**智慧化知识服务技术研究**

**1. 任务要求**

智慧化知识服务技术研究任务第一批（孙宇飞）

任务时间：20250711-20250724

任务详情：

技术项：大语言模型、智能体/代理型AI、场景识别、VR/AR

1. 定义
2. 起源和关键发展阶段
3. 演进趋势、潜在的技术瓶颈
4. 与知识服务场景的融合应用可能性、契合度（关注六大场景：知识图谱、语义搜索、智能问答、个性化推荐、知识推送、可视化服务）

说明：

1. 大语言模型需要按照发展脉络梳理不同功能侧重方向的大模型，并列出典型模型有哪些，包括但不限于：内容创作、对话代理；多模态处理；增强推理；
2. 演进趋势等可以关注Gartner等相关权威报告；
3. 场景识别本阶段可以初步列出涉及到哪些与知识服务相关的技术；
4. 与知识服务场景的融合应用可能性、契合度这一项在本阶段可以先初步做；
5. 所有输出的定义、介绍等内容需要有权威出处，并添加脚注引用；
6. 谨慎使用AI工具，需自行校验所输出结果的准确性和可靠性。

# **2. 大语言模型（Large Language Models）**

## **2.1 定义**

大语言模型（Large Language Models, LLMs）是一种基于海量文本数据训练的超大规模深度学习模型，具备理解和生成自然语言内容的能力，可执行多种任务。[[1]](#footnote-1)

LLM通常采用自监督学习，从大规模语料中学习语言模式，尤其擅长语言生成等自然语言处理任务，其核心在于利用Transformer架构，通过自监督学习捕捉语言的深层规律，从而具备强大的自然语言理解（NLU）和生成（NLG）能力。斯坦福大学以人为本人工智能研究所（HAI）将LLM定义为一种基础模型（Foundation Model），特指那些在广泛的、未标记数据上进行预训练，并能适应（如通过微调）多种下游任务的模型。[[2]](#footnote-2)

这些模型拥有数以十亿乃至万亿计的参数，使其能够存储和处理关于世界的大量“参数化知识”，并执行包括问答、摘要、翻译、内容创作和代码生成在内的复杂任务。[[3]](#footnote-3) 简单的来说，可以把大语言模型定义为“包含数百亿或更多参数的、基于Transformer的神经语言模型”。[[4]](#footnote-4)

这一定义的关键在于，它明确地将LLM与早期的PLM区分开来，其核心区别在于涌现能力的存在，这些能力“在小规模语言模型中是不存在的” 。 值得注意的是，尽管LLM在语言任务上表现出色，但它们本质上是基于概率的文本序列预测器，可能产生不符合事实的“幻觉”（Hallucination），并且其知识截止于训练数据的时间点，这使其在处理需要实时性或高事实性信息的场景时面临挑战。[[5]](#footnote-5) 下表总结了现有研究对LLM的定义要点和突出特征：

| 研究来源 | 定义要点 | 突出特征 |
| --- | --- | --- |
| Zhao等(2303.18223) | 大规模预训练语言模型 | • 涌现能力• 规模效应• 预训练-微调-提示范式 |
| Minaee等(2402.06196) | 基于Transformer的十亿级参数模型 | • 主流模型家族• 涌现能力• 先进训练方法 |
| Stanford CRFM | 基于大规模数据训练的多任务适应模型 | • 能力涌现• 模型同质化 |
| OpenAI | 基于词元预测的文本映射函数 | • 预测机制• 文本学习• 少样本能力 |
| Google | 语言预测与生成模型 | • Transformer结构• 注意力机制• 涌现特性 |

综合以上所有分析，我们可以给出一个全面而严谨的定义：大型语言模型（LLM）是一种由以下三个核心属性共同定义的语言模型：

1.架构基础：它构建于Transformer架构之上，利用自注意力机制的并行化能力，实现了前所未有的训练规模。[[6]](#footnote-6) [[7]](#footnote-7)

2.训练范式：它通过在海量的、多样化的文本数据（并越来越多地包括代码、图像等多模态数据）上进行大规模自监督预训练，学习语言和世界知识的通用表示 。[[8]](#footnote-8) [[9]](#footnote-9)

3.规模诱导的能力：其规模（通常指拥有数百亿乃至更多的参数）足以跨越一个质变阈值，从而涌现出在小型模型中不存在的特殊能力。这些涌现能力，如情境学习和复杂的逐步推理（如思维链），是区分LLM与传统PLM的决定性特征。[[10]](#footnote-10) [[11]](#footnote-11)

## **2.2 起源和关键发展阶段**

自2017年以来，人工智能领域迎来了“大语言模型”（Large Language Models, LLM）的突破性进展。2017年谷歌提出Transformer架构，使得模型可以通过“自注意力”机制高效地建模序列数据，为大语言模型的崛起奠定了技术基石[[12]](#footnote-12) 。

与之前的深度学习模型相比，大语言模型可以在极其庞大且多样的非结构化数据上进行训练，并凭借统一的架构执行多种任务，这是人工智能发展中的一次“跨越式演进”[[13]](#footnote-13) 。 基于Transformer的预训练模型（即“基础模型”）通过自监督方式在海量文本上学习，拥有极其庞大的参数规模，能够在下游以少量示例完成多样化任务[[14]](#footnote-14) 。 得益于此，新一代生成式人工智能应用（如ChatGPT、GitHub Copilot、Stable Diffusion等）展现出前所未有的通用性，几乎任何人都可以利用它们进行交流和创作。这种广泛实用性以及近似与人对话的能力使其受到全球公众的热烈关注，影响力远超此前AlphaGo等专门AI系统[[15]](#footnote-15) 。 例如，ChatGPT自2022年底推出以来，迅速累积了亿万用户，成为技术史上用户增长最快的应用之一，引发了社会各界对生成式AI的高度关注和讨论[[16]](#footnote-16) 。

随着大语言模型在文本生成、对话交互、知识问答等方面展现出强大性能，各大科技公司和研究机构纷纷投入竞争，包括OpenAI、Google DeepMind、Meta、Anthropic、Mistral等在内的机构相继推出一系列具有代表性的模型系统，推动了大语言模型技术的快速演进。 本节将梳理2017年至今大语言模型技术的发展脉络与关键阶段，分析不同模型在内容创作、对话代理、多模态处理、增强推理等功能方向上的侧重与代表性进展。

### **Transformer的诞生与预训练范式的兴起（2017-2018年）**

2017年可以视作大语言模型发展的起点。这一年，谷歌研究员在论文《Attention Is All You Need》中提出了Transformer模型架构[[17]](#footnote-17) 。 Transformer通过多头自注意力机制显著提升了序列建模能力，摆脱了传统循环神经网络在长程依赖建模上的性能桎梏。 Transformer架构的出现，使得训练更大规模的语言模型成为可能，也为后来的预训练-微调范式奠定了基础。

基于Transformer的预训练模型很快展现出卓越的性能。2018年，谷歌发布了BERT（Bidirectional Encoder Representations from Transformers）模型[[18]](#footnote-18) 。 BERT采用Transformer的编码器，对海量文本进行双向特征表征预训练，然后再微调用于下游任务。这一创新在阅读理解、问答、文本分类等自然语言理解任务上取得当时最先进的效果，引发了NLP领域对预训练模型的热情。同年，OpenAI提出了生成式预训练模型GPT的早期版本（通常被称为GPT-1）。与BERT不同，GPT采用Transformer的解码器进行单向语言模型预训练，重点在于生成自然语言文本的能力[[19]](#footnote-19) 。 GPT-1在无监督预训练后，再通过有监督微调适应具体任务的做法，验证了预训练模型在文本生成和理解上的巨大潜力[[20]](#footnote-20) 。

2018年前后还出现了ELMo等预训练模型，它通过双向LSTM从文本中学习词表示。但Transformer架构的优势很快使其成为主流。不论是BERT等偏重理解的模型，还是GPT等偏重生成的模型，都证明了通过在海量语料上的自监督预训练，模型可以掌握广泛的语言知识，并以较少任务特定数据达到优异性能。这一阶段奠定了“基础模型”的概念：即拥有海量参数、在海量数据上自监督训练、可适应多任务的大模型[[21]](#footnote-21) 。 正如Gartner的分析所指出的，这类基础模型代表了深度学习的一次范式转变[[22]](#footnote-22) 。 从技术起源来看，Transformer和自监督预训练是大语言模型演化的原点。

### **模型规模扩张与文本生成能力突破（2019-2020年）**

2019年至2020年间，大语言模型最显著的趋势是参数规模的急剧扩张和生成文本质量的飞跃式提升。OpenAI在2019年发布了GPT-2模型，以15亿参数的规模生成连贯文本，被认为在通用文本生成上取得了突破。GPT-2展示出强大的续写文章和模拟对话的能力，引起学界和业界震惊。OpenAI最初出于安全考虑没有公开GPT-2的全部模型，担心其可能被用于大规模生成虚假信息，这从侧面证明了模型生成能力之强大[[23]](#footnote-23) 。GPT-2的出现标志着“内容创作”型语言模型开始展露锋芒——模型能够根据提示自动续写故事、新闻，甚至模仿特定风格的文本。这为日后诸多文本生成应用（如智能写作助手、自动对话机器人）的兴起打下基础。

同一时期，谷歌等也在探索更大的预训练模型。2019年底，谷歌发布了T5（Text-to-Text Transfer Transformer）模型，它将所有任务统一表示为文本到文本的转换，并以110亿参数在C4海量语料上训练[[24]](#footnote-24) 。T5在翻译、摘要等生成任务上取得当时最好成绩。可以说，2019年的这些模型表明，提升参数规模和训练数据规模，能显著提高模型生成多样化文本的能力。这一认识推动“规模化”成为大语言模型发展的主要方向之一。

2020年是大语言模型史上的里程碑：OpenAI发布了GPT-3模型。GPT-3拥有惊人的1750亿参数[[25]](#footnote-25) （比GPT-2大两个数量级），在通用文本生成和理解任务上展现出跨越性提升。 GPT-3最引人注目之处在于它的**少样本学习**能力：无需任务特定微调，只需给出几个示例，GPT-3就能在翻译、问答、写作等各种任务上产生相当可靠的结果[[26]](#footnote-26) 。 这表明模型参数和训练数据的规模一旦达到一定阈值，预训练模型本身即可蕴含解决多种任务的通用能力，被视作人工智能发展中的重大突破。GPT-3在发布后引起广泛关注和研究热潮，一方面因为其性能优异，另一方面也因为如此规模模型的训练涉及巨大的算力和成本（据估计GPT-3训练成本高达数百万美元）[[27]](#footnote-27) 。 麦肯锡的研究指出，ChatGPT（基于GPT-3.5）的横空出世和GPT-4的推出仅相隔数月，这种技术演进速度在AI史上前所未有[[28]](#footnote-28) 。总的来看，2019-2020阶段，OpenAI的GPT系列和谷歌的T5等模型将大语言模型推向“超大规模”时代，模型在通用**内容创作**方面的能力获得极大提升。

### **功能多元化：对话代理、多模态与推理（2021-2022年）**

经历了早期规模驱动的发展后，2021年至2022年的大语言模型开始在**功能多元化**方向取得进展，包括更好的对话能力、更强的推理计算，以及跨模态的扩展等。同时，多家公司和机构加入竞争，推出各具特色的模型体系，使这一阶段成为大语言模型百花齐放的时期。

**对话代理方面**：谷歌在2021年公布了LaMDA（Language Model for Dialogue Applications）模型，专门针对对话进行优化[[29]](#footnote-29) 。LaMDA具有1370亿参数，通过在大量对话数据上微调，显著提升了上下文对话的连贯性和针对开放性话题的应答能力。谷歌演示了LaMDA就任意话题进行富有创造力的对话，如扮演“冥王星”与用户交流。这代表大语言模型开始从“一次性文本生成”向**持续对话**方向延伸。Facebook（后更名Meta）也在此期间研发对话模型，例如BlenderBot系列，尝试让聊天机器人展现类似人类的个性与长期记忆。然而，这些早期对话模型依然面临上下文缺失、易产生不当言论等问题，提示需要在安全和连贯性上进一步改进。

**推理能力与工具使用**：大型语言模型在2021-2022年展现出初步的推理和复杂任务解决能力。例如，DeepMind在2021年底发布了Gopher模型（2800亿参数）和2022年的Chinchilla模型（700亿参数）[[30]](#footnote-30) 。 特别是Chinchilla通过实验发现，在固定算力下适当减少参数规模、增加训练语料量，可以让模型在推理问答等任务上达到比更大模型（如Gopher）更好的效果[[31]](#footnote-31) 。 这被称为*“Chinchilla 策略”*，说明参数数量并非无限制地越大越好，数据量和模型规模需平衡以提升推理表现。同时，研究者开始探索利用提示工程让模型进行链式推理（Chain-of-Thought）。2022年谷歌等工作的结果显示，给模型示例如何分步思考、逐步推导，可以显著提高模型解决数学推算、逻辑推理等任务的正确率。类似地，Prompting策略ReAct让模型在回答问题时先产生思考步骤、并可调用工具（例如检索互联网或计算），增强了复杂问答的准确性。这些方法丰富了大语言模型的**增强推理**能力，也预示着未来模型与外部工具结合的方向。

**代码生成**也是此阶段的重要突破之一。2021年OpenAI推出了Codex模型，它在GPT-3的基础上继续在海量源代码上训练，使其能够根据自然语言描述生成对应的代码。这一模型被用作GitHub Copilot的底层引擎，为开发者提供自动补全和代码生成功能。DeepMind则研发了AlphaCode，通过生成候选程序并测试筛选，成功在编程竞赛问题上达到中等水平选手的水准[[32]](#footnote-32) 。代码生成模型实质上也是大语言模型的一种特殊应用（将代码视为一种语言），它体现了大模型在**领域专用**方向的威力。

**多模态融合**方面，研究者开始尝试将图像、文本等不同模态的信息结合。虽然严格来说许多图像生成模型（如2021年的DALL-E、2022年的DALL-E 2和Stable Diffusion）并非语言模型，但语言模型技术也被用于图文结合的场景。例如，2022年DeepMind发布Flamingo模型，它能够将视觉信息融入语言模型，以图文对话的形式回答有关给定图像的问题。这预示着未来的大模型将不局限于单一文本模态，而是朝**多模态AI**方向演进。

**开放社区的参与**也是2022年的一大亮点。面对少数巨头掌控超大模型的局面，学术和开源社区发起了“大模型民主化”运动。2022年5月，Meta开源了OPT-175B模型（1750亿参数），这是当时与GPT-3体量相当的模型，虽然性能略有差距但开放获得使用。在7月，多国研究者合作的BigScience项目发布了BLOOM模型（1760亿参数），支持包括中文在内的46种语言并开放研究使用。这些举措降低了研究者接触超大模型的门槛。此外，Anthropic等初创公司在2022年也开始亮相，其关注点在于大模型的**安全和对齐（Alignment）**，他们训练的模型（Claude的早期版本）探索了通过“宪法”引导模型行为的训练方法，以减少有害输出。总的来说，2021-2022阶段，大语言模型在纵向上继续扩大规模、提升性能，在横向上拓展出对话、代码、多模态、推理等多种功能方向，技术生态更加丰富多元。

### **爆发与竞逐：生成式AI浪潮下的多极化竞争（2023年至今）**

2023年被广泛认为是“大语言模型全面走向主流”的年份[[33]](#footnote-33) 。这一年里，大语言模型不仅在技术上继续演进，更通过产品化走进公众视野，行业竞争格局也更加多极化。主要体现在：

**ChatGPT引爆大众市场**：虽然ChatGPT在2022年11月问世，但其影响真正发酵于2023年。ChatGPT基于OpenAI的GPT-3.5模型，并经过人类反馈强化学习（RLHF）的微调，使其能够更贴近人类指令行事。由于ChatGPT可以以对话形式回答几乎任何提问、撰写文章、编写代码，甚至进行一定的推理分析，它迅速风靡全球，短短两个月用户即突破一亿。Gartner报告指出，ChatGPT等对话式大模型的流行使生成式AI在各行业引发了根本性冲击：企业开始重新思考业务流程，人力资源价值也受到重新评估，大模型热潮已达到炒作周期的顶峰[[34]](#footnote-34) 。 各国政府也注意到这股趋势，纷纷研究其影响并准备相应的监管措施[[35]](#footnote-35) 。 ChatGPT的成功标志着**对话代理**型大语言模型的成熟，也证明了通过大规模人机交互微调（如RLHF）可以极大提升模型的实用性和安全性。

**新一代模型的发布**：2023年3月，OpenAI发布了GPT-4模型。这是GPT系列的最新力作，参数规模虽未公开但据推测可能近万亿级。GPT-4在一系列专业和学术基准上达到人类水平表现，包括模拟法学院考试、医师资格考试等，展示出令人惊异的推理和理解能力[[36]](#footnote-36) 。更重要的是，GPT-4首次引入**多模态**能力，能够接受图像和文本输入，并输出文本[[37]](#footnote-37) 。这意味着用户可以给GPT-4提供一张图片，让模型根据图中内容回答问题，实现视觉和语言的融合。OpenAI在技术报告中称GPT-4为“大规模多模态模型”，其推出标志大语言模型进入“文本+视觉”时代[[38]](#footnote-38) 。GPT-4在生成质量、知识范围、推理连贯性等方面较前代有显著提升，被广泛认为是当前最强大的语言模型之一。

谷歌与DeepMind在这一年也加强了投入。谷歌于2023年5月的I/O大会上发布了PaLM 2模型，并将其整合进对话机器人Bard中[[39]](#footnote-39) 。 PaLM 2在代码理解、数学推理、多语言能力上较前代PaLM有改进，支持谷歌全线产品的AI功能升级。同年，DeepMind与谷歌Brain合并为新的Google DeepMind，预告将推出名为Gemini的下一代大模型，被寄望融合AlphaGo系列的强化学习长处与语言能力。据报道，Gemini将是一种从设计上即具备多模态和工具使用能力的模型，定位为GPT-4的竞争者。可见，**OpenAI与谷歌**的“双强”竞争在2023年达到白热化，双方都将大语言模型作为AI战略的核心。

**开源与社区力量**：Meta在2023年2月开放发布了LLaMA模型（共有7B、13B、33B、65B四种规模）。尽管LLaMA仅限研究用途，但其泄漏的权重意外地在社区传播开来，引发了开源界前所未有的创新活力。研究人员和开发者基于LLaMA，迅速开发出各种精调变体，例如Stanford的Alpaca（在LLaMA上进行指令微调），以及诸如Vicuna、ChatGLM等高性能的对话模型。7月，Meta干脆顺势发布了**LLaMA 2**模型，并提供较宽松的开源许可（允许商用），进一步推动了开放生态的发展。LLaMA 2在70B参数规模上据称已接近GPT-3.5水平，并提供了对话优化版。在开源社区努力下，模型蒸馏、量化等技术层出不穷，使得在消费级硬件上运行中等规模的大模型成为现实。Forrester在2024年的报告中指出，开源LLM正在重新定义市场格局，尽管训练和维护顶尖LLM需要庞大数据和基础设施投入，但开放社区的贡献为企业提供了新的选择[[40]](#footnote-40) 。

2023年还涌现了**新创公司**的身影。其中Anthropic公司推出了Claude模型，与OpenAI的ChatGPT直接竞争。Claude采用“宪法式AI”训练策略，强调在不引入人类偏见的情况下通过AI自我调整来实现对齐，其第二版Claude 2在安全性和性能上都有提升，并提供了最高可达10万tokens的超长上下文窗口[[41]](#footnote-41) 。这一100K上下文长度的实现使Claude能够消化一本小说长度的文本，在商业应用中具有吸引力[[42]](#footnote-42) 。此外，创业公司如**Mistral AI**在模型效率上取得进展。Mistral于2023年9月发布了仅有70亿参数的Mistral 7B模型，但通过架构改进实现了对更大模型的性能超越：据称Mistral 7B在各项基准上全面超越了LLaMA 2的130亿参数模型，甚至媲美LLaMA 1的340亿参数模型[[43]](#footnote-43) 。它采用分组查询注意力和滑动窗口机制，不仅提升了推理速度，也可在较小计算开销下处理较长序列[[44]](#footnote-44) 。Mistral 7B的开源发布（Apache 2.0许可）表明，小型高效模型可能成为大模型发展的另一重要方向[[45]](#footnote-45) 。

整体而言，截至2023年，大语言模型领域呈现群雄并立的局面：OpenAI、谷歌（DeepMind）、Meta等科技巨头引领着最前沿的大模型研发与应用落地；Anthropic、Mistral等新锐公司通过差异化路线（安全对齐、效率优化等）参与竞争；开源社区更是以前所未有的热情繁荣发展。在内容创作、对话交互、多模态理解、复杂推理等各方面，都涌现出代表性系统。从技术演进脉络看，Transformer架构引领下的大模型经历了从**规模驱动**到**能力拓展**再到**应用爆发**的阶段性演变。下一步，它将走向何方？面临哪些挑战？以下将深入探讨。

## **2.3 演进趋势与潜在技术瓶颈**

### 2.3.1 演进趋势和展望

**模型规模的扩展与高效化**

自“大规模预训练”范式确立以来，不断增加模型参数和训练数据量是提升语言模型性能的主要途径之一。从GPT-2的15亿参数到GPT-3的1750亿参数，再到GPT-4传闻中近万亿的参数规模，模型大小呈指数级增长。然而，规模扩展带来收益的同时也面临边际效用递减和资源受限的问题。因此未来在“更大”与“更高效”两方面都会出现趋势。

一方面，**追求更大规模**仍将持续。更大的模型有望拥有更强的知识记忆和推理能力，在复杂任务上取得更接近人类的表现。业界普遍认为，参数规模尚未达到认知智能的上限，尤其是当配以足够训练数据时，性能可能进一步提升。例如，OpenAI等在训练GPT-4时就投入了前所未有的计算资源，以期探索规模效应的极限[[46]](#footnote-46) 。可以预见，未来或将出现参数量以万亿计的超大语言模型，具备更深的理解力和更少的统计偏差。

另一方面，“**小而精**”成为新的关注点。正如Forrester报告所强调的，训练和维护顶尖LLM需要巨量数据和基础设施，非少数巨头外的参与者难以企及[[47]](#footnote-47) 。因此业界开始寻求在更小模型上实现接近大模型的性能。这包括通过更高效的架构和训练策略来提升单位参数的能力。例如Mistral 7B模型仅靠70亿参数便达到比肩数十亿参数模型的效果[[48]](#footnote-48) ，证明了模型效率优化的巨大潜力。又如DeepMind的Chinchilla方案，在算力固定情况下缩减模型规模而增加训练量，反而取得更佳效果[[49]](#footnote-49) 。未来，我们将看到更多技术用于提升**参数效率**：模型蒸馏让小模型学习大模型行为，稀疏MoE（专家混合）架构通过条件激活部分网络以节省计算，量化和剪枝降低模型复杂度而性能损失最小化等等。这些方向旨在缓解模型规模带来的高成本壁垒，使大语言模型更加普惠可用。正如麻省理工《科技评论》所述，业界已经出现“**小模型的大作为**”的趋势：较小的模型训练更快、部署更廉价，并可在本地设备上运行[[50]](#footnote-50) 。综上，未来大语言模型的发展在规模上将呈两极：一极冲刺更大、更通用的“巨无霸”模型，另一极打磨更小、更高效的专业模型。两者相辅并存，满足不同应用场景的需求。

**多模态集成**

**多模态人工智能**被广泛认为是AI演进的下一前沿。人类智能的体现往往是多感官协同的，类人AI也需要能理解和生成不同形式的信息。大语言模型正朝这一方向扩展，其核心是将文本、图像、音频乃至视频等模态的信息表示统一融合，从而在一个模型中处理多样输入输出。

目前的探索已经初显成果。OpenAI的GPT-4率先实现了图文多模态能力：该模型能够将图像作为输入，与文本共同处理，然后生成文本输出[[51]](#footnote-51) 。实例如：用户上传一张含复杂内容的图片并提问，GPT-4可以理解图中场景并给出文字回答。这种视觉-语言结合使模型具备基本的“看图说话”能力，也被视为通往通用人工智能（AGI）的重要一步。Google DeepMind正在研发的Gemini据报道将天然支持多模态，有望在单个模型中同时具备视觉、语言甚至动作控制能力。可以想见，不久的将来数字助理不仅能对话，还能“看见”用户所见、“听到”语音，甚至控制机器实体互动，实现更加自然的人机交互。

多模态集成带来的一个直接好处是模型具备**跨领域知识融合**的能力。比如，模型可以读懂一张医学影像并结合病历文字给出诊断建议，或在观看视频后用文字总结其内容。在这一过程中，大语言模型作为核心起到综合推理和生成的作用。麻省理工的研究发现，大型语言模型在处理多模态数据时，会倾向于在内部形成一种抽象的通用语义表示，这种机制与人脑处理不同感官信息的方法类似[[52]](#footnote-52) 。这提示多模态模型有潜力达到对世界更全面的认知。当然，实现真正统一的多模态模型仍有挑战，包括模型架构如何高效处理图像/音频张量，与文本表示对齐训练，以及不同模态间的权衡等。不过总体趋势是明确的：**单一模态的AI模型将让位于多才多艺的综合模型**。

在应用层面，多模态大模型将开创许多新场景。例如，营销领域可借助模型自动生成图文搭配的广告创意；教育领域的AI导师可以分析学生表情肢体（视觉）与口头回答（语音）来提供反馈；机器人领域则能让AI同时处理摄像头画面和指令，从而更智能地感知环境与执行操作。Gartner在2023年将多模态AI视为创新技术趋势的一部分，预期未来2-5年内多模态生成与理解将走向成熟[[53]](#footnote-53) 。因此，多模态集成无疑是大语言模型演进的关键路径之一，未来的模型将不再局限于“语言”，而是真正意义上的多模态大模型。

**上下文扩展与持续对话**

增强**上下文记忆和对话持续性**是提升大语言模型实用性的重要方向。传统语言模型有固定的上下文长度限制（如GPT-3约4096个token），超出窗口的内容将无法被模型记忆。这限制了模型处理长文档或进行长程对话的能力。因此，扩大上下文窗口乃大势所趋。Anthropic公司在2023年率先推出了扩展上下文的Claude模型，将上下文长度从原来的9000 token增加到惊人的100000 token[[54]](#footnote-54) 。这一扩展意味着模型一次可以阅读约75,000字的内容，相当于一本中篇小说的长度，并可在几乎不丢失细节的情况下进行总结和问答[[55]](#footnote-55) 。OpenAI的GPT-4也提供了长上下文版本（最大支持32k tokens），支持用户提交长篇文章、程序代码等进行分析。上下文长度的提升使得**长文档理解**、**多轮长对话**成为可能：模型可以在对话中“记住”用户更早提出的信息或要求，不会像过去那样在对话稍长后就遗忘前文。对于企业应用，这一能力尤为关键，因为商业报告、法律合同等往往篇幅巨大，模型需要有“长记性”才能胜任辅助手的角色。

除了扩展静态的上下文窗口，另一个方向是在架构上赋予模型**持续记忆**或**外部存储**机制，让对话可以真正无限延续。研究者正在探索将神经网络与外部知识库或记忆模块结合，使模型能从之前的对话中不断累积知识。例如，有方法使用向量数据库实时记录对话要点，必要时检索出来提供给模型参考（这被称为检索增强型对话）。还有研究尝试用循环神经网络或Transformer的变体，使模型具备类似RNN那样的持久状态，突破固定窗口限制。但这些尚处于早期实验阶段。

上下文持续性的改进，不仅体现在对话长度上，也意味着模型能够保持**人格一致性**和**多轮推理连贯性**。对于聊天机器人来说，如果模型能“记住”用户以往的喜好和语气，那么交互体验将更趋近真人助手。同时，在需要分步骤思考的问题上，模型能在内部长期保留中间推理结果，避免因为上下文窗口滑动而丢失前提。这些改进都将显著提高大语言模型在对话代理场景下的实用性和智能程度。

**推理能力和工具使用的增强**

**推理能力**是当前大语言模型与真正智能体之间的鸿沟之一。传统的语言模型主要通过模式匹配来生成答案，缺乏可靠的逻辑推导过程。这导致模型有时会产生前后矛盾或不合常理的回应。因此，提升模型的推理和规划能力，使其不仅能“看起来合理”地回答，而且有真正可靠的逻辑依据，成为研究热点。

未来在提升推理方面有多个路径：

* **链式推理（Chain-of-Thought）内化**：在模型训练或推理时，引导其先生成中间推理步骤，再给出最终答案。这一技术在数学和复杂问答任务上已被证明有效，人们希望将其融入模型，使之成为默认行为。OpenAI的GPT-4据称在训练中已经加入了大量这样的示例，从而在推理题上表现出色。进一步，模型可以学习自己反复检查和验证推理过程（自我一致性），以降低错误率。
* **符号融合与程序执行**：所谓*“神经-符号”*融合思路是让大语言模型调用符号计算模块完成其不擅长的精确推理。这包括数学计算、逻辑证明、数据库查询等。实践中，出现了如Toolformer之类的方法，模型可以在回答过程中自动决定调用计算器或搜索引擎等工具[[56]](#footnote-56) 。例如，在需要计算“2025年比2023年多几天”这类问题时，模型可委托外部工具算出准确答案，而非凭内部权重猜测。Gartner将这种AI与外部工具协作的系统称为“复合AI”的一种形式，认为其能拓宽AI解决问题的范围[[57]](#footnote-57) [[58]](#footnote-58) 。
* **内置知识与常识推理**：大语言模型有海量参数，可以储存海量事实。但单纯记忆不足以推理，还需对知识点建立因果或类比联系。未来模型可能通过强化学习或预训练任务，习得基本物理常识、社会常识，从而在推理时更有“常识约束”，不会轻易给出荒谬结论。一些研究正尝试在训练中加入基于知识图谱的约束或监督，让模型显式掌握实体及其关系[[59]](#footnote-59) 。
* **自主Agent与规划**：2023年兴起了让大模型扮演自主智能体的探讨，如AutoGPT等概念，即由模型自我循环产生行动计划并执行，不断反馈直至完成复杂任务。这实际上考验模型的长期规划和上下文管理能力。尽管目前这些自主Agent还不成熟，但随着大模型推理与记忆增强，将来有望胜任一定程度的自主决策任务。

强化推理能力的本质，是希望大语言模型从“概率式写作”进化为“可信赖的思考者”。McKinsey的报告称，当前最先进的大模型依然缺乏对自身知识边界的认知，经常不知道自己何时不懂[[60]](#footnote-60) 。提升推理能力可以部分缓解这一问题，因为模型将更严格地推敲问题而非随意生成。此外，推理增强也能减少模型产生不正确答案（幻觉）的概率，提高回答的可靠性。

值得一提的是，推理能力的提升与**模型解释性**息息相关。如果模型能输出中间推理过程，人类便可以检查其思路是否正确，从而增加信任度。这对在高风险领域应用AI（如医疗、法律）尤为重要。因此，可以预见未来的大模型在回答时可能自带解释或依据引用，使其行为更透明可验。

**实时学习与知识更新**

当前的大语言模型大多属于离线训练范式：模型训练完毕后参数固定，对训练后出现的新知识、新事实无能为力。这就导致所谓*“知识截止”（Knowledge Cutoff）*问题。例如，ChatGPT的基础模型GPT-3.5其知识截止于2021年，因而对之后发生的事件一无所知，需要通过额外检索才能回答。这种静态知识库模式显然无法满足很多实时需求，因此**让模型具备实时更新知识的能力**成为重要研究方向。

一种直接思路是在模型架构中融入外部检索，即**检索增强生成（RAG）**技术[[61]](#footnote-61) 。当模型收到查询时，先通过检索模块从知识库中提取相关最新信息，再将信息与查询一同送入模型生成答案。这样，模型的输出就能够反映训练后发生的新情况[[62]](#footnote-62) 。例如，问“今年诺贝尔奖得主是谁”，模型可以实时搜索新闻，然后基于检索结果作答，而不用受限于旧知识。微软Bing整合GPT-4的方案正是类似思路，通过联网搜索赋予模型实时问答能力。实践证明，RAG能够有效降低模型幻觉率，并显著扩展信息的时效性[[63]](#footnote-63) 。Forrester的报告也指出，企业在规划部署生成式AI时必须考虑训练数据质量和模型偏差，以及如何缓解模型幻觉等问题——融入检索无疑是解决之道之一。[[64]](#footnote-64)

除了检索，另一方向是**持续训练或增量学习**。即让模型在部署阶段继续接受新数据的微调更新，而不必等待下一个大版本。例如，研究者尝试让模型定期吸收新发布的维基百科、新闻等，以小步更新参数。还有一些探索利用在线学习算法，使模型边用边学。然而，大型Transformer模型的在线更新面临灾难性遗忘、分布偏移等挑战，目前效果有限。此外还有混合方法：通过插件或API方式，让模型调用专门维护的最新知识库（如每日财经数据库等），作为回答依据。综上，未来可能出现**“随时可学习”**的大语言模型：基本能力由大规模预训练给出，但同时连接着动态更新的知识源，实现静态能力与动态知识的结合。

实时更新能力还关联到**个性化定制**。如果模型能持续学习，那么它可以逐步学习某个特定用户的偏好与知识背景，提供量身定制的回答。这对于私人助理类应用将是巨大优势。当然，这涉及隐私与安全，需要妥善处理用户数据。此外，模型持续学习还须警惕**错误传播**：如果用户提供了错误信息而模型学了进去，可能会影响后续表现。因此实时学习技术需搭配严格的验证和过滤机制。

总的来说，让大语言模型跳出“一次训练成品”的静态模式，迈向“持续学习演化”的动态模式，将是通往更智能AI的重要一步。MIT科技评论的一篇报道曾探讨了一种从MIT提出的方法，使语言模型能够在推理同时不断调整自身参数，实现持续学习[[65]](#footnote-65) 。虽然此类研究尚处起步，但其愿景符合人工智能长期追求的自适应能力。可以预见，未来成熟的大语言模型将如同一直在线的专家，不断吸收新知、纠正自身误差，始终保持其知识库的时效性与准确性。

### **2.3.2 潜在技术瓶颈分析**

尽管大语言模型前景光明，但当前技术上仍存在诸多瓶颈和挑战，有待深入研究和攻克。以下针对**模型幻觉、能耗成本、训练数据质量、推理效率、社会伦理**五大问题进行分析。

**模型幻觉与准确性问题**

*“幻觉”（Hallucination）*指模型生成了看似合理但实际上错误或凭空捏造的内容。这是大语言模型普遍存在的顽疾之一。举例来说，模型有时会编造不存在的引文、虚构事实，或者在不确定时给出确切但错误的回答。幻觉产生的根源在于语言模型基于统计相关性生成文本，并不真正“理解”事实真相，也无法像人一样具备何时不懂就不乱说的自知之明[[66]](#footnote-66) 。Forrester研究团队在2024年报告中特别提醒企业注意模型幻觉风险，指出训练数据质量欠佳和偏见会放大不准确输出的频率[[67]](#footnote-67) 。麦肯锡的一项调查也发现，在采用生成式AI的早期阶段，许多公司尚未有效缓解模型输出不准确这一最相关的风险[[68]](#footnote-68) 。

幻觉问题带来了信任和安全层面的隐忧。例如，在医疗领域，若AI助手胡乱编造诊断依据，后果不堪设想；在资讯领域，模型可能传播谣言或错误信息。为降低幻觉，大模型研发者做出了多种努力。一是**人类反馈对齐**（RLHF）：通过人工标注，让模型学会拒绝回答不知道的问题或在不确定时给出保留态度，从而减少胡编乱造的冲动。OpenAI的ChatGPT显然在这方面较原始GPT-3有所改善——它更常见地在无法确定时回应“不确定”或建议寻求专业帮助。然而RLHF并非万灵药，模型仍可能在知识盲区自信地给出谬误。二是前述**检索增强**策略，通过让模型查资料或验证，其回答可以引经据典、依据事实[[69]](#footnote-69) 。实践证明，有了检索模块辅助，幻觉率明显下降，因为模型不必凭记忆硬撑，可以引用权威来源。不过检索的有效性取决于查询和来源质量，也无法覆盖所有场景。三是模型内部引入**一致性检查**，例如生成多个答案取交集（self-consistency）或者在回答前让模型自己验证一遍逻辑。这类似于数学计算中双重核对，能一定程度筛除前后不一致或明显错误的输出。DeepMind等研究人员还提出利用外部逻辑推理模块对模型答案进行审核，以发现其中的矛盾。总的来看，幻觉问题远未解决，它是大语言模型可靠落地的最大障碍之一。

学术界也在探索从根本上提高模型准确性的途径，包括改进训练目标使其更加关注事实匹配，以及在预训练语料中注入更多高质量知识源等。另外，提升模型对自身不确定性的评估能力也是关键，让模型“知道自己不知道”。如前所述，强化推理能力和引入解释机制都有助于此——模型在回答时给出依据出处[[70]](#footnote-70) 、展示推理链条，方便用户自行核实。这种透明度可提高用户对模型输出的信任，同时也反过来促进模型减少随意编造，因为其过程需要经得起检查。

**能耗成本与算力瓶颈**

大语言模型的训练和部署都极为耗费计算资源和能源。一方面，训练一套数亿参数的模型，需要海量的浮点计算。以GPT-3为例，其训练耗电量估计约为1287兆瓦时（相当于502吨二氧化碳排放），相当于一辆汽油车连续行驶一年的碳排放[[71]](#footnote-71) 。GPT-3训练过程中使用了数千颗GPU协同计算数周时间，经济成本在数百万美元量级[[72]](#footnote-72) 。如此庞大的能源消耗引发了对AI碳足迹的担忧——如果未来模型继续扩大，这种资源开销将呈指数增长，带来环境不可持续性的问题[[73]](#footnote-73) 。另一方面，**推理阶段**（模型部署供用户使用）同样需要大量算力，因为每次生成都要进行深度神经网络运算。Wharton商学院的分析指出，模型推理目前占AI总能耗的约60%，并且随着用户请求的激增而快速累积[[74]](#footnote-74) 。据估计，ChatGPT每回答一个查询所需能量约为普通谷歌搜索的100倍[[75]](#footnote-75) 。有研究进一步测算，如果ChatGPT全年处理750亿次查询，其耗电量约2.3亿千瓦时，足够充满超过300万部电动汽车[[76]](#footnote-76) 。由此可见，大规模部署大语言模型可能对电网和环境带来不小压力。

算力和能耗瓶颈在一定程度上限制了大语言模型的发展和应用普及。并非所有企业或研究机构都负担得起训练一个GPT-3级别模型的费用与能耗；在应用侧，高昂的计算需求也使模型服务成本居高不下。Forrester报告指出，创建和维护LLM需要显著的数据与基础设施投入，这使许多企业难以自行训练竞争性模型[[77]](#footnote-77) 。因此，提升能效、降低成本成为大模型研究的重要课题。例如，业界正积极采用**专用AI加速芯片**（如谷歌TPU、英伟达A100/H100等）来提高每瓦特算力，从硬件上优化能耗比。同时，软件层面的优化如**混合精度训练**、**张量并行和流水并行**、**高效算子实现**（如FlashAttention）也在减少不必要的计算浪费。据密歇根大学一项研究，优化算法可使训练在同样时间内节省多达30%的能耗[[78]](#footnote-78) 。此外，模型结构优化（如稀疏激活、低秩分解）和**模型压缩**（蒸馏、小模型协作等）也能在推理阶段降低计算开销，从而以更少资源服务更多请求。Mistral 7B采用的Grouped-Query Attention和滑动窗口注意力，即是一种通过限制注意力计算范围来降低复杂度的方法[[79]](#footnote-79) 。

长期来看，算力瓶颈可能通过“更智能的高效算法”与“更强大的计算硬件”双管齐下得到缓解。例如，发展光子计算、量子计算等新技术，或者构建大规模分布式计算网络以分摊训练任务。不过，在可预见的未来，**性价比更高的大模型**将更受青睐：与其盲目扩大参数，不如提高利用率，用单位算力榨取更多AI智能。这既是经济驱动也是环保需要。决策者也开始关注AI的能耗问题，提出绿色AI的理念，要求在追求模型性能的同时衡量其环境成本。可以预见，能耗和算力方面的改进，将决定大语言模型能以多快速度、更大规模走入现实应用。

**训练数据质量与局限**

大语言模型之所以能掌握广博知识，主要仰赖于海量训练数据的喂养。然而，“数据即养料”，其质量高低直接影响模型素质。目前训练大型模型的数据多来自互联网爬取的公开资源，如维基百科、新闻文章、网页文本等。这些数据存在一些内在的问题和局限：

首先，互联网文本良莠不齐，包含大量不准确、不可靠甚至有害的信息。模型在缺乏分辨的情况下机械地学习这些内容，可能继承偏见和错误。比如，若训练语料中某领域谣言频出，模型可能将其当作事实产出。这也是为何大模型常被批评带有种族、性别偏见的原因——训练数据里的历史偏见在模型中被放大重现[[80]](#footnote-80) 。 Forrester报告特别提到模型训练涉及版权材料和数据偏见的问题，提醒企业关注训练数据质量[[81]](#footnote-81) 。 改善数据质量，避免“垃圾进，垃圾出”，是模型性能提升的前提。近期业界兴起“**数据中心主义**”观点，强调通过精选和改进训练数据来提升AI效果，而非一味依赖更复杂模型[[82]](#footnote-82) 。

其次，大模型对训练语料存在依赖和**覆盖盲区**。即使给定海量数据，也有模型学不到的知识，因为某些领域的数据可能很稀少或几乎没有。例如在前沿科学、低资源语言、私有知识库等方面，公开数据无法满足需求。这导致模型在这些领域表现不佳。一个典型例子是法律和医学文本风格独特且专业术语繁多，通用网络语料里相对有限，因而模型在法律咨询、医学问诊等任务上经常出错或不够专业。为此，业界开始构建领域专门的数据集对模型进行补充微调，例如大量医学论文、法律条文等，从而弥补通用预训练的不足。然而，专用数据获取本身不易，往往涉及版权或隐私问题。此外，涉及事实更新的领域（如年度统计数据、当代人物动态等），除非频繁再训练，否则模型知识会滞后甚至过时。

再次，训练数据规模总有极限。随着模型参数数以千亿计增长，所需训练文本近乎天量。目前互联网上高质量、独立的信息文本储量是有限的。OpenAI曾指出，GPT-3基本已经把人类写作的公共文本“读”了个遍，再想在数据量上数量级提升很困难。这意味着未来大模型继续扩大，很可能要依赖**合成数据**（synthetic data）来扩充训练集，例如模型自身生成的新句子、机器翻译的平行语料等。然而，使用模型输出再训练模型可能导致回音室效应，模型逐渐丧失创造性并放大自身错误。这是需要警惕的恶性循环：当网络上充斥模型生成内容，模型又以此为训练素材，会发生何种后果目前尚不明晰。此外还有版权法律层面的挑战——模型训练数据往往未经内容作者授权，由此引发的版权争议开始出现，未来监管可能要求透明公开训练来源，甚至对侵权数据做剔除处理。这些都会对大模型训练提出新的约束。

为应对数据质量瓶颈，未来有几个可能的方向：一是**更精细的数据过滤与标注**。在预处理阶段通过AI和人协作，将明显低质或有毒的数据剔除，提高语料纯度。二是**数据多样性提升**，增加来自不同文化、不同语言、不同观点的数据比例，减轻模型偏见。三是打造**可控数据集**，明确标注各段文本的来源、可信度、时效等，让模型在训练中就学会区分权威信息和普通信息。Gartner提出的数据中心AI理念正是希望通过强化数据侧的投入，来驱动模型性能改进[[83]](#footnote-83) 。 另外，**小样本学习**能力的提高也可以缓解对巨量数据的需求：如果模型能从少量新数据中泛化习得知识，就不必为每项新技能都准备海量训练集。总之，在模型规模红利渐趋饱和后，数据质量将成为决定模型上限的关键因素之一，未来对大模型的数据构建会朝更加**高质、专业、动态**的方向发展。

**推理效率与延迟**

推理效率指模型在生成响应时的速度和资源占用情况。这一指标直接影响用户体验和部署成本。目前大语言模型在推理阶段仍存在**速度较慢、延迟较高**的问题。原因在于Transformer模型每产生一个词都需要执行一次上百亿参数的矩阵计算，生成一段较长回答可能涉及成千上万次计算循环。即使借助GPU并行，响应复杂问题往往需要数秒到数十秒不等，对实时交互来说仍嫌缓慢。此外，由于模型规模庞大，推理时显存占用极高，一台普通机器难以独立承载完整模型，通常需要分布式部署，这进一步加大了调用开销。

推理效率低下导致两个直接后果：用户等待时间长和单位请求成本高。这在商业应用中是痛点。例如客户咨询机器人若滞后良久才答复，用户体验会大打折扣；同时每次交互后台烧掉的GPU算力费用也不容小觑。为此，各大公司都在研究提高推理效率的方法：

* **模型压缩**：包括参数量化（如将模型权重从FP16降低为INT8或INT4表示，可大幅减少内存和计算）、剪枝（移除冗余连接）、知识蒸馏（训练小模型模仿大模型输出）等。Meta的研究表明，适度量化的模型几乎不损失性能却将推理成本降低一半以上，这对于在移动设备等边缘环境运行大模型尤其重要。
* **架构改进**：开发更高效的注意力机制和网络结构。例如使用稀疏注意力让每个token只关注有限邻域而非全局，从而将计算复杂度从平方级降至线性级。先前提到的Mistral 7B应用了滑动窗口注意力，使每一层仅关注固定窗口长度的上下文[[84]](#footnote-84) 。还有一些Transformer替代架构（如Performer, Reformer等）尝试用近似方法加速注意力计算。这些改进可以在保持模型质量的同时，大幅提升推理速度。
* **批量并行和流水线**：在推理服务端，通过并行处理多个请求或对长输出采用流水线生成，最大化硬件利用率。例如OpenAI的服务器可在一张GPU上并行执行多个用户的ChatGPT请求，以减少空闲等待。还有推理引擎如vLLM通过优化缓存管理，实现对大批量请求高效复用计算结果，从而提高整体吞吐量[[85]](#footnote-85) 。
* **专用硬件**：硬件进步也是关键因素。最新的AI芯片针对Transformer运算做了大量优化，如更大显存、更高内存带宽、混合精度支持等，使得同样运算在新GPU/TPU上可比旧设备快数倍。未来FPGA、ASIC等定制加速器甚至光学计算设备，或许将进一步提高推理功耗比，让部署成本降低到可以大规模扩张的地步。

随着上述技术的发展，我们有理由期待大语言模型的响应速度接近实时。例如，目前一些优化后的中等模型已经能在CPU上每秒生成数十个词，这意味着对话级别的应答可以在不到1秒内产出。在高性能GPU上，大模型每秒生成上百词也已实现。不久的将来，通过软硬件协同优化，即使是千亿级参数模型也可能达到毫秒级延迟，为交互式应用铺平道路。此外，推理效率的提升还将使**边缘部署**成为可能——未来智能手机、物联网设备等可能直接运行精简版的大语言模型，在本地为用户提供服务，而不必将所有请求发送云端。这将极大拓展大模型的应用范围。

总之，提高推理效率是大语言模型从研究走向产业的必由之路。只有当用户几乎无感知地获得AI响应，且服务提供方能以合理成本支持海量请求时，大模型才能真正融入日常应用。正因如此，我们看到各大公司都投入大量资源改进推理框架和算法。例如，OpenAI推出了针对推理优化的GPT-3.5 Turbo版本，单位成本相比原始模型降低数十倍，使得将其集成到日常办公软件、搜索引擎中成为可能。可以预见，未来大语言模型的创新不仅在于提升上限性能，更在于提升**单位能耗/算力的性能**，实现既“聪明”又“高效”。

## **2.4 本章结论**

自2017年Transformer问世以来，大语言模型经过短短数年已走过从技术起步、规模跃升到应用爆发的非凡历程。OpenAI、Google DeepMind、Meta、Anthropic、Mistral等领军机构推出的一系列模型，推动了内容创作、对话交互、多模态理解、复杂推理等人工智能关键能力的飞速发展。在这一过程中，大语言模型逐渐成为新一代“基础设施”式的技术——即所谓**基础模型（Foundation Model）**，能够作为通用平台支持千行百业的智能应用[[86]](#footnote-86) 。 据麦肯锡预测，生成式AI有望每年为全球经济增加2.6至4.4万亿美元价值，占2021年全球GDP的比例约达15%之巨[[87]](#footnote-87) 。 可以说，大语言模型已不仅是实验室里的模型，而是释放生产力的战略工具。

展望未来，大语言模型的发展将沿着**“更大更通用”**与**“更小更专精”**两条路径同步演进。一方面，研究者将继续探索模型规模和数据极限，融合多模态信息，增强上下文和推理，努力逼近通用人工智能（AGI）的目标；另一方面，模型的实用化要求我们着力提高效率、降低能耗，使其更加经济高效，并针对具体场景进行定制优化。此外，实时学习能力的引入将令模型保持知识新鲜度，长期陪伴用户成长。在这个过程中，我们需要正视并解决模型幻觉、不透明、高算力消耗等**技术鸿沟**，也需未雨绸缪地制定规范来应对偏见、安全、就业冲击等**社会课题**。正如Forrester报告所言，生成式AI浪潮将深刻重塑各行业，既带来前所未有的机遇，也伴随不容忽视的风险[[88]](#footnote-88) 。

## **2.5 与知识服务场景的融合应用可能性、契合度**

（关注六大场景：知识图谱、语义搜索、智能问答、个性化推荐、知识推送、可视化服务）

# **3. 智能体 Agent / 代理型AI**

自人工智能诞生以来，研究者便致力于创造能够模拟人类智能行为的**智能体（Agent）**。传统AI系统往往只针对单一任务，在复杂环境中缺乏自主性。而智能体概念的提出，旨在设计**能够在特定环境中自主行动、完成目标的计算系统**[[89]](#footnote-89) 。经过数十年发展，智能体研究在不同阶段呈现出不同侧重：早期注重感知-行动规则与专家系统，中期探索多智能体协作与强化学习决策，当代则随着**大规模预训练模型**的出现迈入全新阶段[[90]](#footnote-90) 。特别是2022年以来，大语言模型（LLM）驱动的**代理型AI**蓬勃兴起，如OpenAI的ChatGPT掀起了新一轮技术革命。各行业开始将具有对话、推理能力的智能体嵌入业务流程，希望借此实现**人机协作**的新范式，提高自动化和决策水平。

在这一背景下，**图书馆**作为知识服务的重要场所，也面临着与智能体技术融合的机遇与挑战。图书馆传统上提供文献检索、信息咨询、阅读推广等服务，但随着数字资源爆炸和用户需求多元化，单纯依赖人工已难以及时满足读者个性化的信息需求。如何引入AI智能体来提升知识组织和服务效率，正成为“智慧图书馆”建设的前沿课题之一[[91]](#footnote-91) 。例如，智能问答代理可以7×24小时解答读者问题，语义检索代理能够理解自然语言查询检索馆藏，知识图谱代理可以挖掘领域知识辅助科研，等等。不少高校和公共图书馆已开展试点，将**大模型驱动的智能功能**融入馆藏检索、阅读推荐、参考咨询等场景[[92]](#footnote-92) 。然而，面对涌现的代理型AI技术，图书馆界也需深入审视：当前智能体的能力和局限性如何？其在知识服务中的适用性和成熟度达到什么水平？部署维护成本与潜在风险可否接受？这些都是实际应用前必须评估的问题。

为此，本章围绕“智能体的发展演进与图书馆知识服务融合”展开系统研究。我们首先明确智能体的概念内涵，从学术定义和业界理解两方面进行综述，并分类梳理智能体的类型体系。接着，回顾智能体技术的发展历史与关键演进阶段，分析各阶段功能侧重的变化，并列举典型代表模型以刻画技术演化路径。下一部分，聚焦当前智能体领域的演进趋势与面临的技术瓶颈，引用多份权威报告观点解析智能体未来方向、应用挑战、算力限制、评估机制和安全治理等核心议题。之后重点讨论智能体技术在图书馆知识服务六大核心场景中的融合应用前景，每个场景提供国内外实际案例，分析应用效果、存在问题和改进空间。

## **3.1 智能体的定义与分类**

### **3.1.1 智能体的学术定义与理论基础**

在学术界，“智能体”（Intelligent Agent）通常被定义为**一种能感知其所处环境并采取行动以实现既定目标的计算体**。这一定义的经典表述可追溯至人工智能教材和多智能体系统研究的奠基性工作：例如Jennings等人将智能体描述为“一个**嵌入在特定环境中的计算系统**，能够在该环境中自主行动以达成其设计目标”[[93]](#footnote-93) 。同样地，Russell和Norvig在其教材中开宗明义地指出：“人工智能就是关于**智能代理的研究与设计**”，并将“理性Agent”界定为在感知后采取能最大化其绩效措施的行动的实体。这些定义强调了两个关键特征：其一，智能体具有**感知-决策-行动**的闭环，即通过传感器感知环境状态，经过内部决策机制选择行动，再通过执行器影响环境；其二，智能体具有**自主性（Autonomy）**，能够根据自身目标和环境变化自主做出反应，而非完全依赖预设指令或人为干预[blog.csdn.net](https://blog.csdn.net/Code1994/article/details/142780046" \l ":~:text=%E7%89%B9%E7%82%B9)[blog.csdn.net](https://blog.csdn.net/Code1994/article/details/142780046#:~:text=)。换言之，一个合格的智能体应当在目标驱动下表现出理性的行为，无需步步有人指令，而能自主选择适当的动作。

早期的学术研究还为智能体总结了一些常见属性。例如Wooldridge等提出智能体应具备**自主性、社会性、反应性、主动性**等特性：自主性指智能体能在最小人类干预下控制自身行动；社会性指智能体可与其他智能体或人交互；反应性指能实时感知和响应环境变化；主动性则体现智能体能基于内在目标采取主动行为。值得注意的是，不同智能体可能侧重不同属性组合，这也引出了智能体的多种形态和分类（详见2.3节）。

### **3.1.2 智能体的业界定义与应用理解**

在工业界，随着近年**大模型与工具调用技术**的发展，“代理型AI”（Agentic AI）概念得到广泛关注。业界对智能体的定义更聚焦于其实用功能：IBM的官方释义指出，AI智能体是一种**能够通过设计工作流并利用可用工具，代表用户或其他系统自主执行任务的系统或程序**[ibm.com](https://www.ibm.com/cn-zh/think/topics/ai-agents" \l ":~:text=%E4%BB%80%E4%B9%88%E6%98%AF%20AI%20%E6%99%BA%E8%83%BD%E4%BD%93%EF%BC%9F)。这一定义凸显了代理型AI的几个现实特征：首先，它以**完成特定任务/目标**为导向，往往可以充当用户的数字助理或自动化代理；其次，它通常内嵌了**大语言模型（LLM）**作为核心，以强大的自然语言处理和推理能力为基础来理解用户意图、分解任务并决定行动步骤[ibm.com](https://www.ibm.com/cn-zh/think/topics/ai-agents" \l ":~:text=AI%20%E6%99%BA%E8%83%BD%E4%BD%93%E7%9A%84%E5%B7%A5%E4%BD%9C%E5%8E%9F%E7%90%86)[ibm.com](https://www.ibm.com/cn-zh/think/topics/ai-agents#:~:text=AI%20%E6%99%BA%E8%83%BD%E4%BD%93%E4%BC%9A%E6%A0%B9%E6%8D%AE%E5%AE%83%E4%BB%AC%E6%84%9F%E7%9F%A5%E5%88%B0%E7%9A%84%E4%BF%A1%E6%81%AF%E6%9D%A5%E9%87%87%E5%8F%96%E8%A1%8C%E5%8A%A8%E3%80%82%E9%80%9A%E5%B8%B8%EF%BC%8CAI%20%E6%99%BA%E8%83%BD%E4%BD%93%E5%B9%B6%E4%B8%8D%E5%85%B7%E5%A4%87%E8%A7%A3%E5%86%B3%E5%A4%8D%E6%9D%82%E7%9B%AE%E6%A0%87%E5%86%85%E6%89%80%E6%9C%89%E5%AD%90%E4%BB%BB%E5%8A%A1%E6%89%80%E9%9C%80%E7%9A%84%E5%AE%8C%E6%95%B4%E7%9F%A5%E8%AF%86%E5%BA%93%E3%80%82%E4%B8%BA%E8%A7%A3%E5%86%B3%E6%AD%A4%E9%97%AE%E9%A2%98%EF%BC%8CAI%20%E6%99%BA%E8%83%BD%E4%BD%93%E4%BC%9A%E4%BD%BF%E7%94%A8%E8%87%AA%E5%B7%B1%E7%9A%84%E5%8F%AF%E7%94%A8%E5%B7%A5%E5%85%B7%E3%80%82%E8%BF%99%E4%BA%9B%E5%B7%A5%E5%85%B7%E5%8F%AF%E8%83%BD%E5%8C%85%E6%8B%AC%E5%A4%96%E9%83%A8%E6%95%B0%E6%8D%AE%E9%9B%86%E3%80%81Web%20%E6%90%9C%E7%B4%A2%E3%80%81API,%E7%94%9A%E8%87%B3%E6%98%AF%E5%85%B6%E4%BB%96%E6%99%BA%E8%83%BD%E4%BD%93%E3%80%82%E4%BB%8E%E8%BF%99%E4%BA%9B%E5%B7%A5%E5%85%B7%E6%A3%80%E7%B4%A2%E5%88%B0%E7%BC%BA%E5%A4%B1%E7%9A%84%E4%BF%A1%E6%81%AF%E5%90%8E%EF%BC%8C%E6%99%BA%E8%83%BD%E4%BD%93%E4%BE%BF%E5%8F%AF%E6%9B%B4%E6%96%B0%E5%85%B6%E7%9F%A5%E8%AF%86%E5%BA%93%E3%80%82%E8%BF%99%E6%84%8F%E5%91%B3%E7%9D%80%EF%BC%8C%E6%99%BA%E8%83%BD%E4%BD%93%E5%9C%A8%E5%85%B6%E4%B8%AD%E6%AF%8F%E4%B8%80%E6%AD%A5%E5%9D%87%E4%BC%9A%E9%87%8D%E6%96%B0%E8%AF%84%E4%BC%B0%E5%85%B6%E8%A1%8C%E5%8A%A8%E8%AE%A1%E5%88%92%E5%B9%B6%E8%87%AA%E6%88%91%E6%9B%B4%E6%AD%A3%E3%80%82)；再次，它能够**调用外部工具或服务**来扩展自身能力，例如访问数据库、执行联网搜索、调用其他API，甚至调度其他智能体，从而弥补单纯依赖内部训练知识的不足[ibm.com](https://www.ibm.com/cn-zh/think/topics/ai-agents" \l ":~:text=%E4%BD%BF%E7%94%A8%E5%8F%AF%E7%94%A8%E5%B7%A5%E5%85%B7%E8%BF%9B%E8%A1%8C%E6%8E%A8%E7%90%86)。业界常用“大语言模型智能体”来指代此类系统[ibm.com](https://www.ibm.com/cn-zh/think/topics/ai-agents" \l ":~:text=AI%20%E6%99%BA%E8%83%BD%E4%BD%93%E7%9A%84%E5%B7%A5%E4%BD%9C%E5%8E%9F%E7%90%86)。通俗而言，业界眼中的智能体更像是一个具备一定“自主工作流”能力的AI助手，可以根据用户高层指令自主规划并完成多步任务，而不仅仅是被动回答问题的聊天机器人。

需要指出，**业界对“Agent”一词的使用较为宽松**，目前尚未有统一标准[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00" \l ":~:text=%E5%A6%82%E4%BD%95%E5%AE%9A%E4%B9%89%E2%80%9C%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%93%E2%80%9D%EF%BC%9F)。许多厂商宣传的“AI Agent”能力参差不齐，一些只是包装了预定义脚本的聊天助手即自称代理智能。这种“Agent Washing”（代理混淆）现象引发了 Gartner 等机构的警惕：据统计，当前号称提供代理型AI产品的厂商有上千家，但真正具有自主 Agent 能力的可能只有约130家[reuters.com](https://www.reuters.com/business/over-40-agentic-ai-projects-will-be-scrapped-by-2027-gartner-says-2025-06-25/" \l ":~:text=Many%20vendors%20are%20engaging%20in,agentic%20AI%20vendors%20are%20real)。因此，业界更加关注智能体实际**自主完成复杂目标**的能力，而非名称上的噱头。总体来说，业界定义扩展了学术概念，强调智能体作为**面向任务的自治系统**，能够与环境（软件环境或物理环境）交互并连续地完成用户委派的复杂工作[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00" \l ":~:text=%E4%BA%8B%E5%AE%9E%E4%B8%8A%EF%BC%8C%E4%B8%9A%E7%95%8C%E5%AF%B9%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%93%E7%9A%84%E7%A0%94%E7%A9%B6%E8%BF%98%E5%A4%84%E4%BA%8E%E5%88%9D%E6%9C%9F%E9%98%B6%E6%AE%B5%EF%BC%8C%E8%BF%98%E6%B2%A1%E6%9C%89%E5%AF%B9%E5%85%B6%E8%BF%9B%E8%A1%8C%E6%98%8E%E7%A1%AE%E7%9A%84%E5%AE%9A%E4%B9%89%E3%80%82%E2%80%9C%E7%AE%80%E5%8D%95%E5%9C%B0%E7%90%86%E8%A7%A3%EF%BC%8C%E5%AE%83%E4%BB%AC%E6%9C%AC%E8%B4%A8%E4%B8%8A%E5%B0%B1%E6%98%AF%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E6%A8%A1%E5%9E%8B%E5%92%8C%E7%AE%97%E6%B3%95%EF%BC%8C%E5%8F%AF%E4%BB%A5%E5%9C%A8%E5%8A%A8%E6%80%81%E4%B8%96%E7%95%8C%E4%B8%AD%E8%87%AA%E4%B8%BB%E5%81%9A%E5%87%BA%E5%86%B3%E7%AD%96%E3%80%82%E2%80%9D%E8%8B%B1%E4%BC%9F%E8%BE%BE%E9%AB%98%20%E7%BA%A7%E7%A0%94%E7%A9%B6%E7%A7%91%E5%AD%A6%E5%AE%B6%E3%80%81%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%93%E9%A1%B9%E7%9B%AE%E8%B4%9F%E8%B4%A3%E4%BA%BA%20Jim%20Fan%20%E8%A1%A8%E7%A4%BA%E3%80%82)[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00#:~:text=%E2%80%9C%E5%B9%BF%E4%B9%89%E4%B8%8A%E8%AE%B2%EF%BC%8C%E6%9C%89%E4%B8%A4%E7%B1%BB%E4%B8%8D%E5%90%8C%E7%9A%84%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%93%EF%BC%9A%E8%BD%AF%E4%BB%B6%E6%99%BA%E8%83%BD%E4%BD%93%EF%BC%88Software%20Agents%EF%BC%89%E5%92%8C%E5%85%B7%E8%BA%AB%E6%99%BA%E8%83%BD%E4%BD%93%EF%BC%88Embodied%20Agents%EF%BC%89%E3%80%82%E2%80%9DJim%20Fan%20%E8%AF%B4%E9%81%93%EF%BC%8C%E2%80%9C%E8%BD%AF%E4%BB%B6%E6%99%BA%E8%83%BD%E4%BD%93%E5%9C%A8%E7%94%B5%E8%84%91%E6%88%96%E6%89%8B%E6%9C%BA%E4%B8%8A%E8%BF%90%E8%A1%8C%E5%B9%B6%E4%BD%BF%E7%94%A8%E5%BA%94%E7%94%A8%E7%A8%8B%E5%BA%8F%EF%BC%8C%E8%BF%99%E4%BA%9B%E6%99%BA%E8%83%BD%E4%BD%93%E5%AF%B9%E5%8A%9E%E5%85%AC%E5%AE%A4%E5%B7%A5%E4%BD%9C%E3%80%81%E5%8F%91%E9%80%81%E7%94%B5%E5%AD%90%E9%82%AE%E4%BB%B6%E6%88%96%E5%AE%8C%E6%88%90%E8%BF%99%E4%B8%80%E7%B3%BB%E5%88%97%E6%B4%BB%E5%8A%A8%E9%9D%9E%E5%B8%B8%E6%9C%89%E7%94%A8%E3%80%82%E2%80%9D)。比如，一个电商智能体可以根据用户偏好自动在多家网站比价购物并下单；又如企业RPA智能体可以跨多个业务系统执行审批流程。业界期待的智能体是可以“替人干活”的AI员工，其智能程度超越传统固定流程的机器人。

### **3.1.3 智能体的常见分类方法**

由于智能体概念涵盖范围广泛，从软件机器人到物理机器人，各类系统差异巨大，研究者和从业者提出了多种分类维度来刻画不同智能体类型。以下根据文献梳理几种常见的分类方法：

**（1）按行为决策模式分类：** 这是AI理论教材中经典的划分方式，将智能体分为以下几类[blog.csdn.net](https://blog.csdn.net/Code1994/article/details/142780046#:~:text=)[blog.csdn.net](https://blog.csdn.net/Code1994/article/details/142780046#:~:text=%23%203.%E5%9F%BA%E4%BA%8E%E7%9B%AE%E6%A0%87%E7%9A%84%E4%BB%A3%E7%90%86%EF%BC%88Goal)：

* **简单反射型智能体（Simple Reflex Agent）：** 依据当前感知到的环境状态，直接套用预设的条件-动作规则作出反应，不考虑过去和未来的影响。它只对**当前**刺激作出反射式动作。例如温控器根据当前温度高低立刻开关空调。该类智能体结构最简单，但**缺乏学习和预见能力**。
* **基于模型的反射智能体（Model-Based Reflex Agent）：** 在简单反射基础上，引入对环境的**内部模型**。智能体维护对当前环境的内部状态表示（依据历史感知更新），并以此模型判断当前情境，再按规则行动[blog.csdn.net](https://blog.csdn.net/Code1994/article/details/142780046#:~:text=%23%202.%E5%9F%BA%E4%BA%8E%E6%A8%A1%E5%9E%8B%E7%9A%84%E5%8F%8D%E5%B0%84%E4%BB%A3%E7%90%86%EF%BC%88Model)。这样可在**部分可观察环境**中工作，考虑一定的历史信息。
* **基于目标的智能体（Goal-Based Agent）：** 此类智能体除了环境模型外，还明确**目标**信息。它会根据目标选择能够达成目标的行动序列，涉及搜索和规划。例如路径规划机器人根据目标位置搜索路线[blog.csdn.net](https://blog.csdn.net/Code1994/article/details/142780046" \l ":~:text=%23%203.%E5%9F%BA%E4%BA%8E%E7%9B%AE%E6%A0%87%E7%9A%84%E4%BB%A3%E7%90%86%EF%BC%88Goal)。相比反射型，目标导向代理具有**目的性**，能处理更复杂的问题。
* **基于效用的智能体（Utility-Based Agent）：** 进一步地，引入**效用函数**来评估每种状态/行动的价值。智能体不仅有目标，还追求最大化预期效用，即在可能有多种方案都能实现目标时选择**最优策略**[blog.csdn.net](https://blog.csdn.net/Code1994/article/details/142780046" \l ":~:text=%23%204.%E5%9F%BA%E4%BA%8E%E6%95%88%E7%94%A8%E7%9A%84%E4%BB%A3%E7%90%86%EF%BC%88Utility)。这需要考虑不确定性和概率，常用于决策复杂、需要权衡取舍的情境。
* **学习型智能体（Learning Agent）：** 具备学习能力的智能体，可以从经验中**改进自身行为策略**[blog.csdn.net](https://blog.csdn.net/Code1994/article/details/142780046" \l ":~:text=)。典型结构包括学习模块、绩效模块等，通过不断试验和反馈优化决策（如强化学习算法）。学习型代理能适应动态环境、逐步提升性能。

上述分类体现了智能体从**简单硬编码反应**到**复杂目标优化**再到**自我学习**的渐进关系。在实际系统中，常常将多种机制结合形成**混合智能体（Hybrid Agent）**[blog.csdn.net](https://blog.csdn.net/Code1994/article/details/142780046#:~:text=)。例如自主驾驶系统会融合反射规则（紧急避障）、模型预测（轨迹规划）、学习优化（策略改进）等多个层次。

**（2）按自主程度分类：** 根据智能体在任务执行中所需的人为干预程度，可分为：

* **自主型智能体：** 拥有高度自主决策和行动能力，几乎**无需人类干预**即可完成任务[blog.csdn.net](https://blog.csdn.net/Code1994/article/details/142780046" \l ":~:text=)。如无人驾驶汽车在行驶过程中自主感知环境、规划路线、控制车辆。自主Agent适合太空探测、深海作业等人类难以介入的场景[blog.csdn.net](https://blog.csdn.net/Code1994/article/details/142780046#:~:text=%E8%87%AA%E5%8A%A8%E9%A9%BE%E9%A9%B6%E6%B1%BD%E8%BD%A6%E3%80%81%E8%87%AA%E4%B8%BB%E6%9C%BA%E5%99%A8%E4%BA%BA%E7%AD%89%E3%80%82%E4%BE%8B%E5%A6%82%EF%BC%8C%E8%87%AA%E5%8A%A8%E9%A9%BE%E9%A9%B6%E6%B1%BD%E8%BD%A6%E5%8F%AF%E4%BB%A5%E5%9C%A8%E6%B2%A1%E6%9C%89%E4%BA%BA%E7%B1%BB%E9%A9%BE%E9%A9%B6%E5%91%98%E7%9A%84%E6%83%85%E5%86%B5%E4%B8%8B%EF%BC%8C%E8%87%AA%E4%B8%BB%E5%9C%B0%E6%84%9F%E7%9F%A5%E7%8E%AF%E5%A2%83%E3%80%81%E5%81%9A%E5%87%BA%E5%86%B3%E7%AD%96%E5%92%8C%E6%8E%A7%E5%88%B6%E8%BD%A6%E8%BE%86%E8%A1%8C%E9%A9%B6%E3%80%82)。
* **半自主型智能体：** 具有一定自主性，但**关键时刻需人类监督/指示**[blog.csdn.net](https://blog.csdn.net/Code1994/article/details/142780046" \l ":~:text=)。它可按预定策略自动运行，但遇到复杂或未见过的情况会请求人工决策。例如智能客服机器人平常自动回答常见咨询，但遇到棘手问题时转接人工客服。半自主Agent适合需要人与AI协同工作的领域，可在保证安全前提下提升效率[blog.csdn.net](https://blog.csdn.net/Code1994/article/details/142780046#:~:text=%E5%9C%A8%E4%B8%80%E5%AE%9A%E7%A8%8B%E5%BA%A6%E4%B8%8A%E5%85%B7%E6%9C%89%E8%87%AA%E4%B8%BB%E5%86%B3%E7%AD%96%E5%92%8C%E8%A1%8C%E5%8A%A8%E8%83%BD%E5%8A%9B%EF%BC%8C%E4%BD%86%E4%BB%8D%E9%9C%80%E8%A6%81%E4%BA%BA%E7%B1%BB%E7%9A%84%E7%9B%91%E7%9D%A3%E5%92%8C%E6%8C%87%E5%AF%BC%E3%80%82%E5%AE%83%E4%BB%AC%E5%8F%AF%E4%BB%A5%E6%A0%B9%E6%8D%AE%E9%A2%84%E8%AE%BE%E7%9A%84%E8%A7%84%E5%88%99%E5%92%8C%E7%AD%96%E7%95%A5%E8%BF%9B%E8%A1%8C%E6%93%8D%E4%BD%9C%EF%BC%8C%E4%BD%86%E5%9C%A8%E9%81%87%E5%88%B0%E5%A4%8D%E6%9D%82%E6%83%85%E5%86%B5%E6%88%96%E4%B8%8D%E7%A1%AE%E5%AE%9A%E5%9B%A0%E7%B4%A0%E6%97%B6%EF%BC%8C%E9%9C%80%E8%A6%81%E4%BA%BA%E7%B1%BB%E7%9A%84%E5%B9%B2%E9%A2%84%E5%92%8C%E5%86%B3%E7%AD%96%E3%80%82)。
* **非自主型智能体：** 严格说这类不应称为“智能体”，因为完全由人控制，没有自主能力[blog.csdn.net](https://blog.csdn.net/Code1994/article/details/142780046" \l ":~:text=)。如传统软件工具、机械设备，必须有人下指令才能运行。它们更多是被动工具，不具备智能行为。

显然，全自主智能体是终极追求，但目前许多应用仍采用半自主模式以平衡AI效率与人工把控。

**（3）按存在形态分类：** 即按照智能体所处的载体与环境，可划分为：

* **软件智能体（Software Agent）：** 在计算机、手机等数字环境中运行的纯软件代理[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00#:~:text=%E2%80%9C%E5%B9%BF%E4%B9%89%E4%B8%8A%E8%AE%B2%EF%BC%8C%E6%9C%89%E4%B8%A4%E7%B1%BB%E4%B8%8D%E5%90%8C%E7%9A%84%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%93%EF%BC%9A%E8%BD%AF%E4%BB%B6%E6%99%BA%E8%83%BD%E4%BD%93%EF%BC%88Software%20Agents%EF%BC%89%E5%92%8C%E5%85%B7%E8%BA%AB%E6%99%BA%E8%83%BD%E4%BD%93%EF%BC%88Embodied%20Agents%EF%BC%89%E3%80%82%E2%80%9DJim%20Fan%20%E8%AF%B4%E9%81%93%EF%BC%8C%E2%80%9C%E8%BD%AF%E4%BB%B6%E6%99%BA%E8%83%BD%E4%BD%93%E5%9C%A8%E7%94%B5%E8%84%91%E6%88%96%E6%89%8B%E6%9C%BA%E4%B8%8A%E8%BF%90%E8%A1%8C%E5%B9%B6%E4%BD%BF%E7%94%A8%E5%BA%94%E7%94%A8%E7%A8%8B%E5%BA%8F%EF%BC%8C%E8%BF%99%E4%BA%9B%E6%99%BA%E8%83%BD%E4%BD%93%E5%AF%B9%E5%8A%9E%E5%85%AC%E5%AE%A4%E5%B7%A5%E4%BD%9C%E3%80%81%E5%8F%91%E9%80%81%E7%94%B5%E5%AD%90%E9%82%AE%E4%BB%B6%E6%88%96%E5%AE%8C%E6%88%90%E8%BF%99%E4%B8%80%E7%B3%BB%E5%88%97%E6%B4%BB%E5%8A%A8%E9%9D%9E%E5%B8%B8%E6%9C%89%E7%94%A8%E3%80%82%E2%80%9D)。它们以应用程序形式存在，操作其他软件界面或后台系统，适合处理办公、信息管理等任务。如电子邮件助理可以自动读取邮件内容并安排日程。这类智能体在**虚拟环境**中行动，可大量用于企业流程自动化、用户助理等。
* **具身智能体（Embodied Agent）：** 存在于物理世界，具有某种**实体形态**（机器人、无人机、虚拟角色等）[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00#:~:text=%E2%80%9C%E5%B9%BF%E4%B9%89%E4%B8%8A%E8%AE%B2%EF%BC%8C%E6%9C%89%E4%B8%A4%E7%B1%BB%E4%B8%8D%E5%90%8C%E7%9A%84%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%93%EF%BC%9A%E8%BD%AF%E4%BB%B6%E6%99%BA%E8%83%BD%E4%BD%93%EF%BC%88Software%20Agents%EF%BC%89%E5%92%8C%E5%85%B7%E8%BA%AB%E6%99%BA%E8%83%BD%E4%BD%93%EF%BC%88Embodied%20Agents%EF%BC%89%E3%80%82%E2%80%9DJim%20Fan%20%E8%AF%B4%E9%81%93%EF%BC%8C%E2%80%9C%E8%BD%AF%E4%BB%B6%E6%99%BA%E8%83%BD%E4%BD%93%E5%9C%A8%E7%94%B5%E8%84%91%E6%88%96%E6%89%8B%E6%9C%BA%E4%B8%8A%E8%BF%90%E8%A1%8C%E5%B9%B6%E4%BD%BF%E7%94%A8%E5%BA%94%E7%94%A8%E7%A8%8B%E5%BA%8F%EF%BC%8C%E8%BF%99%E4%BA%9B%E6%99%BA%E8%83%BD%E4%BD%93%E5%AF%B9%E5%8A%9E%E5%85%AC%E5%AE%A4%E5%B7%A5%E4%BD%9C%E3%80%81%E5%8F%91%E9%80%81%E7%94%B5%E5%AD%90%E9%82%AE%E4%BB%B6%E6%88%96%E5%AE%8C%E6%88%90%E8%BF%99%E4%B8%80%E7%B3%BB%E5%88%97%E6%B4%BB%E5%8A%A8%E9%9D%9E%E5%B8%B8%E6%9C%89%E7%94%A8%E3%80%82%E2%80%9D)。它们不仅处理信息，还能通过机械机构影响现实环境。具身Agent需要解决感知、运动控制、物理交互等复杂问题，例如服务机器人需要导航避障、操作物体。这类智能体应用于工业制造、家庭服务、游戏交互等场景，为用户提供“实体化”的智能服务。

两类智能体在技术上有较大差异：软件代理多关注API调用、GUI操作和数据处理，具身代理还需计算机视觉、运动规划等。因此，一些研究将智能体进一步细分至具体领域（如对话智能体、工业机器人智能体、游戏NPC智能体、教育辅导智能体等）[blog.csdn.net](https://blog.csdn.net/Code1994/article/details/142780046" \l ":~:text=)。这些领域划分更多基于应用功能，但背后依然可归为软件或具身两大类。

需要说明的是，以上分类方法彼此并不冲突，同一智能体可以归属多种类别。例如，自动驾驶汽车是“具身+自主+基于效用/目标”的智能体，而客服聊天机器人是“软件+半自主+学习型”智能体等。分类的意义在于帮助我们理解不同智能体在架构和能力上的差异，以及适用的场景边界。总体来看，**智能体是一类涵盖多种形式的智能系统家族**，它们共享的本质是在一定自主性下感知环境并采取行动，但在智能深度、交互方式、应用领域上各有侧重。

本章从概念和分类角度厘清了智能体的内涵。学术定义强调智能体的感知–行动闭环和自主性，将其视为**理性行为体**来研究；业界定义聚焦智能体的实用价值，强调其**自主执行任务**的能力与工具整合。我们也梳理了智能体的多种分类维度，包括决策方式（从简单反射到效用优化、学习型）、自主程度（自主、半自主、非自主）以及存在形态（软件代理 vs. 具身代理）等。这些分类帮助我们认识，不同智能体系统在结构与功能上可能千差万别。需要注意的是，伴随大模型时代的到来，传统分类出现了交融趋势：例如对话式的大语言模型代理兼具学习能力和一定自主性，能通过工具接口影响外部环境，难以简单归类。这也预示着智能体技术正在融合多学科方法向更高级形态演进。理解智能体的定义与分类，为下一章探讨其发展演进奠定了基础。

## **3.2 智能体的发展起源与关键阶段**

### **3.2.1 概念起源与早期发展**

智能体思想的萌芽可以追溯到20世纪中叶人工智能的兴起。早在1980年代，人工智能先驱明斯基在《心灵社会》等著作中提出了“**Agent**”的概念，认为智能可以被视作许多简单代理相互作用的社会360doc.com。这一观点启发了此后对**分布式人工智能**和**多智能体系统**的研究。1990年代，学术界出现了智能体技术的第一次热潮，一系列**基于规则和知识库**的智能代理系统被提出，用于信息检索、自动交易、故障诊断等领域。这一时期的智能体多是**特化的专家系统**和脚本代理，例如早期的“软(bot)”程序能在网络上充当用户代理执行信息查询等任务。然而，由于当时AI技术局限，这些代理往往缺乏学习适应能力，功能相对简单，智能体概念更像是一种设计范式上的尝试。

一个标志性进展是1995年麻省理工学院的Pattie Maes等提出了**智能体的软件结构**，强调智能体具有**持续运行**、**自治**和**与环境交互**的属性，同时开始探索让多个智能体协同工作解决问题的框架。1990-2000年代，多智能体系统（Multi-Agent System, MAS）成为研究热点，涌现出Contract Net协议、博弈协商等经典成果。这一阶段的智能体更多运用于模拟和决策分析，如交通分配模拟中的车辆智能体、供应链仿真中的企业智能体等。但总体而言，由于智能体缺少强大的感知和学习算法支撑，其应用还局限在**规则明确**或**环境封闭**的场景。

### **3.2.2 第一波智能体浪潮：专用智能体的兴起**

真正让“智能体”引发广泛瞩目的是**2010年代中期**随着机器学习和特别是**强化学习**技术的突破所带来的第一波智能体浪潮。其典型代表之一便是2016年震撼业界的**AlphaGo**。AlphaGo虽然通常被称为围棋AI，但本质上可以视为一个高度专用的智能体：它在围棋这个复杂环境中自主与对手对弈，制定策略和动作，最终达到赢棋这一目标。AlphaGo使用了深度神经网络结合蒙特卡罗树搜索的强化学习算法，具备了相当程度的**自主决策和规划**能力，被视为人工智能的里程碑[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00" \l ":~:text=%E5%8D%8E%E7%9B%9B%E9%A1%BF%E5%A4%A7%E5%AD%A6%E8%AE%A1%E7%AE%97%E6%9C%BA%E7%A7%91%E5%AD%A6%E6%95%99%E6%8E%88%20Chirag%20Shah%20%E8%A1%A8%E7%A4%BA%EF%BC%8C%E2%80%9C%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%93%E2%80%9D%E4%B8%80%E8%AF%8D%E5%85%B6%E5%AE%9E%E5%B7%B2%E7%BB%8F%E5%AD%98%E5%9C%A8%E4%BA%86%E5%BE%88%E5%A4%9A%E5%B9%B4%EF%BC%8C%E4%BD%86%E6%98%AF%E5%85%B6%E5%9C%A8%E4%B8%8D%E5%90%8C%E7%9A%84%E6%97%B6%E9%97%B4%E6%84%8F%E5%91%B3%E7%9D%80%E4%B8%8D%E5%90%8C%E7%9A%84%E4%B8%9C%E8%A5%BF%E3%80%82)。正如Nvidia研究人员Jim Fan所指出的，那一波浪潮中的智能体（如AlphaGo）能够**做出决策并制定策略**，但它们过于“**专一**”，每个智能体只擅长完成单一特定任务[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00" \l ":~:text=%E2%80%9C%E5%B7%B2%E7%BB%8F%E5%87%BA%E7%8E%B0%E4%BA%86%E4%B8%A4%E6%B3%A2%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%93%E7%83%AD%E6%BD%AE%EF%BC%8C%E7%9C%BC%E4%B8%8B%E7%9A%84%E8%BF%99%E4%B8%80%E6%B3%A2%E7%83%AD%E6%BD%AE%E4%B8%BB%E8%A6%81%E5%BD%92%E5%8A%9F%E4%BA%8E%E8%AF%AD%E8%A8%80%E6%A8%A1%E5%9E%8B%E7%9A%84%E7%B9%81%E8%8D%A3%E5%92%8C%20ChatGPT%20%E7%9A%84%E5%85%B4%E8%B5%B7%E3%80%82%E2%80%9DJim%20Fan%20%E6%8C%87%E5%87%BA%EF%BC%8C%E2%80%9C%E4%B8%8A%E4%B8%80%E6%B3%A2%E7%83%AD%E6%BD%AE%E6%98%AF%E5%9C%A8,%E5%B9%B4%EF%BC%8C%E5%BD%93%E6%97%B6%E8%B0%B7%E6%AD%8C%20DeepMind%20%E6%8E%A8%E5%87%BA%E4%BA%86%20AlphaGo%EF%BC%8C%E8%BF%99%E6%98%AF%E4%B8%80%E4%B8%AA%E5%BC%BA%E5%A4%A7%E7%9A%84%E5%9B%B4%E6%A3%8B%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E7%B3%BB%E7%BB%9F%EF%BC%8C%E8%83%BD%E5%A4%9F%E5%81%9A%E5%87%BA%E5%86%B3%E7%AD%96%E5%B9%B6%E5%88%B6%E5%AE%9A%E6%88%98%E7%95%A5%E3%80%82AlphaGo%20%E4%B8%BB%E8%A6%81%E4%BE%9D%E8%B5%96%E4%BA%8E%E5%BC%BA%E5%8C%96%E5%AD%A6%E4%B9%A0%EF%BC%8C%E8%BF%99%E6%98%AF%E4%B8%80%E7%A7%8D%E5%A5%96%E5%8A%B1%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E7%AE%97%E6%B3%95%E5%81%9A%E5%87%BA%E7%90%86%E6%83%B3%E8%A1%8C%E4%B8%BA%E7%9A%84%E6%8A%80%E6%9C%AF%E3%80%82%E2%80%9D)。例如AlphaGo只会下围棋，换个环境就束手无策。这一时期其他引人注目的智能体还包括DeepMind的强化学习算法在雅达利游戏中涌现超人表现，OpenAI Five在电子游戏《Dota2》中击败职业选手等。它们展示了AI智能体在**封闭复杂环境**中的强大决策能力，掀起了关于人工通用智能(AGI)前景的新讨论。

然而，第一波智能体浪潮的局限也很明显：这些系统**高度依赖专门训练**，每个智能体只能处理单一领域的问题，缺乏跨领域通用性[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00" \l ":~:text=%E8%B0%B7%E6%AD%8C%20DeepMind%20%E7%A0%94%E7%A9%B6%E5%89%AF%E6%80%BB%E8%A3%81%20Oriol%20Vinyals,%E8%A1%A8%E7%A4%BA%EF%BC%8C%E2%80%9C%E4%BD%86%E8%BF%99%E4%BA%9B%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%93%E5%A4%AA%E8%BF%87%E4%BA%8E%E2%80%98%E4%B8%93%E4%B8%80%E2%80%99%EF%BC%8C%E6%8D%A2%E5%8F%A5%E8%AF%9D%E8%AF%B4%EF%BC%8C%E8%BF%99%E4%BA%9B%E6%99%BA%E8%83%BD%E4%BD%93%E5%8F%AA%E6%98%AF%E4%B8%BA%E5%AE%8C%E6%88%90%E6%9F%90%E7%A7%8D%E7%89%B9%E5%9C%B0%E4%BB%BB%E5%8A%A1%E8%80%8C%E5%88%9B%E5%BB%BA%E7%9A%84%EF%BC%8C%E6%AF%94%E5%A6%82%20AlphaGo%20%E5%B0%B1%E5%8F%AA%E6%98%AF%E4%BC%9A%E4%B8%8B%E5%9B%B4%E6%A3%8B%E3%80%82%E7%9B%B8%E8%BE%83%E4%B9%8B%E4%B8%8B%EF%BC%8C%E6%96%B0%E4%B8%80%E4%BB%A3%E5%9F%BA%E4%BA%8E%E5%9F%BA%E7%A1%80%E6%A8%A1%E5%9E%8B%E7%9A%84%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%BF%E6%99%BA%E8%83%BD%E4%BD%93%E6%9B%B4%E5%8A%A0%E6%99%AE%E9%81%8D%EF%BC%8C%E5%9B%A0%E4%B8%BA%E5%AE%83%E4%BB%AC%E5%8F%AF%E4%BB%A5%E4%BB%8E%E4%BA%BA%E7%B1%BB%E4%BA%92%E5%8A%A8%E7%9A%84%E4%B8%96%E7%95%8C%E4%B8%AD%E8%BF%9B%E8%A1%8C%E5%AD%A6%E4%B9%A0%E3%80%82%E2%80%9D)。换句话说，这时的“智能体”更类似于“**AI+特定任务**”，并没有形成一个可泛化的智能代理框架。尽管如此，这波浪潮证明了通过学习算法，智能体**在复杂动态环境中达到超越人类**表现是可能的，为更进一步的探索打下基础。AlphaGo等成功案例也刺激了产业界投资自治智能体研发的热情，例如自动驾驶汽车、自动化交易代理等都在此后几年取得长足进展。

### **3.2.3 新一代智能体浪潮：大模型驱动的通用代理**

进入2020年代，人工智能的版图发生了巨大变化——**大规模预训练模型**（尤其是大语言模型）的横空出世，为智能体赋予了前所未有的通用知识和推理能力。由此催生了新一波智能体浪潮，其特点是诞生了一批**通用型、面向广域任务**的代理AI雏形。2022年11月OpenAI推出的**ChatGPT**可以视为这一浪潮的引爆点。ChatGPT本质上是基于GPT-3.5大语言模型的对话智能体，它能接受自然语言指令并连续对话，为用户生成文章、代码、答案等多种内容。首次，公众体验到一个AI助手能够如此**灵活地理解各种问题并给出富有内容的回应**，仿佛一个无所不知的智能代理。同时期，Anthropic公司的**Claude**等对话模型也相继出现，它们注重通过“原则”引导提高安全性，旨在成为更可信赖的AI助手。

与第一波浪潮不同，新一代智能体不是为某单一任务专门训练，而是**依托海量数据训练的基础模型**，然后经过指令微调成为对话或多功能助手。它们具备相当广博的知识和语言能力，能够**迁移**到不同任务场景中。这种通用性使业内开始认为这些对话模型已具有初步代理智能的雏形。正如Chirag Shah教授所说：“人工智能体这个词已经存在很多年，但在不同时间意味着不同的东西。现在这一波智能体热潮主要归功于语言模型的繁荣和ChatGPT的兴起”[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00#:~:text=%E5%8D%8E%E7%9B%9B%E9%A1%BF%E5%A4%A7%E5%AD%A6%E8%AE%A1%E7%AE%97%E6%9C%BA%E7%A7%91%E5%AD%A6%E6%95%99%E6%8E%88%20Chirag%20Shah%20%E8%A1%A8%E7%A4%BA%EF%BC%8C%E2%80%9C%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%93%E2%80%9D%E4%B8%80%E8%AF%8D%E5%85%B6%E5%AE%9E%E5%B7%B2%E7%BB%8F%E5%AD%98%E5%9C%A8%E4%BA%86%E5%BE%88%E5%A4%9A%E5%B9%B4%EF%BC%8C%E4%BD%86%E6%98%AF%E5%85%B6%E5%9C%A8%E4%B8%8D%E5%90%8C%E7%9A%84%E6%97%B6%E9%97%B4%E6%84%8F%E5%91%B3%E7%9D%80%E4%B8%8D%E5%90%8C%E7%9A%84%E4%B8%9C%E8%A5%BF%E3%80%82)。也就是说，当今的代理型AI和过去的定义已不同，**大模型让智能体更加普遍通用**，因为它们从人类大量互动世界中学习到了广泛技能[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00" \l ":~:text=%E8%B0%B7%E6%AD%8C%20DeepMind%20%E7%A0%94%E7%A9%B6%E5%89%AF%E6%80%BB%E8%A3%81%20Oriol%20Vinyals,%E8%A1%A8%E7%A4%BA%EF%BC%8C%E2%80%9C%E4%BD%86%E8%BF%99%E4%BA%9B%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%93%E5%A4%AA%E8%BF%87%E4%BA%8E%E2%80%98%E4%B8%93%E4%B8%80%E2%80%99%EF%BC%8C%E6%8D%A2%E5%8F%A5%E8%AF%9D%E8%AF%B4%EF%BC%8C%E8%BF%99%E4%BA%9B%E6%99%BA%E8%83%BD%E4%BD%93%E5%8F%AA%E6%98%AF%E4%B8%BA%E5%AE%8C%E6%88%90%E6%9F%90%E7%A7%8D%E7%89%B9%E5%9C%B0%E4%BB%BB%E5%8A%A1%E8%80%8C%E5%88%9B%E5%BB%BA%E7%9A%84%EF%BC%8C%E6%AF%94%E5%A6%82%20AlphaGo%20%E5%B0%B1%E5%8F%AA%E6%98%AF%E4%BC%9A%E4%B8%8B%E5%9B%B4%E6%A3%8B%E3%80%82%E7%9B%B8%E8%BE%83%E4%B9%8B%E4%B8%8B%EF%BC%8C%E6%96%B0%E4%B8%80%E4%BB%A3%E5%9F%BA%E4%BA%8E%E5%9F%BA%E7%A1%80%E6%A8%A1%E5%9E%8B%E7%9A%84%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%BF%E6%99%BA%E8%83%BD%E4%BD%93%E6%9B%B4%E5%8A%A0%E6%99%AE%E9%81%8D%EF%BC%8C%E5%9B%A0%E4%B8%BA%E5%AE%83%E4%BB%AC%E5%8F%AF%E4%BB%A5%E4%BB%8E%E4%BA%BA%E7%B1%BB%E4%BA%92%E5%8A%A8%E7%9A%84%E4%B8%96%E7%95%8C%E4%B8%AD%E8%BF%9B%E8%A1%8C%E5%AD%A6%E4%B9%A0%E3%80%82%E2%80%9D)。

2023年，被许多人视为“Agentic AI元年”。这一年里，一系列通用代理的探索性应用涌现：

* **多模态智能体：** OpenAI在2023年3月发布了GPT-4，其引入多模态能力，能够处理图像与文本输入。这使智能体朝着“**多模态交互**”迈进一大步。例如在GPT-4演示中，用户给出冰箱内部物品照片，AI能识别物品并建议菜谱。这暗示智能体未来将能同时处理语言、视觉、音频等信息源[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00" \l ":~:text=%E2%80%9C%E5%A4%9A%E6%A8%A1%E6%80%81%E2%80%9D%E6%98%AF%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%93%E7%9A%84%E9%87%8D%E8%A6%81%E6%84%BF%E6%99%AF%E4%B9%8B%E4%B8%80%EF%BC%8C%E8%BF%99%E6%84%8F%E5%91%B3%E7%9D%80%E5%AE%83%E4%BB%AC%E5%8F%AF%E4%BB%A5%E5%A4%84%E7%90%86%E8%AF%AD%E8%A8%80%E3%80%81%E9%9F%B3%E9%A2%91%E5%92%8C%E8%A7%86%E9%A2%91%E7%AD%89%E3%80%82%E4%BE%8B%E5%A6%82%EF%BC%8C%E5%9C%A8%E8%B0%B7%E6%AD%8C%E7%9A%84%E6%BC%94%E7%A4%BA%E4%B8%AD%EF%BC%8C%E7%94%A8%E6%88%B7%E5%8F%AF%E4%BB%A5%E7%94%A8%E6%99%BA%E8%83%BD%E6%89%8B%E6%9C%BA%E6%91%84%E5%83%8F%E5%A4%B4%E5%AF%B9%E5%87%86%E5%90%84%E7%A7%8D%E7%89%A9%E4%BD%93%E5%B9%B6%E5%90%91%20Astra%20%E6%8F%90%E9%97%AE%EF%BC%8C%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%93%E5%8F%AF%E4%BB%A5%E5%AF%B9%E6%96%87%E6%9C%AC%E3%80%81%E9%9F%B3%E9%A2%91%E4%BB%A5%E5%8F%8A%E8%A7%86%E9%A2%91%E8%BE%93%E5%85%A5%E5%81%9A%E5%87%BA%E5%9B%9E%E5%BA%94%E3%80%82)。谷歌DeepMind则研发代号**Gemini**的新模型，被设计为多模态且具有更强推理能力，分为不同规模版本，于2023年底陆续推出[techpoint.africa](https://techpoint.africa/guide/meta-ai-vs-chatgpt-vs-google-gemini/" \l ":~:text=Launched%20in%202023%2C%20Gemini%20is,major%20advantage%20over%20its%20competitors)。Gemini据报道在**实时网页访问、复杂推理和研究类任务**上表现出色，被视为与GPT-4竞争的通用AI[techtarget.com](https://www.techtarget.com/searchenterpriseai/tip/Gemini-vs-ChatGPT-Whats-the-difference#:~:text=In%20contrast%2C%20Google%27s%20Gemini%20excels,competitor%20in%20the%20AI)。到2024年，OpenAI也开放了代号**Sora**的视频生成模型，允许ChatGPT用户生成短视频内容[yicai.com](https://www.yicai.com/news/102393698.html" \l ":~:text=%E6%96%B0%E7%89%88%E5%B7%A5%E5%85%B7Sora%20Turbo%E5%8F%AF%E4%BB%A5%E7%94%9F%E6%88%90%E6%9C%80%E9%95%BF%E8%BE%BE20%E7%A7%92%E7%9A%84%E8%A7%86%E9%A2%91%EF%BC%8C%E5%B9%B6%E4%B8%94%E5%8F%AF%E4%BB%A5%E6%8F%90%E4%BE%9B%E8%BF%99%E4%BA%9B%E8%A7%86%E9%A2%91%E7%9A%84%E5%A4%9A%E7%A7%8D%E5%8F%98%E4%BD%93%E3%80%82)。Sora标志着智能体开始掌握动态图像创作能力，进一步朝多模态通用智能迈进。多模态被认为是人工智能体的重要愿景之一，即未来的Agent可以同时理解和生成语言、图像、音频、视频等各种模态信息[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00" \l ":~:text=%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%93%E7%9A%84%E5%AE%8F%E4%BC%9F%E6%84%BF%E6%99%AF%E6%98%AF%E4%B8%80%E4%B8%AA%E5%8F%AF%E4%BB%A5%E6%89%A7%E8%A1%8C%E5%A4%A7%E9%87%8F%E4%BB%BB%E5%8A%A1%E7%9A%84%E7%B3%BB%E7%BB%9F%EF%BC%8C%E5%B0%B1%E5%A5%BD%E6%AF%94%E6%98%AF%E4%B8%80%E4%B8%AA%E2%80%9C%E4%BA%BA%E7%B1%BB%E5%8A%A9%E6%89%8B%E2%80%9D%E3%80%82%E6%9C%AA%E6%9D%A5%EF%BC%8C%E5%AE%83%E5%8F%AF%E4%BB%A5%E5%B8%AE%E5%8A%A9%E4%BD%A0%E5%88%B6%E5%AE%9A%E5%81%87%E6%9C%9F%E8%AE%A1%E5%88%92%EF%BC%8C%E8%AE%B0%E4%BD%8F%E4%BD%A0%E6%98%AF%E5%90%A6%E6%9B%B4%E5%96%9C%E6%AC%A2%E8%B1%AA%E5%8D%8E%E9%85%92%E5%BA%97%EF%BC%8C%E5%B9%B6%E5%9C%A8%E5%9B%9B%E6%98%9F%E7%BA%A7%E6%88%96%E5%9B%9B%E6%98%9F%E7%BA%A7%E4%BB%A5%E4%B8%8A%E7%9A%84%E9%85%92%20%E5%BA%97%E9%87%8C%E9%A2%84%E8%AE%A2%E5%85%B6%E4%B8%AD%E7%9A%84%E4%B8%80%E5%AE%B6%EF%BC%9B%E7%84%B6%E5%90%8E%EF%BC%8C%E5%AE%83%E8%BF%98%E5%B0%86%E5%BB%BA%E8%AE%AE%E6%9C%80%E9%80%82%E5%90%88%E4%BD%A0%E6%97%A5%E7%A8%8B%E7%9A%84%E8%88%AA%E7%8F%AD%EF%BC%8C%E5%B9%B6%E6%A0%B9%E6%8D%AE%E4%BD%A0%E7%9A%84%E5%96%9C%E5%A5%BD%E8%A7%84%E5%88%92%E8%A1%8C%E7%A8%8B%EF%BC%9B%E5%AE%83%E5%8F%AF%E4%BB%A5%E6%A0%B9%E6%8D%AE%E6%97%85%E8%A1%8C%E8%AE%A1%E5%88%92%E5%92%8C%E5%A4%A9%E6%B0%94%E9%A2%84%E6%8A%A5%E5%88%97%E5%87%BA%E9%9C%80%E8%A6%81%E6%90%BA%E5%B8%A6%E7%9A%84%E4%B8%AA%E4%BA%BA%E7%89%A9%E5%93%81%E6%B8%85%E5%8D%95%EF%BC%9B%E5%AE%83%E7%94%9A%E8%87%B3%E5%8F%AF%E8%83%BD%E6%8A%8A%E4%BD%A0%E7%9A%84%E8%A1%8C%E7%A8%8B%E5%AE%89%E6%8E%92%20%E5%8F%91%E9%80%81%E7%BB%99%E4%BD%A0%E7%9A%84%E5%A5%BD%E6%9C%8B%E5%8F%8B%E5%B9%B6%E9%82%80%E8%AF%B7%E4%BB%96%E4%BB%AC%E4%B8%80%E8%B5%B7%E6%97%85%E8%A1%8C%EF%BC%9B%E8%80%8C%E5%9C%A8%E5%B7%A5%E4%BD%9C%E4%B8%AD%EF%BC%8C%E5%AE%83%E5%8F%AF%E4%BB%A5%E5%88%86%E6%9E%90%E4%BD%A0%E7%9A%84%E5%BE%85%E5%8A%9E%E4%BA%8B%E9%A1%B9%E5%88%97%E8%A1%A8%E5%B9%B6%E6%89%A7%E8%A1%8C%E4%BB%BB%E5%8A%A1%EF%BC%8C%E4%BE%8B%E5%A6%82%E5%8F%91%E9%80%81%E4%BC%9A%E8%AE%AE%E9%82%80%E8%AF%B7%E3%80%81%E5%A4%87%E5%BF%98%E5%BD%95%E4%BB%A5%E5%8F%8A%E7%94%B5%E5%AD%90%E9%82%AE%E4%BB%B6%E7%AD%89%E3%80%82)。
* **自主任务代理：** 随着对大模型“**思维链**”提示技术和工具使用的研究深入，开发者开始尝试让AI代理自主完成更复杂的长程任务。2023年4月，社区开源项目**AutoGPT**引发轰动。AutoGPT将GPT-4接入互联网检索、插件工具，并赋予其设定目标后**自行分解任务、循环执行**的框架[yicai.com](https://www.yicai.com/news/102678941.html" \l ":~:text=%E6%9C%AC%E5%91%A8%E5%8F%88%E6%98%AFAgent%EF%BC%88%E6%99%BA%E8%83%BD%E4%BD%93%EF%BC%89%E7%83%AD%E5%BA%A6%E6%8C%81%E7%BB%AD%E7%9A%84%E4%B8%80%E5%91%A8%E3%80%82%E5%8C%97%E4%BA%AC%E6%97%B6%E9%97%B46%E6%9C%8819%E6%97%A5%EF%BC%8C%E5%9B%BD%E5%86%85%E5%A4%A7%E6%A8%A1%E5%9E%8B%E5%88%9D%E5%88%9B%E4%BC%81%E4%B8%9AMinimax%E5%8F%91%E5%B8%83%E8%83%BD%E5%A4%9F%E5%AE%8C%E6%88%90%E9%95%BF%E7%A8%8B%EF%BC%88Long%20Horizon%EF%BC%89%E5%A4%8D%E6%9D%82%E4%BB%BB%E5%8A%A1%E7%9A%84%E9%80%9A%E7%94%A8%E6%99%BA%E8%83%BD%E4%BD%93MiniMax%20Agent%EF%BC%8C%E5%A4%9A%E6%AD%A5%E8%A7%84%E5%88%92%E4%B8%93%E5%AE%B6%E7%BA%A7%E8%A7%A3%E5%86%B3%E6%96%B9%E6%A1%88%E3%80%81%E6%8B%86%E8%A7%A3%E4%BB%BB%E5%8A%A1%E9%9C%80%E6%B1%82%E3%80%81%E6%89%A7%E8%A1%8C%E5%A4%9A%E4%B8%AA%E5%AD%90%E4%BB%BB%E5%8A%A1%E4%BB%8E%E8%80%8C%E4%BA%A4%E4%BB%98%E6%9C%80%E7%BB%88%E7%BB%93%E6%9E%9C%E3%80%82)。用户只需给出高层目标，AutoGPT就会产生一系列子任务如搜索信息、代码执行、结果分析，不断迭代直至完成目标。这被视为最初的“**自主AI经纪人**”实践。同类的**AgentGPT**提供了Web界面，方便用户创建并行执行的多个子代理处理任务。尽管这类自主代理当时能力仍有限，但展现了AI无需每步人为介入即可连续行动的潜力。Nvidia研究人员也演示了让GPT-4控制游戏中角色持续探索、学习技能的智能体（如Minecraft游戏中的**Voyager**），表现出令人惊叹的自主性[woshipm.com](https://www.woshipm.com/it/6117525.html" \l ":~:text=%E6%B7%B1%E5%BA%A6%EF%BD%9C%E7%BA%A2%E6%9D%89%E7%BE%8E%E5%9B%BD%E5%AF%B9%E8%AF%9D%E8%8B%B1%E4%BC%9F%E8%BE%BE%E7%A7%91%E5%AD%A6%E5%AE%B6Jim%20Fan%EF%BC%9A%E5%8D%81%E5%B9%B4%E5%86%85%E4%BA%BA%E5%BD%A2%E6%9C%BA%E5%99%A8%E4%BA%BA%E5%B0%86%E8%B6%85%E8%B6%8A%20,Fan%E3%80%82Jim%E9%A2%86%E5%AF%BCNvidia%20%E7%9A%84%E5%85%B7%E8%BA%AB%E6%99%BA%E8%83%BD%EF%BC%88Embodied%20AI%20%EF%BC%89%E6%99%BA%E8%83%BD%E4%BD%93%E7%A0%94%E7%A9%B6%E5%B7%A5%E4%BD%9C%EF%BC%8C%E6%B6%89%E5%8F%8A%E7%89%A9%E7%90%86%E4%B8%96%E7%95%8C%E4%B8%AD%E7%9A%84%E6%9C%BA%E5%99%A8%E4%BA%BA%E6%8A%80%E6%9C%AF%E5%92%8C%E8%99%9A%E6%8B%9F%E4%B8%96%E7%95%8C%E4%B8%AD%E7%9A%84%E6%B8%B8%E6%88%8F%E6%99%BA%E8%83%BD%E4%BD%93%E3%80%82)[i-newcar.com](https://i-newcar.com/index.php?m=home&c=View&a=index&aid=3382#:~:text=%E5%9B%BD%E5%86%85%E5%A4%96%E5%85%B8%E5%9E%8B%E7%9A%84%E6%99%BA%E8%83%BD%E6%9C%BA%E5%99%A8%E4%BA%BA%E5%A4%A7%E6%A8%A1%E5%9E%8B%E8%AF%A6%E8%A7%A3%20)。自主代理的出现，使人们开始憧憬更高级的“AI劳动力”：未来智能体或许可自动完成从制定假期计划、预订机票酒店到收发邮件、处理客户投诉等大量日常或业务任务[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00" \l ":~:text=%E7%BA%A7%E7%A0%94%E7%A9%B6%E7%A7%91%E5%AD%A6%E5%AE%B6%E3%80%81%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%93%E9%A1%B9%E7%9B%AE%E8%B4%9F%E8%B4%A3%E4%BA%BA%20Jim%20Fan%20%E8%A1%A8%E7%A4%BA%E3%80%82)[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00#:~:text=%E6%AF%94%E5%A6%82%EF%BC%8C%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%93%E5%8F%AF%E8%83%BD%E8%83%BD%E5%A4%9F%E5%85%85%E5%BD%93%E6%9B%B4%E5%A4%8D%E6%9D%82%E7%9A%84%E5%AE%A2%E6%88%B7%E6%9C%8D%E5%8A%A1%E6%9C%BA%E5%99%A8%E4%BA%BA%E3%80%82%E5%BD%93%E5%89%8D%E5%9F%BA%E4%BA%8E%E8%AF%AD%E8%A8%80%E6%A8%A1%E5%9E%8B%E7%9A%84%E5%8A%A9%E6%89%8B%E5%8F%AA%E8%83%BD%E7%94%9F%E6%88%90%E5%8F%A5%E5%AD%90%E4%B8%AD%E4%B8%8B%E4%B8%80%E4%B8%AA%E5%8F%AF%E8%83%BD%E7%9A%84%E5%8D%95%E8%AF%8D%E5%B9%B6%E7%BB%84%E6%88%90%E5%8F%A5%E5%AD%90%EF%BC%8C%E8%80%8C%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%93%E5%B0%86%E6%9C%89%E8%83%BD%E5%8A%9B%E8%87%AA%E4%B8%BB%E5%A4%84%E7%90%86%E8%87%AA%E7%84%B6%E8%AF%AD%E8%A8%80%E5%91%BD%E4%BB%A4%EF%BC%8C%E5%B9%B6%20%E4%B8%94%E5%9C%A8%E6%B2%A1%E6%9C%89%E7%9B%91%E7%9D%A3%E7%9A%84%E6%83%85%E5%86%B5%E4%B8%8B%E5%A4%84%E7%90%86%E5%AE%A2%E6%88%B7%E6%9C%8D%E5%8A%A1%E4%BB%BB%E5%8A%A1%E3%80%82%E5%86%8D%E6%AF%94%E5%A6%82%EF%BC%8C%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%93%E5%B0%86%E8%83%BD%E5%A4%9F%E5%88%86%E6%9E%90%E5%AE%A2%E6%88%B7%E7%9A%84%E6%8A%95%E8%AF%89%E7%94%B5%E5%AD%90%E9%82%AE%E4%BB%B6%EF%BC%8C%E7%84%B6%E5%90%8E%E7%9F%A5%E9%81%93%E5%A6%82%E4%BD%95%E6%A3%80%E6%9F%A5%E5%AE%A2%E6%88%B7%E7%9A%84%E5%8D%95%E5%8F%B7%EF%BC%8C%E8%AE%BF%E9%97%AE%E5%AE%A2%E6%88%B7%E5%85%B3%E7%B3%BB%E7%AE%A1%E7%90%86%E5%92%8C%E4%BA%A4%E4%BB%98%E7%B3%BB%E7%BB%9F%E7%AD%89%E6%95%B0%E6%8D%AE%E5%BA%93%EF%BC%8C%E4%BB%A5%E6%9F%A5%E7%9C%8B%E6%8A%95%E8%AF%89%20%E6%98%AF%E5%90%A6%E5%90%88%E6%B3%95%EF%BC%8C%E5%B9%B6%E6%A0%B9%E6%8D%AE%E5%85%AC%E5%8F%B8%E7%9A%84%E6%94%BF%E7%AD%96%E8%BF%9B%E8%A1%8C%E5%A4%84%E7%90%86%E3%80%82)。权威人士如OpenAI CEO Sam Altman甚至预言，下一阶段的大模型（GPT-5）将整合跨模态和跨任务能力，构建一个通用的“**执行层**”，赋予模型**强大的干活能力**来大幅提高生产率[yicai.com](https://www.yicai.com/news/102678941.html#:~:text=%E5%90%8C%E6%97%A5%EF%BC%8COpenAI%20CEO%20%E5%B1%B1%E5%A7%86%C2%B7%E5%A5%A5%E5%B0%94%E7%89%B9%E6%9B%BC%EF%BC%88Sam%20Altman%EF%BC%89%E5%9C%A8%E5%AE%98%E6%96%B9%E6%92%AD%E5%AE%A2Podcast%E4%B8%8A%E8%A1%A8%E7%A4%BA%EF%BC%8C%E6%88%96%E5%B0%86%E4%BA%8E%E4%BB%8A%E5%B9%B4%E5%A4%8F%E5%A4%A9%E5%8F%91%E5%B8%83%E7%9A%84GPT,Series%EF%BC%8C%E6%90%AD%E5%BB%BA%E8%B7%A8%E6%A8%A1%E6%80%81%E3%80%81%E8%B7%A8%E4%BB%BB%E5%8A%A1%E7%9A%84%E2%80%9C%E9%80%9A%E7%94%A8%E6%89%A7%E8%A1%8C%E5%B1%82%E2%80%9D%EF%BC%8C%E5%B9%B6%E6%8A%AB%E9%9C%B2%E5%85%AC%E5%8F%B8%E5%9C%A8%E5%BA%95%E5%B1%82%E7%AE%97%E5%8A%9B%E6%96%B9%E9%9D%A2%E7%9A%84%E8%BF%9B%E5%B1%95%E6%83%85%E5%86%B5%EF%BC%8C%E6%98%8E%E7%A1%AE%E4%BA%86%E4%BB%8A%E5%B9%B4%E5%A4%A7%E6%A8%A1%E5%9E%8B%E4%BA%A7%E5%93%81%E5%BC%BA%E6%89%A7%E8%A1%8C%E4%B8%8E%E9%AB%98%E7%AE%97%E5%8A%9B%E9%9C%80%E6%B1%82%E7%9A%84%E5%B1%9E%E6%80%A7%E7%89%B9%E5%BE%81%E3%80%82)[yicai.com](https://www.yicai.com/news/102678941.html#:~:text=%E5%9C%A8%E5%B1%B1%E5%A7%86%C2%B7%E5%A5%A5%E5%B0%94%E7%89%B9%E6%9B%BC%E7%9A%84%E8%A1%A8%E8%BF%B0%E4%B8%AD%EF%BC%8C%E6%96%B0%E6%A8%A1%E5%BC%8F%E7%9A%84%E6%98%BE%E8%91%97%E7%89%B9%E6%80%A7%E5%B7%B2%E4%B8%8D%E6%98%AF%E6%80%A7%E8%83%BD%E5%A4%9A%E4%B9%88%E5%A5%BD%EF%BC%8C%E8%80%8C%E6%98%AF%E5%B9%B2%E6%B4%BB%E5%A4%9A%E4%B9%88%E5%BC%BA%E3%80%82)。

总的来看，新一代浪潮的智能体在**广度**和**灵活性**上远超以往。一方面，大模型充当大脑，提供了强大的语言理解、知识迁移和推理能力；另一方面，工具使用和自主规划模块充当四肢，让智能体可以与外界交互执行具体操作。这使得当今的智能体开始具备“**通用目的智能代理**”的雏形。当然，目前这些探索也暴露出许多问题（在第四章详述），但不可否认的是，AI领域正从“问答机器人”迈向“**任务执行代理**”的转变。这波浪潮引发科技公司巨资投入，Gartner将“Agentic AI”列为2024年前后最重要的技术趋势之一[thejournal.com](https://thejournal.com/articles/2024/10/23/gartner-names-agentic-ai-top-tech-trend-for-2025.aspx" \l ":~:text=Gartner%20Names%20Agentic%20AI%20Top,response)。

### **3.2.4 功能演化阶段与代表性模型**

为了更清晰地理解智能体技术在不同发展阶段侧重的核心功能，我们可以按照演化脉络，将其概括为以下几个阶段，并列举具有代表性的里程碑模型：

* **阶段1：内容创作智能体**（生成式AI崛起，~2014-2020年）：这一阶段以生成内容为主要侧重功能，智能体更多表现为“生成模型”而非完整自主代理。典型代表如**GPT-3**（2020）展示了以一段文本提示生成长篇文章、代码的惊人能力，证明了AI在自然语言内容创作上的潜力。之后的**DALL-E**（2021）等模型则扩展到图像生成领域，实现“以文生图”。这时期的系统严格说不是“智能体”，因为它们不具备主动性，但为后来的代理奠定了关键基础——大模型提供了**语言表达和知识**能力，可用于创作文本、图像、代码等各种内容。[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00" \l ":~:text=%E8%B0%B7%E6%AD%8C%20DeepMind%20%E7%A0%94%E7%A9%B6%E5%89%AF%E6%80%BB%E8%A3%81%20Oriol%20Vinyals,%E8%A1%A8%E7%A4%BA%EF%BC%8C%E2%80%9C%E4%BD%86%E8%BF%99%E4%BA%9B%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%93%E5%A4%AA%E8%BF%87%E4%BA%8E%E2%80%98%E4%B8%93%E4%B8%80%E2%80%99%EF%BC%8C%E6%8D%A2%E5%8F%A5%E8%AF%9D%E8%AF%B4%EF%BC%8C%E8%BF%99%E4%BA%9B%E6%99%BA%E8%83%BD%E4%BD%93%E5%8F%AA%E6%98%AF%E4%B8%BA%E5%AE%8C%E6%88%90%E6%9F%90%E7%A7%8D%E7%89%B9%E5%9C%B0%E4%BB%BB%E5%8A%A1%E8%80%8C%E5%88%9B%E5%BB%BA%E7%9A%84%EF%BC%8C%E6%AF%94%E5%A6%82%20AlphaGo%20%E5%B0%B1%E5%8F%AA%E6%98%AF%E4%BC%9A%E4%B8%8B%E5%9B%B4%E6%A3%8B%E3%80%82%E7%9B%B8%E8%BE%83%E4%B9%8B%E4%B8%8B%EF%BC%8C%E6%96%B0%E4%B8%80%E4%BB%A3%E5%9F%BA%E4%BA%8E%E5%9F%BA%E7%A1%80%E6%A8%A1%E5%9E%8B%E7%9A%84%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%BF%E6%99%BA%E8%83%BD%E4%BD%93%E6%9B%B4%E5%8A%A0%E6%99%AE%E9%81%8D%EF%BC%8C%E5%9B%A0%E4%B8%BA%E5%AE%83%E4%BB%AC%E5%8F%AF%E4%BB%A5%E4%BB%8E%E4%BA%BA%E7%B1%BB%E4%BA%92%E5%8A%A8%E7%9A%84%E4%B8%96%E7%95%8C%E4%B8%AD%E8%BF%9B%E8%A1%8C%E5%AD%A6%E4%B9%A0%E3%80%82%E2%80%9D)
* **阶段2：对话交互智能体**（对话代理兴起，2022年起）：以ChatGPT为标志，各大模型开始通过对话形式与用户交互，注重在多轮对话中理解意图、提供有用答复。**ChatGPT**（2022）和**Claude**（2023）是这一阶段的代表模型。这些对话代理强调**指令遵循**（Instruction Following）和**上下文记忆**，能够根据用户连续提出的问题做出连贯回答，形成类似人类助手的交流体验。对话能力的提升使AI第一次真正以“代理人”身份出现：用户可以像和助手对话般让AI执行各种软性任务（写邮件、翻译、头脑风暴等）。这一阶段体现了智能体的**互动性**和**语言理解**能力大幅增强。
* **阶段3：多模态与世界模型智能体**（感知与生成多模态信息，2023年起）：在对话基础上，智能体开始拥有多模态感知和生成的能力，能够处理文本、图像、音频乃至视频等综合信息。例如**GPT-4**将视觉输入纳入模型，[blog.google](https://blog.google/technology/ai/google-gemini-ai/" \l ":~:text=Introducing%20Gemini%3A%20our%20largest%20and,sizes%3A%20Ultra%2C%20Pro%20and%20Nano)**Gemini**则被设计为同时支持图像、声音输入和输出[blog.google](https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/" \l ":~:text=In%20addition%20to%20supporting%20multimodal,output%20like%20natively%20generated)。OpenAI的**Sora**更是首个面向大众开放的视频生成模型[yicai.com](https://www.yicai.com/news/102393698.html" \l ":~:text=%E6%96%B0%E7%89%88%E5%B7%A5%E5%85%B7Sora%20Turbo%E5%8F%AF%E4%BB%A5%E7%94%9F%E6%88%90%E6%9C%80%E9%95%BF%E8%BE%BE20%E7%A7%92%E7%9A%84%E8%A7%86%E9%A2%91%EF%BC%8C%E5%B9%B6%E4%B8%94%E5%8F%AF%E4%BB%A5%E6%8F%90%E4%BE%9B%E8%BF%99%E4%BA%9B%E8%A7%86%E9%A2%91%E7%9A%84%E5%A4%9A%E7%A7%8D%E5%8F%98%E4%BD%93%E3%80%82)。此外，Meta等公司也在研发让AI拥有“世界模型”（World Model）的智能体，使其能对动态环境建模并预测变化[anquanke.com](https://www.anquanke.com/post/id/294246" \l ":~:text=WorldGPT%E6%9D%A5%E4%BA%86%EF%BC%9A%E6%89%93%E9%80%A0%E7%B1%BBSora%E8%A7%86%E9%A2%91AI%E6%99%BA%E8%83%BD%E4%BD%93%20,%E9%A6%96%E5%85%88%EF%BC%8C%E5%AE%83%E4%BB%A5%E5%A4%A7%E5%9E%8B%E8%AF%AD%E8%A8%80)。具身环境中的探索（如机器人）也在加强智能体的**物理推理**和**连续决策**能力。总之，这一阶段智能体正在从单一的语言领域扩展到对**现实世界的多模态理解与行动**。它预示未来的智能体将更接近人类感官：能看会听、能说会画，在复杂环境中持续感知和决策。
* **阶段4：增强推理与自主执行智能体**（工具使用与自主性，2023年至今）：当前正在发展的前沿，是让智能体具备更强的**自主规划和推理**能力，可以长时间自主运行并完成复杂目标。典型探索包括前述的AutoGPT、AgentGPT，它们实现了AI的**工具调用**和**子任务生成循环**[yicai.com](https://www.yicai.com/news/102678941.html" \l ":~:text=%E6%9C%AC%E5%91%A8%E5%8F%88%E6%98%AFAgent%EF%BC%88%E6%99%BA%E8%83%BD%E4%BD%93%EF%BC%89%E7%83%AD%E5%BA%A6%E6%8C%81%E7%BB%AD%E7%9A%84%E4%B8%80%E5%91%A8%E3%80%82%E5%8C%97%E4%BA%AC%E6%97%B6%E9%97%B46%E6%9C%8819%E6%97%A5%EF%BC%8C%E5%9B%BD%E5%86%85%E5%A4%A7%E6%A8%A1%E5%9E%8B%E5%88%9D%E5%88%9B%E4%BC%81%E4%B8%9AMinimax%E5%8F%91%E5%B8%83%E8%83%BD%E5%A4%9F%E5%AE%8C%E6%88%90%E9%95%BF%E7%A8%8B%EF%BC%88Long%20Horizon%EF%BC%89%E5%A4%8D%E6%9D%82%E4%BB%BB%E5%8A%A1%E7%9A%84%E9%80%9A%E7%94%A8%E6%99%BA%E8%83%BD%E4%BD%93MiniMax%20Agent%EF%BC%8C%E5%A4%9A%E6%AD%A5%E8%A7%84%E5%88%92%E4%B8%93%E5%AE%B6%E7%BA%A7%E8%A7%A3%E5%86%B3%E6%96%B9%E6%A1%88%E3%80%81%E6%8B%86%E8%A7%A3%E4%BB%BB%E5%8A%A1%E9%9C%80%E6%B1%82%E3%80%81%E6%89%A7%E8%A1%8C%E5%A4%9A%E4%B8%AA%E5%AD%90%E4%BB%BB%E5%8A%A1%E4%BB%8E%E8%80%8C%E4%BA%A4%E4%BB%98%E6%9C%80%E7%BB%88%E7%BB%93%E6%9E%9C%E3%80%82)。此外，Anthropic等机构也在研究所谓“**权能智能体**”，让LLM自行指导自己的流程，并调用外部计算或记忆资源[woshipm.com](https://www.woshipm.com/ai/6202918.html" \l ":~:text=AI%20%E6%99%BA%E8%83%BD%E4%BD%93%E8%A7%A3%E9%87%8A%EF%BC%88%E6%99%BA%E8%83%BD%E4%BD%93%E7%9A%843%20%E4%B8%AA%E7%BA%A7%E5%88%AB%EF%BC%89%20OpenAI%EF%BC%9A%E4%B8%80%E4%B8%AA%E5%A4%A7%E5%9E%8B%E8%AF%AD%E8%A8%80%E6%A8%A1%E5%9E%8B%EF%BC%88LLM%EF%BC%89%EF%BC%8C%E9%85%8D%E7%BD%AE%E4%BA%86%E6%8C%87%E4%BB%A4%E5%92%8C%E5%B7%A5%E5%85%B7%EF%BC%9B%20%C2%B7,%E5%8F%AF%E4%BB%A5%E9%80%9A%E8%BF%87%E8%A7%84%E5%88%92%E5%92%8C%E4%BD%BF%E7%94%A8%E5%B7%A5%E5%85%B7%E6%89%A7%E8%A1%8C%E6%9B%B4%E5%A4%8D%E6%9D%82%E7%9A%84%E4%BB%BB%E5%8A%A1%E7%9A%84%E7%B3%BB%E7%BB%9F%EF%BC%9B%20%C2%B7%20Anthropic%EF%BC%9ALLMs%20%E5%8A%A8%E6%80%81%E6%8C%87%E5%AF%BC%E8%87%AA%E5%B7%B1%E7%9A%84%E6%B5%81%E7%A8%8B%E5%92%8C)。微软提出“Jarvis”架构让GPT调用一系列专用模型完成复杂任务[ai.iias.sinica.edu.tw](https://ai.iias.sinica.edu.tw/glossary/intelligent-agent/" \l ":~:text=Network%20ai,Lab%EF%BC%89%E3%80%82%20%E6%99%BA%E6%85%A7%E5%9E%8B%E4%BB%A3%E7%90%86%E4%BA%BA%E4%B8%A6%E7%84%A1%E7%B5%B1%E4%B8%80%E4%B8%94%E6%98%8E%E7%A2%BA%E7%9A%84%E5%AE%9A%E7%BE%A9%E3%80%82%E5%B8%B8%E8%A6%8B%E7%9A%84%E5%AE%9A%E7%BE%A9%E6%98%AF%EF%BC%8C%E4%B8%80%E5%80%8B%E8%83%BD%E7%90%86%E8%A7%A3%E5%85%B6%E5%91%A8%E9%81%AD%E7%92%B0%E5%A2%83%EF%BC%8C%E4%B8%A6%E5%BE%9E%E4%B8%AD%E5%AD%B8%E7%BF%92%E4%B8%94%E8%83%BD)。这些工作旨在突破大模型原本的局限，使其能“**主动地为达到目标而行动**”。增强推理智能体必须解决**任务分解、长期记忆、实时学习**等难题。例如，当前大模型有上下文窗口限制，导致智能体过一段时间可能完全“忘记”自己之前的工作进度[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00" \l ":~:text=Jim%20Fan%20%E5%88%99%E8%A1%A8%E7%A4%BA%EF%BC%9A%E2%80%9C%E6%88%91%E4%BB%AC%E8%BF%98%E8%BF%9C%E8%BF%9C%E6%B2%A1%E6%9C%89%E4%B8%80%E4%B8%AA%E5%8F%AF%E4%BB%A5%E4%B8%BA%E6%88%91%E4%BB%AC%E8%87%AA%E5%8A%A8%E5%8C%96%E6%89%80%E6%9C%89%E8%BF%99%E4%BA%9B%E5%AE%B6%E5%8A%A1%E7%9A%84%E6%99%BA%E8%83%BD%E4%BD%93%E3%80%82%E7%9B%AE%E5%89%8D%E7%9A%84%E7%B3%BB%E7%BB%9F%E5%8F%AF%E8%83%BD%E4%BC%9A%E4%BA%A7%E7%94%9F%E5%B9%BB%E8%A7%89%EF%BC%8C%E8%80%8C%E4%B8%94%E5%AE%83%E4%BB%AC%E4%B9%9F%E4%B8%8D%E6%80%BB%E6%98%AF%E4%B8%A5%E6%A0%BC%E9%81%B5%E5%BE%AA%E6%8C%87%E4%BB%A4%EF%BC%8C%E6%98%BE%E7%84%B6%E8%BF%99%E5%B0%B1%E5%BE%88%E9%BA%BB%E7%83%A6%E4%BA%86%E3%80%82%E2%80%9D)。为此研究者在尝试扩展上下文乃至引入**无限记忆**机制[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00" \l ":~:text=%E2%80%9CChatGPT%20%E5%8F%AF%E4%BB%A5%E6%95%B2%E5%86%99%E4%BB%A3%E7%A0%81%EF%BC%8C%E4%BD%86%E4%B8%8D%E8%83%BD%E5%BE%88%E5%A5%BD%E5%9C%B0%E5%A4%84%E7%90%86%E8%BE%83%E9%95%BF%E7%9A%84%E5%86%85%E5%AE%B9%E3%80%82%E4%BD%86%E5%AF%B9%E4%BA%8E%E4%BA%BA%E7%B1%BB%E5%BC%80%E5%8F%91%E4%BA%BA%E5%91%98%E6%9D%A5%E8%AF%B4%E8%A6%81%E6%9F%A5%E9%98%85%E7%9A%84%E6%98%AF%E6%95%B4%E4%B8%AA%20GitHub%20%E4%BB%A3%E7%A0%81%E5%BA%93%EF%BC%8C%E5%85%B6%E4%B8%AD%E6%9C%89%E5%87%A0%E5%8D%81%E8%A1%8C%E4%B9%83%E8%87%B3%E6%95%B0%E7%99%BE%E8%A1%8C%E4%BB%A3%E7%A0%81%EF%BC%8C%E6%98%BE%E7%84%B6%EF%BC%8C%E8%BF%99%E5%AF%B9%E4%BA%8E%E4%BA%BA%E7%B1%BB%E8%80%8C%E8%A8%80%E5%AE%8C%E5%85%A8%E6%B2%A1%E6%9C%89%E4%BB%BB%E4%BD%95%E5%8E%8B%E5%8A%9B%E3%80%82%E2%80%9DJim%20Fan,%E8%AF%B4%E3%80%82)。增强推理阶段的一个目标是实现真正的“长程自主AI代理”（Long-horizon Autonomous Agent），可以在没有人工干预下执行数小时乃至更长时间，不断根据环境反馈调整策略，直至任务完成。这一方向仍在快速演进，尚未成熟，但已有初步成果让业界感到乐观。例如，Minimax（一家中国AI初创公司）在2025年发布了据称可完成**长程复杂任务**的通用智能体，实现多步规划和多个子任务执行，交付最终结果[yicai.com](https://www.yicai.com/news/102678941.html" \l ":~:text=%E5%A4%A7%E6%A8%A1%E5%9E%8B%E8%A1%8C%E4%B8%9A%E5%86%9B%E5%A4%87%E7%AB%9E%E8%B5%9B%E4%BB%8E%E4%B8%80%E7%BA%A7%E5%B8%82%E5%9C%BA%E7%9A%84%E6%B3%A1%E6%B2%AB%E5%8C%96%E9%AB%98%E9%A2%9D%E4%BC%B0%E5%80%BC%EF%BC%8C%E8%90%BD%E5%9C%B0%E5%88%B0%E6%9B%B4%E5%BA%95%E5%B1%82%E7%9A%84%E7%AE%97%E5%8A%9B%E5%9F%BA%E7%A1%80%E8%AE%BE%E6%96%BD%E5%BB%BA%E8%AE%BE%E3%80%82)。

在每个阶段，我们都能列举代表性模型来体现该阶段的核心进展：从GPT-3到ChatGPT、Claude，再到GPT-4、Gemini、Sora，以及AutoGPT等，它们共同勾勒出智能体从“**能生成**”到“**能对话**”再到“**能行动**”的发展轨迹。需要说明的是，这些阶段并非截然分明、线性替代的关系，而是相互叠加演进：内容生成能力仍是基础，对话交互赋予其人格化接口，多模态扩展感知范围，增强推理则赋予更高自主性。最终的目标形态是集这些能力于一身的**通用智能体**。谷歌前CEO埃里克·施密特近日也表示，AI正朝着拥有认知和执行能力的智能体方向发展，下一个重点就是让AI从被动工具变为主动助手。可以预见，未来几年我们将看到各大公司竞相推出更完善的Agent AI产品，智能体将深入影响诸多行业实践。

本节按照时间线索回顾了智能体技术的发展演进。从概念萌芽时期的试探应用，到2010年代强化学习驱动的第一波专用智能体浪潮，再到近几年大模型驱动的通用代理新时代，智能体的侧重点经历了**从单任务走向多任务、从弱智能走向强认知**的演变。AlphaGo等案例证明了AI智能体在封闭领域可以超越人类，但其通用性不足；而ChatGPT等大模型代理则初步展现了跨领域对话和任务迁移能力，却在行动执行上尚待加强。近年来的种种代表模型（ChatGPT、Claude、GPT-4、Gemini、AutoGPT、Sora等）串联起技术演进脉络，也反映出智能体正从会“**回答**”发展到会“**操作**”。随着多模态感知和自主规划能力的加入，智能体正迈向更全面的智能代理。需要强调的是，当前所谓通用智能体距离真正自主智能仍有差距，其性能不稳定、长期记忆缺失等问题在第四章将详细讨论。但无论如何，智能体技术的演变已经驶入快车道，未来的智能体将可能成为无处不在的数字劳动力和智能助手。理解这一演进脉络有助于我们在下一章辨析当下智能体的真实水平与挑战所在，并为图书馆等领域抓住智能体应用机遇提供历史参照。

## **3.3 当前演进趋势与技术瓶颈**

### **3.3.1 智能体的发展趋势与方向**

权威机构的分析表明，智能体（Agentic AI）正成为未来几年最重要的AI发展方向之一。一系列趋势预示着智能体技术将进一步**深化能力**并**扩大应用**：

首先，**Agentic AI被列为重点技术趋势**。Gartner在2024-2025年的技术趋势报告中将“代理型AI”置于显著位置，认为越来越多的软件应用将内嵌智能体功能。从预测数据看，Gartner预计**到2028年，33%的企业软件应用程序将包含代理型AI（2024年这一比例不到1%）**[gartner.com](https://www.gartner.com/en/articles/intelligent-agent-in-ai#:~:text=Intelligent%20Agents%20in%20AI%20Really,day%20work)[reuters.com](https://www.reuters.com/business/over-40-agentic-ai-projects-will-be-scrapped-by-2027-gartner-says-2025-06-25/#:~:text=Gartner%20predicts%20at%20least%2015,in%202024)；且有至少15%的日常业务决策将在2028年通过智能体自动完成（2024年几乎为0%）[rcrwireless.com](https://www.rcrwireless.com/20250627/business/agentic-ai-gartner#:~:text=Gartner%3A%20More%20than%2040,virtually%20none%20in%202024)[reuters.com](https://www.reuters.com/business/over-40-agentic-ai-projects-will-be-scrapped-by-2027-gartner-says-2025-06-25/#:~:text=Gartner%20predicts%20at%20least%2015,in%202024)。这些预测显示了一个清晰方向：**智能体将快速从稀有试验走向主流应用**，在未来几年内渗透进大量企业流程和决策环节。Forrester亦在其2024年顶尖新兴技术榜单中，将AI智能体列为中期（2-5年内）能带来显著收益的技术之一[apmdigest.com](https://www.apmdigest.com/forrester-top-10-emerging-technologies-2024" \l ":~:text=five%20years)。这表明业界普遍看好智能体在提升自动化、优化用户体验方面的潜力。

其次，智能体的发展**战略重点**正在形成共识。McKinsey在2025年初发布的报告中指出，目前许多企业在部署生成式AI时遭遇“广铺不效”的困境，即广泛试用却未在财务业绩上见效，被称为“生成式AI悖论”[mckinsey.com](https://www.mckinsey.com/capabilities/quantumblack/our-insights/seizing-the-agentic-ai-advantage#:~:text=,the%20transformational%20promise%20of%20gen)。解决之道在于**将关注点从通用聊天助手转向垂直深耕的业务代理**[mckinsey.com](https://www.mckinsey.com/capabilities/quantumblack/our-insights/seizing-the-agentic-ai-advantage" \l ":~:text=,organizational%2C%20data%2C%20and%20cultural%20barriers)。横向的员工助手和聊天机器人虽部署容易但效益分散，而纵向的特定职能智能体由于技术、数据、组织壁垒难以落地。只有突破这些障碍，将AI代理**深度融入核心业务流程**，才能真正释放其价值[mckinsey.com](https://www.mckinsey.com/capabilities/quantumblack/our-insights/seizing-the-agentic-ai-advantage" \l ":~:text=,the%20transformational%20promise%20of%20gen)。这一观点与Forrester的观察一致：真正具竞争力的企业将开发**专门面向特定领域/流程的智能体**，而非仅依赖通用聊天GPT[cdotrends.com](https://www.cdotrends.com/story/4601/true-agentic-ai-needs-you-break-your-architecture#:~:text=While%20most%20organizations%20are%20still,this%20looks%20like%20in%20practice)[cdotrends.com](https://www.cdotrends.com/story/4601/true-agentic-ai-needs-you-break-your-architecture#:~:text=The%20bank%20built%20a%20specialized,tailored%20to%20specific%20client%20profiles)。例如，新加坡华侨银行（OCBC）为私人银行客户经理打造了理财顾问智能体，能分析投资组合并给出定制建议，比通用助手更具实用价值[cdotrends.com](https://www.cdotrends.com/story/4601/true-agentic-ai-needs-you-break-your-architecture" \l ":~:text=The%20bank%20built%20a%20specialized,tailored%20to%20specific%20client%20profiles)[cdotrends.com](https://www.cdotrends.com/story/4601/true-agentic-ai-needs-you-break-your-architecture#:~:text=OCBC%20also%20deployed%20vertical%20AI,specific%20workflows%2C%20not%20generic%20assistance)。类似地，银行还开发了合规监控Agent用于反洗钱审查，法律团队用AI助理加速合同审阅[cdotrends.com](https://www.cdotrends.com/story/4601/true-agentic-ai-needs-you-break-your-architecture" \l ":~:text=%E2%80%9CUsually%2C%20the%20relationship%20manager%20will,%E2%80%9D)。这些案例表明，**垂直领域智能体（Vertical AI Agents）**将是一大趋势，通过行业知识与AI结合来突破通用模型性能瓶颈[cdotrends.com](https://www.cdotrends.com/story/4601/true-agentic-ai-needs-you-break-your-architecture" \l ":~:text=This%20vertical%20approach%20reveals%20why,specialized%20intelligence%20breaks%20through%20it)。总之，发展方向正从“一个AI解决所有问题”转向“**大量专精小Agent协作**”，以求实现在各场景达到最佳效果。

再次，**人机协同和治理**被视为实现智能体价值的关键。McKinsey强调，企业需要从组织层面做好准备，将AI代理的引入视为业务**重塑**而非简单部署[mckinsey.com](https://www.mckinsey.com/capabilities/quantumblack/our-insights/seizing-the-agentic-ai-advantage" \l ":~:text=,functional)[mckinsey.com](https://www.mckinsey.com/capabilities/quantumblack/our-insights/seizing-the-agentic-ai-advantage#:~:text=,to%20bring%20the%20gen%20AI)。领导者必须重新思考决策流程、工作流、人机分工，在构建AI智能体同时**调整组织的操作模式**[mckinsey.com](https://www.mckinsey.com/capabilities/quantumblack/our-insights/seizing-the-agentic-ai-advantage" \l ":~:text=built%20and%20off,experimentation%20to%20industrialized%2C%20scalable%20delivery)。例如，将AI从辅助工具升级为主动代理后，企业需要设计如何让人类与代理共同决策、如何监督代理决策的正确性等。McKinsey提出要建立“**代理化时代的运行模式**”，包括跨职能的AI团队、敏捷的迭代开发以及全面的员工培训[mckinsey.com](https://www.mckinsey.com/capabilities/quantumblack/our-insights/seizing-the-agentic-ai-advantage" \l ":~:text=earning%20trust%2C%20driving%20adoption%2C%20and,upskill%20the%20workforce%2C%20adapt%20the)[mckinsey.com](https://www.mckinsey.com/capabilities/quantumblack/our-insights/seizing-the-agentic-ai-advantage#:~:text=delivery.%20,only%20the%20CEO%20can%20make)。Forrester则特别指出，目前大多数组织**低估了部署智能体的总成本**，除了模型和基础设施，还需投入大量资源于**运营和治理**（如模型评估、合规审查、人员培训等）[cdotrends.com](https://www.cdotrends.com/story/4601/true-agentic-ai-needs-you-break-your-architecture" \l ":~:text=Forrester%E2%80%99s%20research%20reveals%20that%20most,costs%20that%20companies%20can%E2%80%99t%20avoid)[cdotrends.com](https://www.cdotrends.com/story/4601/true-agentic-ai-needs-you-break-your-architecture#:~:text=%E2%80%9CWe%20see%20some%20kind%20of,dwarf%20the%20initial%20technology%20investment)。因此未来智能体的发展将更加强调**可持续的Ops架构和治理框架**，而不仅是模型能力本身。可以预见，一个融合**技术、流程、人的协同网络**（如Forrester提及的“自动化织物”架构[cdotrends.com](https://www.cdotrends.com/story/4601/true-agentic-ai-needs-you-break-your-architecture" \l ":~:text=Translation%3A%20your%20current%20architecture%20wasn%E2%80%99t,was%20built%20for%20solo%20performances)）将逐步成形，以支撑大规模智能体应用。

综上，当前智能体发展的总体趋势可以概括为：**“加速普及应用，聚焦专业场景，重塑组织流程”**。对图书馆等知识服务领域而言，这意味着未来可期望出现许多专门为图书馆业务设计的智能代理，并需要我们做好管理与协作的准备。这一趋势也昭示着智能体将从实验室走向现实，在各行各业深刻改变人与技术互动的方式[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00" \l ":~:text=%E4%BA%8B%E5%AE%9E%E4%B8%8A%EF%BC%8C%E4%B8%9A%E7%95%8C%E5%AF%B9%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%93%E7%9A%84%E7%A0%94%E7%A9%B6%E8%BF%98%E5%A4%84%E4%BA%8E%E5%88%9D%E6%9C%9F%E9%98%B6%E6%AE%B5%EF%BC%8C%E8%BF%98%E6%B2%A1%E6%9C%89%E5%AF%B9%E5%85%B6%E8%BF%9B%E8%A1%8C%E6%98%8E%E7%A1%AE%E7%9A%84%E5%AE%9A%E4%B9%89%E3%80%82%E2%80%9C%E7%AE%80%E5%8D%95%E5%9C%B0%E7%90%86%E8%A7%A3%EF%BC%8C%E5%AE%83%E4%BB%AC%E6%9C%AC%E8%B4%A8%E4%B8%8A%E5%B0%B1%E6%98%AF%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E6%A8%A1%E5%9E%8B%E5%92%8C%E7%AE%97%E6%B3%95%EF%BC%8C%E5%8F%AF%E4%BB%A5%E5%9C%A8%E5%8A%A8%E6%80%81%E4%B8%96%E7%95%8C%E4%B8%AD%E8%87%AA%E4%B8%BB%E5%81%9A%E5%87%BA%E5%86%B3%E7%AD%96%E3%80%82%E2%80%9D%E8%8B%B1%E4%BC%9F%E8%BE%BE%E9%AB%98%20%E7%BA%A7%E7%A0%94%E7%A9%B6%E7%A7%91%E5%AD%A6%E5%AE%B6%E3%80%81%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%93%E9%A1%B9%E7%9B%AE%E8%B4%9F%E8%B4%A3%E4%BA%BA%20Jim%20Fan%20%E8%A1%A8%E7%A4%BA%E3%80%82)[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00#:~:text=%E6%AF%94%E5%A6%82%EF%BC%8C%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%93%E5%8F%AF%E8%83%BD%E8%83%BD%E5%A4%9F%E5%85%85%E5%BD%93%E6%9B%B4%E5%A4%8D%E6%9D%82%E7%9A%84%E5%AE%A2%E6%88%B7%E6%9C%8D%E5%8A%A1%E6%9C%BA%E5%99%A8%E4%BA%BA%E3%80%82%E5%BD%93%E5%89%8D%E5%9F%BA%E4%BA%8E%E8%AF%AD%E8%A8%80%E6%A8%A1%E5%9E%8B%E7%9A%84%E5%8A%A9%E6%89%8B%E5%8F%AA%E8%83%BD%E7%94%9F%E6%88%90%E5%8F%A5%E5%AD%90%E4%B8%AD%E4%B8%8B%E4%B8%80%E4%B8%AA%E5%8F%AF%E8%83%BD%E7%9A%84%E5%8D%95%E8%AF%8D%E5%B9%B6%E7%BB%84%E6%88%90%E5%8F%A5%E5%AD%90%EF%BC%8C%E8%80%8C%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%93%E5%B0%86%E6%9C%89%E8%83%BD%E5%8A%9B%E8%87%AA%E4%B8%BB%E5%A4%84%E7%90%86%E8%87%AA%E7%84%B6%E8%AF%AD%E8%A8%80%E5%91%BD%E4%BB%A4%EF%BC%8C%E5%B9%B6%20%E4%B8%94%E5%9C%A8%E6%B2%A1%E6%9C%89%E7%9B%91%E7%9D%A3%E7%9A%84%E6%83%85%E5%86%B5%E4%B8%8B%E5%A4%84%E7%90%86%E5%AE%A2%E6%88%B7%E6%9C%8D%E5%8A%A1%E4%BB%BB%E5%8A%A1%E3%80%82%E5%86%8D%E6%AF%94%E5%A6%82%EF%BC%8C%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%93%E5%B0%86%E8%83%BD%E5%A4%9F%E5%88%86%E6%9E%90%E5%AE%A2%E6%88%B7%E7%9A%84%E6%8A%95%E8%AF%89%E7%94%B5%E5%AD%90%E9%82%AE%E4%BB%B6%EF%BC%8C%E7%84%B6%E5%90%8E%E7%9F%A5%E9%81%93%E5%A6%82%E4%BD%95%E6%A3%80%E6%9F%A5%E5%AE%A2%E6%88%B7%E7%9A%84%E5%8D%95%E5%8F%B7%EF%BC%8C%E8%AE%BF%E9%97%AE%E5%AE%A2%E6%88%B7%E5%85%B3%E7%B3%BB%E7%AE%A1%E7%90%86%E5%92%8C%E4%BA%A4%E4%BB%98%E7%B3%BB%E7%BB%9F%E7%AD%89%E6%95%B0%E6%8D%AE%E5%BA%93%EF%BC%8C%E4%BB%A5%E6%9F%A5%E7%9C%8B%E6%8A%95%E8%AF%89%20%E6%98%AF%E5%90%A6%E5%90%88%E6%B3%95%EF%BC%8C%E5%B9%B6%E6%A0%B9%E6%8D%AE%E5%85%AC%E5%8F%B8%E7%9A%84%E6%94%BF%E7%AD%96%E8%BF%9B%E8%A1%8C%E5%A4%84%E7%90%86%E3%80%82)。

### **3.3.2 场景应用挑战与行业观察**

尽管前景光明，但现实中智能体项目的推进并非一帆风顺，各类权威报告揭示了不少**应用落地的挑战**和潜在泡沫。根据Gartner在2025年中发布的研究，当前大多数智能体项目还处于**早期试验或概念验证**阶段，很多是受市场炒作驱动而并未找到清晰价值定位[reuters.com](https://www.reuters.com/business/over-40-agentic-ai-projects-will-be-scrapped-by-2027-gartner-says-2025-06-25/" \l ":~:text=,Senior%20Director%20Analyst%20at%20Gartner)。Gartner预言，**超过40%的代理型AI项目将在2027年底前被取消**，原因是不断攀升的成本和不明确的业务价值[reuters.com](https://www.reuters.com/business/over-40-agentic-ai-projects-will-be-scrapped-by-2027-gartner-says-2025-06-25/#:~:text=June%2025%20%28Reuters%29%20,to%20a%20report%20by%20Gartner)。这一高失败率警示业界需要更加审慎地选择应用场景，避免盲目追风投入。事实上，Gartner指出不少厂商存在“**Agent Washing**”行为——将原本功能有限的聊天机器人或流程自动化产品重新包装为“智能体”出售，但实际并无显著的新Agent能力[reuters.com](https://www.reuters.com/business/over-40-agentic-ai-projects-will-be-scrapped-by-2027-gartner-says-2025-06-25/" \l ":~:text=Many%20vendors%20are%20engaging%20in,agentic%20AI%20vendors%20are%20real)。大量这样的伪智能体充斥市场，反而造成用户困惑和期望落空。Gartner分析师Anushree Verma直言：“目前大多数代理型AI项目主要由炒作驱动，很多被错误应用在不合适场景”[reuters.com](https://www.reuters.com/business/over-40-agentic-ai-projects-will-be-scrapped-by-2027-gartner-says-2025-06-25/#:~:text=,Senior%20Director%20Analyst%20at%20Gartner)；且“当前模型尚不具备自主完成复杂业务目标或长期遵循复杂指令的成熟度”[reuters.com](https://www.reuters.com/business/over-40-agentic-ai-projects-will-be-scrapped-by-2027-gartner-says-2025-06-25/#:~:text=concept%20that%20are%20mostly%20driven,Senior%20Director%20Analyst%20at%20Gartner)。换言之，**智能体技术尚未完全成熟，价值兑现常遭高估**。

Forrester的调查同样表明了**认知与现实的落差**。他们发现，很多企业管理者对部署AI智能体所需投入和改变认识不足，往往低估了隐藏的运营成本[cdotrends.com](https://www.cdotrends.com/story/4601/true-agentic-ai-needs-you-break-your-architecture" \l ":~:text=Forrester%E2%80%99s%20research%20reveals%20that%20most,costs%20that%20companies%20can%E2%80%99t%20avoid)。比如，成功运行一个企业级智能体，不仅要有技术架构支撑，还需要持续的模型监控、数据更新、员工培训等。许多组织只预算了模型开发或API使用费用，却忽视了模型效果评估、安全监控、流程改造等开销[cdotrends.com](https://www.cdotrends.com/story/4601/true-agentic-ai-needs-you-break-your-architecture" \l ":~:text=%E2%80%9CWe%20see%20some%20kind%20of,dwarf%20the%20initial%20technology%20investment)。因此智能体一旦从实验转向生产，就面临**维护难、ROI不明**的问题。McKinsey的报告提到，有近80%的公司虽然尝试了生成式AI，但同等比例表示尚未看到显著财务回报[mckinsey.com](https://www.mckinsey.com/capabilities/quantumblack/our-insights/seizing-the-agentic-ai-advantage" \l ":~:text=,organizational%2C%20data%2C%20and%20cultural%20barriers)。这“生成式AI悖论”凸显了AI应用与业务成果间的**断链**。智能体要真正落地，必须打通技术与业务的最后一公里——找到能够**直接创造价值**的用例，并将之融入日常流程。然而目前不少机构仍在摸索有效的智能体场景，难免出现投入与产出不成正比的情况[reuters.com](https://www.reuters.com/business/over-40-agentic-ai-projects-will-be-scrapped-by-2027-gartner-says-2025-06-25/" \l ":~:text=,Senior%20Director%20Analyst%20at%20Gartner)。

此外，**用户信任和采用意愿**也是一大挑战。MIT Technology Review的报道指出，现阶段许多智能体系统还不够可靠，经常会出现错误甚至荒谬的行为，让人无法完全信任它们自主行动[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00" \l ":~:text=%E5%BD%93%E5%89%8D%E8%BF%98%E5%AD%98%E5%9C%A8%E5%93%AA%E4%BA%9B%E9%99%90%E5%88%B6%EF%BC%9F)。Imbue公司的CEO邱侃君将当前智能体状态比作十多年前的自动驾驶汽车：**功能有所展现，但远不够可靠，更谈不上真正自主**[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00" \l ":~:text=%E5%BD%93%E5%89%8D%E8%BF%98%E5%AD%98%E5%9C%A8%E5%93%AA%E4%BA%9B%E9%99%90%E5%88%B6%EF%BC%9F)。例如，一些编码智能体可以自动生成代码，但时常出错，且不会自行测试，仍需人类开发者介入[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00" \l ":~:text=%E7%84%B6%E8%80%8C%E7%8E%B0%E9%98%B6%E6%AE%B5%E4%BB%8D%E6%9C%89%E8%AE%B8%E5%A4%9A%E6%82%AC%E8%80%8C%E6%9C%AA%E5%86%B3%E7%9A%84%E9%97%AE%E9%A2%98%E9%9C%80%E8%A6%81%E5%9B%9E%E7%AD%94%E3%80%82%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E5%88%9D%E5%88%9B%E5%85%AC%E5%8F%B8%20Imbue%20%E7%9A%84%E9%A6%96%E5%B8%AD%E6%89%A7%E8%A1%8C%E5%AE%98%E5%85%BC%E5%88%9B%E5%A7%8B%E4%BA%BA%E9%82%B1%E4%BE%83%E5%90%9B%E8%87%B4%E5%8A%9B%E4%BA%8E%E5%BC%80%E5%8F%91%E8%83%BD%E5%A4%9F%E6%8E%A8%E7%90%86%E5%92%8C%E7%BC%96%E7%A0%81%E7%9A%84%E6%99%BA%E8%83%BD%E4%BD%93%EF%BC%8C%E5%A5%B9%E5%B0%86%E7%9B%AE%E5%89%8D%E6%99%BA%E8%83%BD%E4%BD%93%E7%9A%84%E7%8E%B0%E7%8A%B6%E4%B8%8E%E5%8D%81%E5%A4%9A%E5%B9%B4%20%E5%89%8D%E7%9A%84%E8%87%AA%E5%8A%A8%E9%A9%BE%E9%A9%B6%E6%B1%BD%E8%BD%A6%E7%9B%B8%E6%AF%94%E8%BE%83%E3%80%82%E5%9C%A8%E5%A5%B9%E7%9C%8B%E6%9D%A5%E7%9B%AE%E5%89%8D%E7%9A%84%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%93%E8%99%BD%E7%84%B6%E5%8F%AF%E4%BB%A5%E5%AE%8C%E6%88%90%E4%B8%80%E4%BA%9B%E4%BA%8B%E6%83%85%EF%BC%8C%E4%BD%86%E5%AE%83%E4%BB%AC%E5%B9%B6%E4%B8%8D%E5%8F%AF%E9%9D%A0%EF%BC%8C%E8%80%8C%E4%B8%94%E4%BB%8D%E7%84%B6%E6%B2%A1%E6%9C%89%E7%9C%9F%E6%AD%A3%E7%9A%84%E8%87%AA%E4%B8%BB%E6%80%A7%E3%80%82)。再如，OpenAI插件尝试让ChatGPT执行网页浏览等任务，但实践中**系统仍显笨拙、不可靠，缺乏真正推理能力**[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00" \l ":~:text=%E5%A5%B9%E8%A1%A8%E7%A4%BA%EF%BC%9A%E2%80%9C%E4%BB%8A%E5%A4%A9%E6%88%91%E4%BB%AC%E6%9C%89%E4%BA%86%E8%BF%99%E4%BA%9B%E8%AE%A1%E7%AE%97%E6%9C%BA%EF%BC%8C%E5%AE%83%E4%BB%AC%E7%9C%9F%E7%9A%84%E5%BE%88%E5%BC%BA%E5%A4%A7%EF%BC%8C%E4%BD%86%E6%88%91%E4%BB%AC%E5%BF%85%E9%A1%BB%E5%AF%B9%E5%AE%83%E4%BB%AC%E8%BF%9B%E8%A1%8C%E2%80%98%E5%BE%AE%E8%A7%82%E7%AE%A1%E7%90%86%E2%80%99%E3%80%82%E2%80%9D)。Jim Fan也坦言：“我们离那种能自动帮我们做所有家务的智能体还差得远。当前系统可能会产生幻觉，或者不严格遵守指令，这显然是麻烦的”[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00#:~:text=%E7%8E%B0%E6%8E%A8%E7%90%86%EF%BC%8C%E8%BF%99%E6%98%AF%E5%9C%A8%E5%A4%8D%E6%9D%82%E4%B8%94%E6%A8%A1%E7%B3%8A%E7%9A%84%E4%BA%BA%E7%B1%BB%E4%B8%96%E7%95%8C%E4%B8%AD%E8%BF%90%E8%A1%8C%E7%9A%84%E5%85%B3%E9%94%AE%E4%B8%80%E6%AD%A5%E3%80%82%E2%80%9D)。智能体偶尔的不守规矩和不准确会削弱用户对其信任。同时，一些用户对AI代理会如何处理他们的数据、是否会出错缺乏信心，因而不敢将关键任务托付给它们。这些因素都限制了智能体在敏感场景的应用。如果用户在使用中需要频繁核对和修正智能体输出，那么引入智能体的意义就大打折扣。行业报道指出，要让智能体获得广泛采用，必须**显著提高其可靠性、一致性和可控性**，否则用户宁可相信传统软件或人工。[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00" \l ":~:text=Jim%20Fan%20%E5%88%99%E8%A1%A8%E7%A4%BA%EF%BC%9A%E2%80%9C%E6%88%91%E4%BB%AC%E8%BF%98%E8%BF%9C%E8%BF%9C%E6%B2%A1%E6%9C%89%E4%B8%80%E4%B8%AA%E5%8F%AF%E4%BB%A5%E4%B8%BA%E6%88%91%E4%BB%AC%E8%87%AA%E5%8A%A8%E5%8C%96%E6%89%80%E6%9C%89%E8%BF%99%E4%BA%9B%E5%AE%B6%E5%8A%A1%E7%9A%84%E6%99%BA%E8%83%BD%E4%BD%93%E3%80%82%E7%9B%AE%E5%89%8D%E7%9A%84%E7%B3%BB%E7%BB%9F%E5%8F%AF%E8%83%BD%E4%BC%9A%E4%BA%A7%E7%94%9F%E5%B9%BB%E8%A7%89%EF%BC%8C%E8%80%8C%E4%B8%94%E5%AE%83%E4%BB%AC%E4%B9%9F%E4%B8%8D%E6%80%BB%E6%98%AF%E4%B8%A5%E6%A0%BC%E9%81%B5%E5%BE%AA%E6%8C%87%E4%BB%A4%EF%BC%8C%E6%98%BE%E7%84%B6%E8%BF%99%E5%B0%B1%E5%BE%88%E9%BA%BB%E7%83%A6%E4%BA%86%E3%80%82%E2%80%9D)

总结来说，目前智能体应用面临的挑战包括：**价值错配**（场景选择不当、ROI不明显）、**成本低估**（运维投入被忽视）、**性能不稳**（容易出错和走题）以及**信任不足**（用户采纳意愿不高）。这些问题导致不少试点项目夭折或止步于概念验证。对策方面，行业专家建议采取“小步快跑、务实聚焦”的方法：从细分场景切入，先做出几个可靠的“小智能体”产生实际效益，再逐步扩展规模和领域。同时，在推进过程中要透明评估效果，与用户密切沟通反馈，逐步建立对AI代理的信任。在图书馆领域，我们也应汲取这些经验教训，理性看待智能体的应用边界，既避免一哄而上的炒作，又积极寻找真正能提升服务质量和效率的契合点。

### **3.3.3 算力资源与成本限制**

智能体技术的发展高度依赖**算力和数据**支撑，无论训练大模型还是部署推理，成本之高都对技术演进构成现实瓶颈。当前，大型语言模型背后的算力需求呈指数级增长，OpenAI等公司为训练和运行智能体投入了空前的硬件资源。例如，据报道GPT-4的训练使用了成千上万块GPU，花费数千万美元级别。智能体在推理阶段如果需要长时间自主运行，也会不断消耗算力和内存。Sam Altman曾表示，ChatGPT上线后服务器“被挤爆”，推理需求的激增令算力供给捉襟见肘[yicai.com](https://www.yicai.com/news/102678941.html" \l ":~:text=GPT)。OpenAI为此启动了规模高达5000亿美元的“Stargate”计划，投资自建数据中心以确保未来算力供应[yicai.com](https://www.yicai.com/news/102678941.html" \l ":~:text=%E6%AD%A4%E5%89%8D%E8%B0%88%E5%8F%8AStargate%E8%AE%BE%E7%AB%8B5000%E4%BA%BF%E7%BE%8E%E5%85%83%E8%A7%84%E6%A8%A1%E5%8E%9F%E5%9B%A0%E6%97%B6%EF%BC%8C%E5%B1%B1%E5%A7%86%C2%B7%E5%A5%A5%E5%B0%94%E7%89%B9%E6%9B%BC%E8%A1%A8%E7%A4%BA%EF%BC%8C%E5%9B%A0%E4%B8%BAGPT)[yicai.com](https://www.yicai.com/news/102678941.html#:~:text=%E9%99%A4%E4%BA%86Agent%E5%AF%B9%E7%AE%97%E5%8A%9B%E7%9A%84%E6%8B%89%E5%8A%A8%EF%BC%8C%E5%90%84%E5%AE%B6%E4%BA%92%E8%81%94%E7%BD%91%E5%A4%A7%E5%8E%82%E5%8A%A0%E9%80%9F%E6%8E%A8%E5%8A%A8AI%E4%B8%8E%E5%8E%9F%E6%9C%89%E4%B8%9A%E5%8A%A1%E7%BB%93%E5%90%88%EF%BC%8C%E4%BB%A5%E5%8F%8A%E5%A4%9A%E6%A8%A1%E6%80%81%E6%8A%80%E6%9C%AF%E7%9A%84%E5%8F%91%E5%B1%95%EF%BC%8C%E4%B9%9F%E6%98%AF%E6%8B%89%E5%8A%A8%E7%AE%97%E5%8A%9B%E9%9C%80%E6%B1%82%E5%A2%9E%E9%95%BF%E7%9A%84%E9%87%8D%E8%A6%81%E5%9B%A0%E7%B4%A0%E3%80%82%E5%A4%9A%E6%96%B9%E5%BD%B1%E5%93%8D%E4%B8%8B%EF%BC%8C%E7%AE%97%E5%8A%9B%E5%9F%BA%E7%A1%80%E8%AE%BE%E6%96%BD%E7%9A%84%E6%8A%95%E5%85%A5%E8%A7%84%E6%A8%A1%E9%80%90%20%E6%B8%90%E6%89%A9%E5%A4%A7%E3%80%82)。可见**算力瓶颈已成为阻碍智能体进一步发展的关键因素**：训练更智能的Agent需要更大更贵的模型，推理部署更广泛的Agent需要成倍增加的硬件投入。

有分析指出，**智能体=大模型 + 工具/环境交互**，这意味着除了模型自身的算力消耗，还需考虑工具使用和环境模拟的开销。例如，一个AutoGPT代理在执行任务时可能频繁访问搜索引擎、调用数据库，每一步都带来额外的API或计算成本。当多个智能体并发运行或多轮循环时，成本会进一步放大。Gartner预计，未来几年智能体项目会经历从兴趣高涨到理性回调，其中一个原因正是**成本失控**。他们估计，到2027年，超过四成智能体项目会被放弃，部分因为成本投入远超预期[reuters.com](https://www.reuters.com/business/over-40-agentic-ai-projects-will-be-scrapped-by-2027-gartner-says-2025-06-25/" \l ":~:text=June%2025%20%28Reuters%29%20,to%20a%20report%20by%20Gartner)。Forrester的研究也揭示，大部分组织低估了智能体的总拥有成本，包括软硬件投入以及运营成本[cdotrends.com](https://www.cdotrends.com/story/4601/true-agentic-ai-needs-you-break-your-architecture#:~:text=Forrester%E2%80%99s%20research%20reveals%20that%20most,costs%20that%20companies%20can%E2%80%99t%20avoid)。当智能体需要全天候在线、实时响应时，算力成本将成为持续支出，必须有明确效益支撑方可持续。

除了硬件费用，**能源消耗和碳足迹**也是值得关注的隐性成本。大型AI模型的训练和推理耗电量巨大，据估算训练一次GPT-3的碳排放相当于一辆汽车一生的排放量。随着代理型AI扩展应用，其能源需求可能对环境带来压力。因此业界开始探索更**高效的模型架构**和**加速芯片**，以降低单次推理的能耗。也有人提出是否可以通过**模型压缩、分布式协作**来实现“小模型也能部分代理”的模式，以减少对超大模型的依赖。这些技术路线仍在探索中。

算力限制还体现在**智能体的上下文记忆**方面。当下大模型的上下文窗口有限（GPT-4大约支持数万字），这使得智能体一旦任务链条过长就容易遗忘前面的内容[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00" \l ":~:text=Jim%20Fan%20%E5%88%99%E8%A1%A8%E7%A4%BA%EF%BC%9A%E2%80%9C%E6%88%91%E4%BB%AC%E8%BF%98%E8%BF%9C%E8%BF%9C%E6%B2%A1%E6%9C%89%E4%B8%80%E4%B8%AA%E5%8F%AF%E4%BB%A5%E4%B8%BA%E6%88%91%E4%BB%AC%E8%87%AA%E5%8A%A8%E5%8C%96%E6%89%80%E6%9C%89%E8%BF%99%E4%BA%9B%E5%AE%B6%E5%8A%A1%E7%9A%84%E6%99%BA%E8%83%BD%E4%BD%93%E3%80%82%E7%9B%AE%E5%89%8D%E7%9A%84%E7%B3%BB%E7%BB%9F%E5%8F%AF%E8%83%BD%E4%BC%9A%E4%BA%A7%E7%94%9F%E5%B9%BB%E8%A7%89%EF%BC%8C%E8%80%8C%E4%B8%94%E5%AE%83%E4%BB%AC%E4%B9%9F%E4%B8%8D%E6%80%BB%E6%98%AF%E4%B8%A5%E6%A0%BC%E9%81%B5%E5%BE%AA%E6%8C%87%E4%BB%A4%EF%BC%8C%E6%98%BE%E7%84%B6%E8%BF%99%E5%B0%B1%E5%BE%88%E9%BA%BB%E7%83%A6%E4%BA%86%E3%80%82%E2%80%9D)。为解决此问题，谷歌等正在努力扩展模型的上下文长度，甚至设想未来实现“无限上下文窗口”[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00#:~:text=%E2%80%9CChatGPT%20%E5%8F%AF%E4%BB%A5%E6%95%B2%E5%86%99%E4%BB%A3%E7%A0%81%EF%BC%8C%E4%BD%86%E4%B8%8D%E8%83%BD%E5%BE%88%E5%A5%BD%E5%9C%B0%E5%A4%84%E7%90%86%E8%BE%83%E9%95%BF%E7%9A%84%E5%86%85%E5%AE%B9%E3%80%82%E4%BD%86%E5%AF%B9%E4%BA%8E%E4%BA%BA%E7%B1%BB%E5%BC%80%E5%8F%91%E4%BA%BA%E5%91%98%E6%9D%A5%E8%AF%B4%E8%A6%81%E6%9F%A5%E9%98%85%E7%9A%84%E6%98%AF%E6%95%B4%E4%B8%AA%20GitHub%20%E4%BB%A3%E7%A0%81%E5%BA%93%EF%BC%8C%E5%85%B6%E4%B8%AD%E6%9C%89%E5%87%A0%E5%8D%81%E8%A1%8C%E4%B9%83%E8%87%B3%E6%95%B0%E7%99%BE%E8%A1%8C%E4%BB%A3%E7%A0%81%EF%BC%8C%E6%98%BE%E7%84%B6%EF%BC%8C%E8%BF%99%E5%AF%B9%E4%BA%8E%E4%BA%BA%E7%B1%BB%E8%80%8C%E8%A8%80%E5%AE%8C%E5%85%A8%E6%B2%A1%E6%9C%89%E4%BB%BB%E4%BD%95%E5%8E%8B%E5%8A%9B%E3%80%82%E2%80%9DJim%20Fan,%E8%AF%B4%E3%80%82)。然而超长上下文也意味着指数增长的计算开销，模型需要处理海量的序列数据才能“记住”长期信息，这在现有架构下难以高效实现。因此，有研究转向借助**外部向量数据库**和记忆模块，将较早的信息检索式地供模型调用，以缓解窗口限制。不过这些方案增加了系统复杂度，也带来了新的误差传播风险。

综合来看，**算力与成本壁垒**将在相当长时期内塑造智能体技术的演进路径。大型科技公司凭借资金和算力优势，将继续引领最前沿的大模型智能体开发；而普通组织和开源社群则可能更加注重通过优化和创新，寻求“以小搏大”的方案，如结合专家模型、提升推理效率等。在图书馆等公共机构领域，引入智能体必须充分考虑算力资源：是利用云端大厂的API服务，还是自建本地模型，都需权衡成本、数据安全和性能需求。未来，随着硬件进步和算法改进，单次推理成本有望下降，但如果应用规模暴增，总成本仍会庞大。因此，在拥抱智能体技术时，制定合理的预算和ROI预期，以及探索低成本高效用的方案（如部分任务用小模型代理，大模型只处理关键环节），将是务实之举。

### **3.3.4 评估机制与性能衡量**

评价一个AI智能体的性能和智能水准，是当前的一大难题。传统AI系统（如分类器、对话模型）有明确的评测基准和指标，而智能体往往是**复杂序列决策**，其成功与否难以用单一数值衡量。这导致目前缺乏行业公认的智能体评估机制，各方更多以**定性描述**或**案例展示**来说明智能体能力。Forrester的研究强调了**持续评估**的重要性：他们建议将模型评价融入智能体治理框架中，从部署第一天就开始定期检测AI行为表现[cdotrends.com](https://www.cdotrends.com/story/4601/true-agentic-ai-needs-you-break-your-architecture" \l ":~:text=The%20most%20mature%20organizations%20are,of%20corporate%20values%20and%20ethics)。例如，可以建立多阶段评估：首先由数据科学团队在模拟环境测试智能体完成任务的正确率、效率，其次在有限真实环境观察其决策质量，最后在上线后通过日志分析和用户反馈持续改进[cdotrends.com](https://www.cdotrends.com/story/4601/true-agentic-ai-needs-you-break-your-architecture" \l ":~:text=The%20most%20mature%20organizations%20are,of%20corporate%20values%20and%20ethics)[cdotrends.com](https://www.cdotrends.com/story/4601/true-agentic-ai-needs-you-break-your-architecture#:~:text=%E2%80%9CYou%20cannot%20only%20have%20your,ethics%2C%20and%20responsibilities%2C%E2%80%9D%20he%20explains)。这表明对智能体的评估需要一个**动态的、全生命周期**的过程，而非一锤定音。

具体评估指标方面，业内初步探索了一些方向：

* **任务成功率与收益：** 对于特定目标导向的智能体，可统计其完成任务的频率和质量。例如购物代理成功下单率，导航代理找到最佳路线的比例等。如果能量化任务收益（如用时、成本节约等），也可作为度量指标。
* **决策合理性与鲁棒性：** 分析智能体在各种环境情况下的决策是否合理、一致。当环境发生变化或遇异常，智能体是否依然表现稳健。可通过模拟环境扰动，观测智能体策略变化来评估鲁棒性。
* **资源效率：** 衡量智能体为达成目标所耗费的时间、算力、金钱等资源。一个高效的智能体应以较小代价完成任务。特别在实际应用中，速度和成本都是重要考虑。
* **服从约束度：** 检查智能体是否遵守预设的约束和指令，不偏离人类意图。这可通过设置守则（如不得访问某类数据）然后测试AI在各种诱因下是否违反。
* **用户满意度：** 最终，用户对智能体行为的感受很关键。通过用户调查或反馈打分，评估AI代理的有用性、易用性、可信度等。哪怕客观指标优秀，如果用户体验不佳，智能体也难成功。

然而，要为智能体建立统一评测基准仍有挑战。不同应用场景差异巨大，很难制定通用任务让各类智能体在同一起跑线上比较。目前学界有一些尝试，如使用游戏环境（Atari、Minecraft等）作为标准平台测试智能体的通用决策能力，但这些离真实业务场景尚有距离。另一个问题是**智能体行为的不可预测性**——它可能完成任务的路径与人类截然不同，评估时不能仅以人类预期步骤为准，否则会低估AI创新解法的价值。为此，有研究者提出或许应开发**全新的评估指标**，例如衡量智能体在开放环境中的探索效率、学会新技能的速度等等，这些比传统分类准确率更复杂的概念。

在企业实践中，一个可行的方法是**结合人工审查**与**自动指标**。McKinsey建议将“人类在环”（Human-in-the-loop）纳入治理，在关键决策点由人类对智能体输出进行评判，以确保符合预期[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00#:~:text=%E7%AD%89%E5%B7%A5%E4%BD%9C%E6%B5%81%E8%87%AA%E5%8A%A8%E5%8C%96%E8%BD%AF%E4%BB%B6%E7%AD%89%E7%AD%89%EF%BC%8C%E8%BF%99%E4%BA%9B%E4%B8%8E%E8%83%BD%E5%A4%9F%E5%AE%8C%E6%88%90%E5%A4%8D%E6%9D%82%E4%BB%BB%E5%8A%A1%E7%9A%84%E9%80%9A%E7%94%A8%E5%9E%8B%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%93%E7%9B%B8%E6%AF%94%EF%BC%8C%E8%BF%98%E5%B7%AE%E5%BE%97%E5%BE%88%E8%BF%9C%E3%80%82%E2%80%9D%E5%A5%B9%E8%A1%A5%E5%85%85%E8%AF%B4%E3%80%82)[cdotrends.com](https://www.cdotrends.com/story/4601/true-agentic-ai-needs-you-break-your-architecture#:~:text=The%20most%20mature%20organizations%20are,of%20corporate%20values%20and%20ethics)。同时，通过自动日志分析提取上述量化指标，长期跟踪AI绩效的变化趋势。随着更多智能体落地，各行业也会积累起评估经验，逐步形成领域性的标准。比如，客服智能体或许会建立一套行业标准指标（如问题一次解决率、客服满意度提升等），而工业控制智能体则有另一套指标体系。

对于图书馆知识服务场景，我们同样需要探索有效的评估方式。例如，评估一个图书馆问答智能体，可结合**准确性**（回答正确率）、**覆盖率**（能回答的问题类别占比）、**响应速度**、**读者满意度**等指标。对于个性化推荐智能体，可以看**点击率**、**借阅转化率**以及用户评分等。总之，应针对具体服务目标设定多维度的评估指标，并不断根据实践反馈进行调整优化。

### **3.3.5 控制、安全与治理**

随着智能体变得更自主和强大，对其**安全可控**的担忧也日益增加。如果一个AI代理在执行任务过程中不按照人类预期行动，甚至造成损害，其后果可能比传统AI错误更严重（因为其有行动能力）。因此，**控制与治理**问题是当前智能体技术必须正视的关键瓶颈。

一个基本的安全要求是**防止智能体“走偏”**。MIT Tech Review提到，目前智能体常见的问题包括产生**幻觉**（hallucination，即凭空编造不真实信息）以及**不严格遵守指令**[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00#:~:text=%E7%8E%B0%E6%8E%A8%E7%90%86%EF%BC%8C%E8%BF%99%E6%98%AF%E5%9C%A8%E5%A4%8D%E6%9D%82%E4%B8%94%E6%A8%A1%E7%B3%8A%E7%9A%84%E4%BA%BA%E7%B1%BB%E4%B8%96%E7%95%8C%E4%B8%AD%E8%BF%90%E8%A1%8C%E7%9A%84%E5%85%B3%E9%94%AE%E4%B8%80%E6%AD%A5%E3%80%82%E2%80%9D)。例如某购物代理可能因为误判而下错订单，或一个任务执行Agent因为误解析指令走入无关流程。这类偏差在智能体的自主循环中可能被放大，酿成更大错误。为此，开发者需要在智能体决策链中加入**约束机制**。比如设置每步行动的检查规则，或要求智能体在重要决定前征得人类确认（半自主模式）。Anthropic提出“宪法AI”思路，让AI遵循一系列预先制定的原则来自我指导，减少产生有害行为的可能[woshipm.com](https://www.woshipm.com/ai/6202918.html" \l ":~:text=%C2%B7%20Anthropic%EF%BC%9ALLMs%20%E5%8A%A8%E6%80%81%E6%8C%87%E5%AF%BC%E8%87%AA%E5%B7%B1%E7%9A%84%E6%B5%81%E7%A8%8B%E5%92%8C%20)。OpenAI也为GPT-4加入了系统层面的行为准则，以尽量防止其执行危险操作。然即便如此，完全杜绝不当行为仍很困难——因为智能体面对的开放环境中情况千变万化，规则不可能穷尽。这需要**持续的监控**和**迭代的治理**来完善。

另一个层面的控制是**权限管理**。智能体往往需要访问各种数据和系统完成任务，如果不加限制，可能造成敏感信息泄露或违规操作。因此必须严格控制智能体能获取的资源和能执行的操作。例如，企业可能给AI代理设定访问级别，只允许读取某些公共数据库而不触及机密文件；或对其能调用的工具进行白名单限制[ibm.com](https://www.ibm.com/cn-zh/think/topics/ai-agents" \l ":~:text=,%E4%B8%BA%20%C2%A0AI%20%E6%99%BA%E8%83%BD%E4%BD%93%C2%A0%E6%8F%90%E4%BE%9B%E8%A6%81%E5%AE%8C%E6%88%90%E7%9A%84%E5%85%B7%E4%BD%93%E7%9B%AE%E6%A0%87%E5%B9%B6%E5%BB%BA%E7%AB%8B%E8%A6%81%E4%BD%BF%E7%94%A8%E7%9A%84%E5%8F%AF%E7%94%A8%E5%B7%A5%E5%85%B7%E7%9A%84%E7%94%A8%E6%88%B7%E3%80%82)。在图书馆场景，也需要确保智能体仅在授权范围内使用读者数据、馆藏内容，不侵犯用户隐私或版权。技术上，可以通过**沙盒环境**运行智能体，把它的活动限制在一个隔离的虚拟空间中，即使出错也不危及真实系统。此外，引入**行为审计**机制，记录智能体每一步动作以供事后审查和追责，也非常必要。

治理智能体还涉及**伦理和价值观**的嵌入。McKinsey报告强调，企业在培养AI团队时，不仅要有技术能力，还要让他们充分理解公司的价值观和道德标准[cdotrends.com](https://www.cdotrends.com/story/4601/true-agentic-ai-needs-you-break-your-architecture" \l ":~:text=The%20most%20mature%20organizations%20are,of%20corporate%20values%20and%20ethics)。只有这样，这些原则才能在AI系统设计时就得到贯彻。最终目标是实现“**负责任的智能体**”：其决策不仅考虑效率，也符合人类的伦理规范。为此，组织需要建立专门的**AI治理委员会**或相应流程，将治理贯穿智能体开发、部署的全周期[cdotrends.com](https://www.cdotrends.com/story/4601/true-agentic-ai-needs-you-break-your-architecture" \l ":~:text=The%20most%20mature%20organizations%20are,of%20corporate%20values%20and%20ethics)[cdotrends.com](https://www.cdotrends.com/story/4601/true-agentic-ai-needs-you-break-your-architecture#:~:text=This%20represents%20a%20fundamental%20shift,that%20span%20the%20entire%20organization)。成熟的做法例如：在方案设计阶段进行伦理影响评估，开发阶段进行偏见检测和消歧，部署前通过红队测试发现潜在滥用风险，上线后持续监控和反馈纠正[cdotrends.com](https://www.cdotrends.com/story/4601/true-agentic-ai-needs-you-break-your-architecture" \l ":~:text=The%20most%20mature%20organizations%20are,of%20corporate%20values%20and%20ethics)。可以借鉴近年来AI伦理对话中的很多原则，例如透明性、公平性、问责制等应用到智能体治理上。

值得一提的是，**人机边界**也是治理的重点。Forrester研究员刘孟提出，要让智能体真正发挥作用，企业需要培养“AI翻译官”这样的角色，充当业务和技术之间的桥梁[cdotrends.com](https://www.cdotrends.com/story/4601/true-agentic-ai-needs-you-break-your-architecture#:~:text=,gap)[cdotrends.com](https://www.cdotrends.com/story/4601/true-agentic-ai-needs-you-break-your-architecture#:~:text=companies%2C%E2%80%9D%20Liu%20explains,%E2%80%9D)。这些人了解业务需求又懂AI，实现对智能体行为的监督调整。McKinsey也提出“人类作为共同架构师”的理念，让业务专家深度参与智能体的设计和改进[mckinsey.com](https://www.mckinsey.com/capabilities/quantumblack/our-insights/seizing-the-agentic-ai-advantage#:~:text=to%20deploy%20agents%20in%20ways,be%20living%20and%20working%20with)[mckinsey.com](https://www.mckinsey.com/capabilities/quantumblack/our-insights/seizing-the-agentic-ai-advantage#:~:text=Redefining%20how%20decisions%20are%20made%2C,delivers%20results%20rather%20than%20regrets)。通过这种交互，人类对AI保持影响力，AI也在安全边界内运作。

总体来说，**没有安全就没有智能**。代理型AI越强大，其治理就需越谨慎。当前没有所谓“完全自主且绝对安全”的智能体，大部分成功案例都是**在强人为监督下运行的**。例如，微软给GPT-4接入其产品时，仍设置了诸多人为审核流程，以防AI在联网搜索等环节出问题。业界共识是，在可预见的未来，“人监控AI、AI辅助人”将是主要模式，而非完全无人监管的AI自治[cdotrends.com](https://www.cdotrends.com/story/4601/true-agentic-ai-needs-you-break-your-architecture" \l ":~:text=Despite%20the%20hype%20and%20the,but%20far%20from%20fully%20autonomous)。因此，对图书馆等应用者而言，引入智能体也必须同步建立相应的**使用规程和风险预案**，确保AI助手始终在馆员可控范围内活动。在注重创新的同时，守住安全与伦理底线，是智能体融入知识服务必须遵循的基本准则。

本节围绕当前智能体技术的发展趋势和瓶颈挑战展开讨论。从趋势看，智能体正快速从概念验证走向实际部署，行业预测其将在未来几年广泛嵌入企业软件和决策流程[reuters.com](https://www.reuters.com/business/over-40-agentic-ai-projects-will-be-scrapped-by-2027-gartner-says-2025-06-25/" \l ":~:text=Gartner%20predicts%20at%20least%2015,in%202024)。但落地过程中也出现了明显的**泡沫和挑战**：应用场景选择不当导致价值不彰、成本与算力压力制约大规模推广、缺乏成熟评估机制使性能难以衡量、以及自主性带来的安全不可控风险凸显。权威报告的警示让我们认识到，**智能体技术虽炙手可热，但仍处于早期阶段**，对其能力既不可捧上神坛，也不应一棒打死。Gartner等机构预计，一些盲目跟风的智能体项目将大批失败[reuters.com](https://www.reuters.com/business/over-40-agentic-ai-projects-will-be-scrapped-by-2027-gartner-says-2025-06-25/" \l ":~:text=June%2025%20%28Reuters%29%20,to%20a%20report%20by%20Gartner)，而真正成功的应用需要踏实选准场景、投入资源构建支撑体系，并采取严谨的治理措施。智能体的未来演进将更多聚焦于**行业纵深应用**（如定制领域代理）、**提高可靠性**和**加强人机协同治理**。对于图书馆领域，理解这些趋势与瓶颈有助于我们在引入智能体时避免重蹈覆辙——既看到其提升知识服务的巨大潜力，也充分考虑现实限制和风险防范，以制定稳健的发展策略。在下一章中，我们将结合这些认识，具体探讨智能体与图书馆知识服务融合的可能路径和实践经验。

## **3.4 智能体与图书馆知识服务融合应用**

图书馆作为知识获取与传播的枢纽，正积极寻求人工智能技术赋能以提升服务水平。智能体（Agent/代理型AI）的出现为图书馆实现“智慧服务”提供了新契机。以下我们按照知识服务的六大核心场景，分析智能体技术的融合应用前景，每个场景并辅以国内外案例说明其实际进展、技术成熟度、成本考量和用户体验效果。

### **3.4.1 知识图谱：结构化知识组织与发现**

**场景定义：** 知识图谱（Knowledge Graph）是通过语义关联将知识要素（概念、实体、关系）结构化呈现的一种技术与工具。图书馆可利用知识图谱对馆藏资源和领域知识进行**语义组织**和**可视化表示**，帮助用户直观地探索知识脉络[[94]](#footnote-94) [[95]](#footnote-95) 。代理型AI可以在知识图谱构建和应用中发挥两方面作用：一是**构建阶段**，利用大模型的文本理解能力抽取实体关系、补全图谱；二是**服务阶段**，充当“知识图谱导航员”，根据用户需求在图谱中寻找相关节点、揭示关联，为用户提供**交互式的知识发现**体验。

**案例与实践：** 国内外已有多项探索将智能体与知识图谱结合，丰富图书馆知识服务：

* *国际案例：学术知识图谱助理（Yewno等）☆* – 美国一些大学图书馆引入了Yewno Discover系统，它以AI驱动的知识图谱为用户提供研究探索服务[[96]](#footnote-96) 。例如加州大学河滨分校图书馆部署了Yewno Discover，用户输入一个概念后，系统会生成与之相关的**知识图谱**：中心节点是查询概念，周围节点代表相关子概念或主题[[97]](#footnote-97) 。用户点击任一节点，可浏览概念简介、其关联概念和相关文献[[98]](#footnote-98) 。这种视觉知识网络帮助用户**发散思维**、发现原本未知的关联，大幅缩短资料调研时间[[99]](#footnote-99) 。据馆员反馈，不同于传统线性检索，知识图谱方式更符合探索式学习习惯，让部分用户感到信息获取过程“更有趣高效”[[100]](#footnote-100) [[101]](#footnote-101) 。Yewno背后的AI会从维基百科等权威资料中提取概念关系，不断**扩充更新图谱**[[102]](#footnote-102) 。本例展示了AI智能体辅助知识图谱应用的价值：自动构建并交互呈现知识网络，使用户以**直观地图**形式把握一个领域的整体结构、核心概念和前沿课题[[103]](#footnote-103) 。
* *国内案例：古籍知识服务智能体（深圳图书馆）☆* – 深圳图书馆在智慧图书馆建设中开发了“**方志里的深圳**”微信小程序，以智能技术重新解读清嘉庆年间《新安县志》（一部深圳地方历史文献）[[104]](#footnote-104) 。项目团队将这部两百年前的方志内容数字化后，采用**知识图谱**对人名、地名、物产、制度等进行梳理关联，并辅以生动的交互界面。读者可以在小程序中看到一个围绕“深圳古代生活”的知识网络图，每个节点对应志书中的一个主题（如某位名人、某处古迹），点击节点可查看该主题在志书中的记载，并浏览与其相关的其他节点[[105]](#footnote-105) 。整个体验如同沉浸在一个可视化的史料知识宇宙中，原本枯燥厚重的典籍以**生动、系统**的知识图谱形式重现[[106]](#footnote-106) 。这项服务极大地激发了读者了解本地历史的兴趣，被誉为传统文化活化的成功探索。其背后用到了智能关系抽取技术，将方志文本中的关键实体自动识别并构建关联；同时，智能界面充当交互代理，指引读者在知识图谱中导航。广东省文化厅报道评价该项目“让读者沉浸式了解两百多年前深圳地区的源流沿革”，将零散的历史知识串联成**有机网络**[whly.gd.gov.cn](https://whly.gd.gov.cn/service_newggwhzx/content/post_4061139.html" \l ":~:text=%E7%AB%99%E5%9B%BE%E4%B9%A6%E8%B5%84%E6%BA%90%2C%E8%BF%98%E5%8F%AF%E8%8E%B7%E5%8F%96%E5%85%AC%E5%85%B1%E5%9B%BE%E4%B9%A6%E9%A6%86%E8%97%8F%E4%B9%A6%E6%8C%87%E5%BC%95%2C%E5%AE%9E%E7%8E%B0%E4%BA%86%E5%85%AC%E7%A7%81%E8%97%8F%E4%B9%A6%E7%9A%84%E8%9E%8D%E5%90%88%E5%88%A9%E7%94%A8%2C%E6%8F%90%E9%AB%98%E4%BA%86%E5%9B%BE%E4%B9%A6%E8%8E%B7%E5%8F%96%E6%95%88%E7%8E%87%2C%E6%BB%A1%E8%B6%B3%E4%BA%86%E5%B8%82%E6%B0%91%E9%98%85%E8%AF%BB%E9%9C%80%E6%B1%82%E3%80%82)。该案例表明，智能体对知识图谱的应用不局限于科技领域，对人文古籍等也有显著价值，可**助力中华优秀传统文化的创新传播**[whly.gd.gov.cn](https://whly.gd.gov.cn/service_newggwhzx/content/post_4061139.html" \l ":~:text=%E6%B7%B1%E5%9C%B3%E5%9B%BE%E4%B9%A6%E9%A6%86%E2%80%9C%E6%96%B9%E5%BF%97%E9%87%8C%E7%9A%84%E6%B7%B1%E5%9C%B3%E2%80%9D%E5%BE%AE%E4%BF%A1%E5%B0%8F%E7%A8%8B%E5%BA%8F%E5%AF%B9%E6%B8%85%E5%98%89%E5%BA%86%E3%80%8A%E6%96%B0%E5%AE%89%E5%8E%BF%E5%BF%97%E3%80%8B%E8%BF%9B%E8%A1%8C%E6%99%BA%E6%85%A7%E5%8C%96%E5%BC%80%E5%8F%91%EF%BC%8C%E7%94%A8%E7%94%9F%E5%8A%A8%E3%80%81%E7%B3%BB%E7%BB%9F%E7%9A%84%E7%9F%A5%E8%AF%86%E5%9B%BE%E8%B0%B1%E4%B8%8E%E4%B8%B0%E5%AF%8C%E3%80%81%E8%B6%A3%E5%91%B3%E7%9A%84%E4%BA%92%E5%8A%A8%E7%95%8C%E9%9D%A2%E9%87%8D%E7%8E%B0%E5%8E%9A%E9%87%8D%E7%9A%84%E5%85%B8%E7%B1%8D%2C%E8%AE%A9%E8%AF%BB%E8%80%85%E6%B2%89%E6%B5%B8%E5%BC%8F%E4%BA%86%E8%A7%A3%E4%B8%A4%E7%99%BE%E5%A4%9A%E5%B9%B4%E5%89%8D%E6%B7%B1%20%E5%9C%B3%E5%9C%B0%E5%8C%BA%E7%9A%84%E6%BA%90%E6%B5%81%E6%B2%BF%E9%9D%A9%E3%80%81%E9%A3%8E%E4%BF%97%E7%89%A9%E4%BA%A7%E3%80%81%E5%90%8D%E4%BA%BA%E6%8E%8C%E6%95%85%E3%80%81%E5%B1%B1%E6%B0%B4%E5%9C%B0%E7%90%86%E3%80%81%E5%AE%98%E5%88%B6%E5%AD%A6%E5%88%B6%E3%80%81%E5%8F%A4%E8%BF%B9%E8%89%BA%E6%96%87%E7%AD%89%E7%9F%A5%E8%AF%86%EF%BC%8C%E6%8C%96%E6%8E%98%E5%87%BA%E6%B7%B1%E5%9C%B3%E4%B8%80%E4%BA%8B%E4%B8%80%E7%89%A9%E6%B2%A7%E6%A1%91%E5%8F%98%E8%BF%81%E7%9A%84%E5%9F%BA%E5%9B%A0%E5%AF%86%E7%A0%81%2C%E5%8A%A9%E5%8A%9B%E4%B8%AD%E5%8D%8E%E4%BC%98%E7%A7%80%E4%BC%A0%E7%BB%9F%E6%96%87%E5%8C%96%E5%88%9B%E9%80%A0%E6%80%A7%E8%BD%AC%E5%8C%96%E5%92%8C%E5%88%9B%E6%96%B0%E6%80%A7%E5%8F%91%E5%B1%95%E3%80%82)。

**技术成熟度与成本：** 知识图谱技术本身在图书馆界已运用多年，但传统构建多靠人工和简单工具。引入AI智能体后，可大幅提高构建效率、扩展图谱规模。例如医学图书馆领域，有企业利用“大模型+知识图谱”打造科研情报助理，能从海量医学文献中自动抽取疾病、药物等实体及关系，构建医学知识网络并服务科研检索[ccf.org.cn](https://www.ccf.org.cn/Activities/Training/TF/tfzyxw/2024-11-06/835852.shtml" \l ":~:text=%E5%9F%BA%E4%BA%8E%E7%9F%A5%E8%AF%86%E5%9B%BE%E8%B0%B1%E5%A2%9E%E5%BC%BA%E5%A4%A7%E6%A8%A1%E5%9E%8B%E7%9A%84%E5%8C%BB%E5%AD%A6%E5%9B%BE%E4%B9%A6%E9%A6%86%E6%96%B0%E8%8C%83%E5%BC%8F,%E7%A7%91%E7%A0%94%E6%83%85%E6%8A%A5)。这类智能体需要训练专业NLP模型，成本较高，但对提高专业知识服务水平作用明显。目前来看，在狭窄学科领域构建高质量知识图谱仍需较多专家参与校对，智能体可减少部分体力劳动。服务部署方面，知识图谱可通过Web界面或小程序提供，成本取决于图谱规模和更新频率。如果借助外部平台（如Yewno付费服务），则以订阅形式收费。考虑到知识图谱可复用性强、对用户吸引力大，其投入产出比被认为是较高的。尤其在高校、科研图书馆，知识图谱助理能提升信息素养教育效果，间接带来学术产出提升。

**用户体验提升：** 知识图谱通过**可视化和语义关联**给用户带来全新的信息获取方式。一方面，复杂知识被结构化，有助于用户高效获取可靠信息并理解知识全景[fjlib.net](https://www.fjlib.net/zt/mtxh/stxk/llysj/fjstsgxk201403/201804/t20180414_358197.htm" \l ":~:text=%E5%B9%B4%E5%9C%A8%E3%80%8A%E7%A7%91%E5%AD%A6%E3%80%8B%E6%9D%82%E5%BF%97%E4%B8%8A%E5%8F%91%E8%A1%A8%E7%9A%84%E4%B8%80%E7%AF%87%E8%AE%BA%E6%96%87%E4%B8%AD%EF%BC%8C%E5%B0%B1%E8%AE%A4%E4%B8%BA%E4%BC%A0%E7%BB%9F%E7%9A%84%E5%9F%BA%E4%BA%8E%E5%AD%A6%E7%A7%91%E7%9A%84%E6%A3%80%E7%B4%A2%E5%92%8C%E6%91%98%E8%A6%81%E6%9C%8D%E5%8A%A1%E9%9A%BE%E4%BB%A5%E6%BB%A1%E8%B6%B3%E7%A0%94%E7%A9%B6%E7%9A%84%E9%9C%80%E6%B1%82%E3%80%82)。另一方面，交互式探索满足用户的好奇心，使学习过程更主动有趣[library.ucr.edu](https://library.ucr.edu/news/2022/04/29/ucr-library-implements-yewno-discover-ai-powered-research" \l ":~:text=Unlike%20other%20research%20engines%2C%20Yewno,valuable%20for%20the%20more%20inquisitive)。例如前述方志知识图谱，用户评价其“把书读活了”。再如福建省图书馆的一项研究指出，知识图谱可以将学科领域内**隐性知识显性化**，帮助用户快速发现热点和未知关联[fjlib.net](https://www.fjlib.net/zt/mtxh/stxk/llysj/fjstsgxk201403/201804/t20180414_358197.htm" \l ":~:text=%E5%B9%B4%E5%9C%A8%E3%80%8A%E7%A7%91%E5%AD%A6%E3%80%8B%E6%9D%82%E5%BF%97%E4%B8%8A%E5%8F%91%E8%A1%A8%E7%9A%84%E4%B8%80%E7%AF%87%E8%AE%BA%E6%96%87%E4%B8%AD%EF%BC%8C%E5%B0%B1%E8%AE%A4%E4%B8%BA%E4%BC%A0%E7%BB%9F%E7%9A%84%E5%9F%BA%E4%BA%8E%E5%AD%A6%E7%A7%91%E7%9A%84%E6%A3%80%E7%B4%A2%E5%92%8C%E6%91%98%E8%A6%81%E6%9C%8D%E5%8A%A1%E9%9A%BE%E4%BB%A5%E6%BB%A1%E8%B6%B3%E7%A0%94%E7%A9%B6%E7%9A%84%E9%9C%80%E6%B1%82%E3%80%82)。这些都表明智能体驱动的知识图谱服务能显著提升用户体验：从机械检索转向沉浸探索，从碎片信息获取转向体系知识获取。不过也需注意，不同用户对可视化的接受度不同，一些年长读者可能需要引导才能习惯这种新模式。因此图书馆在提供此类服务时，也要辅以**用户教育**和引导，使各层次用户都能受益。

### 3**.4.2 语义检索：自然语言查询与精准检索**

**场景定义:** 语义检索指利用自然语言处理和语义理解技术改进信息检索，使系统能理解用户查询背后的意图和含义，从而检索出更相关的结果。与传统基于关键词字面匹配的检索不同，语义检索可识别同义词、上下位概念等语义关系，实现**概念级**匹配。例如用户问“含维生素C最多的水果”，语义检索能识别这是寻找“维生素C含量高的水果”，即使数据库里没有这句话，也能通过数据推理找到答案。图书馆引入智能体可为检索提供**智能问答式接口**：读者用口语化的问题查询，AI代理将问题转换为专业检索策略，在馆藏或数据库中寻找答案，再以简洁语言给出结果。

**案例与实践:** 近年来多家图书馆和科技公司合作尝试将大模型驱动的语义检索融入馆藏系统：

* *高校图书馆AI导航助手（清华大学图书馆等）☆* – 2024年4月，清华大学图书馆上线了“AI导航助手”和“AI阅读助手”两项应用，成为国内智慧图书馆融合大模型的标杆案例[calsp.cn](https://www.calsp.cn/2024/08/21/bulletin-202408-05/" \l ":~:text=%E4%BA%BA%E7%9F%A5%E8%AF%86%E4%B8%AD%E5%BF%83%E7%AD%89%E5%88%9B%E6%96%B0%E5%8A%9F%E8%83%BD%E3%80%82%E7%9B%B8%E5%85%B3%E4%BA%A7%E5%93%81%E5%B7%B2%E5%9C%A8%E6%B8%85%E5%8D%8E%E5%A4%A7%E5%AD%A6%E5%9B%BE%E4%B9%A6%E9%A6%86%E3%80%81%E4%B8%8A%E6%B5%B7%E5%9B%BE%E4%B9%A6%E9%A6%86%E3%80%81%E5%8D%97%E4%BA%AC%E5%A4%A7%E5%AD%A6%E5%9B%BE%E4%B9%A6%E9%A6%86%E3%80%81%E8%A5%BF%E5%AE%89%E4%BA%A4%E9%80%9A%E5%A4%A7%E5%AD%A6%E5%9B%BE%E4%B9%A6%E9%A6%86%E7%AD%89%E5%A4%9A%E5%AE%B6%E8%A1%8C%E4%B8%9A%E6%9C%BA%E6%9E%84%E8%BF%9B%E8%A1%8C%E5%BA%94%E7%94%A8%E3%80%82)。其中**AI导航助手**针对馆藏数字资源提供智能检索服务。读者可以像和客服聊天一样输入自然语言问题，例如“请问人工智能伦理方面有何最新研究进展？”系统背后的智能体会将此理解为检索请求，先将问题语义解析提取出核心概念“人工智能伦理”“最新研究”，然后在馆藏数据库中进行**检索增强生成（RAG）**流程[calsp.cn](https://www.calsp.cn/2024/08/21/bulletin-202408-05/" \l ":~:text=%E2%80%9C%E5%B9%BB%E8%A7%89%E2%80%9D%EF%BC%8C%E4%BE%9D%E8%B5%96%E7%9A%84%E4%BF%A1%E6%81%AF%E5%8F%AF%E8%83%BD%E8%BF%87%E6%97%B6%EF%BC%8C%E5%A4%84%E7%90%86%E7%89%B9%E5%AE%9A%E7%9F%A5%E8%AF%86%E6%97%B6%E6%95%88%E7%8E%87%E4%B8%8D%E9%AB%98%EF%BC%8C%E7%BC%BA%E4%B9%8F%E4%B8%93%E4%B8%9A%E9%A2%86%E5%9F%9F%E7%9A%84%E6%B7%B1%E5%BA%A6%E6%B4%9E%E5%AF%9F%EF%BC%8C%E5%9C%A8%E7%89%B9%E5%AE%9A%E9%97%AE%E9%A2%98%E7%9A%84%E6%8E%A8%E7%90%86%E8%83%BD%E5%8A%9B%E4%B8%8A%E4%B9%9F%E6%9C%89%E6%89%80%E6%AC%A0%E7%BC%BA%E3%80%82%E6%AD%A3%E6%98%AF%E5%9C%A8%E8%BF%99%E6%A0%B7%E7%9A%84%E8%83%8C%E6%99%AF%E4%B8%8B%EF%BC%8C%E6%A3%80%E7%B4%A2%E5%A2%9E%E5%BC%BA%E7%94%9F%E6%88%90%E6%8A%80%E6%9C%AF%EF%BC%88Retr%20ieval)[calsp.cn](https://www.calsp.cn/2024/08/21/bulletin-202408-05/#:~:text=RAG%E9%80%9A%E8%BF%87%E6%A3%80%E7%B4%A2%E8%8E%B7%E5%8F%96%E7%9B%B8%E5%85%B3%E7%9A%84%E7%9F%A5%E8%AF%86%E5%B9%B6%E5%B0%86%E5%85%B6%E8%9E%8D%E5%85%A5Prompt%E6%8F%90%E7%A4%BA%E8%AF%8D%EF%BC%8C%E8%AE%A9%E5%A4%A7%E8%AF%AD%E8%A8%80%E6%A8%A1%E5%9E%8B%E8%83%BD%E5%A4%9F%E5%8F%82%E8%80%83%E7%9B%B8%E5%BA%94%E7%9A%84%E7%9F%A5%E8%AF%86%E4%BB%8E%E8%80%8C%E7%BB%99%E5%87%BA%E5%90%88%E7%90%86%E5%9B%9E%E7%AD%94%E3%80%82%E5%9B%A0%E6%AD%A4%EF%BC%8C%E5%8F%AF%E4%BB%A5%E5%B0%86RAG%E7%9A%84%E6%A0%B8%E5%BF%83%E7%90%86%E8%A7%A3%E4%B8%BA%E2%80%9C%E6%A3%80%E7%B4%A2%2B%E7%94%9F%E6%88%90%E2%80%9D%EF%BC%8C%E5%89%8D%E8%80%85%E4%B8%BB%E8%A6%81%20%E6%98%AF%E5%88%A9%E7%94%A8%E5%90%91%E9%87%8F%E6%95%B0%E6%8D%AE%E5%BA%93%E7%9A%84%E9%AB%98%E6%95%88%E5%AD%98%E5%82%A8%E5%92%8C%E6%A3%80%E7%B4%A2%E8%83%BD%E5%8A%9B%EF%BC%8C%E5%8F%AC%E5%9B%9E%E7%9B%AE%E6%A0%87%E7%9F%A5%E8%AF%86%EF%BC%9B%E5%90%8E%E8%80%85%E5%88%99%E6%98%AF%E5%88%A9%E7%94%A8%E5%A4%A7%E8%AF%AD%E8%A8%80%E6%A8%A1%E5%9E%8B%E5%92%8CPrompt%E6%8F%90%E7%A4%BA%E8%AF%8D%E5%B7%A5%E7%A8%8B%EF%BC%8C%E5%B0%86%E5%8F%AC%E5%9B%9E%E7%9A%84%E7%9F%A5%E8%AF%86%E5%90%88%E7%90%86%E5%88%A9%E7%94%A8%EF%BC%8C%E7%94%9F%E6%88%90%E7%9B%AE%E6%A0%87%E7%AD%94%E6%A1%88%E3%80%82)：通过向量数据库召回与该主题相关的最新论文摘要或报告，再让大语言模型基于检索结果生成简明答案给用户[calsp.cn](https://www.calsp.cn/2024/08/21/bulletin-202408-05/" \l ":~:text=RAG%E9%80%9A%E8%BF%87%E6%A3%80%E7%B4%A2%E8%8E%B7%E5%8F%96%E7%9B%B8%E5%85%B3%E7%9A%84%E7%9F%A5%E8%AF%86%E5%B9%B6%E5%B0%86%E5%85%B6%E8%9E%8D%E5%85%A5Prompt%E6%8F%90%E7%A4%BA%E8%AF%8D%EF%BC%8C%E8%AE%A9%E5%A4%A7%E8%AF%AD%E8%A8%80%E6%A8%A1%E5%9E%8B%E8%83%BD%E5%A4%9F%E5%8F%82%E8%80%83%E7%9B%B8%E5%BA%94%E7%9A%84%E7%9F%A5%E8%AF%86%E4%BB%8E%E8%80%8C%E7%BB%99%E5%87%BA%E5%90%88%E7%90%86%E5%9B%9E%E7%AD%94%E3%80%82%E5%9B%A0%E6%AD%A4%EF%BC%8C%E5%8F%AF%E4%BB%A5%E5%B0%86RAG%E7%9A%84%E6%A0%B8%E5%BF%83%E7%90%86%E8%A7%A3%E4%B8%BA%E2%80%9C%E6%A3%80%E7%B4%A2%2B%E7%94%9F%E6%88%90%E2%80%9D%EF%BC%8C%E5%89%8D%E8%80%85%E4%B8%BB%E8%A6%81%20%E6%98%AF%E5%88%A9%E7%94%A8%E5%90%91%E9%87%8F%E6%95%B0%E6%8D%AE%E5%BA%93%E7%9A%84%E9%AB%98%E6%95%88%E5%AD%98%E5%82%A8%E5%92%8C%E6%A3%80%E7%B4%A2%E8%83%BD%E5%8A%9B%EF%BC%8C%E5%8F%AC%E5%9B%9E%E7%9B%AE%E6%A0%87%E7%9F%A5%E8%AF%86%EF%BC%9B%E5%90%8E%E8%80%85%E5%88%99%E6%98%AF%E5%88%A9%E7%94%A8%E5%A4%A7%E8%AF%AD%E8%A8%80%E6%A8%A1%E5%9E%8B%E5%92%8CPrompt%E6%8F%90%E7%A4%BA%E8%AF%8D%E5%B7%A5%E7%A8%8B%EF%BC%8C%E5%B0%86%E5%8F%AC%E5%9B%9E%E7%9A%84%E7%9F%A5%E8%AF%86%E5%90%88%E7%90%86%E5%88%A9%E7%94%A8%EF%BC%8C%E7%94%9F%E6%88%90%E7%9B%AE%E6%A0%87%E7%AD%94%E6%A1%88%E3%80%82)[calsp.cn](https://www.calsp.cn/2024/08/21/bulletin-202408-05/#:~:text=RAG%E6%8A%80%E6%9C%AF%E6%98%AF%E4%B8%80%E7%A7%8D%E7%BB%93%E5%90%88%E4%BF%A1%E6%81%AF%E6%A3%80%E7%B4%A2%E5%92%8C%E6%96%87%E6%9C%AC%E7%94%9F%E6%88%90%E7%9A%84%E6%8A%80%E6%9C%AF%E3%80%82%E5%9C%A8%E5%9B%BE%E4%B9%A6%E9%A6%86%E4%B8%AD%EF%BC%8C%E5%9B%BE%E4%B9%A6%E9%A6%86%E9%A6%86%E8%97%8F%E5%BD%93%E4%B8%AD%E5%AD%98%E5%9C%A8%E5%A4%A7%E9%87%8F%E7%9A%84%E5%AD%A6%E6%9C%AF%E6%96%87%E7%8C%AE%EF%BC%8C%E5%8F%AF%E4%BB%A5%E5%B0%86%E8%BF%99%E4%BA%9B%E5%AD%A6%E6%9C%AF%E6%96%87%E7%8C%AE%E9%80%9A%E8%BF%87%E4%B8%93%E4%B8%9A%E7%9A%84%E5%90%91%E9%87%8F%E5%8C%96%E5%B7%A5%E5%85%B7%E8%BF%9B%E8%A1%8C%E5%88%87%E7%89%87%E5%AD%98%E5%82%A8%E5%9C%A8%E5%90%91%E9%87%8F%E6%95%B0%E6%8D%AE%E5%BA%93%E5%BD%93%E4%B8%AD%EF%BC%8C%20%E5%BD%A2%E6%88%90%E4%B8%80%E4%B8%AA%E6%9C%AC%E5%9C%B0%E7%9F%A5%E8%AF%86%E5%BA%93%EF%BC%8C%E5%86%8D%E5%B0%86%E6%9C%AC%E5%9C%B0%E7%9F%A5%E8%AF%86%E5%BA%93%E4%B8%8E%E5%A4%A7%E8%AF%AD%E8%A8%80%E6%A8%A1%E5%9E%8B%E8%BF%9B%E8%A1%8C%E7%BB%93%E5%90%88%EF%BC%8C%E5%BD%93%E7%94%A8%E6%88%B7%E6%8F%90%E5%87%BA%E9%97%AE%E9%A2%98%E6%97%B6%EF%BC%8C%E7%B3%BB%E7%BB%9F%E8%83%BD%E5%A4%9F%E6%A0%B9%E6%8D%AE%E7%94%A8%E6%88%B7%E7%9A%84%E9%97%AE%E9%A2%98%E6%A3%80%E7%B4%A2%E5%9B%BE%E4%B9%A6%E9%A6%86%E7%9A%84%E6%95%B0%E5%AD%97%E8%B5%84%E6%BA%90%EF%BC%8C%E5%B9%B6%E7%94%9F%E6%88%90%E5%87%86%E7%A1%AE%E7%9A%84%E5%9B%9E%E7%AD%94%E3%80%82%E8%BF%99%E5%8F%AF%E4%BB%A5%E5%A4%A7%E5%A4%A7%E5%87%8F%E5%B0%91%E7%94%A8%E6%88%B7%E6%9F%A5%E6%89%BE%E4%BF%A1%20%E6%81%AF%E7%9A%84%E6%97%B6%E9%97%B4%EF%BC%8C%E5%B9%B6%E6%8F%90%E9%AB%98%E5%9B%9E%E7%AD%94%E7%9A%84%E7%9B%B8%E5%85%B3%E6%80%A7%E3%80%82)。这样，读者无需熟悉复杂的检索语法，也不必人工筛选海量结果，就能直接获得可靠的回答或摘要。清华图书馆表示，这一助手**显著提升了师生的检索与阅读效率**，上线后反馈良好[calsp.cn](https://www.calsp.cn/2024/08/21/bulletin-202408-05/" \l ":~:text=%E6%B8%85%E5%8D%8E%E5%A4%A7%E5%AD%A6%E5%9B%BE%E4%B9%A6%E9%A6%86%E4%B8%BA%E5%BF%AB%E9%80%9F%E6%8E%A8%E8%BF%9B%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E5%9C%A8%E5%9B%BE%E4%B9%A6%E9%A6%86%E7%9A%84%E8%90%BD%E5%9C%B0%E5%BA%94%E7%94%A8%EF%BC%8C%E4%BA%8E2024%E5%B9%B44%E6%9C%8828%E6%97%A5%E6%AD%A3%E5%BC%8F%E4%B8%8A%E7%BA%BF%E4%BA%86AI%E9%98%85%E8%AF%BB%E5%8A%A9%E6%89%8B%E3%80%81AI%E5%AF%BC%E8%88%AA%E5%8A%A9%E6%89%8B%E5%BA%94%E7%94%A8%E3%80%82%E9%80%9A%E8%BF%87%E6%99%BA%E8%83%BD%E5%88%86%E6%9E%90PDF%E6%96%87%E6%A1%A3%E5%92%8C%E6%95%B0%E6%8D%AE%E5%BA%93%E5%86%85%E5%AE%B9%EF%BC%8C%E6%8F%90%E5%8D%87%E4%BA%86%E5%B8%88%20%E7%94%9F%E7%9A%84%E9%98%85%E8%AF%BB%E5%92%8C%E7%A0%94%E7%A9%B6%E6%95%88%E7%8E%87%E3%80%82%E4%B8%A4%E9%A1%B9%E5%BA%94%E7%94%A8%E5%9D%87%E9%9D%A2%E5%90%91%E6%B8%85%E5%8D%8E%E5%85%A8%E6%A0%A1%E5%B8%88%E7%94%9F%E6%8F%90%E4%BE%9B%E6%9C%8D%E5%8A%A1%EF%BC%8C%E4%B8%8A%E7%BA%BF%E5%90%8E%E6%94%B6%E5%88%B0%E4%BA%86%E8%AF%BB%E8%80%85%E7%9A%84%E8%89%AF%E5%A5%BD%E5%8F%8D%E9%A6%88%E3%80%82)。类似地，上海图书馆专业服务中心也融入了多个AI辅助功能，包括快速分析文献检索结果、生成主题综述等[calsp.cn](https://www.calsp.cn/2024/08/21/bulletin-202408-05/" \l ":~:text=%E4%B8%8A%E6%B5%B7%E5%9B%BE%E4%B9%A6%E9%A6%86%E4%B8%93%E4%B8%9A%E6%9C%8D%E5%8A%A1%E4%B8%AD%E5%BF%83%E5%9C%A8%E4%B8%8A%E5%9B%BE%E4%B8%93%E4%B8%9A%E6%9C%8D%E5%8A%A1%E9%97%A8%E6%88%B7%E4%B8%AD%EF%BC%8C%E4%B9%9F%E8%9E%8D%E5%85%A5%E4%BA%86%E5%A4%9A%E4%B8%AAAI%E6%99%BA%E8%83%BD%E8%BE%85%E5%8A%A9%E5%8A%9F%E8%83%BD%E3%80%82%E5%88%9B%E6%96%B0%E6%80%A7%E7%9A%84%E5%B0%86%E5%A4%A7%E8%AF%AD%E8%A8%80%E6%A8%A1%E5%9E%8B%E8%9E%8D%E5%85%A5%E5%9B%BE%E4%B9%A6%E9%A6%86%E8%B5%84%E6%BA%90%E5%8F%91%E7%8E%B0%E6%9C%8D%E5%8A%A1%EF%BC%8C%E5%AE%9E%E7%8E%B0%E4%BA%86%E5%AF%B9%E6%96%87%E7%8C%AE%E6%A3%80%E7%B4%A2%E7%BB%93%E6%9E%9C%E7%9A%84%E5%BF%AB%E9%80%9F%E5%88%86%E6%9E%90%E5%8A%9F%E8%83%BD%E3%80%82%E5%B9%B6%E9%80%9A%E8%BF%87%20%E6%89%93%E9%80%A0%E4%B8%AA%E6%80%A7%E5%8C%96%E7%9A%84%E8%AF%BB%E8%80%85%E4%B8%AA%E4%BA%BA%E7%9F%A5%E8%AF%86%E4%B8%AD%E5%BF%83%EF%BC%8C%E6%8F%90%E4%BE%9B%E6%99%BA%E8%83%BD%E9%97%AE%E7%AD%94%E7%AD%89%E8%BE%85%E5%8A%A9%E9%98%85%E8%AF%BB%E6%9C%8D%E5%8A%A1%EF%BC%8C%E5%85%A8%E9%9D%A2%E6%8F%90%E5%8D%87%E8%AF%BB%E8%80%85%E8%B5%84%E6%BA%90%E6%A3%80%E7%B4%A2%E4%B8%8E%E9%98%85%E8%AF%BB%E4%BD%93%E9%AA%8C%E3%80%82)。这些实践表明，在学术信息检索场景，智能体能充当知识检索的“**中间代理**”，既懂用户语言又懂数据库语言，把两者桥接起来，提供比传统检索更友好的体验。
* *商业引擎与图书馆合作☆* – 一些商业学术搜索平台也推出AI助手功能，开始与图书馆系统对接。例如知名期刊数据库EBSCO Discovery在其Summon检索中集成了OpenAI的GPT模型，以实现**对用户查询的语义扩展和结果总结**。读者搜索后，系统可以给出一段基于结果的总结性回答或下一步建议，而不仅是一串文献列表。这类似为读者配备一个信息助理，帮助理解检索结果的内容精华。国内，维普资讯开发了“智图”智慧图书馆解决方案，与一些高校合作，提供“一框式”语义检索服务[vipslib.com](https://www.vipslib.com/case.html" \l ":~:text=%E9%80%9A%E8%BF%87%E4%B8%8E%E7%BB%B4%E6%99%AE%E6%99%BA%E5%9B%BE%E5%90%88%E4%BD%9C%E5%BB%BA%E8%AE%BE%E6%99%BA%E6%85%A7%E5%9B%BE%E4%B9%A6%E9%A6%86%EF%BC%8C%E5%8E%86%E7%BB%8F%E4%B8%89%E5%B9%B4%E5%BB%BA%E8%AE%BE%EF%BC%8C%E5%9C%A8%E6%9C%AC%E6%A0%A1%E5%9B%BE%E4%B9%A6%E9%A6%86%E5%BD%A2%E6%88%90%E4%BA%86%E4%B8%80%E5%A5%97%E2%80%9C%E5%85%A8%E5%BC%80%E6%94%BE%E3%80%81%E4%B8%80%E6%A1%86%E5%BC%8F%E2%80%9D%E6%95%B0%E5%AD%97%E8%B5%84%E6%BA%90%E6%9C%8D%E5%8A%A1%E6%A8%A1%E5%BC%8F%EF%BC%8C%E5%B9%B6%E5%8F%96%E5%BE%97%E4%BA%8695)。例如内蒙古工业大学图书馆通过智图建设了全开放的一站式检索门户，用户无论输入自然语言问题还是关键字，系统都能调用后端多源数据进行语义处理，返回融合排序的结果，资源服务满意度超过95%[vipslib.com](https://www.vipslib.com/case.html#:~:text=%E9%80%9A%E8%BF%87%E4%B8%8E%E7%BB%B4%E6%99%AE%E6%99%BA%E5%9B%BE%E5%90%88%E4%BD%9C%E5%BB%BA%E8%AE%BE%E6%99%BA%E6%85%A7%E5%9B%BE%E4%B9%A6%E9%A6%86%EF%BC%8C%E5%8E%86%E7%BB%8F%E4%B8%89%E5%B9%B4%E5%BB%BA%E8%AE%BE%EF%BC%8C%E5%9C%A8%E6%9C%AC%E6%A0%A1%E5%9B%BE%E4%B9%A6%E9%A6%86%E5%BD%A2%E6%88%90%E4%BA%86%E4%B8%80%E5%A5%97%E2%80%9C%E5%85%A8%E5%BC%80%E6%94%BE%E3%80%81%E4%B8%80%E6%A1%86%E5%BC%8F%E2%80%9D%E6%95%B0%E5%AD%97%E8%B5%84%E6%BA%90%E6%9C%8D%E5%8A%A1%E6%A8%A1%E5%BC%8F%EF%BC%8C%E5%B9%B6%E5%8F%96%E5%BE%97%E4%BA%8695)。这表明，智能检索代理不仅可以回答问题，还能整合不同数据库的结果，为用户提供统一的检索体验。

**技术成熟度与成本:** 语义检索所需的核心技术包括：大语言模型的自然语言理解，向量语义索引/搜索，以及检索增强的生成。这些技术单项已较成熟，但端到端系统实现仍有一定门槛。幸运的是，不少AI公司推出了成套解决方案（如Microsoft Azure Cognitive Search, OpenAI API等），图书馆可通过API调用实现语义检索功能，而无需自行训练大型模型[calsp.cn](https://www.calsp.cn/2024/08/21/bulletin-202408-05/" \l ":~:text=%E4%BB%80%E4%B9%88%E6%98%AFRAG)。成本方面，大模型API目前按调用计费，对每次查询消耗的Token收费。若用户量较大，则需要考虑较高的预算。国内一些解决方案采用本地部署小型模型结合开源向量数据库，如仰格信息科技为多所高校构建的RAG系统中，使用自研或开源模型以控制成本[calsp.cn](https://www.calsp.cn/2024/08/21/bulletin-202408-05/" \l ":~:text=%E5%9B%BE%E4%B9%A6%E9%A6%86%E5%BA%94%E7%94%A8RAG%E5%AE%9E%E9%99%85%E6%A1%88%E4%BE%8B)。例如南京大学、西安交大等图书馆上线的AI助手皆源于这套方案[calsp.cn](https://www.calsp.cn/2024/08/21/bulletin-202408-05/" \l ":~:text=%E5%8D%97%E4%BA%AC%E4%BB%B0%E6%A0%BC%E4%BF%A1%E6%81%AF%E7%A7%91%E6%8A%80%E6%9C%89%E9%99%90%E5%85%AC%E5%8F%B8%E5%9F%BA%E4%BA%8E%E4%B8%8A%E6%96%87%E7%9A%84%E6%80%9D%E8%B7%AF%EF%BC%8C%E8%AE%BE%E8%AE%A1%E7%A0%94%E5%8F%91%E4%BA%86RAG%E6%A3%80%E7%B4%A2%E5%A2%9E%E5%BC%BA%E7%94%9F%E6%88%90%E6%8A%80%E6%9C%AF%E7%9B%B8%E5%85%B3%E7%9A%84%E4%BA%A7%E5%93%81%E3%80%82%E5%AE%9E%E7%8E%B0%E4%BA%86PDF%E6%99%BA%E8%83%BD%E8%BE%85%E5%8A%A9%E9%98%85%E8%AF%BB%E5%92%8C%E5%88%86%E6%9E%90%E3%80%81%E6%95%B0%E6%8D%AE%E5%BA%93AI%E5%AF%BC%E8%88%AA%E5%8A%A9%E6%89%8B%E3%80%81%E8%B5%84%E6%BA%90%E5%8F%91%E7%8E%B0AI%E6%A3%80%E7%B4%A2%E5%8A%A9%E6%89%8B%E3%80%81%E4%B8%AA%20%E4%BA%BA%E7%9F%A5%E8%AF%86%E4%B8%AD%E5%BF%83%E7%AD%89%E5%88%9B%E6%96%B0%E5%8A%9F%E8%83%BD%E3%80%82%E7%9B%B8%E5%85%B3%E4%BA%A7%E5%93%81%E5%B7%B2%E5%9C%A8%E6%B8%85%E5%8D%8E%E5%A4%A7%E5%AD%A6%E5%9B%BE%E4%B9%A6%E9%A6%86%E3%80%81%E4%B8%8A%E6%B5%B7%E5%9B%BE%E4%B9%A6%E9%A6%86%E3%80%81%E5%8D%97%E4%BA%AC%E5%A4%A7%E5%AD%A6%E5%9B%BE%E4%B9%A6%E9%A6%86%E3%80%81%E8%A5%BF%E5%AE%89%E4%BA%A4%E9%80%9A%E5%A4%A7%E5%AD%A6%E5%9B%BE%E4%B9%A6%E9%A6%86%E7%AD%89%E5%A4%9A%E5%AE%B6%E8%A1%8C%E4%B8%9A%E6%9C%BA%E6%9E%84%E8%BF%9B%E8%A1%8C%E5%BA%94%E7%94%A8%E3%80%82)。经验表明，在限定领域问答上，中型模型配合知识库也能取得不错效果，而费用远低于调用GPT-4等大模型。因此，语义检索智能体可根据馆藏规模和预算灵活选型。需要一提的是，维护成本不容忽视——语义索引需要定期更新新文献，模型也需根据用户反馈持续优化。

**用户体验提升:** 语义检索智能体带来的直观好处是**检索门槛大大降低**。许多读者不会构造复杂的布尔检索式，而智能体支持用日常语言发问，从而让信息检索更贴近自然思维方式。这对本科生、新手科研人员尤其友好[calsp.cn](https://www.calsp.cn/2024/08/21/bulletin-202408-05/" \l ":~:text=RAG%E6%8A%80%E6%9C%AF%E6%98%AF%E4%B8%80%E7%A7%8D%E7%BB%93%E5%90%88%E4%BF%A1%E6%81%AF%E6%A3%80%E7%B4%A2%E5%92%8C%E6%96%87%E6%9C%AC%E7%94%9F%E6%88%90%E7%9A%84%E6%8A%80%E6%9C%AF%E3%80%82%E5%9C%A8%E5%9B%BE%E4%B9%A6%E9%A6%86%E4%B8%AD%EF%BC%8C%E5%9B%BE%E4%B9%A6%E9%A6%86%E9%A6%86%E8%97%8F%E5%BD%93%E4%B8%AD%E5%AD%98%E5%9C%A8%E5%A4%A7%E9%87%8F%E7%9A%84%E5%AD%A6%E6%9C%AF%E6%96%87%E7%8C%AE%EF%BC%8C%E5%8F%AF%E4%BB%A5%E5%B0%86%E8%BF%99%E4%BA%9B%E5%AD%A6%E6%9C%AF%E6%96%87%E7%8C%AE%E9%80%9A%E8%BF%87%E4%B8%93%E4%B8%9A%E7%9A%84%E5%90%91%E9%87%8F%E5%8C%96%E5%B7%A5%E5%85%B7%E8%BF%9B%E8%A1%8C%E5%88%87%E7%89%87%E5%AD%98%E5%82%A8%E5%9C%A8%E5%90%91%E9%87%8F%E6%95%B0%E6%8D%AE%E5%BA%93%E5%BD%93%E4%B8%AD%EF%BC%8C%20%E5%BD%A2%E6%88%90%E4%B8%80%E4%B8%AA%E6%9C%AC%E5%9C%B0%E7%9F%A5%E8%AF%86%E5%BA%93%EF%BC%8C%E5%86%8D%E5%B0%86%E6%9C%AC%E5%9C%B0%E7%9F%A5%E8%AF%86%E5%BA%93%E4%B8%8E%E5%A4%A7%E8%AF%AD%E8%A8%80%E6%A8%A1%E5%9E%8B%E8%BF%9B%E8%A1%8C%E7%BB%93%E5%90%88%EF%BC%8C%E5%BD%93%E7%94%A8%E6%88%B7%E6%8F%90%E5%87%BA%E9%97%AE%E9%A2%98%E6%97%B6%EF%BC%8C%E7%B3%BB%E7%BB%9F%E8%83%BD%E5%A4%9F%E6%A0%B9%E6%8D%AE%E7%94%A8%E6%88%B7%E7%9A%84%E9%97%AE%E9%A2%98%E6%A3%80%E7%B4%A2%E5%9B%BE%E4%B9%A6%E9%A6%86%E7%9A%84%E6%95%B0%E5%AD%97%E8%B5%84%E6%BA%90%EF%BC%8C%E5%B9%B6%E7%94%9F%E6%88%90%E5%87%86%E7%A1%AE%E7%9A%84%E5%9B%9E%E7%AD%94%E3%80%82%E8%BF%99%E5%8F%AF%E4%BB%A5%E5%A4%A7%E5%A4%A7%E5%87%8F%E5%B0%91%E7%94%A8%E6%88%B7%E6%9F%A5%E6%89%BE%E4%BF%A1%20%E6%81%AF%E7%9A%84%E6%97%B6%E9%97%B4%EF%BC%8C%E5%B9%B6%E6%8F%90%E9%AB%98%E5%9B%9E%E7%AD%94%E7%9A%84%E7%9B%B8%E5%85%B3%E6%80%A7%E3%80%82)。另外，智能体可直接给出汇总答案或推荐，让用户无需自行比对大量文献，提高了效率和体验。例如过去用户查一个专业问题，可能要阅读多篇论文才能总结出答案；现在AI助手能基于馆藏自动汇总出要点，然后列出参考文献来源[calsp.cn](https://www.calsp.cn/2024/08/21/bulletin-202408-05/" \l ":~:text=RAG%E9%80%9A%E8%BF%87%E6%A3%80%E7%B4%A2%E8%8E%B7%E5%8F%96%E7%9B%B8%E5%85%B3%E7%9A%84%E7%9F%A5%E8%AF%86%E5%B9%B6%E5%B0%86%E5%85%B6%E8%9E%8D%E5%85%A5Prompt%E6%8F%90%E7%A4%BA%E8%AF%8D%EF%BC%8C%E8%AE%A9%E5%A4%A7%E8%AF%AD%E8%A8%80%E6%A8%A1%E5%9E%8B%E8%83%BD%E5%A4%9F%E5%8F%82%E8%80%83%E7%9B%B8%E5%BA%94%E7%9A%84%E7%9F%A5%E8%AF%86%E4%BB%8E%E8%80%8C%E7%BB%99%E5%87%BA%E5%90%88%E7%90%86%E5%9B%9E%E7%AD%94%E3%80%82%E5%9B%A0%E6%AD%A4%EF%BC%8C%E5%8F%AF%E4%BB%A5%E5%B0%86RAG%E7%9A%84%E6%A0%B8%E5%BF%83%E7%90%86%E8%A7%A3%E4%B8%BA%E2%80%9C%E6%A3%80%E7%B4%A2%2B%E7%94%9F%E6%88%90%E2%80%9D%EF%BC%8C%E5%89%8D%E8%80%85%E4%B8%BB%E8%A6%81%20%E6%98%AF%E5%88%A9%E7%94%A8%E5%90%91%E9%87%8F%E6%95%B0%E6%8D%AE%E5%BA%93%E7%9A%84%E9%AB%98%E6%95%88%E5%AD%98%E5%82%A8%E5%92%8C%E6%A3%80%E7%B4%A2%E8%83%BD%E5%8A%9B%EF%BC%8C%E5%8F%AC%E5%9B%9E%E7%9B%AE%E6%A0%87%E7%9F%A5%E8%AF%86%EF%BC%9B%E5%90%8E%E8%80%85%E5%88%99%E6%98%AF%E5%88%A9%E7%94%A8%E5%A4%A7%E8%AF%AD%E8%A8%80%E6%A8%A1%E5%9E%8B%E5%92%8CPrompt%E6%8F%90%E7%A4%BA%E8%AF%8D%E5%B7%A5%E7%A8%8B%EF%BC%8C%E5%B0%86%E5%8F%AC%E5%9B%9E%E7%9A%84%E7%9F%A5%E8%AF%86%E5%90%88%E7%90%86%E5%88%A9%E7%94%A8%EF%BC%8C%E7%94%9F%E6%88%90%E7%9B%AE%E6%A0%87%E7%AD%94%E6%A1%88%E3%80%82)。这种“一站式解答+出处”的模式极大地方便了用户。当然，也有挑战：用户可能对AI给出的答案过度信任而不去验证，特别当AI偶有出错时可能误导。因此图书馆在上线此类服务时，通常会在界面提示“本回答由AI生成，仅供参考”，并附上引用文献供用户自行核实[calsp.cn](https://www.calsp.cn/2024/08/21/bulletin-202408-05/" \l ":~:text=RAG%E6%8A%80%E6%9C%AF%E6%98%AF%E4%B8%80%E7%A7%8D%E7%BB%93%E5%90%88%E4%BF%A1%E6%81%AF%E6%A3%80%E7%B4%A2%E5%92%8C%E6%96%87%E6%9C%AC%E7%94%9F%E6%88%90%E7%9A%84%E6%8A%80%E6%9C%AF%E3%80%82%E5%9C%A8%E5%9B%BE%E4%B9%A6%E9%A6%86%E4%B8%AD%EF%BC%8C%E5%9B%BE%E4%B9%A6%E9%A6%86%E9%A6%86%E8%97%8F%E5%BD%93%E4%B8%AD%E5%AD%98%E5%9C%A8%E5%A4%A7%E9%87%8F%E7%9A%84%E5%AD%A6%E6%9C%AF%E6%96%87%E7%8C%AE%EF%BC%8C%E5%8F%AF%E4%BB%A5%E5%B0%86%E8%BF%99%E4%BA%9B%E5%AD%A6%E6%9C%AF%E6%96%87%E7%8C%AE%E9%80%9A%E8%BF%87%E4%B8%93%E4%B8%9A%E7%9A%84%E5%90%91%E9%87%8F%E5%8C%96%E5%B7%A5%E5%85%B7%E8%BF%9B%E8%A1%8C%E5%88%87%E7%89%87%E5%AD%98%E5%82%A8%E5%9C%A8%E5%90%91%E9%87%8F%E6%95%B0%E6%8D%AE%E5%BA%93%E5%BD%93%E4%B8%AD%EF%BC%8C%20%E5%BD%A2%E6%88%90%E4%B8%80%E4%B8%AA%E6%9C%AC%E5%9C%B0%E7%9F%A5%E8%AF%86%E5%BA%93%EF%BC%8C%E5%86%8D%E5%B0%86%E6%9C%AC%E5%9C%B0%E7%9F%A5%E8%AF%86%E5%BA%93%E4%B8%8E%E5%A4%A7%E8%AF%AD%E8%A8%80%E6%A8%A1%E5%9E%8B%E8%BF%9B%E8%A1%8C%E7%BB%93%E5%90%88%EF%BC%8C%E5%BD%93%E7%94%A8%E6%88%B7%E6%8F%90%E5%87%BA%E9%97%AE%E9%A2%98%E6%97%B6%EF%BC%8C%E7%B3%BB%E7%BB%9F%E8%83%BD%E5%A4%9F%E6%A0%B9%E6%8D%AE%E7%94%A8%E6%88%B7%E7%9A%84%E9%97%AE%E9%A2%98%E6%A3%80%E7%B4%A2%E5%9B%BE%E4%B9%A6%E9%A6%86%E7%9A%84%E6%95%B0%E5%AD%97%E8%B5%84%E6%BA%90%EF%BC%8C%E5%B9%B6%E7%94%9F%E6%88%90%E5%87%86%E7%A1%AE%E7%9A%84%E5%9B%9E%E7%AD%94%E3%80%82%E8%BF%99%E5%8F%AF%E4%BB%A5%E5%A4%A7%E5%A4%A7%E5%87%8F%E5%B0%91%E7%94%A8%E6%88%B7%E6%9F%A5%E6%89%BE%E4%BF%A1%20%E6%81%AF%E7%9A%84%E6%97%B6%E9%97%B4%EF%BC%8C%E5%B9%B6%E6%8F%90%E9%AB%98%E5%9B%9E%E7%AD%94%E7%9A%84%E7%9B%B8%E5%85%B3%E6%80%A7%E3%80%82)。总体而言，大多数用户乐于接受AI辅助检索这一创新服务，因为它**节省时间**且**交互自然**。一些评价研究表明，AI检索助手可将用户找到满意答案的时间平均缩短30-50%，对于复杂课题可能节省更多。随着用户逐渐熟悉这一模式，对图书馆的信息咨询满意度有望提升。

### **3.4.3 智能问答：对话式信息咨询服务**

**场景定义：** 图书馆传统的信息咨询和参考问答服务，指由馆员解答读者各类信息需求，例如帮助查找资料、解答主题领域问题等。引入智能体后，可提供**对话式的智能问答**服务：读者通过聊天界面向AI提问，AI利用其知识和馆藏资源即时给出回答。这实际上扩展了图书馆的“虚拟参考服务”，使之24小时可用，并能处理高并发的咨询请求。智能问答代理既可以回答常见的馆务问题（例如开放时间、借阅规则），也可以在领域知识许可范围内回答学术或科普问题，甚至充当读者的研究助手。

**案例与实践：** 各类型图书馆均开始探索用ChatGPT等大模型构建本馆的智能问答系统：

* *公共图书馆虚拟助手☆* – 武汉市图书馆早在2021年就推出了微信公众号上的“**小图AI助手**”，最初基于规则库回答常见问题。2023年起，他们引入了大语言模型能力，使AI助手能够理解更复杂的自然语言提问并给出灵活作答。例如读者问“能不能推荐几本关于长江文化的书？”，以前的系统可能无法理解，现在的小图AI可以结合馆藏知识推荐相关书籍并简述每本书的简介。这背后利用的是知识库问答技术：将图书馆书目及FAQ等数据接入向量数据库，然后由大模型实时匹配生成回复。一些公共馆还尝试让AI助手提供生活信息咨询，如某读者问“附近有什么适合周末带孩子去的文化活动？”，AI可根据市民文化数据库给出建议清单。虽然这些功能尚在完善，但初步调查显示用户对图书馆AI助手的**响应速度和便利性**表示认可。
* *高校图书馆学术问答☆* – 前述清华大学图书馆上线的**AI阅读助手**就是学术问答的实例[calsp.cn](https://www.calsp.cn/2024/08/21/bulletin-202408-05/" \l ":~:text=%E6%B8%85%E5%8D%8E%E5%A4%A7%E5%AD%A6%E5%9B%BE%E4%B9%A6%E9%A6%86%E4%B8%BA%E5%BF%AB%E9%80%9F%E6%8E%A8%E8%BF%9B%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E5%9C%A8%E5%9B%BE%E4%B9%A6%E9%A6%86%E7%9A%84%E8%90%BD%E5%9C%B0%E5%BA%94%E7%94%A8%EF%BC%8C%E4%BA%8E2024%E5%B9%B44%E6%9C%8828%E6%97%A5%E6%AD%A3%E5%BC%8F%E4%B8%8A%E7%BA%BF%E4%BA%86AI%E9%98%85%E8%AF%BB%E5%8A%A9%E6%89%8B%E3%80%81AI%E5%AF%BC%E8%88%AA%E5%8A%A9%E6%89%8B%E5%BA%94%E7%94%A8%E3%80%82%E9%80%9A%E8%BF%87%E6%99%BA%E8%83%BD%E5%88%86%E6%9E%90PDF%E6%96%87%E6%A1%A3%E5%92%8C%E6%95%B0%E6%8D%AE%E5%BA%93%E5%86%85%E5%AE%B9%EF%BC%8C%E6%8F%90%E5%8D%87%E4%BA%86%E5%B8%88%20%E7%94%9F%E7%9A%84%E9%98%85%E8%AF%BB%E5%92%8C%E7%A0%94%E7%A9%B6%E6%95%88%E7%8E%87%E3%80%82%E4%B8%A4%E9%A1%B9%E5%BA%94%E7%94%A8%E5%9D%87%E9%9D%A2%E5%90%91%E6%B8%85%E5%8D%8E%E5%85%A8%E6%A0%A1%E5%B8%88%E7%94%9F%E6%8F%90%E4%BE%9B%E6%9C%8D%E5%8A%A1%EF%BC%8C%E4%B8%8A%E7%BA%BF%E5%90%8E%E6%94%B6%E5%88%B0%E4%BA%86%E8%AF%BB%E8%80%85%E7%9A%84%E8%89%AF%E5%A5%BD%E5%8F%8D%E9%A6%88%E3%80%82)。该助手支持读者与之多轮对话来获取文献中的信息：读者可以上传PDF论文，然后向AI提问论文内容，如“这篇文章的核心结论是什么？”、“作者用的研究方法有何优势？”等，AI会基于对全文的理解作答[calsp.cn](https://www.calsp.cn/2024/08/21/bulletin-202408-05/" \l ":~:text=%E6%89%93%E9%80%A0%E4%B8%AA%E6%80%A7%E5%8C%96%E7%9A%84%E8%AF%BB%E8%80%85%E4%B8%AA%E4%BA%BA%E7%9F%A5%E8%AF%86%E4%B8%AD%E5%BF%83%EF%BC%8C%E6%8F%90%E4%BE%9B%E6%99%BA%E8%83%BD%E9%97%AE%E7%AD%94%E7%AD%89%E8%BE%85%E5%8A%A9%E9%98%85%E8%AF%BB%E6%9C%8D%E5%8A%A1%EF%BC%8C%E5%85%A8%E9%9D%A2%E6%8F%90%E5%8D%87%E8%AF%BB%E8%80%85%E8%B5%84%E6%BA%90%E6%A3%80%E7%B4%A2%E4%B8%8E%E9%98%85%E8%AF%BB%E4%BD%93%E9%AA%8C%E3%80%82)。这等于是一个实时的文献解读助手，帮助读者快速提炼论文要点甚至生成图表总结[calsp.cn](https://www.calsp.cn/2024/08/21/bulletin-202408-05/" \l ":~:text=PDF%E6%99%BA%E8%83%BD%E8%BE%85%E5%8A%A9%E9%98%85%E8%AF%BB%E5%92%8C%E5%88%86%E6%9E%90)。此外，在不提供文献时，读者也能直接提学术问题，AI会据馆藏资料和自身训练知识给出解答，并引用相应文献来源。这与传统参考咨询台的作用相似：学生可以直接在线问“人工智能伦理有哪些主要问题？”，AI就像一位经验丰富的馆员，先给出要点梳理，然后附上几篇重要论文来源[calsp.cn](https://www.calsp.cn/2024/08/21/bulletin-202408-05/" \l ":~:text=RAG%E9%80%9A%E8%BF%87%E6%A3%80%E7%B4%A2%E8%8E%B7%E5%8F%96%E7%9B%B8%E5%85%B3%E7%9A%84%E7%9F%A5%E8%AF%86%E5%B9%B6%E5%B0%86%E5%85%B6%E8%9E%8D%E5%85%A5Prompt%E6%8F%90%E7%A4%BA%E8%AF%8D%EF%BC%8C%E8%AE%A9%E5%A4%A7%E8%AF%AD%E8%A8%80%E6%A8%A1%E5%9E%8B%E8%83%BD%E5%A4%9F%E5%8F%82%E8%80%83%E7%9B%B8%E5%BA%94%E7%9A%84%E7%9F%A5%E8%AF%86%E4%BB%8E%E8%80%8C%E7%BB%99%E5%87%BA%E5%90%88%E7%90%86%E5%9B%9E%E7%AD%94%E3%80%82%E5%9B%A0%E6%AD%A4%EF%BC%8C%E5%8F%AF%E4%BB%A5%E5%B0%86RAG%E7%9A%84%E6%A0%B8%E5%BF%83%E7%90%86%E8%A7%A3%E4%B8%BA%E2%80%9C%E6%A3%80%E7%B4%A2%2B%E7%94%9F%E6%88%90%E2%80%9D%EF%BC%8C%E5%89%8D%E8%80%85%E4%B8%BB%E8%A6%81%20%E6%98%AF%E5%88%A9%E7%94%A8%E5%90%91%E9%87%8F%E6%95%B0%E6%8D%AE%E5%BA%93%E7%9A%84%E9%AB%98%E6%95%88%E5%AD%98%E5%82%A8%E5%92%8C%E6%A3%80%E7%B4%A2%E8%83%BD%E5%8A%9B%EF%BC%8C%E5%8F%AC%E5%9B%9E%E7%9B%AE%E6%A0%87%E7%9F%A5%E8%AF%86%EF%BC%9B%E5%90%8E%E8%80%85%E5%88%99%E6%98%AF%E5%88%A9%E7%94%A8%E5%A4%A7%E8%AF%AD%E8%A8%80%E6%A8%A1%E5%9E%8B%E5%92%8CPrompt%E6%8F%90%E7%A4%BA%E8%AF%8D%E5%B7%A5%E7%A8%8B%EF%BC%8C%E5%B0%86%E5%8F%AC%E5%9B%9E%E7%9A%84%E7%9F%A5%E8%AF%86%E5%90%88%E7%90%86%E5%88%A9%E7%94%A8%EF%BC%8C%E7%94%9F%E6%88%90%E7%9B%AE%E6%A0%87%E7%AD%94%E6%A1%88%E3%80%82)[calsp.cn](https://www.calsp.cn/2024/08/21/bulletin-202408-05/#:~:text=%E6%B8%85%E5%8D%8E%E5%A4%A7%E5%AD%A6%E5%9B%BE%E4%B9%A6%E9%A6%86%E4%B8%BA%E5%BF%AB%E9%80%9F%E6%8E%A8%E8%BF%9B%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E5%9C%A8%E5%9B%BE%E4%B9%A6%E9%A6%86%E7%9A%84%E8%90%BD%E5%9C%B0%E5%BA%94%E7%94%A8%EF%BC%8C%E4%BA%8E2024%E5%B9%B44%E6%9C%8828%E6%97%A5%E6%AD%A3%E5%BC%8F%E4%B8%8A%E7%BA%BF%E4%BA%86AI%E9%98%85%E8%AF%BB%E5%8A%A9%E6%89%8B%E3%80%81AI%E5%AF%BC%E8%88%AA%E5%8A%A9%E6%89%8B%E5%BA%94%E7%94%A8%E3%80%82%E9%80%9A%E8%BF%87%E6%99%BA%E8%83%BD%E5%88%86%E6%9E%90PDF%E6%96%87%E6%A1%A3%E5%92%8C%E6%95%B0%E6%8D%AE%E5%BA%93%E5%86%85%E5%AE%B9%EF%BC%8C%E6%8F%90%E5%8D%87%E4%BA%86%E5%B8%88%20%E7%94%9F%E7%9A%84%E9%98%85%E8%AF%BB%E5%92%8C%E7%A0%94%E7%A9%B6%E6%95%88%E7%8E%87%E3%80%82%E4%B8%A4%E9%A1%B9%E5%BA%94%E7%94%A8%E5%9D%87%E9%9D%A2%E5%90%91%E6%B8%85%E5%8D%8E%E5%85%A8%E6%A0%A1%E5%B8%88%E7%94%9F%E6%8F%90%E4%BE%9B%E6%9C%8D%E5%8A%A1%EF%BC%8C%E4%B8%8A%E7%BA%BF%E5%90%8E%E6%94%B6%E5%88%B0%E4%BA%86%E8%AF%BB%E8%80%85%E7%9A%84%E8%89%AF%E5%A5%BD%E5%8F%8D%E9%A6%88%E3%80%82)。上海图书馆也建设了**读者个人知识中心**，读者可在其中使用智能问答服务获取个性化阅读建议或知识解答[calsp.cn](https://www.calsp.cn/2024/08/21/bulletin-202408-05/" \l ":~:text=%E4%B8%8A%E6%B5%B7%E5%9B%BE%E4%B9%A6%E9%A6%86%E4%B8%93%E4%B8%9A%E6%9C%8D%E5%8A%A1%E4%B8%AD%E5%BF%83%E5%9C%A8%E4%B8%8A%E5%9B%BE%E4%B8%93%E4%B8%9A%E6%9C%8D%E5%8A%A1%E9%97%A8%E6%88%B7%E4%B8%AD%EF%BC%8C%E4%B9%9F%E8%9E%8D%E5%85%A5%E4%BA%86%E5%A4%9A%E4%B8%AAAI%E6%99%BA%E8%83%BD%E8%BE%85%E5%8A%A9%E5%8A%9F%E8%83%BD%E3%80%82%E5%88%9B%E6%96%B0%E6%80%A7%E7%9A%84%E5%B0%86%E5%A4%A7%E8%AF%AD%E8%A8%80%E6%A8%A1%E5%9E%8B%E8%9E%8D%E5%85%A5%E5%9B%BE%E4%B9%A6%E9%A6%86%E8%B5%84%E6%BA%90%E5%8F%91%E7%8E%B0%E6%9C%8D%E5%8A%A1%EF%BC%8C%E5%AE%9E%E7%8E%B0%E4%BA%86%E5%AF%B9%E6%96%87%E7%8C%AE%E6%A3%80%E7%B4%A2%E7%BB%93%E6%9E%9C%E7%9A%84%E5%BF%AB%E9%80%9F%E5%88%86%E6%9E%90%E5%8A%9F%E8%83%BD%E3%80%82%E5%B9%B6%E9%80%9A%E8%BF%87%20%E6%89%93%E9%80%A0%E4%B8%AA%E6%80%A7%E5%8C%96%E7%9A%84%E8%AF%BB%E8%80%85%E4%B8%AA%E4%BA%BA%E7%9F%A5%E8%AF%86%E4%B8%AD%E5%BF%83%EF%BC%8C%E6%8F%90%E4%BE%9B%E6%99%BA%E8%83%BD%E9%97%AE%E7%AD%94%E7%AD%89%E8%BE%85%E5%8A%A9%E9%98%85%E8%AF%BB%E6%9C%8D%E5%8A%A1%EF%BC%8C%E5%85%A8%E9%9D%A2%E6%8F%90%E5%8D%87%E8%AF%BB%E8%80%85%E8%B5%84%E6%BA%90%E6%A3%80%E7%B4%A2%E4%B8%8E%E9%98%85%E8%AF%BB%E4%BD%93%E9%AA%8C%E3%80%82)。这些实践标志着学术图书馆正引入“AI参考馆员”，随时解答读者的研学问题，延伸了人工咨询服务的能力。

**技术成熟度与成本：** 智能问答是目前大模型应用最成熟的方向之一，ChatGPT的出现即为此。图书馆构建自己的问答代理，可以基于通用大模型细调或通过插件接入馆藏数据。例如，OpenAI提供的ChatGPT Plugins允许接入自定义知识库，因此一些图书馆已尝试制作插件，让ChatGPT能够查询本馆OPAC或数据库，为读者提供有馆藏结合的答案[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00" \l ":~:text=%E5%A5%B9%E8%A1%A8%E7%A4%BA%EF%BC%9A%E2%80%9C%E4%BB%8A%E5%A4%A9%E6%88%91%E4%BB%AC%E6%9C%89%E4%BA%86%E8%BF%99%E4%BA%9B%E8%AE%A1%E7%AE%97%E6%9C%BA%EF%BC%8C%E5%AE%83%E4%BB%AC%E7%9C%9F%E7%9A%84%E5%BE%88%E5%BC%BA%E5%A4%A7%EF%BC%8C%E4%BD%86%E6%88%91%E4%BB%AC%E5%BF%85%E9%A1%BB%E5%AF%B9%E5%AE%83%E4%BB%AC%E8%BF%9B%E8%A1%8C%E2%80%98%E5%BE%AE%E8%A7%82%E7%AE%A1%E7%90%86%E2%80%99%E3%80%82%E2%80%9D)。不过目前插件模式尚不稳定，需要技术团队调试维护。另一途径是使用现有对话模型如ChatGLM等，在其上叠加图书馆知识库检索功能（即前述RAG方法）。从效果看，若问题涉及馆藏知识，检索增强能提升准确性；若是一般性知识，模型自身即可回答。成本上，调用通用API按对话长度收费，常规问答成本不高，但长篇总结或连续对话费用会增加。高校图书馆通常有一定经费可支撑初期试点，而公共图书馆资源有限，可能更依赖厂商提供的定制服务（如一些公司把ChatGPT能力包装到图书馆微信公众号插件中销售）。模型维护和内容审查也是成本项，需管理员定期检查AI回答的准确性和不当内容。总的来说，在问答场景应用AI，技术上**门槛较低**、投入可控，但**运营要跟上**，否则错误答案或胡言乱语可能引发负面影响。

**用户体验提升：** 智能问答的优势在于：**即时、便利、交互自然**。与等待馆员邮件答复或亲临咨询台相比，读者更乐于在网页或APP上直接提问并立刻收到回答。对于常见简易问题，AI助手减少了馆员重复劳动；对于复杂问题，AI能提供初步答案，再由馆员跟进深化。用户调查显示，读者喜欢AI助手的**7×24可用性**，甚至半夜遇到问题也能获得帮助，这是人工服务难以做到的[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00" \l ":~:text=%E5%BD%93%E5%89%8D%E8%BF%98%E5%AD%98%E5%9C%A8%E5%93%AA%E4%BA%9B%E9%99%90%E5%88%B6%EF%BC%9F)[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00#:~:text=%E5%AE%9E%E9%99%85%E4%B8%8A%EF%BC%8C%E4%BB%8E%E6%9F%90%E7%A7%8D%E7%A8%8B%E5%BA%A6%E6%9D%A5%E8%AF%B4%EF%BC%8C%E4%BD%A0%E5%BE%88%E5%8F%AF%E8%83%BD%E5%B7%B2%E7%BB%8F%E4%BD%93%E9%AA%8C%E8%BF%87%E5%AE%83%E4%BB%AC%E7%9A%84%E6%97%A9%E6%9C%9F%E5%8E%9F%E5%9E%8B%EF%BC%8C%E6%AF%94%E5%A6%82%20OpenAI%20%E7%9A%84%20ChatGPT%20%E5%92%8C,4%E3%80%82%E2%80%9C%E5%A6%82%E6%9E%9C%E4%BD%A0%E6%AD%A3%E5%9C%A8%E4%B8%8E%E6%84%9F%E8%A7%89%E5%BE%88%E6%99%BA%E8%83%BD%E7%9A%84%E8%BD%AF%E4%BB%B6%E8%BF%9B%E8%A1%8C%E4%BA%A4%E4%BA%92%EF%BC%8C%E9%82%A3%E5%AE%83%E5%85%B6%E5%AE%9E%E5%B0%B1%E6%98%AF%E4%B8%80%E7%A7%8D%E6%99%BA%E8%83%BD%E4%BD%93%E3%80%82%E2%80%9D%E9%82%B1%E4%BE%83%E5%90%9B%E8%AF%B4%E3%80%82)。此外，AI可以不停询问追问，实现**多轮对话**厘清需求，而人工咨询有时受限于时间难以深入。需要注意的是，用户体验也取决于AI回答质量。目前AI问答容易出现的信息幻觉或不精确之处[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00" \l ":~:text=%E9%82%B1%E4%BE%83%E5%90%9B%E8%AF%B4%EF%BC%8C%E2%80%9C%E4%B8%BE%E4%B8%AA%E4%BE%8B%E5%AD%90%EF%BC%8C%E7%BC%96%E7%A0%81%E6%99%BA%E8%83%BD%E4%BD%93%E5%8F%AF%E4%BB%A5%E7%94%9F%E6%88%90%E4%BB%A3%E7%A0%81%EF%BC%8C%E4%BD%86%E6%9C%89%E6%97%B6%E4%BC%9A%E5%87%BA%E9%94%99%EF%BC%8C%E8%80%8C%E4%B8%94%E5%AE%83%E4%B9%9F%E4%B8%8D%E7%9F%A5%E9%81%93%E5%A6%82%E4%BD%95%E6%B5%8B%E8%AF%95%E6%AD%A3%E5%9C%A8%E5%88%9B%E5%BB%BA%E7%9A%84%E4%BB%A3%E7%A0%81%E3%80%82%E5%9B%A0%E6%AD%A4%EF%BC%8C%E4%BA%BA%E7%B1%BB%E4%BB%8D%E7%84%B6%E9%9C%80%E8%A6%81%E5%8F%82%E4%B8%8E%E5%88%B0%E8%BF%99%E4%B8%80%E8%BF%87%E7%A8%8B%E4%B8%AD%E3%80%82%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E7%B3%BB%E7%BB%9F%E4%BB%8D%E7%84%B6%E6%97%A0%E6%B3%95%E5%AE%8C%E5%85%A8%E5%AE%9E%20%E7%8E%B0%E6%8E%A8%E7%90%86%EF%BC%8C%E8%BF%99%E6%98%AF%E5%9C%A8%E5%A4%8D%E6%9D%82%E4%B8%94%E6%A8%A1%E7%B3%8A%E7%9A%84%E4%BA%BA%E7%B1%BB%E4%B8%96%E7%95%8C%E4%B8%AD%E8%BF%90%E8%A1%8C%E7%9A%84%E5%85%B3%E9%94%AE%E4%B8%80%E6%AD%A5%E3%80%82%E2%80%9D)。图书馆通过将AI回答与可信来源关联，可以提高用户信任。例如AI回答后列出相关书目索引或数据库链接，用户看到依据就更信服[calsp.cn](https://www.calsp.cn/2024/08/21/bulletin-202408-05/" \l ":~:text=RAG%E9%80%9A%E8%BF%87%E6%A3%80%E7%B4%A2%E8%8E%B7%E5%8F%96%E7%9B%B8%E5%85%B3%E7%9A%84%E7%9F%A5%E8%AF%86%E5%B9%B6%E5%B0%86%E5%85%B6%E8%9E%8D%E5%85%A5Prompt%E6%8F%90%E7%A4%BA%E8%AF%8D%EF%BC%8C%E8%AE%A9%E5%A4%A7%E8%AF%AD%E8%A8%80%E6%A8%A1%E5%9E%8B%E8%83%BD%E5%A4%9F%E5%8F%82%E8%80%83%E7%9B%B8%E5%BA%94%E7%9A%84%E7%9F%A5%E8%AF%86%E4%BB%8E%E8%80%8C%E7%BB%99%E5%87%BA%E5%90%88%E7%90%86%E5%9B%9E%E7%AD%94%E3%80%82%E5%9B%A0%E6%AD%A4%EF%BC%8C%E5%8F%AF%E4%BB%A5%E5%B0%86RAG%E7%9A%84%E6%A0%B8%E5%BF%83%E7%90%86%E8%A7%A3%E4%B8%BA%E2%80%9C%E6%A3%80%E7%B4%A2%2B%E7%94%9F%E6%88%90%E2%80%9D%EF%BC%8C%E5%89%8D%E8%80%85%E4%B8%BB%E8%A6%81%20%E6%98%AF%E5%88%A9%E7%94%A8%E5%90%91%E9%87%8F%E6%95%B0%E6%8D%AE%E5%BA%93%E7%9A%84%E9%AB%98%E6%95%88%E5%AD%98%E5%82%A8%E5%92%8C%E6%A3%80%E7%B4%A2%E8%83%BD%E5%8A%9B%EF%BC%8C%E5%8F%AC%E5%9B%9E%E7%9B%AE%E6%A0%87%E7%9F%A5%E8%AF%86%EF%BC%9B%E5%90%8E%E8%80%85%E5%88%99%E6%98%AF%E5%88%A9%E7%94%A8%E5%A4%A7%E8%AF%AD%E8%A8%80%E6%A8%A1%E5%9E%8B%E5%92%8CPrompt%E6%8F%90%E7%A4%BA%E8%AF%8D%E5%B7%A5%E7%A8%8B%EF%BC%8C%E5%B0%86%E5%8F%AC%E5%9B%9E%E7%9A%84%E7%9F%A5%E8%AF%86%E5%90%88%E7%90%86%E5%88%A9%E7%94%A8%EF%BC%8C%E7%94%9F%E6%88%90%E7%9B%AE%E6%A0%87%E7%AD%94%E6%A1%88%E3%80%82)。在实践中，大部分用户将AI问答视为**辅助工具**：先得到AI解答，对思路有了方向，再结合参考资料进行验证。可以预见，随着模型水平提高和知识库融合增强，AI问答将在图书馆咨询服务中扮演愈发重要的角色，让知识获取更**即时化和个性化**。

### **3.4.4 个性化推荐：用户画像驱动的资源推荐**

**场景定义：** 图书馆累积了丰富的用户借阅、浏览数据，可以通过分析用户兴趣和行为，为其**个性化推荐**可能感兴趣的书籍、文章、活动等，实现“千人千面”的服务。传统推荐多基于规则或协同过滤，而智能体可引入更智能的**用户画像**和推荐策略：通过大模型理解用户的显性需求和隐性偏好，从馆藏和外部资源中主动挑选匹配的内容推送给用户。在图书馆阅读推广中，个性化推荐能提高读者与资源的匹配度，促进资源利用和阅读量提升。

**案例与实践：** 许多高校和公共图书馆已经在门户网站或App中上线了智能推荐功能：

* *高校图书馆阅读推荐☆* – 南京理工大学图书馆开发了“**悦见书苑**”微信小程序，利用用户画像为每位用户提供个性化阅读推荐[nju.edu.cn](https://www.nju.edu.cn/info/1056/340981.htm" \l ":~:text=%E9%AB%98%E6%A0%A1%E5%9B%BE%E4%B9%A6%E9%A6%86%E5%A6%82%E4%BD%95%E5%BC%80%E5%B1%95%E9%98%85%E8%AF%BB%E6%8E%A8%E5%B9%BF%EF%BC%9F%E6%B1%9F%E8%8B%8F%E8%BF%99%E6%89%B9%E7%B2%BE%E5%93%81%E6%A1%88%E4%BE%8B%E7%BB%99%E4%BD%A0%E7%AD%94%E6%A1%88%20)。该系统收集用户的借阅历史、浏览点击、点赞评价等信息，结合读者专业年级等属性，构建**读者模型**。然后，通过混合推荐算法（协同过滤+内容推荐+深度学习模型），每天为用户推送定制书目清单。例如，一名机械工程专业的大三学生，系统可能推荐“机器学习”“工业4.0”等领域的新书，再辅以根据其平时借阅的小说类型推送相关文学作品。江苏省智慧图书馆论坛上，该案例获得优秀奖，认为其“用有深度的数据，做有温暖的服务”[nju.edu.cn](https://www.nju.edu.cn/info/1056/340981.htm#:~:text=%E9%AB%98%E6%A0%A1%E5%9B%BE%E4%B9%A6%E9%A6%86%E5%A6%82%E4%BD%95%E5%BC%80%E5%B1%95%E9%98%85%E8%AF%BB%E6%8E%A8%E5%B9%BF%EF%BC%9F%E6%B1%9F%E8%8B%8F%E8%BF%99%E6%89%B9%E7%B2%BE%E5%93%81%E6%A1%88%E4%BE%8B%E7%BB%99%E4%BD%A0%E7%AD%94%E6%A1%88%20)。背后的技术亮点在于融合了多种推荐方法、引入标签语义分析，显著提升了推荐准确率[manu44.magtech.com.cn](https://manu44.magtech.com.cn/Jwk_infotech_wk3/article/2012/1003-3513/1003-3513-28-6-1.html" \l ":~:text=%E5%9F%BA%E4%BA%8E%E5%A4%9A%E5%85%B4%E8%B6%A3%E7%89%B9%E5%BE%81%E5%88%86%E6%9E%90%E7%9A%84%E5%9B%BE%E4%B9%A6%E9%A6%86%E4%B8%AA%E6%80%A7%E5%8C%96%E5%9B%BE%E4%B9%A6%E6%8E%A8%E8%8D%90%E6%96%B9%E6%B3%95%20%E5%9F%BA%E4%BA%8E%E5%86%85%E5%AE%B9%E6%8E%A8%E8%8D%90%E6%96%B9%E6%B3%95%E6%98%AF%E9%80%9A%E8%BF%87%E8%AF%BB%E8%80%85%E4%BB%A5%E5%BE%80%E5%80%9F%E9%98%85%E8%BF%87%E7%9A%84%E4%B9%A6%E7%B1%8D%E7%9A%84%E7%9B%B8%E5%85%B3%E4%BF%A1%E6%81%AF%E4%B8%8E%E7%8E%B0%E6%9C%89%E4%B9%A6%E7%B1%8D%E4%B9%8B%E9%97%B4%E7%9A%84%E5%8C%B9%E9%85%8D%E7%A8%8B%E5%BA%A6%E8%BF%9B%E8%A1%8C%E8%BF%87%E6%BB%A4%E6%8E%A8%E8%8D%90%E3%80%82%E5%9F%BA%E4%BA%8E%E5%86%85%E5%AE%B9%E6%8E%A8%E8%8D%90%E7%9A%84%E4%BC%98%E7%82%B9%E4%B8%BB%E8%A6%81%E6%98%AF%E5%8F%AF%E4%BB%A5%E5%A4%84%E7%90%86%E6%96%B0%E7%94%A8%E6%88%B7%E5%92%8C%E6%96%B0%E9%A1%B9%E7%9B%AE%E9%97%AE%E9%A2%98%2C%E5%8D%B3%E5%86%B7%E5%90%AF%E5%8A%A8%E9%97%AE%E9%A2%98%E3%80%82%E4%BD%86%E6%98%AF%20)。据统计，上线后推荐书单的点击率和后续借阅率有明显增长，说明读者确实从中发现了感兴趣的资源。
* *公共图书馆精准推送☆* – 佛山市图书馆建设的“易本书”平台不仅用于共享民间藏书，也具备个性化推荐功能[whly.gd.gov.cn](https://whly.gd.gov.cn/service_newggwhzx/content/post_4061139.html" \l ":~:text=%E4%BE%8B%E3%80%82%E5%B9%BF%E4%B8%9C%E7%9C%81%E6%88%90%E5%8A%9F%E9%80%89%E9%80%815%E4%B8%AA%E7%9C%81%E5%86%85%E9%A1%B9%E7%9B%AE%EF%BC%8C%E5%88%86%E5%88%AB%E6%98%AF%E5%B9%BF%E4%B8%9C%E7%9C%81%E7%AB%8B%E4%B8%AD%E5%B1%B1%E5%9B%BE%E4%B9%A6%E9%A6%86%E7%9A%84%E2%80%9C%E3%80%8A%E5%8D%8E%E5%95%86%E6%8A%A5%E3%80%8B%E6%8A%A5%E7%BA%B8%E6%95%B0%E5%AD%97%E5%8C%96%E5%B1%95%E7%A4%BA%E5%B9%B3%E5%8F%B0%E2%80%9D%E5%92%8C%E2%80%9C%E5%9B%BE%E4%B9%A6%E9%87%87%E5%88%86%E7%BC%96%E6%99%BA%E8%83%BD%E4%BD%9C%E4%B8%9A%E7%B3%BB%E7%BB%9F%E2%80%94%E2%80%94%E9%87%87%E7%BC%96%E5%9B%BE%E7%81%B5%E2%80%9D%E3%80%81%E4%BD%9B%E5%B1%B1%E5%B8%82%E5%9B%BE%E4%B9%A6%E9%A6%86%E7%9A%84%E2%80%9C%E6%98%93%E6%9C%AC%E4%B9%A6%E2%80%9D%E5%AE%B6%E8%97%8F%20%E5%9B%BE%E4%B9%A6%E5%85%B1%E4%BA%AB%E5%B9%B3%E5%8F%B0%E3%80%81%E6%B7%B1%E5%9C%B3%E5%9B%BE%E4%B9%A6%E9%A6%86%E7%9A%84%E2%80%9C%E6%96%B9%E5%BF%97%E9%87%8C%E7%9A%84%E6%B7%B1%E5%9C%B3%E2%80%9D%E5%B0%8F%E7%A8%8B%E5%BA%8F%E3%80%81%E5%B9%BF%E5%B7%9E%E5%B8%82%E5%8D%97%E6%B2%99%E5%8C%BA%E5%9B%BE%E4%B9%A6%E9%A6%86%E7%9A%84%E2%80%9C%E5%9C%B0%E7%A3%81%E5%AF%BC%E8%88%AA%E2%80%9D%E3%80%82)[whly.gd.gov.cn](https://whly.gd.gov.cn/service_newggwhzx/content/post_4061139.html#:~:text=%E4%BD%9B%E5%B1%B1%E5%B8%82%E5%9B%BE%E4%B9%A6%E9%A6%86%E2%80%9C%E6%98%93%E6%9C%AC%E4%B9%A6%E2%80%9D%E5%B9%B3%E5%8F%B0%E6%98%AF%E5%9F%BA%E4%BA%8E%E5%8C%BA%E5%9D%97%E9%93%BE%E6%8A%80%E6%9C%AF%E7%9A%84%E5%AE%B6%E8%97%8F%E5%9B%BE%E4%B9%A6%E6%B5%81%E9%80%9A%E5%B9%B3%E5%8F%B0%2C%E6%97%A8%E5%9C%A8%E8%AE%A9%E5%AE%B6%E5%BA%AD%E8%97%8F%E4%B9%A6%E5%92%8C%E5%85%AC%E5%85%B1%E5%9B%BE%E4%B9%A6%E9%A6%86%E8%97%8F%E4%B9%A6%E2%80%94%E8%B5%B7%E6%B5%81%E5%8A%A8%E8%B5%B7%E6%9D%A5%2C%E4%B8%B0%E5%AF%8C%E8%B5%84%E6%BA%90%E4%BE%9B%E7%BB%99%2C%E5%AE%9E%E7%8E%B0%E5%85%A8%E7%A4%BE%E4%BC%9A%E5%AE%B6%E5%BA%AD%E8%97%8F%E4%B9%A6%E7%9A%84%E5%85%B1%E4%BA%AB%E4%B8%8E%E6%B5%81%E9%80%9A%E3%80%82%E9%80%9A%E8%BF%87%E8%AF%A5%E5%B9%B3%20%E5%8F%B0%2C%E5%B8%82%E6%B0%91%E5%8F%AF%E4%BB%A5%E4%B8%8A%E4%BC%A0%E5%AE%B6%E4%B8%AD%E9%97%B2%E7%BD%AE%E7%9A%84%E5%9B%BE%E4%B9%A6%E4%BE%9B%E4%BA%BA%E5%80%9F%E9%98%85%2C%E4%B9%9F%E5%8F%AF%E4%BB%A5%E5%80%9F%E9%98%85%E5%88%AB%E4%BA%BA%E4%B8%8A%E4%BC%A0%E7%9A%84%E8%97%8F%E4%B9%A6%E3%80%82%E4%B8%8A%E4%BC%A0%E7%9A%84%E5%9B%BE%E4%B9%A6%E5%8F%AF%E4%BB%A5%E9%80%89%E6%8B%A9%E9%99%90%E6%9C%9F%E5%80%9F%E9%98%85%E3%80%81%E6%BC%82%E6%B5%81%E3%80%81%E8%B5%A0%E9%80%81%E3%80%81%E5%B1%95%E7%A4%BA%E5%92%8C%E6%99%BA%E8%83%BD%E5%8C%96%E7%AE%A1%E7%90%86%E7%AD%89%E5%A4%9A%E7%A7%8D%E6%B5%81%E9%80%9A%E6%A8%A1%E5%BC%8F%E3%80%82%E5%B8%82%E6%B0%91%E4%B8%8D%E4%BB%85%E5%8F%AF%E4%BB%A5%E4%B8%80%E9%94%AE%E6%90%9C%E7%B4%A2%E6%9C%AC%20%E7%AB%99%E5%9B%BE%E4%B9%A6%E8%B5%84%E6%BA%90%2C%E8%BF%98%E5%8F%AF%E8%8E%B7%E5%8F%96%E5%85%AC%E5%85%B1%E5%9B%BE%E4%B9%A6%E9%A6%86%E8%97%8F%E4%B9%A6%E6%8C%87%E5%BC%95%2C%E5%AE%9E%E7%8E%B0%E4%BA%86%E5%85%AC%E7%A7%81%E8%97%8F%E4%B9%A6%E7%9A%84%E8%9E%8D%E5%90%88%E5%88%A9%E7%94%A8%2C%E6%8F%90%E9%AB%98%E4%BA%86%E5%9B%BE%E4%B9%A6%E8%8E%B7%E5%8F%96%E6%95%88%E7%8E%87%2C%E6%BB%A1%E8%B6%B3%E4%BA%86%E5%B8%82%E6%B0%91%E9%98%85%E8%AF%BB%E9%9C%80%E6%B1%82%E3%80%82)。系统根据用户兴趣爱好抽象出**读者模型**，将预设的用户个性标签与图书标签关联起来，通过计算相关度权重值向用户自动推荐书籍，实现精准推荐[whly.gd.gov.cn](https://whly.gd.gov.cn/gkmlpt/content/4/4404/post_4404325.html" \l ":~:text=%E4%BD%9B%E5%B1%B1%E2%80%9C%E6%98%93%E6%9C%AC%E4%B9%A6%E2%80%9D%E9%A1%B9%E7%9B%AE%E5%85%A5%E9%80%89%E6%96%87%E5%8C%96%E5%92%8C%E6%97%85%E6%B8%B8%E9%83%A8%E6%99%BA%E6%85%A7%E5%9B%BE%E4%B9%A6%E9%A6%86%E5%88%9B%E6%96%B0%E5%BA%94%E7%94%A8%E4%BC%98%E7%A7%80%E6%A1%88%E4%BE%8B%20%E6%A0%B9%E6%8D%AE%E7%94%A8%E6%88%B7%E5%85%B4%E8%B6%A3%E7%88%B1%E5%A5%BD%E7%AD%89%E4%BF%A1%E6%81%AF%EF%BC%8C%E6%8A%BD%E8%B1%A1%E5%87%BA%E8%AF%BB%E8%80%85%E6%A8%A1%E5%9E%8B%EF%BC%8C%E5%B0%86%E9%A2%84%E8%AE%BE%E7%9A%84%E7%94%A8%E6%88%B7%E4%B8%AA%E6%80%A7%E5%8C%96%E6%A0%87%E7%AD%BE%E5%92%8C%E5%9B%BE%E4%B9%A6%E6%A0%87%E7%AD%BE%E5%85%B3%E8%81%94%EF%BC%8C%E8%87%AA%E5%8A%A8%E6%A0%B9%E6%8D%AE%E9%A2%84%E8%AE%BE%E7%9B%B8%E5%85%B3%E5%BA%A6%E6%9D%83%E9%87%8D%E5%80%BC%E5%90%91%E7%94%A8%E6%88%B7%E6%8E%A8%E8%8D%90%E5%9B%BE%E4%B9%A6%EF%BC%8C%E5%AE%9E%E7%8E%B0%E7%B2%BE%E5%87%86%E5%8C%96%E3%80%81%E4%B8%AA%E6%80%A7%E5%8C%96%E6%8E%A8%E8%8D%90%E3%80%82)。简单说，如果某读者在平台上表现出对“粤菜烹饪”“本地历史”的兴趣（通过其浏览、关注行为），系统就会定期推送相关的新书信息或活动通知。这是以用户标签匹配图书标签的思路。再如深圳图书馆在其借阅系统中应用了借阅记录驱动的“猜你喜欢”功能，当读者在线选书时，界面会提示“看过这本书的人还借过哪些书”，借助大数据关联实现关联推荐。这些都属于个性化推荐的实践形式，充分调动馆藏长尾资源供给潜力，让冷门资源找到对的人。

**技术成熟度与成本：** 个性化推荐在电商等行业已非常成熟，图书馆领域相对起步稍晚。目前实现途径主要有：基于专业推荐平台（如 Ex Libris 的 bX Recommender 专为学术文献推荐[knowledge.exlibrisgroup.com](https://knowledge.exlibrisgroup.com/Summon/Product_Materials/Summon%3A_bX_Recommender_Integration" \l ":~:text=bX%20is%20a%20service%20available,based%20scholarly%20article%20recommendations)）、或自建模型系统。Ex Libris bX利用全球学术资源的**链接解析使用数据**，为文章提供“相关文献推荐”，很多图书馆通过其API在发现系统中呈现文章推荐列表[carli.illinois.edu](https://www.carli.illinois.edu/products-services/i-share/discovery-interface/bxrecommender" \l ":~:text=CARLI%20www,and%20ebooks%20in%20Primo)[stephenfrancoeur.com](https://www.stephenfrancoeur.com/beatingthebounds/#:~:text=Ex%20Libris%20provides%20this%20quick,data%20from%20many%20academic)。这属于协同过滤类方法，学术界验证其可有效提升参考文献发现。对于通俗读物，国内一些厂商也推出了图书推荐引擎。自建方面，借助开源算法库（如 Surprise 库）和大模型（用于标签抽取、摘要分析）也能搭建个性推荐系统。但挑战在于图书馆用户行为数据相对有限，且有**冷启动**问题（新用户、冷门书籍缺少历史数据）。AI智能体可通过**知识迁移和内容分析**部分缓解此问题：例如用大模型根据图书简介和评论推测该书适合的人群标签，从而给无历史记录的新书找到潜在读者。成本上，个性化推荐系统需要持续维护用户和资源特征库，计算开销不大但需投入数据清洗和策略调优的人力。许多馆选择购买成熟服务，如bX服务按年订阅，价格视馆藏规模和用户数而定，一般可承受。

**用户体验提升：** 个性化推荐的好处是让读者**惊喜地发现**自己可能喜欢但不知道的资源，提高用户粘性和满意度。调查显示，当推荐准确时，用户乐于接受AI推荐并觉得图书馆服务更贴心。例如通过推荐功能，读者A在看完一本书后得知馆里还有类似风格的书B，于是顺利续接阅读兴趣。这种体验在数字内容平台很常见，现在图书馆也能提供，帮助用户**挖掘长尾知识**。另一方面，错误或无关推荐会降低体验。因此推荐质量很关键，需要一定的调整试验。幸运的是，图书馆的推荐相对不涉及商业利益，可以更纯粹地以读者兴趣为导向，不像电商那样充斥广告，这使读者更愿意相信和使用。国内实践（如三峡大学图书馆的读者画像推荐）表明，好的推荐系统能让读者感觉服务更懂自己，图书借阅率也有所提升[nju.edu.cn](https://www.nju.edu.cn/info/1056/340981.htm" \l ":~:text=%E9%AB%98%E6%A0%A1%E5%9B%BE%E4%B9%A6%E9%A6%86%E5%A6%82%E4%BD%95%E5%BC%80%E5%B1%95%E9%98%85%E8%AF%BB%E6%8E%A8%E5%B9%BF%EF%BC%9F%E6%B1%9F%E8%8B%8F%E8%BF%99%E6%89%B9%E7%B2%BE%E5%93%81%E6%A1%88%E4%BE%8B%E7%BB%99%E4%BD%A0%E7%AD%94%E6%A1%88%20)[journal12.magtechjournal.com](https://journal12.magtechjournal.com/jwk3_qbkx/CN/abstract/abstract10300.shtml#:~:text=%E5%88%9B%E6%96%B0%2F%E5%B1%80%E9%99%90%20%E5%9F%BA%E4%BA%8E%E7%94%A8%E6%88%B7%E7%94%BB%E5%83%8F%E7%9A%84%E5%9B%BE%E4%B9%A6%E6%8E%A8%E8%8D%90%E4%BB%8E%E5%A4%9A%E4%B8%AA%E7%BB%B4%E5%BA%A6%E8%BF%9B%E8%A1%8C%E7%BB%84%E5%90%88%E6%8E%A8%E8%8D%90%EF%BC%8C%E5%AE%9E%E7%8E%B0%E4%BA%86%E6%99%BA%E6%85%A7%E5%9B%BE%E4%B9%A6%E9%A6%86%E8%99%9A%E6%8B%9F%E5%9B%BE%E4%B9%A6%E5%92%8C%E5%AE%9E%E4%BD%93%E5%9B%BE%E4%B9%A6%E7%9A%84%E4%B8%AA%E6%80%A7%E5%8C%96%E6%8E%A8%E8%8D%90%EF%BC%8C%E6%8F%90%E9%AB%98%E4%BA%86%E6%8E%A8%E8%8D%90%E8%B4%A8%E9%87%8F%EF%BC%8C%E4%B8%BA%E6%8F%90%E5%8D%87%E6%99%BA%E6%85%A7%E5%9B%BE%E4%B9%A6%E9%A6%86%E4%B8%AA%E6%80%A7%E5%8C%96%E6%9C%8D%E5%8A%A1%E5%85%B7%E6%9C%89%E4%B8%80%E5%AE%9A%E7%9A%84%20%E5%80%9F%E9%89%B4%20)。当然，隐私顾虑需考虑：用户可能担心AI搜集了自己的阅读隐私。因此图书馆必须**透明告知**推荐所用数据、保护隐私的措施，让用户放心使用并可选择退出。如果处理得当，个性化推荐将成为智慧图书馆中“有温度”的服务，让每位读者感受到被关注、被理解，从而更积极地参与阅读和利用馆藏。

### **3.4.5 知识推送：主动式知识按需服务**

**场景定义：** 知识推送指图书馆根据用户需求偏好，主动将最新的、相关的知识内容发送给用户，而非等用户来检索。它类似精准的信息“订阅”和“通报”服务，是传统选择传播(SDI)服务在AI时代的升级。智能体可充当信息经纪人，持续监测用户感兴趣的领域动态，在适当时机推送**新书通告、学科热点综述、定制资讯简报**等到用户终端（邮件、微信等），实现从“人找信息”到“信息找人”的转变。

**案例与实践：** 知识推送在科研情报服务中早已有之，如课题新文献通报等。AI的加入让推送更个性化、智能化：

* *科研知识快报服务☆* – 中国科学院国家科学图书馆多年来为科研人员提供定制的信息通报服务。现在，他们开发了智能信息监测系统，研究人员可以指定自己的研究主题或课题关键词，系统每日爬取全球文献数据库、新闻网站，**自动生成定制资讯简报**发送用户邮箱。例如某学者研究石墨烯材料，他设置相关主题后，每周将收到一份AI整合的报告，包括本周石墨烯领域发表的论文清单、专利信息，甚至会议动态。AI在其中负责对海量信息进行筛选、分类和简洁呈现，用户无需手动检索多个库即可及时掌握进展。这类服务已有一些第三方提供者（如Dimension的Alerts功能），图书馆通过AI可进一步提升本地化和深度定制程度。
* *公共文化知识推送☆* – 疫情期间，北京、上海等多地图书馆曾开展“数字资源免费推送”服务[cwu.edu.cn](https://www.cwu.edu.cn/tsg/info/1342/6831.htm" \l ":~:text=%E7%AC%AC%E4%BA%8C%E6%9C%9F%E6%8E%A8%E9%80%81)。馆员精选适合居家学习的电子书、网络课程，定期通过网站或公众号推送给公众。这可以视为广义的知识推送。未来，智能体可根据每个用户的职业、教育背景等提供**差异化推送**。比如针对教师用户定期推送教学参考资料，针对创业者推送行业报告。广东图书馆联盟也鼓励各馆利用读者数据实施精准推送服务[whly.gd.gov.cn](https://whly.gd.gov.cn/gkmlpt/content/4/4404/post_4404325.html" \l ":~:text=%E4%BD%9B%E5%B1%B1%E2%80%9C%E6%98%93%E6%9C%AC%E4%B9%A6%E2%80%9D%E9%A1%B9%E7%9B%AE%E5%85%A5%E9%80%89%E6%96%87%E5%8C%96%E5%92%8C%E6%97%85%E6%B8%B8%E9%83%A8%E6%99%BA%E6%85%A7%E5%9B%BE%E4%B9%A6%E9%A6%86%E5%88%9B%E6%96%B0%E5%BA%94%E7%94%A8%E4%BC%98%E7%A7%80%E6%A1%88%E4%BE%8B%20%E6%A0%B9%E6%8D%AE%E7%94%A8%E6%88%B7%E5%85%B4%E8%B6%A3%E7%88%B1%E5%A5%BD%E7%AD%89%E4%BF%A1%E6%81%AF%EF%BC%8C%E6%8A%BD%E8%B1%A1%E5%87%BA%E8%AF%BB%E8%80%85%E6%A8%A1%E5%9E%8B%EF%BC%8C%E5%B0%86%E9%A2%84%E8%AE%BE%E7%9A%84%E7%94%A8%E6%88%B7%E4%B8%AA%E6%80%A7%E5%8C%96%E6%A0%87%E7%AD%BE%E5%92%8C%E5%9B%BE%E4%B9%A6%E6%A0%87%E7%AD%BE%E5%85%B3%E8%81%94%EF%BC%8C%E8%87%AA%E5%8A%A8%E6%A0%B9%E6%8D%AE%E9%A2%84%E8%AE%BE%E7%9B%B8%E5%85%B3%E5%BA%A6%E6%9D%83%E9%87%8D%E5%80%BC%E5%90%91%E7%94%A8%E6%88%B7%E6%8E%A8%E8%8D%90%E5%9B%BE%E4%B9%A6%EF%BC%8C%E5%AE%9E%E7%8E%B0%E7%B2%BE%E5%87%86%E5%8C%96%E3%80%81%E4%B8%AA%E6%80%A7%E5%8C%96%E6%8E%A8%E8%8D%90%E3%80%82)。以深圳图为例，他们有大量高科技企业员工读者，AI可帮助监测这些领域的新标准、新报告，并主动发送给相关读者群。这种按需推送把图书馆的信息服务融入用户日常工作流。

**技术成熟度与成本：** 知识推送涉及信息监测、内容过滤加工、用户匹配等环节。AI擅长的在于**信息筛选摘要**，如利用大模型生成简报摘要，让推送内容精炼易读。难点则在实时监测需要持续爬取外部信息，这需一定IT投入和版权考虑。一些馆选择与商业情报数据库合作，由后者提供定制RSS流再经AI加工。成本上，推送服务是增值服务，需要编制推送主题知识库，AI调用的次数按订阅用户计，需控制频率以节约费用。另外，要注意避免信息过载和打扰用户：推送过勤可能引反感，因此智能体应学习用户反馈，调整推送频率和内容。总体说，目前AI知识推送在图书馆还处于初步试水阶段，技术上已经可行，但运营模式（如收费或免费、用户获取等）尚在探索。

**用户体验提升：** 对忙碌的研究者或无法常来图书馆的公众来说，知识推送提供了极大便利：他们可以**坐等最新信息送达**。这有助于保持对领域的掌握，避免遗漏重要进展。在快节奏时代，用户往往没有时间每天搜集信息，图书馆替其打理信息源，提供**私人情报助手**的感觉，会增加用户黏性。另外，推送服务能提高馆藏数字资源的利用率。比如图书馆购买了许多数据库，但一般用户不知道，通过推送文章给他们阅读，就有效提升了这些资源的曝光和使用。有案例显示，自从实施定制推送后，一些专门数据库的下载量增长显著，因为AI推送激发了用户点击查全文的行为。需要照顾的是不同用户的信息需求层次，有人喜欢简报摘要，有人愿意浏览全清单，因此推送内容可以多层次展现。同时，应提供易于取消或调整订阅的途径，让用户掌控接收偏好。只要推送做到**相关且不骚扰**，大多数用户会欢迎这种贴心服务，因为它**节时高效**且有一种被服务的尊贵感。在知识过载的背景下，图书馆智能推送有潜力成为区分服务水平的亮点。

### 3.4.6 可视化服务：知识可视化与交互展示

**场景定义：** 可视化服务指利用图表、地图、网络等可视化方式呈现知识和数据，帮助用户更直观理解信息。图书馆积累了大量书目、引文、使用数据，借助可视化可以展示比如学科知识结构、科研合作网络、借阅趋势等。智能体可以自动生成、更新这些可视化图谱，并与用户交互，形成图书馆的新型信息服务形态。例如构建区域历史事件的时空地图、学科知识演化的动态知识图谱，让用户以沉浸方式探索知识。

**案例与实践：** 可视化在图书馆已有一些应用，AI的引入让其更易实现和互动：

* *科学知识图谱与分析☆* – 中国科学技术信息研究所和国家图书馆等单位曾开发“数字学术地图”等服务，利用CiteSpace等软件绘制国际科研合作网络、主题演化图等[dag.pku.edu.cn](https://www.dag.pku.edu.cn/academic_details/14.html" \l ":~:text=,%E5%9B%BE3%E4%B8%AD%E5%9C%86%E5%9C%88%E8%8A%82%E7%82%B9%E5%A4%A7%E5%B0%8F%E4%B8%8E%E4%BD%9C%E8%80%85%E8%A2%AB%E5%BC%95)。这些以前需要信息分析员操作，现在智能体可接管相当部分工作。例如在馆员提交一组文献数据后，AI分析引文关系自动生成知识图谱并高亮核心作者群体[whly.gd.gov.cn](https://whly.gd.gov.cn/service_newggwhzx/content/post_4061139.html" \l ":~:text=%E3%80%8A%E5%8D%8E%E5%95%86%E6%8A%A5%E3%80%8B%E6%8A%A5%E7%BA%B8%E6%95%B0%E5%AD%97%E5%8C%96%E5%B1%95%E7%A4%BA%E5%B9%B3%E5%8F%B0%E4%BB%A5%E9%A6%86%E8%97%8F%E4%B8%8A%E4%B8%96%E7%BA%AA40%E5%B9%B4%E4%BB%A3%E3%80%8A%E5%8D%8E%E5%95%86%E6%8A%A5%E3%80%8B%E4%B8%BA%E5%88%87%E5%85%A5%E7%82%B9%EF%BC%8C%E5%B0%86%E7%9F%A5%E8%AF%86%E6%88%90%E6%9E%9C%E9%80%9A%E8%BF%87%E4%B8%BB%E9%A2%98%E5%8C%96%E3%80%81%E4%B8%93%E9%A2%98%E5%8C%96%E6%96%B9%E5%BC%8F%EF%BC%8C%E5%BC%95%E5%85%A5%E2%80%9C%E7%9F%A5%E8%AF%86%E6%98%9F%E7%90%83%E2%80%9D%E7%9A%84%E6%A6%82%E5%BF%B5%EF%BC%8C%E5%88%A9%E7%94%A8%E7%9F%A5%E8%AF%86%E5%85%B3%E7%B3%BB%E5%9B%BE%E8%B0%B1%E5%AF%B9%E6%8A%A5%E7%BA%B8%E5%86%85%E5%AE%B9%E8%BF%9B%E8%A1%8C%E6%A2%B3%E7%90%86%20%E7%BB%84%E7%BB%87%EF%BC%8C%E7%BB%93%E5%90%88%E6%96%B0%E9%A2%96%E7%9A%84%E4%BA%A4%E4%BA%92%E6%8A%80%E6%9C%AF%EF%BC%8C%E4%B8%BA%E8%AF%BB%E8%80%85%E6%8F%90%E4%BE%9B%E5%9F%BA%E4%BA%8E%E6%99%BA%E6%85%A7%E5%9B%BE%E4%B9%A6%E9%A6%86%E7%9F%A5%E8%AF%86%E7%BB%84%E7%BB%87%E3%80%81%E7%9F%A5%E8%AF%86%E5%9B%BE%E8%B0%B1%E5%92%8C%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0%E7%AD%89%E7%90%86%E5%BF%B5%E7%9A%84%E5%85%A8%E6%96%B0%E8%80%81%E6%8A%A5%E7%BA%B8%E9%98%85%E8%AF%BB%E4%BD%93%E9%AA%8C%E3%80%82)[fjlib.net](https://www.fjlib.net/zt/mtxh/stxk/llysj/fjstsgxk201403/201804/t20180414_358197.htm#:~:text=%E8%8B%B1%E5%9B%BD%E8%91%97%E5%90%8D%E7%9A%84%E6%83%85%E6%8A%A5%E5%AD%A6%E5%AE%B6%E5%B8%83%E9%B2%81%E5%85%8B%E6%96%AF%E6%9B%BE%E6%8C%87%E5%87%BA%EF%BC%9A%E6%83%85%E6%8A%A5%E5%AD%A6%E7%9A%84%E7%9C%9F%E6%AD%A3%E4%BB%BB%E5%8A%A1%E5%BA%94%E8%AF%A5%E6%98%AF%E7%BB%84%E7%BB%87%E3%80%81%E5%8A%A0%E5%B7%A5%E5%92%8C%E6%95%B4%E7%90%86%E4%BA%BA%E7%B1%BB%E5%AE%A2%E8%A7%82%E7%9F%A5%E8%AF%86%EF%BC%8C%E5%B0%86%E7%9F%A5%E8%AF%86%E7%BB%93%E6%9E%84%E7%BB%98%E5%88%B6%E6%88%90%E4%BB%A5%E5%90%84%E4%B8%AA%E7%9F%A5%E8%AF%86%E5%8D%95%E5%85%83%E6%A6%82%E5%BF%B5%E4%B8%BA%E8%8A%82%E7%82%B9%E7%9A%84%E5%AD%A6%E7%A7%91%E2%80%9C%E7%9F%A5%E8%AF%86%E5%9C%B0%E5%9B%BE%E2%80%9D%EF%BC%8C%E9%80%9A%E8%BF%87%E5%AF%B9%E6%96%87%E7%8C%AE%E4%B8%AD%E7%9A%84%20%E9%80%BB%E8%BE%91%E5%86%85%E5%AE%B9%E8%BF%9B%E8%A1%8C%E5%88%86%E6%9E%90%EF%BC%8C%E6%89%BE%E5%88%B0%E4%BA%BA%E4%BB%AC%E6%80%9D%E6%83%B3%E7%9A%84%E7%9B%B8%E4%BA%92%E5%BD%B1%E5%93%8D%E5%8F%8A%E8%81%94%E7%B3%BB%E7%9A%84%E7%BB%93%E5%90%88%E7%82%B9%EF%BC%8C%E4%BB%8E%E8%80%8C%E4%B8%BA%E7%94%A8%E6%88%B7%E6%8F%90%E4%BE%9B%E7%9F%A5%E8%AF%86%E4%B9%8B%E9%97%B4%E5%85%B3%E7%B3%BB%E7%9A%84%E4%B8%80%E7%A7%8D%E7%9F%A5%E8%AF%86%E7%BB%84%E7%BB%87%E7%9A%84%E7%90%86%E6%83%B3%E7%8A%B6%E6%80%81%E3%80%82)。北京大学图书馆的信息分析师通过CiteSpace绘制了“国际LAM（图书馆-档案-博物馆融合研究）领域作者共被引知识图谱”供决策参考[opaclib.lut.edu.cn](https://opaclib.lut.edu.cn/opac/book/de9ead4dc59d51259106b144167cb3dd" \l ":~:text=graph%20and%20application%20cases%20)。将来用户自己也可调用AI生成类似可视化，比如一位研究生上传自己收集的100篇文献，AI即可输出该领域的研究热点图谱和演进路径。
* *文化资源可视化展示☆* – 上海图书馆在智慧图书馆建设中尝试数字人文项目，如“红色文化资源知识图谱”[slcp.zzuli.edu.cn](https://slcp.zzuli.edu.cn/_t2316/2024/0124/c21538a305868/page.htm#:~:text=%E6%88%91%E6%A0%A1%E7%94%B5%E5%AD%90%E4%BF%A1%E6%81%AF%E5%B7%A5%E7%A8%8B%E5%AD%A6%E9%99%A2%E8%92%8B%E9%80%A2%E6%98%A5%E8%80%81%E5%B8%88%E3%80%8A%E5%A4%A7%E5%AD%A6%E7%89%A9%E7%90%86C%E3%80%8B%E5%B7%B2%E6%88%90%E5%8A%9F%E6%9E%84%E5%BB%BA%E7%9F%A5%E8%AF%86%E5%9B%BE%E8%B0%B1%EF%BC%8C%E5%B9%B6%E5%8F%96%E5%BE%97%E8%89%AF%E5%A5%BD%E7%9A%84%E6%95%88%E6%9E%9C%E3%80%82%E4%B8%8B%E9%9D%A2%E6%98%AF%E8%AF%A5%E8%AF%BE%E7%A8%8B%E7%9F%A5%E8%AF%86%E5%9B%BE%E8%B0%B1%E7%9A%84%E6%95%88%E6%9E%9C%E5%B1%95%E7%A4%BA%EF%BC%9A)。研究者以红军长征相关的图像、文献为素材，构建本体模型并用Neo4j图数据库存储，实现了一个长征历史知识图谱[slcp.zzuli.edu.cn](https://slcp.zzuli.edu.cn/_t2316/2024/0124/c21538a305868/page.htm" \l ":~:text=%E6%88%91%E6%A0%A1%E7%94%B5%E5%AD%90%E4%BF%A1%E6%81%AF%E5%B7%A5%E7%A8%8B%E5%AD%A6%E9%99%A2%E8%92%8B%E9%80%A2%E6%98%A5%E8%80%81%E5%B8%88%E3%80%8A%E5%A4%A7%E5%AD%A6%E7%89%A9%E7%90%86C%E3%80%8B%E5%B7%B2%E6%88%90%E5%8A%9F%E6%9E%84%E5%BB%BA%E7%9F%A5%E8%AF%86%E5%9B%BE%E8%B0%B1%EF%BC%8C%E5%B9%B6%E5%8F%96%E5%BE%97%E8%89%AF%E5%A5%BD%E7%9A%84%E6%95%88%E6%9E%9C%E3%80%82%E4%B8%8B%E9%9D%A2%E6%98%AF%E8%AF%A5%E8%AF%BE%E7%A8%8B%E7%9F%A5%E8%AF%86%E5%9B%BE%E8%B0%B1%E7%9A%84%E6%95%88%E6%9E%9C%E5%B1%95%E7%A4%BA%EF%BC%9A)。该图谱用图节点表示事件、人物、地点等，用户可通过可视界面探索长征历史要素间的关联。这与深圳方志知识图谱有异曲同工之妙（5.1节已述）。未来，AI可以帮助自动从历史文献中提取实体，降低构建这类可视化知识库的人力。用户则可以像玩策略游戏一样在触摸屏上浏览历史事件的关系网，或在大屏幕上观看动态演化演示。微软研究院与学界合作的“甲骨文智能解读”项目就是利用计算机视觉把甲骨文文字和释文对应，可视化展示商代占卜活动，实现传统文献现代可视化[microsoft.com](https://www.microsoft.com/en-us/research/articles/bofeng-mo-zhirong-wu/" \l ":~:text=%E4%B8%BA%E2%80%9C%E5%86%B7%E9%97%A8%E7%BB%9D%E5%AD%A6%E2%80%9D%E7%94%B2%E9%AA%A8%E6%96%87%E7%A0%94%E7%A9%B6%E6%8F%92%E4%B8%8A%E7%A7%91%E6%8A%80%E4%B9%8B%E7%BF%BC)。

**技术成熟度与成本：** 知识可视化需要底层数据支持和图形渲染技术。开源工具和商业软件都有很多选择。难点在于**数据预处理**，而这正是AI所长。例如生成科研知识图谱，需要抽取文献关键词、计算共现关系，这些AI能胜任。渲染方面，已有成熟的可视化库(D3.js等)可与AI集成，由AI负责数据、库负责呈现。成本取决于可视化规模和实时性要求。如果只是静态图生成，成本较低；若要实现实时交互，需要考虑服务器支持并发渲染。这方面或可利用云服务。总的来说，目前知识可视化更多是项目制开发，不是常规服务，因此投入需项目预算支持。但一旦开发成功，可作为展览或数字人文教育的长期平台。

**用户体验提升：** 可视化的优势在于**图形胜千言**。复杂的信息通过形象呈现更易被理解和记忆[fjlib.net](https://www.fjlib.net/zt/mtxh/stxk/llysj/fjstsgxk201403/201804/t20180414_358197.htm" \l ":~:text=)[fjlib.net](https://www.fjlib.net/zt/mtxh/stxk/llysj/fjstsgxk201403/201804/t20180414_358197.htm#:~:text=%E5%B9%B4%E5%9C%A8%E3%80%8A%E7%A7%91%E5%AD%A6%E3%80%8B%E6%9D%82%E5%BF%97%E4%B8%8A%E5%8F%91%E8%A1%A8%E7%9A%84%E4%B8%80%E7%AF%87%E8%AE%BA%E6%96%87%E4%B8%AD%EF%BC%8C%E5%B0%B1%E8%AE%A4%E4%B8%BA%E4%BC%A0%E7%BB%9F%E7%9A%84%E5%9F%BA%E4%BA%8E%E5%AD%A6%E7%A7%91%E7%9A%84%E6%A3%80%E7%B4%A2%E5%92%8C%E6%91%98%E8%A6%81%E6%9C%8D%E5%8A%A1%E9%9A%BE%E4%BB%A5%E6%BB%A1%E8%B6%B3%E7%A0%94%E7%A9%B6%E7%9A%84%E9%9C%80%E6%B1%82%E3%80%82)。用户尤其是学生群体，对互动图形通常兴趣更高，比阅读长篇文字更有吸引力。以知识图谱来说，可视化能展示学科发展全貌，让用户在几秒内抓住该领域核心人物和热点[fjlib.net](https://www.fjlib.net/zt/mtxh/stxk/llysj/fjstsgxk201403/201804/t20180414_358197.htm" \l ":~:text=%E5%B9%B4%E5%9C%A8%E3%80%8A%E7%A7%91%E5%AD%A6%E3%80%8B%E6%9D%82%E5%BF%97%E4%B8%8A%E5%8F%91%E8%A1%A8%E7%9A%84%E4%B8%80%E7%AF%87%E8%AE%BA%E6%96%87%E4%B8%AD%EF%BC%8C%E5%B0%B1%E8%AE%A4%E4%B8%BA%E4%BC%A0%E7%BB%9F%E7%9A%84%E5%9F%BA%E4%BA%8E%E5%AD%A6%E7%A7%91%E7%9A%84%E6%A3%80%E7%B4%A2%E5%92%8C%E6%91%98%E8%A6%81%E6%9C%8D%E5%8A%A1%E9%9A%BE%E4%BB%A5%E6%BB%A1%E8%B6%B3%E7%A0%94%E7%A9%B6%E7%9A%84%E9%9C%80%E6%B1%82%E3%80%82)。读者也能自己发现原先未注意的关联（比如看到两位作者有共被引关系，从而认识到他们研究的关联性）。这是一种**探索式学习**，主动性强，效果往往比被动阅读更好。而在文化宣传上，可视化展示能将静态文献变为动态体验，增强公众参与度。例如观众通过触屏浏览红色文化知识图谱，比单纯看图文展板更具沉浸感和互动性。可以说，可视化服务让图书馆资源**“看得见、摸得着”**，降低了知识门槛，扩大了影响力。不过也要考虑不要用力过猛：并非所有信息都适合可视化，如果强行可视化可能弄巧成拙。因此智能体应辅助馆员选择合适内容场景应用可视化。只要选题对路，可视化几乎总能带来耳目一新的用户体验，体现图书馆的技术创新形象。

本节围绕图书馆六大知识服务场景，探讨了智能体融合应用的具体方式和案例。可以看到，在**知识图谱、语义检索、智能问答、个性化推荐、知识推送、可视化服务**各方面，国内外均已有探索尝试，并初步取得了积极成效：知识图谱为用户提供了结构化、可视的知识发现工具[library.ucr.edu](https://library.ucr.edu/news/2022/04/29/ucr-library-implements-yewno-discover-ai-powered-research" \l ":~:text=How%20does%20Yewno%20Discover%20work%3F,the%20user%20in%20their%20research)[whly.gd.gov.cn](https://whly.gd.gov.cn/service_newggwhzx/content/post_4061139.html#:~:text=%E7%AB%99%E5%9B%BE%E4%B9%A6%E8%B5%84%E6%BA%90%2C%E8%BF%98%E5%8F%AF%E8%8E%B7%E5%8F%96%E5%85%AC%E5%85%B1%E5%9B%BE%E4%B9%A6%E9%A6%86%E8%97%8F%E4%B9%A6%E6%8C%87%E5%BC%95%2C%E5%AE%9E%E7%8E%B0%E4%BA%86%E5%85%AC%E7%A7%81%E8%97%8F%E4%B9%A6%E7%9A%84%E8%9E%8D%E5%90%88%E5%88%A9%E7%94%A8%2C%E6%8F%90%E9%AB%98%E4%BA%86%E5%9B%BE%E4%B9%A6%E8%8E%B7%E5%8F%96%E6%95%88%E7%8E%87%2C%E6%BB%A1%E8%B6%B3%E4%BA%86%E5%B8%82%E6%B0%91%E9%98%85%E8%AF%BB%E9%9C%80%E6%B1%82%E3%80%82)；语义检索让读者能用自然语言查询并得到精准结果[calsp.cn](https://www.calsp.cn/2024/08/21/bulletin-202408-05/" \l ":~:text=RAG%E6%8A%80%E6%9C%AF%E6%98%AF%E4%B8%80%E7%A7%8D%E7%BB%93%E5%90%88%E4%BF%A1%E6%81%AF%E6%A3%80%E7%B4%A2%E5%92%8C%E6%96%87%E6%9C%AC%E7%94%9F%E6%88%90%E7%9A%84%E6%8A%80%E6%9C%AF%E3%80%82%E5%9C%A8%E5%9B%BE%E4%B9%A6%E9%A6%86%E4%B8%AD%EF%BC%8C%E5%9B%BE%E4%B9%A6%E9%A6%86%E9%A6%86%E8%97%8F%E5%BD%93%E4%B8%AD%E5%AD%98%E5%9C%A8%E5%A4%A7%E9%87%8F%E7%9A%84%E5%AD%A6%E6%9C%AF%E6%96%87%E7%8C%AE%EF%BC%8C%E5%8F%AF%E4%BB%A5%E5%B0%86%E8%BF%99%E4%BA%9B%E5%AD%A6%E6%9C%AF%E6%96%87%E7%8C%AE%E9%80%9A%E8%BF%87%E4%B8%93%E4%B8%9A%E7%9A%84%E5%90%91%E9%87%8F%E5%8C%96%E5%B7%A5%E5%85%B7%E8%BF%9B%E8%A1%8C%E5%88%87%E7%89%87%E5%AD%98%E5%82%A8%E5%9C%A8%E5%90%91%E9%87%8F%E6%95%B0%E6%8D%AE%E5%BA%93%E5%BD%93%E4%B8%AD%EF%BC%8C%20%E5%BD%A2%E6%88%90%E4%B8%80%E4%B8%AA%E6%9C%AC%E5%9C%B0%E7%9F%A5%E8%AF%86%E5%BA%93%EF%BC%8C%E5%86%8D%E5%B0%86%E6%9C%AC%E5%9C%B0%E7%9F%A5%E8%AF%86%E5%BA%93%E4%B8%8E%E5%A4%A7%E8%AF%AD%E8%A8%80%E6%A8%A1%E5%9E%8B%E8%BF%9B%E8%A1%8C%E7%BB%93%E5%90%88%EF%BC%8C%E5%BD%93%E7%94%A8%E6%88%B7%E6%8F%90%E5%87%BA%E9%97%AE%E9%A2%98%E6%97%B6%EF%BC%8C%E7%B3%BB%E7%BB%9F%E8%83%BD%E5%A4%9F%E6%A0%B9%E6%8D%AE%E7%94%A8%E6%88%B7%E7%9A%84%E9%97%AE%E9%A2%98%E6%A3%80%E7%B4%A2%E5%9B%BE%E4%B9%A6%E9%A6%86%E7%9A%84%E6%95%B0%E5%AD%97%E8%B5%84%E6%BA%90%EF%BC%8C%E5%B9%B6%E7%94%9F%E6%88%90%E5%87%86%E7%A1%AE%E7%9A%84%E5%9B%9E%E7%AD%94%E3%80%82%E8%BF%99%E5%8F%AF%E4%BB%A5%E5%A4%A7%E5%A4%A7%E5%87%8F%E5%B0%91%E7%94%A8%E6%88%B7%E6%9F%A5%E6%89%BE%E4%BF%A1%20%E6%81%AF%E7%9A%84%E6%97%B6%E9%97%B4%EF%BC%8C%E5%B9%B6%E6%8F%90%E9%AB%98%E5%9B%9E%E7%AD%94%E7%9A%84%E7%9B%B8%E5%85%B3%E6%80%A7%E3%80%82)；智能问答扩大了图书馆咨询服务的时空范围，提高了响应效率[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00" \l ":~:text=%E5%BD%93%E5%89%8D%E8%BF%98%E5%AD%98%E5%9C%A8%E5%93%AA%E4%BA%9B%E9%99%90%E5%88%B6%EF%BC%9F)[calsp.cn](https://www.calsp.cn/2024/08/21/bulletin-202408-05/#:~:text=%E6%B8%85%E5%8D%8E%E5%A4%A7%E5%AD%A6%E5%9B%BE%E4%B9%A6%E9%A6%86%E4%B8%BA%E5%BF%AB%E9%80%9F%E6%8E%A8%E8%BF%9B%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E5%9C%A8%E5%9B%BE%E4%B9%A6%E9%A6%86%E7%9A%84%E8%90%BD%E5%9C%B0%E5%BA%94%E7%94%A8%EF%BC%8C%E4%BA%8E2024%E5%B9%B44%E6%9C%8828%E6%97%A5%E6%AD%A3%E5%BC%8F%E4%B8%8A%E7%BA%BF%E4%BA%86AI%E9%98%85%E8%AF%BB%E5%8A%A9%E6%89%8B%E3%80%81AI%E5%AF%BC%E8%88%AA%E5%8A%A9%E6%89%8B%E5%BA%94%E7%94%A8%E3%80%82%E9%80%9A%E8%BF%87%E6%99%BA%E8%83%BD%E5%88%86%E6%9E%90PDF%E6%96%87%E6%A1%A3%E5%92%8C%E6%95%B0%E6%8D%AE%E5%BA%93%E5%86%85%E5%AE%B9%EF%BC%8C%E6%8F%90%E5%8D%87%E4%BA%86%E5%B8%88%20%E7%94%9F%E7%9A%84%E9%98%85%E8%AF%BB%E5%92%8C%E7%A0%94%E7%A9%B6%E6%95%88%E7%8E%87%E3%80%82%E4%B8%A4%E9%A1%B9%E5%BA%94%E7%94%A8%E5%9D%87%E9%9D%A2%E5%90%91%E6%B8%85%E5%8D%8E%E5%85%A8%E6%A0%A1%E5%B8%88%E7%94%9F%E6%8F%90%E4%BE%9B%E6%9C%8D%E5%8A%A1%EF%BC%8C%E4%B8%8A%E7%BA%BF%E5%90%8E%E6%94%B6%E5%88%B0%E4%BA%86%E8%AF%BB%E8%80%85%E7%9A%84%E8%89%AF%E5%A5%BD%E5%8F%8D%E9%A6%88%E3%80%82)；个性化推荐和知识推送体现了以读者为中心的主动服务，使阅读推广更加精准温馨[whly.gd.gov.cn](https://whly.gd.gov.cn/gkmlpt/content/4/4404/post_4404325.html" \l ":~:text=%E4%BD%9B%E5%B1%B1%E2%80%9C%E6%98%93%E6%9C%AC%E4%B9%A6%E2%80%9D%E9%A1%B9%E7%9B%AE%E5%85%A5%E9%80%89%E6%96%87%E5%8C%96%E5%92%8C%E6%97%85%E6%B8%B8%E9%83%A8%E6%99%BA%E6%85%A7%E5%9B%BE%E4%B9%A6%E9%A6%86%E5%88%9B%E6%96%B0%E5%BA%94%E7%94%A8%E4%BC%98%E7%A7%80%E6%A1%88%E4%BE%8B%20%E6%A0%B9%E6%8D%AE%E7%94%A8%E6%88%B7%E5%85%B4%E8%B6%A3%E7%88%B1%E5%A5%BD%E7%AD%89%E4%BF%A1%E6%81%AF%EF%BC%8C%E6%8A%BD%E8%B1%A1%E5%87%BA%E8%AF%BB%E8%80%85%E6%A8%A1%E5%9E%8B%EF%BC%8C%E5%B0%86%E9%A2%84%E8%AE%BE%E7%9A%84%E7%94%A8%E6%88%B7%E4%B8%AA%E6%80%A7%E5%8C%96%E6%A0%87%E7%AD%BE%E5%92%8C%E5%9B%BE%E4%B9%A6%E6%A0%87%E7%AD%BE%E5%85%B3%E8%81%94%EF%BC%8C%E8%87%AA%E5%8A%A8%E6%A0%B9%E6%8D%AE%E9%A2%84%E8%AE%BE%E7%9B%B8%E5%85%B3%E5%BA%A6%E6%9D%83%E9%87%8D%E5%80%BC%E5%90%91%E7%94%A8%E6%88%B7%E6%8E%A8%E8%8D%90%E5%9B%BE%E4%B9%A6%EF%BC%8C%E5%AE%9E%E7%8E%B0%E7%B2%BE%E5%87%86%E5%8C%96%E3%80%81%E4%B8%AA%E6%80%A7%E5%8C%96%E6%8E%A8%E8%8D%90%E3%80%82)[nju.edu.cn](https://www.nju.edu.cn/info/1056/340981.htm#:~:text=%E9%AB%98%E6%A0%A1%E5%9B%BE%E4%B9%A6%E9%A6%86%E5%A6%82%E4%BD%95%E5%BC%80%E5%B1%95%E9%98%85%E8%AF%BB%E6%8E%A8%E5%B9%BF%EF%BC%9F%E6%B1%9F%E8%8B%8F%E8%BF%99%E6%89%B9%E7%B2%BE%E5%93%81%E6%A1%88%E4%BE%8B%E7%BB%99%E4%BD%A0%E7%AD%94%E6%A1%88%20)；可视化服务则通过直观呈现和交互增强了用户对知识的理解和兴趣[fjlib.net](https://www.fjlib.net/zt/mtxh/stxk/llysj/fjstsgxk201403/201804/t20180414_358197.htm" \l ":~:text=%E5%B9%B4%E5%9C%A8%E3%80%8A%E7%A7%91%E5%AD%A6%E3%80%8B%E6%9D%82%E5%BF%97%E4%B8%8A%E5%8F%91%E8%A1%A8%E7%9A%84%E4%B8%80%E7%AF%87%E8%AE%BA%E6%96%87%E4%B8%AD%EF%BC%8C%E5%B0%B1%E8%AE%A4%E4%B8%BA%E4%BC%A0%E7%BB%9F%E7%9A%84%E5%9F%BA%E4%BA%8E%E5%AD%A6%E7%A7%91%E7%9A%84%E6%A3%80%E7%B4%A2%E5%92%8C%E6%91%98%E8%A6%81%E6%9C%8D%E5%8A%A1%E9%9A%BE%E4%BB%A5%E6%BB%A1%E8%B6%B3%E7%A0%94%E7%A9%B6%E7%9A%84%E9%9C%80%E6%B1%82%E3%80%82)。这些案例表明，智能体技术与图书馆服务具有很高的契合度，能够不同程度地**提升用户体验**、**提高服务效率**和**挖掘资源价值**。

当然，实际应用中也遇到了一些问题需要重视。例如智能问答的准确性和可信度需要通过引用来源和人工监管来保障[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00" \l ":~:text=%E9%82%B1%E4%BE%83%E5%90%9B%E8%AF%B4%EF%BC%8C%E2%80%9C%E4%B8%BE%E4%B8%AA%E4%BE%8B%E5%AD%90%EF%BC%8C%E7%BC%96%E7%A0%81%E6%99%BA%E8%83%BD%E4%BD%93%E5%8F%AF%E4%BB%A5%E7%94%9F%E6%88%90%E4%BB%A3%E7%A0%81%EF%BC%8C%E4%BD%86%E6%9C%89%E6%97%B6%E4%BC%9A%E5%87%BA%E9%94%99%EF%BC%8C%E8%80%8C%E4%B8%94%E5%AE%83%E4%B9%9F%E4%B8%8D%E7%9F%A5%E9%81%93%E5%A6%82%E4%BD%95%E6%B5%8B%E8%AF%95%E6%AD%A3%E5%9C%A8%E5%88%9B%E5%BB%BA%E7%9A%84%E4%BB%A3%E7%A0%81%E3%80%82%E5%9B%A0%E6%AD%A4%EF%BC%8C%E4%BA%BA%E7%B1%BB%E4%BB%8D%E7%84%B6%E9%9C%80%E8%A6%81%E5%8F%82%E4%B8%8E%E5%88%B0%E8%BF%99%E4%B8%80%E8%BF%87%E7%A8%8B%E4%B8%AD%E3%80%82%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E7%B3%BB%E7%BB%9F%E4%BB%8D%E7%84%B6%E6%97%A0%E6%B3%95%E5%AE%8C%E5%85%A8%E5%AE%9E%20%E7%8E%B0%E6%8E%A8%E7%90%86%EF%BC%8C%E8%BF%99%E6%98%AF%E5%9C%A8%E5%A4%8D%E6%9D%82%E4%B8%94%E6%A8%A1%E7%B3%8A%E7%9A%84%E4%BA%BA%E7%B1%BB%E4%B8%96%E7%95%8C%E4%B8%AD%E8%BF%90%E8%A1%8C%E7%9A%84%E5%85%B3%E9%94%AE%E4%B8%80%E6%AD%A5%E3%80%82%E2%80%9D)；个性化推荐与知识推送必须平衡隐私和干扰，取得用户信任和授权；知识图谱与可视化构建仍需投入相当人力策划关键内容；大模型调用的成本对一些馆是负担，需寻求更经济方案等。但总体而言，这些挑战是可以通过技术迭代和管理措施逐步解决的。许多服务在初期也采取了**人机结合**模式：让AI承担繁重的处理工作，同时保留人工审核与高端咨询，以确保服务质量平稳过渡。例如清华等馆上线AI助手时都强调其答案仅供参考，鼓励读者结合馆员指导使用[calsp.cn](https://www.calsp.cn/2024/08/21/bulletin-202408-05/" \l ":~:text=RAG%E6%8A%80%E6%9C%AF%E6%98%AF%E4%B8%80%E7%A7%8D%E7%BB%93%E5%90%88%E4%BF%A1%E6%81%AF%E6%A3%80%E7%B4%A2%E5%92%8C%E6%96%87%E6%9C%AC%E7%94%9F%E6%88%90%E7%9A%84%E6%8A%80%E6%9C%AF%E3%80%82%E5%9C%A8%E5%9B%BE%E4%B9%A6%E9%A6%86%E4%B8%AD%EF%BC%8C%E5%9B%BE%E4%B9%A6%E9%A6%86%E9%A6%86%E8%97%8F%E5%BD%93%E4%B8%AD%E5%AD%98%E5%9C%A8%E5%A4%A7%E9%87%8F%E7%9A%84%E5%AD%A6%E6%9C%AF%E6%96%87%E7%8C%AE%EF%BC%8C%E5%8F%AF%E4%BB%A5%E5%B0%86%E8%BF%99%E4%BA%9B%E5%AD%A6%E6%9C%AF%E6%96%87%E7%8C%AE%E9%80%9A%E8%BF%87%E4%B8%93%E4%B8%9A%E7%9A%84%E5%90%91%E9%87%8F%E5%8C%96%E5%B7%A5%E5%85%B7%E8%BF%9B%E8%A1%8C%E5%88%87%E7%89%87%E5%AD%98%E5%82%A8%E5%9C%A8%E5%90%91%E9%87%8F%E6%95%B0%E6%8D%AE%E5%BA%93%E5%BD%93%E4%B8%AD%EF%BC%8C%20%E5%BD%A2%E6%88%90%E4%B8%80%E4%B8%AA%E6%9C%AC%E5%9C%B0%E7%9F%A5%E8%AF%86%E5%BA%93%EF%BC%8C%E5%86%8D%E5%B0%86%E6%9C%AC%E5%9C%B0%E7%9F%A5%E8%AF%86%E5%BA%93%E4%B8%8E%E5%A4%A7%E8%AF%AD%E8%A8%80%E6%A8%A1%E5%9E%8B%E8%BF%9B%E8%A1%8C%E7%BB%93%E5%90%88%EF%BC%8C%E5%BD%93%E7%94%A8%E6%88%B7%E6%8F%90%E5%87%BA%E9%97%AE%E9%A2%98%E6%97%B6%EF%BC%8C%E7%B3%BB%E7%BB%9F%E8%83%BD%E5%A4%9F%E6%A0%B9%E6%8D%AE%E7%94%A8%E6%88%B7%E7%9A%84%E9%97%AE%E9%A2%98%E6%A3%80%E7%B4%A2%E5%9B%BE%E4%B9%A6%E9%A6%86%E7%9A%84%E6%95%B0%E5%AD%97%E8%B5%84%E6%BA%90%EF%BC%8C%E5%B9%B6%E7%94%9F%E6%88%90%E5%87%86%E7%A1%AE%E7%9A%84%E5%9B%9E%E7%AD%94%E3%80%82%E8%BF%99%E5%8F%AF%E4%BB%A5%E5%A4%A7%E5%A4%A7%E5%87%8F%E5%B0%91%E7%94%A8%E6%88%B7%E6%9F%A5%E6%89%BE%E4%BF%A1%20%E6%81%AF%E7%9A%84%E6%97%B6%E9%97%B4%EF%BC%8C%E5%B9%B6%E6%8F%90%E9%AB%98%E5%9B%9E%E7%AD%94%E7%9A%84%E7%9B%B8%E5%85%B3%E6%80%A7%E3%80%82)。

从技术成熟度看，语义检索、问答、推荐这些用例已经相对成熟，可在更多馆推广；知识图谱、可视化等属于锦上添花型应用，可根据需要重点打造特色项目；知识推送等还需探索运营模式。部署成本上，大模型的API费用可能随技术进步而下降，同时开源模型性能提升也会降低门槛，图书馆联盟共建共享也是可行途径。因此，智能体赋能图书馆知识服务的**可持续性前景良好**。真正关键的是馆员观念和能力的提升：馆员需拥抱AI工具，学会训练和调校智能体，更要发挥创造性将AI技术与服务需求对接，设计出读者喜闻乐见的新服务形式。

总之，通过本节的案例分析，可以确信智能体在图书馆知识服务融合方面**大有可为**。它不仅可以提高效率，更重要的是有望革新服务模式，使图书馆从被动文献提供者转变为主动智慧信息助手，为用户创造更大价值。下一章我们将在此基础上对全文进行总结，提出对未来的展望和建议。

## 3.5 结论

本章围绕智能体（Agent/代理型AI）技术的发展演进及其与图书馆知识服务的融合展开了系统探讨。通过梳理文献和案例，我们获得以下主要结论和认识：

**第一，智能体概念在学术界和业界存在融合但侧重不同的定义。** 学术上，智能体被视为感知-决策-行动的自主智能实体，强调理性行为和自主性[m.thepaper.cn](https://m.thepaper.cn/kuaibao_detail.jsp?contid=30475035&from=kuaibao" \l ":~:text=%E6%B5%8B%E8%83%BD%E5%8A%9B%EF%BC%8C%20)[blog.csdn.net](https://blog.csdn.net/Code1994/article/details/142780046#:~:text=)；业界则将其理解为能代表用户自主执行任务的AI系统，突出其实用工具化特征[ibm.com](https://www.ibm.com/cn-zh/think/topics/ai-agents" \l ":~:text=%E4%BB%80%E4%B9%88%E6%98%AF%20AI%20%E6%99%BA%E8%83%BD%E4%BD%93%EF%BC%9F)[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00#:~:text=%E5%A6%82%E4%BD%95%E5%AE%9A%E4%B9%89%E2%80%9C%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%93%E2%80%9D%EF%BC%9F)。二者并不矛盾，而是从理论本质和应用功能两个层面刻画了智能体。我们总结了智能体的多种分类，包括行为决策模式（从简单反射到学习型）[blog.csdn.net](https://blog.csdn.net/Code1994/article/details/142780046" \l ":~:text=%23%203.%E5%9F%BA%E4%BA%8E%E7%9B%AE%E6%A0%87%E7%9A%84%E4%BB%A3%E7%90%86%EF%BC%88Goal)[blog.csdn.net](https://blog.csdn.net/Code1994/article/details/142780046#:~:text=)、自主程度（自主/半自主）[blog.csdn.net](https://blog.csdn.net/Code1994/article/details/142780046" \l ":~:text=%E7%89%B9%E7%82%B9)[blog.csdn.net](https://blog.csdn.net/Code1994/article/details/142780046#:~:text=)、存在形态（软件代理 vs. 具身代理）[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00" \l ":~:text=%E2%80%9C%E5%B9%BF%E4%B9%89%E4%B8%8A%E8%AE%B2%EF%BC%8C%E6%9C%89%E4%B8%A4%E7%B1%BB%E4%B8%8D%E5%90%8C%E7%9A%84%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%93%EF%BC%9A%E8%BD%AF%E4%BB%B6%E6%99%BA%E8%83%BD%E4%BD%93%EF%BC%88Software%20Agents%EF%BC%89%E5%92%8C%E5%85%B7%E8%BA%AB%E6%99%BA%E8%83%BD%E4%BD%93%EF%BC%88Embodied%20Agents%EF%BC%89%E3%80%82%E2%80%9DJim%20Fan%20%E8%AF%B4%E9%81%93%EF%BC%8C%E2%80%9C%E8%BD%AF%E4%BB%B6%E6%99%BA%E8%83%BD%E4%BD%93%E5%9C%A8%E7%94%B5%E8%84%91%E6%88%96%E6%89%8B%E6%9C%BA%E4%B8%8A%E8%BF%90%E8%A1%8C%E5%B9%B6%E4%BD%BF%E7%94%A8%E5%BA%94%E7%94%A8%E7%A8%8B%E5%BA%8F%EF%BC%8C%E8%BF%99%E4%BA%9B%E6%99%BA%E8%83%BD%E4%BD%93%E5%AF%B9%E5%8A%9E%E5%85%AC%E5%AE%A4%E5%B7%A5%E4%BD%9C%E3%80%81%E5%8F%91%E9%80%81%E7%94%B5%E5%AD%90%E9%82%AE%E4%BB%B6%E6%88%96%E5%AE%8C%E6%88%90%E8%BF%99%E4%B8%80%E7%B3%BB%E5%88%97%E6%B4%BB%E5%8A%A8%E9%9D%9E%E5%B8%B8%E6%9C%89%E7%94%A8%E3%80%82%E2%80%9D)等，这些分类有助于理解不同类型智能体的特征。随着大模型时代的到来，传统分类出现交融，一个智能体往往兼具多方面特性，未来或出现更统一的分类框架。

**第二，智能体技术经历了重要的发展阶段，并正迈入通用代理的新纪元。** 我们回顾了智能体起源于多智能体系统等早期探索，2010年代中期因强化学习突破而迎来第一波专用智能体高潮（如AlphaGo）[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00#:~:text=%E5%8D%8E%E7%9B%9B%E9%A1%BF%E5%A4%A7%E5%AD%A6%E8%AE%A1%E7%AE%97%E6%9C%BA%E7%A7%91%E5%AD%A6%E6%95%99%E6%8E%88%20Chirag%20Shah%20%E8%A1%A8%E7%A4%BA%EF%BC%8C%E2%80%9C%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%93%E2%80%9D%E4%B8%80%E8%AF%8D%E5%85%B6%E5%AE%9E%E5%B7%B2%E7%BB%8F%E5%AD%98%E5%9C%A8%E4%BA%86%E5%BE%88%E5%A4%9A%E5%B9%B4%EF%BC%8C%E4%BD%86%E6%98%AF%E5%85%B6%E5%9C%A8%E4%B8%8D%E5%90%8C%E7%9A%84%E6%97%B6%E9%97%B4%E6%84%8F%E5%91%B3%E7%9D%80%E4%B8%8D%E5%90%8C%E7%9A%84%E4%B8%9C%E8%A5%BF%E3%80%82)，进入2020年代又因大模型驱动出现通用对话代理的新浪潮（如ChatGPT）[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00#:~:text=%E5%8D%8E%E7%9B%9B%E9%A1%BF%E5%A4%A7%E5%AD%A6%E8%AE%A1%E7%AE%97%E6%9C%BA%E7%A7%91%E5%AD%A6%E6%95%99%E6%8E%88%20Chirag%20Shah%20%E8%A1%A8%E7%A4%BA%EF%BC%8C%E2%80%9C%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%93%E2%80%9D%E4%B8%80%E8%AF%8D%E5%85%B6%E5%AE%9E%E5%B7%B2%E7%BB%8F%E5%AD%98%E5%9C%A8%E4%BA%86%E5%BE%88%E5%A4%9A%E5%B9%B4%EF%BC%8C%E4%BD%86%E6%98%AF%E5%85%B6%E5%9C%A8%E4%B8%8D%E5%90%8C%E7%9A%84%E6%97%B6%E9%97%B4%E6%84%8F%E5%91%B3%E7%9D%80%E4%B8%8D%E5%90%8C%E7%9A%84%E4%B8%9C%E8%A5%BF%E3%80%82)。当前，智能体正从能够**内容创作**(生成式模型)到善于**对话交互**(ChatGPT)、进一步扩展**多模态**感知与输出(GPT-4、Gemini)[blog.google](https://blog.google/technology/ai/google-gemini-ai/#:~:text=Introducing%20Gemini%3A%20our%20largest%20and,sizes%3A%20Ultra%2C%20Pro%20and%20Nano)[techpoint.africa](https://techpoint.africa/guide/meta-ai-vs-chatgpt-vs-google-gemini/#:~:text=Launched%20in%202023%2C%20Gemini%20is,major%20advantage%20over%20its%20competitors)，并朝着**增强推理和自主执行**方向演进(AutoGPT等自主代理)[yicai.com](https://www.yicai.com/news/102678941.html#:~:text=%E6%9C%AC%E5%91%A8%E5%8F%88%E6%98%AFAgent%EF%BC%88%E6%99%BA%E8%83%BD%E4%BD%93%EF%BC%89%E7%83%AD%E5%BA%A6%E6%8C%81%E7%BB%AD%E7%9A%84%E4%B8%80%E5%91%A8%E3%80%82%E5%8C%97%E4%BA%AC%E6%97%B6%E9%97%B46%E6%9C%8819%E6%97%A5%EF%BC%8C%E5%9B%BD%E5%86%85%E5%A4%A7%E6%A8%A1%E5%9E%8B%E5%88%9D%E5%88%9B%E4%BC%81%E4%B8%9AMinimax%E5%8F%91%E5%B8%83%E8%83%BD%E5%A4%9F%E5%AE%8C%E6%88%90%E9%95%BF%E7%A8%8B%EF%BC%88Long%20Horizon%EF%BC%89%E5%A4%8D%E6%9D%82%E4%BB%BB%E5%8A%A1%E7%9A%84%E9%80%9A%E7%94%A8%E6%99%BA%E8%83%BD%E4%BD%93MiniMax%20Agent%EF%BC%8C%E5%A4%9A%E6%AD%A5%E8%A7%84%E5%88%92%E4%B8%93%E5%AE%B6%E7%BA%A7%E8%A7%A3%E5%86%B3%E6%96%B9%E6%A1%88%E3%80%81%E6%8B%86%E8%A7%A3%E4%BB%BB%E5%8A%A1%E9%9C%80%E6%B1%82%E3%80%81%E6%89%A7%E8%A1%8C%E5%A4%9A%E4%B8%AA%E5%AD%90%E4%BB%BB%E5%8A%A1%E4%BB%8E%E8%80%8C%E4%BA%A4%E4%BB%98%E6%9C%80%E7%BB%88%E7%BB%93%E6%9E%9C%E3%80%82)。这一演进体现了AI系统从单点智能向连续自主智能的升级。典型代表模型如ChatGPT、Claude开启了通用代理雏形，GPT-4、Gemini展示了多模态理解力，AutoGPT、AgentGPT探索了自主规划能力。可以预见，未来几年将涌现更强大的智能体，将目前割裂的能力融合成更统一的**通用智能代理**。

**第三，当前智能体发展既充满机遇又面临显著瓶颈，需要理性审视。** 趋势上，权威机构预测Agentic AI将迅速普及于软件和工作流程，到2028年可自动做出15%的日常决策[reuters.com](https://www.reuters.com/business/over-40-agentic-ai-projects-will-be-scrapped-by-2027-gartner-says-2025-06-25/" \l ":~:text=Gartner%20predicts%20at%20least%2015,in%202024)。企业也认识到必须将AI从辅助工具深化为流程中枢，这将带来效率和创新的大幅提升[mckinsey.com](https://www.mckinsey.com/capabilities/quantumblack/our-insights/seizing-the-agentic-ai-advantage" \l ":~:text=,functional)[mckinsey.com](https://www.mckinsey.com/capabilities/quantumblack/our-insights/seizing-the-agentic-ai-advantage#:~:text=earning%20trust%2C%20driving%20adoption%2C%20and,upskill%20the%20workforce%2C%20adapt%20the)。然而挑战同样突出：许多初期项目由于缺乏明确价值和驾驭经验而失败，泡沫炒作需警惕[reuters.com](https://www.reuters.com/business/over-40-agentic-ai-projects-will-be-scrapped-by-2027-gartner-says-2025-06-25/" \l ":~:text=,Senior%20Director%20Analyst%20at%20Gartner)；算力和成本压力巨大，当前模型自主执行仍受限于上下文和性能瓶颈[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00" \l ":~:text=Jim%20Fan%20%E5%88%99%E8%A1%A8%E7%A4%BA%EF%BC%9A%E2%80%9C%E6%88%91%E4%BB%AC%E8%BF%98%E8%BF%9C%E8%BF%9C%E6%B2%A1%E6%9C%89%E4%B8%80%E4%B8%AA%E5%8F%AF%E4%BB%A5%E4%B8%BA%E6%88%91%E4%BB%AC%E8%87%AA%E5%8A%A8%E5%8C%96%E6%89%80%E6%9C%89%E8%BF%99%E4%BA%9B%E5%AE%B6%E5%8A%A1%E7%9A%84%E6%99%BA%E8%83%BD%E4%BD%93%E3%80%82%E7%9B%AE%E5%89%8D%E7%9A%84%E7%B3%BB%E7%BB%9F%E5%8F%AF%E8%83%BD%E4%BC%9A%E4%BA%A7%E7%94%9F%E5%B9%BB%E8%A7%89%EF%BC%8C%E8%80%8C%E4%B8%94%E5%AE%83%E4%BB%AC%E4%B9%9F%E4%B8%8D%E6%80%BB%E6%98%AF%E4%B8%A5%E6%A0%BC%E9%81%B5%E5%BE%AA%E6%8C%87%E4%BB%A4%EF%BC%8C%E6%98%BE%E7%84%B6%E8%BF%99%E5%B0%B1%E5%BE%88%E9%BA%BB%E7%83%A6%E4%BA%86%E3%80%82%E2%80%9D)；评估机制不完善，智能体效果和风险难以量化，需要建立持续评估和治理框架[cdotrends.com](https://www.cdotrends.com/story/4601/true-agentic-ai-needs-you-break-your-architecture" \l ":~:text=The%20most%20mature%20organizations%20are,of%20corporate%20values%20and%20ethics)；安全控制是重中之重，智能体自主性越强越需严防不当行为，要求在技术和伦理上同步加强**可控性**[view.inews.qq.com](https://view.inews.qq.com/a/20240710A02T1K00" \l ":~:text=%E7%8E%B0%E6%8E%A8%E7%90%86%EF%BC%8C%E8%BF%99%E6%98%AF%E5%9C%A8%E5%A4%8D%E6%9D%82%E4%B8%94%E6%A8%A1%E7%B3%8A%E7%9A%84%E4%BA%BA%E7%B1%BB%E4%B8%96%E7%95%8C%E4%B8%AD%E8%BF%90%E8%A1%8C%E7%9A%84%E5%85%B3%E9%94%AE%E4%B8%80%E6%AD%A5%E3%80%82%E2%80%9D)[cdotrends.com](https://www.cdotrends.com/story/4601/true-agentic-ai-needs-you-break-your-architecture#:~:text=The%20most%20mature%20organizations%20are,of%20corporate%20values%20and%20ethics)。整体来看，我们正处于智能体技术的“**理性回调期**”：不再无限夸大其万能，但也坚定其长远价值。成功之道在于**小步试错、场景聚焦、加强治理**，逐步积累成熟度，而非一蹴而就。

**第四，智能体与图书馆知识服务的融合拥有广阔前景，已经在多个场景展现出赋能效果。** 我们分析的六大场景案例表明：

* 在**知识组织**方面，智能体协助构建知识图谱和可视化网络，使文献资源以直观知识脉络呈现[library.ucr.edu](https://library.ucr.edu/news/2022/04/29/ucr-library-implements-yewno-discover-ai-powered-research" \l ":~:text=How%20does%20Yewno%20Discover%20work%3F,the%20user%20in%20their%20research)[whly.gd.gov.cn](https://whly.gd.gov.cn/service_newggwhzx/content/post_4061139.html#:~:text=%E7%AB%99%E5%9B%BE%E4%B9%A6%E8%B5%84%E6%BA%90%2C%E8%BF%98%E5%8F%AF%E8%8E%B7%E5%8F%96%E5%85%AC%E5%85%B1%E5%9B%BE%E4%B9%A6%E9%A6%86%E8%97%8F%E4%B9%A6%E6%8C%87%E5%BC%95%2C%E5%AE%9E%E7%8E%B0%E4%BA%86%E5%85%AC%E7%A7%81%E8%97%8F%E4%B9%A6%E7%9A%84%E8%9E%8D%E5%90%88%E5%88%A9%E7%94%A8%2C%E6%8F%90%E9%AB%98%E4%BA%86%E5%9B%BE%E4%B9%A6%E8%8E%B7%E5%8F%96%E6%95%88%E7%8E%87%2C%E6%BB%A1%E8%B6%B3%E4%BA%86%E5%B8%82%E6%B0%91%E9%98%85%E8%AF%BB%E9%9C%80%E6%B1%82%E3%80%82)，增强了用户探索和认知体验。
* 在**信息获取**方面，智能体驱动的语义检索和智能问答让读者能以自然语言提问并获得精准答案[calsp.cn](https://www.calsp.cn/2024/08/21/bulletin-202408-05/" \l ":~:text=RAG%E6%8A%80%E6%9C%AF%E6%98%AF%E4%B8%80%E7%A7%8D%E7%BB%93%E5%90%88%E4%BF%A1%E6%81%AF%E6%A3%80%E7%B4%A2%E5%92%8C%E6%96%87%E6%9C%AC%E7%94%9F%E6%88%90%E7%9A%84%E6%8A%80%E6%9C%AF%E3%80%82%E5%9C%A8%E5%9B%BE%E4%B9%A6%E9%A6%86%E4%B8%AD%EF%BC%8C%E5%9B%BE%E4%B9%A6%E9%A6%86%E9%A6%86%E8%97%8F%E5%BD%93%E4%B8%AD%E5%AD%98%E5%9C%A8%E5%A4%A7%E9%87%8F%E7%9A%84%E5%AD%A6%E6%9C%AF%E6%96%87%E7%8C%AE%EF%BC%8C%E5%8F%AF%E4%BB%A5%E5%B0%86%E8%BF%99%E4%BA%9B%E5%AD%A6%E6%9C%AF%E6%96%87%E7%8C%AE%E9%80%9A%E8%BF%87%E4%B8%93%E4%B8%9A%E7%9A%84%E5%90%91%E9%87%8F%E5%8C%96%E5%B7%A5%E5%85%B7%E8%BF%9B%E8%A1%8C%E5%88%87%E7%89%87%E5%AD%98%E5%82%A8%E5%9C%A8%E5%90%91%E9%87%8F%E6%95%B0%E6%8D%AE%E5%BA%93%E5%BD%93%E4%B8%AD%EF%BC%8C%20%E5%BD%A2%E6%88%90%E4%B8%80%E4%B8%AA%E6%9C%AC%E5%9C%B0%E7%9F%A5%E8%AF%86%E5%BA%93%EF%BC%8C%E5%86%8D%E5%B0%86%E6%9C%AC%E5%9C%B0%E7%9F%A5%E8%AF%86%E5%BA%93%E4%B8%8E%E5%A4%A7%E8%AF%AD%E8%A8%80%E6%A8%A1%E5%9E%8B%E8%BF%9B%E8%A1%8C%E7%BB%93%E5%90%88%EF%BC%8C%E5%BD%93%E7%94%A8%E6%88%B7%E6%8F%90%E5%87%BA%E9%97%AE%E9%A2%98%E6%97%B6%EF%BC%8C%E7%B3%BB%E7%BB%9F%E8%83%BD%E5%A4%9F%E6%A0%B9%E6%8D%AE%E7%94%A8%E6%88%B7%E7%9A%84%E9%97%AE%E9%A2%98%E6%A3%80%E7%B4%A2%E5%9B%BE%E4%B9%A6%E9%A6%86%E7%9A%84%E6%95%B0%E5%AD%97%E8%B5%84%E6%BA%90%EF%BC%8C%E5%B9%B6%E7%94%9F%E6%88%90%E5%87%86%E7%A1%AE%E7%9A%84%E5%9B%9E%E7%AD%94%E3%80%82%E8%BF%99%E5%8F%AF%E4%BB%A5%E5%A4%A7%E5%A4%A7%E5%87%8F%E5%B0%91%E7%94%A8%E6%88%B7%E6%9F%A5%E6%89%BE%E4%BF%A1%20%E6%81%AF%E7%9A%84%E6%97%B6%E9%97%B4%EF%BC%8C%E5%B9%B6%E6%8F%90%E9%AB%98%E5%9B%9E%E7%AD%94%E7%9A%84%E7%9B%B8%E5%85%B3%E6%80%A7%E3%80%82)[calsp.cn](https://www.calsp.cn/2024/08/21/bulletin-202408-05/#:~:text=%E6%B8%85%E5%8D%8E%E5%A4%A7%E5%AD%A6%E5%9B%BE%E4%B9%A6%E9%A6%86%E4%B8%BA%E5%BF%AB%E9%80%9F%E6%8E%A8%E8%BF%9B%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E5%9C%A8%E5%9B%BE%E4%B9%A6%E9%A6%86%E7%9A%84%E8%90%BD%E5%9C%B0%E5%BA%94%E7%94%A8%EF%BC%8C%E4%BA%8E2024%E5%B9%B44%E6%9C%8828%E6%97%A5%E6%AD%A3%E5%BC%8F%E4%B8%8A%E7%BA%BF%E4%BA%86AI%E9%98%85%E8%AF%BB%E5%8A%A9%E6%89%8B%E3%80%81AI%E5%AF%BC%E8%88%AA%E5%8A%A9%E6%89%8B%E5%BA%94%E7%94%A8%E3%80%82%E9%80%9A%E8%BF%87%E6%99%BA%E8%83%BD%E5%88%86%E6%9E%90PDF%E6%96%87%E6%A1%A3%E5%92%8C%E6%95%B0%E6%8D%AE%E5%BA%93%E5%86%85%E5%AE%B9%EF%BC%8C%E6%8F%90%E5%8D%87%E4%BA%86%E5%B8%88%20%E7%94%9F%E7%9A%84%E9%98%85%E8%AF%BB%E5%92%8C%E7%A0%94%E7%A9%B6%E6%95%88%E7%8E%87%E3%80%82%E4%B8%A4%E9%A1%B9%E5%BA%94%E7%94%A8%E5%9D%87%E9%9D%A2%E5%90%91%E6%B8%85%E5%8D%8E%E5%85%A8%E6%A0%A1%E5%B8%88%E7%94%9F%E6%8F%90%E4%BE%9B%E6%9C%8D%E5%8A%A1%EF%BC%8C%E4%B8%8A%E7%BA%BF%E5%90%8E%E6%94%B6%E5%88%B0%E4%BA%86%E8%AF%BB%E8%80%85%E7%9A%84%E8%89%AF%E5%A5%BD%E5%8F%8D%E9%A6%88%E3%80%82)，极大降低了信息检索门槛、提高了咨询效率。
* 在**用户服务**方面，智能体支持的个性化推荐和知识推送实现了按用户兴趣主动送达知识，图书馆服务从被动响应转为主动关怀[whly.gd.gov.cn](https://whly.gd.gov.cn/gkmlpt/content/4/4404/post_4404325.html" \l ":~:text=%E4%BD%9B%E5%B1%B1%E2%80%9C%E6%98%93%E6%9C%AC%E4%B9%A6%E2%80%9D%E9%A1%B9%E7%9B%AE%E5%85%A5%E9%80%89%E6%96%87%E5%8C%96%E5%92%8C%E6%97%85%E6%B8%B8%E9%83%A8%E6%99%BA%E6%85%A7%E5%9B%BE%E4%B9%A6%E9%A6%86%E5%88%9B%E6%96%B0%E5%BA%94%E7%94%A8%E4%BC%98%E7%A7%80%E6%A1%88%E4%BE%8B%20%E6%A0%B9%E6%8D%AE%E7%94%A8%E6%88%B7%E5%85%B4%E8%B6%A3%E7%88%B1%E5%A5%BD%E7%AD%89%E4%BF%A1%E6%81%AF%EF%BC%8C%E6%8A%BD%E8%B1%A1%E5%87%BA%E8%AF%BB%E8%80%85%E6%A8%A1%E5%9E%8B%EF%BC%8C%E5%B0%86%E9%A2%84%E8%AE%BE%E7%9A%84%E7%94%A8%E6%88%B7%E4%B8%AA%E6%80%A7%E5%8C%96%E6%A0%87%E7%AD%BE%E5%92%8C%E5%9B%BE%E4%B9%A6%E6%A0%87%E7%AD%BE%E5%85%B3%E8%81%94%EF%BC%8C%E8%87%AA%E5%8A%A8%E6%A0%B9%E6%8D%AE%E9%A2%84%E8%AE%BE%E7%9B%B8%E5%85%B3%E5%BA%A6%E6%9D%83%E9%87%8D%E5%80%BC%E5%90%91%E7%94%A8%E6%88%B7%E6%8E%A8%E8%8D%90%E5%9B%BE%E4%B9%A6%EF%BC%8C%E5%AE%9E%E7%8E%B0%E7%B2%BE%E5%87%86%E5%8C%96%E3%80%81%E4%B8%AA%E6%80%A7%E5%8C%96%E6%8E%A8%E8%8D%90%E3%80%82)[nju.edu.cn](https://www.nju.edu.cn/info/1056/340981.htm#:~:text=%E9%AB%98%E6%A0%A1%E5%9B%BE%E4%B9%A6%E9%A6%86%E5%A6%82%E4%BD%95%E5%BC%80%E5%B1%95%E9%98%85%E8%AF%BB%E6%8E%A8%E5%B9%BF%EF%BC%9F%E6%B1%9F%E8%8B%8F%E8%BF%99%E6%89%B9%E7%B2%BE%E5%93%81%E6%A1%88%E4%BE%8B%E7%BB%99%E4%BD%A0%E7%AD%94%E6%A1%88%20)。
* 在**阅读推广和教育**方面，可视化、交互式的AI应用激发了用户兴趣，让学习过程更具沉浸感和趣味性[fjlib.net](https://www.fjlib.net/zt/mtxh/stxk/llysj/fjstsgxk201403/201804/t20180414_358197.htm" \l ":~:text=%E5%B9%B4%E5%9C%A8%E3%80%8A%E7%A7%91%E5%AD%A6%E3%80%8B%E6%9D%82%E5%BF%97%E4%B8%8A%E5%8F%91%E8%A1%A8%E7%9A%84%E4%B8%80%E7%AF%87%E8%AE%BA%E6%96%87%E4%B8%AD%EF%BC%8C%E5%B0%B1%E8%AE%A4%E4%B8%BA%E4%BC%A0%E7%BB%9F%E7%9A%84%E5%9F%BA%E4%BA%8E%E5%AD%A6%E7%A7%91%E7%9A%84%E6%A3%80%E7%B4%A2%E5%92%8C%E6%91%98%E8%A6%81%E6%9C%8D%E5%8A%A1%E9%9A%BE%E4%BB%A5%E6%BB%A1%E8%B6%B3%E7%A0%94%E7%A9%B6%E7%9A%84%E9%9C%80%E6%B1%82%E3%80%82)[whly.gd.gov.cn](https://whly.gd.gov.cn/service_newggwhzx/content/post_4061139.html#:~:text=%E7%AB%99%E5%9B%BE%E4%B9%A6%E8%B5%84%E6%BA%90%2C%E8%BF%98%E5%8F%AF%E8%8E%B7%E5%8F%96%E5%85%AC%E5%85%B1%E5%9B%BE%E4%B9%A6%E9%A6%86%E8%97%8F%E4%B9%A6%E6%8C%87%E5%BC%95%2C%E5%AE%9E%E7%8E%B0%E4%BA%86%E5%85%AC%E7%A7%81%E8%97%8F%E4%B9%A6%E7%9A%84%E8%9E%8D%E5%90%88%E5%88%A9%E7%94%A8%2C%E6%8F%90%E9%AB%98%E4%BA%86%E5%9B%BE%E4%B9%A6%E8%8E%B7%E5%8F%96%E6%95%88%E7%8E%87%2C%E6%BB%A1%E8%B6%B3%E4%BA%86%E5%B8%82%E6%B0%91%E9%98%85%E8%AF%BB%E9%9C%80%E6%B1%82%E3%80%82)。

试点结果显示这些应用普遍**提升了用户满意度**和**资源利用率**。例如，清华大学图书馆的AI助手上线后广受师生好评，被认为大幅节省了文献查找分析时间[calsp.cn](https://www.calsp.cn/2024/08/21/bulletin-202408-05/" \l ":~:text=%E6%B8%85%E5%8D%8E%E5%A4%A7%E5%AD%A6%E5%9B%BE%E4%B9%A6%E9%A6%86%E4%B8%BA%E5%BF%AB%E9%80%9F%E6%8E%A8%E8%BF%9B%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E5%9C%A8%E5%9B%BE%E4%B9%A6%E9%A6%86%E7%9A%84%E8%90%BD%E5%9C%B0%E5%BA%94%E7%94%A8%EF%BC%8C%E4%BA%8E2024%E5%B9%B44%E6%9C%8828%E6%97%A5%E6%AD%A3%E5%BC%8F%E4%B8%8A%E7%BA%BF%E4%BA%86AI%E9%98%85%E8%AF%BB%E5%8A%A9%E6%89%8B%E3%80%81AI%E5%AF%BC%E8%88%AA%E5%8A%A9%E6%89%8B%E5%BA%94%E7%94%A8%E3%80%82%E9%80%9A%E8%BF%87%E6%99%BA%E8%83%BD%E5%88%86%E6%9E%90PDF%E6%96%87%E6%A1%A3%E5%92%8C%E6%95%B0%E6%8D%AE%E5%BA%93%E5%86%85%E5%AE%B9%EF%BC%8C%E6%8F%90%E5%8D%87%E4%BA%86%E5%B8%88%20%E7%94%9F%E7%9A%84%E9%98%85%E8%AF%BB%E5%92%8C%E7%A0%94%E7%A9%B6%E6%95%88%E7%8E%87%E3%80%82%E4%B8%A4%E9%A1%B9%E5%BA%94%E7%94%A8%E5%9D%87%E9%9D%A2%E5%90%91%E6%B8%85%E5%8D%8E%E5%85%A8%E6%A0%A1%E5%B8%88%E7%94%9F%E6%8F%90%E4%BE%9B%E6%9C%8D%E5%8A%A1%EF%BC%8C%E4%B8%8A%E7%BA%BF%E5%90%8E%E6%94%B6%E5%88%B0%E4%BA%86%E8%AF%BB%E8%80%85%E7%9A%84%E8%89%AF%E5%A5%BD%E5%8F%8D%E9%A6%88%E3%80%82)；南京理工大学图书馆的个性推荐小程序成功地将读者从社交媒体引流回馆，实现借阅量增长[nju.edu.cn](https://www.nju.edu.cn/info/1056/340981.htm" \l ":~:text=%E9%AB%98%E6%A0%A1%E5%9B%BE%E4%B9%A6%E9%A6%86%E5%A6%82%E4%BD%95%E5%BC%80%E5%B1%95%E9%98%85%E8%AF%BB%E6%8E%A8%E5%B9%BF%EF%BC%9F%E6%B1%9F%E8%8B%8F%E8%BF%99%E6%89%B9%E7%B2%BE%E5%93%81%E6%A1%88%E4%BE%8B%E7%BB%99%E4%BD%A0%E7%AD%94%E6%A1%88%20)。这说明智能体技术完全可以“接地气”地融入图书馆日常服务，为读者创造实实在在的价值。当然，我们也需注意应用中的问题，如AI回答的准确性和可信度仍需人工监督兜底，隐私保护需贯穿始终，技术成本投入需要评估产出。

**第五，要发挥智能体在图书馆的更大作用，还需在人力和组织上做好准备。** 技术本身只是手段，图书馆要真正实现智慧化转型，还需要**馆员角色和技能的转变**。馆员应积极学习掌握AI工具的使用，理解其局限和原理，才能有效引导读者并优化AI服务。这方面的培训和团队建设尤为重要。某种意义上，馆员将从“信息的提供者”转变为“**人与AI协作的服务设计师**”。另一方面，图书馆应制定相应的**服务规范和伦理准则**，明确AI服务边界和应急预案。例如当AI出现错误时如何介入补救，当用户依赖AI导致学习能力下降时如何平衡等。这需要管理层的前瞻规划和与IT部门的紧密合作。此外，不同规模和类型的图书馆在采用智能体时也应结合自身定位，不必追求面面俱到，而应选择**契合本馆读者需求**的应用切入点。总之，技术赋能必须与人文关怀相结合，图书馆应保持以读者为中心的服务初心，利用智能体锦上添花，而非喧宾夺主。

综上所述，智能体/代理型AI作为人工智能发展的前沿方向，正在逐步走向实用，并为图书馆行业带来了**前所未有的创新机遇**。通过本文的研究，我们既看到了智能体技术曲折演进、蓄势待发的恢宏图景，也深入认识了其现实短板和应对之策。对于图书馆而言，智能体并非洪水猛兽或万能灵药，而是一个强有力的新工具，需要我们以专业精神去掌握和驾驭。只要坚持理性务实的态度，充分发挥图书馆人对知识和用户的深刻理解，与新技术相辅相成，我们有理由相信：未来的图书馆将蜕变为**智慧信息服务中枢**，在智能体的助力下，以更精准、高效和个性化的方式将知识传递给每一位有需要的人。这既是技术发展的必然趋势，也是图书馆事业与时俱进、再创辉煌的崭新篇章。

# 4. 场景识别（Scene Recognition）

当今信息环境下，图书馆的知识服务正在经历深刻变革。用户需求日益多样化且情境化，传统的“以用户检索为中心”的服务模式面临挑战[[107]](#footnote-107) 。 用户不仅希望获取所需信息，更期望在**正确的时间、地点和情境**下获得个性化、智能化的知识支持。这就引出了“场景识别”（情境感知）技术在图书馆中的潜在应用：通过识别用户所处的场景（例如地点、时间、身份、任务等上下文），动态调整知识服务内容与方式，从而提升服务相关性和用户体验。

场景识别源自计算领域的情境感知（Context-Awareness）理念。早在20世纪90年代，中移动计算和泛在计算兴起，一些学者便提出让计算机自动感知环境变化并做出相应适应的愿景[[108]](#footnote-108) 。 图书馆界逐渐认识到，这一技术同样可以用于改进知识服务的效率和智慧化程度。**场景识别在图书馆知识服务中的演进与融合**已成为智慧图书馆研究的重要方向之一[[109]](#footnote-109) 。 据调查，超过六成的图书馆正计划将人工智能等技术整合到服务中，以实现个性化推荐、智能空间管理等高级应用[[110]](#footnote-110) 。这些趋势体现出，将场景识别融入知识服务已是大势所趋。

本章报告对场景识别在图书馆知识服务领域的发展进行了系统梳理和展望。首先，我们界定场景识别的含义及其关键要素；其次，回顾场景识别技术的起源和演进脉络，尤其关注其在图书馆知识服务中的应用历程；然后，分析当前场景识别的发展趋势与可能遭遇的技术瓶颈，重点讨论其在图书馆领域适配时面临的难点与挑战；最后，深入探讨场景识别与图书馆知识服务融合的可行性和契合度，并聚焦六大典型应用场景——知识图谱、语义检索、智能问答、个性化推荐、知识推送、可视化服务，结合实际案例展望未来演进方向。报告中将广泛引用国内外权威研究成果与实践案例，涵盖自然语言处理、深度学习、上下文建模、用户建模、图数据库、可视化交互设计等相关技术，以期为智慧图书馆的建设提供有价值的学术参考。

## 4.1 场景识别的定义

**场景识别**（Scene Recognition）在图书馆领域一般指对用户所处情境的自动感知与理解，从而辅助系统提供**情境相关**的知识服务。这一概念与计算机科学中的**情境感知**(Context Awareness)高度重合。Dey等人对“情境”（context）做出的经典定义是：\_“情境是能够用来刻画某实体（人、地点、对象等）所处情形的任何信息”[[111]](#footnote-111) 。 换言之，场景/情境包括用户、应用以及交互相关的各种要素，如当前的时间、地点、环境、设备、用户身份和需求状态等等[[112]](#footnote-112) 。 基于此，Abowd等人将**情境感知系统**定义为：“能够利用情境信息为用户提供相应的服务或信息的系统，其中服务/信息的相关性取决于用户当前的任务”[[113]](#footnote-113) 。 由此可见，场景识别的本质就是获取和理解用户周围及自身的各种上下文信息，并据此调整系统行为以保证服务内容与用户情境匹配[[114]](#footnote-114) 。

在图书馆知识服务背景下，场景识别关注的问题包括：**是谁在寻求信息、何时何地提出需求、目的为何、使用什么设备、偏好如何**等。比如，同一读者在工作日白天于图书馆阅览室检索文献，与在深夜通过手机远程访问数据库，其信息需求和期望的服务形式可能截然不同。情境感知技术正是要捕捉这些差异。Wei Gao等（2022）的研究指出，移动图书馆环境下用户的信息活动具有碎片化、临时性和实用性的特征，对应的信息需求往往与时间（何时需要）、空间（何地需要）、用户个人（是谁，需要什么）以及社交情境（他人评价与共享）等维度相关[[115]](#footnote-115) 。 因此，系统需要**感知用户的情境信息，实时分析判断其所处情境，并快速响应**，为用户提供方便快捷、贴合当下情境的知识服务[[116]](#footnote-116) 。 具体而言，情境信息可包括：当前时间（如学期周、节假日）、地点位置（馆内区域或馆外地点）、用户特征（身份角色、学科背景、历史借阅和偏好）、设备终端特性（手机/电脑）以及社交行为（点评、分享）等[[117]](#footnote-117) 。 场景识别技术通过传感器、日志分析、用户画像等手段获取这些数据，经由上下文模型进行理解与推理，最终将结果作用于服务策略上，使系统具备“感知环境、即时适应”的能力。

简而言之，本报告讨论的场景识别指的就是**情境感知智能**在图书馆知识服务中的应用。它要求图书馆系统像一个经验丰富的馆员一样，能够“察言观色”——感知读者所处的情境，揣摩其潜在需求，并提供恰如其分的知识支撑。例如，一个情境感知的参考咨询系统可以根据读者的专业背景和当前提问主题，自动调整回答的深度与专业术语程度[[118]](#footnote-118) ； 一个情境感知的推荐系统可以考虑读者当前所在地是馆内还是家中，从而决定推荐纸质书还是电子资源。这样的**场景自适应**能力正是智慧图书馆的重要特征之一[[119]](#footnote-119) 。

需要注意的是，计算机视觉领域也有“场景识别”概念（如识别图像中的场景类型），但在本报告中，我们聚焦的并非视觉场景分析，而是**用户信息行为情境**的识别，即更偏重语义和使用情境的理解。在英文文献中常用“Context Awareness（情境感知）”来表述这一概念。总之，场景识别（情境感知）为图书馆提供了一种**以用户为中心、以情境为线索**的服务新范式，其核心在于让系统理解“此时此地的这个用户真正需要什么”，从而动态调整知识服务内容与交互方式[[120]](#footnote-120) 。

## 4.2 场景识别技术的起源和关键发展阶段

**场景识别（情境感知）技术**的思想由来已久，其起源可以追溯到20世纪末计算机领域的前瞻性研究。1991年Mark Weiser提出“泛在计算”（Ubiquitous Computing）的理念，畅想计算将无处不在、主动感知环境[[121]](#footnote-121) 。 随后在1994年，Schilit等人在移动计算研讨会上首次明确提出“Context-Aware Computing”（情境感知计算）概念，描述了这样一种软件：\_“能够根据使用的地点、附近的人和设备集合，以及这些因素随时间的变化来自动调整自身”[[122]](#footnote-122) 。 这被视为场景识别技术的开端。早期的情境感知研究主要集中在**定位感知**（Location-Aware）应用上，例如根据用户位置提供相应服务[[123]](#footnote-123) 。 典型实例是卫星导航系统：它依据当前位置这一主要情境参数，自动调整地图显示并规划路线[[124]](#footnote-124) 。 随着研究深入，人们认识到情境不仅限于位置，还包括时间、环境光线、邻近人物等诸多方面[[125]](#footnote-125) 。 1999年Schmidt等提出了情境要素空间模型，将情境划分为环境、用户、时间等层次结构，强调应用需针对具体情境特征进行设计[[126]](#footnote-126) 。

进入21世纪，情境感知计算成为热门课题，相关定义和框架逐步完善。2000年前后，Dey和Abowd等对情境感知做了权威综述和定义[[127]](#footnote-127) 。 Chen & Kotz (2000)发表技术报告对移动情境感知研究进行了全面调查[[128]](#footnote-128) ， 梳理了当时情境建模与推理的各种方法（如基于键值对、基于面向对象、基于ontology的模型等）[[129]](#footnote-129) 。 这一时期的标志性成果还包括Dey于2001年开发的**Context Toolkit**，为情境感知应用提供了通用的软件支持框架，以及随后出现的各种情境**本体（Ontology）**和中间件，旨在标准化情境信息的表示与交换[[130]](#footnote-130) 。 例如，有研究者提出用OWL本体来表达RFID传感器捕获的情境数据，并基于语义推理实现动态场景管理[[131]](#footnote-131) 。这些工作为场景识别从理念走向实用奠定了基础。

**图书馆领域对场景识别技术的关注**则在21世纪的第二个十年逐渐兴起，与数字图书馆和智慧图书馆建设的演进紧密相关。早期的数字图书馆侧重资源数字化与网络获取，但服务模式大多仍是被动响应式的。随着Web2.0和移动互联网的发展，图书馆开始探索个性化和泛在服务。2010年前后，“**图书馆2.0**”推动用户参与和社交互联，而进一步的“**图书馆3.0**”概念则引入了**语义网和情境感知**（又称“泛在图书馆”或“智慧图书馆”）[[132]](#footnote-132) 。 韩国学者Noh在2013年的研究中系统展望了**下一代数字图书馆**的发展方向，明确提出应引入情境感知技术[[133]](#footnote-133) 。 他认为，未来的智慧图书馆应能够识别进入图书馆的用户身份（新用户或老用户），并针对不同情境提供最佳服务；情境感知型图书馆可以实现情境感知的参考咨询、情境感知的借阅服务，以及在阅览空间提供满足当下环境需求的支持[[134]](#footnote-134) 。 例如，当用户走近某一书架时系统自动推送相关资料指南，或监测用户行为以在紧急情况下提供援助[[135]](#footnote-135) 。 Noh的研究还指出，当时真正落实情境感知技术的图书馆实例还不多，但情境感知有望极大提升用户便利性和服务品质[[136]](#footnote-136) 。这一观点鼓舞了随后智慧图书馆实践对场景识别的探索。

2010年代中期，一些具体的场景识别在图书馆应用的原型系统相继出现。例如，芬兰奥卢市图书馆的“**UbiLibrary**”项目（2014）就开发了一种结合**语义信息**与**情境感知**的大型公共显示屏服务[[137]](#footnote-137) 。 该系统聚合了图书馆馆藏数据库和外部网络资源，通过元数据语义增强技术丰富了书目信息，并在图书馆大厅的电子屏上以标签云形式动态呈现[[138]](#footnote-138) 。 值得一提的是，UbiLibrary能够利用计算机视觉识别读者的大致年龄和性别，并据此自适应地调整图书推荐结果[[139]](#footnote-139) 。 比如，当系统检测到屏幕前是一位年轻读者，可能优先推荐其年龄段常读的书目；若是一位年长男性，则推荐不同类别的内容[[140]](#footnote-140) 。 这是场景识别（用户特征情境）用于图书馆个性化推荐的早期尝试，证明了上下文数据（如读者属性）对改进服务的价值。同一时期还有一些移动图书馆应用尝试利用GPS定位提供附近图书馆分馆信息、馆内导航服务等，也属于情境感知的实践范畴。

进入2020年代，随着**物联网（IoT）**、**大数据**和**人工智能**的发展，场景识别在图书馆知识服务中的应用更加可行且丰富。大量智能终端和传感器为采集情境数据提供了条件，读者的行为日志、偏好数据经由机器学习模型分析可以推断深层次需求。中国的移动图书馆研究显示，在智能物联网环境下构建情境感知的个性化服务体系已成为热点课题[[141]](#footnote-141) 。Wei Gao等设计了基于UML的情境感知移动图书馆服务模型，通过实验证明用户情境（时间、地点、个人偏好、终端等）与服务接受度存在显著相关[[142]](#footnote-142) 。 他们提出的体系架构包括情境数据获取层、处理层、推荐服务层和用户交互层等，以分层方式实现情境感知功能[[143]](#footnote-143) 。 另一方面，图书馆实践界开始引入**人工智能驱动**的服务，如智能问答系统和推荐算法。最新的调查表明，全球范围内许多图书馆正积极规划将AI融入服务流程，包括基于AI的个性化推荐和空间利用优化等[[144]](#footnote-144) 。 **大型语言模型**（如OpenAI的ChatGPT）在2022年掀起热潮，其**“上下文对话”**能力使智能问答服务取得突破[[145]](#footnote-145) 。 一些高校图书馆已经试验让读者用自然语言提问，由AI检索馆藏并给出参考答案，这实际上利用了模型对语义和上下文的理解来提升知识服务[[146]](#footnote-146) 。 可以预见，随着AI技术的发展，场景识别将进一步与知识图谱、深度学习等结合，推动图书馆知识服务进入更高智能化阶段。

纵观其发展历程，场景识别技术经历了从**理论提出（1990s）→基础设施奠定（2000s）→小规模试点（2010s）→融合AI加速发展（2020s）**的进程。在图书馆领域，从最初意识到用户情境重要性，到逐步试验情境感知服务，再到如今将之作为智慧图书馆建设的重要标志，我们看到了明显的演进轨迹[[147]](#footnote-147) 。 这一演进也是多学科技术协同的结果：物联网为情境数据采集提供了手段，知识组织和语义网技术为情境信息的表达和利用提供了语义支持，机器学习使从海量行为数据中识别情境模式成为可能，自然语言处理则让系统更好地理解用户语义上下文。可以说，场景识别正日益成为智慧图书馆不可或缺的关键技术之一，为知识服务的创新拓展了广阔空间。

## 4.3 当前场景识别的发展趋势和技术瓶颈

### 4.3.1 发展趋势：从被动响应到主动智能

**发展趋势1：更加主动的情境感知服务。**传统图书馆服务多为被动式，即读者明确提出请求，系统再响应。而场景识别的引入，使服务逐渐走向**主动推送**和**预判需求**。现代用户习惯了智能助手、推荐引擎那样的体验，期望图书馆服务也能“懂我”。因此，越来越多图书馆尝试基于场景触发提供服务：例如当读者走进图书馆馆舍时，手机自动收到欢迎信息和馆内活动推介（基于定位情境）；当用户在OPAC检索某主题文献时，系统即时推荐相关的数据库或咨询服务（基于行为情境）等。这种由**上下文事件触发服务呈现**的模式，在其他领域已有成功实践。例如旅游领域的CAIPS模型提出通过检测游客位置或行为等上下文事件来触发信息推送[[148]](#footnote-148) 。在图书馆，同理可以构建“情境-事件-响应”规则，实现用户达到某情境就自动提供对应服务。情境感知服务从过去的实验性应用，正走向实际部署。尤其移动端App与馆内物联网基础设施结合，为主动情境服务奠定基础。

**发展趋势2：多源数据融合与深度学习。**当前场景识别技术呈现出融合多模态数据、借助深度学习模型提升精度的趋势[[149]](#footnote-149) 。 过去，情境获取主要依赖单一来源（如RFID定位或用户配置的偏好），而如今传感器网络和大数据使我们能够获取**丰富的上下文线索**：包括地理位置、环境传感器读数、社交媒体痕迹、用户历史行为、文本交互记录等。通过将这些异构数据融合，能够更全面地刻画用户场景。例如，一套情境识别模型可能综合：GPS定位判断用户是否在图书馆内，日历时间推断其可能从事的活动（考试周=复习情境），借阅历史和当前检索关键词分析其研究主题，以及社交平台上关注的话题推测其兴趣热点。深度学习模型（如RNN、Transformer等）可以在这样的多模态输入下学习复杂的情境模式，提高识别准确率[[150]](#footnote-150) 。 有研究利用**循环神经网络**融合室内传感和室外GPS数据，实现高精度的用户场景（室内/室外等活动场景）识别[[151]](#footnote-151) 。 在图书馆语境中，深度学习也可用于用户意图识别和需求预测。例如，通过对大量查询日志训练模型，预测用户在某情境下最可能需要哪类资源，以便提前准备或推荐。

**发展趋势3：知识图谱与语义情境的结合。**随着知识图谱（Knowledge Graph）技术的发展，利用图谱丰富的语义关联来辅助情境理解成为一大趋势[[152]](#footnote-152) 。 知识图谱可视为连接实体及概念的网络，当用户的行为与某些知识节点相关时，我们可以拓展出其语义“上下文”。例如，若系统识别用户当前研究主题是“机器学习”，通过知识图谱联想，可以得知其相关概念还有“人工智能、深度学习、数据挖掘”等，那么服务可扩展推荐这些相关领域的新资源。在智能制造领域已有将知识图谱用于生成情境感知服务的案例[[153]](#footnote-153) 。同样地，在图书馆，可构建**读者知识图谱**或**书目知识图谱**，将用户背景、兴趣与文献语义网连接，用于增强场景识别的语义层面。例如，郑州大学的研究者构建了“图书知识图谱”并开发知识服务系统，实现了智能语义检索、知识问答等功能[[154]](#footnote-154) 。这提示我们，通过知识图谱的推理能力，系统可以更好地理解用户询问背后的语义场景，并提供关联知识的推送。知识图谱还可以积累情境与服务效果的关联数据，供日后分析和改进情境模型。

**发展趋势4：大型语言模型助力情境对话和问答。**近年来的生成式预训练模型（如GPT系列）展现出强大的**上下文理解和生成**能力，对图书馆智能服务产生重大影响[[155]](#footnote-155) 。 ChatGPT等模型可以依据对话上下文连贯地回答复杂问题，具备一定的**对话情境记忆**和推理能力。这类模型引入图书馆后，一方面在用户前端可以提供更为自然的对话式问答体验，理解用户长问句甚至连续提问的意图[[156]](#footnote-156) ； 另一方面，在系统后台也可用于情境推理——模型可以从非结构化的描述中提取情境信息。例如用户发来邮件咨询“我在准备一篇关于机器学习伦理的综述，目前手头资料不多，有什么推荐吗？”，传统系统很难直接处理，而大型语言模型可以分析出用户身份（可能是研究生）、目的（写综述）、主题（机器学习伦理）等情境，并据此与知识库匹配最佳资源进行回答[[157]](#footnote-157) 。各大图书馆系统厂商也在将LLM整合进发现系统，如Ex Libris的Primo研究助手能够接受自然语言查询并给出引用了馆藏文献的回答[[158]](#footnote-158) 。 可以预见，未来大型语言模型将成为情境感知的重要引擎之一，让机器对用户所处场景“领会得更灵敏”。

**发展趋势5：用户隐私与伦理考量。**在技术快速推进的同时，越来越多关注点落在用户隐私和伦理上。场景识别不可避免地涉及对用户行为和环境的广泛数据收集，这引发了隐私保护的讨论。未来的发展趋势之一是**隐私敏感的情境感知**设计，即在保障服务智能化的同时，严格控制对用户私密数据的使用。例如，通过在终端设备上进行情境计算，避免将原始数据上传云端，或采用差分隐私技术对数据进行模糊处理。图书馆作为公共机构，更需注重在智慧服务中保护读者隐私和数据安全。此外，还有**算法透明度**和**公平性**的问题：情境感知算法如何做出决策应当有迹可循，以便获得用户信任；算法不应因利用某些情境数据而产生对特定群体的歧视性结果（例如过度推送某类信息）。这些都成为趋势中需要平衡的重要方面[[159]](#footnote-159) 。 未来，图书馆情境感知服务的设计将融入更多伦理规范，确保技术应用符合“以人为本、公平公正”的原则。

### 4.3.2 潜在技术瓶颈与挑战

尽管场景识别在图书馆知识服务中的前景令人期待，但在实际落地过程中仍面临诸多**技术瓶颈和挑战**：

* **数据获取与整合困难：**情境感知依赖多源数据，但许多图书馆目前的数据基础仍相对分散、孤立。例如，读者借阅数据在集成库系统(ILS)中，数字资源使用数据在数据库平台，门禁出入数据在安防系统，用户在线行为在网站或App日志。这些数据彼此缺乏联通，情境分析需要的**全局用户画像**难以建立。此外，有些情境信息（如读者当前位置、实时行为）需通过传感器或移动设备获取，但部署这些硬件对于经费有限的图书馆而言也是挑战。即使获取了多种数据，如何将异构数据清洗、关联，形成统一的情境表示也是一大技术难点[[160]](#footnote-160) 。 没有高质量的数据，“巧妇难为无米之炊”，场景识别算法将无从发挥。
* **情境建模与推理复杂：**即便有了数据，如何对情境进行有效的**建模与推理**也是瓶颈之一。情境具有高度的动态性和不确定性，同样的信号在不同情况下可能意义迥异。例如，用户深夜登录数据库可能意味着他正在加班写论文，也可能只是失眠随便浏览。要让系统正确“读懂”情境，需要强大的推理机制和上下文知识。尽管语义网技术和机器学习提供了部分解决方案，但目前的情境建模仍不够成熟，缺乏统一标准和易用的开发框架[[161]](#footnote-161) 。Wei Gao等指出当前对于情境感知服务模型的研究仍显不足[[162]](#footnote-162) 。 如果模型不准确，可能出现**情境误判**，导致提供不合适的服务，反而降低用户体验。例如系统错将某用户识别为教师从而推送学术资源，但其实是个本科生，收到太深奥的信息会适得其反。
* **跨领域知识融合挑战：**图书馆情境感知涉及图书情报学知识与计算技术的深度融合。这要求馆员和开发者既理解读者服务，又掌握AI、NLP等技术。然而现实中，这样的**复合型人才**相对缺乏，团队协作成本高。很多图书馆依赖外部厂商提供智能服务解决方案，但厂商对图书馆特定业务场景未必有深入理解，导致系统难以**贴合馆情**。例如，一个通用的推荐算法可能不了解图书馆的分类体系和学术资源特点，推荐结果不理想。此外，情境感知系统需要持续调优，本地化定制，传统图书馆IT人员编程或数据分析能力不足也是瓶颈。如何打造**图书馆与技术专家合作**的机制，一起打磨适用的情境模型，是目前的一大挑战。
* **系统集成与实时性能：**在图书馆环境中引入场景识别，需要将情境感知模块与现有业务系统（OPAC、数字图书馆门户、移动App等）进行集成。这可能牵涉不同厂商的软件接口兼容问题，增加实现难度。另外，情境感知往往要求**实时性**：捕捉情境-分析-响应必须在短时间内完成，用户才有良好体验。例如读者走到书架前几秒内就应该收到导引信息。如果情境识别和服务触发耗时太久，用户可能已经离开。这对系统性能、算法效率提出了高要求。大规模并发用户情况下如何保障实时响应，也是技术瓶颈之一。相关硬件投入（如定位基站、边缘计算节点）以及高效算法研究都需要跟进。
* **用户接受度与反馈机制：**从读者角度看，情境感知服务是一柄双刃剑。如果做得好，会被视为体贴聪明的助手；做得不好，可能引起反感甚至隐私担忧。一大挑战在于**用户接受度**：并非所有读者都愿意让系统“随时跟踪”他们的行为。有人会担心个人阅读隐私泄露，或觉得系统干预过多不胜其烦。因此在实现场景识别时需设计**透明、可控**的机制，例如允许用户自主设定愿意开放的情境数据、提供一键开启/关闭情境服务的选项等。此外，需要建立**用户反馈机制**：当系统的情境推断出错或服务不符合预期时，用户能方便地纠正或反馈，这样系统才能持续学习改进。如果缺乏反馈，系统可能一直固守错误的情境模型。用户教育也是瓶颈之一——图书馆需要向读者解释情境感知服务的益处及数据使用边界，争取信任和支持。
* **安全与隐私合规：**前面提到隐私保护，这里强调其带来的技术和制度挑战。情境数据可能包含敏感信息（如身份、行为轨迹），一旦泄露后果严重。图书馆必须确保自身及合作厂商在收集、存储、处理这些数据时遵循严格的安全标准（如数据加密、访问控制）和隐私法规（如《GDPR》等）。技术上，需要实现**数据匿名化和最小化**原则，仅保留服务所需的最低限度信息。例如采用代号而非真实身份记录行为。制度上，要制定明确的隐私政策与应急预案。一旦发生数据泄漏或算法偏见事件，能够及时响应和补救。这些要求对传统图书馆IT是新的课题，需要与法律、安全专家合作解决。

综上，场景识别在图书馆知识服务应用中正处于**机遇与挑战并存**的阶段。一方面，新技术为实现更智能的情境服务提供了可能，另一方面，数据、算法、人才、用户等多方面的瓶颈需一一突破。正如有研究指出的，尽管情境感知的研究成果日益丰富，但真正用于图书馆服务的还不多，主要原因就在于上述困难[[163]](#footnote-163) [[164]](#footnote-164) 。 未来能否成功大规模部署场景识别，取决于我们在技术上逐步完善并在实践中积累经验，找到**平衡智能服务与用户信任、系统复杂度与可靠性的最佳方案**。

## 4.4 场景识别与图书馆知识服务融合的应用场景

场景识别技术的最终价值在于与具体的图书馆知识服务**场景（应用领域）**相融合，产生协同增效的智能服务模式。图书馆知识服务涵盖多个方面，下文将聚焦六大典型应用场景，探讨情境感知（场景识别）如何在每个场景中发挥作用，以及二者融合所带来的可能性与契合度。这六个领域分别是：**知识图谱**、**语义检索**、**智能问答**、**个性化推荐**、**知识推送**和**可视化服务**。每一项都是当前智慧图书馆建设的热点，也是场景识别大显身手的舞台。

### 4.4.1 知识图谱：情境语义关联与知识组织

**知识图谱**是一种以语义网络形式组织知识的技术，在图书馆领域通常用于关联馆藏资源、学术概念和实体。将场景识别融入知识图谱，有助于实现**情境化的知识组织与获取**。其契合点主要体现在两个方面：

一是利用知识图谱增强**情境语义理解**。当系统识别出用户所处的情境（如研究主题或任务），可以在知识图谱中找到与该情境相关的实体和关联关系，作为扩展的语义上下文。例如，若检测到用户在查找“机器学习伦理”的资料，图谱能够提供该主题相关的概念（算法透明度、AI法规、数据隐私等）的网络[[165]](#footnote-165) 。系统据此可以向用户推荐更全面的参考资源，而不是仅局限于字面匹配的结果。实际上，有研究已将领域知识图谱用于提升情境感知服务的智能性，比如在智能制造中构建工业知识图谱来产生情境感知的知识服务[[166]](#footnote-166) 。在图书馆，同理可建立学科知识图谱或馆藏知识图谱，使系统具备“知识背景”，更好地理解用户查询背后的意图和所需知识范围。

二是利用情境信息动态调节**知识图谱的查询与显示**。知识图谱往往规模庞大，包含错综复杂的关系。场景识别可以帮助**过滤和聚焦**图谱中与当前情境最相关的部分，从而提高知识获取效率。例如，对于一位医学院学生，系统识别其身份和偏好后，在医学知识图谱中优先呈现临床医学相关的知识路径，而对计算机科学领域的节点予以淡化。这就相当于根据用户情境对知识图谱做了**个性化裁剪**。再如，北京大学等机构开发的一些学科知识图谱系统，会根据用户选择的研究主题高亮相关概念和文献网络，让用户在该情境下快速浏览关联知识点。类似地，场景识别可用于**排序**知识图谱的查询结果：如果用户当前处于写毕业论文的场景，系统可优先显示学术性更强、引用次数高的关联文献节点；如果用户只是科普兴趣，则优先显示通俗易懂的概念节点。

知识图谱与场景识别的融合还体现在新型服务模式上。例如，创建**情境化知识导航**系统：当用户进入某个专题场景时，系统自动生成一张该专题的知识图谱视图，让用户直观看到主要概念和文献的关联脉络（这也是可视化服务的一种体现）。一项针对数字人文遗产的研究已经展示了这种思路：通过知识图谱系统的可视化界面支持用户浏览数字资源[[167]](#footnote-167) 。在图书馆，我们可针对某些热门研究课题预先构建知识脉络图，配合情境感知在用户需要时推送展现。又如**情境问答**（IQ&A）系统，也可以以知识图谱为知识基底，通过情境识别确定用户提问涉及的图谱子域，从而在小范围内精准推理答案，提高问答准确率[[168]](#footnote-168) 。

当然，这一领域的挑战在于构建和维护高质量的图书馆知识图谱，以及开发高效的情境到语义的映射算法。但总体而言，知识图谱提供了情境感知一个极好的语义支撑平台，使图书馆知识服务更好地“知其然并知其所以然”。情境感知可以赋予知识图谱以“智能滤镜”，让庞杂的知识网络在合适的场景下呈现出恰当的片段与视角。两者结合，将促成一种**语义驱动的智慧知识服务**：既有深厚的知识关联，又有敏锐的情境判断，从而极大提升用户获取知识的效率和体验。

### 4.4.2 语义检索：理解意图的情境搜索

**语义检索**旨在超越传统基于关键词的检索，通过理解用户查询的语义和意图来提供更准确的结果[[169]](#footnote-169) 。在图书馆知识服务中，引入场景识别可以使语义检索更具“语境意识”，从而提升检索精准度和用户满意度。二者融合的契合点包括：

* **情境增强的意图识别：**语义检索引擎通常利用自然语言处理来解析查询背后的意图和含义。例如，它会将查询词映射到概念、扩展同义词等。然而，仅凭查询词本身有时不足以判断用户真正所求。这时，如果结合用户当前情境，理解就会更准确[[170]](#footnote-170) 。 举例来说，用户输入“java 安全”进行检索，如果知道此用户是计算机专业背景且在校，语义检索应倾向于理解为编程语言Java的安全性；若用户是生态学者，可能是寻找爪哇岛的生态安全研究。**情境信息（领域、身份）**帮助消除歧义。再如，同样是搜索“鼠标”，程序员想找计算机鼠标信息，生物学者则想找实验小鼠资料。通过场景识别预先判断用户所属领域，语义检索就能做出正确的意图识别，从而检索到相关度更高的结果[[171]](#footnote-171) 。国内一些数据库已经开始融合大模型的意图识别能力，比如知网的AI增强检索能够捕捉用户检索意图，简化复杂检索流程[[172]](#footnote-172) 。 这些实践表明，情境感知的意图识别将成为语义检索的重要环节。
* **动态结果排序与过滤：**传统检索结果排序大多根据文本相关度或全局学术影响力排序，而场景融合的语义检索可以根据用户当前任务情境动态调整排序。例如，当识别出用户正处于撰写综述的场景，可以在结果中优先显示该领域的综述性文章或高被引论文；若识别出用户是初学者入门场景，则优先展示基础教材或导论级别的材料。情境信息还可用于**结果过滤**。比如，当用户在馆内终端检索，系统可优先显示馆藏纸本可供借阅的结果（匹配地点情境）；当用户使用手机并且在校外，系统则滤掉无法远程访问的资源，避免用户点击后遭遇权限限制的挫败感。这种情境感知的排序与过滤能极大提升检索体验，让用户更快找到适合自己当下情境的资料。正如Google等搜索引擎在通用领域已经根据用户位置、历史记录调整结果一样，图书馆语义搜索也应因人因时而变。
* **对话式检索与上下文记忆：**情境感知还表现在多轮检索对话中。语义检索越来越多地支持对话查询，即用户可以逐步细化或更改查询。在这种场景下，前文上下文就是用户的情境之一。情境感知技术让检索引擎能够“记住”用户前面的提问，从而正确解析省略词或代词指代的对象。例如：用户先问“有关于人工智能伦理的文献吗？”得到结果后接着问“那它在医疗领域的应用呢？”第二问中的“它”指代人工智能伦理，此类承接需要情境记忆能力。现代AI驱动的检索助手（如Primo Research Assistant）正是通过大型语言模型实现对话理解，能够提供带引用的回答并链接全文[[173]](#footnote-173) 。 这也可以看作语义检索在更高层次上的情境融合——把整个对话视为情境，理解用户逐步澄清需求的过程并及时调整检索策略，给出连贯相关的结果。未来，图书馆用户可能更倾向于这种自然的对话式信息检索，因此情境感知在其中的作用将更加关键。

值得一提的是，语义检索和知识图谱常常结合实现，即利用知识图谱做语义扩展与精排。这一点在上一节已讨论。这里强调的是，无论内部实现如何，引入场景识别都会使检索更加“懂你所需”。图书馆的一些智能检索系统已开始号称支持“深度理解用户检索意图”[[174]](#footnote-174) ，其实质就是情境感知能力的体现——既理解查询，也了解查询者。总而言之，场景识别赋予语义检索以**“因人而异、因境而变”**的智慧，使检索从冷冰冰的匹配行为变成富有人性化的交互过程，为用户提供更**贴合语境的检索体验**[[175]](#footnote-175) 。

### 4.4.3 智能问答：情境驱动的参考咨询

**智能问答（Intelligent Q&A）**服务是图书馆知识服务的一项重要内容，旨在解答用户各类信息需求和问题咨询。传统的图书馆参考咨询由馆员人工完成，而借助自然语言处理和知识库的智能问答系统正在兴起。场景识别在其中的融合，可以使问答更加**上下文相关和个性化**：

* **理解提问背后的场景：**用户在提问时往往不会把背景交代清楚，但馆员通常会通过经验猜测提问者的场景，从而给出合适的答案。智能问答系统若具备这种情境理解能力，将更接近真人水平。比如，用户提问：“请问怎么引用参考文献？” 如果识别用户是大一新生且晚上在自习室提问，这可能是基本的学术写作问题，回答可提供引用格式指南；如果提问者是研究生且白天在实验室，很可能需要高级参考管理工具推荐。场景识别通过用户身份（新生/研究生）、时间地点（晚自习/日间科研）等推断提问背景，进而帮助系统**选择回答策略**。再如用户问：“这本书有电子版吗？” 如果检测到其身处馆外且时间是周末，那么更可能希望获得电子版链接；若在馆内，则可能只是没找到纸本，需要馆员帮助。这些情境使得问答系统能“听懂弦外之音”。
* **基于情境的答案定制：**对于同一个问题，不同用户在不同情境下可能需要不同深度或形式的答案。情境感知允许系统**定制回答**。例如，问“什么是量子计算？” 初学者需要通俗易懂的解释[[176]](#footnote-176) ，专业学者则希望听到更严谨甚至数学推导层面的说明。这种差异可以通过用户模型（背景知识水平）作为情境输入来调整答案的专业程度。同样地，馆员回答问题时也会考虑对方是谁。智能系统则可利用情境信息在知识库中**选取适配的答案**：比如对于大众读者引用Wikipedia、科普文章的表述，对于专家则引用期刊文献定义。又如答复形式上，如果用户在手机上询问路线，回答可以直接给地图定位（视觉化的答案）；如果在PC上，可以给详细文字说明和链接。**情境驱动**确保答案对用户而言是实用而友好的[[177]](#footnote-177) 。
* **多轮问答中的上下文关联：**类似前述对话式检索，在智能问答中情境感知也体现在保持多轮对话的上下文。用户可能连续提多个相关问题，此时之前的问题和回答构成了新的情境。系统需要“记住”用户已获得的信息，不重复回答，并据此推断接下来更深层的需求。例如：用户先问“如何查找某期刊论文？”，系统指导了数据库检索。紧接着用户问“这个数据库能调出全文吗？”，系统应该意识到“这个数据库”指的是上文提及的某数据库。这要求系统具备**对话情境记忆**，不然会把第二问当作孤立问题回答不好。大型语言模型令这一能力显著增强[[178]](#footnote-178) 。当前一些图书馆已上线基于GPT的智能助手，可以连续对话解答读者问题，并根据对话进展调整措辞和提供进一步帮助。例如OpenAI模型的上下文关联能力被用于构建馆员对话机器人，可让用户就一个主题逐步深入提问而系统保持连贯。
* **知识库的情境调用：**许多智能问答系统背后有庞大的知识库（FAQ库、知识图谱等）。场景识别可以优化知识库的检索利用。例如，在企业知识库问答中，有学者提出根据提问情境动态选择子知识库以提高准确率。同理，图书馆可能有多个知识库（文献信息、读者服务、规章制度等），情境感知可帮助**路由问题**到最相关的知识源。例如检测到问题涉及馆藏资源，就在馆藏FAQ中查找；涉及图书馆规定，则在规章库中查找。这避免了全库搜索的干扰，提高问答精度。

目前，一些**图书馆智能问答系统**（如ChatLibrary等）开始体现情境元素，如能根据用户提问语言自动切换中英文回答，根据提问类别提供相应风格的答复等[[179]](#footnote-179) 。展望未来，更深入的情境融合将让图书馆智能问答达到“类人”的参考咨询效果。当用户评价这些系统“就像在和真人馆员聊天”时，正是情境感知成功运作的体现[[180]](#footnote-180) 。不过，需要注意的是，问答系统的内容质量和可信度也必须有保证。引入情境后系统变得更复杂，可能出现新的错误类型（例如情境判断失误导致答非所问）。因此在利用情境定制答案的同时，仍需建立**人工审阅**或**用户校正**机制，以确保最终答案准确可靠。这也是图书馆专业精神在智能时代的坚守。

### 4.4.4 个性化推荐：情境感知的资源推介

**个性化推荐**服务旨在根据用户的兴趣和行为记录，向其推荐可能感兴趣的书籍、文章、数据库等资源。在图书馆中，推荐系统常用于OPAC的“借阅此书的读者还借了…”或数字图书馆的关联文献推荐等。场景识别的融合使推荐从静态的“千人一面”走向动态的“**因情境而异**”，主要表现在：

* **Context-Aware 推荐模型：**推荐系统领域早已提出**情境感知推荐系统(CARS)**的概念，即在传统用户-项目矩阵基础上加入情境维度[[181]](#footnote-181) 。其核心思想是：考虑用户在不同情境下对同一资源的喜好可能不同，通过将时间、地点、心情、目的等上下文纳入，生成更相关的推荐[[182]](#footnote-182) 。在图书馆，一个读者平时可能喜欢阅读专业论文，但周末休闲时更倾向借阅小说。如果推荐能识别出当前是周末非工作情境，就可以调整策略推荐轻松读物，而非像工作日那样推专业书。这种做法提高了推荐的接受度。情境感知还能避免不合时宜的推荐，例如凌晨使用移动端的用户，多半不方便读长篇PDF，可推荐短文章或音频资源。通过算法上将情境作为额外约束，推荐结果将对当下场景更友好。研究表明，引入上下文信息的推荐系统能够产生**更有针对性的推荐结果**，提高用户满意度[[183]](#footnote-183) 。
* **实时场景的短期偏好捉取：**传统推荐多基于长期历史偏好，而场景识别可以捕捉用户**短期的瞬时兴趣**并纳入推荐。例如，一个平时借阅历史学的读者某天突然检索了多篇人工智能论文，这可能表示其近期对AI有需求（比如跨学科研究）。情境感知系统可识别出这一“当前兴趣”场景，及时在推荐列表中加入AI领域的热门文献，而不被其长期历史偏好所完全束缚。同样地，季节和事件也是情境：毕业季时，很多读者关注就业和论文写作，推荐系统应顺应这个集体情境多推相关资源。通过情境信号，系统能够对**用户动态兴趣**做出快速反应，令推荐更加鲜活贴切。
* **多源情境提升冷启动：**图书馆常遇到新读者（无历史数据）或老读者涉足新领域的情形，传统推荐难以奏效。情境感知可利用其他信息缓解“冷启动”问题。例如，新读者的专业和年级本身就是强情境，可据此推荐该专业热门教材、基础读物等（因为大多数处于这个学业阶段的人都有相似需求）。又如当读者开始一个新课题，可以根据其检索和浏览行为情境，迅速构建该课题的知识关系图，从而推荐项目相关文献。甚至用户所处的社群情境（比如与其相似背景的用户群阅读趋势）也可用作间接依据。场景识别提供了**额外的侧信息**，帮助推荐系统在缺乏直接偏好数据时仍能给出较合理的推介。实证研究显示，融合情境因素的模型在冷启动场景下效果优于纯协同过滤模型[[184]](#footnote-184) 。
* **Library Use Case实例：**奥卢市图书馆的UbiLibrary项目中，书籍推荐模块正是结合了用户的性别和年龄情境进行过滤[[185]](#footnote-185) 。结果表明，不同年龄、性别组对推荐书目接受度有所差异，通过情境调节后推荐更受欢迎[[186]](#footnote-186) 。另一个例子是许多图书馆移动App在用户进入馆内Wi-Fi范围后，会在首页推荐馆内新到图书或当前热门借阅书籍，这也是利用位置情境提升推荐相关性的一种方式。此外，部分高校图书馆的系统会根据学科情境做推荐：理工科读者登录后首页展示最新SCI论文推荐，人文学科读者则展示人文社科领域的新书通知。这些都体现了**“推荐因场景而变”**的理念。

当然，实现情境感知的个性化推荐也要注意避免“信息茧房”和**过度个性化**。一方面，过于依赖情境可能忽略用户潜在的跨情境兴趣，错失多样性；另一方面，需要在推荐解释上向用户传达为何推荐，以增加信任（比如注明“根据您的当前位置，我们推荐附近馆藏的以下书籍”）。总体而言，场景识别赋予推荐系统**更高的上下文敏感度**和**服务温度**——用户会感觉推荐“恰是时候”地出现。这将有助于图书馆更有效地引导读者发现资源，提升资源利用率和用户粘性，真正实现知识服务的**“千人千面”**[[187]](#footnote-187) [[188]](#footnote-188) 。

### 4.4.5 知识推送：场景驱动的主动信息服务

**知识推送**指图书馆主动将有价值的信息内容发送给用户的服务模式，例如新书通报、学科动态提醒、定题服务等。传统推送多是周期性或根据预设兴趣进行，场景识别的引入则可以实现**更精细、更实时的触发式推送**[[189]](#footnote-189) 。融合的契合点包括：

* **基于情境事件的触发机制：**情境感知允许知识推送采用**“Event-Condition-Action (ECA)”**规则，即一旦检测到特定情境事件发生，就触发相应推送[[190]](#footnote-190) 。例如，当识别到用户进入图书馆某专题阅览室时，立即向其推送该专题最新资源指南；当检测到用户首次使用某数据库后，推送该数据库使用攻略或培训信息（事件=首次使用情境）。这类似旅游领域的情境信息推送服务，在检测到游客到达某景点时自动推送讲解[[191]](#footnote-191) 。图书馆可以设置许多情境触发，如\_时间事件\_（每晚10点向仍在学习的用户推送休息提醒或相关轻松读物推荐）、*位置事件*（用户经过新书架时推送新书介绍）、*行为事件*（检索多次未果时推送寻求参考咨询帮助的提示）等。通过场景识别实现这些ECA规则，让知识服务更加主动及时。
* **推送内容的情境相关性：**除了触发，更重要的是确保推送内容与用户当下情境紧密相关，否则容易被视为垃圾信息。情境感知可帮助**精准匹配推送内容**。例如，系统得知某用户正在撰写毕业论文（可根据其近来大量下载文献且使用文献管理软件的情境推断），那么推送的内容应聚焦于论文写作辅导、学术规范提醒等[[192]](#footnote-192) 。又如深夜使用电子资源的用户可能需要的是在线服务支持而非馆内活动通知。情境信息可以为推送内容打标签，只有当标签与用户场景吻合时才发送推送。这提高了推送的**命中率**和**价值**。近期有技术实现了利用情境规则过滤推送信息的尝试，如针对旅游者的系统会依据游客当前活动（开车/步行等）调整推送的信息量和形式[[193]](#footnote-193) 。类比地，图书馆可对不同场景设定不同推送策略：学习场景推送学习资源，娱乐场景推送休闲阅读等等。
* **交互与反馈闭环：**场景识别还可以完善推送服务的反馈机制。当用户接收到推送后，其后续行为（如是否点击、是否忽略）又形成新的情境信号，系统可以学习这一反馈，优化日后的推送策略。例如某用户经常无视凌晨推送，则系统应减少他深夜场景下的推送频率，或调整内容以更符合其潜在需求。这样形成**感知-推送-再感知-调整**的闭环，不断提高推送服务的有效性。

当前许多图书馆已经有一定的知识推送服务，例如新书通报邮件、微信服务号定阅推送等。但这些推送往往是批量同质的，没有考虑个体情境差异，因而信息噪音较大。应用场景识别后，推送将朝着**精准化**和**即时化**方向演进。例如，“选择性推送”代替“全体群发”，只把医学新书通知推送给医学领域相关的读者；“即时推送”代替“定时推送”，根据情境需要在恰当时机送达，而不是一刀切的固定时间。需要强调的是，虽然情境驱动推送威力巨大，但也要有**频度控制和尊重用户意愿**的机制，否则可能引起用户反感甚至选择退订。这就要求系统能识别过度推送的负面情境（如用户长时间未响应任何推送），从而自动降低推送频率，或提供简便的推送偏好设置供用户调整。只要把握好度，场景识别赋能的知识推送将成为图书馆服务的一大利器，将**合适的知识在合适的场景主动送达**给需要的读者，真正实现“知识的及时雨”。

### 4.4.6 可视化服务：情境感知的交互呈现

**可视化服务**在图书馆中指利用信息可视化技术来呈现知识、数据或服务界面，以帮助用户更直观地理解和使用信息。场景识别可以与可视化手段相结合，实现**情境自适应的界面与信息呈现**：

* **界面布局随情境自适应：**根据用户的设备、环境亮度、使用偏好等情境，动态调整图书馆系统的界面显示，是提升可用性的重要途径。比如，当检测到用户使用手机且带宽较低时，系统可视化界面应切换到简洁模式，减少高分辨率图片展示，突出核心检索框和结果列表；反之在桌面宽屏上，则可以展示更多图形化内容如知识图谱可视化或数据分析图表。再如，夜间使用时，可自动切换界面为暗色模式以保护视力，这也是根据时间情境进行可视化调整的一例。情境感知还能针对特殊人群优化界面，比如识别出用户可能有色觉障碍，则调整配色方案以提高对比度。这些**无障碍和自适应设计**都属于场景识别和可视化融合的应用，让每个用户都能在其情境下获得最佳视觉体验。
* **情境驱动的信息可视化内容：**图书馆拥有丰富的数据（借阅量、浏览趋势、学科热点等），通过可视化可以揭示模式、提供洞察。而场景识别可以帮助确定在何时何地向何人展示哪些可视化内容。例如，在学科馆员讨论会上（情境：馆员群体、场合专业讨论），系统可以投影显示近期馆藏利用统计图、读者满意度调查图等专业数据；而在普通读者进入大厅时（情境：读者群体、场合欢迎），电子屏幕上显示的可视化内容应是动态的新书推荐封面墙或活动宣传海报。这体现了**可视化内容的情境适配**。又如读者个人使用时，如果识别出其在进行科研选题，可以提供某学科知识领域图谱的交互可视化界面供其探索[[194]](#footnote-194) ；而如果用户只是随便浏览小说，则不需要复杂知识图，可简单以封面瀑布流方式可视化展示热门小说。IBM等公司曾提出“**情境可视化**”概念，强调根据用户任务阶段实时更新可视化呈现。图书馆完全可以借鉴，将可视化服务做得更加敏捷智能。
* **环境和空间的可视化互动：**随着AR/VR技术的发展，图书馆开始探索混合现实的服务，如AR导览、VR虚拟展厅等。这些应用极大依赖情境感知。例如AR导览需要知道用户所在位置和朝向，然后在其视野中叠加方向指引或书架信息。已有智慧图书馆案例中，当读者携带PDA靠近某书架，系统可提供该区域藏书的地图和导引，这是融合了定位情境的服务cflms.lib.sjtu.edu.cn。在这种场景下，可视化内容（地图、箭头）仅在用户需要时才出现，并且与物理空间对齐cflms.lib.sjtu.edu.cn。情境识别确保AR信息的正确时间和地点呈现。另一方面，VR展厅则可以根据用户交互行为情境调整场景，比如用户多次凝视某件展品，可自动弹出该展品的背景知识图谱以可视化形式呈现，供其深入了解。可以预见，未来智慧图书馆的物理与数字空间将高度融合，环境本身成为交互界面，而情境感知就是驱动这种**空间可视化交互**的关键，使之做到“你来即显，你走即隐”，增强用户的沉浸体验。

简言之，可视化服务与场景识别的融合使图书馆界面从静态走向动态，从“一视同仁”走向“因境制宜”。它让系统界面和展示内容像水一样适应容器，适应环境。**信息可视化本身是一种增强认知的手段，加入情境因素后，其认知辅助作用将更为显著**[[195]](#footnote-195) 。当然，实现情境自适应可视化也面临挑战，如需要设计多种界面模板、高效的实时渲染等。不过其收益显而易见：用户将更舒适高效地获取信息，图书馆的数据价值得以更充分地传达。例如，以前静态的数据报告可能无人问津，但现在通过一个根据读者兴趣实时变化的可视化仪表板，大家随时可以了解图书馆动态。可以说，情境感知为图书馆的信息可视化注入了“生命力”，让其随用户的脚步和需要而舞动起来，为知识服务增色不少。

## 4.5 总结

场景识别（情境感知）作为人工智能时代的重要技术，在图书馆知识服务领域展现出广阔的应用前景。本章通过对场景识别的定义、发展历程、趋势瓶颈和六大典型应用场景的深入探讨，可以看到情境感知技术正在引领图书馆从传统服务范式向智慧服务范式演进。情境感知让图书馆系统具备了一定程度的“**洞察力**”与“**适应力**”：能够洞察用户所处的具体情境与隐含需求，并动态适应服务内容与方式，使知识供给更加贴合用户实际[[196]](#footnote-196) 。

从历史演进看，情境感知理念从计算机领域萌芽，经由数字图书馆的融合逐步渗透进图书馆服务。早期的理论和模型为我们奠定了基础，近期的发展和实践又为情境感知在图书馆落地扫清障碍。当前，移动互联和物联网技术提供了丰富的数据源，深度学习和知识图谱提升了情境理解的深度，生成式AI拓展了情境交互的边界。这些进步共同推动场景识别成为智慧图书馆的重要支柱之一。然而，我们也必须清醒认识到，技术瓶颈和应用挑战依然存在，包括数据整合难题、模型准确率和实时性能、用户隐私保护等方面。只有通过持续的研发投入、跨学科合作和实践反馈迭代，才能不断完善情境感知服务的可靠性和用户体验。

六大典型应用场景的分析表明，场景识别与**知识图谱、语义检索、智能问答、个性化推荐、知识推送、可视化服务**等技术模块均有高度融合的契合点，融合后能够产生“1+1>2”的效应。例如，情境感知让知识图谱真正活起来，为不同用户呈现恰如其分的知识网络；让语义检索读懂弦外之音，返回更符合用户意图的结果；让智能问答有了类人人情味，回答因人而异；让推荐服务更懂场合，推送恰逢其时；让信息推送变得精准不扰民，做到润物细无声；也让可视化界面随环境变化而优雅自适应。这一系列融合的成果，最终指向一个共同的目标：**以用户为中心、以知识为本体的智慧服务**。

可以预见，在未来的高水平图书馆中，场景识别技术将无处不在地融入服务链条。当读者跨入图书馆的一刻，系统已识别其身份和意图，定制化的知识导航悄然展开；当研究者在线检索文献，智能助手已根据其课题背景筛选优化结果；当管理者决策馆藏发展，数据仪表板实时地根据场景展示关键指标。一切服务将更加**主动、精准、人性化**，而这背后正是情境感知的强大支撑。诚然，实现这一愿景仍有大量工作要做，但趋势已不可逆转。正如Noh在其研究中所指出的，引入情境感知的下一代数字图书馆将能为用户提供**最佳可能的服务**，极大提升图书馆服务的便利性和效能[[197]](#footnote-197) 。

# 5. 增强现实（AR）与混合现实（MR）

## 5.1 定义：AR与MR的概念、原理和界定标准

**增强现实（Augmented Reality, AR）**通常是指一种将虚拟信息叠加到真实世界的技术，使用户能够同时感知现实环境和叠加其上的数字内容[[198]](#footnote-198)

* **虚实结合**：将计算机生成的虚拟对象与现实场景融合同步呈现。
* **实时交互**：系统能在用户操作时实时更新，实现对虚拟内容的即时响应。
* **三维配准**：虚拟对象与真实环境中的位置和姿态精确对齐（注册）在三维空间中。

这一定义避免将AR限制于特定设备（如头戴显示器），强调的是功能特征而非硬件形式[[199]](#footnote-199) [[200]](#footnote-200)

**混合现实（Mixed Reality, MR）**概念最早由Milgram和Kishino在1994年提出，用于描述介于完全现实和完全虚拟之间的连续统[[201]](#footnote-201) [[202]](#footnote-202) [[203]](#footnote-203) [[204]](#footnote-204)

需要注意的是，AR与MR概念存在交叉和演变。许多文献曾将AR与MR近似同义使用[[205]](#footnote-205) [[206]](#footnote-206)

**核心原理方面**，AR/MR系统一般包含以下关键技术：1）**显示技术**：如光学透视式头戴显示器（透过式眼镜）或视频透视技术，将虚拟影像叠加到用户视野[[207]](#footnote-207) [[208]](#footnote-208) [[209]](#footnote-209)

## 5.2 起源与关键发展阶段：AR与MR的发展历程

增强现实与混合现实技术的发展可以追溯数十年，其间经历了从概念萌芽、实验室原型到商业应用的多个阶段，涌现出许多重要**里程碑事件**和**代表性产品**。下面按照时间顺序梳理AR/MR发展的关键阶段和突破：

* **1968年：基础奠定** – Ivan Sutherland开发了第一台头戴式显示器（HMD）原型“Sword of Damocles”，可在用户视野中显示简单的计算机图形，被视为AR/VR的开端[en.wikipedia.org](https://en.wikipedia.org/wiki/Augmented\_reality](h](https://www.frontiersin.org/journals/virtual-reality/articles/10.3389/frvir.2021.578080/full
* **1990年：术语诞生** – 美国波音公司的研究员Thomas Caudell首次提出“Augmented Reality”概念，用于描述一种利用头戴式显示器辅助飞机装配工人工作的方法 [[210]](#footnote-210)
* **1992年：首批AR系统** – 路易斯·罗森伯格（Louis Rosenberg）在美国空军研究实验室开发出**Virtual Fixtures**系统，利用机械臂和立体显示将虚拟指导信息叠加到操作任务中，被认为是第一个功能完善的AR原型[[211]](#footnote-211)
* **1994年：混合现实概念** – Paul Milgram发表经典论文《现实-虚拟连续统的显示分类》，系统提出了**Mixed Reality（混合现实）**的概念框架，将AR和**增强虚拟**（Augmented Virtuality）统一在混合现实范畴内[[212]](#footnote-212)
* **后1990年代：AR技术奠基** – 这一时期的学术界和工业界相继攻克了AR的关键技术难题，包括**跟踪注册**（如UMass 1995年用计算机视觉改进物体跟踪[[213]](#footnote-213) [[214]](#footnote-214) [[215]](#footnote-215)
* **2000年代初：军工与企业应用** – 2000年左右，AR开始在工业和军事领域试点应用。例如2000年，美国海军研究实验室启动了“战场增强现实系统 (BARS)”项目，研制可穿戴AR系统用于士兵态势感知训练[en.wikipedia.org](https://en.wikipedia.org/wiki/Augmented\_reality](https://en.wikipedia.org/wiki/Augmented\_reality
* **2008年：移动AR兴起** – 随着智能手机的普及，AR跨入大众消费领域。**2008年10月**，全球首款商用移动AR应用**Wikitude AR Travel Guide**在Android手机（G1）上发布，它利用手机GPS和摄像头，在旅游者手机屏幕上叠加兴趣点信息[[216]](#footnote-216)
* **2012年：谷歌眼镜** – Google推出了开创性的头戴式设备**Google Glass**（谷歌眼镜），将智能手机功能集成进眼镜框架，实现基本的“眼前”信息提示。尽管Google Glass更接近于头戴HUD（抬头显示）而非严格的沉浸式AR，但它引发了大众对AR可穿戴设备的想象，也暴露出隐私等社会问题，成为AR发展历程中具有标志意义的产品。
* **2013年：AR游戏初露锋芒** – Niantic公司发布了基于地理位置的AR手游**Ingress**，玩家需要在现实地图上争夺虚拟据点[[217]](#footnote-217)
* **2015年：MR设备问世** – **微软在2015年宣布HoloLens**，这是一款完全自含的透视式头戴显示器，配备深度传感器和空间映射功能，能够将全息影像叠加于现实并进行交互[[218]](#footnote-218)
* **2016年：AR现象级应用** – **任天堂与Niantic合作推出手游《Pokémon GO》**，将AR与LBS（基于位置服务）结合，在全球引发空前热潮[[219]](#footnote-219)
* **2017-2018年：生态完善** – 主要移动平台相继推出原生AR开发框架：**苹果发布ARKit (2017)**，**谷歌推出ARCore (2018)**，为iOS和Android设备提供了强大的SLAM（即时定位与地图构建）跟踪和环境理解功能。这些SDK降低了AR应用开发门槛，催生了大量移动AR应用。硬件方面，2018年创业公司Magic Leap发布了**Magic Leap One**头显，这是继HoloLens之后又一款备受关注的MR眼镜[[220]](#footnote-220)
* **2019年：第二代MR头显** – 微软发布**HoloLens 2**，改进了视场角和佩戴舒适度等问题[[221]](#footnote-221)
* **2020-2021年：元宇宙热潮** – 随着“元宇宙”（Metaverse）概念兴起，AR/MR重新成为科技投资热点。Facebook（后更名Meta）等公司投入巨资开发AR/MR技术，认为其将是未来元宇宙的重要接口。2021年前后，Niantic等提出了“AR云”理念，即构建全球共享的AR坐标和内容空间，使多人设备可以协同感知同一虚拟对象。这些趋势推动AR从单机应用向**大规模联网、多人交互**方向发展。
* **2022年：政策与行业推动** – 中国政府工信部等五部门发布《虚拟现实与行业应用融合发展行动计划（2022—2026年）》，将**增强现实、混合现实**纳入虚拟现实产业重点，提出到2026年实现关键技术突破、终端产品丰富和规模化应用的目标[[222]](#footnote-222)
* **2023年：消费级MR设备登场** – 元宇宙概念进一步落实到产品：Meta发布了支持彩色**透视AR的Quest 3**混合现实VR头显，同年**苹果发布Apple Vision Pro**，定位为高端MR头显[[223]](#footnote-223)

综上，增强现实技术从概念提出到如今，已经走过了半个多世纪的演进。**里程碑式事件**如AR术语的提出（1990年）、Virtual Fixtures系统（1992年）、ARToolKit开源发布（1999年）、移动AR应用问世（2008年）、HoloLens发布（2015年）、Pokemon GO爆红（2016年）等，每一步都推动AR/MR从科幻走向现实[[224]](#footnote-224) [[225]](#footnote-225)

## 5.3 演进趋势与技术瓶颈：现状分析与未来展望

### 5.3.1 AR/MR发展的现状与态势

当前，AR和MR技术正处于**快速发展与产业布局期**。根据最新市场预测，全球AR/VR行业收入将从2022年的约138亿美元增长到2026年的约509亿美元，年均复合增长率超过30%[[226]](#footnote-226)

从消费市场看，智能手机依然是AR应用最主要的平台，数以亿计的智能终端已经支持AR功能（通过ARKit/ARCore等）[[227]](#footnote-227)

产业方面，各大科技公司纷纷加码AR/MR。**苹果**于2023年发布Vision Pro标志其正式进军MR领域，后续可能推出面向大众的轻量级AR眼镜；**Meta**和**微软**则持续优化各自的MR硬件和生态（如Meta Reality系列、HoloLens系列）。**中国**也出现了多家AR硬件创业公司，诸如亮风台、蚂可、雷鸟创新等，推出面向工业和消费者的AR眼镜产品。此外，内容生态正在丰富：越来越多开发者和内容创作者利用Unity、Unreal等平台为AR/MR开发应用，涵盖游戏、教育、导览、社交媒体等各个类别。这种**软硬件生态的完善**为AR/MR大规模应用奠定了基础。

### 5.3.2 未来演进趋势展望

**（1）硬件演进：轻量化与高性能并重。**未来AR/MR设备将朝着更轻便时尚的形态发展，实现从目前笨重的头显向普通眼镜外观的过渡[frontiersin.org](https://www.frontiersin.org/journals/virtual-reality/articles/10.3389/frvir.2021.578080/full](https://www.frontiersin.org/journals/virtual-reality/articles/10.3389/frvir.2021.578080/full](https://www.frontiersin.org/journals/virtual-reality/articles/10.3389/frvir.2021.578080/full](https://www.rinf.tech/mixed-reality-use-cases-and-challenges-in-2024/](https://www.rinf.tech/mixed-reality-use-cases-and-challenges-in-2024/](https://ronaldazuma.com/papers/Presence\_AR\_challenge.pdf](https://ronaldazuma.com/papers/Presence\_AR\_challenge.pdf

**（2）软件与内容：AR云与AI赋能。**未来AR/MR体验的重要方向是**环境智能化**，即设备对真实世界具有深度的理解和记忆能力。这离不开人工智能（AI）和语义技术的融合。目前AR应用需要持续采集用户周围环境的各种数据（图像、深度、语音等），并借助复杂算法加以理解[radiantvisionsystems.com](https://www.radiantvisionsystems.com/blog/future-looks-bright-ar/vr/mr-2023-beyond](https://www.rinf.tech/mixed-reality-use-cases-and-challenges-in-2024/](https://www.sciencedirect.com/science/article/pii/S2468502X20300012](https://www.rinf.tech/mixed-reality-use-cases-and-challenges-in-2024/](https://daolan.info/?p=1228](https://ronaldazuma.com/papers/Presence\_AR\_challenge.pdf

**（3）应用场景拓展：从专用到大众。**当前AR/MR已在游戏娱乐和营销领域率先取得成功，未来将进一步向**专业和日常**领域扩展。一方面，在**工业制造、医疗、教育培训**等专业领域，AR/MR因其直观可视化和解放双手的优势，将发挥更大作用，如远程专家指导、手术辅助、虚拟装配验证等[rinf.tech](https://www.rinf.tech/mixed-reality-use-cases-and-challenges-in-2024/](https://www.rinf.tech/mixed-reality-use-cases-and-challenges-in-2024/](https://www.rinf.tech/mixed-reality-use-cases-and-challenges-in-2024/](https://www.rinf.tech/mixed-reality-use-cases-and-challenges-in-2024/

### 5.3.3 技术瓶颈与挑战

尽管前景诱人，AR和MR技术目前仍面临诸多**技术瓶颈**和**难题**亟待解决[rinf.tech](https://www.rinf.tech/mixed-reality-use-cases-and-challenges-in-2024/](https://www.rinf.tech/mixed-reality-use-cases-and-challenges-in-2024/

* **硬件限制**：现有AR/MR头显设备普遍存在**视场角（FOV）小、分辨率有限、体积笨重**等问题[frontiersin.org](https://www.frontiersin.org/journals/virtual-reality/articles/10.3389/frvir.2021.578080/full](https://www.rinf.tech/mixed-reality-use-cases-and-challenges-in-2024/](https://www.rinf.tech/mixed-reality-use-cases-and-challenges-in-2024/](https://www.rinf.tech/mixed-reality-use-cases-and-challenges-in-2024/](https://www.frontiersin.org/journals/virtual-reality/articles/10.3389/frvir.2021.578080/full](https://www.rinf.tech/mixed-reality-use-cases-and-challenges-in-2024/](https://www.rinf.tech/mixed-reality-use-cases-and-challenges-in-2024/
* **功耗与续航**：AR/MR设备的能耗高、续航短也是实际应用的掣肘[rinf.tech](https://www.rinf.tech/mixed-reality-use-cases-and-challenges-in-2024/](https://www.rinf.tech/mixed-reality-use-cases-and-challenges-in-2024/](https://www.rinf.tech/mixed-reality-use-cases-and-challenges-in-2024/
* **跟踪与交互精度**：高精度、鲁棒的**位置追踪**和**交互控制**始终是AR系统的关键技术挑战[frontiersin.org](https://www.frontiersin.org/journals/virtual-reality/articles/10.3389/frvir.2021.578080/full](https://www.frontiersin.org/journals/virtual-reality/articles/10.3389/frvir.2021.578080/full](https://ronaldazuma.com/papers/Presence\_AR\_challenge.pdf](https://ronaldazuma.com/papers/Presence\_AR\_challenge.pdf](https://www.frontiersin.org/journals/virtual-reality/articles/10.3389/frvir.2021.578080/full](https://www.frontiersin.org/journals/virtual-reality/articles/10.3389/frvir.2021.578080/full](https://www.frontiersin.org/journals/virtual-reality/articles/10.3389/frvir.2021.578080/full
* **内容制作与生态**：AR/MR的价值很大程度上依赖于**丰富的应用内容**，但目前内容生态仍显不足[rinf.tech](https://www.rinf.tech/mixed-reality-use-cases-and-challenges-in-2024/](https://www.rinf.tech/mixed-reality-use-cases-and-challenges-in-2024/](https://www.rinf.tech/mixed-reality-use-cases-and-challenges-in-2024/
* **标准化与兼容性**：由于AR/MR产业仍处早期，各家公司生态相对封闭，缺乏统一的**技术标准和协议**。这导致不同设备和平台之间内容互通困难，用户体验不连贯[rinf.tech](https://www.rinf.tech/mixed-reality-use-cases-and-challenges-in-2024/](https://www.rinf.tech/mixed-reality-use-cases-and-challenges-in-2024/
* **用户体验与接受度**：技术之外，**用户层面的挑战**也不可忽视。首先是**舒适度**和**安全性**，长时间佩戴头显可能引起眩晕、视觉疲劳等（尤其光学透视AR容易造成焦距调节冲突问题）[frontiersin.org](https://www.frontiersin.org/journals/virtual-reality/articles/10.3389/frvir.2021.578080/full](https://www.frontiersin.org/journals/virtual-reality/articles/10.3389/frvir.2021.578080/full](https://www.frontiersin.org/journals/virtual-reality/articles/10.3389/frvir.2021.578080/full](https://www.frontiersin.org/journals/virtual-reality/articles/10.3389/frvir.2021.578080/full
* **隐私与伦理**：AR/MR涉及对现实世界信息的采集和呈现，带来了新的**隐私、安全和伦理**问题[rinf.tech](https://www.rinf.tech/mixed-reality-use-cases-and-challenges-in-2024/](https://www.frontiersin.org/journals/virtual-reality/articles/10.3389/frvir.2021.578080/full](https://www.frontiersin.org/journals/virtual-reality/articles/10.3389/frvir.2021.578080/full](https://www.frontiersin.org/journals/virtual-reality/articles/10.3389/frvir.2021.578080/full](https://www.frontiersin.org/journals/virtual-reality/articles/10.3389/frvir.2021.578080/full](https://www.frontiersin.org/journals/virtual-reality/articles/10.3389/frvir.2021.578080/full](https://www.rinf.tech/mixed-reality-use-cases-and-challenges-in-2024/

概括而言，AR/MR要真正成熟落地，必须突破**“硬件—软件—内容—用户”**全链条上的瓶颈[rinf.tech](https://www.rinf.tech/mixed-reality-use-cases-and-challenges-in-2024/](https://www.rinf.tech/mixed-reality-use-cases-and-challenges-in-2024/](https://www.frontiersin.org/journals/virtual-reality/articles/10.3389/frvir.2021.578080/full

## 5.4 AR/MR与知识服务场景融合应用的可能性评估

图书馆等知识服务机构正积极探索将增强现实和混合现实技术融入其服务与环境中，以提升用户体验和知识获取效率。以下结合**六大典型场景**（知识图谱、语义搜索、智能问答、个性化推荐、知识推送、可视化服务），评估AR/MR在其中的技术适配性、现有案例、未来前景和潜在落地模式。

### 5.4.1 知识图谱场景

**知识图谱**是以语义关联构建的知识网络，用于揭示概念与实体之间的关系。在图书馆和知识服务中，知识图谱被广泛用于组织馆藏知识、辅助信息检索和推理。AR/MR技术与知识图谱融合，有望实现对**抽象知识网络的直观可视化**和交互展示。

**技术适配性：**AR/MR擅长将不可见的数字信息形象地呈现于物理空间，非常适合展示知识图谱这种复杂的关联数据[daolan.info](https://daolan.info/?p=1228

**现有案例：**目前直接针对知识图谱的AR应用案例还不多见，但相关研究和概念验证已有所探索。例如，有研究提出将**语义网本体**与AR结合，使AR应用可以访问知识图谱数据，在现实场景中提供知识性叠加信息[sciencedirect.com](https://www.sciencedirect.com/science/article/pii/S2468502X20300012](https://www.sciencedirect.com/science/article/pii/S2468502X20300012

**未来前景：**随着知识图谱在图书情报领域的深入应用，我们可以预见**AR知识图谱可视化**将成为智慧图书馆的一项创新服务。一方面，知识图谱可以作为AR内容的重要来源，为每一本书、每个主题提供结构化的背景知识。另一方面，AR提供了一个突破屏幕限制的展示维度，使复杂的知识网络以更符合人脑认知的方式呈现（如3D关系图谱、动态漫游等）。未来的智慧图书馆中，读者可能戴上AR眼镜漫步书库，只需凝视某本书的封面或书脊，书本相关的知识网络就会在眼前铺展开来，包括作者生平、相关主题分类、推荐书目等；读者可进一步点击虚拟节点查看详细信息或借阅其他相关资料。这种沉浸式的知识图谱导航将极大增强探索性学习的趣味和效率[daolan.info](https://daolan.info/?p=1228

**契合度评估：**总体而言，AR/MR与知识图谱场景**高度契合**。知识图谱解决“有什么知识、它们如何关联”的问题，AR/MR解决“如何直观呈现和交互这些知识”的问题。两者结合，可将传统上静态的知识网络转化为动态的“增强知识空间”，符合读者的认知习惯，激发探索兴趣。从技术上看，实现AR知识图谱需要解决大规模知识数据的可视化布局和交互效率问题，但这些在大数据可视化领域已有研究基础。可以预期，随着三维可视化技术和图形硬件的进步，数以千计节点规模的知识图谱都可在AR中流畅呈现。特别是MR设备能允许用户“走进”知识图谱，在空间中环顾知识网络的各个部分，无疑将把知识服务提升到一个全新的维度。

### 5.4.2 语义搜索场景

**语义搜索**旨在超越传统基于关键词匹配的检索方式，通过理解用户查询的语义和意图，更精准地找到相关信息。在图书馆的数字资源查找中，语义搜索利用本体和知识关联来提高检索的精准度和智能性。将AR/MR融入语义搜索，可以创造**更加自然直观的人机检索交互**，提升搜索效率和用户体验。

**技术适配性：**AR/MR可以为搜索提供全新的交互界面和反馈形式。传统检索是在屏幕上以文本列表呈现结果，而在AR环境下，检索过程和结果都可通过**虚实融合的方式**展现。例如，用户可以直接对着馆藏实体（如某个书架区域）提出语音或手势查询，AR系统理解其意图后，将相关结果以可视化方式在现实空间中显示——如把匹配的书籍虚拟高亮在书架上，或在用户眼前浮现与查询相关的主题词云、3D图表等。这种**“所见即所得”的语义搜索**让用户无需苦思检索词或浏览冗长的结果列表，而是通过与周围实物和AR界面的互动，自然地缩小搜索范围[daolan.info](https://daolan.info/?p=1228

**现有案例：**早期已有图书馆尝试将AR用于搜索导航。中国美术学院图书馆开发过基于AR的**图书导航系统**[daolan.info](https://daolan.info/?p=1228](https://daolan.info/?p=1228](https://daolan.info/?p=1228

**未来前景：**在未来智慧图书馆，AR/MR有望与语义搜索深度融合，实现“**所问即所见**”的知识获取体验。具体展望如下：

* **自然语言与多模态查询：**用户可以直接用语音向AR系统提问，例如“请给我推荐关于量子计算的最新论文”，AR系统将解析语义后在视野中显示一组论文的封面缩略图，甚至按主题、年份在空间中分布排列，供用户进一步筛选[daolan.info](https://daolan.info/?p=1228](https://daolan.info/?p=1228
* **可视化检索结果：**利用AR的图形渲染能力，检索结果不再局限于文本列表。用户可以选择多种可视化形式来浏览结果——比如以**知识图谱**形式展现检索到的知识点关联（结合4.1节所述的知识网络），或用图表/3D模型呈现统计信息（如检索结果按年代、类别的分布）[daolan.info](https://daolan.info/?p=1228](https://daolan.info/?p=1228
* **动态语境感知：**AR语义搜索还可以结合用户的**位置、环境和上下文**，提供契合场景的检索体验[daolan.info](https://daolan.info/?p=1228](https://daolan.info/?p=1228

**落地模式：**短期内，可通过馆内安装的AR导览终端或读者自备手机AR应用实现语义搜索。例如，读者在图书馆APP中使用AR摄像头扫描书架标签，直接对APP说出需求；APP即时在画面上标记出相关书籍的位置，或显示数字资源的链接。长远看，随着AR眼镜普及，读者戴着智能眼镜即可随时随地调用图书馆的语义搜索服务。图书馆需要确保自身OPAC系统、数字馆藏支持**语义查询接口**（如支持SPARQL查询知识库），并与AR交互模块对接。

**契合度评估：**AR/MR为语义搜索提供了**革命性的交互方式**，使搜索过程更符合人类自然的交流和认知习惯。语义技术负责理解用户意图、挖掘隐性知识，而AR负责友好地呈现和交互反馈[daolan.info](https://daolan.info/?p=1228](https://daolan.info/?p=1228

### 5.4.3 智能问答场景

**智能问答**（智能参考咨询）是图书馆为用户提供解答问题、提供信息的一项传统且重要的服务。随着AI的发展，越来越多图书馆引入了自动问答系统或聊天机器人来承担部分咨询工作。AR/MR技术可以将智能问答从二维屏幕延伸到三维空间，构建沉浸式、互动性的**虚拟咨询服务**。

**技术适配性：**AR/MR可以为智能问答提供一个**拟人化、具身化**（Embodied）的交互界面，例如以“虚拟馆员”的形象出现[daolan.info](https://daolan.info/?p=1228

**现有案例：**一些图书馆已尝试构建AR虚拟角色来辅助服务。国外有博物馆利用AR投影出历史人物的虚拟讲解员，回答游客问题。在图书馆领域，国内有研究提出“**虚拟交互式咨询空间**”概念：用户通过AR设备进入一个由虚拟馆员服务的咨询界面，系统能识别和分析用户的问题，从后台知识库检索答案即时反馈[daolan.info](https://daolan.info/?p=1228](https://daolan.info/?p=1228

**未来前景：**可以预见，未来的智慧图书馆中，**AR智能问答**将成为读者咨询服务的重要一环：

* **24小时虚拟馆员：**通过MR眼镜或手机AR，无论读者何时何地都能调出一个虚拟馆员形象获取帮助。这相当于一个随身的图书馆助手。深夜在家写论文的学生，如遇资料瓶颈，可唤出图书馆的虚拟助理寻求推荐相关文献。虚拟馆员依托AI问答引擎和图书馆知识库，能够理解口语化的问题并给出专业解答，大大突破了实体图书馆时间和空间的限制。
* **场景化咨询：**AR智能问答还能结合场景，提供所见即所问的咨询服务。例如读者在馆内看到一幅画作，不知道其背景，可以用AR设备指向画作提问“这幅画是谁画的？有什么典故？”，虚拟馆员立刻出现在画旁边，讲解作品信息并显示相关资料。这种场景感知功能让知识获取更加即时、直观，满足用户随时随地的求知欲。
* **主动推送与个性化：**未来AR馆员会越来越“聪明”，通过分析用户的查询历史和行为偏好，能够**预判用户需求**并主动提供帮助[daolan.info](https://daolan.info/?p=1228

**实现模式：**构建AR智能问答需要**AI技术与AR界面**的结合。关键组件包括：自然语言理解与生成（用于解析问题、生成回答），后端知识库/FAQ数据库（提供权威答案），以及前端AR渲染（用于显示虚拟形象和多模态内容）。目前已经比较成熟的聊天机器人系统可以作为后端基础，再开发AR前端接入。对于硬件，可以在移动App中实现基础版本，例如用户通过手机摄像头看到一个AR馆员3D模型，与之语音对话。随着MR眼镜普及，可升级为真正的全息虚拟助手悬浮在用户视野中。关键是保证**知识准确性**和**自然交互**：图书馆应维护高质量的问答知识库，AI应能引用权威出处，必要时给出文献来源[slideshare.net](https://www.slideshare.net/slideshow/smart-libraries-and-information-institutions-karen-makkeh-m1pptx/267436197

**契合度评估：**AR/MR与智能问答场景**高度匹配**，能极大丰富图书馆参考咨询服务的形式和效果。从提升**可获得性**（随时随地咨询）、**交互性**（面对面交流体验）、**主动性**（提前满足需求）等方面来看，AR虚拟馆员都比传统网页问答有优势[daolan.info](https://daolan.info/?p=1228

### 5.4.4 个性化推荐场景

**个性化推荐**是基于用户的兴趣偏好和行为历史，为其推荐可能需要的资源或服务。在数字图书馆中，常见如向读者推荐相关书籍、论文，或根据借阅历史推送新书讯等。AR/MR技术可将个性化推荐嵌入用户的现实情境，提供**情境感知的动态推荐**服务。

**技术适配性：**AR的突出优势在于能够将数字信息叠加到现实场景中。对于个性化推荐，这意味着系统可以**在恰当的时间和地点，将恰当的信息直接呈现给用户**，而无需用户主动去检索[daolan.info](https://daolan.info/?p=1228](https://daolan.info/?p=1228

**现有案例：**当前个性化推荐主要在网页/App界面上呈现（如图书馆主页的推荐列表）。但已有研究设想了AR环境下的推荐方式。如上文提到的虚拟馆员不仅用于问答，也能**分析用户行为实时推送感兴趣内容**[daolan.info](https://daolan.info/?p=1228

**未来前景：**AR/MR有望使个性化推荐达到**“润物细无声”**的效果，即推荐内容**精准**且**时机合宜**，用户接受度高：

* **环境感知推荐：**未来图书馆的AR系统会综合考虑用户所处的**物理环境**和**上下文**，提供动态推荐[daolan.info](https://daolan.info/?p=1228
* **跨场景连续推荐：**MR设备可以让推荐服务贯穿用户在馆内外的整个过程。例如用户在家通过AR眼镜阅读文献，当看到有用内容做了标记；第二天来到图书馆时，AR系统已经为他将相关主题的馆藏书籍准备好路线指引或座位上的数字推送。又或者用户在馆内借出一本书，离馆后在公交车上通过手机AR收到一段与该书内容相关的短视频推荐。这种跨场景的延续，使图书馆的服务无缝融入用户的日常学习生活。
* **深度个性化与社交推荐：**AR允许引入更多用户偏好维度，比如结合用户的**社交圈**或群体行为进行推荐。如果某用户的朋友最近都在阅读某个话题，AR眼镜或应用可以在合适的时候提醒他“您的3位好友近期都在关注XX主题，图书馆有新到的一本该主题热门书籍，您可能也感兴趣”。这种带有社交元素的推荐通过AR推送，可能比冷冰冰的系统推荐更容易引起用户共鸣，因为其形式更贴近人与人交流的感觉（AR界面可以显示好友头像或读后评价）。

**落地模式：**实现AR个性化推荐需建立**用户模型**并实时匹配内容。图书馆已积累用户借阅史、检索史等数据，可用于偏好分析；结合即时的环境感知（通过用户定位、浏览对象识别等），就能判断何时何地推荐何种内容最合适。实际落地时，可在移动端通过推送通知+AR界面的方式先试验。例如，当用户经过某书架，手机振动并在AR取景中显示“这里有一本可能对你有用的书”。待AR眼镜普及，可升级为更自动化的提示（用户视野边缘出现提示图标等）。需要注意**隐私保护**：个性化意味着对用户行为的跟踪分析，必须确保这些数据的使用透明和用户可控，AR界面上的推荐也应支持反馈机制（用户可以标记不感兴趣，避免以后类似推送）。

**契合度评估：**AR/MR可以赋能个性化推荐，使之达到**“对的内容以对的方式在对的时刻”**触达用户的理想状态[daolan.info](https://daolan.info/?p=1228

### 5.4.5 知识推送场景

**知识推送**广义上指向用户主动发送其可能需要的知识信息，包括通知、警报、更新等服务。在图书馆，典型如新书通告、预约到书提醒、学科动态推送等。AR/MR可以通过全新的感官渠道来执行知识推送，使信息传达更**直观及时**且**上下文相关**。

**技术适配性：**与个性化推荐类似，知识推送在AR/MR中不再局限于文字消息，而可以利用视觉、听觉、甚至触觉提示来传达信息。AR知识推送的一个关键优势是**上下文关联**：推送内容可以与用户当前所见的现实对象绑定在一起。这让信息获取的语境更加明确、减少认知开销。例如，当馆员在后台更新了某本图书的数字化版本，读者正好在浏览这本书的实体书架位置时，AR眼镜可以叠加一个提示在该书位置：“本书有最新数字版，可线上阅读”。读者一目了然，立刻明白推送所指，无需再从短信或邮件中费力关联。

**现有案例：**目前图书馆常用的知识推送手段包括电子邮件通知、手机APP推送等，尚未见专门AR化的实践。但在其他领域有AR推送的雏形。例如商场使用AR导航时，会在用户路过某店铺时弹出优惠券信息，算是一种情境推送。又如汽车AR抬头显示，会在驾驶员眼前弹出道路危险警告，也是将信息直接推送在环境中。移植到图书馆情境，我们可以想象：当用户走近馆内公告栏时，戴着AR眼镜会自动显示“本周新书通告”的摘要，不必驻足阅读纸质公告；或者某研讨会将在会议室召开，参会用户进入图书馆时就收到AR弹出的提醒卡片。这些都是AR用于知识推送的场景，有些馆舍已经通过数字标牌实现类似功能，但AR能让提示**以用户为中心**，更加个性化。

**未来前景：**在未来智慧知识服务中，AR/MR将使知识推送更具**实时交互性**和**环境融合感**：

* **实时学术资讯推送：**对于科研型用户，图书馆可结合AR推送最新的学科资讯和动态。例如，某用户是人工智能领域研究者，当他步入实验室或图书馆时，AR眼镜自动推送一条当天AI领域重要论文发表或会议截稿日期的信息，以弹窗或语音形式呈现。这相当于一个随身的“学术秘书”，时刻用AR提醒用户关注领域进展。
* **物理空间触发的推送：**图书馆的每个物理空间都可以与数字信息关联。用户走进某主题阅览室，AR推送该主题最新的馆藏推荐书目；在自习室久坐后，AR建议用户休息并推送馆内咖啡吧优惠信息等。这种“智慧空间”与知识推送结合，被认为是未来图书馆的重要特征[calsp.cn](https://www.calsp.cn/2023/06/20/bulletin-202306-01/](https://www.lis.ac.cn/CN/article/downloadArticleFile.do?attachType=PDF&id=23200
* **跨平台全方位推送：**MR设备有潜力整合视觉、听觉提示。例如重要通知（如借阅到期警告）时，AR眼镜不仅弹出文字，还可伴随轻声提示音或语音说明，确保用户注意到。甚至未来AR戒指等可穿戴可以提供轻微震动，形成多模态的提醒体系。在图书馆这样强调安静的场所，视觉AR提示尤其有用，避免了广播喇叭的打扰。总之，知识推送将从二维的消息框走向**环绕用户360度**的信息环境。

**落地模式：**图书馆需建立**推送规则和用户偏好管理**机制，以免过犹不及。AR推送可分层级：极重要的信息（如安全疏散警报）则强制显现在用户视野中心，其它一般通知则以小图标或角落弹窗方式呈现，由用户选择是否展开查看。这些在AR界面设计时要考虑周全。初期，可通过手机AR应用实现地理围栏式的推送（用户进出某区域触发），或者扫描式获取（用户用AR扫某二维码以订阅对应板块通知）。随着设备智能化，可转为后台自动触发。一个关键是**精准投递**：利用大数据分析，避免向不需要的人推送无关信息。例如儿童读者不会收到学术讲座通知，而研究人员不被打扰推送少儿活动讯息。AR推送结合用户画像，可实现**定向播报**而非“大喇叭”。

**契合度评估：**知识推送与AR/MR融合能显著提升信息分发的**效率**和**用户体验**。相比传统邮件公告，AR推送的信息**位置感**和**即时性**更强——它在用户眼前出现，就发生在用户所在情境，这本身就是一种筛选，保证了相关性。对于知识服务机构而言，这是提供**“精准知识供给”**的重要手段，有助于防止信息遗漏和提高服务触达率。当然也要慎防信息过载，一旦AR推送过于频繁，用户可能视而不见甚至反感。因此需要借助语义和AI制定推送策略，确保**少而精**。综合来看，这一场景与AR/MR的技术能力高度吻合，也是智慧图书馆建设的一个亮点方向，能让读者真正感受到“处处有知识,时时受启发”的氛围。

### 5.4.6 可视化服务场景

**可视化服务**指利用图形化方式呈现数据和信息，帮助用户更直观地理解知识内容。在图书馆，典型的可视化包括数据仪表盘（如馆藏利用率图）、知识地图、学科发展图谱等。AR/MR为可视化服务提供了一个**三维、沉浸**的新媒体，可以突破屏幕，将信息直接呈现在用户周围环境中。

**技术适配性：**AR/MR在可视化方面具有两个显著优势：**空间维度**和**交互维度**。一方面，传统可视化受屏幕尺寸限制，而AR可利用整个现实空间来排列和展现信息。例如，在MR眼镜中，用户查询某主题的发展史，可以看到时间轴沿着房间地板延伸，每个节点上竖立着对应年份的代表文献封面3D模型，走近某节点时，还能弹出详览。这种空间化展示比平面图表更具沉浸感和可探索性。另一方面，AR可视化是可交互的：用户可以移动、缩放可视化内容，甚至用手势“抓取”某一数据点以获得细节，或通过注视一个元素触发关联变化。这使得可视化从被动浏览变为主动探索，更契合**探索式数据分析**的需求。

**现有案例：**许多研究已将AR用于科学数据可视化和教育。如Nature等报道过利用AR展示复杂分子结构，学生戴AR设备可以从各个角度观察分子3D模型，加强理解。同理，图书馆领域有尝试用AR来可视化馆藏分布或利用情况。例如通过AR应用在地图上重现图书借阅热力图：用户站在大厅中央，用AR看到整个楼层的借阅密度以不同颜色覆盖在各区域上，红色表示热门区。这给馆员直观的空间统计认识。另一个例子是知识关联的3D可视化：早年有项目使用AR投影出“三维知识树”，让用户用手柄选择展开哪个分支，对应地墙上投射出该主题的文献列表。虽然这些案例还比较基础，但展现了AR在增强可视化理解上的潜力。

**未来前景：**AR/MR可视化服务有广阔的前景：

* **复杂数据的沉浸分析：**未来图书馆积累的大数据（如海量学术文献、用户行为日志等）可以通过MR进行可视分析。决策者可以走进一个“数据作战室”，四周墙面和桌面都是AR投影的数据可视化：某面墙显示本馆各学科文献增长曲线，另一面墙显示读者流量热图，桌面上悬浮着预算分配3D饼图。管理者可通过手势点击某年份曲线，所有相关图表动态联动更新。这种沉浸式可视化有助于多维数据关联洞察，比看报告或屏幕切换高效得多。
* **知识结构可视化：**对于读者，很多深奥知识如果以3D形式展现会更易理解。例如历史学科的谱系，如果用MR在空中展示一棵3D树状图（时间为纵轴，地域为横轴，人物事件为节点），读者可以“走”进这棵知识树细看细枝末节。这比平面年表更生动。再如文献引用网络，用AR展示成节点-连线的三维网络，重要论文在中心高亮，引用关系如光带连接，读者可以拖拽节点查看邻近网络结构。这其实就是知识图谱的AR可视化，是4.1场景的具体延伸应用。
* **教育培训可视化：**图书馆常有信息素养教育任务。AR可用于教学，比如教新生如何使用馆藏检索系统，不再停留于PPT讲解，而是让学生戴AR眼镜实际演练：眼前浮现一个虚拟OPAC终端，导师演示检索过程，同时墙上出现检索流程图和结果排名可视化。这比传统投影更具参与感，也能增加趣味性和理解力。

**落地模式：**实现AR可视化需要良好的软件支持。目前Unity等引擎已提供3D图表的制作能力，一些数据可视化库也开始支持VR/AR输出。图书馆应探索将自身数据接口（如借阅统计、馆藏元数据）与AR前端结合，开发一些**试验性应用**。起步可以从简单的如“AR馆藏导航图”“AR借阅统计投影”等入手，以增强馆员和读者对馆藏和服务的理解。未来，如有条件，可设置专门的AR可视化体验室或在馆内举办AR数据可视化展览，让用户亲身感受知识以新颖方式展现的魅力。

**契合度评估：**可视化服务是AR/MR的天然用武之地。信息可视化旨在借助人类视觉直觉来传达复杂数据，与AR将数字内容嵌入现实的理念不谋而合。AR的三维显示、互动操控和上下文融合能力可以将可视化的效果提升到**前所未有**的高度，使抽象知识“看得见，摸得着”[daolan.info](https://daolan.info/?p=1228

增强现实（AR）和混合现实（MR）技术正逐步从科幻走进现实，并展现出与知识服务领域深度融合的巨大潜力。本文通过对AR/MR的定义厘清和发展脉络回顾，分析了当前技术演进趋势与瓶颈挑战，进而结合图书馆六大典型知识服务场景，全面讨论了AR/MR的融合应用可能性与价值。可以看到，**AR/MR能够为知识服务带来更直观的感知、更智能的交互和更主动的服务**：从虚实结合的知识图谱、沉浸对话的智能问答，到情境驱动的推荐推送、三维交互的数据可视化，AR/MR有望革新传统图书馆服务模式，提升信息获取的效率和体验。

1. https://www.ibm.com/think/topics/large-language-models [↑](#footnote-ref-1)
2. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., … & Liang, P. (2021). On the Opportunities and Risks of Foundation Models. Stanford University Human-Centered Artificial Intelligence (HAI). [↑](#footnote-ref-2)
3. “What are large language models (LLMs)?”. IBM Technology. [↑](#footnote-ref-3)
4. http://arxiv.org/pdf/2303.18223 [↑](#footnote-ref-4)
5. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., … & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38. [↑](#footnote-ref-5)
6. https://papers.neurips.cc/paper/7181-attention-is-all-you-need.pdf [↑](#footnote-ref-6)
7. https://papers.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf [↑](#footnote-ref-7)
8. https://arxiv.org/html/2402.06196v3 [↑](#footnote-ref-8)
9. https://www.elastic.co/what-is/large-language-models [↑](#footnote-ref-9)
10. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf [↑](#footnote-ref-10)
11. https://research.google/blog/characterizing-emergent-phenomena-in-large-language-models/ [↑](#footnote-ref-11)
12. https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2023-gartner-hype-cycle [↑](#footnote-ref-12)
13. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier [↑](#footnote-ref-13)
14. https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2023-gartner-hype-cycle [↑](#footnote-ref-14)
15. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier [↑](#footnote-ref-15)
16. https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2023-gartner-hype-cycle [↑](#footnote-ref-16)
17. https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2023-gartner-hype-cycle [↑](#footnote-ref-17)
18. https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2023-gartner-hype-cycle [↑](#footnote-ref-18)
19. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier [↑](#footnote-ref-19)
20. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier [↑](#footnote-ref-20)
21. https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2023-gartner-hype-cycle [↑](#footnote-ref-21)
22. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier [↑](#footnote-ref-22)
23. https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2023-gartner-hype-cycle [↑](#footnote-ref-23)
24. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier [↑](#footnote-ref-24)
25. https://www.sciencedirect.com/science/article/pii/S0268401223000233 [↑](#footnote-ref-25)
26. https://www.sciencedirect.com/science/article/pii/S0268401223000233 [↑](#footnote-ref-26)
27. https://thegrizzlynews.org/2361/news/large-language-models-carry-enormous-energy-consumption-and-cost [↑](#footnote-ref-27)
28. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier [↑](#footnote-ref-28)
29. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier [↑](#footnote-ref-29)
30. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier [↑](#footnote-ref-30)
31. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier [↑](#footnote-ref-31)
32. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier [↑](#footnote-ref-32)
33. https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2023-gartner-hype-cycle [↑](#footnote-ref-33)
34. https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2023-gartner-hype-cycle [↑](#footnote-ref-34)
35. https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2023-gartner-hype-cycle [↑](#footnote-ref-35)
36. https://arxiv.org/abs/2303.08774 [↑](#footnote-ref-36)
37. https://arxiv.org/abs/2303.08774 [↑](#footnote-ref-37)
38. https://arxiv.org/abs/2303.08774 [↑](#footnote-ref-38)
39. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier [↑](#footnote-ref-39)
40. https://www.linkedin.com/pulse/top-10-insights-from-forresters-state-generative-ai-2024-columbus-dqtvc [↑](#footnote-ref-40)
41. https://www.anthropic.com/news/100k-context-windows [↑](#footnote-ref-41)
42. https://www.anthropic.com/news/100k-context-windows [↑](#footnote-ref-42)
43. https://mistral.ai/news/announcing-mistral-7b [↑](#footnote-ref-43)
44. https://mistral.ai/news/announcing-mistral-7b [↑](#footnote-ref-44)
45. https://mistral.ai/news/announcing-mistral-7b [↑](#footnote-ref-45)
46. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier [↑](#footnote-ref-46)
47. https://www.linkedin.com/pulse/top-10-insights-from-forresters-state-generative-ai-2024-columbus-dqtvc [↑](#footnote-ref-47)
48. https://mistral.ai/news/announcing-mistral-7b [↑](#footnote-ref-48)
49. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier [↑](#footnote-ref-49)
50. https://medium.com/@writerdotcom/small-language-models-the-next-big-thing-for-solo-developers-and-entrepreneurs-6dc520fb3bb8 [↑](#footnote-ref-50)
51. https://arxiv.org/abs/2303.08774 [↑](#footnote-ref-51)
52. https://news.mit.edu/2025/large-language-models-reason-about-diverse-data-general-way-0219 [↑](#footnote-ref-52)
53. https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2023-gartner-hype-cycle [↑](#footnote-ref-53)
54. https://www.anthropic.com/news/100k-context-windows [↑](#footnote-ref-54)
55. https://www.anthropic.com/news/100k-context-windows [↑](#footnote-ref-55)
56. https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2023-gartner-hype-cycle [↑](#footnote-ref-56)
57. https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2023-gartner-hype-cycle [↑](#footnote-ref-57)
58. https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2023-gartner-hype-cycle [↑](#footnote-ref-58)
59. https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2023-gartner-hype-cycle [↑](#footnote-ref-59)
60. https://www.linkedin.com/posts/mit-technology-review\_large-language-models-can-do-jaw-dropping-activity-7276879508077514752-xn6l [↑](#footnote-ref-60)
61. https://milvus.io/ai-quick-reference/how-does-retrievalaugmented-generation-help-with-the-issue-of-an-llms-static-knowledge-cutoff-or-memory-limitations [↑](#footnote-ref-61)
62. https://milvus.io/ai-quick-reference/how-does-retrievalaugmented-generation-help-with-the-issue-of-an-llms-static-knowledge-cutoff-or-memory-limitations [↑](#footnote-ref-62)
63. https://milvus.io/ai-quick-reference/how-does-retrievalaugmented-generation-help-with-the-issue-of-an-llms-static-knowledge-cutoff-or-memory-limitations [↑](#footnote-ref-63)
64. https://www.linkedin.com/pulse/top-10-insights-from-forresters-state-generative-ai-2024-columbus-dqtvc [↑](#footnote-ref-64)
65. https://www.wired.com/story/this-ai-model-never-stops-learning [↑](#footnote-ref-65)
66. https://www.linkedin.com/posts/mit-technology-review\_large-language-models-can-do-jaw-dropping-activity-7276879508077514752-xn6l [↑](#footnote-ref-66)
67. https://www.linkedin.com/pulse/top-10-insights-from-forresters-state-generative-ai-2024-columbus-dqtvc [↑](#footnote-ref-67)
68. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year [↑](#footnote-ref-68)
69. https://milvus.io/ai-quick-reference/how-does-retrievalaugmented-generation-help-with-the-issue-of-an-llms-static-knowledge-cutoff-or-memory-limitations [↑](#footnote-ref-69)
70. https://www.linkedin.com/pulse/top-10-insights-from-forresters-state-generative-ai-2024-columbus-dqtvc [↑](#footnote-ref-70)
71. https://knowledge.wharton.upenn.edu/article/the-hidden-cost-of-ai-energy-consumption [↑](#footnote-ref-71)
72. https://thegrizzlynews.org/2361/news/large-language-models-carry-enormous-energy-consumption-and-cost [↑](#footnote-ref-72)
73. https://knowledge.wharton.upenn.edu/article/the-hidden-cost-of-ai-energy-consumption [↑](#footnote-ref-73)
74. https://knowledge.wharton.upenn.edu/article/the-hidden-cost-of-ai-energy-consumption [↑](#footnote-ref-74)
75. https://knowledge.wharton.upenn.edu/article/the-hidden-cost-of-ai-energy-consumption [↑](#footnote-ref-75)
76. https://balkangreenenergynews.com/chatgpt-consumes-enough-power-in-one-year-to-charge-over-three-million-electric-cars [↑](#footnote-ref-76)
77. https://www.linkedin.com/pulse/top-10-insights-from-forresters-state-generative-ai-2024-columbus-dqtvc [↑](#footnote-ref-77)
78. https://news.engin.umich.edu/2024/11/up-to-30-of-the-power-used-to-train-ai-is-wasted-heres-how-to-fix-it [↑](#footnote-ref-78)
79. https://mistral.ai/news/announcing-mistral-7b [↑](#footnote-ref-79)
80. https://www.linkedin.com/pulse/top-10-insights-from-forresters-state-generative-ai-2024-columbus-dqtvc [↑](#footnote-ref-80)
81. https://www.linkedin.com/pulse/top-10-insights-from-forresters-state-generative-ai-2024-columbus-dqtvc [↑](#footnote-ref-81)
82. https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2023-gartner-hype-cycle [↑](#footnote-ref-82)
83. https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2023-gartner-hype-cycle [↑](#footnote-ref-83)
84. https://mistral.ai/news/announcing-mistral-7b [↑](#footnote-ref-84)
85. https://mistral.ai/news/announcing-mistral-7b [↑](#footnote-ref-85)
86. https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2023-gartner-hype-cycle [↑](#footnote-ref-86)
87. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier [↑](#footnote-ref-87)
88. https://www.linkedin.com/pulse/top-10-insights-from-forresters-state-generative-ai-2024-columbus-dqtvc [↑](#footnote-ref-88)
89. https://m.thepaper.cn/kuaibao\_detail.jsp?contid=30475035&from=kuaibao#:~:text=%E6%B5%8B%E8%83%BD%E5%8A%9B%EF%BC%8C%20 [↑](#footnote-ref-89)
90. https://view.inews.qq.com/a/20240710A02T1K00#:~:text=%E5%8D%8E%E7%9B%9B%E9%A1%BF%E5%A4%A7%E5%AD%A6%E8%AE%A1%E7%AE%97%E6%9C%BA%E7%A7%91%E5%AD%A6%E6%95%99%E6%8E%88%20Chirag%20Shah%20%E8%A1%A8%E7%A4%BA%EF%BC%8C%E2%80%9C%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BD%93%E2%80%9D%E4%B8%80%E8%AF%8D%E5%85%B6%E5%AE%9E%E5%B7%B2%E7%BB%8F%E5%AD%98%E5%9C%A8%E4%BA%86%E5%BE%88%E5%A4%9A%E5%B9%B4%EF%BC%8C%E4%BD%86%E6%98%AF%E5%85%B6%E5%9C%A8%E4%B8%8D%E5%90%8C%E7%9A%84%E6%97%B6%E9%97%B4%E6%84%8F%E5%91%B3%E7%9D%80%E4%B8%8D%E5%90%8C%E7%9A%84%E4%B8%9C%E8%A5%BF%E3%80%82 [↑](#footnote-ref-90)
91. https://whly.gd.gov.cn/service\_newggwhzx/content/post\_4061139.html#:~:text=%E8%87%AA%E5%8A%A9%E5%AF%BC%E8%88%AA%E5%AE%9A%E4%BD%8D%E6%9C%8D%E5%8A%A1%E3%80%82%E7%94%B1%E6%AD%A4%2C%E8%AF%BB%E8%80%85%E8%83%BD%E5%A4%9F%E5%9C%A8%E9%A6%86%E5%86%85%E4%B8%8D%E5%90%8C%E5%8C%BA%E5%9F%9F%E6%9F%A5%E8%AF%A2%E5%9B%BE%E4%B9%A6%2C%E9%80%9A%E8%BF%87%E5%AF%BC%E8%88%AA%E5%8F%AF%E7%9B%B4%E6%8E%A5%E6%89%BE%E5%88%B0%E5%9B%BE%E4%B9%A6%E6%9E%B6%E4%BD%8D%EF%BC%8C%E5%88%87%E5%AE%9E%E8%A7%A3%E5%86%B3%E4%BA%86%E2%80%9C%E6%89%BE%E4%B9%A6%E9%9A%BE%E2%80%9D%E8%BF%99%E4%B8%80%E5%9B%BE%E4%B9%A6%E9%A6%86%E7%BA%BF%E4%B8%8B%E6%9C%8D%E5%8A%A1%E9%9A%BE%E9%A2%98%E3%80%82 [↑](#footnote-ref-91)
92. https://www.calsp.cn/2024/08/21/bulletin-202408-05/#:~:text=%E5%8D%97%E4%BA%AC%E4%BB%B0%E6%A0%BC%E4%BF%A1%E6%81%AF%E7%A7%91%E6%8A%80%E6%9C%89%E9%99%90%E5%85%AC%E5%8F%B8%E5%9F%BA%E4%BA%8E%E4%B8%8A%E6%96%87%E7%9A%84%E6%80%9D%E8%B7%AF%EF%BC%8C%E8%AE%BE%E8%AE%A1%E7%A0%94%E5%8F%91%E4%BA%86RAG%E6%A3%80%E7%B4%A2%E5%A2%9E%E5%BC%BA%E7%94%9F%E6%88%90%E6%8A%80%E6%9C%AF%E7%9B%B8%E5%85%B3%E7%9A%84%E4%BA%A7%E5%93%81%E3%80%82%E5%AE%9E%E7%8E%B0%E4%BA%86PDF%E6%99%BA%E8%83%BD%E8%BE%85%E5%8A%A9%E9%98%85%E8%AF%BB%E5%92%8C%E5%88%86%E6%9E%90%E3%80%81%E6%95%B0%E6%8D%AE%E5%BA%93AI%E5%AF%BC%E8%88%AA%E5%8A%A9%E6%89%8B%E3%80%81%E8%B5%84%E6%BA%90%E5%8F%91%E7%8E%B0AI%E6%A3%80%E7%B4%A2%E5%8A%A9%E6%89%8B%E3%80%81%E4%B8%AA%20%E4%BA%BA%E7%9F%A5%E8%AF%86%E4%B8%AD%E5%BF%83%E7%AD%89%E5%88%9B%E6%96%B0%E5%8A%9F%E8%83%BD%E3%80%82%E7%9B%B8%E5%85%B3%E4%BA%A7%E5%93%81%E5%B7%B2%E5%9C%A8%E6%B8%85%E5%8D%8E%E5%A4%A7%E5%AD%A6%E5%9B%BE%E4%B9%A6%E9%A6%86%E3%80%81%E4%B8%8A%E6%B5%B7%E5%9B%BE%E4%B9%A6%E9%A6%86%E3%80%81%E5%8D%97%E4%BA%AC%E5%A4%A7%E5%AD%A6%E5%9B%BE%E4%B9%A6%E9%A6%86%E3%80%81%E8%A5%BF%E5%AE%89%E4%BA%A4%E9%80%9A%E5%A4%A7%E5%AD%A6%E5%9B%BE%E4%B9%A6%E9%A6%86%E7%AD%89%E5%A4%9A%E5%AE%B6%E8%A1%8C%E4%B8%9A%E6%9C%BA%E6%9E%84%E8%BF%9B%E8%A1%8C%E5%BA%94%E7%94%A8%E3%80%82 [↑](#footnote-ref-92)
93. https://m.thepaper.cn/kuaibao\_detail.jsp?contid=30475035&from=kuaibao#:~:text=%E6%B5%8B%E8%83%BD%E5%8A%9B%EF%BC%8C%20 [↑](#footnote-ref-93)
94. https://www.fjlib.net/zt/mtxh/stxk/llysj/fjstsgxk201403/201804/t20180414\_358197.htm#:~:text=%E7%9F%A5%E8%AF%86%E5%9B%BE%E8%B0%B1%28Mapping%20Knowledge%20Domain%EF%BC%89%E5%9C%A8%E5%9B%BE%E4%B9%A6%E6%83%85%E6%8A%A5%E7%95%8C%E4%B9%9F%E7%A7%B0%E4%B8%BA%E7%9F%A5%E8%AF%86%E5%9F%9F%E5%8F%AF%E8%A7%86%E5%8C%96%E6%88%96%E7%9F%A5%E8%AF%86%E9%A2%86%E5%9F%9F%E6%98%A0%E5%B0%84%E5%9C%B0%E5%9B%BE%EF%BC%8C%E6%98%AF%E6%98%BE%E7%A4%BA%E7%9F%A5%E8%AF%86%E5%8F%91%E5%B1%95%E8%BF%9B%E7%A8%8B%E4%B8%8E%E7%BB%93%E6%9E%84%E5%85%B3%E7%B3%BB%E7%9A%84%E4%B8%80%E7%B3%BB%E5%88%97%E5%90%84%E7%A7%8D%E4%B8%8D%E5%90%8C%E7%9A%84%E5%9B%BE%E5%BD%A2%20%E3%80%82%E8%AF%A6%E7%BB%86%E5%9C%B0%E8%AF%B4%EF%BC%8C%E7%9F%A5%E8%AF%86%E5%9B%BE%E8%B0%B1%E5%B0%B1%E6%98%AF%E6%8A%8A%E5%9B%BE%E5%BD%A2%E5%AD%A6%E3%80%81%E4%BF%A1%E6%81%AF%E5%8F%AF%E8%A7%86%E5%8C%96%E3%80%81%E4%BF%A1%E6%81%AF%E7%A7%91%E5%AD%A6%E7%AD%89%E5%AD%A6%E7%A7%91%E7%9A%84%E7%90%86%E8%AE%BA%E5%92%8C%E6%96%B9%E6%B3%95%E4%B8%8E%E7%A7%91%E5%AD%A6%E8%AE%A1%E9%87%8F%E5%AD%A6%E4%B8%AD%E7%9A%84%E5%BC%95%E6%96%87%E5%88%86%E6%9E%90%E3%80%81%E5%85%B1%E7%8E%B0%E5%88%86%E6%9E%90%E7%AD%89%E6%96%B9%E6%B3%95%E7%BB%93%E5%90%88%E8%B5%B7%E6%9D%A5%EF%BC%8C%E7%94%A8%E5%8F%AF%E8%A7%86%E5%8C%96%E6%8A%80%E6%9C%AF%E6%9D%A5%E5%BD%A2%E5%AE%B9%E6%8F%8F%E8%BF%B0%E7%9F%A5%E8%AF%86%E8%B5%84%E6%BA%90%E5%8F%8A%E5%85%B6%E8%81%94%E7%B3%BB%20%EF%BC%8C%E6%8C%96%E6%8E%98%E3%80%81%E5%88%86%E6%9E%90%E3%80%81%E5%BB%BA%E6%9E%84%E3%80%81%E6%B5%8B%E7%BB%98%E5%8F%8A%E6%98%BE%E7%A4%BA%E7%9F%A5%E8%AF%86%E5%8F%8A%E5%AE%83%E4%BB%AC%E4%B9%8B%E9%97%B4%E7%9A%84%E5%85%B3%E8%81%94%EF%BC%8C%E5%B1%95%E7%A4%BA%E4%B8%80%E9%97%A8%E5%AD%A6%E7%A7%91%E7%9A%84%E6%A0%B8%E5%BF%83%E7%BB%93%E6%9E%84%E3%80%81%E5%8F%91%E5%B1%95%E5%8E%86%E5%8F%B2%E3%80%81%E5%89%8D%E6%B2%BF%E7%83%AD%E7%82%B9%E4%BB%A5%E5%8F%8A%E6%95%B4%E4%BD%93%E7%9F%A5%E8%AF%86%E7%BB%93%E6%9E%84%E7%9A%84%E5%B0%86%E5%A4%9A%E7%A7%8D%E5%AD%A6%E7%A7%91%E7%9F%A5%E8%AF%86%E3%80%81%E6%8A%80%E6%9C%AF%E8%BF%9B%E8%A1%8C%E8%9E%8D%E5%90%88%E7%9A%84%E4%B8%80%E7%A7%8D%E7%A0%94%E7%A9%B6%E6%96%B9%E6%B3%95%E3%80%82, [↑](#footnote-ref-94)
95. https://www.fjlib.net/zt/mtxh/stxk/llysj/fjstsgxk201403/201804/t20180414\_358197.htm#:~:text=%E5%B9%B4%E5%9C%A8%E3%80%8A%E7%A7%91%E5%AD%A6%E3%80%8B%E6%9D%82%E5%BF%97%E4%B8%8A%E5%8F%91%E8%A1%A8%E7%9A%84%E4%B8%80%E7%AF%87%E8%AE%BA%E6%96%87%E4%B8%AD%EF%BC%8C%E5%B0%B1%E8%AE%A4%E4%B8%BA%E4%BC%A0%E7%BB%9F%E7%9A%84%E5%9F%BA%E4%BA%8E%E5%AD%A6%E7%A7%91%E7%9A%84%E6%A3%80%E7%B4%A2%E5%92%8C%E6%91%98%E8%A6%81%E6%9C%8D%E5%8A%A1%E9%9A%BE%E4%BB%A5%E6%BB%A1%E8%B6%B3%E7%A0%94%E7%A9%B6%E7%9A%84%E9%9C%80%E6%B1%82%E3%80%82 [↑](#footnote-ref-95)
96. https://library.ucr.edu/news/2022/04/29/ucr-library-implements-yewno-discover-ai-powered-research#:~:text=Artificial%20intelligence,a%20short%20amount%20of%20time [↑](#footnote-ref-96)
97. https://library.ucr.edu/news/2022/04/29/ucr-library-implements-yewno-discover-ai-powered-research#:~:text=How%20does%20Yewno%20Discover%20work%3F,the%20user%20in%20their%20research [↑](#footnote-ref-97)
98. https://library.ucr.edu/news/2022/04/29/ucr-library-implements-yewno-discover-ai-powered-research#:~:text=Once%20a%20primary%20or%20secondary,relevant%20parts%20of%20the%20text [↑](#footnote-ref-98)
99. https://library.ucr.edu/news/2022/04/29/ucr-library-implements-yewno-discover-ai-powered-research#:~:text=For%20certain%20users%2C%20what%20may,research%20session%20on%20Yewno%20Discover [↑](#footnote-ref-99)
100. https://library.ucr.edu/news/2022/04/29/ucr-library-implements-yewno-discover-ai-powered-research#:~:text=Unlike%20other%20research%20engines%2C%20Yewno,valuable%20for%20the%20more%20inquisitive [↑](#footnote-ref-100)
101. https://library.ucr.edu/news/2022/04/29/ucr-library-implements-yewno-discover-ai-powered-research#:~:text=%E2%80%9CYewno%20Discover%20is%20particularly%20useful,%E2%80%9D [↑](#footnote-ref-101)
102. https://library.ucr.edu/news/2022/04/29/ucr-library-implements-yewno-discover-ai-powered-research#:~:text=How%20does%20Yewno%20Discover%20work%3F,the%20user%20in%20their%20research [↑](#footnote-ref-102)
103. https://www.fjlib.net/zt/mtxh/stxk/llysj/fjstsgxk201403/201804/t20180414\_358197.htm#:~:text=%E5%B9%B4%E5%9C%A8%E3%80%8A%E7%A7%91%E5%AD%A6%E3%80%8B%E6%9D%82%E5%BF%97%E4%B8%8A%E5%8F%91%E8%A1%A8%E7%9A%84%E4%B8%80%E7%AF%87%E8%AE%BA%E6%96%87%E4%B8%AD%EF%BC%8C%E5%B0%B1%E8%AE%A4%E4%B8%BA%E4%BC%A0%E7%BB%9F%E7%9A%84%E5%9F%BA%E4%BA%8E%E5%AD%A6%E7%A7%91%E7%9A%84%E6%A3%80%E7%B4%A2%E5%92%8C%E6%91%98%E8%A6%81%E6%9C%8D%E5%8A%A1%E9%9A%BE%E4%BB%A5%E6%BB%A1%E8%B6%B3%E7%A0%94%E7%A9%B6%E7%9A%84%E9%9C%80%E6%B1%82%E3%80%82 [↑](#footnote-ref-103)
104. https://whly.gd.gov.cn/service\_newggwhzx/content/post\_4061139.html#:~:text=%E7%AB%99%E5%9B%BE%E4%B9%A6%E8%B5%84%E6%BA%90%2C%E8%BF%98%E5%8F%AF%E8%8E%B7%E5%8F%96%E5%85%AC%E5%85%B1%E5%9B%BE%E4%B9%A6%E9%A6%86%E8%97%8F%E4%B9%A6%E6%8C%87%E5%BC%95%2C%E5%AE%9E%E7%8E%B0%E4%BA%86%E5%85%AC%E7%A7%81%E8%97%8F%E4%B9%A6%E7%9A%84%E8%9E%8D%E5%90%88%E5%88%A9%E7%94%A8%2C%E6%8F%90%E9%AB%98%E4%BA%86%E5%9B%BE%E4%B9%A6%E8%8E%B7%E5%8F%96%E6%95%88%E7%8E%87%2C%E6%BB%A1%E8%B6%B3%E4%BA%86%E5%B8%82%E6%B0%91%E9%98%85%E8%AF%BB%E9%9C%80%E6%B1%82%E3%80%82 [↑](#footnote-ref-104)
105. https://whly.gd.gov.cn/service\_newggwhzx/content/post\_4061139.html#:~:text=%E7%AB%99%E5%9B%BE%E4%B9%A6%E8%B5%84%E6%BA%90%2C%E8%BF%98%E5%8F%AF%E8%8E%B7%E5%8F%96%E5%85%AC%E5%85%B1%E5%9B%BE%E4%B9%A6%E9%A6%86%E8%97%8F%E4%B9%A6%E6%8C%87%E5%BC%95%2C%E5%AE%9E%E7%8E%B0%E4%BA%86%E5%85%AC%E7%A7%81%E8%97%8F%E4%B9%A6%E7%9A%84%E8%9E%8D%E5%90%88%E5%88%A9%E7%94%A8%2C%E6%8F%90%E9%AB%98%E4%BA%86%E5%9B%BE%E4%B9%A6%E8%8E%B7%E5%8F%96%E6%95%88%E7%8E%87%2C%E6%BB%A1%E8%B6%B3%E4%BA%86%E5%B8%82%E6%B0%91%E9%98%85%E8%AF%BB%E9%9C%80%E6%B1%82%E3%80%82 [↑](#footnote-ref-105)
106. https://whly.gd.gov.cn/service\_newggwhzx/content/post\_4061139.html#:~:text=%E7%AB%99%E5%9B%BE%E4%B9%A6%E8%B5%84%E6%BA%90%2C%E8%BF%98%E5%8F%AF%E8%8E%B7%E5%8F%96%E5%85%AC%E5%85%B1%E5%9B%BE%E4%B9%A6%E9%A6%86%E8%97%8F%E4%B9%A6%E6%8C%87%E5%BC%95%2C%E5%AE%9E%E7%8E%B0%E4%BA%86%E5%85%AC%E7%A7%81%E8%97%8F%E4%B9%A6%E7%9A%84%E8%9E%8D%E5%90%88%E5%88%A9%E7%94%A8%2C%E6%8F%90%E9%AB%98%E4%BA%86%E5%9B%BE%E4%B9%A6%E8%8E%B7%E5%8F%96%E6%95%88%E7%8E%87%2C%E6%BB%A1%E8%B6%B3%E4%BA%86%E5%B8%82%E6%B0%91%E9%98%85%E8%AF%BB%E9%9C%80%E6%B1%82%E3%80%82 [↑](#footnote-ref-106)
107. https://www.techscience.com/iasc/v33n3/47092/html [↑](#footnote-ref-107)
108. https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/context-aware-computing-context-awareness-context-aware-user-interfaces-and-implicit-interaction?srsltid=AfmBOorXfODm3\_bxP8bKSjXQyHyy2qXny9XjwWifLUxTP6rF7AHz-pMg [↑](#footnote-ref-108)
109. https://www.ingentaconnect.com/content/mcb/238/2013/00000031/00000002/art00004 [↑](#footnote-ref-109)
110. https://www.libraryjournal.com/story/ais-role-in-the-future-of-library-services-250501 [↑](#footnote-ref-110)
111. https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/context-aware-computing-context-awareness-context-aware-user-interfaces-and-implicit-interaction?srsltid=AfmBOorXfODm3\_bxP8bKSjXQyHyy2qXny9XjwWifLUxTP6rF7AHz-pMg [↑](#footnote-ref-111)
112. https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/context-aware-computing-context-awareness-context-aware-user-interfaces-and-implicit-interaction?srsltid=AfmBOorXfODm3\_bxP8bKSjXQyHyy2qXny9XjwWifLUxTP6rF7AHz-pMg [↑](#footnote-ref-112)
113. https://oulurepo.oulu.fi/bitstream/handle/10024/39921/nbnfioulu-201406091704.pdf?sequence=1&isAllowed=y [↑](#footnote-ref-113)
114. https://oulurepo.oulu.fi/bitstream/handle/10024/39921/nbnfioulu-201406091704.pdf?sequence=1&isAllowed=y [↑](#footnote-ref-114)
115. https://www.techscience.com/iasc/v33n3/47092/html [↑](#footnote-ref-115)
116. https://www.techscience.com/iasc/v33n3/47092/html [↑](#footnote-ref-116)
117. https://www.techscience.com/iasc/v33n3/47092/html [↑](#footnote-ref-117)
118. https://www.ingentaconnect.com/content/mcb/238/2013/00000031/00000002/art00004 [↑](#footnote-ref-118)
119. https://www.ingentaconnect.com/content/mcb/238/2013/00000031/00000002/art00004 [↑](#footnote-ref-119)
120. https://oulurepo.oulu.fi/bitstream/handle/10024/39921/nbnfioulu-201406091704.pdf?sequence=1&isAllowed=y [↑](#footnote-ref-120)
121. https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/context-aware-computing-context-awareness-context-aware-user-interfaces-and-implicit-interaction?srsltid=AfmBOorXfODm3\_bxP8bKSjXQyHyy2qXny9XjwWifLUxTP6rF7AHz-pMg [↑](#footnote-ref-121)
122. https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/context-aware-computing-context-awareness-context-aware-user-interfaces-and-implicit-interaction?srsltid=AfmBOorXfODm3\_bxP8bKSjXQyHyy2qXny9XjwWifLUxTP6rF7AHz-pMg [↑](#footnote-ref-122)
123. https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/context-aware-computing-context-awareness-context-aware-user-interfaces-and-implicit-interaction?srsltid=AfmBOorXfODm3\_bxP8bKSjXQyHyy2qXny9XjwWifLUxTP6rF7AHz-pMg [↑](#footnote-ref-123)
124. https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/context-aware-computing-context-awareness-context-aware-user-interfaces-and-implicit-interaction?srsltid=AfmBOorXfODm3\_bxP8bKSjXQyHyy2qXny9XjwWifLUxTP6rF7AHz-pMg [↑](#footnote-ref-124)
125. https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/context-aware-computing-context-awareness-context-aware-user-interfaces-and-implicit-interaction?srsltid=AfmBOorXfODm3\_bxP8bKSjXQyHyy2qXny9XjwWifLUxTP6rF7AHz-pMg [↑](#footnote-ref-125)
126. https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/context-aware-computing-context-awareness-context-aware-user-interfaces-and-implicit-interaction?srsltid=AfmBOorXfODm3\_bxP8bKSjXQyHyy2qXny9XjwWifLUxTP6rF7AHz-pMg [↑](#footnote-ref-126)
127. https://oulurepo.oulu.fi/bitstream/handle/10024/39921/nbnfioulu-201406091704.pdf?sequence=1&isAllowed=y [↑](#footnote-ref-127)
128. https://paginaspersonales.deusto.es/dipina/papers/ucami2011\_submission\_55.pdf [↑](#footnote-ref-128)
129. https://paginaspersonales.deusto.es/dipina/papers/ucami2011\_submission\_55.pdf [↑](#footnote-ref-129)
130. https://www.techscience.com/iasc/v33n3/47092/html [↑](#footnote-ref-130)
131. https://www.techscience.com/iasc/v33n3/47092/html [↑](#footnote-ref-131)
132. https://www.ingentaconnect.com/content/mcb/238/2013/00000031/00000002/art00004 [↑](#footnote-ref-132)
133. https://www.ingentaconnect.com/content/mcb/238/2013/00000031/00000002/art00004 [↑](#footnote-ref-133)
134. https://www.ingentaconnect.com/content/mcb/238/2013/00000031/00000002/art00004 [↑](#footnote-ref-134)
135. https://www.ingentaconnect.com/content/mcb/238/2013/00000031/00000002/art00004 [↑](#footnote-ref-135)
136. https://www.ingentaconnect.com/content/mcb/238/2013/00000031/00000002/art00004 [↑](#footnote-ref-136)
137. https://oulurepo.oulu.fi/bitstream/handle/10024/39921/nbnfioulu-201406091704.pdf?sequence=1&isAllowed=y [↑](#footnote-ref-137)
138. https://oulurepo.oulu.fi/bitstream/handle/10024/39921/nbnfioulu-201406091704.pdf?sequence=1&isAllowed=y [↑](#footnote-ref-138)
139. https://oulurepo.oulu.fi/bitstream/handle/10024/39921/nbnfioulu-201406091704.pdf?sequence=1&isAllowed=y [↑](#footnote-ref-139)
140. https://oulurepo.oulu.fi/bitstream/handle/10024/39921/nbnfioulu-201406091704.pdf?sequence=1&isAllowed=y [↑](#footnote-ref-140)
141. https://www.techscience.com/iasc/v33n3/47092/html [↑](#footnote-ref-141)
142. https://www.techscience.com/iasc/v33n3/47092/html [↑](#footnote-ref-142)
143. https://www.techscience.com/iasc/v33n3/47092/html [↑](#footnote-ref-143)
144. https://www.libraryjournal.com/story/ais-role-in-the-future-of-library-services-250501 [↑](#footnote-ref-144)
145. https://www.libraryjournal.com/story/ais-role-in-the-future-of-library-services-250501 [↑](#footnote-ref-145)
146. https://www.libraryjournal.com/story/ais-role-in-the-future-of-library-services-250501 [↑](#footnote-ref-146)
147. https://www.ingentaconnect.com/content/mcb/238/2013/00000031/00000002/art00004 [↑](#footnote-ref-147)
148. https://link.springer.com/article/10.1007/s007790200028 [↑](#footnote-ref-148)
149. https://www.sciencedirect.com/science/article/abs/pii/S0263224119312874 [↑](#footnote-ref-149)
150. https://www.sciencedirect.com/science/article/abs/pii/S0263224119312874 [↑](#footnote-ref-150)
151. https://www.sciencedirect.com/science/article/abs/pii/S0263224119312874 [↑](#footnote-ref-151)
152. https://www.sciencedirect.com/science/article/pii/S1474034621002433 [↑](#footnote-ref-152)
153. https://www.sciencedirect.com/science/article/pii/S1474034621002433 [↑](#footnote-ref-153)
154. https://dl.acm.org/doi/10.1145/3677389.3702556 [↑](#footnote-ref-154)
155. https://www.libraryjournal.com/story/ais-role-in-the-future-of-library-services-250501 [↑](#footnote-ref-155)
156. https://www.libraryjournal.com/story/ais-role-in-the-future-of-library-services-250501 [↑](#footnote-ref-156)
157. https://www.libraryjournal.com/story/ais-role-in-the-future-of-library-services-250501 [↑](#footnote-ref-157)
158. https://www.libraryjournal.com/story/ais-role-in-the-future-of-library-services-250501 [↑](#footnote-ref-158)
159. https://www.libraryjournal.com/story/ais-role-in-the-future-of-library-services-250501 [↑](#footnote-ref-159)
160. https://paginaspersonales.deusto.es/dipina/papers/ucami2011\_submission\_55.pdf [↑](#footnote-ref-160)
161. https://www.techscience.com/iasc/v33n3/47092/html [↑](#footnote-ref-161)
162. https://www.techscience.com/iasc/v33n3/47092/html [↑](#footnote-ref-162)
163. https://www.techscience.com/iasc/v33n3/47092/html [↑](#footnote-ref-163)
164. https://www.ingentaconnect.com/content/mcb/238/2013/00000031/00000002/art00004 [↑](#footnote-ref-164)
165. https://www.sciencedirect.com/science/article/pii/S1474034621002433 [↑](#footnote-ref-165)
166. https://www.sciencedirect.com/science/article/pii/S1474034621002433 [↑](#footnote-ref-166)
167. https://ital.corejournals.org/index.php/ital/article/view/16719 [↑](#footnote-ref-167)
168. https://ascelibrary.org/doi/10.1061/JCEMD4.COENG-15230 [↑](#footnote-ref-168)
169. https://cloud.google.com/discover/what-is-semantic-search [↑](#footnote-ref-169)
170. https://cloud.google.com/discover/what-is-semantic-search [↑](#footnote-ref-170)
171. https://cloud.google.com/discover/what-is-semantic-search [↑](#footnote-ref-171)
172. https://www.lib.dicp.ac.cn/info/1063/2062.htm [↑](#footnote-ref-172)
173. https://www.libraryjournal.com/story/ais-role-in-the-future-of-library-services-250501 [↑](#footnote-ref-173)
174. https://chatlibrary.newacademic.net/ [↑](#footnote-ref-174)
175. https://www.libraryjournal.com/story/ais-role-in-the-future-of-library-services-250501 [↑](#footnote-ref-175)
176. https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/context-aware-computing-context-awareness-context-aware-user-interfaces-and-implicit-interaction?srsltid=AfmBOorXfODm3\_bxP8bKSjXQyHyy2qXny9XjwWifLUxTP6rF7AHz-pMg [↑](#footnote-ref-176)
177. https://www.ingentaconnect.com/content/mcb/238/2013/00000031/00000002/art00004 [↑](#footnote-ref-177)
178. https://www.libraryjournal.com/story/ais-role-in-the-future-of-library-services-250501 [↑](#footnote-ref-178)
179. https://chatlibrary.newacademic.net/ [↑](#footnote-ref-179)
180. https://www.libraryjournal.com/story/ais-role-in-the-future-of-library-services-250501 [↑](#footnote-ref-180)
181. https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2364 [↑](#footnote-ref-181)
182. https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2364 [↑](#footnote-ref-182)
183. https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2364 [↑](#footnote-ref-183)
184. https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2364 [↑](#footnote-ref-184)
185. https://oulurepo.oulu.fi/bitstream/handle/10024/39921/nbnfioulu-201406091704.pdf?sequence=1&isAllowed=y [↑](#footnote-ref-185)
186. https://oulurepo.oulu.fi/bitstream/handle/10024/39921/nbnfioulu-201406091704.pdf?sequence=1&isAllowed=y [↑](#footnote-ref-186)
187. https://www.libraryjournal.com/story/ais-role-in-the-future-of-library-services-250501 [↑](#footnote-ref-187)
188. https://www.libraryjournal.com/story/ais-role-in-the-future-of-library-services-250501 [↑](#footnote-ref-188)
189. https://link.springer.com/article/10.1007/s007790200028 [↑](#footnote-ref-189)
190. https://link.springer.com/article/10.1007/s007790200028 [↑](#footnote-ref-190)
191. https://link.springer.com/article/10.1007/s007790200028 [↑](#footnote-ref-191)
192. https://www.libraryjournal.com/story/ais-role-in-the-future-of-library-services-250501 [↑](#footnote-ref-192)
193. https://www.sciencedirect.com/science/article/pii/S1474034624007377 [↑](#footnote-ref-193)
194. https://ital.corejournals.org/index.php/ital/article/view/16719 [↑](#footnote-ref-194)
195. https://ital.corejournals.org/index.php/ital/article/view/16719 [↑](#footnote-ref-195)
196. https://www.techscience.com/iasc/v33n3/47092/html [↑](#footnote-ref-196)
197. https://www.ingentaconnect.com/content/mcb/238/2013/00000031/00000002/art00004 [↑](#footnote-ref-197)
198. https://www.cs.unc.edu/~azuma/ARpresence.pdf [↑](#footnote-ref-198)
199. https://www.cs.unc.edu/~azuma/ARpresence.pdf [↑](#footnote-ref-199)
200. https://en.wikipedia.org/wiki/Augmented\_reality [↑](#footnote-ref-200)
201. https://zh.wikipedia.org/zh-hans/%E6%93%B4%E5%A2%9E%E5%AF%A6%E5%A2%83 [↑](#footnote-ref-201)
202. https://en.wikipedia.org/wiki/Augmented\_reality [↑](#footnote-ref-202)
203. https://arinsider.co/2022/11/18/staging-reality-the-ar-vr-mr-continuum/ [↑](#footnote-ref-203)
204. https://www.researchgate.net/publication/254057138\_Here\_We\_Are\_Where\_Are\_We\_Locating\_Mixed\_Reality\_in\_The\_Age\_of\_the\_Smartphone [↑](#footnote-ref-204)
205. https://en.wikipedia.org/wiki/Augmented\_reality [↑](#footnote-ref-205)
206. https://arinsider.co/2022/11/18/staging-reality-the-ar-vr-mr-continuum/ [↑](#footnote-ref-206)
207. https://www.cs.unc.edu/~azuma/ARpresence.pdf [↑](#footnote-ref-207)
208. https://www.frontiersin.org/journals/virtual-reality/articles/10.3389/frvir.2021.578080/full [↑](#footnote-ref-208)
209. https://arinsider.co/2022/11/18/staging-reality-the-ar-vr-mr-continuum/ [↑](#footnote-ref-209)
210. ttps://www.frontiersin.org/journals/virtual-reality/articles/10.3389/frvir.2021.578080/full [↑](#footnote-ref-210)
211. https://en.wikipedia.org/wiki/Augmented\_reality [↑](#footnote-ref-211)
212. https://zh.wikipedia.org/zh-hans/%E6%93%B4%E5%A2%9E%E5%AF%A6%E5%A2%83 [↑](#footnote-ref-212)
213. https://en.wikipedia.org/wiki/Augmented\_reality [↑](#footnote-ref-213)
214. https://www.gerardfriel.com/ar/the-history-of-ar/ [↑](#footnote-ref-214)
215. https://fekki.io/blogs/a-historical-overview-of-ar-vr-development-and-key-milestones/ [↑](#footnote-ref-215)
216. https://en.wikipedia.org/wiki/Augmented\_reality [↑](#footnote-ref-216)
217. https://en.wikipedia.org/wiki/Augmented\_reality [↑](#footnote-ref-217)
218. https://en.wikipedia.org/wiki/Augmented\_reality [↑](#footnote-ref-218)
219. https://en.wikipedia.org/wiki/Augmented\_reality [↑](#footnote-ref-219)
220. https://en.wikipedia.org/wiki/Augmented\_reality [↑](#footnote-ref-220)
221. https://en.wikipedia.org/wiki/Augmented\_reality [↑](#footnote-ref-221)
222. https://www.gov.cn/zhengce/2022-11/01/content\_5723274.htm [↑](#footnote-ref-222)
223. https://en.wikipedia.org/wiki/Augmented\_reality [↑](#footnote-ref-223)
224. https://en.wikipedia.org/wiki/Augmented\_reality [↑](#footnote-ref-224)
225. https://www.frontiersin.org/journals/virtual-reality/articles/10.3389/frvir.2021.578080/full [↑](#footnote-ref-225)
226. https://www.radiantvisionsystems.com/blog/future-looks-bright-ar/vr/mr-2023-beyond [↑](#footnote-ref-226)
227. https://www.frontiersin.org/journals/virtual-reality/articles/10.3389/frvir.2021.578080/full [↑](#footnote-ref-227)