索口腔里的神秘世界》（E1）	《宇宙侦探手册：为什么外星搜寻要从显微镜下开始》（E2）

注：上述编号为 A–D 的 8 篇文章均于 2025 年 1 月 13 日生成，编号为 E 的 2 篇文章均于 2025 年 3 月 8 日生成。

2.2 实验过程

2.2.1 人类科普作品的选取

人类科普作品的选取方式为在“科普中国”网站中，选择“中国科普博览”科普号发表的作品①，以“微生物”为关键词检索，排除视频音频类科普作品和需图片传达重要信息的科普作品，并将字数限定在 1 500~1 800 字。最终得到 8 篇符合条件的人类作品，其中“微生物知识”主题的作品 3 篇、“微生物与社会”主题的作品 5 篇。考虑到人工智能已具备较强写作能力，选择人类作者的高质量作品与人工智能作品比较能够更好地回应研究问题，3 位研究者通过评价量表对 8 篇文章分别评价，选择各主题下得分最高的科普文章，作为本次实验的人类科普作品。

2.2.2 生成式人工智能科普作品的生成

对生成式人工智能的提示要消除其具有特征性的形式，并保持与人类作品主题的一

① “科普中国”是中国科协打造的提供科学、权威、准确的科普信息内容和相关资讯的平台，而“中国科普博览”在其介绍中被描述为“中国科协、中国科学院携手‘互联网 + 科普’平台，深耕科普内容创作”。

致性。实验采取以下提示策略:(1)呈现形式，根据同主题人类作品的文体特征，要求其不出现大小标题或生成若干小标题，以与人类创作版本一致;(2)内容主题，参照人类作品主题并自拟题目;(3)文字体量，1 500~1 800字;(4)写作标准，初步提示版本无提示，深度提示版本要求“内容科学准确、有趣、有吸引力、语言简洁通俗易懂不依赖专业术语”。

实验采用盲评方式，研究者告知评价者，其评价的每篇文章为随机抽取，可能为人类或生成式人工智能创作。在实验中，每位一般评价者对“微生物与社会”或“微生物知识”主题的5篇作品给出评分，专家评价者为全部10篇作品评分。为排除连续阅读对评价造成干扰，实验过程分5天完成。最终，每篇文章的科学性指标得到3个评价结果，其他指标得到20个评价结果，取均值作为最终结果。

3 实验结果分析

3.1 人机科普写作能力的总体评价结果

为直观了解各版本文章的整体表现水平与文章间的差异性，研究采取描述性统计、方差分析、独立样本t检验等方法分析不同科普作品的得分情况与显著性水平。结果表明(见表3)，在5类科普作品的总体表现方面，DeepSeek深度提示版本得分最高且标准差最小，说明经深度提示的DeepSeek在科普文章创作方面具有较高水平且不同评价者的评分间有较好的稳定性。人类创作版本的得分虽然高于文心一言初步提示版本，但低于ChatGPT初步提示与深度提示版本。经方差分析(p=0.003**)和每两组间的t检验，发现文心一言初步提示版本分别和ChatGPT初步提示版本(p=0.047*)、ChatGPT深度提示版本(p=0.0023**)以及DeepSeek深度提示版本(p=0.0004***)之间存在显著差异。

表3 科普文章得分汇总表

	N	M	SD	方差分析
E DeepSeek深度提示版本	40	73.03	1.128459329	
D ChatGPT深度提示版本	40	70.85	1.261990215	F=4.10
C ChatGPT初步提示版本	40	67.70	1.647297736	F0=2.65
A 人类创作版本	40	66.06	1.640248304	P=0.003**
B 文心一言初步提示版本	40	60.98	1.790997907	

注：* 表示p<0.05，** 表示p<0.01，*** 表示p<0.001。

进一步考察“微生物与社会”和“微生物知识”两个子主题的结果(见表4)，发现:(1)在“微生物与社会”主题下，DeepSeek深度提示版本表现最优，人类创作版本所得均分仅高于文心一言初步提示版本。经方差分析和每两组间的t检验，发现文心一言初步提示版本得分显著低于其余4篇实验文章，其他文章之间均无显著差异。(2)在“微生物知识”主题下，不同版本的人工智能创作文章在均分上均高于人类创作版本，其中DeepSeek深度提示版本表现最优。经方差分析和独立样本t检验，DeepSeek深度提示版本得分显著高于人类创作版本(p=0.0374*)，其余实验文章之间不存在显著差异。

表4 “微生物与社会”“微生物知识”两主题人机得分情况表

	“微生物与社会”科普作品				“微生物知识”科普作品			
	N	M	SD	方差分析	N	M	SD	方差分析
A 人类创作版本	20	67.96	1.857		20	63.48	1.542	
B 文心一言初步提示版本	20	56.66	1.750	F=4.13	20	65.16	1.930	F=1.42
C ChatGPT初步提示版本	20	68.67	1.625	F0=2.47	20	64.46	1.861	F0=2.47
D ChatGPT深度提示版本	20	70.07	1.358	P=0.004**	20	71.03	1.297	P=0.235
E DeepSeek深度提示版本	20	71.78	1.170		20	73.79	1.347	

注：* 表示p<0.05，** 表示p<0.01，*** 表示p<0.001。

3.2 人机科普写作能力的分指标评价结果

3.2.1 “易读性”评价结果

在“易读性”评价指标中(见表5)，人类创作版本得分最低，ChatGPT深度提示版本得分最高。经显著性检验(p=0.03*)，ChatGPT深度提示版本在“易读性”上显著优于人类创作版本。

3.2.2 “趣味性”评价结果

在“趣味性”评价指标中（见表6），DeepSeek深度提示版本得分最高。经显著性检验，DeepSeek深度提示版本得分显著优于人类创作版本（p=0.0018**）、文心一言初步提示版本（p=0.0009***）和ChatGPT初步提示版本（p=0.0005***），ChatGPT深度提示版本与人类创作版本之间差异不明显，但显著优于文心一言和ChatGPT初步提示版本（p=0.011*，p=0.047*），这说明不同提示程度对“趣味性”得分有显著影响。

3.2.3 “科学性”评价结果

在“科学性”表现中（见表7），人类创作版本优于ChatGPT深度提示版本、DeepSeek深度提示版本和文心一言初步提示版本。但经显著性检验，人类创作版本与其他人工智能创作版本之间不存在显著差异。

3.2.4 “传播效果”评价结果

在“传播效果”表现中（见表8），人类创作版本得分低于DeepSeek深度提示版本和ChatGPT深度提示版本，高于ChatGPT初步提示版本和文心一言初步提示版本。经显著性检验，DeepSeek深度提示版本在“传播效果”上显著优于人类创作版本（p=0.042*）、文心一言初步提示版本（p=0.046*）和ChatGPT初步提示版本（p=0.037*）。

表5 “易读性”指标单项得分汇总表

	“微生物”主题全部科普作品		“微生物与社会”主题科普作品		“微生物知识”主题科普作品	
	N	M	N	M	N	M
D ChatGPT深度提示版本	40	79.250	20	80.75	20	77.75
E DeepSeek深度提示版本	40	76.875	20	77.50	20	76.25
B 文心一言初步提示版本	40	73.250	20	73.00	20	73.50
C ChatGPT初步提示版本	40	71.875	20	75.75	20	68.00
A 人类创作版本	40	70.250	20	71.75	20	69.25

表6 “趣味性”指标单项得分汇总表

	“微生物”主题全部科普作品		“微生物与社会”主题科普作品		“微生物知识”主题科普作品	
	N	M	N	M	N	M
E DeepSeek深度提示版本	40	76.750	20	78.00	20	75.50
D ChatGPT深度提示版本	40	68.635	20	65.75	20	71.50
A 人类创作版本	40	61.000	20	63.00	20	59.00
C ChatGPT初步提示版本	40	58.375	20	59.00	20	57.75
B 文心一言初步提示版本	40	54.750	20	46.75	20	62.75

表7 “科学性”指标单项得分汇总表

	“微生物”主题全部科普作品		“微生物与社会”主题科普作品		“微生物知识”主题科普作品	
	N	M	N	M	N	M
C ChatGPT初步提示版本	6	84.167	3	76.67	3	91.67
A 人类创作版本	6	82.500	3	93.33	3	71.67
D ChatGPT深度提示版本	6	81.667	3	85.00	3	78.33
E DeepSeek深度提示版本	6	75.000	3	73.34	3	76.67
B 文心一言初步提示版本	6	61.667	3	56.67	3	66.67

表8 “传播效果”指标单项得分汇总表

	“微生物”主题全部科普作品		“微生物与社会”主题科普作品		“微生物知识”主题科普作品	
	N	M	N	M	N	M
E DeepSeek深度提示版本	40	65.375	20	64.00	20	66.75
D ChatGPT深度提示版本	40	56.375	20	53.75	20	59.00
A 人类创作版本	40	54.500	20	52.25	20	56.75
C ChatGPT初步提示版本	40	54.375	20	59.25	20	49.50
B 文心一言初步提示版本	40	54.250	20	50.25	20	58.25

3.3 人机作品的甄别结果

评价者对实验文章创作者（人类或人工智能）的判断结果显示（见表 9），人类创作版本的人类倾向性（即判断系人类创作的比例）仅为 52.17%，这说明评价者无法准确区分人机作品。从甄别正确率来看，4 篇出自人工智能之手的科普文章分别成功“欺骗”了 45.65%、54.35%、63.04% 和 52.17% 的读者。

表 9 人工智能创作者与人类创作者作品甄别效果汇总表

	全部评价者		
	N	人类倾向性 /%	甄别正确率 /%
A 人类创作版本	46	52.17	52.17
B 文心一言初步提示版本	46	45.65	54.35
C ChatGPT 初步提示版本	46	54.35	45.65
D ChatGPT 深度提示版本	46	63.04	36.96
E DeepSeek 深度提示版本	46	52.17	47.83

为了解评价者的甄别结果与文章评分的对应分布情况，通过交叉分析，发现那些被判断为人类所作的文章，整体得分高于被判断为人工智能所作的文章，皮尔逊相关系数与斯皮尔曼秩相关系数计算结果同样显示“甄别结果”与“均分”之间存在较强的负相关关系。这表明评价者倾向于给他们视作人类创作的文章打高分，而给视作人工智能创作的文章打低分，即评价者对人类创作者有更高的期待。为了解释这一现象，研究进一步分析了评价者在进行人机甄别时给出的原因，通过对原因内容进行聚类分析发现，评价者主要将易读性和趣味性（合计占比超过 50%）作为甄别人机作品的依据。一篇语言自然、内容生动又有吸引力的科普作品，会更容易被评价者认为是由人类创作，而人工智能创作的科普作品则更容易被评价者认为是格式化、生硬和缺乏情感的。有趣的是，这一主观认知与客观结果恰好相反，如 ChatGPT 深度提示版本在“易读性”“趣味性”上的得分均高于人类创作版本，这也解释了为什么 63.04% 的评价者将 ChatGPT 深度提示版本的作者判定为人类。

4 结论

本文基于实验数据，从 4 个维度探查了生成式人工智能的科普写作能力，分析了评价者对人机科普作品的甄别情况，得出以下结论。

第一，生成式人工智能具备替代人类科普创作者的潜力。ChatGPT 创作的科普作品得分（不论是初步提示还是深度提示版本）在趣味性、科学性和传播效果 3 个指标上均与人类创作版本无统计学上的显著差异。上述结果显示，机器作品能够很轻易地“迷惑”读者，评价者对所有生成式人工智能创作科普作品的甄别正确率均在 55% 以下，且 ChatGPT 深度提示版本能够“欺骗”63.04% 的评价者。换言之，人工智能已具有与人类近似的科普写作能力，读者无法对二者的作品作出区分。

进一步，ChatGPT 深度提示版本在“易读性”指标上显著优于人类创作版本，且 DeepSeek 深度提示版本在“趣味性”和“传播效果”两个维度的得分上显著优于人类创作版本。人类科普创作者在过去之所以不可或缺，是因为科普作者与科技工作者相比，具有写作语言通俗易懂、讲述方式直观形象、想象力丰富、知识面更宽等特征 [17–18]，然而上述结果不仅证明生成式人工智能挑战科普作者具有现实的可能性，也有力地回应了麦克对人工智能在科学传播中可能会产生“大规模错误”的担忧 [3]。但是需要特别指出的是，大型语言模型的核心原理是通过统计关联预测下一个词或句子，而不是真正理解科学概念或逻辑 [19–20]。例如，如果训练数据中“量子力学”和“平行宇宙”经常一起出现，

模型就可能会生成“量子力学证明了平行宇宙的存在”的表述，即使这种说法在科学上并不准确。

第二，人机合作科普创作是一种可行的创作模式。除了人工智能在科学问题“理解”上的不足，人类尚有人工智能不具备的优势，如人类创作者在“微生物与社会”主题上的表现整体优于文心一言，这说明人类创作者在科普写作的“价值”维度上能够更好地把握读者需求，在情感共鸣、伦理判断等方面展现专长，而已有研究表明，这些方面正是人工智能的不足之处[21-22]。未来在“人机合作科普创作”的模式下，人类创作者应“扬长避短”，更多将社会期待与人文价值融入科普作品的创作过程中，承担唤醒公众的科学理性意识的社会责任，促进和构建科学合理的社会价值体系[23]，避免只做科学知识的“搬运工”，而是做科学精神的“传播者”，如此方能发挥科普作者的独特价值。

需要注意的是，不同的生成式人工智能在科普写作方面表现出较大差异。在本文的研究范围内，DeepSeek 和 ChatGPT 比文心一言表现更优，文心一言在整体得分、趣味性、科学性等指标上均表现最差，其原因也许可以从在训练数据、模型架构、技术路线等方面的差异得到解释[24-25]。因此，“人机合作科普创作”的有效实现需要选择合适的生成式人工智能模型。同时，运用适当的提示工程（Prompt Engineering）亦可提升生成式人工智能的科普写作能力，如经过更细致提示的 DeepSeek 深度提示版本创作的科普作品在全部科普作品中整体表现最优。

第三，“人类能力幻觉”现象需引起关注。研究发现，评价者先验地认为人类的科普写作能力强于生成式人工智能，即倾向于将高分作品认定为人类创作。本文将这种人类的认知偏差现象称为“人类能力幻觉”。如果说“AI 幻觉”展现出 AI 本身的能力局限和人们对 AI 的不信任与担忧[26]，那么“人类能力幻觉”则展现出人类对生成式人工智能能力的认识不足——尽管有学者通过实证研究等方法证明目前生成式人工智能的写作在创造力等方面仍难以和人类媲美[27]，但是就科普写作而言，不得不承认这种“人类能力幻觉”已在事实上产生。这种“幻觉”的产生可能受人类中心主义以及对生成式人工智能抱有误解和偏见的影响，另一方面也说明生成式人工智能的科普写作能力并不为公众所知。

5 结语

本研究基于“微生物”主题，通过人机对比实验，从易读性、趣味性、科学性、传播效果以及人机作品的甄别等多个维度评估生成式人工智能的科普写作能力。研究发现，生成式人工智能具备替代人类科普创作者的潜力，人机合作科普创作是一种可行的创作模式，并提示关注“人类能力幻觉”现象。

本研究虽力求严谨，但仍存在局限。第一，专家评价者与一般评价者数量相对有限。第二，被评价的科普作品只涉及“微生物”主题的纯文字科普短文，难以反映其他科普主题、图文结合作品或长篇科普文章等类型作品。第三，本文的科普作品内容主要来自经典生物学理论，不能反映人工智能对前沿科技成果的科普能力。这些局限导致上述结论的可推广性有待进一步证实。

未来研究可进一步考察生成式人工智能在前沿科技或其他学科领域的科普写作，从而更全面地理解生成式人工智能与人类科普作者的能力差异，为新时代科普人才培养提供有益借鉴。

参考文献

[1] Markowitz D M. From Complexity to Clarity：How AI Enhances Perceptions of Scientists and the Public's Understanding of Science[J]. PNAS Nexus，2024，3(9)：387.

[2] Alvarez A，Caliskan A，Crockett M J，et al. Science Communication with Generative AI[J].Nature Human Behaviour，2024，8(4)：625-627.

[3] Schäfer M S. The Notorious GPT：Science Communication in the Age of Artificial Intelligence[J]. JCOM：Journal of Science Communication，2023，22(2)：1-15.

[4] 中国科协之声 . AI 大模型之下，科普何为？ [EB/OL].（2024-09-23）[2025-02-06]. http://www.kepu.gov.cn/sci-fi/2024-09/23/content_233498.html.

[5] Volk S C，Schäfer M S，Lombardi D，et al. How Generative Artificial Intelligence Portrays Science：Interviewing ChatGPT from the Perspective of Different Audience Segments[J]. Public Understanding of Science，2024：09636625241268910.

[6] Bui N M，Barrot J S. ChatGPT as an Automated Essay Scoring Tool in the Writing Classrooms：How it Compares with Human Scoring[J]. Education and Information Technologies，2024：1-18.

[7] Lin S，Crosthwaite P. The Grass is not always Greener：Teacher VS. GPT-Assisted Written Corrective Feedback[J]. System，2024，127：103529.

[8] Huang Y，Wu R，He J，et al. Evaluating ChatGPT-4.0's Data Analytic Proficiency in Epidemiological Studies：A Comparative Analysis with SAS，SPSS，and R[J]. Journal of Global Health，2024，14：04070.

[9] Revell T，Yeadon W，Cahilly-Bretzin G，et al. ChatGPT versus Human Essayists：An Exploration of the Impact of Artificial Intelligence for Authorship and Academic Integrity in the Humanities[J]. International Journal for Educational Integrity，2024，20：18.

[10] Dergaa I，Saad H B，Glenn J M，et al. A Thorough Examination of ChatGPT-3.5 Potential Applications in Medical Writing：A Preliminary Study[J]. Medicine，2024，103(40)：e39757.

[11] Kong X，Liu C. A Comparative Genre Analysis of AI-Generated and Scholar-Written Abstracts for English Review Articles in International Journals[J]. Journal of English for Academic Purposes，2024，71：101432.

[12] Li J，Huang J，Wu W，et al. Evaluating the Role of ChatGPT in Enhancing EFL Writing Assessments in Classroom Settings：A Preliminary Investigation[J]. Humanities and Social Sciences Communications，2024，11(1)：1-9.

[13] Lozić E，Štular B. Fluent but not Factual：A Comparative Analysis of ChatGPT and Other AI Chatbots' Proficiency and Originality in Scientific Writing for Humanities[J]. Future Internet，2023，15(10)：336.

[14] Kim J，Yu S，Detrick R，et al. Exploring Students' Perspectives on Generative AI-Assisted Academic Writing[J]. Education and Information Technologies，2025，30(1)：1265-1300.

[15] 和鸿鹏 . 生成式人工智能学术应用引发的伦理问题及其应对 [J]. 伦理学研究，2025(02)：115-122.

[16] 斯科特·L. 蒙哥马利 . 芝加哥科学传播指南 [M]. 杨文源，赵博，译 . 北京：科学出版社，2021：315.

[17] 李国昌，王凤林，龙昭月 . 科普工作新需求下作者队伍建设的对策 [J]. 出版科学，2020，28(1)：48-52.

[18] 周丽，王德福，姜华 . 大学出版社的作者培训模式和实现途径——以科普出版物的作者培训为鉴 [J]. 科技与出版，2012(2)：35-37.

[19] 陈小平 . 大模型关联度预测的形式化和语义解释研究 [J]. 智能系统学报，2023，18(4)：894-900.

[20] 陈小平 . 大模型：人工智能思想及其社会实验 [J]. 文化纵横，2023(3)：70-77.

[21] Zhang X，Zhang P，Shen Y，et al. A Systematic Literature Review of Empirical Research on Applying Generative Artificial Intelligence in Education[J]. Frontiers of Digital Education，2024，1(3)：223-245.

[22] 殷杰 . 生成式人工智能的主体性问题 [J]. 中国社会科学，2024(8)：124-145，207.

[23] 郑念 . 科普的社会责任及实现路径 [J]. 科学与社会，2011，1(4)：79-87.

[24] Yuan X，Shao C，Zhang Z，et al. Comparing the Performance of ChatGPT and ERNIE Bot in Answering Questions Regarding Liver Cancer Interventional Radiology in Chinese and English Contexts：A Comparative Study[J]. Digital Health，2025，11：20552076251315511.

[25] Wei X. The Use of Large Language Models for Translating Buddhist Texts from Classical Chinese to Modern English：An Analysis and Evaluation with ChatGPT 4，ERNIE Bot 4，and Gemini Advanced[J]. Religions，2024，15(12)：1559.

[26] Onder I，McCabe S. How AI Hallucinations Threaten Research Integrity in Tourism[J]. Annals of Tourism Research，2025，111：103900.

[27] Franceschelli G，Musolesi M. On the Creativity of Large Language Models[J]. AI & SOCIETY，2024：1-11.

（编辑 颜　燕　和树美）

there is a paucity of sufficient online textual data for GenAI training.Secondly，these topics often entail relatively higher uncertainty，necessitating a more cautious approach to content formulation. This study conducts a comparative analysis of science popularization content on frontier scientific achievements generated by traditional science popularizers and that produced by GenAI models. The results indicate no significant disparities between the two in terms of scientific accuracy and source reliability. However，the content generated by GenAI models tends to exhibit a more positive emotional tone compared to that produced by traditional science popularizers. Based on these results，this study suggests that while the risk of GenAI disseminating misinformation may be limited，it has the potential to reinforce existing stereotypes. Consequently，GenAI can not entirely supplant traditional science popularizers within the current framework of science popularization.

Keywords：generative artificial intelligence；frontier scientific achievements；science popularization
CLC Numbers：N4；TP18　Document Code：A　DOI：10.19293/j.cnki.1673-8357.2025.02.002

Evaluation on Science Popularization Writing Ability of Generative Artificial Intelligence：Science Popularization Creation Based on the Theme of “Microbiology”

He Hongpeng　Zhang Yaxin　Zhang Likai

（Institute for Advanced Studies in Humanities and Social Sciences，Beihang University，Beijing 100191）

Abstract：Generative artificial intelligence（AI）has emerged as a pivotal tool in the creation of science popularization literature; however，its capability in this domain remains subject to skepticism when compared to human authors. To evaluate the sciencepopularization writing competence of generative artificial intelligence，a four-dimensional evaluation framework was developed，comprising readability，engagement，scientific accuracy，and dissemination effectiveness. Using “microbiology” as the thematic focus，sciencepopularization works produced by human creators，DeepSeek，ChatGPT，and ERNIE Bot were systematically examined. The results indicate that both the initial-prompt and deep-prompt versions generated by ChatGPT showed no significant difference in overall scores relative to the human-authored version，with the deep-prompt version exhibiting a significantly superior performance in terms of readability. Moreover，the deep-prompt version of DeepSeek outperformed the human version in the dimensions of engagement and dissemination effectiveness. Notably，evaluators correctly distinguished AI-generated works in less than 55% of cases and showed a tendency to attribute high-scoring works to human creators. These experimental findings suggest that generative artificial intelligence holds the potential to substitute for humanscience popularization creators，thereby prompting the proposal of a novel “human-machine collaborative sciencepopularization creation” model. The study further calls upon the academic community to critically examine the phenomenon of the “Hallucination of Human’s Ability”.

Keywords：generativeartificial intelligence；science popularization writing；science communication；creation ability of artificial intelligence
CLC Numbers：N4；TP18　Document Code：A　DOI：10.19293/j.cnki.1673-8357.2025.02.003