知识生成之谜：来自 AI 专利文本与引文信息的经验证据

董直庆　张　潇　余淼杰*

摘　要：AI 生成知识易出现“幻觉”，学术界尚未清晰识别创新背后 AI 扮演的角色。本文利用专利文本与引文数据，探讨 AI 对知识生成的影响，比较预测式与生成式 AI 的适用场景，并分析其在知识应用和唤醒“沉睡知识”方面的作用。研究表明，AI 不仅能通过知识新构与旧知识淘汰推动新知识产生，还能通过知识延伸与拓展激发后续创新，并提高知识新颖度和影响力。然而，AI 的效果受企业知识关联结构的制约。此外，AI 有助于推动知识整合与泛化，缩短知识应用周期，并唤醒有创新价值的“沉睡知识”。

关键词：预测式 AI；生成式 AI；知识生成过程
DOI：10.13821/j.cnki.ceq.2026.03.14

一、引　　言

AI 重塑了人类发现、表征与重组知识的方式，正引发一场创新范式的变革。在计算机等工具尚未普及的时代，研发人员主要依赖数学推演和重复实验，实现知识推断与验证。随着数字技术的发展，许多复杂问题能够通过建模和计算方法进行分析，从而为知识的量化分析与深度探索创造了条件。然而，当样本数据逐渐呈现出大规模以及多模态特征时，传统分析方式已难以精准刻画其内在规律。相比之下，AI 可以借助多层参数结构进行深度分析，揭示复杂信息中潜在的知识关联。例如，研发人员可以借助深度 Q 网络模型调整实验参数，减少实验次数并提升精度；也可以利用生成对抗网络缓解研发任务面临的信息缺失问题（Iwana and Uchida, 2021）。可以预见，伴随算法的持续迭代，AI 可以实现更高效的数据处理，并辅助人类进行实验设计和决策。

已有研究开始关注到 AI 对创新的积极影响，发现 AI 不仅可以帮助企业识

* 董直庆，浙江财经大学管理学院；张潇，华东师范大学经济与管理学院；余淼杰，辽宁大学。通信作者及地址：余淼杰，辽宁省沈阳市皇姑区崇山中路 66 号，110036；电话：13810003055；E-mail：mjyu@nsd.pku.edu.cn。本研究得到国家社会科学基金重点项目（24AJY029）的资助。感谢匿名审稿人和期刊编辑的宝贵建议，当然文责自负。

别潜在消费市场并推动新产品和新技术创新(Rammer et al., 2022; Babina et al., 2024),还可以通过持续数据反馈加速既有市场中技术的迭代与优化(Wang and Wu, 2026)。以药物研发为例,AI 可以快速分析大量生物数据,促进新药和新治疗方案的产生。借助高精度预测模型,企业能更加高效地识别关键治疗靶点和评估待选药物的效果,进而加快新药研发进程(Lou and Wu, 2021)。然而,现有文献多聚焦于 AI 对创新产出以及效率的影响,缺乏对创新内在过程的探讨。实际上,创新来源于知识生成过程,涉及知识识别、重组和再创造等多个阶段。AI 如何重塑这些过程,在哪些阶段产生作用,尚缺乏清晰的经验验证。此外,不同类型 AI 算法的任务目标以及训练逻辑均存在差别,对知识生成过程也会产生不同影响。但现有研究通常将不同类型 AI 视为同质化技术,忽视了其内在差异(Raisch and Fomina, 2024)。因此,探究 AI 对知识生成过程的影响,并明晰各类技术的应用场景,将有利于打开知识生成的"黑箱",助力企业实现关键技术的突破。

基于此,本文首先利用 2005—2022 年专利文本与引文信息数据,通过分析中心专利与其前后向引用专利之间的内在知识联系,从知识新构、知识接续、知识弃用、知识延伸及知识拓展维度,诠释 AI 在知识生成过程中的作用。其次,利用 BERT 模型[①]以及 LDA-SVC 模型[②]构建知识新颖性与影响力的指标,分析在知识生成过程中 AI 对知识属性的影响。再次,采用 LLM-BERT 分类方法[③]将 AI 划分为预测式和生成式两类,比较两者在知识生成过程中的作用以及在不同研发环境下的影响差异。最后,将研究视角延伸到知识应用过程,进一步探讨 AI 在缩短知识应用周期、唤醒企业"沉睡知识"等方面的可能性影响。

本文有三个主要发现:第一,AI 会通过知识新构和旧知识淘汰促进新知识的产生,也会以知识延伸和知识拓展方式诱发后续知识创新;在这个过程中,AI 会提升知识新颖性和影响力;相较于预测式 AI,生成式 AI 作用更明显。第二,在知识领域较为集中、核心知识一致性较高的情况下,预测式 AI 有助于推动知识生成,而在知识领域较为分散且核心知识一致性较低的条件下,生成式 AI 更能推动知识生成。此外,当企业内部知识与外部知识距离远且在知识网络中处于边缘位置时,AI 对知识生成的促进作用受限。第三,AI 有利于推动知识的整

① BERT 模型(Bidirectional Encoder Representations from Transformers, BERT)是一种基于深度语义表征的预训练语言模型,能够从专利文本中提取语义信息,并将文本转化为可计算的向量表示,可以用于文本相似度计算、文本分类等任务。

② LDA-SVC 模型(Latent Dirichlet Allocation-Support Vector Classifier, LDA-SVC)是一种将主题模型与支持向量机结合的分类模型,其中 LDA 用于识别文本中潜在的主题,SVC 则基于这些主题特征实现样本分类。

③ LLM-BERT 分类方法(Large Language Model-Bidirectional Encoder Representations from Transformers, LLM-BERT)是一种将大语言模型与 BERT 模型相结合的方法,借助大语言模型进行标签辅助生成,之后通过 BERT 模型对大规模文本进行稳定分类。

合泛化，缩短知识应用周期，并唤醒那些具有创新价值的“沉睡知识”。

本文对现有文献的贡献主要有三：一是从知识生成过程的视角重新审视AI在创新中的作用。本文不再囿于文献普遍关注AI对创新产出以及效率的影响，而是从知识新构、知识接续、知识弃用、知识延伸、知识拓展等维度量化知识生成过程，评估AI在知识生成各阶段中的作用。二是本文从企业知识关联结构的视角出发，拓展了AI创新效应差异的研究。相较于以往主要关注企业劳动力构成（姚加权等，2024；Brynjolfsson et al.，2025）、行业特征（李玉花等，2024）等因素，本文聚焦于知识多样性、核心知识一致性、知识距离及知识网络等微观维度，这一视角更有助于揭示知识创新的适用场景与条件。同时，比较预测式与生成式AI在不同知识关联结构下的表现，为理解其在创新过程中可能出现的“幻觉”现象提供经验启示。三是现有文献多集中于AI带来的创新效应，而对其在知识应用环节的影响关注不足。本文进一步剖析了知识的价值实现时间，探讨AI在缩短知识应用周期与唤醒“沉睡知识”方面的可能性作用，为企业加速释放知识价值打开了新视角。

二、文献综述与理论假说

AI正广泛用于研发实践，改变了现有的研发模式。传统研发过程主要依赖研究人员从现有理论、数据结构或偶然发现的信息关联中汲取灵感并进行实验验证。然而，人类在信息搜寻过程中不可避免地会受到认知能力的限制（Csaszar and Levinthal，2016）。而研发人员的认知局限性，使其在探索中易形成路径依赖而缺乏突破性创新。AI可以不受人类记忆以及思维惯性的影响，消除由人类偏好导致的认知偏差，实现在知识空间内的全局搜索。Weinan（2021）认为数理推演和数据分析是研发人员实现创新的重要手段，但随着技术复杂度的增加，完全依赖数理演绎已难以解决研发过程中的大部分问题，且数据集规模的增加使得个体认知局限愈发明显，研发越来越依赖高性能算法发掘有潜力的创新方向。Privett et al.（2012）利用AI进行材料搜寻，通过模拟合成大量蛋白质变体，发现了多种被人类忽视的催化活性酶。这表明，AI参与研发可改变传统知识生成路径并推动创新。

AI技术引发知识生成范式的变革，实现了知识的创新性重构。借助异构数据融合与特征提取技术，AI技术能够高效地将文字、图片、传感器信号等多模态数据转化为二进制数字信息，实现信息的高度整合（Chen et al.，2026）。在这个过程中，AI打破了不同领域知识的信息壁垒，促进了知识的跨领域创新性重构。Lipkova et al.（2022）构建了超声影像、MRI数据与电子病历的多模态知识图谱，揭示了患者耐药性差异的机制，发现了药物设计的新方向。此外，AI

可以识别和提取高价值的知识，实现对过时知识的淘汰。由于创新过程涉及众多因素，传统研究方法难以从海量信息中识别出有价值的内容（Khosravi et al.，2023）。AI 可以高效提取数据中的潜在信息，并通过提供可验证的假设为知识接续提供更精确的目标领域，并淘汰过时知识。Sendek et al.（2019）将 AI 与密度泛函理论结合，通过计算离子电导率，从 12 000 多种已有材料里筛选出具有创新价值的晶体材料。

此外，AI 驱动的创新是一个可持续改进、可延展的知识创新过程。AI 技术的可重复编程性，使其能够根据新的数据持续优化推理过程。在后续研发过程中，企业可以通过不断改进已有的算法模型，深化对特定领域的理解，促进知识延伸、拓展与深化（Yoo et al.，2010）。Tabor et al.（2018）认为 AI 的优势在于利用最近迭代的信息，为下一轮实验提供最优的设计方案，并强调了其在清洁能源领域后续知识延伸过程中的重要作用。此外，AI 还促进了新领域的知识拓展。AI 算法具有跨任务的通用性，基于标准接口的程序模块可以方便地与其他遵循相同标准的模块进行整合，参数也可在不同任务间迁移。这使得特定研发任务中形成的知识能够被应用于相似技术领域，实现知识的跨领域拓展。例如，Deepak and Ameer（2019）使用迁移学习算法，将通用的图像识别技术应用于肿瘤分类任务，实现了识别技术在健康医学领域的应用。

假说 1　AI 可以重塑知识的生成过程，促进知识的创新性重构，推动过时知识的淘汰，并促进后续创新过程中的知识延伸与拓展。

一般而言，AI 可区分为预测式与生成式 AI。其中，预测式 AI 的训练过程主要基于已有的标注数据，这些数据反映过去的实践经验和认知。数据标注的误差会导致预测式 AI 的结果存在“幻觉”，特别是在标注数据与实际场景存在差异时，这种偏差会降低知识创新的效率。此外，预测式模型的训练过程需要研发人员持续参与，通过设定具体的优化任务以确保输出结果的准确性（Choudhury et al.，2020）。相比之下，生成式 AI 主要基于自监督学习过程，无需工程师进行大量数据标记，而是通过学习数据的分布特征，深入挖掘数据中的信息，为创新活动提供广阔的知识搜索空间。

企业内部的知识关联结构会影响预测式与生成式 AI 知识生成。当企业知识多样性较高时，不同知识之间的联系往往更加复杂。由于预测式 AI 主要依赖标注数据进行学习，这种训练方式可能会丢失部分情境下的信息关联，使那些难以被编码的知识在学习过程中被边缘化，这会增加知识生成偏误的风险（Messeri and Crockett，2024）。相比之下，生成式 AI 对标注数据的依赖程度较低，更容易从多样化的知识中获取深层信息关联，从而促进新知识的生成。除了知识多样性外，核心知识一致性也会影响预测式和生成式 AI 的知识创新效应。预测式 AI 通常依赖研发人员对目标函数的设定，因此更适合技术路径

稳定的情境。当企业核心知识一致性较高时，稳定的技术方向有助于减少预测式模型优化目标的调整，从而提升其学习效率和输出质量。而生成式 AI 大多基于自监督训练过程，其生成的解决方案可能偏离既有的技术路径。在核心知识一致性较高的企业中，这种偏离会增加研发人员验证其可行性的负担。在这种情况下，模型输出结果易受到研发人员“算法厌恶”和“选择性遵从”[①]行为的影响，从而削弱其在知识生成中的实际作用（Alon-Barkat and Busuioc，2023）。

企业内部知识与外部知识的关联结构也会影响预测式与生成式 AI 的知识生成。当企业自有的内部知识与外部知识库距离较远时，预测式与生成式 AI 的效果会受到制约。具体而言，在学习外部知识的过程中，AI 可能出现应用场景受限、泛化能力减弱以及关键知识缺失等问题（Luo et al.，2019）。由于 AI 学习需要大量的训练数据，并且这些数据需要准确反映目标领域的知识结构和知识特点，在知识库距离较远的应用场景中，预测式与生成式 AI 可能均难以发挥预期的效果。此外，企业在知识网络中的位置也会显著影响 AI 应用的效果。通常而言，处于知识网络中心位置的企业能够为研发人员提供更丰富、多元的信息获取渠道，有助于提升企业的创新能力和推动知识创新。由于研发人员在面对海量信息时，需要高效地判断信息的价值并进行筛选，过多的无效信息可能会增加知识负担，甚至导致创新方向的混乱（Dong and Yang，2016）。预测式与生成式 AI 技术可以通过自动化数据分析和优化实验流程，提升信息处理效率，从而有效减轻个体在知识获取与处理中的负担。特别是那些处于知识网络中心位置的企业，凭借其广泛的连接性，可以借助 AI 从海量信息中提取有价值的知识，挖掘潜在的创新方向，从而加速创新进程。与之相反，处于网络边缘位置的企业，AI 的应用效果可能有限。

假说 2.1 预测式和生成式 AI 适用的场景不同，在知识多样性和核心知识一致性存在差异的情境下，两类 AI 对知识生成的作用表现迥异。

假说 2.2 当企业内部知识与外部知识的距离较远且在知识网络中处于边缘位置时，预测式与生成式 AI 的作用受到限制。

AI 技术的通用性可以助力多领域知识整合和知识泛化。AI 通过信息检索和匹配，实现不同领域的知识融合，并生成具体问题的解决方案，也可以通过迁移学习的方式，将原始模型中积累的知识迁移至新的模型或任务中，从而促进知识在不同任务之间的应用（Siriwardhana et al.，2023）。在研发过程中，AI 还能够作为知识整合与泛化的中介，促进不同专业知识背景的研发人员之间的沟通协调。特别是在分布式创新环境中，不同领域研发人员对问题认知的差异会增加沟通成本并降低创新效率（Seidel and O’Mahony，2014）。而 AI 功能模块

① 算法厌恶是指研发人员对算法或算法决策的抵触情绪或不信任感；选择性遵从是指研发人员在面对 AI 的建议时，有意识或无意识地只选择那些符合自己偏好、利益或价值观的部分来接受和执行。

的独立性和接口程序的标准化，降低了大规模研发团队的沟通协调成本。在项目开发过程中，研发人员可在已有代码模块基础上协作，不必深究代码模块的内部细节，这进一步提升了知识整合以及泛化的效率(Becker et al., 2021)。

然而，知识生成仅仅是创新链条中的起点，还需要将新知识转化为解决实际问题的技术方案，才能真正推动创新的落地。从新知识的生成到技术的实际应用，或者说从研发阶段的知识创造到技术创新的产业化应用，这一过程往往呈现出明显的双峰周期特征。一般而言，新知识的涌现会引发创新的首次繁荣，而随着时间的推移，关键问题的解决为知识应用开辟了更广阔的市场前景，产业化应用推动了创新的二次繁荣。在创新周期中这两个峰值间的时间差亦被称为知识应用周期(Dedehayir and Steinert, 2016)。然而，若市场需求快速变化，而创新主体不能迅捷地响应，易导致知识创新与实际应用之间的脱节。AI 能够通过对大规模市场数据进行分析，帮助企业更有效地识别市场需求，发现新知识可能对应的多种应用场景，缩短知识的应用周期(Cooper, 2024)。同时，AI 能够结合实时数据，对技术开发和应用过程中的关键决策进行优化。通过不断修正技术路线，企业可以更快完成技术迭代并推动知识向可实施技术方案的转化，从而加速知识的产业化应用。

AI 也有助于发现“沉睡知识”。新技术应用效益往往具有高度不确定性，因此企业在知识创新和应用时，倾向于重复使用经过验证的知识，以此来降低不确定性。通过知识重用能够最大限度地降低因盲目尝试而产生的风险，进而增强创新的可控性(Lee et al., 2024)。事实上，一些长期未被充分利用的知识也可能蕴含着丰富的信息，随着应用条件的成熟，这些处于“沉睡”状态的知识资源可以被激活并应用于研发过程。知识重用需要对过往信息进行系统总结，并把握新的应用场景。在此过程中，AI 具备天然的优势，能够高效分析历史数据，精准定位新场景所需的关键资源，并从历史信息中筛选出最相关和有价值的部分，为解决问题提供有力支持。Sourati and Evans(2023)利用 AI 对 4 000 余种已上市药物进行深入研究，发现它们可以治疗 100 多种新出现的疾病。

假说 3　AI 可助力知识整合泛化、缩短知识应用周期并唤醒有价值的“沉睡知识”。

三、变量指标设计、计量模型选择与数据来源说明

(一) 变量指标设计

现有研究主要采用两类方法刻画知识生成过程：一是采用搜索路径技术(Search Path Link Count, SPLC)等引文网络分析方法来追踪知识生产过程

（Huenteler et al.，2016）。这类研究通常以国际专利分类号（IPC）为依据界定技术范围与知识领域。IPC 体系根据专利所涉及的技术主题及其应用领域进行分层编码，从而实现对不同领域专利的系统化归类与整合。通过对专利 IPC 的分布、共现及其随时间的变化进行分析，可以识别特定技术领域的知识生成路径。二是基于自然语言处理等机器学习模型的语义相似度分析方法。这类研究通过深入理解专利文本中的语义内容，分析专利之间的语义关联，借助语义内容的演变来刻画知识生成过程。本文参考第一类方法，使用专利引文数据构建被解释变量。将知识生成过程分解为现有知识与前后向知识的联系。其中，后向联系主要从知识新构、知识接续与知识弃用三个维度刻画；前向联系则主要通过知识延伸与知识拓展两个维度刻画。

现有知识与后向引用知识的联系：①知识新构。如果知识元素 k 在中心专利 V_{ft} 中，但不在其引用的专利集合 V_b 中，表示中心专利在已有专利的基础上，发现新的知识。知识新构的量化指标为 $\sum_{V_{ft}=1}^{q_{ft}} \frac{\sum_{k\in V_{ft}}(1-k_{V_b})k_{V_{ft}}}{q_{ft}}$。其中，$V_{ft}$ 表示 f 企业 t 年授权的专利，q_{ft} 表示 f 企业 t 年授权专利的总量，V_b 表示 V_{ft} 专利引用的专利集合。k_{V_b} 取值为 0 或 1，若知识 k 在 V_{ft} 引用的专利集合 V_b 中，则取 1，否则为 0。同理，若知识 k 在中心专利 V_{ft} 中，则 $k_{V_{ft}}$ 取 1，否则为 0。②知识接续。若知识元素 k 同时在中心专利与其引用的专利中，表示知识接续过程，即现有知识是对已有知识的积累和传承。知识接续的量化指标为 $\sum_{V_{ft}=1}^{q_{ft}} \frac{\sum_{k\in V_b}k_{V_b}k_{V_{ft}}}{q_{ft}}$。③知识弃用。如果知识元素 k 不在中心专利中，但在其引用的专利中，表示知识弃用，即在创新过程中对旧知识进行了淘汰。知识弃用的量化指标为 $\sum_{V_{ft}=1}^{q_{ft}} \frac{\sum_{k\in V_b}k_{V_b}(1-k_{V_{ft}})}{q_{ft}}$。

现有知识与前向引用知识的联系：①知识延伸。如果知识元素 k 同时出现在中心专利 V_{ft} 与引用其的专利集合 V_d 中，表明知识在未来创新过程中得到了延续与发展，此过程为知识延伸。知识延伸的量化指标为 $\sum_{V_{ft}=1}^{q_{ft}} \frac{\sum_{k\in V_d}k_{V_{ft}}k_{V_d}}{q_{ft}}$。其中，$V_d$ 表示引用 V_{ft} 的专利集合，k_{V_d} 取值为 0 或 1，若知识 k 在引用 V_{ft} 的专利集合 V_d 中，则取 1，否则为 0。②知识拓展。若知识元素 k 不在中心专利中，但在引用其的专利中，表示在现有知识的基础上，通过进一步探索与实践，实现了知识边界的突破，即知识拓展。知识拓展的量化指标为 $\sum_{V_{ft}=1}^{q_{ft}} \frac{\sum_{k\in V_d}(1-k_{V_{ft}})k_{V_d}}{q_{ft}}$。本文将专利视为由多项技术知识构成的知识集合，并以专利所属的 IPC 小组代

码表示其所包含的知识元素。在构建被解释变量时，本文剔除了包含 AI 领域 IPC 的专利样本，保证知识生成过程的度量主要反映企业在非 AI 领域的创新活动，尽可能缓解由指标构建引发的内生性问题。

核心解释变量是企业的 AI 水平，部分文献使用企业年报测算该指标（姚加权等，2024）。然而，年报往往代表企业对 AI 的关注度，而非真实的 AI 水平。由于专利数据常用于刻画企业对特定技术的掌握程度，本文使用企业当年 AI 相关专利占所有专利的比重来度量。同时，参考王林辉等（2022），依据世界知识产权组织《2019 年人工智能技术趋势报告》（*Technology Trends* 2019）及国家知识产权局《关键数字技术专利分类体系（2023）》中公布的 AI 相关 IPC 分类号，筛选 AI 专利。当然，企业所使用的 AI 技术并非全部源于自主研发，也可能通过外部购买获得。若仅考虑企业自主研发的 AI 专利，则可能低估其实际技术水平。为此，结合专利转移数据，识别企业从外部获取的专利类型。具体而言：首先，利用专利转让数据库中有关专利权人变更的记录，追踪专利在企业间的转移路径；其次，识别在样本期内转让至该企业的 AI 专利；最后，将外部获取的 AI 专利与自主研发的相关专利合并，构建修正后的 AI 水平指标，并将其用于稳健性检验。

（二）计量模型选择与数据来源说明

为检验 AI 的知识生成效应，构建如下计量模型：

$$KTC_{ft}=\beta_0+\beta_1 AI_{ft}+\boldsymbol{\gamma}\boldsymbol{X}_{ft}+\mu_f+\theta_h+\sigma_t+\varepsilon_{ft}, \tag{1}$$

其中，KTC_{ft} 表示知识生成过程，下标 f、t 分别表示企业和年份。核心解释变量 AI_{ft} 表示企业 AI 水平，$\boldsymbol{X}_{ft}$ 包括了可能影响知识生成的一系列控制变量。ε_{ft} 为随机扰动项。参考 Rammer et al.（2022），选取以下控制变量：研发人员比例，用研发人员与总员工数量之比表征；研发支出比例，用研发投资在总支出中的份额表征；企业规模，用企业总资产的对数值表征；企业年龄，用企业存续时间对数值表征；企业负债率，用企业负债与总资产之比表征；净资产收益率，用企业净利润与平均净资产之比表征；流动资产周转率，用营业收入与流动资产总额的比值表征；固定资产占比，用固定资产净额与总资产的比值表征；无形资产占比，用无形资产净额与总资产的比值表征。同时，为保证估计结果的稳健性，回归模型控制了企业固定效应 μ_f、行业固定效应 θ_h 和时间固定效应 σ_t。

历年专利数据来自中国企业大数据平台（RESSET）的知识产权数据库以及中国研究数据服务平台（CNRDS）的专利引用与被引用数据库。按照授权公告号将专利引文信息与其摘要文本信息进行匹配。企业层面的数据来自国泰安数据库（CSMAR），将处理后的专利数据与上市企业财务数据进行匹配，并剔除金融行业以及当年处于 ST 和*ST 状态的样本。对于财务数据缺失的样本，按照行业均值或者零值进行填补。为了消除极端值的影响，本文对变量

在上下 1%的水平上进行缩尾处理。专利转移数据来源于中国开放数据平台(CnOpenData)的专利申请权和专利权转移数据库。

四、实证检验结果与评价

（一）基准回归

表 1 展示了 AI 对知识生成的影响，其中列(1)—(5)的被解释变量分别是知识新构、知识接续、知识弃用、知识延伸和知识拓展。列(1)和列(3)回归结果显示，AI 对知识新构和知识弃用的影响均显著为正，表明 AI 的应用促进了新知识的产生，也推动了旧知识的淘汰。同时，列(4)和列(5)的回归结果显示，AI 对知识延伸与知识拓展具有显著的正向影响，说明 AI 的应用也促进了后续创新过程中知识的延伸与跨领域拓展。但列(2)的结果显示，AI 对知识接续的影响并不显著，原因可能在于知识接续本身具有较强的路径依赖性。对于那些在研发过程中已被反复应用的知识，研发人员能够较为容易地沿既有技术路径推进创新。在这种情形下，AI 所提供的信息支持可能与研发人员自身认知之间的重合度较高，其作用相对有限，从而在统计上可能难以表现为显著的结果。

表 1 基准回归

	现有知识与后向引用知识的联系			现有知识与前向引用知识的联系	
	知识新构	知识接续	知识弃用	知识延伸	知识拓展
	(1)	(2)	(3)	(4)	(5)
AI	0.2107***	0.0209	0.0877***	0.0220***	0.1581***
	(0.0470)	(0.0279)	(0.0327)	(0.0080)	(0.0446)
控制变量	是	是	是	是	是
年份固定效应	是	是	是	是	是
行业固定效应	是	是	是	是	是
企业固定效应	是	是	是	是	是
样本量	35 071	35 071	35 071	35 071	35 071
R^2	0.4532	0.3101	0.5110	0.3246	0.5385

注：括号内为企业层面的聚类稳健标准误。***、**、* 分别表示在 1%、5%、10%的水平上显著，下表同。

（二）稳健性检验

为准确识别 AI 对知识生成的影响，本文采用多种方法进行稳健性检验。第一，剔除了样本期间新增的 IPC。参考国家知识产权局公布的《2006.01 版 IPC 分类表》—《2022.01 版 IPC 分类表》可以发现，IPC 分类规则并非静态不变，其分类标准会根据技术演化进行定期更新与细化。这一动态调整可能导致

类似的知识内容在不同年份被归入不同的 IPC 类别，从而造成测量误差。为避免这一测量误差，在计算被解释变量时删除了因规则变化而增加的 IPC，排除因修订规则变化带来的偏误。第二，借助专利转移数据，重新计算了企业的 AI 水平并重新回归。第三，在测度核心解释变量时排除主分类号为 AI 领域的专利。主分类号为 AI 领域的专利技术，通常聚焦于 AI 底层算法的优化。这些底层算法的改进虽然在技术层面具有重要意义，但可能并未融入企业的实际研发流程。相比之下，副分类号中涉及 AI 领域的专利表明，AI 相关技术可能已被纳入某一特定领域的知识创新过程。排除主分类号为 AI 领域的专利可能会减少高估偏误。第四，由于被解释变量可能不符合正态分布假设，本文使用面板泊松模型重新回归。第五，参考姚加权等(2024)，在基准回归中剔除信息传输、软件和信息技术服务业的样本，以排除 AI 技术优势行业的影响。第六，将其他数字技术作为控制变量纳入模型。第七，在涉及企业知识与其后向知识关联的分析中，引入核心解释变量的滞后项。稳健性检验结果均显示，AI 显著促进了知识生成。

（三）内生性处理

AI 与知识创新关系的识别中不可避免地会出现内生性问题，从而影响因果推断的准确性和可靠性。例如，知识创新能力强的企业会不断积累创新经验，并反馈到 AI 模型的训练和程序优化中，从而提升 AI 技术水平。对此，本文构建 Bartik 工具变量处理内生性。具体而言，本文采用企业样本初期各技术领域的专利占比作为初始份额；并以全国层面（剔除目标企业后）各技术领域 AI 技术增长率作为外生冲击。最后，将企业各技术领域的初始份额与对应领域的 AI 专利增长率相乘，加总得到 Bartik 工具变量。该工具变量和企业的 AI 水平相关，并且由于全国层面 AI 技术增长率比较外生，因而满足工具变量的外生性假设。经过内生性处理后 AI 对知识新构、知识弃用、知识延伸和知识拓展的影响仍显著为正。①

五、知识属性、后续创新来源以及 AI 分类检验

前述检验结果表明，AI 显著影响了知识生成过程。那么在创新过程中，知识的新颖性与影响力如何变化？在知识延伸以及知识拓展过程中，AI 是推动了新知识的进一步发展，还是揭示了原有知识体系中潜在但未被充分发现的信息？预测式和生成式 AI 的效果是否存在区别？本文围绕上述问题进一步展开分析。

① 稳健性检验以及内生性检验回归结果见附录Ⅰ表 A1 和表 A2。篇幅所限，附录未在正文列示，感兴趣的读者可在《经济学》(季刊)官网(https://ceq.ccer.pku.edu.cn)下载。

（一）知识属性

本文采用 BERT 模型和 LDA-SVC 模型分析创新过程中知识新颖性和知识影响力的变化。首先，本文对专利摘要文本进行向量化，计算中心专利与其引用专利之间的文本相似度；通过取相似度均值的倒数，衡量专利所包含知识的新颖性，见式(2)：

$$Nove_{ft}=\frac{1}{q_{ft}}\sum_{V_{ft}=1}^{q_{ft}}\frac{N_{V_b}}{\sum_{j\in V_b}\rho_{V_{ft},j}}, \tag{2}$$

其中，V_{ft} 表示 f 企业 t 年授权的专利，q_{ft} 表示 f 企业 t 年授权专利的总量，V_b 表示专利 V_{ft} 引用的专利集合，j 为 V_b 中的专利，$\rho_{V_{ft},j}$ 是专利 V_{ft} 与其引用专利 j 之间的文本相似度，N_{V_b} 代表 V_b 集合内的专利数量。$Nove_{ft}$ 值越大，表明在知识创新过程中与原有知识相比，该知识加入了更多新概念、新方法、新内容或新观点，从而在知识层面上表现出更高的新颖性。

其次，利用 LDA－SVC 模型分析知识影响力的变化。具有重大影响力的研究会引发研究焦点转移，并推动后续针对同一知识主题的持续研究。若一项专利涉及未来知识主题较多，而涉及过去知识主题较少，则该专利更有可能对后续同类创新活动产生影响，知识影响力越大。本文使用 LDA 模型提取专利中出现的知识主题及各专利对应的主题分布，并在此基础上训练多类线性支持向量机(SVC)，以预测不同知识主题的专利在各年份出现的概率(Savov et al.，2020)，见式(3)：

$$SVC_{ft}=\frac{1}{q_{ft}}\sum_{V_{ft}=1}^{q_{ft}}\sum_{t=t_0}^{t_n}\left(conf(V_{ft},t)\,(t-t_r)-ePr(u_t)\right), \tag{3}$$

其中，$conf(V_{ft},t)$ 表示专利 V_{ft} 在 t 年出现的概率，t_r 表示专利 V_{ft} 实际出现的年份，t_0 代表样本期初，t_n 代表样本期末。然而，由于多类线性支持向量机对早期的预测会产生正向偏误，往往更多用于同期比较。为进行不同年份对比，在此减去预测误差 $ePr(u_n)$。SVC_{ft} 越大，表示专利 V_{ft} 在未来出现的概率越大，其知识影响力越高。在基准回归的基础上，筛选出存在知识生成的企业样本进行回归。表 2 列(1)和列(2)结果显示，AI 不仅推动了知识生成，也提高了知识的新颖性和增强其在后续创新中的影响力。

（二）后续创新来源

AI 推动了知识延伸与拓展，而后续创新的知识来源有两种途径：一是新知创造，即通过对新知识的进一步探究，推动形成新的知识体系；二是旧知新解，即揭示旧知识中隐藏的信息或潜在应用领域，尤其是旧知识在新情境中的应用。如果前向引用专利集合中更多地包含中心专利的知识而非后向引用专利

的知识，或者没有包含后向引用专利的知识，表明知识延伸及拓展偏向于对新知识的进一步探究，将其称为“新知创造”过程。如果前向引用专利集合中更多地包含后向引用专利知识而非中心专利特有的知识，或者前向引用知识集合中包含了中心专利所忽视的后向引用专利的知识，表明企业后续创新过程偏向于揭示旧知识原隐藏的信息，将其称为“旧知新解”过程（Funk and Owen-Smith，2017）。本文对上述两类知识创新来源进行了检验，结果见表 2 列（3）—列（6）。结果显示，AI 主要促进了新知识的延伸以及拓展，而对旧知识延伸的影响比较有限。这表明，借助 AI 新知识不仅能够快速涌现，还在后续的知识创新活动中成为主导力量；相对而言，旧知识可能难以成为后续知识创新的主导方向。

表 2　知识属性和后续创新来源

	知识属性		后续创新来源			
	新颖性	影响力	新知创造		旧知新解	
	（1）	（2）	（3）	（4）	（5）	（6）
AI	0.0073***	0.0256***	0.0155***	0.1535***	0.0059	0.0061*
	（0.0015）	（0.0022）	（0.0060）	（0.0435）	（0.0042）	（0.0032）
控制变量	是	是	是	是	是	是
年份固定效应	是	是	是	是	是	是
行业固定效应	是	是	是	是	是	是
企业固定效应	是	是	是	是	是	是
样本量	12 692	12 692	35 071	35 071	35 071	35 071
R^2	0.5082	0.9733	0.2681	0.5380	0.2689	0.2066

（三）分类检验：预测式与生成式 AI

本文根据技术特性将 AI 分为预测式（PAI）与生成式（GAI）两种类型，分析二者对知识生成的影响差异。其中，预测式 AI 主要执行判别式任务，目标是识别输入数据与输出结果之间条件概率分布并预测相应的结果；而生成式 AI 主要执行生成式任务，通过学习整体数据的联合概率分布，生成新的或从未见过的数据或样本（Raisch and Fomina，2024）。为有效区分这两类技术，采用融合 DeepSeek-V3.2 与 BERT 模型的 AI 专利识别与分类流程。步骤如下：首先，从 AI 专利中随机抽取 5 000 个样本，依据上述分类标准设计提示词，使用 DeepSeek-V3.2 模型分析专利的标题与摘要；借助模型判断结果，将 5 000 份专利归类为预测式 AI、生成式 AI 或二者皆非，由此构建出初始标注数据集。其次，以该标注结果作为训练集，对 BERT 模型进行微调，使其能够识别不同类型 AI 的语义特征。最后，将训练完成的 BERT 分类模型应用于全部 AI 专利样本，实现

对预测式与生成式 AI 专利的分类。表 3 的列(1)—列(5)结果为预测式与生成式 AI 对知识生成过程的影响，我们借鉴 Flannery and Rangan(2006)的方法，进行了相对重要性检验。结果显示，生成式 AI 的效果普遍高于预测式 AI，且这种优势在后续知识创新过程中表现得更为明显。这表明生成式 AI 的应用，可能更有助于研发人员突破既有认知框架，并在未来研发活动中推动相关知识的延伸与拓展。

表 3 预测式与生成式 AI 分类检验

	现有知识与后向引用知识的联系			现有知识与前向引用知识的联系	
	知识新构	知识接续	知识弃用	知识延伸	知识拓展
	(1)	(2)	(3)	(4)	(5)
GAI	0.0303**	−0.0041	0.0135*	0.0069***	0.0316***
	(0.0123)	(0.0081)	(0.0077)	(0.0024)	(0.0098)
PAI	0.0149**	0.0066	0.0092*	0.0016	0.0111
	(0.0075)	(0.0052)	(0.0055)	(0.0014)	(0.0079)
控制变量	是	是	是	是	是
年份固定效应	是	是	是	是	是
行业固定效应	是	是	是	是	是
企业固定效应	是	是	是	是	是
样本量	35 071	35 071	35 071	35 071	35 071
R^2	0.4533	0.3101	0.5110	0.3248	0.5385
相对重要性分析	2.25%	—	1.35%	3.19%	2.02%
	1.11%	—	0.92%	0.73%	0.71%

注：为了控制两类 AI 可能存在的交互影响，加入了预测式与生成式 AI 交互项作为控制变量，并对变量进行标准化处理。

六、约束条件检验：内外部知识关联结构

上文实证结果表明，生成式 AI 对知识生成的影响更显著。下面从知识多样化、核心知识一致性、知识距离以及知识网络视角，探究预测式与生成式 AI 促进知识生成的适用场景。

（一）内部知识关联结构

1. 知识多样性

在创新过程中，不同研发任务需要不同技术领域的专业知识，多元化的知识结构更有助于企业运用 AI 进行创新。但是，知识多元化也会增加企业在知

识整合、吸收和应用方面的难度。参考 Kim et al.(2022)的方法，采用熵指数测算企业的知识多样性：

$$Dive_{ft}=\sum_{t'=t-5}^{t-1}\sum_{k=1}^{n_{t'}}\frac{K_{kft'}}{K_{ft'}}\ln\left(\frac{K_{ft'}}{K_{kft'}}\right)\ ,\tag{4}$$

其中，$t'\in[t-5,t-1]$，$K_{kft'}$ 表示 f 企业在 t' 时期拥有的知识元素 k 的数量；$K_{ft'}$ 表示 f 企业在 t' 时期的知识总量，$n_{t'}$ 为企业在 t' 时期拥有的知识种类数量。$Dive_{ft}$ 越大，表明 f 企业的知识多样性水平越高。特别地，当企业所有的知识都来源于同一分类号时，$Dive_{ft}$ 为 0。分组检验的回归系数和置信区间见图 1。结果显示，预测式和生成式 AI 对知识多样性水平不同的企业知识创新存在差异化影响。其中，预测式 AI 显著促进低知识多样性企业的知识生成，而其影响在知识多样性高的企业中并不显著。与预测式 AI 不同，生成式 AI 对知识生成的影响不受企业知识多样性的制约，但对知识多样性较高企业的影响更为明显。考虑到直接比较子样本系数的大小可能存在偏差，本文进行费舍尔组合检验，以评估两组系数估计值是否存在差异。[①] 检验结果表明，生成式 AI 对知识新构、知识弃用、知识延伸以及知识拓展的影响在两组之间存在显著差异。

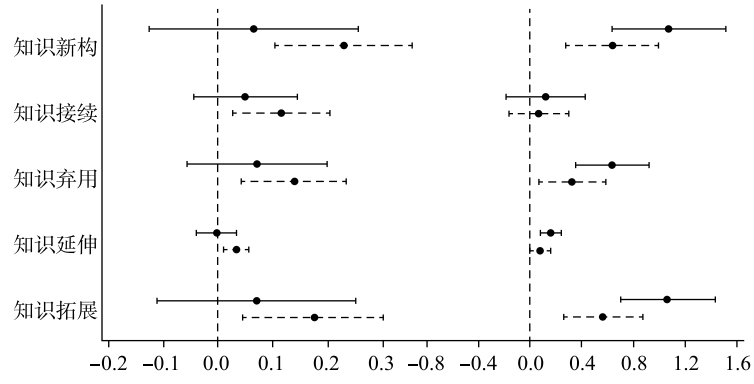

图 1　知识多样性分组检验

注：左侧为预测式 AI，右侧为生成式 AI，根据知识多样性的中位数对企业进行分组，实线为高知识多样性的分组，虚线为低知识多样性的分组。其余各图标注与之类似，实线为高组别，虚线为低组别。

2. 核心知识一致性

核心知识一致性反映了企业技术研发方向的稳定程度。较高的一致性意味着企业能够沿既有轨迹持续积累知识，但同时也可能强化对原有技术路径的依赖。参考 Dosi et al.(2022)设计核心知识一致性指标；首先，基于专利分类号的出现频率识别企业不同时期的核心知识；之后通过分析企业核心知识前五年的共现概率，衡量其核心知识的一致性程度。图 2 结果显示，在核心知识一致

① 约束性条件检验的具体系数与组间差异检验结果详见附录Ⅰ表 A3 至表 A6。

性较高的企业中，预测式 AI 促进了知识新构、知识弃用以及知识拓展过程，而该影响在核心知识一致性较低的企业中并不显著。但在核心知识一致性较低的情况下，生成式 AI 显著促进了知识新构以及知识拓展，并且组间差异检验结果表明，生成式 AI 对核心知识一致性较高企业的知识延伸促进作用更为显著。

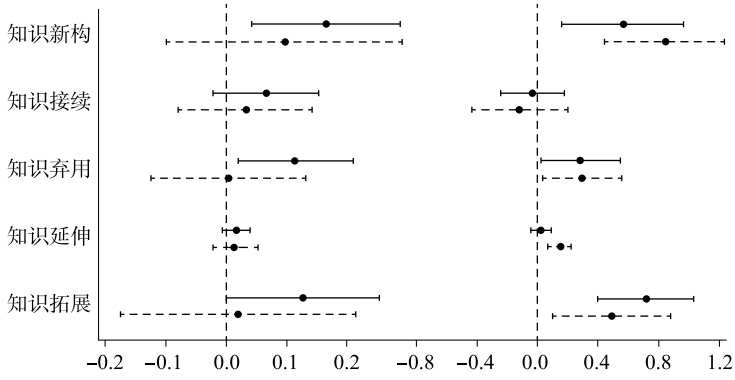

图 2 核心知识一致性分组检验

（二）外部知识关联结构

1. 知识距离

企业在追求创新的过程中，若受到内部知识的局限，则需要通过寻求协作创新等方式以弥补知识不足。在这个过程中，企业之间的知识距离会对知识创新产生重要影响。与知识距离较大的企业进行合作，可以为企业技术研发提供互补性知识和激发新思路，但也可能存在沟通障碍，导致创新效率下降。为探讨在不同知识距离的情况下 AI 对知识生成的差异化影响，本文计算企业前五年内部知识与外部知识的对称差集来衡量企业间的知识距离：

$$Dis_{ft}=\sum_{t'=t-5}^{t-1}\frac{\mathrm{card}(S_{ft'}\oplus S_{t'})}{\mathrm{card}(S_{ft'}\bigcup S_{t'})}, \tag{5}$$

其中，$S_{ft'}$ 表示 f 企业 t' 年知识的集合，而 $S_{t'}$ 表示 t' 年所有企业知识的集合。$S_{ft'}\oplus S_{t'}$ 表示企业内部与外部知识集合的差异，$S_{ft'}\bigcup S_{t'}$ 代表 t' 时期的知识总集合。$\mathrm{card}(\cdot)$ 表示集合中元素的个数。Dis_{ft} 数值越大，表明 f 企业在 t 年与其他企业的知识距离越远。回归结果见图 3，实线表示知识距离大，虚线表示距离小。结果显示，在企业内部知识与外部知识距离较近的情况下，预测式与生成式 AI 均显著促进了知识新构、知识弃用与知识拓展，而该影响在企业知识距离较远时并不显著。这表明若企业与其他企业知识距离较远时，企业往往难以利用 AI 提取有效信息，从而限制了两类技术在知识生成过程中的应用效果。

2. 知识网络中心度

在合作研发的网络结构中，企业所处的网络位置直接影响其知识获取与整

合的效率。其中，处于网络中心位置的企业，因其连接着更广泛的知识节点，能高效汇聚与整合内外部知识资源，促进自身知识体系的更新。为了探讨在不同知识网络位置的企业中 AI 的差异化影响，参考 Guan et al.(2016)设计接近中心度指标：

$$Close_{ft}=\frac{1}{n_{t-1}}\sum_{k=1}^{n_{t-1}}\frac{1}{\sum_j d(k_{fk,t-1},k_{j,t-1})}, \tag{6}$$

其中，$d(k_{fk,t-1},k_{j,t-1})$ 表示 f 企业在 $t-1$ 时期的知识 $k_{fk,t-1}$ 和 $t-1$ 时期网络中其他知识 $k_{j,t-1}$ 之间建立联系的最短路径长度，n_{t-1} 为在 $t-1$ 时期企业的知识种类数量。该指标反映了企业与其他企业之间知识流动路径的连通性，中心度越高，表明企业在网络中越处于知识交汇的关键节点。图 4 中实线表示网络中心度较高，虚线表示网络中心度较低。结果表明，AI 显著推动了知识中心度较高企业的知识新构、知识弃用、知识延伸和知识拓展。而对于处于边缘位置的企业来说，AI 的正向影响并不明显。

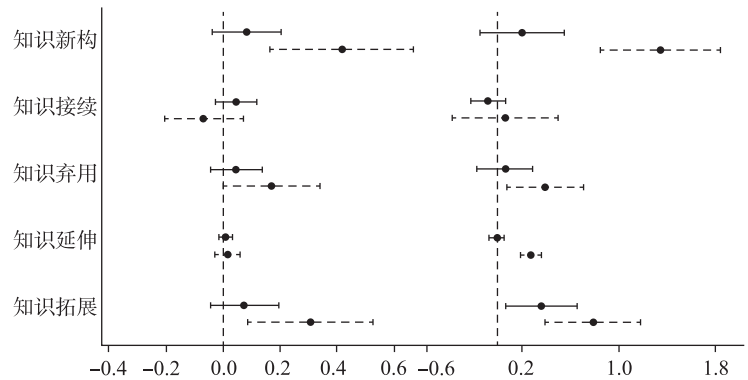

图 3　知识距离分组检验

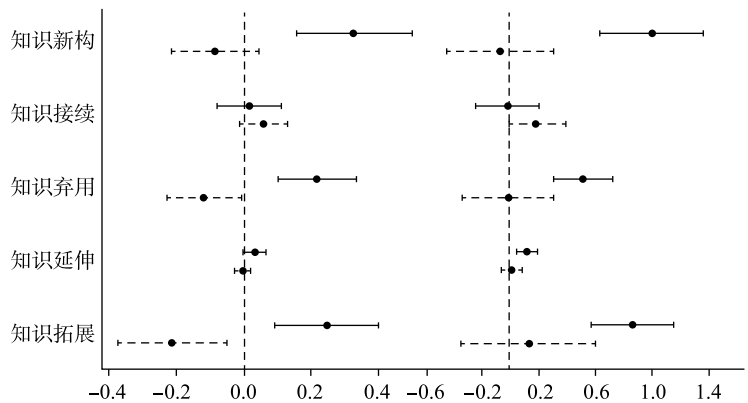

图 4　知识网络中心度分组检验

七、进一步分析

知识生成仅是企业创新过程的起点，知识价值的真正实现还依赖于企业能否将新生成的知识嵌入实际的应用场景之中，实现从知识生产向知识应用的转化。基于此，下面进一步关注知识的价值实现过程，探讨 AI 在促进知识整合与泛化、缩短知识应用周期以及激活“沉睡知识”等方面的可能性影响。

（一）知识整合和知识泛化

知识整合是指将不同来源、不同领域的知识元素进行汇集和融合的过程。而知识泛化是指从特定的知识或经验中抽象出普遍适用的规律或原则，使其能够应用于更广泛场景的过程。本文参考 Capello and Lenzi(2013)设计知识整合和泛化指标：

$$Integ_{ft}=1-\sum_{V_{ft}=1}^{q_{ft}}\sum_{k\in V_b}\left(\frac{K_{k,V_b}}{K_{V_b}}\right)^2, \tag{7}$$

$$Gener_{ft}=1-\sum_{V_{ft}=1}^{q_{ft}}\sum_{k\in V_d}\left(\frac{K_{k,V_d}}{K_{V_d}}\right)^2, \tag{8}$$

其中，q_{ft} 表示 f 企业 t 年专利的总量；K_{k,V_b} 表示中心专利 V_{ft} 对应的后向引用专利 V_b 中包含的知识 k 的数量，K_{V_b} 表示 V_b 包含的知识元素总量；K_{k,V_d} 表示 V_{ft} 对应的前向引用专利 V_d 中包含的知识 k 的数量，K_{V_d} 表示 V_d 包含的知识元素总量。$Integ_{ft}$ 越高，表示知识来源分布越均衡，即该技术的知识来源于多个不同的知识领域，没有依赖某一特定的知识领域，体现了该专利对多领域的知识的整合。$Gener_{ft}$ 越高，表示知识在各个领域均有应用，体现了其在多领域的泛化效果。表 4 列(1)和列(2)结果显示，AI 显著促进了知识整合和泛化。

（二）知识应用周期

前文证实，AI 显著提升了企业的知识整合和泛化能力，那么 AI 是否可以缩短知识应用周期？本文参考 Bianchini et al.(2022)，测算专利被引量第一个峰值与第二个峰值的时间差距来表征知识应用周期。[①] 由于专利可能有多个引用高峰，本文忽略间隔 2 年以下的时间差距。此外，由于市场需求的变化，知识可能不会出现第二个被引高峰，故本文也对专利被引量是否存在第二高峰进行检验。第二高峰和知识应用周期的检验都排除了专利自引。表 4 列(3)结果显示，AI 有助于形成第二个被引高峰，即促进了知识的产业化应用。列(4)的结果

① 附录Ⅰ图 A1 展示了这种周期特征。

表 4　知识整合泛化、知识应用周期以及唤醒“沉睡知识”检验

	知识整合	知识泛化	是否存在第二个峰值	知识应用周期	“沉睡知识”是否唤醒		“沉睡知识”唤醒时值		“沉睡知识”重用价值	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
AI	0.0218***	0.0331***	0.7967***	−0.1212**	1.3669***	1.1571***	−0.3472**	−0.3543***	0.0611**	0.0392**
	(0.0079)	(0.0087)	(0.2510)	(0.0521)	(0.4589)	(0.4110)	(0.1323)	(0.1109)	(0.0237)	(0.0169)
控制变量	是	是	是	是	是	是	是	是	是	是
年份固定效应	是	是	是	是	是	是	是	是	是	是
行业固定效应	是	是	是	是	是	是	是	是	是	是
企业固定效应	是	是	否	是	否	否	是	是	是	是
样本量	35 071	35 071	30 299	11 378	30 686	28 093	3 207	2 408	3 207	2 408
R^2/Pseudo R^2	0.5446	0.6038	0.5521	0.4670	0.2021	0.1961	0.6220	0.6474	0.4973	0.5002

注：表中被解释变量为虚拟变量时，采用面板 Logit 模型估计。在 Logit 估计过程中控制行业和年份固定效应。回归过程中采用行业层面的聚类稳健标准误。表 4 列(5)、列(7)和列(9)对应间隔 3 年以上的“沉睡知识”，列(6)、列(8)和列(10)对应间隔 5 年以上的“沉睡知识”。

表明，AI 缩短了知识的应用周期。这表明企业应用 AI，有助于缩短新知识发现到商业化落地的时间，从而提高了创新转化效率。

（三）唤醒“沉睡知识”

表 1 和表 2 结果显示，AI 驱动并未显著促进知识接续，但会对旧知新解产生一定程度的影响。那么，AI 是否会激活以往未被企业充分利用的知识，即唤醒具有价值的“沉睡知识”？参考 Kok et al.(2019)的方法，在企业层面构建了“沉睡知识”重用的指标。具体步骤如下：首先，统计企业各类专利分类号出现的历史年份，并筛选出现间隔超过 3 年或 5 年以上分类号对应的专利，将其界定为“沉睡知识”；然后，计算该类知识在企业不同年份专利中出现的时间差距，将其均值作为“沉睡知识”的唤醒时间；最后，计算该分类号在企业未来专利授权中被重复使用的概率表征“沉睡知识”重用价值。表 4 列(5)和列(6)结果显示，AI 可以唤醒企业长久未用的“沉睡知识”；列(7)和列(8)表明，AI 缩短了“沉睡知识”的唤醒时间；列(9)和列(10)表明，AI 可能增加了“沉睡知识”唤醒后的重用价值。整体而言，AI 可以识别潜在可利用的知识，缩短高价值知识的重用时间，激活未被企业充分利用的“沉睡知识”，这体现了 AI 在知识重用过程中的重要作用。

八、政 策 含 义

本文研究表明，AI 正在重塑企业的知识生成过程，推动了创新过程中的知识新构、知识弃用、知识延伸与知识拓展；也会增加企业知识整合和泛化能力，缩短知识应用周期，并有助于唤醒有价值的“沉睡知识”。本研究为企业创新战略规划和政府政策设计提供了重要的决策参考：第一，企业应重视 AI 对知识创新的驱动作用，推动其在实验设计、技术研发和模拟测试等关键环节的深化应用，以智能化手段提高创新效率，加速知识生产和技术迭代。第二，建议政府加大算力基础设施建设，通过搭建公共算力资源平台和跨区域算力调度平台，保障算力需求的快速响应，并通过补贴等方式激励企业应用与开发大模型。第三，鼓励企业开发可解释性强的算法，通过模块化设计和标准化接口，使其能够适配不同应用场景，推动 AI 技术的产业化发展和应用场景落地。

参 考 文 献

[1] Alon-Barkat, S., and M. Busuioc, “Human-AI Interactions in Public Sector Decision Making: ‘Automation Bias’ and ‘Selective Adherence’ to Algorithmic Advice”, *Journal of Public Administra-*

tion Research and Theory, 2023, 33(1), 153-169.

[2] Babina, T., A. Fedyk, A. He, and J. Hodson, "Artificial Intelligence, Firm Growth, and Product Innovation", *Journal of Financial Economics*, 2024, 151, 103745.

[3] Becker, M. C., F. Rullani, and F. Zirpoli, "The Role of Digital Artefacts in Early Stages of Distributed Innovation Processes", *Research Policy*, 2021, 50(10), 104349.

[4] Bianchini, S., M. Müller, and P. Pelletier, "Artificial Intelligence in Science: An Emerging General Method of Invention", *Research Policy*, 2022, 51(10), 104604.

[5] Brynjolfsson, E., D. Li, and L. Raymond, "Generative AI at Work", *The Quarterly Journal of Economics*, 2025, 140(2), 889-942.

[6] Capello, R., and C. Lenzi, "Territorial Patterns of Innovation: A Taxonomy of Innovative Regions in Europe", *The Annals of Regional Science*, 2013, 51(1), 119-154.

[7] Chen, J., M. Wu, Q. Liu, and Y. Zhang, "Explainable Prediction of Knowledge Recombination: A Synergized Method with Heterogeneous Hypergraph Learning and Large Language Models", *Information Processing & Management*, 2026, 63(1), 104336.

[8] Choudhury, P., E. Starr, and R. Agarwal, "Machine Learning and Human Capital Complementarities: Experimental Evidence on Bias Mitigation", *Strategic Management Journal*, 2020, 41(8), 1381-1411.

[9] Cooper R. G., "The AI Transformation of Product Innovation", *Industrial Marketing Management*, 2024, 119, 62-74.

[10] Csaszar, F. A., and D. A. Levinthal, "Mental Representation and the Discovery of New Strategies", *Strategic Management Journal*, 2016, 37(10), 2031-2049.

[11] Dedehayir, O., and M. Steinert, "The Hype Cycle Model: A Review and Future Directions", *Technological Forecasting and Social Change*, 2016, 108, 28-41.

[12] Deepak, S., and P. M. Ameer, "Brain Tumor Classification Using Deep CNN Features via Transfer Learning", *Computers in Biology and Medicine*, 2019, 111, 103345.

[13] Dong, J. Q., and C.-H. Yang, "Being Central Is a Double-Edged Sword: Knowledge Network Centrality and New Product Development in U.S. Pharmaceutical Industry", *Technological Forecasting and Social Change*, 2016, 113, 379-385.

[14] Dosi, G., N. Mathew, and E. Pugliese, "What a Firm Produces Matters: Processes of Diversification, Coherence and Performances of Indian Manufacturing Firms", *Research Policy*, 2022, 51(8), 104152.

[15] Flannery, M. J., and K. P. Rangan, "Partial Adjustment toward Target Capital Structures", *Journal of Financial Economics*, 2006, 79(3), 469-506.

[16] Funk, R. J., and J. Owen-Smith, "A Dynamic Network Measure of Technological Change", *Management Science*, 2017, 63(3), 791-817.

[17] Guan, J., K. Zuo, K. Chen, and R. C. M. Yam, "Does Country-Level R&D Efficiency Benefit from the Collaboration Network Structure?", *Research Policy*, 2016, 45(4), 770-784.

[18] Huenteler, J., J. Ossenbrink, T. S. Schmidt, and V. H. Hoffmann, "How a Product's Design Hierarchy Shapes the Evolution of Technological Knowledge—Evidence from Patent-Citation Networks in Wind Power", *Research Policy*, 2016, 45(6), 1195-1217.

[19] Iwana, B. K., and S. Uchida "An Empirical Survey of Data Augmentation for Time Series Classifi-

cation with Neural Networks", *PLOS ONE*, 2021, 16(7), e0254841.

[20] Khosravi B., A. D. Weston, F. Nugen, J. P. Mickley, H. M. Kremers, C. C. Wyles, R. E. Carter, and M. J. Taunton, "Demystifying Statistics and Machine Learning in Analysis of Structured Tabular Data", *The Journal of Arthroplasty*, 2023, 38(10), 1943-1947.

[21] Kim, J., T. Kollmann, A. Palangkaraya, and E. Webster, "Does Local Technological Specialisation, Diversity and Dynamic Competition Enhance Firm Creation?", *Research Policy*, 2022, 51(7), 104557.

[22] Kok, H., D. Faems, and P. De Faria, "Dusting Off the Knowledge Shelves: Recombinant Lag and the Technological Value of Inventions", *Journal of Management*, 2019, 45(7), 2807-2836.

[23] Lee, K. Y., H. J. Jung, and Y. Kwon, "Boundary-Spanning Technology Search, Product Component Reuse, and New Product Innovation: Evidence from the Smartphone Industry", *Research Policy*, 2024, 53(4), 104959.

[24] 李玉花、林雨昕、李丹丹，"AI 技术应用如何影响企业创新"，《中国工业经济》，2024 年第 10 期，第 155—173 页。

[25] Lipkova, J., R. J. Chen, B. Chen, M. Y. Lu, M. Barbieri, D. Shao, A. J. Vaidya, C. Chen, L. Zhuang, D. F. K. Williamson, M. Shaban, T. Y. Chen, and F. Mahmood, "Artificial Intelligence for Multimodal Data Integration in Oncology", *Cancer Cell*, 2022, 40(10), 1095-1110.

[26] Lou, B., and L. Wu, "AI on Drugs: Can Artificial Intelligence Accelerate Drug Development? Evidence from a Large-Scale Examination of Bio-Pharma Firms", *MIS Quarterly*, 2021, 45(3), 1451-1482.

[27] Luo, Y., Y. Wen, T. Liu, and D. Tao, "Transferring Knowledge Fragments for Learning Distance Metric from a Heterogeneous Domain", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, 41(4), 1013-1026.

[28] Messeri, L., and M. J. Crockett, "Artificial Intelligence and Illusions of Understanding in Scientific Research", *Nature*, 2024, 627(8002), 49-58.

[29] Privett, H. K., G. Kiss, T. M. Lee, R. Blomberg, R. A. Chica, L. M. Thomas, D. Hilvert, K. N. Houk, and S. L. Mayo, "Iterative Approach to Computational Enzyme Design", *Proceedings of the National Academy of Sciences*, 2012, 109(10), 3790—3795.

[30] Raisch, S., and K. Fomina, "Combining Human and Artificial Intelligence: Hybrid Problem-Solving in Organizations", *Academy of Management Review*, 2024, amr.2021.0421.

[31] Rammer, C., G. P. Fernández, and D. Czarnitzki, "Artificial Intelligence and Industrial Innovation: Evidence from German Firm-Level Data", *Research Policy*, 2022, 51(7), 104555.

[32] Savov, P., A. Jatowt, and R. Nielek, "Identifying Breakthrough Scientific Papers", *Information Processing & Management*, 2020, 57(2), 102168.

[33] Seidel, V. P., and S. O'Mahony, "Managing the Repertoire: Stories, Metaphors, Prototypes, and Concept Coherence in Product Innovation", *Organization Science*, 2014, 25(3), 691-712.

[34] Sendek, A. D., E. D. Cubuk, E. R. Antoniuk, G. Cheon, Y. Cui, and E. J. Reed, "Machine Learning-Assisted Discovery of Solid Li-Ion Conducting Materials", *Chemistry of Materials*, 2019, 31(2), 342-352.

[35] Siriwardhana, S., R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, and S. Nanayakkara, "Im-

proving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering", *Transactions of the Association for Computational Linguistics*, 2023, 11, 1-17.

[36] Sourati, J., and J. A. Evans, "Accelerating Science with Human-Aware Artificial Intelligence", *Nature Human Behaviour*, 2023, 7(10), 1682-1696.

[37] Tabor, D. P., L. M. Roch, S. K. Saikin, C. Kreisbeck, D. Sheberla, J. H. Montoya, S. Dwaraknath, M. Aykol, C. Ortiz, H. Tribukait, C. Amador-Bedolla, C. J. Brabec, B. Maruyama, K. A. Persson, and A. Aspuru-Guzik, "Accelerating the Discovery of Materials for Clean Energy in the Era of Smart Automation", *Nature Reviews Materials*, 2018, 3(5), 5-20.

[38] 王林辉、姜昊、董直庆,"工业智能化会重塑企业地理格局吗",《中国工业经济》,2022 年第 2 期,第 137—155 页。

[39] Wang, X., and L. Wu, "Artificial Intelligence, Lean Startup Method, and Product Innovations", *Management Science*, 2026, 72(1), 756-782.

[40] Weinan, E., "The Dawning of a New Era in Applied Mathematics", *Notices of the American Mathematical Society*, 2021, 68(4), 1.

[41] 姚加权、张锟澎、郭李鹏、冯绪,"AI 如何提升企业生产效率?——基于劳动力技能结构调整的视角",《管理世界》,2024 年第 2 期,第 101—116 页。

[42] Yoo, Y., O. Henfridsson, and K. Lyytinen, "Research Commentary—The New Organizing Logic of Digital Innovation: An Agenda for Information Systems Research", *Information Systems Research*, 2010, 21(4), 724-735.

The Puzzle of Knowledge Generation: Evidence from AI Patent Texts and Citation Data

DONG Zhiqing

(Zhejiang University of Finance and Economics)

ZHANG Xiao

(East China Normal University)

YU Miaojie*

(Liaoning University)

Abstract: Knowledge generated by AI is prone to model hallucinations, and the role of AI in knowledge generation remains unclear. Using patent texts and citation data, we examine

* Corresponding Author: YU Miaojie, Liaoning University, No. 66 Chongshan Middle Road, Shenyang, Liaoning 110036, China; Tel: 86-13810003055; E-mail: mjyu@nsd.pku.edu.cn.

AI's impact on knowledge generation, comparing predictive and generative models, and analyze their roles in knowledge application and in awakening knowledge. Results show that AI promotes new knowledge creation by recombining and abandoning existing knowledge, and fosters further innovation through extending and expanding knowledge, enhancing the novelty and impact of knowledge. However, the effects vary with firms' knowledge structures. AI supports knowledge integration and generalization, accelerates knowledge application, and reactivates dormant knowledge.

Keywords: predictive AI; generative AI; knowledge generation

JEL Classification: L86, O32, O33