构建 BERT模型：团队动态与研究重点如何塑造人工智能模型的影响力 t

Likun Cao（普渡大学社会学系；芝加哥大学知识实验室；0000-0001-5234-3855）**Kai Li**（田纳西大学诺克斯维尔分校，美国田纳西州诺克斯维尔；0000-0002-7264-365X）

通讯作者：Kai Li，kli16@utk.edu

摘要

以BERT系列模型为代表的人工智能技术迅速发展，改变了科学研究，但我们对这些技术在科学体系中的生产与认可动态知之甚少。本研究考察BERT系列模型的发展与影响，重点分析模型背后的团队规模、主题专业化和引文模式。我们使用来自Papers with Code（PWC）数据集的4,208篇BERT相关论文，分析BERT系列模型如何跨越不同的方法学世代演进，以及模型的新颖性与其生产和认可之间的相关性。研究结果表明，较新的BERT模型由规模更大、经验更丰富且机构背景更多元的团队开发，反映出人工智能研究日益增长的复杂性。此外，这些模型呈现出更高的主题专业化程度，面向细分应用，这与科学专业化的广泛趋势相符。然而，较新的模型获得的引文较少，尤其是从长期来看，这表明存在“先发优势”：BERT等早期模型获得了不成比例的认可。这些发现凸显了建立公平评估框架的必要性，使基础性创新和渐进式创新都得到重视。本研究揭示了人工智能研究中合作、专业化与认可之间不断演变的相互作用。

关键词：知识生产；人工智能；BERT；引文

该论文已被《定量科学研究》接受发表。

1 引言

科学与技术长期以来一直处于共同演化的关系之中，技术进步为科学探究提供工具与能力，拓展了科学探究的边界（Brooks, 1994; Rip, 1992）。进入21世纪，随着科学日益由数据驱动且计算密集，技术已成为推动科学发展的主导力量。这种日益加剧的不对称性体现在《AI4Science》这一概念（Stevens et al., 2020; Xie et al., 2024），以及以“计算”作为前缀的领域不断涌现，例如计算生物学和计算社会科学（Lazer et al., 2009; Nobel, 2002）。随着技术格局迅速演变，理解新兴工具——尤其是人工智能技术——的发展与应用，对于认识科学体系及其与技术进步之间的互动至关重要（Gao & Wang, 2024; Wang et al., 2023 ）。

过去十年间，人工智能领域最具变革性的发展之一，是 BERT（基于变换器的双向编码器表示）的问世。作为一种预训练语言模型，BERT 通过引入深度双向上下文理解，彻底改变了自然语言处理（NLP）（Devlin 等，2019）。与 ELMo（语言模型嵌入）和 LSTM（长短期记忆网络）等早期模型按顺序处理文本不同，BERT 采用基于变换器的架构，同时考虑左右两侧的上下文，从而形成更丰富的语言表示。这一创新使 BERT 成为现代自然语言处理和人工智能研究的基础模型，并在科学研究中得到广泛应用（Beltagy 等，2019）。

BERT 发布后，催生了一系列衍生模型，并逐渐发展为如今所称的 BERT家族。这些变体在功能、效率和领域适配方面均有所改进。值得注意的例子包括 RoBERTa（Liu 等，2019）、ALBERT（Lan 等，2019）和 BioBERT（Lee 等，2020）。这些模型的广泛涌现，凸显了 BERT 对人工智能领域的奠基性影响，也展现了人工智能工具在科学协作推动下的动态演进。

尽管 BERT系列模型在人工智能研究中意义重大，且自最初发布 BERT 以来已开发出众多变体，但人们对其发展历程如何随时间演变，以及它们在科学共同体中如何获得认可，知之甚少。本研究聚焦于这些 BERT系列模型的投入（团队规模和资历）与产出（以引文衡量的影响力），考察模型生命周期内的生产与认可条件。

本研究探讨三个关键问题：

1. 更新的 BERT模型在开发过程中是否需要规模更大、构成更多元的研究团队？2. 更新的 BERT模型是否聚焦于范围更窄的研究主题？3. 这些较新的模型获得的引文认可度是否相当?

通过探究这些动态，我们旨在深入了解人工智能发展与研究不断演变的结构，以及其对科学共同体如何认可科学领域技术进步所产生的更广泛影响。我们认为，这为理解人工智能发展及其在科学研究中的应用格局提供了新颖的视角。

根据我们的研究问题，我们提出了以下三个核心假设：

假设1：较新的BERT系列模型由规模更大、构成更多元的团队构建。这一假设基于一种广为接受的观点：科学研究越来越多地通过大规模合作开展（Cronin, 2021年；Larivière等人, 2015年；Wuchty等人, 2007年）。人工智能研究日益增长的复杂性和专业化程度，使机器学习、数据工程和特定领域知识（例如生物医学或法律应用）等领域的多元专业知识成为必要。此外，人工智能研究的全球性促进了国际合作，进一步推动团队构成多元化（AlShebli等人, 2024年；Hu等人, 2020年）。具体而言，基于这些观察，我们针对作者和机构两个层级提出了以下两个子假设：

假设1A：较新的BERT系列模型涉及更多合作者。随着人工智能模型日益复杂，其开发需要整合个体研究人员所具备的多样化技能，例如算法设计、大规模数据处理，以及针对特定应用进行微调。在这一子假设中，我们使用合作者数量这一指标来衡量科学合作的规模和复杂程度（Newman，2001年；Sonnenwald，2007年）。

假设1B：**较新的BERT系列模型涉及更多机构**。同样，我们假设人工智能模型日益需要来自不同机构的投入，这是《科学计量学》中使用的另一个重要指标（Gazni等，2012年）。

假设2：关于较新 BERT系列模型 的出版物具有更狭窄的概念范围。这一假设基于科学必然走向专业化这一事实：随着时间推移，科学领域会不断形成更新、更小的细分领域（Ben-David & Collin, 1966; Price, 1969; Wray, 2005）。此外，渐进式发展往往针对特定应用和现实场景量身打造，这一观察也支持了该假设（Palmer & Brookes, 2002）。据此，我们提出，较新的模型与较窄的主题覆盖范围相关，因为它们往往面向更具体的应用场景。

假设3：较新的 BERT系列模型 获得的引文更少。这一假设基于将引文视为可引用对象所获关注度指标的概念（Parolo et al., 2015; Petersen et al., 2014）。按照这一观点，可访问性更高（例如开放获取或提供公开数据集）或传播更广泛（例如通过社交媒体传播）的出版物，更有可能吸引引文（Langham-Putrow et al., 2021; Piwowar & Vision, 2013; Wang et al., 2015）。此外，提出新颖思想的研究往往会获得更多关注和引文（Bornmann & Daniel, 2008; Foster et al., 2015; Newman, 2009）。本研究假设，随着时间推移，较新的出版物获得的引文更少，因为与原始 BERT模型 相比，它们相对更具“衍生性”。

3 数据与方法

3.1 数据集

本研究使用了以下两个数据集：Papers with Code（PWC）和 OpenAlex 中的研究论文。PWC 是一个综合性的开源存储库，将学术论文与相应的代码实现关联起来（Kardas 等，2020年）。它主要涵盖机器学习和人工智能研究，并提供结构化元数据，例如论文标题、摘要、发表场所、研究任务、方法和性能基准。选择这一数据源，是因为它对“人工智能论文”的界定相对清晰明确。我们将 PWC 收录的所有出版物映射至 Unpaywall（https://unpaywall.org/），以获取论文 DOI，并利用检索到的所有出版物 DOI，从 OpenAlex 收集元数据。OpenAlex 是科学学和科学计量研究中最常用的研究数据库之一（Priem 等，2022年）。

PWC-OpenAlex数据集共包含477,718篇论文，其中9,485篇的摘要中提到了“BERT”一词。本研究聚焦于后者这一样本。在该样本中，有些“BERT”并非指 BERT 模型，而是具有其他含义，例如数学术语或人名。为了区分与 BERT 相关的术语（如 RoBERTa）和无关名词（如希尔伯特空间），并准确识别对真正的 BERT 概念的提及，我们使用深度学习模型进行了分类。

我们首先使用所有 Papers with Code 论文摘要对 BERT-base-uncased 模型进行微调。随后，我们为每篇摘要中的每个 BERT 相关概念生成上下文词嵌入。上下文词嵌入使我们能够根据同一个词在不同语境中的用法，为其生成不同的向量。例如，在句子“He sat on the river bank and watched the water flow”和另一句“She went to the bank to deposit her paycheck”中，“bank”一词会对应不同的嵌入向量：前者更接近水和《自然》等概念，而后者则与金融术语的关联更紧密。由于我们的数据主要

聚焦于 BERT 这一主题，因此预计大多数“BERT”实例指的是 BERT模型。因此，如果某个概念与“BERT”平均嵌入的余弦相似度超过预设阈值，就会被归类为真正的 BERT 概念。根据人工评估，我们将该阈值设为 0.3。为检验分类器的准确率，我们使用人工标注的验证集（样本量=500）对其进行验证（见附录 A）。模型的准确率达到 0.97，F1分数达到 0.95，证实了其在区分 BERT 相关概念方面的有效性。总计有 4,208 篇论文被识别为与“BERT”算法相关，而非与人名或数学术语相关。

我们还采用另一种样本纳入标准进行了稳健性检验。如附录B所示，结果依然一致。

3.2 变量

自变量：*BERT* 模型的代际。对于计算机科学研究而言，模型的代际可能与客观发表时间交织在一起。为区分 BERT 模型这两种“新颖性”的体现，我们使用引文关系量化了 BERT 论文的方法学代际。引文关系是衡量知识流动的常用代理指标（Alcácer & Gittelman, 2006；Kang 等人，2024；Price，1969年）。

我们首先基于引文链接构建了BERT相关术语的概念网络（例如，BERT支撑下游概念 BioBERT，因此从BERT向BioBERT绘制一条弧）。接下来，我们应用图层次算法（ Moutsinas 等人，2021年），为每个BERT 相关概念估计代际得分。最后，我们通过计算一篇论文摘要中出现的所有BERT 相关概念的平均代际得分，确定该论文的方法学世代。该流程在图 1的图 A中以概念图的形式加以说明。

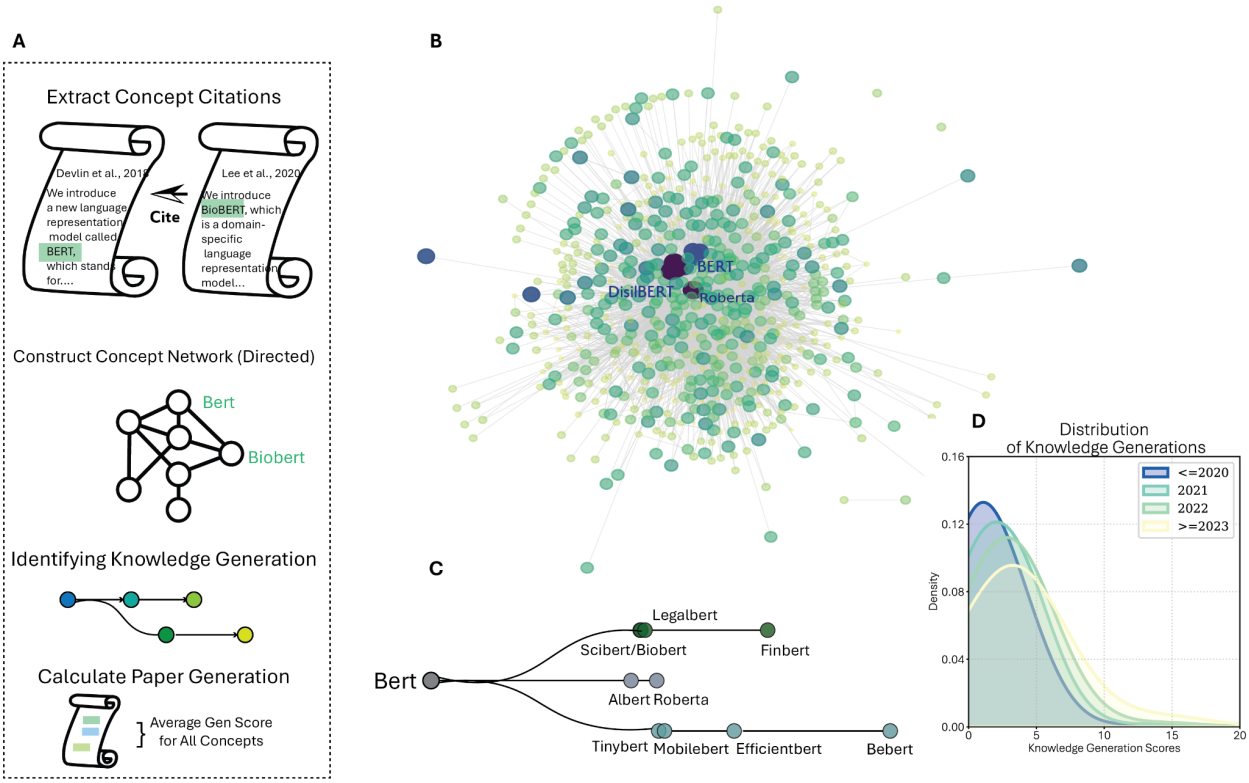

图 1. BERT模型的生成分数：示例与分布

面板B使用力导向布局（采用NetworkX软件包中的sfdp算法）展示了BERT相关术语之间的引文网络，而面板C则沿时间线呈现了若干知名BERT系列模型的代际得分。这些图表明，知识从较早的概念（如BERT、RoBERTa和DistilBERT）扩散到下游概念（如FinBERT和EfficientBERT）。图 1中的面板D显示了不同年份发表的论文的代际得分分布。

尽管代际得分与客观发表时间呈正相关，但这两项指标反映了知识生命周期的不同方面。当两项变量同时纳入模型时，代际得分反映的是目标 BERT 系列模型在知识发展轨迹中的位置——也就是说，在控制按时间顺序发表的影响后，该模型是否充当前驱并为其他模型提供支持。我们认为，代际效应在理论上十分重要，因为客观发表时间往往与时期效应及更广泛的历史趋势交织在一起，而代际得分则能捕捉我们旨在识别的内在动态：由知识生产的内部机制以及知识系统中的知识依赖关系结构所塑造的研究投入与产出变化。

因变量组 ***1***：研究投入。我们采用三组变量衡量研究投入，每组变量分别从作者和机构的角度反映项目所涉及的研究资源数量。**对于作者**，我们考察了：(1) 每项研究的作者数量；(2) 作者的平均职业年龄，按其首篇发表论文至今的年数计算；(3) 截至目标项目发表年份，每位作者此前获得的平均累计引用次数（经对数转换以降低过度离散）。这些变量反映了作者的经验和专业知识，以及他们的综合能力。**对于机构**，我们衡量了：(1) 每项研究涉及的不同机构总数；(2) 目标项目发表年份各机构发表的论文平均数量（经对数转换）；(3) 作者在各机构间的离散度，以 1 - 赫芬达尔指数衡量（根据作者在各机构间的分布计算）。这些变量既反映了各机构的研究能力，也体现了跨机构合作的程度。对于**国际合作**，我们考察了：(1) 根据机构所在地确定的参与研究项目的国家数量；(2) 作者在各国间的离散度，以 1 - 赫芬达尔指数衡量（根据作者在各国间的分布计算）。这些变量反映了特定项目涉及国际合作的程度。

因变量集合 ***2***：研究概念广度。我们采用最简单、最直观的指标之一——概念数量，来衡量每篇出版物的知识广度。OpenAlex 概念源自 Microsoft Academic Graph，因此可以与现有研究进行比较，并确保方法上的延续性。该分类体系将学术知识划分为三个层级。0级代表宽泛的学科，例如生物学、计算机科学、和社会学；而1级和2级则提供更具体的类别，例如机器学习（1级）和神经网络（2级）。研究概念的数量通常反映一篇论文涉及多个学科的程度。

本研究纳入了 OpenAlex API 提供的所有概念，未设置任何阈值。在附录中，我们进一步采用不同的截断分数（0.05、0.1 和 0.2），以确定纳入概念所需达到的最低主题相关性。我们使用这些阈值重新估计模型，并在附录表 22 中报告结果；结果显示，阈值对结果没有显著影响。

因变量集 ***3***：研究影响。鉴于计算机科学领域发展迭代迅速，且 BERT 研究历史较短（首篇论文发表于 2018 年），本文将论文发表后一年、两年和三年内获得的引文数量作为衡量短期至长期影响的指标。计算方法如下：

$$I_{1y} = \sum_{i=0}^{1} C_i \quad , \quad I_{2y} = \sum_{i=0}^{2} C_i \quad , \quad I_{3y} = \sum_{i=0}^{3} C_i \quad .$$

其中, C_i 表示每篇论文自发表以来第 i 年获得的引文数量。引文数量越多，说明论文的研究影响越大。

控制变量。 所有模型均控制了以下因素：(1) 发表渠道特征：会议（1 = 是，0 = 否）、核心合集（发表渠道是否被 OpenAlex 数据集标记为“核心来源”[1]）（1 = 是，0 = 否），以及期刊或会议的影响因子。 (2) 作者隶属与经验：任何工业关联（1 =是，0 = 否），以及作者在焦点发表之前发表的 BERT 相关论文的平均数量。 (3) 主要领域：纳入知识领域固定效应，以控制广泛学术领域之间的差异。领域分类依据 OpenAlex 数据中的 0 级领域索引，共包含 19 个不同领域。(4) 发表时间：还纳入了发表时间固定效应，以控制发表趋势随时间的变化。

对于研究投入模型，我们控制了一年前向被引次数，将其作为研究质量的代理变量，并控制了 2 级 概念的数量。对于知识广度模型，我们控制了一年前向被引次数、作者数量和机构数量。对于研究影响模型，我们控制了作者数量、机构数量以及 2 级 概念。

3.3 模型设定

为了估计方法学世代对研究资源和影响力的作用，我们进行了多元回归，并控制了发表场所特征、子领域和发表年份等关键混杂因素。由于因变量主要是计数数据（例如引文），我们采用了负二项回归模型。对于平均职业年龄等连续型因变量，我们采用了OLS回归。

完整模型设定如下：

$$ResearchInput_i = \beta_1 \times Generation_i + \beta_2 \times Controls_i + F_i + Period_i + \epsilon_i \quad (1)$$

$$KnowledgeBreadth_i = \beta_1 \times Generation_i + \beta_2 \times Controls_i + F_i + Period_i + \epsilon_i \quad (2)$$

$$ResearchImpact_i = \beta_1 \times Generation_i + \beta_2 \times Controls_i + F_i + Period_i + \epsilon_i \quad (3)$$

$ResearchInput_i$ 表示论文 i 所涉及的研究资源，包括与作者和机构相关的变量；$Controls_i$ 表示分配给焦点论文 i 概念的数量；$KnowledgeBreadth_i$ 表示论文 i 在给定时期（一年、两年或三年）内获得的引文数量；$ResearchImpact_i$ 表示论文 i 的

$Generation_i$

[1] 核心来源列表由莱顿大学科学与技术研究中心编制，并已用于莱顿开放排名，对世界各地的大学进行排名。

代际得分。此外，我们将论文 i 的一组特定特征作为控制变量（$Controls_i$），并在模型中纳入领域固定效应（F_i）。

为确保结果的稳健性，我们采用前向选择进行了逐步回归，以观察随着变量加入模型，系数如何变化。

我们还进行了额外的稳健性检验，以确保研究结果的可靠性，相关结果见附录B。所有分析均使用 Stata 18.0 完成。

3.4 描述性统计

所有变量的均值和标准差列于表1。所有变量的相关矩阵见表2。尽管同一类别中的变量（例如一年期、两年期和三年期引用）彼此相关，但大多数变量之间仅呈弱相关。

表1. 研究投入与研究影响的统计描述

	BERT相关论文（N=4208）			
	Mean	SD	Min	Max
作者数量	4.196	2.113	1.000	27.000
平均职业年龄	6.737	2.318	0.000	12.000
累计引用次数的对数	6.512	1.899	0.000	11.149
机构数量	1.787	1.155	1.000	15.000
论文总数的对数	8.631	3.742	0.000	12.849
机构分散度（1-赫芬达尔-赫希曼指数）	0.279	0.080	0.000	0.893
国家数量	1.298	0.352	1.000	7.000
国家分散度（1-赫芬达尔-赫希曼指数）	0.123	0.043	0.000	0.820
0级概念数量	3.197	1.256	1.000	8.000
1级概念数量	5.737	1.970	0.000	13.000
二级概念数量	5.420	2.476	0.000	14.000
一年引用次数	3.330	4.031	0.000	25.000
两年期引文数	6.368	6.095	0.000	25.000
三年引用次数	8.579	7.736	0.000	25.000
论文生成分数	2.350	2.125	0.000	14.880
出版年份	2021.477	1.368	2018.000	2024.000
发表场所：会议	0.755	0.430	0.000	1.000
发表场所：核心合集	0.177	0.382	0.000	1.000
发表场所：影响因子	1.977	3.365	0.000	27.041
任何工业关联（虚拟变量）	0.146	0.353	0.000	1.000
先前的 BERT 出版物	0.400	0.816	0.000	8.000

表 2. 所有变量的相关矩阵

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)
(1) 作者数量	1																					
(2) 平均职业年龄	0.14	1																				
(3) 累计引用次数的对数	0.25	0.61	1																			
(4) 机构数量	0.39	0.14	0.20	1																		
(5) 论文总数的对数	0.11	0.11	0.22	0.32	1																	
(6) 机构分散度（1-赫芬达尔-赫希曼指数）	0.39	0.15	0.20	0.87	0.17	1																
(7) 国家数量	0.26	0.11	0.16	0.65	0.23	0.59	1															
(8) 国家分散度（1-赫芬达尔-赫希曼指数）	0.19	0.11	0.14	0.53	0.09	0.60	0.93	1														
(9) 0级概念数量	-0.03	-0.09	-0.04	-0.02	0.04	-0.03	-0.02	-0.03	1													
(10) 1级概念数量	-0.03	-0.07	-0.01	0.01	0.09	0.00	0.02	0.01	0.70	1												
(11) 二级概念数量	-0.09	-0.06	-0.02	-0.01	0.08	-0.02	0.01	0.00	0.54	0.71	1											
(12) 一年引用次数	0.17	0.10	0.20	0.13	0.11	0.10	0.13	0.11	0.00	0.02	0.02	1										
(13) 两年期引文数	0.17	0.07	0.22	0.14	0.12	0.12	0.13	0.12	0.02	0.07	0.08	0.85	1									
(14) 三年引用次数	0.15	0.01	0.21	0.12	0.10	0.11	0.11	0.10	0.03	0.09	0.11	0.73	0.94	1								
(15) 论文生成分数	0.07	0.13	0.06	0.05	0.03	0.05	0.04	0.03	-0.03	-0.02	-0.06	-0.03	-0.06	-0.11	1							

（16）出版年份	0.10	0.40	0.13	0.07	0.05	0.04	0.06	0.05	-0.16	-0.18	-0.23	-0.06	-0.20	-0.36	0.29	1					
（17）发表场所：会议	-0.03	-0.15	-0.03	-0.07	-0.06	-0.03	-0.04	-0.00	0.06	0.04	0.03	-0.11	-0.07	-0.01	-0.02	-0.27	1				
(18) 发表场所：核心合集	0.05	0.14	0.05	0.09	0.10	0.03	0.04	-0.01	-0.05	-0.00	-0.00	0.12	0.10	0.06	0.05	0.22	-0.81	1			
(19) 发表场所：影响因子	0.11	0.16	0.18	0.12	0.14	0.08	0.11	0.08	-0.02	0.01	0.03	0.30	0.30	0.25	0.05	0.15	-0.33	0.33	1		
(20) 任何工业关联	0.20	0.09	0.10	0.17	0.08	0.16	0.14	0.12	-0.00	0.01	0.03	0.13	0.15	0.16	-0.02	-0.08	0.07	-0.08	0.04	1	
(21) 先前的 BERT 出版物	0.02	0.15	0.26	0.05	0.09	0.08	0.08	0.10	0.03	0.02	0.04	0.16	0.13	0.09	0.06	0.11	0.05	-0.03	0.10	0.06	1
(22) 虚拟变量：仅提及BERT	-0.05	-0.09	-0.06	-0.05	-0.04	-0.05	-0.04	-0.03	0.02	0.01	0.03	0.02	0.02	0.06	-0.83	-0.25	0.04	-0.06	-0.05	0.01	-0.06	1

4 结果

4.1 规模更大的合作团队和经验更丰富的作者参与了后续世代模型的开发

我们首先检验了假设1A（较新的BERT系列模型涉及更多合作者），重点关注研究团队中作者的数量与经验。这些模型报告于表3。针对我们的三个变量，我们分别估计了三个模型。除主要自变量（代际得分）外，第一个模型仅控制论文的宽泛领域；第二个模型纳入除客观发表时间之外的所有主要控制变量；第三个模型加入客观发表时间，而客观发表时间是代际得分最重要的混杂因素之一。 我们的研究结果表明，较晚世代的BERT模型通常由规模更大、且包含经验更丰富研究者的团队开发。根据我们最严格模型（表3中的模型3、6和9）的估计，代际得分从最小值（0）增加到最大值（14.88），与作者人数增加0.31人相关（$14.88 \times 0.021 = 0.31$），且每位作者的研究经验增加1.22年（$14.88 \times 0.082 = 1.22$）。这些结论得到多组控制变量的支持，证明了我们结论的稳健性。

我们的发现与研究领域中合作团队规模不断扩大的趋势一致，相关趋势已有充分记录（Cronin，2021年；Larivière等人，2015年；Wuchty等人，2007年）。在最严格的模型（模型3、6和9）中纳入出版年份固定效应后，我们证明，即使考虑到学术文化随时间发生的广泛变化，这些效应依然存在。为了进一步比较代际效应与时期效应（即出版年份）的大小，我们在模型中同时纳入这两个变量，并比较其标准化回归系数。结果见表8。请注意，表8报告的是标准化系数，而其他表格呈现的是非标准化系数，因此数值可能有所不同。标准化系数可以直接比较各变量的相对重要性，因为它们表示自变量每变化一个标准差，因变量将变化多少个标准差。

表8显示，在考虑出版年份后，我们的系数在显著性和方向上均保持一致。在预测作者数量时，代际得分的效应量略大于出版年份；而在预测作者经验时，出版年份的效应量更大。

一个可能的担忧是，代际得分可能捕捉到未被出版年份吸收的时间效应（例如，同一年1月与10月发表的模型之间的差异）。为此，我们采用不同的时间窗口来控制客观发表时间，进行了一系列稳健性检验，结果见附录表18–20。效应

量在这些设定中保持一致，这表明模型世代的影响在很大程度上独立于客观时间效应。

关于控制变量，较大且经验更丰富的团队也与更高的论文影响力相关，这一点体现为质量：一年期引用次数的正系数；同时，这类团队也更可能在影响力更高的发表场所发表论文，这一点体现为发表场所：影响因子。此外，我们的结果表明，即使考虑工业参与，核心发现依然成立（附录中的表18–20）。

接下来，我们检验了假设1B（较新的BERT系列模型涉及更多机构），结果见表4。原始模型（模型1、4和7）显示出正向且显著的系数，但在加入额外控制变量后，这些效应变得不显著（模型3、6和9）。在最严格的模型中，代际得分仅对机构分散度这一因变量表现出边际显著性（ β =0。008+ 模型9中）。这表明，代际得分从最小值（0）增加到最大值（14.88），与机构分散度指数提高0.119点相关（14。88 × 0。008 = 0.119）。关于控制变量，表4中的模型也表明，在高影响因子期刊或会议上发表的高质量论文，其作者往往更均匀地分布于多个机构。

最后，我们考察了不同世代的 BERT 算法之间国际合作的演变情况。人工智能研究领域的国际合作是一个颇具意义且日益受到关注的议题（AlShebli等人，2024年；Hu等人，2020年）。我们采用与表3和表4相[illegible]否更有可能纳入来自多个国家的作者，以及作者在这些国家之间的分布是否更加均衡。有趣的是，表5中的结果并不支持人工智能模型开发中的国际合作有所增加这一说法。尽管所有系数均为正，但没有一个具有统计显著性。表8显示，近年发表的 BERT 论文往往涉及更多国际合作，但这一效应并未体现在代际得分上。

简要总结如下：我们的模型提供了明确证据，支持假设1A；对假设1B仅提供部分且边际性的支持。此外，我们没有发现国际合作增加的证据。

T表3：世代对研究投入的影响的回归估计：作者 ors

	因变量：作者数量			因变量：平均职业年龄			因变量：log（累计引用次数）		
	模型1	模型2	模型3	模型4	模型5	模型6	模型7	模型8	模型9
论文生成	0.015***	0.026***	0.021**	0.129***	0.185***	0.082**	0.052***	0.038	0.022
	(0.004)	(0.006)	(0.006)	(0.017)	(0.029)	(0.028)	(0.014)	(0.024)	(0.024)
质量：一年引用次数		0.014***	0.014***		0.017+	0.033***		0.057***	0.063***
		(0.002)	(0.002)		(0.009)	(0.009)		(0.007)	(0.008)
二级概念数量		-0.019***	-0.019***		-0.025	0.035*		-0.026+	-0.013
		(0.004)	(0.004)		(0.017)	(0.017)		(0.014)	(0.014)
发表场所：会议		0.066*	0.132***		-0.499***	0.101		0.201+	0.400***
		(0.032)	(0.034)		(0.142)	(0.143)		(0.115)	(0.121)
发表场所：核心合集		0.091*	0.130***		0.186	0.264+		0.226+	0.281*
		（0.035）	（0.037）		（0.158）	（0.155）		（0.128）	（0.132）
发表场所：影响因子		0.006*	0.006*		0.052***	0.042***		0.055***	0.062***
		(0.002)	(0.003)		(0.011)	(0.011)		(0.009)	(0.010)
任何工业关联		0.235***	0.238***		0.564***	0.686***		0.365***	0.386***
		(0.020)	(0.020)		(0.098)	(0.093)		(0.079)	(0.079)
先前的 BERT 出版物		-0.003	-0.007		0.388***	0.307***		0.511***	0.527***
		(0.009)	(0.010)		(0.043)	(0.042)		(0.035)	(0.036)
领域固定效应	YES	YES	YES	YES	YES	YES	YES	YES	YES
发表时间固定效应			YES			YES			YES
常数	1.371***	1.260***	1.135***	6.322***	6.185***	4.280***	6.190***	5.767***	5.212***
	(0.029)	(0.051)	(0.063)	(0.137)	(0.229)	(0.267)	(0.114)	(0.186)	(0.227)
调整后的 R 平方/伪 R 平方	0.007	0.022	0.026	0.041	0.098	0.199	0.025	0.125	0.139
N	4130	4130	4130	4121	4121	4121	4121	4121	4121

注：标准误在括号 括号。$^{+}p < 0.1$，$^{*}p < 0.05$，$^{**}p < 0.01$，$^{***}p < 0.001$。所有检验均为双尾检验 .

表 4. 代际效应对研究投入的回归估计：机构

	因变量：机构数量			因变量：log（论文总数）			因变量：机构分散度		
	模型1	模型2	模型3	模型4	模型5	模型6	模型7	模型8	模型9
论文生成	0.012*	0.012	0.011	0.034**	0.015	0.006	0.006**	0.007+	0.008+
	(0.006)	(0.010)	(0.010)	(0.012)	(0.020)	(0.020)	(0.002)	(0.004)	(0.004)
质量：一年引用次数		0.010***	0.010***		0.024***	0.026***		0.004***	0.004**
		(0.003)	(0.003)		(0.006)	(0.006)		(0.001)	(0.001)
二级概念数量		-0.001	-0.002		-0.016	-0.008		-0.000	-0.001
		(0.006)	(0.006)		(0.012)	(0.012)		(0.002)	(0.002)
发表场所：会议		0.047	0.086+		0.043	0.170		0.017	0.028
		(0.049)	(0.052)		(0.101)	(0.107)		(0.020)	(0.021)
发表场所：核心合集		0.141**	0.163**		-0.001	0.023		0.032	0.039+
		(0.054)	(0.056)		(0.110)	(0.114)		(0.022)	(0.023)
发表场所：影响因子		0.009*	0.010*		0.029***	0.030***		0.003+	0.003*
		(0.004)	(0.004)		(0.008)	(0.008)		(0.001)	(0.002)
任何工业关联		0.274***	0.274***		-0.931***	-0.926***		0.116***	0.115***
		(0.031)	(0.031)		(0.064)	(0.064)		(0.013)	(0.013)
先前的 BERT 出版物		0.022	0.020		0.175***	0.174***		0.020***	0.020***
		(0.014)	(0.015)		(0.029)	(0.030)		(0.006)	(0.006)
领域固定效应	YES	YES	YES	YES	YES	YES	YES	YES	YES
发表时间固定效应			YES			YES			YES
常数	0.534***	0.374***	0.275**	10.052***	10.175***	9.719***	0.252***	0.187***	0.169***
	(0.046)	(0.078)	(0.098)	(0.097)	(0.162)	(0.197)	(0.019)	(0.032)	(0.039)
调整后的 R 平方/伪 R 平方	0.004	0.015	0.016	0.013	0.085	0.095	0.013	0.049	0.049
N	4130	4130	4130	3550	3550	3550	3550	3550	3550

注：括号内为标准误。$^{+}p < 0.1$,$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$。所有检验均为双尾检验。

表5. 代际效应对研究投入的回归估计：国际合作

	因变量：参与国家数量			因变量：国家分散度		
	模型1	模型2	模型3	模型4	模型5	模型6
论文生成	0.007	0.006	0.004	0.002	0.002	0.002
	(0.006)	(0.012)	(0.012)	(0.002)	(0.003)	(0.003)
质量：一年引用次数		0.009*	0.010**		0.004***	0.004***
		(0.003)	(0.004)		(0.001)	(0.001)
二级概念数量		0.003	0.004		0.001	0.001
		(0.007)	(0.007)		(0.002)	(0.002)
发表场所：会议		0.010	0.038		-0.000	0.009
		(0.056)	(0.060)		(0.015)	(0.016)
发表场所：核心合集		0.030	0.042		-0.011	-0.008
		(0.062)	(0.065)		(0.016)	(0.017)
发表场所：影响因子		0.007+	0.007		0.003*	0.003*
		(0.004)	(0.005)		(0.001)	(0.001)
任何工业关联		0.151***	0.154***		0.057***	0.058***
		(0.037)	(0.037)		(0.009)	(0.009)
先前的 BERT 出版物		0.029+	0.024		0.019***	0.017***
		(0.016)	(0.017)		(0.004)	(0.004)
领域固定效应	YES	YES	YES	YES	YES	YES
发表时间固定效应			YES			YES
常数	0.239***	0.143	0.052	0.095***	0.060*	0.035
	(0.053)	(0.091)	(0.114)	(0.014)	(0.024)	(0.029)
调整后的 R 平方/伪 R 平方	0.001	0.006	0.006	0.013	0.042	0.041
N	4130	4130	4130	3550	3550	3550

注：标准误见括号内。 $^{+}p < 0.1$，$^{*}p < 0.05$，$^{**}p < 0.01$，$^{***}p < 0.001$。所有检验均为双尾检验.

4.2 后世代的 BERT模型更加专业化。

接下来，我们讨论模型世代如何与 BERT相关论文中 概念的广度相关联。我们在控制所有控制变量的情况下，以代际得分为自变量，对 概念 数量进行回归分析。表6所报告的结果显示，后世代模型平均涉及的 概念更少。这些结论适用于大多数严格模型（ = -0.008* 模型3中的0级概念，以及 = -0.058* 模型9中的2级 概念）。因此，与早期 BERT模型相比，后世代模型往往更加专业化，专注于地方社群内部更为狭窄的问题。 假设2得到支持，即关于较新 BERT系列模型的出版物具有更狭窄的概念范围。

表8中的标准化回归系数显示，专业化趋势不仅适用于后世代研究，也适用于发表于更近年份的研究。事实上，对于这一结果，出版年份的效应量大于代际得分的效应量。

知识广度可被视为打破学科边界或整合来自不同领域洞见的信号（Rafols & Meyer，2010年）。从这个角度看，我们的发现与如下理论论点一致：开创性著作往往挑战既有的知识范式，而后续研究则倾向于遵循现有科学学科的规范与实践（Kuhn，1996年）。这一模式也反映了科学专业化程度的不断提高，这一趋势在科学发展的经典研究中被广泛讨论（Ben-David & Collins，1966年；Price，1969年；Wray，2005年）。它也呼应了近期的实证证据，即科学研究随时间推移变得不那么具有颠覆性（Park等人，2023年）。

表6. 代际效应对知识广度的回归估计

	因变量：0级概念			因变量：1级概念			因变量：2级概念		
	模型1	模型2	模型3	模型4	模型5	模型6	模型7	模型8	模型9
论文生成	-0.003	-0.011**	-0.008*	-0.008	-0.023	-0.005	-0.055***	-0.107***	-0.058*
	(0.002)	(0.004)	(0.004)	(0.010)	(0.019)	(0.019)	(0.015)	(0.027)	(0.026)
质量：一年引用次数		0.002	0.001		0.012*	0.009		0.018*	0.009
		(0.001)	(0.001)		(0.006)	(0.006)		(0.008)	(0.008)
作者数量		0.001	0.002		-0.010	-0.009		-0.097***	-0.091***
		(0.002)	(0.002)		(0.012)	(0.012)		(0.017)	(0.017)
机构数量		0.004	0.004		0.045*	0.045*		0.037	0.034
		(0.004)	(0.004)		(0.021)	(0.021)		(0.030)	(0.029)
发表场所：会议		0.058**	0.045*		0.293**	0.326***		0.281*	0.274*
		(0.018)	(0.019)		(0.090)	(0.095)		(0.128)	(0.133)
发表场所：核心合集		0.012	0.011		0.333***	0.476***		0.333*	0.566***
		(0.020)	(0.020)		(0.101)	(0.103)		(0.143)	(0.144)
发表场所：影响因子		-0.001	-0.001		0.006	0.001		0.015	0.013
		(0.001)	(0.001)		(0.007)	(0.008)		(0.010)	(0.011)
任何工业关联		0.016	0.011		0.082	0.026		0.237**	0.123
		(0.013)	(0.013)		(0.065)	(0.064)		(0.092)	(0.090)
先前的 BERT 出版物		0.012*	0.012*		0.006	-0.018		0.050	0.035
		(0.005)	(0.006)		(0.028)	(0.028)		(0.039)	(0.039)
领域固定效应	YES	YES	YES	YES	YES	YES	YES	YES	YES
发表时间固定效应			YES			YES			YES
常数	2.783***	2.756***	2.763***	5.178***	4.868***	4.683***	5.193***	5.355***	5.381***
	(0.017)	(0.028)	(0.034)	(0.086)	(0.141)	(0.172)	(0.122)	(0.200)	(0.241)
调整后的 R 平方/伪 R 平方	0.951	0.952	0.952	0.483	0.485	0.499	0.335	0.343	0.381
N	4130	4130	4130	4130	4130	4130	4130	4130	4130

注：括号中的标准误括号。$^{+}p < 0.1$，$^{*}p < 0.05$，$^{**}p < 0.01$，$^{***}p < 0.001$。所有检验均为双尾。

4.3 后续世代模型的引用影响力较低，这由较长期的引用影响力驱动

最后，我们考察代际得分与研究影响之间的关系。我们基于负二项回归，考察不同引文时间窗口（一至三年）下 BERT 研究获得的引用影响力。结果见表7。

表7中最严谨的模型显示，较新的模型平均而言往往具有较低的研究影响（$\beta = -0.034^{*}$，一年期引用的模型3中；$\beta = -0.075^{***}$，两年期引用的模型6中；以及$\beta = -0.102^{***}$，三年期引用的模型9中）。这些系数表明，随着时间推移，新旧模型之间的引用差距会扩大。代际得分从经验最小值（0）增加到最大值（14.88）时，一年期引用仅减少0.51（–0.034 × 14.88 = –0.51），但会导致三年期引用减少1.52（–0.102× 14.88 = –1.52）。比较表8中出版年份和代际得分的效应量后，我们发现代际得分的影响更大。这表明，早期世代模型的引用优势更多源于其在知识发展轨迹中的先发地位，而不仅仅是因为它们较早发表。因此，结果验证了假设3（较新的 BERT系列模型获得的引文更少）。

这一模式也得到了图2中描述性结果的支持。在该图中，我们根据代际得分将所有样本划分为十个规模相等的分位数分组，并计算其平均引文，从而能够在不加入额外控制变量的情况下直观呈现引用差异。尽管BERT论文的一年期引用影响力相对接近，但在三年期引用中差距更为明显：最早的模型获得的引文比最新世代分位数中的模型大约多50%。这一梯度在回归结果中也同样明显，如附录B中的表16所示。

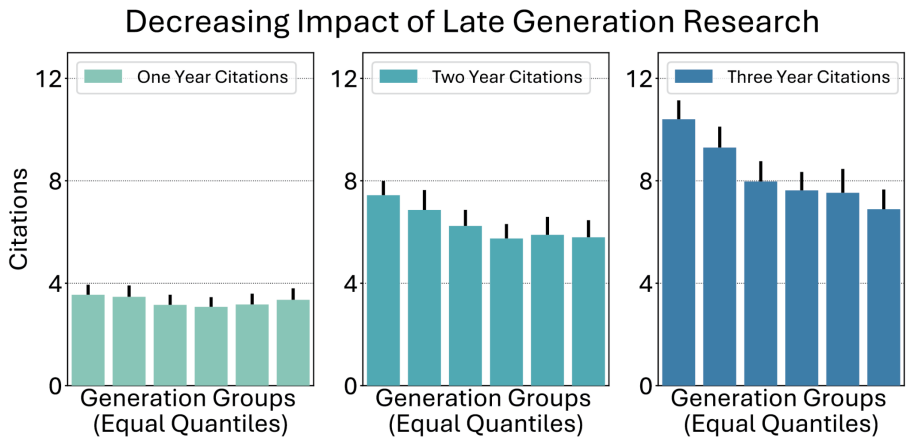

图2. 描　　述早期和后期BERT模型之间的引用差距

Table 7. 代际效应对研究影响的回归估计：引文 s

	因变量：一年期前向引用（2023年及之前的论文）			因变量：两年期前向引用（2022年及之前的论文）			因变量：三年期前向引用 2021年及之前的论文		
	模型1	模型2	模型3	模型4	模型5	模型6	模型7	模型8	模型9
论文生成	-0.009	-0.039*	-0.034*	-0.015+	-0.093***	-0.075***	-0.021*	-0.117***	-0.102***
	(0.009)	(0.016)	(0.016)	(0.008)	(0.015)	(0.015)	(0.009)	(0.019)	(0.018)
作者数量		0.072***	0.072***		0.055***	0.054***		0.051***	0.046***
		(0.010)	(0.010)		(0.009)	(0.009)		(0.010)	(0.010)
机构数量		0.013	0.014		0.037*	0.042**		0.022	0.020
		(0.018)	(0.017)		(0.015)	(0.015)		(0.017)	(0.017)
二级概念数量		0.004	0.005		0.020*	0.018*		0.017*	0.009
		(0.009)	(0.009)		(0.008)	(0.008)		(0.009)	(0.009)
发表场所：会议		0.065	0.245**		0.160*	0.357***		0.157+	0.354***
		(0.080)	(0.084)		(0.072)	(0.080)		(0.082)	(0.094)
发表场所：核心合集		0.203*	0.409***		0.239**	0.450***		0.225*	0.451***
		(0.087)	(0.090)		(0.080)	(0.086)		(0.091)	(0.101)
发表场所：影响因子		0.072***	0.071***		0.054***	0.060***		0.038***	0.044***
		(0.006)	(0.006)		(0.005)	(0.005)		(0.006)	(0.006)
任何工业关联		0.228***	0.183***		0.206***	0.167***		0.184***	0.167***
		(0.050)	(0.049)		(0.042)	(0.042)		(0.045)	(0.044)
先前的 BERT 出版物		0.133***	0.160***		0.085***	0.122***		0.090***	0.155***
		(0.021)	(0.021)		(0.020)	(0.020)		(0.025)	(0.026)
领域固定效应	YES	YES	YES	YES	YES	YES	YES	YES	YES
发表时间固定效应			YES			YES			YES
常数	1.241***	0.718***	1.266***	2.057***	1.632***	1.604***	1.805***	1.561***	1.517***
	(0.073)	(0.131)	(0.145)	(0.059)	(0.115)	(0.130)	(0.390)	(0.402)	(0.403)
调整后的 R 平方/伪 R 平方	0.004	0.026	0.044	0.004	0.023	0.030	0.004	0.017	0.024

N	3835	3835	3835	3109	3109	3109	2159	2159	2159

注：括号内为标准误。$^{+}p < 0.1, ^{*}p < 0.05, ^{**}p < 0.01, ^{***}p < 0.001$。所有检验均为双尾检验。

表8. 出版年份和世代的标准化回归系数

自变量\因变量	数量 作者	Ave. 职业年龄	log(累计 引用次数)	数量 机构	log(论文 数量)	机构 分散度	数量 国家	国家 分散度
出版年份	0.021***	0.372***	0.096***	0.025*	0.100***	0.007	0.046	0.038*
	(4.87)	(23.26)	(5.77)	(2.10)	(5.50)	(0.40)	(1.73)	(2.03)
论文生成	0.022**	0.075**	0.018	0.017	-0.001	0.053	0.010	0.015
	(3.27)	(2.92)	(0.69)	(0.92)	(-0.04)	(1.74)	(0.23)	(0.49)

自变量\因变量	0级 概念	1级 概念	2级 概念	一年 引文	两年 引文	三年 引文
出版年份	-0.017***	-0.079***	-0.156***	-0.032***	-0.026***	-0.019***
	(-4.29)	(-6.26)	(-11.03)	(-6.36)	(-8.72)	(-8.05)
论文生成	-0.013*	-0.004	-0.051*	-0.009	-0.026***	-0.026***
	(-2.13)	(-0.20)	(-2.22)	(-1.07)	(-5.00)	(-5.44)

注：括号内为标准误。$^{+}p < 0.1$，$^{*}p < 0.05$，$^{**}p < 0.01$，$^{***}p < 0.001$。所有检验均为双尾检验。控制变量与表*3-7*中的处于同一层级。该模型在同一模型中估计了发表时间效应和代际效应，并报告了标准化回归系数，因此可以直接比较各系数。

我们的发现与以下观察一致：突破性研究往往比渐进式研究具有更大的科学影响力（Schilling 与 Green，2011年），尤其是在长期影响方面（Cao 与 Evans，2024年）。更重要的是，新旧模型之间不同的引文模式，似乎反映了 Leydesdorff 等人（2016年）所描述的两种引文功能：短期引文表明与当前研究前沿的相关性，而长期引文则意味着知识被编码为概念符号。这一联系为理解人工智能研究的认可模式提供了新的见解。RoBERTa 等早期模型获得引用，不仅是因为其技术贡献，也因为它们在知识生产过程中具有象征意义。

4.4 稳健性检验

如果一篇论文在摘要中只提到“BERT”，那么它可能与 BERT 算法仅有边缘关联（例如，将 BERT 用作基准）。此类研究的性质可能不同于那些更直接参与 BERT 开发的研究。为了解决这一潜在问题并减少歧义，我们通过将这些论文排除在分析之外进行了稳健性检验，得到一个包含 2,743 条记录的子样本。随后，我们采用与主分析相同的设定估计回归模型。

附录B中呈现的结果与主要发现一致：在该子样本上进行的回归中，所有假设均得到支持。因此，我们认为我们的结论是稳健的，且不受样本选择的影响。

我们还通过对测量构建和时间窗口进行小幅调整，开展了几项额外的稳健性检验，以考察模型结果在不同条件下是否仍保持一致。这些检验的详情及结果见附录B。总体而言，我们的发现具有很强的稳健性，在不同的变量构建和模型设定下均观察到一致的实证模式。

5 讨论

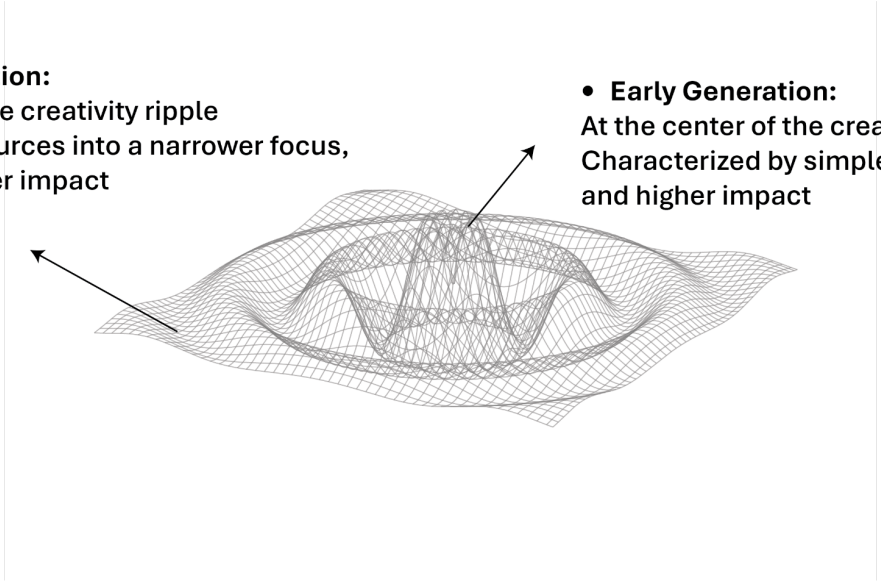

图 3. 人工智能发展中研究投入的收益递减的概念示意图

本研究以BERT系列模型为例，考察人工智能技术的产出及其影响，重点关注其开发所投入的资源——以参与作者和机构的数量及性质衡量——以及模型发表后三年内获得的引文。较新的BERT模型开发所需的资源更多，但其获得的长期引用往往少于前代模型。这与一种传统观点相悖：增量改进通常较为简单，因此只需要规模更小的合作团队（Beaver & Rosen，1978）。此外，本研究也为科学学研究中的一项发现提供了有价值的数据，即研究团队规模与创新之间总体上呈负相关（Tang等人，2024；Wu等人，2019）。此外，我们的结果还表明，该系列中的初代模型获得的引文也多于其后续模型。这一发现验证了Newman（2009）提出的“先发优势”在BERT系列模型领域的适用性。综合这些发现，我们可以得出结论：某一主题领域内的早期发展具有更高的投资回报。我们在图3中阐释了这一观点。

然而，上述观点从根本上不利于增量研究，而增量研究在知识生产与传播的生态中发挥着至关重要的作用（Wu等人，2019）。就我们的研究而言，许多较新的BERT模型都侧重于改进、应用和扩展经典模型，聚焦于更具体的应用场景，从而将方法学进展与更广泛的科学共同体联系起来。正如Newman（2009）所指出的，被高频引用并不意味着某一主题的首篇发表论文始终具有最高质量。我们认为，如果增量

研究持续得不到足够认可，可能会抑制那些推动人工智能技术演进与应用的必要且有效的改进和应用。

与此同时，BERT模型之间引文数量的差异，主要由长期引文驱动，而非由模型在第一年内获得的引文所致。这一发现为我们的论述增添了更多细微层次：所有模型最初都被视为研究前沿的一部分（Leydesdorff 等人，2016年），并得到积极讨论，融入学术话语；然而，较早出现的模型更有可能被作为研究主题的象征性代表而被引用，仅仅由于其历史先发优势，便累积了不成比例的引文份额（Small，1978）。这一现象体现了技术系列中学术关注分配的马太效应：早期模型持续累积引文，而这些引文往往与其具体贡献无关；相比之下，较新的模型尽管通常代表着实质性的改进与完善，却难以在长期内获得类似的认可。与此同时，较早的模型往往具有高得多的失败率（Franco与Filson，2006；Markides与Sosa，2013年），这进一步表明，我们所观察到的马太效应也可能是对这些先行者所付出冒险努力的 [illegible]。

我们的两项发现凸显了人工智能领域渐进式创新所面临的挑战，并表明亟需建立更公平的研究评估框架，在技术发展中既认可基础性突破，也重视渐进式进步。当前以引文为基础的指标虽有用处，却可能低估持续创新在人工智能领域的作用，从而可能使研究人员不愿从事必要的改进和领域特定应用。政策制定者、资助机构和学术机构应探索互补的评估机制，例如影响案例研究、定性同行评审和产业采纳评估，以更全面地了解这些模型的研究贡献。

6 结论

本研究全面分析了BERT系列模型的产出与影响，揭示了人工智能研究中的关键动态。我们的发现证实，较新的BERT模型由规模更大、经验更丰富且机构背景更多元的团队开发，这反映出人工智能发展日益增长的复杂性与协作性（Cronin，2001年；Larivière等人，2015年）。此外，这些模型在主题上表现出更强的专业化，获得的引文也更少，尤其是从长远来看；这凸显了“先发优势”——早期模型（如BERT）获得了不成比例的认可（Newman，2009）。

这些发现凸显了人工智能研究需要更公平的评估框架。基础模型推动着变革性发展，而渐进式进步对于实际应用和特定领域的发展同样不可或缺。政策制定者和学术机构应考虑采用互补指标，例如行业应用情况或定性影响评估，以更好地认可较新模型的贡献。此外，资源密集型的

现代人工智能发展的特点要求采取开放获取数据集和共享计算资源等举措，以促进包容性并减少全球科研参与方面的差距。值得注意的是，BERT模型各个世代在推出后的第一年获得的引文数量相近。尽管引文语境并非本研究的重点，但一年期引用与三年期引用之间差距的一种合理解释是，较早的模型获得了更多象征性引用（Small，1978），而这未必能直接体现其科学价值或实用价值。

尽管本研究有所贡献，但仍存在局限性。首先，它仅聚焦于BERT系列模型。尽管我们认为这些发现反映了人工智能研究与发展中更广泛的模式，但未来研究应比较不同的人工智能技术，以得出关于其在科学中的产出与认可的更具可推广性、数据驱动的结论。其次，PWC和OpenAlex数据集存在已知局限，例如OpenAlex中机构信息不完整（Zhang等人，2024年）以及引文不准确和缺失（Culbert等人，2025；Rodrigues等人，2025），这可能影响我们的分析。未来研究应利用更全面的数据来源以及在出版物中识别人工智能模型的先进方法（Gao和Wang，2024年），以构建更稳健的数据集，从而表征人工智能发展及其对科学的影响。这些步骤将增进我们对人工智能发展与认可如何嵌入科学体系的理解，而这是科学技术关系的一个关键方面。

声明

作者不存在需要披露的相关财务或非财务利益。

参考文献

Alcacer, J., & Gittelman, M. (2006). 专利引文作为知识流动的衡量指标：审查员引文的影响

s, J.

科

ragy, I., Lo, K., & Conan,

arX

ollins, R. (1966). 新兴科学形成过程中的社会因素：引文

》, *64*(1), 45-80. Brooks,

Cao, L., & Evans, J. (2024). 技术层级重组中的深度与广度。管理学会会议论文集,

202 tract Cronin, B. (2001). 超作

者 结构性转变的证据？美国信息科学与

Sch (25). OpenAlex与Web of Science及Scopus的

Lee 。载于

201 义论文集：人类语言技术，第*1*卷(pp. 4171-4186).

新。

《 企业：通

过 (12),

228 绘制全球科学合作图谱：

Der 35至2019年的文献计量学与网络分析。

(20 Kardas, M., Czapla,

P., Stenetorp, P., Ruder, S., Riedel, S., Taylor, R., & Stojnic, R. (2020). AllenAI：从机

器学 B. Webber, T. Cohn, Y. He, & Y. Liu (编), 2020年自

然 (*EMNLP*) (pp. 8580–8594). 计算语言学协会。

htt 命的

Bar 计算社会科学。《科学》,

Sor

arX

.11942.

Langham-Putrow, A., Bakker, C., & Riegelman, A. (2021). 开放获取引用优势是否真实存在[illegible] Larivière, V., Ghirlanda, Y., Sugimoto, C. R., & Tsou, A. (2015). 团队规模很重要：自 [illegible] J., Yoo[illegible]n, S., So, C. H., & Kang, J. (2020). BioBERT：一种用于生[illegible] Ley[illegible]文：质量指标？影响[illegible] *1*, 1. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., [illegible] *arX*[illegible]与先发优势：[illegible] C., Guo, W[illegible]生》 *11*([illegible]构。《美国国家科学院院刊》，[illegible] *86*([illegible]，[illegible]营销与[illegible] Leahey, E., & Funk, R. J. (2023). 论文和专[illegible] Par[illegible]sh, R., Huberman, B. A., Kaski, K., & Fortunato, S. (2015). 科学领域的[illegible]《信息计量学杂志》，*9*(4), 734-745. Petersen, A. M., Fortunato, S., [illegible]中的声[illegible] 21. Piwowar, H. A., & [illegible] (19[illegible]（第2次印刷）。哥伦比亚大学出版社。Priem, J., [illegible]场所、机构[illegible] Meyer, M. (2010). 多样[illegible] *82*([illegible]

Rip, A.（1992）。科学与技术作为共舞伙伴。载于工业时代的技术发展与科学：科学技术关[illegible]
Lo[illegible]个文献计量平台中的不一致引文数据：[illegible]
Sch[illegible]2011）。重组式搜索与突破性创意生成：对社会科学领[illegible]
（2[illegible]

Ste[illegible]s, R., Taylor, V., Nichols, J., Maccabe, A. B., Yelick, K., & Brown, D.（2020）。[illegible]
AN[illegible]0/17号）。阿贡国家实验室（ANL）。美国伊利诺伊州阿贡。Sugimoto, C. R., Sug[illegible]龄分层与[illegible]《信息处理[illegible]接人工[illegible]的大规模文献分析所揭示的启示。*arXiv*预印本[illegible]
Y.（2021）。科研国际[illegible]《信息处理与管理》，[illegible]（1），103[illegible]Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., [illegible] & Zitnik, M.（2023）。人工[illegible]Liu, C., Mao, W[illegible]取优势。[illegible]《科[illegible]，*35*（1），151-164。Wu, L., Wang, D., & Evans, J. A.（2019）。[illegible]44），378[illegible]生产中的主导地位[illegible]Di, Z., Wang, Y., & Hay[illegible]《自然：人类行为》[illegible]5（[illegible]），1314-1322。

Zhang, L., Cao, Z., Shang, Y., Sivertsen, G.和Huang, Y. （2024）。OpenAlex中缺失的机

附录：*BERT* 如何蓬勃发展

附录 A. 识别真实 BERT 概念的方法细节

我们采用深度学习模型来区分与 BERT 相关的术语（例如 RoBERTa）和无关名词（例如希尔伯特空间）。为实现这一目标，我们首先使用所有 Papers with Code 摘要对 bert-base-uncased 模型进行微调，使算法能够更好地捕捉信息，并识别计算机科学出版物中表达的逻辑关系。利用这个微调后的模型，我们为每篇摘要中的每个 BERT 相关概念生成上下文词嵌入。上下文嵌入会根据同一个词出现的不同语境，为其生成不同的向量。如果某个概念与“BERT”的平均嵌入之间的余弦相似度超过 0.3 的阈值，就会被归类为对 BERT 的真实指涉；该阈值是根据对语料库的人工阅读选定的。

为评估这一简单分类器的性能，我们使用人工标注数据集对其进行验证。具体而言，我们从 PwC 摘要中提取了 500 个包含 BERT 相关短语的句子，并对其进行人工标注。表 9 展示了若干示例。

表 9. 人工标注的 BERT 术语示例

句子	标签
● 相比之下，sapBERT 在其余数据集上的表现优于 xlinker。	True
● 在 vfas 中，向量既可以表示单个数据点，也可以表示函数空间中的元素，即再生核希尔伯特空间。	否
● 我们使用预训练模型 pubmedBERT 进行了实验，该模型针对这项任务在特定语料库上进行了微调。	True

我们应用分类方法，将真正的 BERT 术语与其他拼写相似的术语区分开来。与基于人工标注的真实标签相比，我们的方法准确率达到 0.97，F1分数达到 0.95，表明其在识别 BERT 相关概念方面效果显著。其他性能指标见表10。

表10. BERT分类器的分类报告

	精确率	召回率	F1分数	N
0（负样本）	0.97	0.87	0.92	83
1（正样本）	0.97	1.00	0.98	417

准确率			0.97	500
宏平均	0.97	0.93	0.95	500
加权平均	0.97	0.97	0.97	500

附录B. 稳健性检验

为评估研究结果的稳健性，我们进行了多项额外的稳健性检验。

首先，我们剔除了摘要中仅提及“BERT”、而未提及任何其他与 BERT 相关术语的论文。在原有的4,208条记录中，有2,711条符合这一纳入标准。表11至表15所报告的结果与基准模型的结果保持一致。

其次，我们用有序变量替代连续的代际得分，以考察其对论文影响力的影响。表16所示结果表明，随着代际得分增加，影响力逐渐下降。

第三，我们采用另一种方法计算代际得分。虽然我们继续使用正文中描述的图层次算法，但我们直接根据每篇论文的引文网络估算其代际得分。表17所报告的结果在研究影响模型中呈现出相似的模式，进一步强化了我们的结论。正如预期，由于该指标源自引文数据，对引文的影响更为显著。尽管研究投入模型中的代际效应大多不显著，但其方向仍保持一致。

第四，我们使用不同的客观时间窗口作为固定效应，重新估计了模型。这些稳健性检验旨在证明，代际效应代表一种独立的影响，无论如何衡量客观时间，这种影响都持续存在。表18–20所报告的结果与正文中的结果保持一致。

第五，我们通过删除所有自引，对引文模型进行了稳健性检验，以确保较早模型的引文数量较高并非由同一研究团队引用自身成果所致。表 21所报告的结果呈现出与表7一致的模式，证实了研究发现的稳健性。

第六，我们采用不同的概念纳入阈值进行了稳健性检验，分别将筛选阈值设为 0.1、0.2 和 0.5。我们依据这些不同标准构建知识广度，并重新估计模型。表 22 所报告的结果总体稳定，与正文中的结果一致。

最后，我们进行了额外分析，将工业参与作为因变量，以考察这一因素如何随模型发展而变化。结果见表 23。

表11. 稳健性检验1：回归估计　　基于选择性样本的代际效应对作者特征的回归估计　　e

	因变量：作者数量			因变量：平均职业年龄			因变量：累计引用次数的对数		
	模型1	模型2	模型3	模型4	模型5	模型6	模型7	模型8	模型9
论文生成	0.026***	0.027***	0.022***	0.186***	0.189***	0.082**	0.038	0.042+	0.021
	(0.006)	(0.006)	(0.007)	(0.030)	(0.030)	(0.029)	(0.025)	(0.024)	(0.024)
质量：一年引用次数		0.017***	0.017***		0.035**	0.042***		0.063***	0.068***
		(0.002)	(0.002)		(0.012)	(0.011)		(0.009)	(0.009)
二级概念数量		-0.023***	-0.023***		-0.019	0.051*		-0.015	0.003
		(0.005)	(0.005)		(0.022)	(0.022)		(0.017)	(0.018)
发表场所：会议		0.051	0.105*		-0.542**	0.019		-0.119	0.091
		(0.041)	(0.043)		(0.188)	(0.186)		(0.148)	(0.154)
发表场所：核心合集		0.078+	0.107*		0.116	0.148		-0.103	-0.061
		(0.045)	(0.046)		(0.206)	(0.199)		(0.163)	(0.165)
发表场所：影响因子		0.001	0.001		0.039**	0.036*		0.050***	0.057***
		(0.003)	(0.003)		(0.014)	(0.014)		(0.011)	(0.012)
任何工业关联		0.212***	0.218***		0.397**	0.572***		0.243*	0.286**
		(0.026)	(0.026)		(0.126)	(0.119)		(0.100)	(0.099)
先前的 BERT 出版物		0.000	-0.002		0.347***	0.293***		0.493***	0.508***
		(0.011)	(0.012)		(0.053)	(0.051)		(0.042)	(0.042)
领域固定效应	YES	YES	YES	YES	YES	YES	YES	YES	YES
发表时间固定效应			YES			YES			YES
常数	1.359***	1.316***	1.141***	6.342***	6.439***	4.332***	6.324***	6.108***	5.238***
	(0.040)	(0.059)	(0.087)	(0.188)	(0.272)	(0.366)	(0.153)	(0.215)	(0.304)
调整后的 R 平方/伪 R 平方	0.008	0.023	0.026	0.045	0.093	0.198	0.025	0.123	0.143
N	2633	2633	2633	2629	2629	2629	2629	2629	2629

注：括号内为标准误。$^{+}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$。所有检验均为双尾检验。

表12。稳健性检验2：.....的回归估计　　基于选择性样本的代际效应对机构特征的影响

	因变量：机构数量			因变量：log（论文总数）			因变量：机构分散度		
	模型1	模型2	模型3	模型4	模型5	模型6	模型7	模型8	模型9
论文生成	0.009	0.013	0.013	0.026	0.017	0.010	0.006	0.008+	0.009*
	(0.010)	(0.010)	(0.010)	(0.020)	(0.020)	(0.020)	(0.004)	(0.004)	(0.004)
质量：一年引用次数		0.012**	0.012**		0.021**	0.025**		0.004*	0.004*
		(0.004)	(0.004)		(0.008)	(0.008)		(0.002)	(0.002)
二级概念数量		0.001	-0.000		-0.011	-0.000		0.002	0.000
		(0.007)	(0.008)		(0.015)	(0.015)		(0.003)	(0.003)
发表场所：会议		0.015	0.026		0.104	0.182		0.005	0.006
		(0.063)	(0.066)		(0.125)	(0.131)		(0.025)	(0.027)
发表场所：核心合集		0.138*	0.142*		0.056	0.039		0.037	0.040
		(0.068)	(0.070)		(0.135)	(0.139)		(0.027)	(0.028)
发表场所：影响因子		0.006	0.007		0.032***	0.034***		0.002	0.003
		(0.005)	(0.005)		(0.009)	(0.009)		(0.002)	(0.002)
任何工业关联		0.283***	0.284***		-0.956***	-0.937***		0.131***	0.130***
		(0.038)	(0.039)		(0.078)	(0.078)		(0.016)	(0.016)
先前的 BERT 出版物		0.010	0.009		0.150***	0.160***		0.016*	0.017*
		(0.017)	(0.018)		(0.035)	(0.036)		(0.007)	(0.007)
领域固定效应	YES	YES	YES	YES	YES	YES	YES	YES	YES
发表时间固定效应			YES			YES			YES
常数	0.558***	0.388***	0.373**	10.124***	10.161***	9.607***	0.254***	0.183***	0.192***
	(0.061)	(0.091)	(0.131)	(0.127)	(0.183)	(0.255)	(0.025)	(0.037)	(0.052)
调整后的 R 平方/伪 R 平方	0.004	0.016	0.018	0.014	0.089	0.100	0.010	0.049	0.051
N	2633	2633	2633	2283	2283	2283	2283	2283	2283

注：括号内为标准误。$^{+}p < 0.1$，$^{*}p < 0.05$，$^{**}p < 0.01$，$^{***}p < 0.001$。所有检验均为双尾检验。

表13. 稳健性检验3：基于选择性样本的代际效应对国际合作的回归估计

	因变量：参与国家数量			因变量：国家分散度		
	模型1	模型2	模型3	模型4	模型5	模型6
论文生成	0.004	0.007	0.005	0.002	0.003	0.002
	(0.012)	(0.012)	(0.012)	(0.003)	(0.003)	(0.003)
质量：一年引用次数		0.009*	0.009*		0.004**	0.003**
		(0.004)	(0.005)		(0.001)	(0.001)
二级概念数量		0.005	0.006		0.002	0.002
		(0.009)	(0.009)		(0.002)	(0.002)
发表场所：会议		-0.025	-0.026		-0.015	-0.017
		(0.072)	(0.076)		(0.019)	(0.020)
发表场所：核心合集		0.020	0.011		-0.013	-0.016
		(0.079)	(0.082)		(0.021)	(0.021)
发表场所：影响因子		0.005	0.005		0.002	0.002
		(0.005)	(0.006)		(0.001)	(0.001)
任何工业关联		0.175***	0.178***		0.069***	0.070***
		(0.046)	(0.047)		(0.012)	(0.012)
先前的 BERT 出版物		0.022	0.018		0.016**	0.016**
		(0.020)	(0.021)		(0.005)	(0.005)
领域固定效应	YES	YES	YES	YES	YES	YES
发表时间固定效应			YES			YES
常数	0.268***	0.170	0.186	0.107***	0.073**	0.094*
	(0.072)	(0.106)	(0.152)	(0.019)	(0.028)	(0.039)
调整后的 R 平方/伪 R 平方	0.001	0.006	0.007	0.012	0.040	0.038
N	2633	2633	2633	2283	2283	2283

注：括号内为标准误。$^{+}p < 0.1$，$^{*}p < 0.05$，$^{**}p < 0.01$，$^{***}p < 0.001$。所有检验均为双尾检验。表14。稳健性检验4：基于选择性样本的代际效应对知识广度的回归估计

	因变量：0级概念			因变量：1级概念			因变量：2级概念		
	模型1	模型2	模型3	模型4	模型5	模型6	模型7	模型8	模型9
论文生成	-0.012***	-0.012***	-0.011***	-0.027	-0.025	-0.006	-0.123***	-0.109***	-0.058*
	(0.003)	(0.003)	(0.003)	(0.019)	(0.019)	(0.019)	(0.026)	(0.026)	(0.026)
质量：一年引用次数		0.002	0.001		0.013+	0.009		0.015	0.004
		(0.001)	(0.001)		(0.007)	(0.008)		(0.011)	(0.010)
作者数量		0.003	0.003		-0.025+	-0.023		-0.115***	-0.108***
		(0.003)	(0.003)		(0.015)	(0.015)		(0.021)	(0.021)
机构数量		0.003	0.004		0.055*	0.055*		0.055	0.047
		(0.004)	(0.004)		(0.025)	(0.025)		(0.036)	(0.035)
发表场所：会议		0.050*	0.044*		0.445***	0.408***		0.323*	0.265
		(0.019)	(0.020)		(0.117)	(0.121)		(0.164)	(0.167)
发表场所：核心合集		0.003	0.003		0.398**	0.490***		0.249	0.465**
		(0.021)	(0.022)		(0.128)	(0.130)		(0.181)	(0.179)
发表场所：影响因子		-0.002	-0.002		0.008	0.007		0.021	0.021
		(0.002)	(0.002)		(0.009)	(0.009)		(0.013)	(0.013)
任何工业关联		0.023+	0.021		0.063	-0.000		0.260*	0.124
		(0.013)	(0.014)		(0.081)	(0.080)		(0.114)	(0.111)
先前的 BERT 出版物		0.003	0.001		-0.005	-0.023		0.048	0.024
		(0.006)	(0.006)		(0.033)	(0.034)		(0.046)	(0.046)
领域固定效应	YES	YES	YES	YES	YES	YES	YES	YES	YES
发表时间固定效应			YES			YES			YES
常数	2.618***	2.561***	2.521***	5.035***	4.597***	4.484***	5.187***	5.143***	5.162***
	(0.019)	(0.027)	(0.039)	(0.115)	(0.163)	(0.233)	(0.163)	(0.230)	(0.321)
调整后的 R 平方/伪 R 平方	0.966	0.966	0.966	0.493	0.496	0.509	0.353	0.362	0.406
N	2633	2633	2633	2633	2633	2633	2633	2633	2633

注：标准误在括　　号。$^{+}p < 0.1$，$^{*}p < 0.05$，$^{**}p < 0.01$，$^{***}p < 0.001$。所有检验均为双尾　。

表15. 稳健性检验5：基于选择性样本的代际效应对研究影响的回归估计

	因变量：一年期前向引用（2023年及之前的论文）			因变量：两年期前向引用（2022年及以前的论文）			因变量：三年期前向引用（2021年及以前的论文）		
	模型1	模型2	模型3	模型4	模型5	模型6	模型7	模型8	模型9
论文生成	-0.029+	-0.043**	-0.044**	-0.078***	-0.096***	-0.080***	-0.111***	-0.121***	-0.109***
	(0.017)	(0.016)	(0.016)	(0.015)	(0.015)	(0.015)	(0.019)	(0.019)	(0.019)
作者数量		0.086***	0.085***		0.066***	0.064***		0.066***	0.060***
		(0.013)	(0.012)		(0.011)	(0.011)		(0.013)	(0.013)
机构数量		0.012	0.014		0.037+	0.050**		0.026	0.023
		(0.022)	(0.021)		(0.020)	(0.019)		(0.023)	(0.023)
二级概念数量		0.000	0.004		0.022*	0.018+		0.017	0.008
		(0.012)	(0.011)		(0.011)	(0.011)		(0.012)	(0.012)
发表场所：会议		0.004	0.219*		0.049	0.255*		0.098	0.342**
		(0.104)	(0.107)		(0.100)	(0.106)		(0.116)	(0.130)
发表场所：核心合集		0.203+	0.387***		0.132	0.332**		0.168	0.416**
		(0.112)	(0.115)		(0.108)	(0.114)		(0.129)	(0.141)
发表场所：影响因子		0.074***	0.070***		0.056***	0.062***		0.040***	0.044***
		(0.007)	(0.007)		(0.006)	(0.006)		(0.008)	(0.008)
任何工业关联		0.291***	0.250***		0.226***	0.176***		0.198***	0.185**
		(0.062)	(0.061)		(0.054)	(0.053)		(0.059)	(0.059)
先前的 BERT 出版物		0.144***	0.143***		0.084***	0.101***		0.094**	0.130***
		(0.026)	(0.025)		(0.024)	(0.024)		(0.033)	(0.035)
领域固定效应	YES	YES	YES	YES	YES	YES	YES	YES	YES
发表时间固定效应			YES			YES			YES
常数	1.251***	0.631***	1.029***	2.266***	1.660***	1.452***	2.484***	1.725***	1.544***
	(0.100)	(0.157)	(0.192)	(0.082)	(0.144)	(0.174)	(0.434)	(0.449)	(0.456)

调整后的 R 平方/伪 R 平方	0.005	0.033	0.047	0.008	0.028	0.035	0.009	0.023	0.026
N	2396	2396	2396	1852	1852	1852	1174	1174	1174

注：括号内为标准误。$^{+}p < 0.1$，$^{*}p < 0.05$，$^{**}p < 0.01$，$^{***}p < 0.001$。所有检验均为双尾检验。

表16。稳健性检验5：序数代际效应对研究影响的回归估计

	因变量：一年期前向引用（2023年及之前发表的论文）			因变量：两年期前向引用（2022年及之前发表的论文）			因变量：三年期前向引用（2021年及之前发表的论文）		
	模型1	模型2	模型3	模型4	模型5	模型6	模型7	模型8	模型9
论文生成：第2组	-0.001	-0.092	-0.057	-0.101	-0.175**	-0.124*	-0.120+	-0.180**	-0.137*
	(0.068)	(0.064)	(0.063)	(0.058)	(0.055)	(0.055)	(0.064)	(0.061)	(0.061)
论文生成：第 3 组	-0.071	-0.123+	-0.142*	-0.162**	-0.203***	-0.188***	-0.186**	-0.209***	-0.195**
	(0.070)	(0.066)	(0.064)	(0.060)	(0.057)	(0.056)	(0.063)	(0.061)	(0.060)
论文生成：第 4 组	-0.038	-0.095	-0.073	-0.222***	-0.285***	-0.217***	-0.332***	-0.384***	-0.326***
	(0.066)	(0.062)	(0.063)	(0.058)	(0.055)	(0.056)	(0.066)	(0.064)	(0.065)
控制变量同表15	YES	YES	YES	YES	YES	YES	YES	YES	YES
领域固定效应	YES	YES	YES	YES	YES	YES	YES	YES	YES
出版年份固定效应			YES			YES			YES
常数	1.181***	0.554***	0.967***	2.124***	1.501***	1.325***	2.186***	1.396**	1.235**
	(0.090)	(0.151)	(0.189)	(0.073)	(0.141)	(0.173)	(0.427)	(0.442)	(0.448)
调整后的 R 平方/伪 R 平方	0.005	0.033	0.047	0.007	0.027	0.034	0.008	0.022	0.025
N	2396	2396	2396	1852	1852	1852	1174	1174	1174

注：括号内为标准误。$^{*}p < 0.05$，$^{**}p < 0.01$，$^{***}p < 0.001$。所有检验均为双尾检验。我们将代际得分划分为四个等量的分位数分组，并将第一组作为该变量的参照组。

表17。稳健性检验6：基于替代测量的代际效应对研究影响的回归估计

	因变量：一年期前向引用（2023年及之前发表的论文）			因变量：两年期前向引用（2022年及之前发表的论文）			因变量：三年期前向引用（2021年及之前的论文）		
	模型1	模型2	模型3	模型4	模型5	模型6	模型7	模型8	模型9
论文生成（替代方案）	-0.488***	-0.507***	-0.578***	-0.499***	-0.525***	-0.556***	-0.532***	-0.536***	-0.553***
	(0.034)	(0.034)	(0.039)	(0.025)	(0.026)	(0.031)	(0.028)	(0.028)	(0.033)
作者数量		0.057***	0.048***		0.033***	0.029***		0.029***	0.024***
		(0.009)	(0.009)		(0.007)	(0.007)		(0.007)	(0.007)
机构数量		0.020	0.019		0.029*	0.032**		0.010	0.006
		(0.018)	(0.016)		(0.013)	(0.012)		(0.012)	(0.012)
二级概念数量		-0.004	0.004		0.010	0.009		0.004	-0.000
		(0.010)	(0.009)		(0.007)	(0.007)		(0.006)	(0.007)
发表场所：会议		-0.227**	0.034		-0.025	0.171*		0.026	0.190*
		(0.075)	(0.080)		(0.063)	(0.071)		(0.063)	(0.076)
发表场所：核心合集		0.040	0.207*		0.101	0.251***		0.082	0.224**
		(0.083)	(0.085)		(0.069)	(0.076)		(0.068)	(0.078)
发表场所：影响因子		0.055***	0.048***		0.044***	0.041***		0.030***	0.027***
		(0.004)	(0.005)		(0.003)	(0.003)		(0.004)	(0.004)
任何工业关联		0.149**	0.155***		0.115***	0.101**		0.124***	0.120***
		(0.047)	(0.044)		(0.035)	(0.034)		(0.031)	(0.030)
先前的 BERT 出版物		0.145***	0.132***		0.101***	0.096***		0.093***	0.091***
		(0.018)	(0.016)		(0.015)	(0.015)		(0.018)	(0.018)
论文组成部分固定效应	YES	YES	YES	YES	YES	YES	YES	YES	YES
领域固定效应	YES	YES	YES	YES	YES	YES	YES	YES	YES
出版年份固定效应			YES			YES			YES
常数	2.302***	2.049***	2.047***	3.054***	2.762***	2.631***	3.603***	3.337***	3.253***
	(0.088)	(0.138)	(0.145)	(0.064)	(0.100)	(0.112)	(0.063)	(0.100)	(0.114)

调整后的 R 平方/伪 R 平方	0.069	0.150	0.197	0.098	0.162	0.175	0.130	0.168	0.174
N	3103	3103	3103	2569	2569	2569	1793	1793	1793

注：括号内为标准误。* $p < 0.05$，** $p < 0.01$，*** $p < 0.001$。所有检验均为双尾检验。

表 18. 时间固定效应变化下代际效应对作者特征的回归估计

	因变量：作者数量			因变量：平均职业年龄			因变量：log（累计引用次数）		
	模型1	模型2	模型3	模型4	模型5	模型6	模型7	模型8	模型9
FE周期时间窗口	**1M**	**2M**	**3M**	**1M**	**2M**	**3M**	**1M**	**2M**	**3M**
论文生成	0.021**	0.022***	0.021***	0.085**	0.079**	0.084**	0.025	0.025	0.023
	(0.007)	(0.006)	(0.006)	(0.028)	(0.028)	(0.028)	(0.024)	(0.024)	(0.024)
质量：一年引用次数	0.015***	0.014***	0.014***	0.041***	0.037***	0.037***	0.068***	0.065***	0.065***
	(0.002)	(0.002)	(0.002)	(0.009)	(0.009)	(0.009)	(0.008)	(0.008)	(0.008)
二级概念数量	-0.020***	-0.020***	-0.020***	0.044**	0.038*	0.034*	-0.008	-0.012	-0.014
	(0.004)	(0.004)	(0.004)	(0.017)	(0.017)	(0.017)	(0.014)	(0.014)	(0.014)
发表场所：会议	0.142***	0.143***	0.136***	0.178	0.194	0.172	0.486***	0.489***	0.432***
	(0.037)	(0.036)	(0.035)	(0.155)	(0.150)	(0.145)	(0.133)	(0.127)	(0.124)
发表场所：核心合集	0.143***	0.140***	0.131***	0.319+	0.337*	0.314*	0.379**	0.385**	0.316*
	(0.040)	(0.039)	(0.037)	(0.166)	(0.162)	(0.157)	(0.142)	(0.138)	(0.134)
发表场所：影响因子	0.006*	0.005*	0.006*	0.039***	0.039***	0.040***	0.059***	0.058***	0.060***
	(0.003)	(0.003)	(0.003)	(0.011)	(0.011)	(0.011)	(0.010)	(0.010)	(0.010)
任何工业关联	0.239***	0.240***	0.240***	0.674***	0.676***	0.670***	0.389***	0.388***	0.383***
	(0.021)	(0.020)	(0.020)	(0.093)	(0.093)	(0.093)	(0.079)	(0.079)	(0.079)
先前的 BERT 出版物	-0.005	-0.005	-0.006	0.280***	0.296***	0.297***	0.524***	0.524***	0.526***
	(0.010)	(0.010)	(0.010)	(0.042)	(0.042)	(0.042)	(0.036)	(0.036)	(0.036)
领域/时间固定效应	YES	YES	YES	YES	YES	YES	YES	YES	YES
常数	0.462	0.468	0.475	4.621*	4.662*	4.679*	5.650**	5.661**	5.744**
	(0.709)	(0.709)	(0.709)	(2.069)	(2.076)	(2.075)	(1.768)	(1.770)	(1.770)
调整后的 R 平方/伪 R 平方	0.029	0.027	0.027	0.208	0.202	0.202	0.142	0.140	0.140

N	4130	4130	4130	4121	4121	4121	4121	4121	4121

注：括号内为标准误。$^{+}p < 0.1$，$^{*}p < 0.05$，$^{**}p < 0.01$，$^{***}p < 0.001$。所有检验均为双尾检验。

表19. 时间固定效应变化下代际效应对知识广度的回归估计

	因变量：0级概念			因变量：1级概念			因变量：2级概念		
	模型1	模型2	模型3	模型4	模型5	模型6	模型7	模型8	模型9
FE周期时间窗口	**1M**	**2M**	**3M**	**1M**	**2M**	**3M**	**1M**	**2M**	**3M**
论文生成	-0.008*	-0.008*	-0.008*	-0.004	-0.007	-0.005	-0.048+	-0.055*	-0.054*
	(0.004)	(0.004)	(0.004)	(0.019)	(0.019)	(0.019)	(0.026)	(0.026)	(0.026)
质量：一年引用次数	0.001	0.001	0.001	0.005	0.007	0.007	0.004	0.007	0.007
	(0.001)	(0.001)	(0.001)	(0.006)	(0.006)	(0.006)	(0.008)	(0.008)	(0.008)
作者数量	0.002	0.002	0.002	-0.007	-0.007	-0.008	-0.089***	-0.093***	-0.091***
	(0.002)	(0.002)	(0.002)	(0.012)	(0.012)	(0.012)	(0.017)	(0.017)	(0.017)
机构数量	0.003	0.003	0.004	0.036+	0.041+	0.041*	0.017	0.029	0.029
	(0.004)	(0.004)	(0.004)	(0.021)	(0.021)	(0.021)	(0.029)	(0.029)	(0.029)
发表场所：会议	0.050*	0.046*	0.043*	0.327**	0.307**	0.326***	0.245+	0.229+	0.250+
	(0.020)	(0.020)	(0.019)	(0.103)	(0.100)	(0.097)	(0.143)	(0.139)	(0.134)
发表场所：核心合集	0.023	0.015	0.013	0.517***	0.479***	0.507***	0.542***	0.523***	0.560***
	(0.022)	(0.021)	(0.021)	(0.110)	(0.108)	(0.105)	(0.153)	(0.150)	(0.146)
发表场所：影响因子	-0.001	-0.000	-0.000	0.001	0.002	0.001	0.013	0.015	0.016
	(0.002)	(0.002)	(0.001)	(0.008)	(0.008)	(0.008)	(0.011)	(0.011)	(0.011)
任何工业关联	0.012	0.010	0.010	0.028	0.026	0.024	0.115	0.131	0.114
	(0.013)	(0.013)	(0.013)	(0.064)	(0.064)	(0.064)	(0.089)	(0.089)	(0.089)
先前的 BERT 出版物	0.013*	0.013*	0.013*	-0.008	-0.020	-0.024	0.040	0.029	0.019
	(0.006)	(0.006)	(0.006)	(0.028)	(0.028)	(0.028)	(0.039)	(0.039)	(0.039)
领域/年份固定效应	YES	YES	YES	YES	YES	YES	YES	YES	YES
常数	2.959***	2.955***	2.960***	3.112*	3.159*	3.123*	4.430*	4.488*	4.426*

	(0.275)	(0.276)	(0.276)	(1.387)	(1.394)	(1.393)	(1.925)	(1.937)	(1.938)
调整后的 R 平方/伪 R 平方	0.953	0.952	0.952	0.507	0.502	0.502	0.398	0.390	0.389
N	4130	4130	4130	4130	4130	4130	4130	4130	4130

注：括号内为标准误。$^{+}p < 0.1$，$^{*}p < 0.05$，$^{**}p < 0.01$，$^{***}p < 0.001$。所有检验均为双尾。

表20. 代际效应对研究影响的回归估计：变化的时间固定效应

	因变量：一年期前向引用（2023年及以前的论文）			因变量：两年期前向引用（2022年及以前的论文）			因变量：三年期前向引用（2021年及以前的论文）		
	模型1	模型2	模型3	模型4	模型5	模型6	模型7	模型8	模型9
FE周期时间窗口	**1M**	**2M**	**3M**	**1M**	**2M**	**3M**	**1M**	**2M**	**3M**
论文生成	-0.027+	-0.027+	-0.027+	-0.069***	-0.068***	-0.069***	-0.097***	-0.097***	-0.099***
	(0.016)	(0.016)	(0.016)	(0.014)	(0.014)	(0.015)	(0.018)	(0.018)	(0.018)
作者数量	0.069***	0.069***	0.069***	0.051***	0.053***	0.052***	0.043***	0.044***	0.043***
	(0.010)	(0.010)	(0.010)	(0.009)	(0.009)	(0.009)	(0.010)	(0.010)	(0.010)
机构数量	0.015	0.017	0.018	0.044**	0.041**	0.044**	0.024	0.021	0.024
	(0.017)	(0.017)	(0.017)	(0.015)	(0.015)	(0.015)	(0.017)	(0.017)	(0.017)
二级概念数量	0.002	0.004	0.004	0.017*	0.017*	0.017*	0.008	0.010	0.010
	(0.009)	(0.009)	(0.009)	(0.008)	(0.008)	(0.008)	(0.009)	(0.009)	(0.009)
发表场所：会议	0.382***	0.332***	0.306***	0.572***	0.473***	0.407***	0.661***	0.505***	0.386***
	(0.095)	(0.092)	(0.086)	(0.098)	(0.092)	(0.083)	(0.120)	(0.109)	(0.099)
发表场所：核心合集	0.505***	0.477***	0.447***	0.652***	0.561***	0.493***	0.746***	0.603***	0.484***
	(0.100)	(0.098)	(0.091)	(0.103)	(0.098)	(0.089)	(0.126)	(0.114)	(0.105)
发表场所：影响因子	0.066***	0.069***	0.069***	0.061***	0.061***	0.061***	0.043***	0.046***	0.046***
	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)	(0.005)	(0.006)	(0.006)	(0.006)
任何工业关联	0.167***	0.178***	0.183***	0.166***	0.166***	0.174***	0.172***	0.173***	0.178***
	(0.048)	(0.048)	(0.048)	(0.041)	(0.041)	(0.041)	(0.044)	(0.044)	(0.044)
先前的 BERT 出版物	0.177***	0.171***	0.172***	0.143***	0.132***	0.134***	0.169***	0.160***	0.162***
	(0.021)	(0.021)	(0.021)	(0.020)	(0.020)	(0.020)	(0.026)	(0.026)	(0.026)
领域/年份固定效应	YES	YES	YES	YES	YES	YES	YES	YES	YES
常数	-0.012	-0.015	0.021	1.579*	1.677*	1.754*	1.637	1.852*	1.926*

	(1.128)	(1.137)	(1.138)	(0.793)	(0.798)	(0.800)	(0.837)	(0.828)	(0.831)
调整后的 R 平方/伪 R 平方	0.052	0.049	0.049	0.036	0.034	0.033	0.028	0.026	0.025
N	3835	3835	3835	3109	3109	3109	2159	2159	2159

注：括号内为标准误。$^{+}p < 0.1, ^{*}p < 0.05, ^{**}p < 0.01, ^{***}p < 0.001$。所有检验均为双尾检验。

表 21. 代际效应对研究影响的回归估计：引文（删除自我引用）

	因变量：一年期前向引用（2023年及之前发表的论文）			因变量：两年期前向引用（2022年及之前发表的论文）			因变量：三年期前向引用（2021年及之前发表的论文）		
	模型1	模型2	模型3	模型4	模型5	模型6	模型7	模型8	模型9
论文生成	-0.020*	-0.051**	-0.046**	-0.020*	-0.101***	-0.084***	-0.026**	-0.130***	-0.114***
	(0.010)	(0.018)	(0.018)	(0.009)	(0.016)	(0.016)	(0.010)	(0.020)	(0.020)
作者数量		0.065***	0.066***		0.048***	0.048***		0.045***	0.040***
		(0.011)	(0.011)		(0.009)	(0.009)		(0.010)	(0.010)
机构数量		-0.012	-0.010		0.021	0.026		0.004	0.002
		(0.019)	(0.019)		(0.017)	(0.016)		(0.018)	(0.018)
二级概念数量		0.007	0.007		0.022*	0.020*		0.019*	0.011
		(0.010)	(0.010)		(0.009)	(0.009)		(0.009)	(0.009)
发表场所：会议		0.068	0.221*		0.168*	0.356***		0.169+	0.371***
		(0.087)	(0.092)		(0.078)	(0.087)		(0.087)	(0.101)
发表场所：核心合集		0.176+	0.363***		0.237**	0.439***		0.242*	0.476***
		(0.095)	(0.098)		(0.085)	(0.093)		(0.096)	(0.107)
发表场所：影响因子		0.075***	0.074***		0.056***	0.063***		0.037***	0.043***
		(0.007)	(0.007)		(0.005)	(0.006)		(0.007)	(0.007)
任何工业关联		0.259***	0.212***		0.223***	0.184***		0.200***	0.181***
		(0.055)	(0.053)		(0.045)	(0.045)		(0.048)	(0.047)
先前的 BERT 出版物		0.117***	0.148***		0.074***	0.112***		0.085**	0.151***
		(0.023)	(0.023)		(0.021)	(0.022)		(0.026)	(0.028)
领域固定效应	YES	YES	YES	YES	YES	YES	YES	YES	YES
发表时间固定效应			YES			YES			YES
常数	1.137***	0.678***	1.312***	1.942***	1.561***	1.582***	1.388**	1.195**	1.161*

	(0.080)	(0.142)	(0.159)	(0.064)	(0.123)	(0.140)	(0.433)	(0.449)	(0.451)
调整后的 R 平方/伪 R 平方	0.004	0.024	0.041	0.005	0.021	0.028	0.004	0.016	0.022
N	3835	3835	3835	3109	3109	3109	2159	2159	2159

注：括号内为标准误。$^{+}p < 0.1$，$^{*}p < 0.05$，$^{**}p < 0.01$，$^{***}p < 0.001$。所有检验均为双尾检验。

表 22. 代际效应对知识广度的回归估计

	因变量：0级概念			因变量：1级概念			因变量：2级概念		
	模型1	模型2	模型3	模型4	模型5	模型6	模型7	模型8	模型9
概念纳入阈值	**0.05**	**0.1**	**0.2**	**0.05**	**0.1**	**0.2**	**0.05**	**0.1**	**0.2**
论文生成	0.008	0.003	0.002	0.002	0.012	-0.005	-0.074**	-0.072**	-0.069**
	(0.008)	(0.008)	(0.005)	(0.015)	(0.014)	(0.013)	(0.026)	(0.026)	(0.026)
质量：一年引用次数	-0.004	-0.002	-0.005**	0.005	0.006	0.006	0.015+	0.016+	0.015+
	(0.003)	(0.002)	(0.001)	(0.005)	(0.004)	(0.004)	(0.008)	(0.008)	(0.008)
作者数量	0.017**	0.021***	0.012***	0.011	0.011	0.007	-0.087***	-0.085***	-0.087***
	(0.006)	(0.005)	(0.003)	(0.010)	(0.009)	(0.008)	(0.017)	(0.017)	(0.017)
机构数量	-0.013	0.006	0.015**	0.013	0.023	0.018	0.036	0.035	0.039
	(0.009)	(0.009)	(0.005)	(0.017)	(0.016)	(0.014)	(0.029)	(0.029)	(0.029)
发表场所：会议	-0.212***	-0.224***	-0.153***	0.001	-0.011	0.115+	0.322*	0.314*	0.324**
	(0.041)	(0.038)	(0.023)	(0.071)	(0.067)	(0.061)	(0.126)	(0.126)	(0.125)
发表场所：核心合集	-0.244***	-0.186***	-0.099***	0.077	0.090	0.167*	0.350*	0.354*	0.355*
	(0.045)	(0.042)	(0.026)	(0.080)	(0.075)	(0.068)	(0.141)	(0.141)	(0.140)
发表场所：影响因子	-0.006+	-0.002	0.002	-0.005	-0.006	-0.005	0.013	0.013	0.011
	(0.003)	(0.003)	(0.002)	(0.006)	(0.006)	(0.005)	(0.010)	(0.010)	(0.010)
任何工业关联	-0.051+	-0.028	-0.027	-0.043	-0.054	0.014	0.260**	0.250**	0.266**
	(0.029)	(0.027)	(0.016)	(0.051)	(0.048)	(0.044)	(0.090)	(0.090)	(0.090)
先前的 BERT 出版物	0.018	0.005	0.005	-0.017	-0.021	-0.021	0.044	0.042	0.042
	(0.012)	(0.012)	(0.007)	(0.022)	(0.021)	(0.019)	(0.038)	(0.038)	(0.038)
领域固定效应	YES	YES	YES	YES	YES	YES	YES	YES	YES
发表时间固定效应	YES	YES	YES	YES	YES	YES	YES	YES	YES
常数	1.357***	1.266***	1.151***	3.245***	3.030***	2.562***	4.690***	4.619***	4.468***

	(0.064)	(0.059)	(0.036)	(0.111)	(0.106)	(0.095)	(0.197)	(0.197)	(0.195)
调整后的 R 平方/伪 R 平方	0.513	0.362	0.256	0.097	0.080	0.049	0.307	0.305	0.305
N	4130	4130	4130	4130	4130	4130	4130	4130	4130

注：括号内为标准误。$^{+}p < 0.1$，$^{*}p < 0.05$，$^{**}p < 0.01$，$^{***}p < 0.001$。所有检验均为双尾。表 23. 代际效应对工业参与的回归估计

	因变量：工业参与（完整样本）		因变量：工业参与（选择性样本）	
	模型1	模型2	模型3	模型4
论文生成	0.005	-0.061	-0.017	-0.046
	(0.022)	(0.043)	(0.041)	(0.043)
出版年份	-0.173***	-0.190***	-0.235***	-0.244***
	(0.035)	(0.040)	(0.045)	(0.052)
作者数量		0.227***		0.201***
		(0.024)		(0.030)
机构数量		0.249***		0.264***
		(0.038)		(0.047)
质量：一年引用次数		0.048***		0.060***
		(0.011)		(0.014)
二级概念数量		0.057*		0.059+
		(0.024)		(0.030)
发表场所：会议		0.188		0.205
		(0.200)		(0.262)
发表场所：核心合集		-0.688**		-0.728*
		(0.233)		(0.303)
发表场所：影响因子		0.027*		0.034*
		(0.014)		(0.017)
先前的 BERT 出版物		0.143**		0.106
		(0.053)		(0.066)
领域固定效应	YES	YES	YES	YES
常数	348.573***	381.554***	473.906***	489.282***
	(71.465)	(81.381)	(91.904)	(104.688)
调整后的 R 平方/伪 R 平方	0.019	0.104	0.031	0.115
N	4128	4128	2631	2631

注：括号内为标准误。$^{+}p < 0.1$，$^{*}p < 0.05$，$^{**}p < 0.01$，$^{***}p < 0.001$。所有检验均为双尾检验。