生成式人工智能十大趋势与公共文化机构的应对策略*

刘　炜　刘倩倩

[摘　要] DeepSeek 的问世标志着人工智能（AI）发展进入全新阶段，即从以大模型训练为主向以推理应用为主的过渡，预示着 AI 应用即将迎来爆发式增长。对于美术馆、图书馆、档案馆和博物馆等公共文化机构而言，大模型正逐步成为核心驱动力，推动内容生产、知识重组及服务智能化进入新阶段。生成式人工智能近期发展的十个关键趋势为 AI 驱动的科学研究得到普及，具身智能机器人提升服务体验，多模态大模型走向实用化，合成数据与数据治理挑战凸显，世界模型与因果推理能力突破，AI 算力与模型优化协同发展，智能体技术普及带来产品爆发，资本投入与产业整合加速，开源生态与小模型应用扩展，AI 伦理与治理框架完善。通过深入分析这些趋势，公共文化机构可以更好地把握 AI 带来的机遇，提升服务效率与质量，实现智能化转型。

[关键词] 生成式人工智能　公共文化机构　大模型　智能体　多模态
[中图分类号] G250.7　[文献标志码] A
[DOI] 10.19764/j.cnki.tsgjs.20250294
[本文引用格式] 刘炜，刘倩倩．生成式人工智能十大趋势与公共文化机构的应对策略 [J]. 图书馆建设，2025(1):4–14.

Top 10 Trends in Generative Artificial Intelligence and Response Strategies for Public Cultural Institutions

Liu Wei, Liu Qianqian

[Abstract] The emergence of DeepSeek marks a new stage in the development of Artificial Intelligence (AI), transitioning from a focus on training large models to a focus on inference applications, indicating that AI applications are about to experience explosive growth. For public cultural institutions such as galleries, libraries, archives, and museums, large models are gradually becoming the core driving force, promoting content production, knowledge recombination, and service intelligence into a new stage. The ten key trends in the recent development of Generative Artificial Intelligence are the popularization of AI for science, the improvement of service experience by embodied intelligent robots, the practical application of Multimodal Large Models, the prominent challenges of synthetic data and data governance, breakthroughs in World Models and causal reasoning capabilities, the coordinated development of AI computing power and model optimization, the acceleration of capital investment and industry integration brought about by the popularization of Agent technology, the expansion of open source ecology and small model applications, and the improvement of AI ethics and governance frameworks. By analyzing these trends in depth, public cultural institutions can better grasp the opportunities brought by AI, improve service efficiency and quality, and achieve intelligent transformation.

[Keywords] Generative Artificial Intelligence; Public cultural institution; Large model; Agent; Multimodal

* 本文系国家社会科学基金重大项目“智能时代提升全民数字素养的理论与实践研究”的成果之一，项目编号：24&ZD180。

2025 年伊始，DeepSeek 的突破性进展 [1] 为生成式人工智能（Generative Artificial Intelligence，简称 GAI）的发展定下了基调：扩展定律（Scaling Law）依然有效，但是换了一种方式，模型训练的重点从预训练阶段转向后训练阶段，推动力由单纯的算力投入转变为模型微调、数据蒸馏和系统对齐的协同进化；技术普及化趋势正在挑战少数大型科技公司的主导地位，尽管大型科技企业仍在探索技术极限，但随着算力壁垒的降低和开源模型的兴起，中小型企业得以在云端、边缘计算及具身智能领域蓬勃发展，生态系统迅速扩展，形成“中心化基础设施 + 分布式应用生态”的新格局，核心竞争力正从算力垄断逐渐转向算法创新和数据质量；企业间的竞争进一步升级为地缘政治的博弈，主要体现在美中之间的多维度对抗，包括技术、资本和人才等方面。这些进展似乎共同指向一个结论：通用人工智能（Artificial General Intelligence，简称 AGI）的成败将在 2025 年见分晓。

GAI 的迅猛发展正在重塑各行各业，公共文化服务领域亦不例外。作为知识传播和文化传承的重要载体，图书馆等公共文化机构正面临新一轮 AI（Artificial Intelligence，人工智能）技术发展带来的机遇与挑战。本文从技术创新、模型训练、硬件升级、资本投入、产品应用等多个角度对 GAI 发展的十大趋势进行了分析和总结，重点探讨了这些趋势对图书馆等公共文化机构的影响。开源大模型生态的繁荣和垂直领域大模型的应用扩展，将使得公共文化机构能够迅速受益于技术进步，从而带来整个行业的变革与重塑。

1　AI 驱动的科学研究得到普及

大模型与深度学习的迅猛发展为科学研究带来了革命性的变化，催生了“人工智能助力科学研究”（AI for Science，简称 AI4S）的全新模式 [2]。2024 年，大型语言模型在语义理解、逻辑推理及跨语言生成等领域取得了重大进展，这主要得益于上下文窗口的扩展、知识密度的增强以及混合专家架构的应用。OpenAI 推出的 o3 推理模型在数学、编程及科学问答方面的展现超越了部分人类专家 [3]，而谷歌的 Gemini 2.0 则致力于解决复杂的科学问题 [4]。2025 年初，DeepSeek R1 推理模型的问世标志着又一重要里程碑。该模型在性能上直逼 OpenAI 的同时，显著降低了模型训练和推理成本，降幅达数十倍。这一突破使得多步推理等高资源消耗的任务得以普及，普通学者也能负担得起。

这一模式正逐步成为继实验科学、理论科学、计算科学和数据密集型科学之后的第五种范式 [5]。通过 AI 技术的应用，科学研究实现了从理论假设、数据收集与处理、实验路径设计到实验数据分析以及复杂问题建模的全流程革新。例如，DeepMind 开发的 AlphaFold 模型在蛋白质结构预测方面的突破性进展 [6]，充分展示了机器学习和 AI 技术在加速科学发现方面的巨大潜力。未来，AI 将深度参与科学研究的多维数据融合、实验设计及理论验证等关键环节，推动科学研究流程向自动化和智能化方向发展。AI 系统与科研人员之间的关系将超越传统的辅助角色，演变为相互启发、共同进步的协同伙伴关系。这种新型的合作模式不仅提升了科研效率，也为解决复杂科学问题开辟了新的路径。

在数字人文研究领域，AI 正逐步革新其研究方法 [7]。AI4DH（AI for Digital Humanities，人工智能驱动的数字人文）通过运用 AI 技术，对大量文本、图像及音频数据进行分析，从而揭示历史、文化及社会的深层模式和内在联系 [8]。许多研究者已开始探索 GAI 在元数据创建、资源描述及知识加工流程中的应用。AI 技术在古籍数字化修复方面也取得了显著进展，如合合信息旗下的扫描全能王与华南理工大学团队合作开发的 AI 古籍修复模型，已成功应用于敦煌遗书的数字化修复工作 [9]。

GAI 在模拟历史场景方面同样展现出巨大潜力 [10]。加州大学圣克鲁斯分校的历史学教授 Benjamin Breen 在其历史课程教学中，利用 GPT-4 和 DELL-3 等 AI 模型，模拟特定历史时期和地点的场景，以此鼓励学生进行深入的历史研究 [11]。这一创新应用不仅丰富了教学手段，也为历史研究提供了新的视角和方法。

2023 年 11 月 20 日，IFLA 人工智能特别兴趣小组发布了题为《图书馆对人工智能的战略响应》[12] 的报告，该报告详细阐述了 AI 技术在图书馆领域的应用前景，涵盖了馆藏资源的规模化描述、AI 技术对元数据的增强或创建、智能用户咨询服务、文献发现服务以及后端业务系统的 AI 优化等多个方面。

随着自博弈训练和慢思考推理等技术的进步，大型语言模型正变得日益强大和高效。2025 年初，国产推理模型（如 DeepSeek R1 和 Kimi 1.5 等）的崛起，进一步扩大了大模型的应用范围。这将有效降低 AI4S 和 AI4DH 的成本，使更多研究机构和学者能够利用这些技术，从而推动 AI4S 和 AI4DH 的普及，为科学发现和人文研究开辟新的途径。

图书馆、博物馆等公共文化机构可借助这一趋势，进一步优化传统文献及实物馆藏的管理方式，完善和构建智能数字档案库，尝试统一资源平台和知识库的建设。通过提供各类智能化工具，这些机构能够实现自动化数据挖掘，提升文献、馆藏和艺术品的加工、检索与策展效率。这将有力推动文化遗产数字化、历史文献分析和艺术创作辅助等领域的研究，为科研、教学和公众文化传播提供更为精准的支持。

2 具身智能机器人提升服务体验

2024 年被誉为人形机器人技术的“应用元年”，这一年见证了多家企业在具身智能领域的重大进展。尤其是中国的杭州宇树科技有限公司，以高性能四足机器人与低成本人形机器人占据了全球近 70% 的市场份额，深度融合大语言模型（Large Language Model，简称 LLM）与具身智能算法，通过自研运动控制架构与 AI 多模态交互系统（如集成 OpenAI 的 ChatGPT 接口），实现了从消费级陪伴到工业级应用的全面覆盖 [13]。特斯拉于 2024 年 3 月发布了 Optimus 2.0 的最新研发成果。与前代产品相比，Optimus 2.0 在动作灵敏度、目标识别和自主决策能力方面均有显著提升。该机型能够在复杂的生产环境中实现自主移动，并进行零部件的抓取与检测，错误率较之前降低了 40%[14]。波士顿动力公司在 2024 年 8 月举行的国际机器人与自动化大会（IEEE/RSJ International Conference on Intelligent Robots and Systems，简称 IROS）上展示了新一代 Atlas-Adv 机器人。该机型融合了深度学习和强化学习技术，能够根据环境变化实时进行路径规划，并对人类的手势指令作出精准响应 [15]。中国企业优必选（UBTECH）等也在 2024 年发布了具备更强自主导航和任务规划能力的人形机器人。这些机器人主要面向商用客服和养老陪伴等应用场景 [16]，展现了人形机器人在服务领域的巨大潜力。

日常生活中，各种具身智能机器人也开始广泛应用。例如，Agility Robotics 公司研发的 Digit 机器人已在亚马逊仓库中开展测试，用于搬运空载货框 [17]。此外，辅助攀登泰山的机器狗以及智慧交通系统中的巡检机器人也逐渐得到普及。

具身智能的研究得到了学术机构的更多重视 [18]。卡耐基梅隆大学的 Robotics Institute 为移动机器人创建一个强化学习系统，并设计一个框架，让机器人能够像人类一样通过独立的试错来学习任务 [19]；北京大学 HMI Lab 提出了一个名为 RoboMamba 的高效视觉—语言—动作（Vision-Language-Action，简称 VLA）模型，旨在为机器人提供推理和操作能力 [20]。

2025 年，特斯拉的 Optimus 和国内智元机器人等产品的批量生产，标志着具身智能技术的重大突破。与此同时，在机器人智能“大脑”的构建方面，DeepSeek R1 模型的推出进一步证实了 AI 技术在提升机器人的性能和降低成本方面将高速迭代，2025 年将因此成为“具身智能元年”，人工智能不再局限于数字领域，而是与物理实体深度融合。智能机器人、机器车、自动导览、仓储管理系统、实体交互设备等将日益普及，推动传统自动化向智能代理、具身机器人及自主系统的转型。

图书馆领域的专业人士对上述机器人应用场景并不陌生。随着技术进步，机器人的智能水平、交互体验及操作流畅度均将得到显著提升。当前，图书馆已采用基于 RFID（Radio Frequency Identification，射频识别技术）和传感器技术的无感借还系统，为具身智能的物理交互奠定了坚实基础。这些更为先进的机器人将具备更强的环境感知能力，并拥有自主决策和行动能力，实质上成为硬件智能体。它们可广泛应用于场馆导览、文物盘点、资料整理与传递、环境监控以及互动展示等多个方面，从而大幅提升图书馆的工作效率和服务质量，解放馆员生产力，成为未来智慧图书馆的重要组成部分。

展望未来，具身智能机器人将承担图书馆内的重复性工作，为用户提供更加个性化和智能化的服务体验。与此同时，实体服务的自动化将有效提高资源利用效率，提升管理效率并降低运营成本，从而推动图书馆服务的全面升级。这一趋势标志着图书馆行业正迈向一个更加智能化的智慧图书馆 2.0 时代。

3　多模态大模型走向实用化

随着 2024 年多模态技术的持续爆发，视频生成与理解模型迎来了“GPT 时刻”。在多模态领域，有几个备受瞩目的视频模型：Sora 是 OpenAI 于 2024 年 2 月推出的一款革命性的 AI 视频生成模型，能够根据用户的文本提示生成高质量视频；Kling 则是由快手科技开发的 AI 视频生成工具，于 2024 年 6 月推出，主打视频生成与创作；MiniMax 团队的海螺 AI 通过直观的控制和精确度提供一流的视频生成。随着模型能力和技术的进步，更多统一多模态大模型（Unified Multimodal Large Models）出现并在跨模态推理与交互能力上实现质的飞跃，如 DeepSeek 推出的 Janus-Pro[21]，原生融合文本、图像、音频、视频等多模态数据，不仅能“文生图”，还能对图片进行描述、识别地标景点、识别图像中的文字，并能对图片中的知识进行介绍。这些模型的崛起预示着 AI 在视觉、听觉、文本的综合处理能力上进一步提升，多模态融合作为新的增长点正成为行业焦点，也是 2025 年大模型技术发展的核心方向之一。

随着统一多模态大模型的发展，AI 向着更接近人类的多模态感知与表达能力迈进。未来，图书馆服务平台可借助全方位的 AI 进行升级，帮助图书馆等公共文化服务机构对馆藏资源进行更好地揭示和管理，实现跨媒体数字化展示。

多模态大模型具备同时处理文字、图像、视频、音频等多种数据输入，实现端到端的信息理解和生成。这种“下一个 Token 预测”技术将大幅提升模型效率和应用场景的广度。例如，通过统一模型实现文本、图像、音频等数据的综合检索，提高图书馆资源检索和内容推荐的精准度；通过自动生成的文字说明、图片解析和视频介绍，助力图书馆打造互动虚拟展览和智慧文创平台，提升观众沉浸体验；通过多模态模型提升图书馆数字资源管理效率，让图书馆在数字化时代更具竞争力和吸引力。

4 合成数据与数据治理挑战凸显

大模型的参数量与训练数据的 Token 数量密切相关。随着大模型公司在这一领域的竞争日趋激烈，模型的参数量不断增加，几乎已耗尽可获取的所有数据资源。在此背景下，行业开始转向使用人工合成数据作为补充。合成数据是通过算法模拟真实数据特征而生成的人工数据。与收集和标注真实数据相比，生成合成数据的成本通常较低，在某些领域能够有效弥补真实数据量的不足。此外，合成数据不包含真实个体的敏感信息，有助于规避数据隐私问题。2024 年下半年，多个先进模型如 Llama3.1[22]、OpenAI o1[23] 和 DeepSeek V3[24] 等，在其发布的论文或技术报告中均表明已采用合成数据。

在当前数据获取过程中，真实数据面临着隐私保护、版权限制以及高昂成本等挑战，合成数据技术正逐步成为推动大模型迭代与落地的重要驱动力。在中国，大模型技术的持续发展及其在本土化应用场景中的深入探索，合成数据技术展现出巨大的应用潜力。通过生成高质量、标注详尽的中文合成数据，不仅可以有效降低对真实数据的依赖，还能有效解决数据隐私和版权问题，从而加速模型的迭代和应用范围的扩展。

然而，合成数据的应用也面临着数据质量、安全性及合规性等多重挑战。为充分发挥合成数据的优势，必须强化数据治理体系的建设。

对于图书馆等公共文化机构而言，数据治理尤为重要，它直接关系到文化资源的保护与公众信任的维护，是文化机构的核心能力之一。在数字化转型过程中，图书馆可利用合成数据技术丰富馆藏数据库，自动填补信息空白。例如，通过 GAI 对古籍、地方特色文化等历史文化资源进行深度开发和利用，根据需求生成并整合文本、图像、音频、视频、3D 模型等多种类型的数据，提升资源的表现力和吸引力。同时，合成数据还可用于训练和优化检索算法，提高数据挖掘和个性化推荐的效果。

此外，文化机构需借助数据治理技术确保数字资源的质量和合规性。通过数据标注、数据清洗、数据验证等手段，提高合成数据的准确性和可靠性。在利用合成数据时，应高度重视数据安全、隐私保护和版权管理，确保数据的安全和合法使用，提供可信的资源服务。图书馆应建立健全安全机制和风险评估体系，确保合成数据在生成、存储、传输和使用过程中的安全性，保障文化信息资源在开放共享的同时得到有效保护，从而提升公众对数字文化服务的信任度。

5 世界模型与因果推理能力突破

2024 年，多模态大模型技术的进步以及 Sora 的推出，使得世界模型（World Models）成为人工智能领域的焦点议题。世界模型被广泛认为是迈向 AGI 的重要途径，其展现出的捕捉世界知识的新能力备受瞩目。然而，尽管 Sora 生成的视频在表面上似乎完美遵循物理定律，但其是否真正符合全面世界模型的标准仍存在争议。Meta 公司科学家杨立昆（Yann Lecun）对此提出批评，认为“仅通过生成像素来建模世界可能存在局限性”[25]。

世界模型是指机器对世界运作方式的理解和内部表示，其目标是深入理解世界动态并预测未来场景 [26]。世界模型的核心特征包括以下三个方面 [26]：

（1）物理世界建模：通过多模态数据（视频、音频、传感器数据等）构建对现实世界的内部表征，理解物体运动、力作用等物理规律。

（2）因果推理能力：超越传统大模型的统计关联，实现反事实推理（Counterfactual Reasoning），即回

答“如果发生某种变化，结果会如何？”的问题。

（3）动态场景生成：可生成逼真的虚拟环境，用于训练机器人、自动驾驶系统等需要物理交互的场景。

用概率统计的相关性无法精确表达物理定律的因果性，而因果推理是世界模型的核心能力，若大模型能从复杂的多模态数据中提炼出真正的因果关系，并结合因果推理能力，将可能真正实现超越人类的智能。目前，有几家公司在紧锣密鼓推进研发世界模型技术，特斯拉推出通用世界模型，通过历史视频片段预测未来场景；Wayve 发布 GAIA-1 模型，生成分钟级驾驶视频以训练自动驾驶系统；英伟达（NVIDIA）在 CES（International Consumer Electronics Show，国际消费类电子产品展览会）发布 Cosmos 世界基础模型（World Foundation Model，简称 WFM），支持自动驾驶与机器人应用；蔚来推出智能驾驶世界模型 NWM（NIO World Model），实现多场景推演与决策优化；AI“教母”李飞飞也创立了 World Labs 专门致力于空间智能和大世界模型的研发。世界模型作为 AGI 的核心技术，正从理论探索迈向产业落地，未来，具备因果推理能力的 AI 系统将能预测未来动态，解决跨任务协同等复杂问题。

世界模型可模拟复杂场景，在数字博物馆和智慧图书馆中，世界模型技术可用于辅助图书馆空间规划和 AR 导航系统，并构建沉浸式虚拟展览和智能互动系统，让观众跨越时空体验文化历史。这些技术尤其对历史数据的揭示、描述和推理服务提供新的可能。同时，因果推理技术可优化读者需求预测，并辅助决策者优化信息服务和内容传播策略，推动图书馆等公共文化机构在数字化时代的创新发展。

6 AI 算力与模型优化协同发展

在 GAI 技术的飞速发展中，算力和模型优化正呈现出协同发展、互相促进的态势。一方面，算力的提升为更大规模、更复杂模型的训练提供了硬件基础；另一方面，模型优化则能有效降低对算力的需求，提高计算效率。这种良性互动是推动 AI 技术不断进步的关键动力。有人说 DeepSeek 的出现降低了对算力的需求，“杰文斯悖论”（Jevons Paradox）① 告诉我们，这其实是不对的，技术效率的提升反而会扩大资源消耗总量，因此对算力硬件的需求其实是一个刺激。

2024 年，英伟达等公司在 GPU（Graphics Processing Unit，图形处理单元）算力提供方面持续创新。英伟达在 GTC（Global Trade Center，全球商品交易中心） 2024 上正式推出了 H200 和 H200X GPU，新款芯片相较于前代 H100 具备更低的功耗和更高的算力密度，并特别针对大模型推理进行了优化 [27]；AMD 也于 2024 年 7 月推出了 Radeon Instinct MI300 系列，使用 Chiplet 设计以实现多 GPU 高效协作，据称在部分基准测试中性能直追 NVIDIA H200[28]；英特尔则宣布在 2024 年底发布面向数据中心的 Gaudi 3 芯片，在混合精度计算上具有较高的性价比 [29]；Google 也在自家 TPU V6 上进一步升级了片上内存与自动混合精度的功能，可在大规模并行训练时减少数据移动开销 [30]。这些硬件升级为训练万亿参数级别的模型提供了坚实的硬件基础。根据 IDC（International Data Corporation，国际数据公司）[31] 与 Gartner 的预测报告 [32]，2024 年的全球 AI 芯片市场规模将达到 800 亿美元左右，硬件算力的井喷式发展直接推动了大模型的普及与应用，为 AI 生态的持续繁荣提供了稳固的基石。为应对日益庞大的 AI 模型计算需求，2025 年将见证 GPU 等专用 AI 硬件、低功耗计算

① 杰文斯悖论是 1865 年经济学家威廉·斯坦利·杰文斯提出的，核心观点为当技术进步提高了某种资源的使用效率后，非但不会减少该资源的总体消耗，反而可能导致其总需求激增。具体可参见 https://philosophyterms.com/jevons-paradox/。

设备以及高效数据中心的进一步发展（如英伟达 Digits）。新一代芯片和液冷系统、无碳能源数据中心等技术的应用，将推动 AI 基础设施向更高效、节能和可持续方向发展。

2024 年多项关于模型参数规模与性能的研究发现，单纯依靠无上限地扩充数据和增加算力的做法，边际收益已经明显下降 [33]。当然这并不意味着缩放定律就此失效，扩展模型在融入推理策略及结构化知识后，依旧能够获取显著的收益 [34]。DeepSeek R1 通过底层优化和软硬件协同创新，实现了 FP8 低比特训练出高质量模型。未来模型训练将从简单扩大规模转向重视数据质量、后训练（Fine-Tuning）和强化学习（Reinforcement Learning，简称 RL）技术的应用，从而实现模型泛化能力的提升与更高的性价比。

AI 算力与模型优化是相辅相成的，二者的协同发展能够加速推动 AI 的成熟并促进领域内落地。对于公共文化机构而言，有利于更快落地大模型应用，促进智慧升级，为公众提供更优质、更智能的文化体验。图书馆等机构在推进智慧化转型时，可以结合本地算力的配置，通过更大范围地依托高性能硬件，支持大规模数据存储、智能检索和提供互动体验，内置算力的各类自助服务机或支持本地化数据处理的图书馆智能终端将能够更高效快捷地提供全方位的智慧服务，同时保护自身的数据资产和用户的隐私。硬件升级不仅能提高服务响应速度，还能保障系统长期稳定运行，助力智慧化资源平台的建设和服务。

7 智能体技术普及带来产品爆发

智能体（Agent）是指基于大模型技术，通过整合感知、记忆、规划及工具等模块，实现对外部环境主动交互并自主执行复杂任务的软硬件系统。尽管智能体这一概念早已存在，但在大模型出现后，其内涵才得以真正实现。2025 年被视为智能体发展的元年，随着智能体框架的日益繁荣与标准化，智能体应用预计将迎来爆发式增长。

当前，智能体发展的主要趋势是将大语言模型与任务执行机制相结合，形成“理解—推理—行动”的完整闭环。例如，OpenAI 推出的 Swarm 智能体框架是一个通过智能体和交接机制（Handoff）实现多智能体协同工作的轻量级开源框架，专为复杂任务的高效编排与执行而设计 [35]。2025 年 1 月，OpenAI 又发布了其首款智能体应用——Operator[36]。该系统集成了实时检索、视觉感知、推理与规划、行动执行、自我纠错与学习等多项功能，能够在多种复杂场景中自动执行操作。Operator 无需通过 API（Application Program Interface，应用程序接口）即可直接操作浏览器，这种通用性更强的智能体模式有望将大模型应用推向新的技术高度，并加速 AI 在企业级场景中的深度应用。未来，机器人流程自动化（Robotic Process Automation，简称 RPA）、个人助理、客户服务和数据分析等领域将基于智能体技术开发新一代应用，AI 将从单纯的问答工具转变为具备高度自治能力的智能代理（Agentic AI）。这些智能代理不仅能够自动执行多步骤任务，还能在企业内部承担复杂的流程管理工作，从而彻底改变工作方式与业务流程。智能体的应用使得产品能够更准确地理解用户意图，进行自主决策和行动，从而提供更加个性化和智能化的服务。这种自主性和高效性将显著提升用户体验，用户只需通过自然语言与智能体进行交互，即可完成复杂任务，无需烦琐的操作。

这种新兴的 AI 应用模式正在深刻重塑产品形态，其在垂直领域的渗透、与实体物体的结合以及与物联网的深入融合，将为各行各业带来新的机遇。图书馆等公共文化服务机构可以利用智能体对现有服务系统进行升级。图书馆服务平台的各个模块可以利用智能体实现自动化客服、全天候智能问答和个性化推荐，从而提升用户互动体验。同时，在服务流程中，智能代理将优化流程管理、读者 / 用户数据分析与精准推荐、评估评价及咨询报告服务，如

自动化处理借阅、咨询等流程，甚至参与图书馆运营决策，实现效率与体验的双重提升。

8 资本投入与产业整合加速

2024 年，AI 技术发展显著，多款产品已在实际应用场景中取得广泛应用。在 AI 辅助编程领域，Copilot X Pro、Visual Studio Code AI Companion、Code Whisperer G2 以及 AliGen Code 等工具的推出，极大提升了开发者的编程效率，显著缩短了开发和调试周期，成为代码生产力的关键推动力 [37]。与此同时，大模型的应用范围已从网页端和移动端逐步扩展至智能硬件领域，包括手机、耳机、眼镜和智能音箱等。软硬件的深度融合推动了 AI 原生应用流量的增长。

全球主要科技公司和投资机构正加大对 AI 领域的投入，除硬件算力外，底层技术研发、数据中心建设、智能应用开发以及应用场景的拓展等方面也吸引了大量资本。资本市场的积极布局促使技术巨头与初创公司在各自的优势领域展开激烈竞争，推动 AI 产业生态的整合与升级。大型科技公司凭借雄厚的资金和技术积累，在 AGI 技术方面占据主导地位，并通过合作与收购巩固其市场地位。而初创公司则凭借其灵活性，专注于特定领域的创新应用。

预计 2025 年，AI 搜索及各类 AI 原生应用的多样化和深度化将延续 2024 年的繁荣态势，各企业在技术与商业模式上的竞争将更加激烈。资本与产业的整合将加速智慧图书馆、智慧文化馆和数字博物馆等新型应用的落地。图书馆应抓住资本投入带来的机遇，与科技公司合作以获取技术支持。市场力量将推动专业服务商的更新迭代，市场格局将发生改变，新的参与者将获得进入机会。未来，将有更多资金流入数字化改造项目，促进文化资源的共享与传播，同时也将催生新的商业模式和服务创新。

9 开源生态与小模型应用扩展

开源生态系统通过共享代码、数据和模型，降低了 AI 技术的学习和使用门槛，使得更多开发者和企业能够快速上手并进行创新，随着开源模型的影响力显著提升，中国开源社区的活跃度明显增强，并加速了技术的创新和迭代。以阿里巴巴的通义千问 Qwen 系列模型为例，据不完全统计，全球已有超 9 万基于 Qwen 的衍生模型 [38]。2025 年初，DeepSeek 带来的模型后训练与推理优化新模式展现了强大威力，新的训练模式和推理优化技术将使模型在特定场景下表现更优，它的开源直接推动了算力护城河的倒塌，其 AI 开源生态可望一举超越 Meta 成为 AI 领域的安卓。未来，算法创新和数据质量的竞争逐渐取代了算力垄断。这不仅深化了开源生态，也使得小模型的性能逐步增强，有研究表明，小型策略模型通过特别的优化技术（Test-Time Scaling，简称 TTS），在解决数学问题等特定任务时能在计算预算有限的情况下实现对大型模型的超越 [39]。今后，更多高级 AI 开始转向在个人设备上运行，这一趋势不仅降低了云端推理成本，还增强了用户隐私控制。

开源生态的繁荣和小型模型的优化为图书馆等公共文化服务机构提供了更多低成本、高效益的智能化解决方案。中国开源模型推动初创企业孵化、降低技术门槛，图书馆等机构不再受限于资金不足和应用不成熟而无法普及大模型应用，更多小公司能够以极低成本为图书馆提供智慧创新服务，这些公司可以利用经过优化后的小模型实现领域数据库的开发和工具定制，在自动化文本摘要、图像分类和元数据生成等各方面的个性化需求都可以用到 AIGC（Artificial Intelligence Generated Content，人工智能生成内容）的能力，从而提高数字资源的整理效率和信息可访问性，提供精准的信息推荐和个性化的学习方案制订等，降低服务人员的人力成本。有条件的机构还可以利用开源基座模型训练自己机构的领域模型，如可以通过私有部署 DeepSeek R1 构建本地 AI 服务，并

通过 DeepSeek Janus Pro 协同提供基于资源内容的智能体服务。

10 AI 伦理与治理框架完善

随着 GAI 技术的迅猛发展和广泛应用，其潜在风险和不确定性日益受到关注。为确保 AI 技术在可控、安全及符合伦理的框架内发展，完善 AI 伦理与治理框架已成为国际社会的普遍共识。当前，各国政府、产业界及学术界正积极探索并制定相关政策、标准和规范，以应对 AI 发展带来的诸多挑战。

2024 年，联合国通过多项决议 [40-41]，强调利用 AI 推动可持续发展，并呼吁各国建立监管和治理框架，以确保 AI 系统的人类中心性、可靠性和伦理合规性。世界卫生组织亦发布了针对多模态大模型的 AI 伦理和治理指南，提出了 40 多项建议，以确保此类模型的合理使用 [42]。中国信息通信研究院于 12 月发布了《人工智能治理蓝皮书（2024 年）》[43]，提出了“1244”AI 治理总体框架，即 1 组概念、2 类风险、4 类主体和 4 组议题，强调负责任创新、可持续发展、伦理先行和安全底线的重要性。

随着 AI 模型能力的不断提升和广泛应用，其潜在风险和不确定性也在增加，包括模型“幻觉”、数据泄露、伦理问题、劳动市场变化和能源消耗等。未来，各国政府和产业界将加快制定相关政策和标准，建立完善的 AI 安全治理体系，确保技术发展在可控范围内进行。AI 伦理与治理方面的国际合作将成为关键，制定统一或兼容的 AI 伦理标准，促进国际间的交流与合作，共同应对风险。

图书馆等公共文化服务机构是数据治理的前哨阵地，在数字化与智慧化转型过程中，尤其需要关注数据安全、隐私保护与版权管理，制定数据使用规范，避免算法歧视，确保 AI 服务的公平透明，并提供可信的资源服务。同时，通过图书馆的 AI 素养教育，将 AI 安全意识传递给读者。此外，还需建立健全安全机制和风险评估体系，确保文化信息资源在开放共享的同时得到有效保护，提升公众对数字文化服务的信任度。

2025 年，AI 正通过多模态融合、智能代理、硬件创新、数据治理及安全监管等多重趋势推动技术变革。无论是在科研、工业、商业还是文化传播领域，AI 都将在提升效率、优化服务及激发创新方面发挥日益重要的作用。对于图书馆、博物馆、美术馆等公共文化机构而言，AI 技术将深度重塑其服务模式。顺应这一趋势，不仅能够实现业务流程的再造和服务水平的提升，更能为公众提供全新的、沉浸式的数字文化体验。

未来，公共文化机构应优先考虑提升数字资源管理效率，探索具身智能与合成数据应用，以优化物理空间的服务能力。同时，需加强 AI 伦理与治理，规避潜在风险，并通过开源合作降低技术成本。智慧图书馆将成为未来发展的必然方向，而 AI 的加持将为智慧图书馆的发展奠定坚实基础。这需要各方共同努力，以期在各个方面取得里程碑式的进展。

参考文献：

[1] DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning[EB/OL].[2025-01-25].https://github.com/deepseek-ai/DeepSeek-R1/blob/main/DeepSeek_R1.pdf.

[2] OECD.Artificial intelligence in science: challenges, opportunities and the future of research[EB/OL].[2025-01-25].https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/06/artificial-intelligence-in-science_4f3d6efd/a8d820bd-en.pdf.

[3] OpenAI.Competitive programming with large reasoning models[EB/OL].[2025-01-25].https://arxiv.org/pdf/2502.06807.

[4] MALLICK S B,KILPATRICK L.Gemini 2.0: flash,flash-lite and pro [EB/OL].[2025-01-25].https://developers.googleblog.com/en/gemini-2-family-expands/.

[5] 周代数 , 魏杉汀 . 人工智能驱动的科学研究第五范式 : 演进、机制与影响 [J]. 中国科技论坛 ,2024(12):97–107.DOI:10.13580/j.cnki.fstc.2024.12.014.
[6] ABRAMSON J,ADLER J,DUNGER J. et al.Accurate structure prediction of biomolecular interactions with AlphaFold 3[J/OL]. Nature, 2024,630:493–500[2025–01–25].https://www.nature.com/articles/s41586–024–07487–w#citeas.
[7] 刘炜 , 嵇婷 . 数字人文的未来 20 年 [J]. 信息与管理研究 ,2024,9(6):1–11.
[8] 范炜 , 曾蕾 .AI 新时代面向文化遗产活化利用的智慧数据生成路径探析 [J]. 中国图书馆学报 ,2024,50(2):4–29.DOI:10.13530/j.cnki.jlis.2024010.
[9] 2024 世界人工智能大会 | AIGC"黑科技"助力敦煌遗书数字化修复 [EB/OL].[2025–01–25].https://tech.gmw.cn/2024–07/04/content_37421044.htm.
[10] MARCHANT J.How AI is unlocking ancient texts—and could rewrite history[J/OL].Nature,2025,637:14–17[2025–01–25]. https://doi.org/10.1038/d41586–024–04161–z.
[11] BREEN B.Simulating history with ChatGPT[EB/OL].[2025–01–25].https://resobscura.substack.com/p/simulating–history–with–chatgpt.
[12] IFLA.Developing a library strategic response to artificial intelligence [EB/OL].[2025–01–25].https://www.ifla.org/g/ai/developing–a–library–strategic–response–to–artificial–intelligence/.
[13] GGII:2024年中国四足机器人行业发展报告[EB/OL].[2025–01–25].https://news.qq.com/rain/a/20241005A03SNM00?suid=&media_id=.
[14] 国投证券 .2024 年特斯拉研究报告 :Optimus 开辟第二战场 , 推动特斯拉再进阶 [EB/OL].[2025–01–25].https://www.vzkoo.com/read/20241008abf469e1a62b49d399c1dbc2.html.
[15] BostonDynamics.Atlas® and beyond: the world's most dynamic robots[EB/OL].[2025–01–25].https://bostondynamics.com/atlas/.
[16] 优必选科技 .WalkerX[EB/OL].[2025–01–25].https://www.ubtrobot.com/cn/humanoid/products/WalkerX.
[17] BADASIE C.Digit,the humanoid robot taking over amazon warehouses[EB/OL].[2025–01–25].https://www.giantfreakinrobot.com/sci/digit–amazon.html.
[18] LIU Y,CHEN W,BAI Y, et al.Aligning cyber space with physical world: a comprehensive survey on embodied AI[EB/OL].[2025–01–25].https://arxiv.org/pdf/2407.06886.
[19] Innovative framework drives autonomous learning and task mastery[EB/OL].[2025–01–25].https://www.cmu.edu/news/stories/archives/2024/october/innovative–framework–drives–autonomous–learning–and–task–mastery.
[20] LIU J M,LIU M Z,WANG Z Y,et al.RoboMamba: efficient vision–language–action model for robotic reasoning and manipulation[EB/OL].[2025–01–25].https://arxiv.org/pdf/2406.04339.
[21] CHEN X K,WU Z Y,LIU X C,et al.Janus–Pro: unified multimodal understanding and generation with data and model scaling[EB/OL].[2025–01–25].https://arxiv.org/pdf/2501.17811.
[22] Meta.Introducing Llama 3.1: our most capable models to date [EB/OL].[2025–01–25].https://ai.meta.com/blog/meta–llama–3–1/.
[23] OpenAI.OpenAI o1 system card[EB/OL].[2025–01–25].https://openai.com/index/openai–o1–system–card/.
[24] DeepSeek–V3 technical report [EB/OL].[2025–01–25].https://github.com/deepseek–ai/DeepSeek–V3/blob/main/DeepSeek_V3.pdf.
[25] YANN L C.A path towards autonomous machine intelligence[EB/OL].[2025–01–25].https://openreview.net/forum?id=BZ5a1r–kVsf.
[26] DING J T,ZHANG Y K,SHANG Y,et al.Understanding world or predicting future? A comprehensive survey of world models[EB/OL].[2025–01–25].https://arxiv.org/pdf/2411.14499.
[27] Hopper scales new heights,accelerating AI and HPC applications for mainstream enterprise servers[EB/OL].[2025–01–25].https://blogs.nvidia.com/blog/hopper–h200–nvl/.
[28] AMD Instinct ™ MI300 series accelerators[EB/OL].[2025–01–25].https://www.amd.com/en/products/accelerators/instinct/mi300.html.
[29] Intel® Gaudi® 3 AI accelerators[EB/OL].[2025–01–25].https://www.intel.com/content/www/us/en/products/details/

processors/ai-accelerators/gaudi3.html.
[30] 隆重推出第 6 代 Google Cloud TPU:Trillium[EB/OL].[2025-01-25].https://blog.google/intl/zh-tw/products/cloud/2024_05_introducing-trillium-6th-gen-tpus/.
[31] Semiconductor market outlook: AI investment supercycle [EB/OL].[2025-01-25].https://www.gsaglobal.org/events/wp-content/uploads/sites/10/2024/03/sp-1v1-mario-morales-idc-gsa-presentation-october-30th-2024.pdf.
[32] Gartner predicts worldwide AI chip revenue will gain 33% in 2024[EB/OL]. [2025-01-25].https://www.techrepublic.com/article/gartner-ai-chip-revenue-2024/.
[33] ZENG Z Y,CHENG Q Y,YIN Z Y,et al.Scaling of search and learning: a roadmap to reproduce o1 from reinforcement learning perspective[EB/OL].[2025-01-25].https://arxiv.org/abs/2412.14135.
[34] TRAN H,YAO Z H,WANG J D,et al.RARE:retrival-augmented reasoning enhancement for large language models [EB/OL].[2025-01-25].https://arxiv.org /abs/2412.02830.
[35] Openai/swarm[EB/OL].[2025-01-25].https://github.com/openai/swarm.
[36] 可联网自主完成任务！OpenAI 发布智能体 Operator, 给 AI Agent 又添了把火！[EB/OL].[2025-01-25].https://finance.sina.com.cn/roll/2025-01-24/doc-inefzsec5281898.shtml.
[37] DENIZ B K. Unleashing developer productivity with generative AI[EB/OL].[2025-01-25].https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/unleashing-developer-productivity-with-generative-ai#/.
[38] Huggingface 开源大模型排行榜 : 通义千问 Qwen 成开源翘楚 [EB/OL].[2025-01-25].https://ai-kit.cn/7083.html.
[39] LIU R Z,GAO J Q,ZHAO J,et al.Can 1B LLM surpass 405B LLM? Rethinking compute-optimal test-time scaling[EB/OL].[2025-01-25].https://arxiv.org/abs/2502.06703.
[40] 抓住安全、可靠和值得信赖的人工智能系统带来的机遇 , 促进可持续发展 [EB/OL].[2025-01-25].https://documents.un.org/doc/undoc/ltd/n24/065/91/pdf/n2406591.pdf.
[41] 加强人工智能能力建设方面的国际合作 [EB/OL].[2025-01-25].https://documents.un.org/doc/undoc/gen/n24/197/25/pdf/n2419725.pdf.
[42] World Health Organization.Ethics and governance of artificial intelligence for health: guidance on large multi-modal models[EB/OL].[2025-01-25].https://www.who.int/publications/i/item/9789240084759.
[43] 中国信息通信研究院.人工智能治理蓝皮书(2024年)[EB/OL].[2025-01-25].https://www.caict.ac.cn/kxyj/qwfb/bps/202412/P020241227660032159191.pdf.

[作者简介]

刘 炜 上海社会科学院信息研究所所长、研究员，研究方向为数字人文、智慧图书馆、知识组织、人工智能应用和 Web3.0 等。E-mail：wliu@sass.org.cn。
刘倩倩 上海图书馆 (上海科学技术情报研究所) 数据馆员，研究方向为数字人文、数据处理与平台建设，本文通信作者。E-mail：qqliu@libnet.sh.cn。

[收稿日期：2025-01-25]