智慧图书馆大模型创新与应用白皮书

上海图书馆(上海科学技术情报研究所)
上海人工智能研究院
智慧图书馆技术应用联盟（筹）

2023 年 9 月

前　言

当前，大模型技术加快创新发展，掀起人工智能创新发展的新一轮浪潮。为推动智慧图书馆建设，我国政府积极出台了一系列政策，加强智慧图书馆的战略部署。经过多年的发展，智慧图书馆已经积累了良好的平台和数据基础。与此同时，信息技术的飞速发展和数字人文研究也对图书馆提出高价值知识服务能力的新需求。在这个人工智能科技创新日新月异的时代，智慧图书馆作为知识传播的重要载体，迎来了前所未有的发展机遇和契机。

本白皮书深入探讨智慧图书馆领域大模型应用，通过研究大模型对于智慧图书馆的赋能作用，并从场景创新、落地实践、生态建设三方面提供相关建议和指导，以促进智慧图书馆领域大模型技术的创新应用。首先勾勒**大模型赋能智慧图书馆全景视图：**

一是提出大模型技术赋能智慧图书馆路径，即大模型基于自然语言处理技术优势和涌现能力，或通过模型微调训练、对接外部系统以满足体验类、专业类不同的功能和性能需求。

二是提出两类智慧图书馆大模型应用总体架构，分别为“智慧图书馆+大模型”“大模型+智慧图书馆”。“智慧图书馆+大模型”围绕图书馆业务流程嵌入大模型技术，以图书馆业务需求为导向建设大模型应用，提升馆员业务管理和工作的效率以及智能化程度。“大模型+智慧图书馆”以大模型为核心延伸智慧图书馆服务应用，面向读者打造统一服务入口，提供咨询问答、检索推荐、新型阅读等多元服务。此外提出智慧图书馆大模型应用将逐渐由“+大模型”向“大模型+”范式转变的演变特征。

三是形成智慧图书馆大模型应用场景全景视图，梳理智慧管理、智慧服务、智慧业务、智慧空间四个方面的智慧图书馆大模型应用场景。并分析提出智慧图书馆大模型应用将逐渐由内部管理、对外服务过渡到未来体验的发展态势。

以大模型赋能智慧图书馆全景视图为引领，**白皮书随后从场景创新路径、落地实践路径、生态建设路径三方面提出重点方向、技术参考、建议举措，**以加快促进智慧图书馆领域大模型技术的创新

应用：

一是场景创新路径，白皮书梳理了智慧管理、智慧业务、智慧服务、智慧空间四个方向典型的智慧图书馆大模型应用场景，为智慧图书馆领域大模型技术创新应用提供借鉴和参考。智慧管理对图书馆馆藏资源进行语义化管理和深度分析挖掘，充分释放图书馆信息资源的潜在价值。智慧业务依托大模型技术面向馆员的日常工作和业务提供智能化辅助，以智慧化手段提高馆员信息素养和能力。智慧服务依托大模型技术面向读者提供智能、便捷、人性化、个性化的新型阅读体验和高价值的知识服务。智慧空间运用大模型技术优化图书馆空间和读者的交互体验，打造基于元宇宙图书馆的虚实融合交互体验。

二是落地实践路径，白皮书首先从智慧图书馆大模型应用的规划设计角度提供了包括现状调研、需求分析、总体设计及实施路径设计等方面的工作指导，其次针对智慧图书馆大语言模型的微调训练，白皮书围绕模型开发训练全过程，梳理了包括模型选型、模型调优、模型评估、模型部署、模型使用等环节的技术指南，以为智慧图书馆大模型应用创新实践提供借鉴和参考。

三是生态建设路径，白皮书围绕建立完善智慧图书馆大模型创新生态，提出加强数据开放共享和分析挖掘、搭建多元服务集聚的开放平台、完善大模型创新应用标准规范、依托联盟营造开放的创新氛围、开展行业人才培养与交流互动等建议举措，推动图书馆行业开放合作，共建智慧图书馆大模型创新生态，为智慧图书馆大模型创新应用营造开放包容、协同创新的发展环境。

鉴于大模型技术的快速发展和行业应用仍处在创新探索阶段，以及对相关行业和业务的理解不够深入，我们深知白皮书存在诸多不足之处，可能仍然是完善版本前的 0.9 版本。因此，我们也诚挚邀请各界人士进行批评指正，我们将借助各方经验和智慧对白皮书进行修改和完善，从而为智慧图书馆大模型创新应用提供有益参考。

目 录

一、 智慧图书馆发展环境与机遇 1
（一） 政策布局和需求驱动加快智慧图书馆建设 1
（二） 图书馆紧跟数字化发展步伐并积累良好基础 4
（三） 大模型技术赋能智慧图书馆具有广阔前景 7
二、 大模型赋能智慧图书馆全景视图 15
（一） 大模型技术赋能智慧图书馆路径 15
（二） 智慧图书馆领域大模型应用总体架构 17
（三） 智慧图书馆大模型应用场景全景视图 19
三、 智慧图书馆大模型应用场景创新路径 23
（一） 智慧管理实现图书馆资源的语义化管理 23
（二） 智慧业务打造辅助图情业务的智能助手 25
（三） 智慧服务提供新型阅读体验和知识服务 26
（四） 智慧空间打造虚实融合的智能交互体验 28
四、 智慧图书馆大模型应用落地实践路径 30
（一） 智慧图书馆大模型创新应用规划设计 30
（二） 智慧图书馆大模型创新开发落地实施 36
五、 智慧图书馆大模型创新生态建设路径 55
（一） 加强数据开放共享和分析挖掘 55
（二） 搭建多元服务集聚的开放平台 56
（三） 完善大模型创新应用标准规范 56
（四） 依托联盟营造开放的创新氛围 57
（五） 开展行业人才培养与交流互动 58

一、智慧图书馆发展环境与机遇

（一）政策布局和需求驱动加快智慧图书馆建设

1、公共文化与技术创新政策叠加为智慧图书馆带来新机遇

我国陆续出台推进公共文化服务数字化的相关政策，智慧图书馆建设的战略部署持续深化。

2021 年 3 月，我国正式发布了《中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要》，明确提出积极发展智慧图书馆，提供智慧便捷的公共服务。建设智慧图书馆成为加快数字社会建设步伐当中非常重要的战略任务。

同年 3 月，国家发展改革委联合多部门印发《关于推动公共文化服务高质量发展的意见》，提出加快推进公共文化服务数字化，明确加强智慧图书馆体系建设，建立覆盖全国的图书馆智慧服务和管理架构，大力发展基于 5G 等新技术应用的数字服务类型，拓宽数字文化服务应用场景。

2021 年 4 月，文化和旅游部印发《“十四五”文化和旅游发展规划》，提出加快公共数字文化建设，部署全国智慧图书馆体系建设重点任务，即以全国智慧图书馆体系建设为核心，搭建一套支撑智慧图书馆运行的云基础设施，形成国家层面知识内容集成仓储，建设和运行智慧图书馆管理系统，在全国各级图书馆及其基层服务网点普遍建立实体智慧服务空间。

2021 年 6 月，文化和旅游部发布《“十四五”公共文化服务体系建设规划》，提出推动公共文化数字化、网络化、智能化建设，以全国智慧图书馆体系建设项目和公共文化云项目为依托，加强数字文化内容资源建设，建设公共文化网络平台，以及拓展公共文化服务智慧应用场景。

2022 年 5 月，中共中央办公厅、国务院办公厅印发了《关于推进实施国家文化数字化战略的意见》，提出统筹推进国家文化大数据

体系、全国智慧图书馆体系和公共文化云建设，增强公共文化数字内容的供给能力，提升公共文化服务数字化水平。

此外，中国国家图书馆在2021年10月发布了《国家图书馆“十四五”发展规划》，提出实施“智慧转型”战略，基于5G网络、大数据、云计算、物联网、区块链、人工智能等关键技术，推动国家图书馆在资源、服务、设施、管理等领域全面实现智慧化转型，并部署加强信息化基础设施建设、构建智慧图书馆管理系统、推进线下服务空间智慧化升级、建设数字孪生国家图书馆、构建开放知识服务平台等重点举措。

以深化人工智能等新型数字技术创新应用为主线，我国先后发布了多个政策文件鼓励引导人工智能在公共文化服务领域的创新应用，为智慧图书馆建设营造了良好的政策环境。

2021年12月，国务院印发《“十四五”数字经济发展规划》，提出充分运用新型数字技术持续提升公共服务数字化普惠水平，加快优秀文化的数字化转化和开发，推动文化教育等领域公共服务资源的数字化供给和网络化服务。

2022年7月，国家科技部、工业和信息化部等六部门联合引发《关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见》，明确提出围绕高水平科研活动加强人工智能应用场景创新，充分发挥人工智能技术在文献数据获取、实验预测、结果分析等方面的作用，推动人工智能技术成为解决科学问题的新范式，为图书馆加强文献知识资源的整合开发利用、优化面向科研机构的知识服务等带来了有利的政策条件。

地方层面，结合大模型技术创新发展热潮，多个城市谋划了促进大模型创新应用的相关政策布局，为探索图书馆行业领域的大模型落地应用提供了良好契机。

北京市率先发布《北京市促进通用人工智能创新发展的若干措

施》，围绕加强算力资源统筹供给、加强高质量数据要素供给、系统构建通用人工智能技术体系、创新通用人工智能技术场景应用等方面出台二十一条政策措施。

上海市推出《上海市推动人工智能大模型创新发展的若干措施》，围绕大模型创新能力、创新要素、创新应用、创新环境四大方向实施大模型创新扶持计划、智能算力加速计划、示范应用推进计划，并发布“模都”倡议，全力推进卓越引领的“模”都上海建设。

深圳市发布《深圳市加快推动人工智能高质量发展高水平应用行动方案（2023—2024年）》，从强化智能算力集群供给、增强关键核心技术与产品创新能力、提升产业集聚水平、打造全域全时场景应用、强化数据和人才要素供给、保障措施等六个方面提出十八项具体举措。

成都市发布《成都市加快大模型创新应用推进人工智能产业高质量发展的若干措施》，具体围绕强化智能算力供给、提升创新策源能力、提升产业发展能级、构建全域场景体系、加强生态要素聚集等方面出台二十条政策措施。

2、信息技术发展和数字人文研究对图书馆服务能力提出新需求

信息技术发展倒逼图书馆提升信息服务生态位，提高知识服务和空间伴随能力。一方面，伴随着互联网、大数据等技术的发展，信息资源呈现爆炸式的增长，信息内容的体量、复杂度不断变高，使得单个图书馆管理的信息资源占比变得越来越小。另一方面，ChatGPT等大语言模型应用的创新也使得人们对于信息、知识的获取、搜索、生产方式也发生了显著变化，人们可直接通过对ChatGPT提问来高效地获取相关知识。过去一段时间，图书馆作为提供资源和服务的门户而存在，管理重在书目、纸本等实体资源的收集、处理、组织和服务以及实体资源的数字化管理。如果图书馆长期停留在以资源为中心、以机构为中心的管理理念，而不重视以

用户为中心并利用数字技术提高信息服务价值，图书馆将难以适应数字时代下信息技术的发展潮流、难以满足数字时代下读者用户个性化、高质量的阅读需求，可能会沦为信息资源价值没有得到充分发挥的信息孤岛。因此，图书馆需要提升在信息服务中的生态位，提高信息资源收集、加工、流通和知识分析挖掘、服务能力，从提供资源检索的传统情报向提供高价值知识服务的智慧情报转变，并打造促进用户交流互动、学习分享、价值创造的空间，这样方能保持与时俱进，充分融入信息社会下的知识服务生态。

数字人文范式转变要求图书馆加快提高知识生产和服务能力。数字人文是数字技术与人文学科的交叉领域，伴随计算机技术的发展，数字人文已从早期利用计算机实现人文资源的复刻镜像、全文检索排序，逐步发展至基于自然语言处理、知识图谱、数据可视化、AR/VR 等数字技术进行数据加工、知识生产、内容服务等高阶形态，数字人文研究的精细度、价值挖掘深度显著提高。图书馆作为文化记忆机构，具有大量的家谱、古籍、地方志、碑帖等极具人文研究价值的特藏资源，在文化传承、参与数字人文研究、赋能人文学科建设方面具有重要地位和作用。因此，图书馆需要加快提高运用数字技术进行内容挖掘和知识生产、服务的能力，通过运用数字技术对特藏资源进行数字化加工，提取细粒度知识信息，或进行数字化再现兼顾古籍的“藏”与“用”，从而加强图书馆特藏资源的开发利用，以满足人文学科建设和研究、文化传承的需要。

（二）图书馆紧跟数字化发展步伐并积累良好基础

1、图书馆经历从自动化、数字化到智慧化的发展历程

图书馆自动化时期对应 Web1.0 阶段。这一时期，公众的信息来源主要为 PGC（Professional Generated Content），指的是由专业人员生成、编辑和发布的内容。图书馆的机读目录(MARC)和由此发展出的元数据（metadata）是由 PGC 方式产生内容的典型代表。1969

年公布发行的MARC标准奠定了图书馆行业在全球范围内对信息技术应用的前沿地位。在图书馆1.0时期，图书馆自动化的业务主要是结合计算机技术和已有的MARC标准和元数据来自动化管理纸质馆藏。

图书馆数字化时期伴随着Web2.0的发展。在这一时期，互联网上的社交媒体为用户提供了一个开放的、丰富的平台，其中UGC（User Generated Content）成为了重要的内容来源。UGC指的是由普通用户创建和发布的内容，这种内容通常没有经过专业人员的审核。这一时期，图书馆开始建立数字图书馆项目，保存数字资源。除了保存与PGC相关的实体书的数字化内容外，数字图书馆还开始保存门户检索网页上的用户书评，以及采用众包模式下用户提供的内容。

随着Web3.0兴起,图书馆进入智慧化的发展阶段。Web3.0时代，虚实融合技术、智能交互技术等新技术逐渐兴起，智能系统不仅会响应用户的搜索请求和组合呈现信息，还能像人类一样读懂信息，并根据用户需求生成个性化的内容，用户从互联网上获得的信息将包括PGC、UGC，以及由人工智能生成的内容，即AIGC（Artificially Intelligence Generated Content）。面对Web3.0的发展，图书馆原有的集成管理平台已不能满足IT应用需求和多元灵活的内容需求。于是，图书馆开始建立智慧图书馆项目，搭建第三代图书馆服务平台，这一新平台具有融合多种平台、多种数据类型的知识整合能力。随着智能技术的广泛应用和新一代图书馆服务平台的发展，图书馆进入智慧化阶段。

2、图书馆数字化发展积累良好基础

图书馆数字化发展过程中逐渐积累了良好的数据和平台基础。在数据方面，一方面，随着互联网的普及，图书馆的部分服务场景从实体馆藏转向了线上数字资源，数字图书馆开始着重于对纸质资源的数字化，丰富了图书馆的数字馆藏，为后续的服务奠定了内容

基础。另一方面，Web2.0 技术的兴起使得图书馆与读者的交互变得更加密切，用户生成的内容（UGC）也为图书馆提供了新的内容来源，其中包括众包平台上的标引数据和家谱数据等宝贵的高质量数据。**在平台方面，**图书馆早期建立了集成管理平台，平台能够整合各种资源，如电子书、期刊、数据库、多媒体资源等，为用户提供统一的检索和访问入口，并且许多图书馆都提供了移动应用或移动优化的网站，使用户可以随时随地访问图书馆的资源和服务。随着图书馆发展逐步从传统的数字化转向智慧化，图书馆开始搭建第三代图书馆服务平台，这些平台具有融合多种平台、多种数据类型的知识整合能力，并能够根据用户的实际需求提供个性化、高质量的服务，具有自我迭代和演化的能力。新一代的图书馆服务平台采用微服务架构，每个功能或服务作为一个独立的模块存在，提高了系统的灵活性和可扩展性，使平台能够灵活地组合和拼装智慧应用，满足不同图书馆的应用需求。总的来说，图书馆数字化发展得益于早期的技术和内容基础。目前，图书馆正在迅速适应新技术和数据来源，不断优化其服务和系统。

3、智慧图书馆发展取得成就和存在问题

如前所述，图书馆在数字化过程中积累了大量的数据资源，通过运用数字化技术将纸质资源数字化并进行管理，使其便于在线访问和搜索。然而这些管理和服务方式的核心在于资源的数字化管理，使得用户可以随时随地访问图书和文献资源，与智慧图书馆所强调的个性化、智慧化服务相比仍有不少差距。智慧图书馆的内涵不再以资源类型来定义，而是以服务特点来定义，意味着智慧图书馆不仅仅是一个资源的集合，更应该是一个智能的、以用户为中心的服务平台，实现以人为本的智慧化服务，通过使用大数据、人工智能等技术提供更加个性化、智能化的服务，例如通过用户行为分析为用户推荐合适的读物。

当前，智慧图书馆已经从概念阶段发展到实际应用阶段，国家图书馆等机构启动了一系列的智慧图书馆项目，加速整个行业的转型，包括硬件和软件的更新，以及业务流程和服务模式的重塑。但是，智慧图书馆的发展也面临着许多挑战。例如，如何确保数字资源的长期保存；数据的隐私和安全问题；以及如何确保智慧图书馆真正满足用户的需求，而不仅仅是技术的展示等问题，需要实施智慧化转型建设的图书馆进行综合考虑和应对。

（三）大模型技术赋能智慧图书馆具有广阔前景

1、基于 Transformer 的大型语言模型加速创新

2017 年谷歌发布的 Transformer 神经网络是大模型发展的源头技术，该模型在机器翻译任务上的表现大幅超越已有的模型，打破了传统 CNN 和 RNN 结构在自然语言处理领域的垄断地位。Transformer 最先是作为机器翻译的 Seq2Seq 模型而提出的，相比于传统网络结构，Transformer 的自注意力机制能从输入预料中捕获词与词之间的相关性，即对于每一个输入词，自注意力机制能对其它词赋予权重，权重越大表示词与词越相关。这种自注意力机制使得模型本身具有良好的语义理解和文本生成能力。同时，Transformer 神经网络具有更低的层计算复杂度，更强的并行操作性，更好的长距离依赖学习能力。这使得 Transformer 架构能够从数据中学习更多知识的同时，具有更好的计算效率。至此，Transformer 神经网络为大模型奠定了良好基础，并在自然语言处理、计算机视觉、智能语音、多模态等多个方向得到应用，基于 Transformer 架构的大模型开始加速涌现。

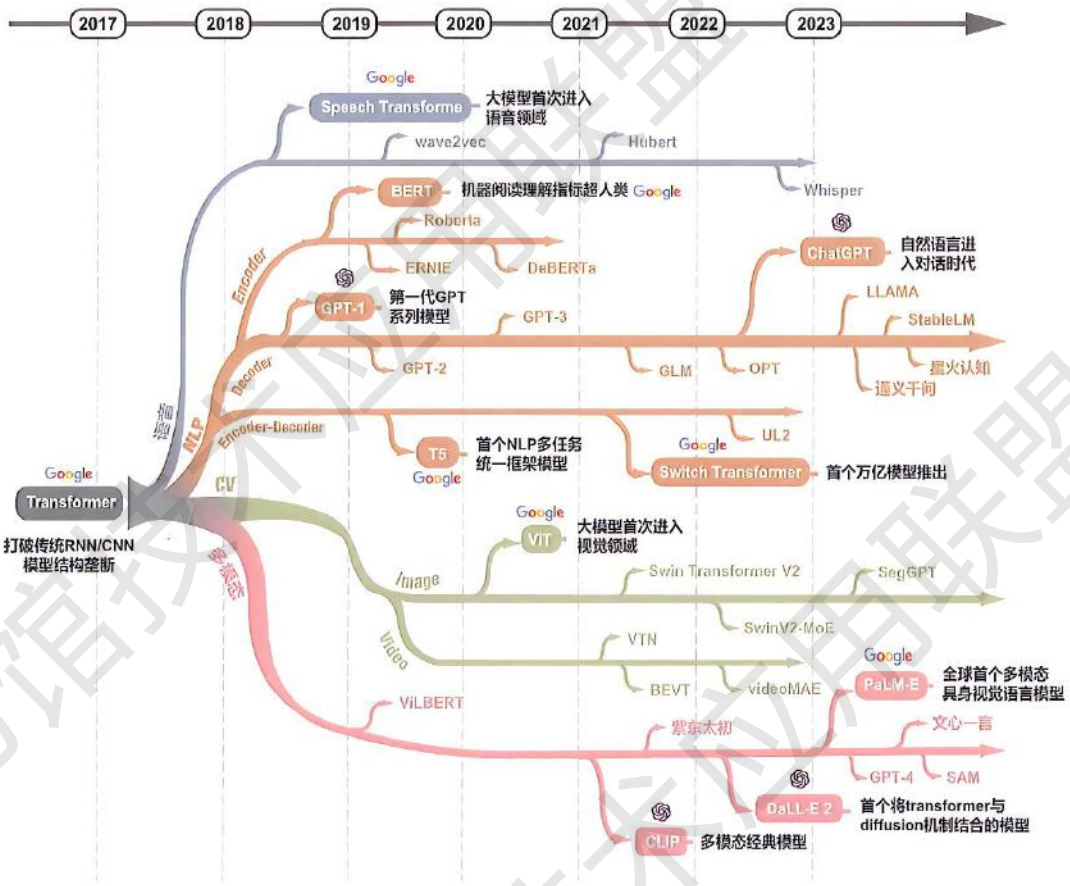

图 1：Transformer 大模型演进过程[1]

起源于机器翻译的 Transformer 由于在最具备“先验”通识和世界解释模型的数据类型——大规模文本语料库上得到了预训练，于是在自然语言处理领域取得突破性的进展并得到广泛应用，涌现出很多大型语言模型（通常，大型语言模型 LLM 是指包含数千亿或更多参数的语言模型）。**其中，令人瞩目的当属于仅采用 Transformer 解码器训练的 GPT 系列。**2022 年 11 月，OpenAI 发布了 ChatGPT。ChatGPT 采用 GPT3.5 架构，使用大量的语料库进行训练，具有语言理解和文本生成能力，可根据聊天上下文进行流畅自然的互动，做到与真人几乎无异的对话交流。而后，OpenAI 发布 GPT4，在 ChatGPT 的基础上增加了视觉模态，具备图文理解和处理能力，在数学、编码等能力上大幅超越 ChatGPT，表现出接近人类甚至超

[1] 资料来源：《中国人工智能大模型地图》

越人类的水平。

2、大语言模型的技术特性

大语言模型通常以 Transformer 作为基础架构，其核心能力在于通过大量的文本数据训练，学习语言的结构和规律，从而能够准确地理解和生成文本。此外，大语言模型还具有涌现能力，能够根据上下文生成新颖、富有创造力的内容。这些技术特点使得大语言模型在自然语言处理领域具有广泛的应用前景，广泛用于机器翻译、问答系统、文本摘要等多种应用场景。

（1）核心能力

大型语言模型基于自注意力机制将词与词之间的关联度蕴藏在深度神经网络模型中，具有较强的语言理解和生成能力，可根据给定的上下文（例如 prompt）生成高质量的文本，使得自然语言处理技术获得显著提升。

语义理解。基于 transformer 的大语言模型在语义理解方面具有很大的优势。它采用了注意力机制，能够有效地捕捉文本中不同位置之间的关联性，从而更好地理解文本的语义、句子的结构、上下文关系以及词汇间的相互联系。这种机制能够让模型在处理长文本时，仍然能够保持较高的准确性和效率。此外，transformer 模型还采用了多头注意力机制，能够同时捕捉多种不同类型的关联性，进一步提高了模型的语义理解能力。

信息匹配。信息匹配是指大语言模型能够根据用户的需求，从大量的数据中找出与查询相关的信息。在信息匹配方面，基于 transformer 的大语言模型也表现出色。它能够通过对文本进行深入分析，找到文本中的关键信息，并将其与其他文本进行匹配，根据上下文、文本结构和语义关系来提高信息匹配的准确性。这种匹配能力使得大语言模型能够在问答系统、推荐系统等应用场景中发挥重要作用。此外，transformer 大语言模型还具有很强的自适应能力，

能够根据不同的应用场景调整匹配策略，进一步提高匹配效果。

语言生成。语言生成是指大语言模型能够根据给定的上下文自动产生新的文本。大语言模型可以根据用户的需求生成描述、总结、扩展等不同类型的文本，根据上下文预测下一个词从而生成连贯、通顺的文本。在生成文本时，模型会考虑词汇、语法、风格等多个因素，确保生成的文本符合语言规则，并与给定的上下文保持一致。这种生成能力使得大语言模型能够用于机器翻译、文本摘要、自动写作等多种应用场景。此外，transformer 大语言模型还具有很强的泛化能力，能够在面对新颖、未见过的文本时仍然保持较高的生成质量。

（2）涌现能力

大型语言模型的性能大致遵循随着模型大小的增加而增加的规律，然而大模型某些能力是不可预测的，只有当模型大小超过某个水平时才能观察到，即大模型的涌现现象。大语言模型的涌现能力指的是当语言模型规模增加到一定程度时，会出现较小模型不具备的能力。这种能力并非随着模型规模的增加而线性增长，而是存在一个临界点。只有当模型规模超过这个临界值时，才会涌现出新的能力。这种涌现能力与模型规模大小（模型参数量）有一定的关联关系，但也可能受到其他因素的影响，例如训练数据量、数据质量等。目前，大模型主要包括以下典型的涌现能力。

上下文学习。上下文学习指的是大型语言模型根据给定的少量示例理解并执行任务的能力。通过使用示例来构建演示上下文，通常以自然语言模板的形式编写。模型将查询与上下文演示连接起来，形成一个带有提示的输入，并基于此输入进行预测。上下文学习不需要参数更新，直接使用预训练的语言模型进行预测。这种方法在许多零样本条件下被证明是有效的，引起了学术界和工业界的关注。

指令跟随。指令跟随指的是大语言模型理解并遵循自然语言指

令的能力。这种能力是通过在高质量指令数据上对模型进行微调来实现的。模型能够理解指令的要求并生成适当的响应。指令跟随已被证明能够提高大型语言模型在各种任务上的性能。

逐步推理。逐步推理指的是大型语言模型在复杂问题上进行多步推理的能力。通过使用“思维链”提示等技术，引导大型语言模型进行详细的中间推理步骤。模型能够生成自己的推理链并提供更准确的答案。逐步推理已被证明能够显著提高大型语言模型在数学问题和常识推理等任务上的性能。

知识承载。大语言模型能够从大量的文本数据中学习语言知识，并将这些知识储存在庞大的参数空间中，使得它们能够在执行任务时利用这些知识，从而实现信息浓缩和知识承载。例如，在文本总结任务中，大语言模型通过信息浓缩能够从一篇长文中提取出关键信息以简洁的方式呈现给用户。在用户对话过程中，大语言模型能够将训练过程中提取的知识以自然语言的形式表达出来。

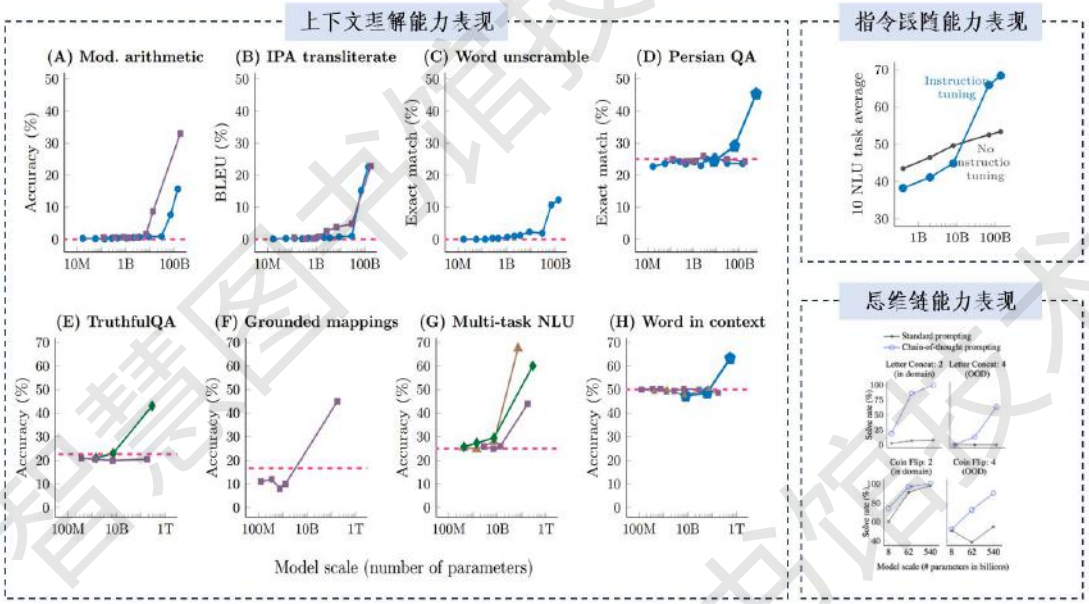

图 2：大模型涌现能力表现

（3）大模型的局限性

大模型在具备上述核心能力和涌现能力的同时，也存在一些局限性，主要体现在：

大模型具有“幻觉”，可能会胡编乱造，产生虚假不良信息。训练数据有偏、理解推理不足、监督训练误导、细分领域知识有限等原因，可能会导致模型捏造没有事实来源的答案，或给出具有偏向性的观点。因此，在对信息准确性和可靠性要求严格、容错率低的领域中，如医疗、金融等，企业应审慎防范信息偏误带来的高风险。倾向于在没有警告的情况下产生错误，包括数学、编程、归因和更高层次的概念错误。这样的错误通常被称为幻觉，因为它们倾向于显得合理或与真实的推论相一致。幻觉，如错误的参考文献、内容和陈述，可能与正确的信息交织在一起，并以一种有说服力和自信的方式呈现，使得它们在没有仔细检查和努力检查事实的情况下难以被识别。

大模型与人类价值观不够对齐，可解释性不足，可能会存在偏见或歧视。模型在响应前的分析步骤具有“黑箱”性质，呈现不透明、不可解释性。生成式语言模型的底层输出逻辑是推测句子中最有可能出现的下一个单词进行“填空”，而随着数十亿甚至千亿级别参数大模型的出现，运算过程变得十分复杂并且难以解释，最终导致模型决策行为难以评估并施加控制。

大模型动态学习、知识更新能力不足。大模型的实时更新问题也在一定程度上限制了生成式人工智能的应用扩展。大语言模型的“智力”依赖于大型数据集和高性能算力，而数据集不具备自我更新的机制，因此模型的升级需要更新训练数据集，呈现阶段性和滞后性特征。例如，GPT-4 的知识库更新截止 2021 年 9 月，后续信息无法被用于学习，可能出现推理错误的情况。

大模型系统规划能力不足、逻辑推理并不严密。大模型在系统规划和逻辑推理上仍然有所不足。原因是大模型通常是通过在大量数据上进行训练来学习知识的，它的输出更多地是基于数据中的模式，而不是基于严格的逻辑规则。因此，当面对需要高度逻辑推理

的问题时，大模型可能不能给出完全正确或者最优的答案。此外，它在处理复杂的、多步骤的问题时，可能无法进行深入的规划或预测未来的行动。

大模型技术稳定性不足。尽管大模型的规模巨大、参数众多，但它们在某些特定场景或任务上可能仍会出现不稳定的行为。例如，在某些输入下，模型的响应可能是不可预测的，或者与训练数据中的模式不一致。这种不稳定性可能源于训练数据的不足、模型架构的限制或其他未知因素。这意味着，在某些应用中，特别是高风险或需要高可靠性的场景下，依赖大模型可能存在风险。

大模型的“记忆力”有限，可能会出现灾难性遗忘。当大模型在学习新的信息或任务时，它可能会遗忘之前学到的知识。这是由于模型在训练过程中需要不断地调整其内部参数，以适应新的数据和任务，这可能导致它“忘记”之前的知识。这种遗忘现象对于需要长时间、持续学习的应用，如生涯学习或某些类型的在线学习，可能会成为一个严重的问题。

3、大模型技术为智慧图书馆发展提供新机遇

具体而言，大模型技术对于智慧图书馆的赋能重点体现在以下方面：

信息理解和分析。大模型技术有助于智慧图书馆进行信息理解和分析。大语言模型能够快速、准确地理解用户输入的自然语言，理解用户需求和意图，从而更加精准地为用户提供个性化的信息资源服务。此外，图书馆需要处理大量的文本信息，包括图书、期刊、报纸等各类文献资料，大模型技术能够对这些内部信息资源进行深入分析和挖掘，提取关键信息发掘其中蕴含的知识价值，将其整合到知识库中，并将其呈现给读者，为读者提供符合其需求的高价值、高质量知识服务，大幅提高用户信息资源检索查询和知识加工的效率。

信息检索和推荐。大语言模型技术有助于智慧图书馆进行信息检索匹配。大语言模型能够根据用户输入的关键词或短语，快速检索出相关的信息资源，并根据相关性和重要性对检索结果进行排序，将结果以清晰易懂的方式呈现给用户，为用户提供更加精准、高效的检索服务。此外，大语言模型技术还能够根据用户的检索历史和偏好，为用户个性化推荐相关的信息资源。

信息交互和展示。大语言模型技术有助于智慧图书馆进行信息交互展示。大语言模型能够以自然语言的方式与读者进行交流，通过对读者需求的深入理解，为读者提供更加个性化、人性化的知识服务。此外，大语言模型技术还能够通过多种方式展示信息，如文字、图片、音频、视频等，为用户提供更加生动灵活的阅读体验，通过更生动、可视化的信息展示方式呈现信息和知识，从而提升交互效果，使用户能够更直观地理解和获取知识。

总之，大语言模型技术作为一种先进的人工智能技术，能够为智慧图书馆建设提供新型的技术路径和赋能支撑。它能够帮助智慧图书馆更好地整合和加工信息资源、分析和挖掘文本信息，并为读者提供更加高价值、人性化的知识服务。随着人工智能技术的不断发展，大语言模型技术在智慧图书馆建设中将发挥越来越重要的作用。

二、大模型赋能智慧图书馆全景视图

（一）大模型技术赋能智慧图书馆路径

综合来看，智慧图书馆面向读者服务、业务助手以及未来体验的智能化需求分别在功能和性能方面具有不同的特点，可以归纳为不同的需求类型，即体验功能需求、专业功能需求、体验性能需求和专业性能需求。结合大模型技术在自然语言处理方面的优势和特点，以及大模型经过微调后表现出专业性和稳定性提升的潜力优势，大模型技术基于不同的技术实现路径可以有效匹配和满足不同的需求类型，即在赋能智慧图书馆业务方面主要呈现四种路径：

大模型结合自身技术优势满足体验类功能需求。图书馆智能化应用场景对于人机交互、内容生成、多轮对话等交互类、体验类功能有着显著需求，例如读者基于 AIGC 的内容创作、面向读者的咨询解答、面向馆员的办公助手等智能化应用，需要智能化系统提供内容生成、自然语言交互、通识问题解答等类 ChatGPT 功能。对于这类体验类功能需求，**大语言模型基于自然语言处理方面的技术优势以及涌现出的上下文学习、多轮对话、思维链推理能力可以有效的技术支撑，精准匹配自然语言对话、内容生成等功能需求，**为智慧图书馆交互服务类应用场景提供赋能支撑。

大模型经过微调训练或调用外部系统满足专业类功能需求。图书馆智能化应用场景对于内容存储、专业问答、资源检索和更新、资源推荐等专业化、个性化服务功能有着显著需求，例如读者阅读助手可能结合读者的借阅历史等数据个性化推荐新书名目，同样可能基于读者阅读助手实现信息资源的精准检索，面向读者的交互问答可能需要满足图书馆业务方面的问答需求，面向读者、馆员的专业知识服务也需要智能化系统具备精准匹配、归纳总结、知识推理等专业服务能力。对于这类专业类功能需求，**大语言模型通过专业数据微调训练提升模型的专业问答和推理求解能力，或对接外部系**

统（例如搜索引擎、用户管理系统、知识库系统）提升大模型的信息检索、匹配以及存储记忆能力，可以有效满足和匹配个性化推荐、专业知识问答、知识分析服务等专业功能需求，为智慧图书馆的高价值信息内容服务提供有效的技术支撑。

大模型结合自身技术优势满足体验类性能需求。从性能需求来看，部分图书馆智能化应用场景的性能需求主要侧重互动性、体验感和创造性等方面，例如面向读者的咨询问答、人机交互，以及基于文生图等多模态的新型阅读体验，智慧图书馆空间元宇宙虚实交互体验等场景更加看重人机交互的流畅性、互动性，内容生成的丰富性、创意性，以实现更好的体验和交互效果。对于这类体验性能需求，**大语言模型基于自然语言处理方面的技术优势以及涌现出的上下文学习、多轮对话、思维链推理能力能够在智能交互过程提供流畅自然且有创意性的互动体验，从而满足智慧图书馆交互服务类应用场景的体验性能需求，**为智慧图书馆交互服务类应用场景尤其是面向读者的智能阅读服务提供赋能支撑。

大模型经过微调训练或调用外部系统满足专业类性能需求。在部分图书馆智能化应用场景具有体验性能需求的同时，部分应用场景的性能需求主要侧重内容准确、稳定可靠等方面，例如面向读者的信息资源检索查询、图书馆业务知识问答、专业知识服务，以及面向馆员的图书辅助编目、情报研究和人文研究辅助等应用场景更加看重智能化系统的稳定性、可靠性，以及在内容生成方面准确可信、有据可查，杜绝虚假信息的产生。对于这类专业性能需求，**大语言模型通过专业数据微调训练，或者通过对接外部系统等方式增强模型自身的稳定性、可靠性，以及内容的准确性和真实性，能够保障其在智能交互过程中满足图书馆高价值信息服务以及智能业务辅助等应用场景的专业性能需求，**从而为专业服务、业务辅助等应用场景提供专业可靠的性能保障。

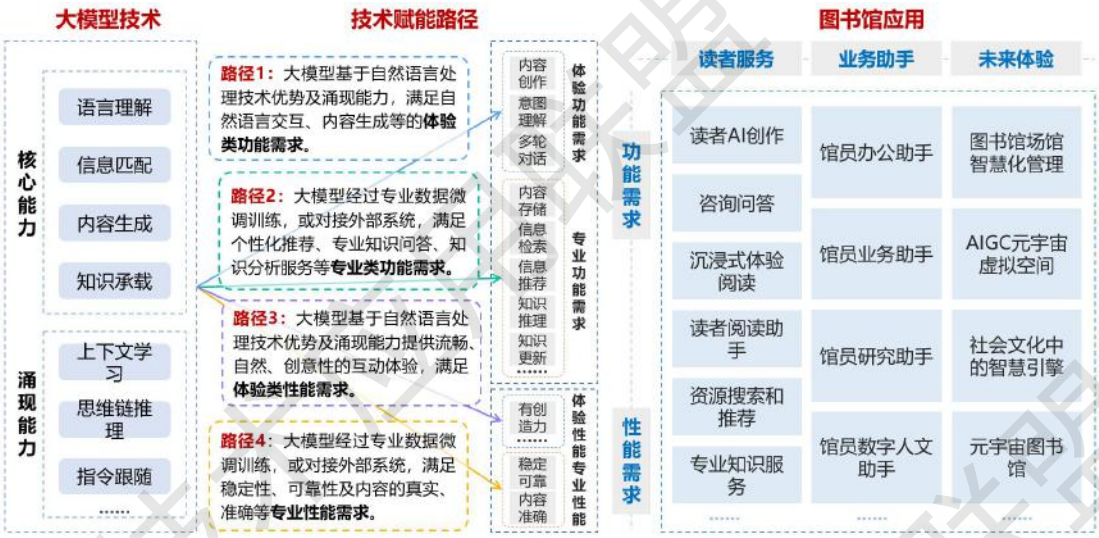

图 3：大模型技术赋能智慧图书馆路径

（二）智慧图书馆领域大模型应用总体架构

智慧图书馆领域的大模型应用总体架构主要包括两类，分别为“智慧图书馆+大模型”“大模型+智慧图书馆”。

“智慧图书馆+大模型”。“智慧图书馆+大模型”即围绕图书馆业务嵌入大模型技术。“智慧图书馆+大模型”以图书馆业务场景为核心，面向图书馆馆员业务打造大模型及大模型应用场景，通过接入图书馆大模型的智能化能力以及场景模型的专业化服务能力提升馆员的业务管理和办公效率。围绕馆员的日常办公、图书推荐、图书采编业务、情报科研、数字人文研究等业务的智能化、个性化需求，依托由行业内已发布的成熟的大语言模型如 GPT-4、“文心一言”、或者基于基础模型微调训练的智慧图书馆大语言模型、以及在智慧图书馆大语言模型基础上经过模型调优、量化等处理的图书馆细分业务场景模型，例如文献服务模型、人文研究模型、图书采编模型等所构成的模型体系，通过插件安装或 API 接口调用等多种方式将大模型能力接入到图书馆业务系统、办公系统等业务场景中，对现有业务系统进行智能化升级改造，从而为馆员日常工作和业务提供智能化辅助，提高馆员业务办理的效率。

“大模型+智慧图书馆”。“大模型+智慧图书馆”即以大模型为

核心衍生服务应用。“大模型+智慧图书馆”以智慧图书馆大语言模型为核心，通过调用知识库等外部系统提高智能代理能力以及图书馆领域任务的认知推理、问答能力，面向读者用户打造统一服务入口，提供咨询问答、检索推荐、新型阅读等多元服务。在开源的预训练大语言模型的基础上，结合图书馆业务领域的训练数据对模型进行训练微调，得到图书馆领域的大语言模型，并以智慧图书馆大语言模型作为决策智能中枢，通过对接图书馆已有的业务系统或外部信息系统，例如搜索引擎、用户管理系统、翻译系统、预约系统、知识库等，扩展大模型对于问题或任务的认知理解、决策推理求解、内容生成以及可视化呈现、智能交互、任务自动执行的闭环能力，同时也弥补大语言模型在记忆能力、专业能力、稳定性方面的缺陷和不足，以智慧图书馆大语言模型为智能中枢打造具备智能客服、检索推荐、新型阅读、阅读助手及知识服务等功能的新一代智慧图书馆服务平台。

智慧图书馆大模型应用将逐渐由“+大模型”向“大模型+”范式转变。在大模型落地应用初期，“+大模型”是智慧图书馆大模型主要的应用范式，智慧图书馆大模型应用将以图书馆业务场景为核心，以业务需求为驱动使用大模型技术，或使用插件、或调用API接口，在延续现有业务形态的基础上通过使用大模型技术促进图书馆业务管理和服务降本增效、提升智能化程度、优化服务体验。随着大模型技术不断创新发展，并在图书馆领域数据得到充分的训练和学习，“大模型+”成为主要的应用范式，大模型将逐渐成为智慧图书馆大模型应用的核心驱动因素，依托大模型承载的海量知识以及表现出的核心能力、涌现能力，并通过对接外部系统提高智能代理和任务执行能力，大模型将重构和颠覆现有的图书馆业务和服务形态，打造面向未来的一体化、个性化、泛在可及的新一代智慧图书馆服务。

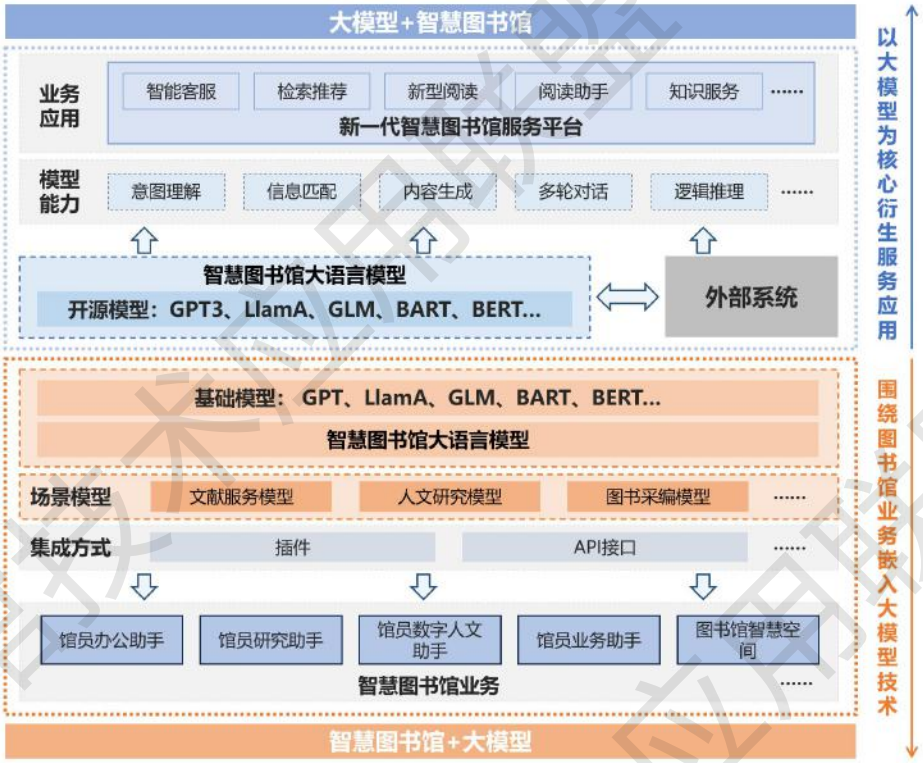

图 4：智慧图书馆领域大模型总体架构

（三）智慧图书馆大模型应用场景全景视图

智慧图书馆大模型应用场景主要体现在四个方面，智慧管理、智慧服务、智慧业务、智慧空间。

智慧管理。大模型能够对图书馆馆藏资源进行语义化管理和深度分析挖掘，充分释放图书馆信息资源的潜在价值。大模型在智慧管理方面的应用主要包括智慧数据、智慧馆藏、智慧化知识组织等。如智慧馆藏基于大模型技术对图书馆馆藏资源进行语义识别、理解，通过生成知识图谱、数据标签、摘要简介等方式精准地描述和管理信息资源并实现信息资源的交叉融合，辅助不同主题下图书馆馆藏资源的开发利用和知识服务。智慧化知识组织可提高知识组织整合能力，利用大模型技术自动识别和标注资源，在知识图谱中生成新的节点和关系，从而减轻构建知识图谱的工作量。

智慧服务。智慧服务基于大模型技术面向读者提供更具人性化、智能化的阅读体验服务，由大模型作为读者用户接入到图书馆服务

资源的统一入口，可提供咨询问答、空间预约、图书检索与个性化推荐等交互服务以及多模态、沉浸式的新型阅读体验，同时也可基于自然语言处理能力为读者提供智慧情报、研究助手等方面的专业服务。例如打造智慧图书馆服务平台为用户提供咨询问答、线下预约、检索查询等便捷服务，或根据读者的阅读需求和阅读记录等数据，准确理解读者的个性化需求并推荐图书资源，提供有情感的、专业的个性化服务，或基于自然语言交互和AIGC技术为读者提供图书领航、元宇宙创意阅读等沉浸式新型阅读体验。

智慧业务。智慧业务服务对象主要为图书馆馆员，通过运用大模型技术提高馆员的日常工作和业务办理效能。大模型在馆员业务中的应用主要包括日常办公、业务辅助、科研辅助以及人文研究辅助四个方面。日常办公方面，大模型作为插件接入到办公软件中提高馆员办公效率；业务辅助方面，大模型可用于图书的采购推荐、辅助编目等业务提供智能化辅助；科研辅助方面，大模型可为馆员的情报研究、学术研究提供翻译、摘要、概括等方面的自然语言处理能力；人文研究方面，大模型也可基于多模态的识别和处理能力以及自然语言处理能力提供信息检索、匹配、OCR识别等辅助支撑。

智慧空间。大模型技术在图书馆建设智慧空间方面的应用主要体现在基于AIGC能力快速高效的构建元宇宙图书馆，以及基于元宇宙图书馆的虚实交互体验。典型的大模型应用场景包括：基于AIGC技术构建图书馆元宇宙虚拟空间；基于大模型的自然语言交互能力并结合数字孪生图书馆构建元宇宙图书馆虚实交互体验，例如基于AR的图书馆智慧导航以及人机交互服务，帮助读者用户在图书馆内找到所需的资源或服务；或利用大模型作为人机交互入口，与读者进行自然语言交互并通过AIGC生成数字内容呈现元宇宙图书馆虚实融合的交互体验，或打造虚拟数字人馆员与读者进行自然语言对话交互等。

智慧管理、智慧服务、智慧业务、智慧空间这些基于大模型的智慧图书馆应用场景都可以依托下一代图书馆服务平台（如云瀚平台）来承载。云瀚平台是基于开源技术和云原生架构研发的下一代图书馆服务平台，同时也是一套可以替代传统图书馆集成管理系统的应用组件。云瀚平台的开放接口、微服务模式和模块化的结构，使其能够与其他系统和应用无缝集成，并具备灵活的可扩展性。依托云瀚平台，图书馆可以根据自己的特定需求选择和整合各种应用和服务，智慧阅读推广、数字资源管理等不同的应用场景，都可以基于云瀚平台的微服务架构和开放接口进行研发部署和集成对接，从而将各种应用系统和场景服务统一地接入并整合到云瀚平台中，实现智慧图书馆的多功能和多场景服务，面向读者提供更加丰富和个性化、智能化的服务体验。

智慧图书馆大模型应用场景将呈现由内部管理、对外服务过渡到未来体验的发展路径。考虑到大语言模型具有幻觉、或价值对齐不足，仍可能会产生不负责任的输出，以及智慧图书馆大语言模型需要经过一定时间的训练调优过程，而且图书馆作为公共文化服务机构需要针对提供给读者的内容的事实性、准确性、安全性进行严格把关，因此，**从短期来看，**智慧图书馆大模型应用场景建设可能以图书馆内部馆藏资源的语义化管理和分析挖掘即智慧管理以及面向馆员业务提供智能化辅助的智慧业务为切入点，通过应用大模型技术提升馆员对于图书馆信息资源的分析挖掘、知识组织管理能力，提高馆员的业务办理和工作效率，并由馆员作为智慧图书馆大模型面向读者提供智能化、个性化阅读服务的把关、审核中间环节。**从中期来看，**随着智慧图书馆大语言模型经过图书馆领域数据的微调，并通过对接知识库等外部系统以实现更加专业、准确的交互效果，提升图书馆馆藏资源的语义理解、信息匹配组织、内容生成等 NLP 能力，智慧图书馆大模型应用场景将向提供智能化阅读体验以及知

识服务的智慧服务领域拓展，即利用大模型技术同读者进行自然语言交互，提供咨询解答、查询预约、资源检索与个性化推荐等阅读助手服务，以及定制主题的文献、情报等高价值知识服务，同时可以面向读者提供虚拟场景动态呈现、图书游览等新型阅读体验。**从长期来看，**随着沉浸式交互、全息通信、扩展现实、元宇宙等技术进一步发展和成熟，以及大模型技术与元宇宙的无缝集成应用，智慧图书馆大模型应用场景将不仅仅停留在线上 web 端、移动端，也将向线上线下虚实融合的元宇宙体验场景拓展延伸，届时智慧图书馆大模型将呈现更多智慧空间应用场景，大模型将成为读者与图书馆知识海洋互动的智能入口，带领读者在由 AIGC 等技术打造的元宇宙图书馆中进行沉浸式的知识探索、体验、分享与创造，从而营造极具未来感的知识探索体验。

图 5：智慧图书馆大模型应用场景全景视图

三、智慧图书馆大模型应用场景创新路径

智慧图书馆通过大模型等技术的创新应用，对图书馆的管理、数据、服务、空间、馆藏、馆员等方面进行优化和创新，以实现“知识作为一种服务”，将图书馆的服务水平提升到智慧层次，更好地满足读者的需求，提高服务质量，实现智慧化管理和服务。在智慧管理方面，大模型能够帮助图书馆实现对藏书、数据等资源的语义化、精细化管理。在智慧业务方面，大模型能够支持图书馆开展智能推荐、智能检索等业务。在智慧服务方面，大模型可为读者提供个性化阅读推荐、在线咨询等服务。在智慧空间方面，大模型能够帮助图书馆实现空间布局优化、环境监控等功能。本章将具体介绍大模型在智慧图书馆领域的创新应用场景，为智慧图书馆大模型应用场景建设提供借鉴和参考。

（一）智慧管理实现图书馆资源的语义化管理

智慧管理实现对图书馆馆藏、数据等要素的语义化、智慧化管理。通过运用大语言模型技术对图书馆海量优质的馆藏、数据等资源进行细粒度的语义识别、理解和分析，并自动生成知识图谱、数据标签实现对馆藏资源的结构化、语义化表示，有效支撑不同主题、场景的图书馆资源开发利用，可通过数据分析挖掘、文本生成等实现数字资源的深度挖掘和高效利用，为主题馆、特色馆等场景提供高质量的数据资源和知识辅助。

1、智慧馆藏

智慧馆藏依托大模型技术，实现对图书馆馆藏资源的智慧化、精细化管理和深度分析挖掘，提高图书馆的馆藏质量和丰富度。智慧馆藏可以实现纸质资源的数字化转换和保存，以及精选、采购、编目、流通等环节的自动化、智能化辅助。此外，智慧馆藏还可以实现数字馆藏资源的深度挖掘和高效利用。通过基于自然语言处理的结构化、语义化的数据表示，实现对数字馆藏资源的语义理解和

知识挖掘，有利于图书馆馆藏资源的开发利用和知识服务。

2、智慧数据

智慧数据依托大模型技术对海量的全媒体信息资源进行深度的自然语言处理，充分挖掘图书馆的数据价值。智慧数据依托大模型技术实现对图书馆的数据资源的语义理解、分析和加工处理，将海量优质的全媒体信息资源提炼出高质量、高价值的知识，提高图书馆的数据价值和利用率，并支持以自然语言交互等人性化的方式满足读者多样化、泛在的阅读需求和知识获取需求，有利于提高图书馆的数据影响力和价值，促进读者的知识创新和传播。智慧数据的应用可以包括通过人工智能辅助生成知识图谱，实现对数字资源的结构化和语义化表示，实现对数字资源的深度挖掘和高效利用等。

3、智慧化知识组织

图书馆通过知识图谱和大模型技术的结合，在提升大模型性能的同时提高图书馆的知识组织能力。图书馆长期致力于知识的组织和表示，具有丰富的经验和技术积累。知识组织技术也从机读目录发展到关系数据库，再到关联开放数据和提供知识发现和推理功能的知识图谱。知识图谱和大模型在图书馆领域的应用中具有显著的互补性。图书馆拥有丰富的、经过验证的知识资源，这些资源可以用来构建知识图谱，进一步提升大模型的性能和可靠性。知识图谱可以为大模型提供结构化、关系丰富的知识，当大模型与知识图谱结合时，可以实现对幻觉问题的有效缓解，为用户提供更准确、更可信的服务。大模型也可以在训练过程中直接访问知识图谱，以获取额外的背景知识或上下文信息，从而增强模型的推理能力。而大模型的技术能力也可以提高图书馆的知识组织能力，可以自动识别和标注资源，在知识图谱中生成新的节点和关系，并通过预先训练的知识库进行对齐和验证，从而减轻构建知识图谱的工作量。

（二）智慧业务打造辅助图情业务的智能助手

智慧业务依托大模型技术面向馆员的日常工作和业务提供智能化辅助，以智慧化手段提高馆员信息素养和能力。例如通过自然语言处理技术，帮助馆员快速撰写报告、发送邮件等，提高工作效率。同时，大模型技术通过分析用户需求、阅读习惯和借阅记录，为馆员提供更精准的图书推荐，帮助馆员更好地选择和采购图书。此外，大模型技术也能够提供文献检索、智能阅读、辅助编辑等功能，帮助研究人员快速阅读学术文献、获取所需信息，更高效地开展研究工作。总之，大模型技术在辅助馆员业务方面具有巨大的潜力和应用价值。

1、智慧馆员助手

智慧馆员助手依托大模型技术为图书馆馆员的办公、图书采编及学术研究等方面提供智能化的辅助，提高图书馆馆员的信息素养和工作能力。在日常办公方面，大模型技术可以通过自然语言处理技术，帮助馆员快速撰写报告、发送邮件等，帮助馆员更快速地处理文书工作，更有效地管理日常事务。在图书采编等业务方面，大模型技术通过分析用户需求、阅读习惯和借阅记录，为馆员提供更精准的图书推荐，帮助馆员更好地选择和采购图书。在学术研究方面，大模型技术通过提供文献检索、智能阅读、辅助编辑等功能，帮助研究人员快速阅读学术文献、获取所需信息，帮助馆员更有效地开展研究工作。

2、智能化资源采购

智能化资源采购通过运用大模型技术帮助馆员更好地利用采购资金，采购更符合读者阅读需求的资源。馆员可以用自然语言与大模型进行交互，结合本地和云端的多个采访数据源，大模型通过自然语言处理技术来分析图书内容，帮助图书馆工作人员快速了解图书的主题和风格，从而更好地进行图书采购决策。同时，大模型根

据图书馆的采购历史数据、读者阅读行为，提供当前的采购建议。此外，大模型能够回顾历年采购数据和规则，制定和调整采购评估标准，帮助馆员更科学地构建资源架构，提高采购资金的使用效率。

3、智能化阅读推广

智能化阅读推广基于大模型技术显著提高阅读推广的精准性、科学性以及工作效率。首先，图书馆可以利用各类活动管理平台和宣传推广平台的业务数据，对用户进行画像分析，更加精准地了解读者的需求和喜好，从而提供更加精准的阅读推荐服务。其次，大模型能够基于多模态对齐技术将多种数据类型的读者需求、偏好数据统一使用自然语言进行表达和描述，再结合语义理解和检索匹配，精准推荐图书馆已有的资源，从而提高阅读推广的科学性和精准性。此外，图书馆也可运用 AIGC 等技术，创造出“数智人”和“数字馆员”等新的服务形式，自动生成与图书内容相关的图片、视频等文案资源，极大提高阅读推广的效率，并为读者提供更加便捷、高效、智能化的阅读体验。

4、智慧化数字人文研究和服务

大模型技术可助力数字人文研究进行细粒度、语义化的知识提取和挖掘。首先，在数据处理方面，大模型技术能够进行图像处理、文本处理、知识库构建和实体提取等工作，可辅助生成和完善知识图谱，为数字人文研究提供更加丰富、全面的数据支持。其次，在数据分析和读者服务方面，大模型技术可通过文本分析、情感分析等技术面向读者提供文风介绍、情感和主题词等阅读服务，帮助读者更好地了解数字人文研究成果，为用户提供更加优质的知识服务。

（三）智慧服务提供新型阅读体验和知识服务

智慧服务依托大模型技术面向读者提供智能、便捷、人性化、个性化的新型阅读体验和高价值的知识服务，通过自然语言交互对读者咨询问题进行精准回答和引导，并基于语义理解和信息匹配等

技术根据读者需求对图书馆资源进行精准检索和匹配，从而满足读者对阅读体验和学术研究的多样化方式和内容需求，提升信息检索能力和效率，增强读者的阅读体验和满意度，并丰富读者的文化生活。

1、智慧图书馆服务平台

大模型技术在智慧图书馆服务平台中可应用于读者的咨询问答、线上预约、检索查询等业务领域，优化阅读服务体验。大模型技术可以用于咨询问答匹配，提供人机问答服务，可通过语义理解准确理解和识别用户意图后结合馆内数据库（书目检索系统、馆藏管理系统、读者证管理系统、预约系统等）检索进行精准匹配，快速准确地给出答案，提高读者的满意度。大模型技术也可为图书馆空间、资源、服务等线上预约提供导引支持，通过与读者进行自然语言交互并利用大模型的语义分析能力，准确理解用户意图，并根据读者语言意图、结合用户历史记录推荐相应的预约方案，向读者推荐最有意向预约的时间/预约座位、读者意图查询的最相关的结果等，最终通过对接预约系统自动化执行预约操作，为用户提供智能便捷的线上预约服务。大模型也可优化阅读推广机器人的交互体验，提高参考咨询服务的人性化、智能化程度，通过语义分析理解用户意图，结合用户借阅数据、馆区停留数据、馆所资源数据，综合上下文改进机器人的回复效果，并提供有情感的、实时专业的个性化服务，如根据读者历史行为信息，推荐符合的图书、音乐等。

2、智慧化学术资源检索平台

大模型技术有利于提高学术资源检索效率和匹配精准性，可提供学术资源的检索匹配、辅助阅读、知识服务等功能。一方面，它可以基于现有的检索平台和学术资源数据库，为科研机构及相关工作人员提供学术信息资源的检索、语义搜索、语义分析和成果评价等功能，使研究人员能够更有效地获取所需的学术信息，从而推动

学术研究的进展。另一方面，大模型技术还可以支持学术论文的自然语言交互，通过生成摘要、提炼总结等辅助功能使得研究人员更轻松地理解论文内容，有助于弥补语言障碍，使全球范围内的研究人员都能够从学术文献中获取知识。

3、元宇宙沉浸式知识体验

图书馆可以利用大模型、VR/AR 和元宇宙等技术，构建丰富的、沉浸式的知识体验环境。每本书、每个故事、每个知识领域都可以基于 AIGC 技术构建一个独立的元宇宙，用户可以在其中自由地探索、交互和学习。大模型所扮演的虚拟数字人也将成为用户的领航员，和读者进行生动、自然的对话和交互，带领读者在元宇宙空间中穿梭和体验。这不仅仅是一个虚拟的数字空间，而是一个真实与虚拟结合的、知识交流和互动的、充满智慧和创意的新世界。在这个世界中，知识不再是静态的、孤立的，而是动态的、相互连接的。用户可以自由地探索、创造和分享知识，体验无限的可能性。

（四）智慧空间打造虚实融合的智能交互体验

智慧空间运用大模型技术优化图书馆空间和读者的交互体验，通过接入 APP、AR 眼镜等智能终端，或打造虚拟数字人客服等多种方式将大模型接入图书馆与读者互动的交互入口，依托大模型技术的自然语言交互和对话能力，优化读者与空间之间的人机交互、虚实互动体验，打造自然流畅、内容丰富、智能便捷的空间互动体验，面向读者提供泛在、智能的空间服务和数字文化服务，提高读者的体验感和满意度。同时也可运用 AIGC 等技术赋能数字孪生图书馆的建设。

1、图书馆智慧导航服务

大模型技术可发挥自然语言对话、多模态交互能力赋能智慧图书馆提供空间导航导引及咨询交互服务。通过将大模型接入 APP、AR 眼镜等智能终端带动人机交互入口的智能化升级，读者可以通过

语音或文本等多模态输入与大模型进行交互，询问图书馆内部的位置信息，例如某本书的位置、某个阅览室的位置等，大模型通过自然语言理解分析，根据读者的需求，以自然语言交互方式提供精确的室内导航信息，或通过对接 AR 系统进行内容的增强显示，帮助读者快速找到目标位置。此外，大模型还可以根据读者的需求，为其提供个性化的推荐服务。例如，如果读者询问某个主题相关的书籍，大模型也可以通过资源检索提供相关书籍的位置信息，还可以根据读者的阅读喜好，为其推荐其他相关主题的书籍。图书馆智慧导航服务可为读者提供精确、便捷、个性化的室内导航服务，极大地提高了读者在图书馆内部的体验感和满意度。

2、元宇宙图书馆虚实交互体验

大模型技术可基于自然语言交互和内容自动生成的技术优势赋能智慧图书馆打造元宇宙虚实交互体验。大模型可以接入 AR 眼镜、移动端等流量入口并促进其智能化升级，随后读者便可以自然语言交互的方式同元宇宙图书馆进行互动交流，虚实融合的互动性和体验感得到大幅提升。此外，大模型技术帮助元宇宙图书馆的智能交互系统更加精准地理解读者意图，并且能够识别多模态的数据输入，结合大模型技术的语义分析理解和内容生成给出精准、高质量的回答，同时可根据读者的个性化需求、特点使用 AIGC 技术自动生成新颖生动、有创意的数字内容，随后叠加在交互终端上增强显示，从而在整个过程可以大幅提升元宇宙图书馆的虚实交互体验，提高读者的体验感。

四、智慧图书馆大模型应用落地实践路径

（一）智慧图书馆大模型创新应用规划设计

智慧图书馆建设不能一哄而上，需要有清晰明确的战略规划和行动方案。在开展大模型创新应用之前，图书馆需要进行统筹规划，通过现状调研、需求分析以及总体设计，明确图书馆应用大模型技术的战略目标、应用场景、技术路径、数据资源、基础设施和建设运营模式、进度计划等内容，并系统谋划组织架构、人员培训、标准规范、安全管理等方面的配套保障体系。

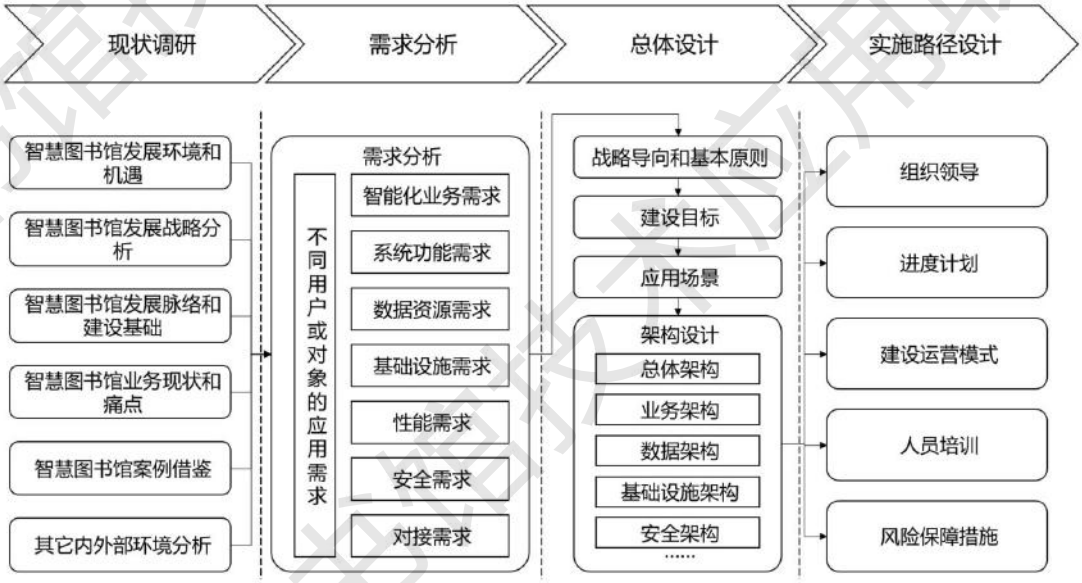

图 6：智慧图书馆大模型应用场景建设统筹规划框架

1、现状调研

在现状调研阶段，图书馆可通过问卷调查、座谈交流、实地访谈、资料分析等多种调研方式分析解读图书馆数字化发展面临的形势和机遇，梳理图书馆以往的发展战略和脉络，并摸排图书馆信息化建设基础，分析图书馆在公共服务、运营管理等方面的业务现状和痛点，总结智慧图书馆建设面临的挑战，同时也可参考借鉴其他图书馆应用大模型技术、建设智慧图书馆的典型案例，以系统全面总结智慧图书馆应用大模型技术的内外部环境。

2、需求分析

在需求分析阶段，图书馆需要在现状调研的基础上，从用户角度出发梳理读者、馆员、空间等不同用户或对象对于大模型技术的应用需求，总结图书馆应用大模型技术的智能化业务需求，进而分析系统功能需求、信息资源需求、信息共享和业务协同需求、基础设施需求，并明确系统性能需求、安全需求和接口需求等。

3、总体设计

在总体设计阶段，图书馆需要结合现状调研和需求分析，对图书馆应用大模型技术的战略导向、建设目标、基本原则、应用场景、总体架构、技术路径、数据架构、基础设施架构以及安全体系、标准体系等方面进行一体化统筹规划设计。其中，本白皮书针对战略导向、基本原则和应用场景提出如下指南建议。

（1）战略导向

图书馆需要坚持数字化、智慧化转型战略，转变以信息储存为主导的管理思维，主动顺应数字经济环境下知识服务生态的新发展，创新服务方式，通过运用数字化、智能化技术挖掘信息资源价值，产生价值增值效益，建立开放共享、互联互通的知识服务体系，打造读者交流互动分享的开放平台，以高价值信息服务赋能地区经济、社会和文化高质量发展。

（2）基本原则

图书馆应注重把握因地制宜、体现特色的原则，不宜千篇一律地盲目开展图书馆大模型开发建设，应结合自身优势和特色，尤其是本地馆藏资源或独特服务，应当充分利用大模型技术发挥和挖掘图书馆独特的资源价值；此外，图书馆应注重以人为本，推动以资源为中心的管理理念向以用户为中心的服务理念转变，关注用户体验，并从用户体验角度出发优化业务流程，面向读者提供高价值、个性化的信息内容服务。

（3）应用场景

图书馆可参考图5作为智慧图书馆大模型应用场景的全景视图，结合实际的业务需求、技术成熟度、资金预算规模等因素综合考虑，明确短期内大模型技术应用场景的切入点或建设重点。之后，图书馆应制定大模型应用场景建设的路线图，滚动推进大模型应用场景建设，不断拓宽大模型技术在图书馆的应用领域。从大模型技术特点来看，短期内，面向读者的智能客服、资源检索和推荐、阅读推广机器人、新型阅读、读者AI创作等创意类、体验类、交互类应用场景适合作为智慧图书馆大模型应用场景建设的切入点。从短期来看，运用大模型技术对图书馆馆藏资源进行精细化、语义化管理和分析挖掘以提高知识组织管理能力、面向馆员工作业务提供智能化辅助这些应用场景可能更适合作为智慧图书馆领域大模型应用场景建设的切入点。

（4）架构设计

在明确大模型应用场景的基础上，图书馆应参考图4结合业务现状等因素厘清计划建设的大模型应用场景是属于“大模型+智慧图书馆”还是“智慧图书馆+大模型”，以此来确定大模型应用总体架构搭建逻辑。之后，图书馆可综合考虑应用场景的功能需求及性能需求、技术可行性、投入成本等因素，参考本白皮书第二章大模型技术赋能智慧图书馆路径分析明确大模型技术需求及实现路径，例如直接调用大模型（如GPT-4）的API接口适配业务流程，或基于开源预训练语言模型训练智慧图书馆大语言模型以提供有关功能和服务，或采用基于预训练大语言模型结合本地向量知识库的方式，如果选择训练智慧图书馆大语言模型这条路径则需要对开源预训练大语言模型进行选型（大模型选型参考第四章第二节），进而细化明确大模型应用场景的模型基础、场景模型、集成对接方式、外部对接系统、服务场景等基本要素构成及其相互关系，从而形成大模型

应用总体架构。之后，图书馆应根据大模型应用总体架构进一步梳理形成数据架构、业务架构、基础设施架构、安全架构、标准体系等。

4、实施路径设计

在实施路径设计阶段，图书馆应从落地实施角度出发，对智慧图书馆大模型应用场景建设的组织领导、进度计划、建设运营模式、人员培训、风险保障措施等进行统筹规划与设计。常见的建设运营模式如下表。

表 1：智慧图书馆大模型应用场景建设运营模式

建设主体	运营主体	技术提供方	建设运营模式	详情	优势	劣势
图书馆	图书馆	图书馆	**图书馆自行投资建设和运营**	由图书馆全程主导项目的规划、立项、投资、建设和运营，**图书馆自行承担模型开发训练、应用场景建设等建设内容并承担项目长效运营**（算力资源可能采购第三方的云服务）；**由财政资金提供资金保障。**	1）图书馆掌握项目主导权； 2）无数据安全风险。	1）**资金压力较大**； 2）项目建设运营专业性较强、技术难度较高，**技术、人才方面比较缺乏**； 3）适应技术发展、个性化服务方面稍显不足。
	图书馆（购买服务模式下运营主体为企业）	企业	**图书馆投资建设，企业提供技术支持或企业运营**	由**图书馆负责统筹规划和指导**，通过招标采购引入社会资本，委托**中标企业承担图书馆大模型的落地建设和运营**；**由财政资金支持**	1）图书馆掌握项目主导权（但存在一定的外购/外包管理风险）； 2）引入专业技术团队，专业技术实力和人才方面有保障。	1）**资金压力较大**； 2）合作企业变更后容易出现系统或数据对接困难、**推倒重来、重复投入的情况**； 3）图书馆采购服务模式下可能存在一定的数据安全风险。
图书馆、合作企业	图书馆、合作企业	图书馆、合作企业	**图书馆和企业联合投资建设和运营**	图书馆与相关企业达成合作协议，**由图书馆与合作企业联合开**	1）可发挥双方优势，综合管理和服务能力较强，较好的专业和	1）项目管理存在一定风险，协调工作量大，统筹管理能力要

				展项目的建设运营，双方基于各自优势和角色定位开展分工合作，共同承担图书馆大模型应用场景的建设运营，实现资源互补、合作共赢。	技术实力； 2）减轻资金压力，具有一定的自我盈利能力，有利于可持续运营； 3）创新活力和动力较强，及时响应技术发展和用户需求变化提升服务质量	求较高； 2）可能面临一定的数据安全风险

（二）智慧图书馆大模型创新开发落地实施

结合大模型行业应用加速涌现的发展态势，为针对图书馆领域大模型应用场景建设提供较为及时的技术参考和建议，本白皮书按照基于开源预训练大语言模型微调智慧图书馆大模型以提供相关功能和服务的技术路线，针对智慧图书馆大模型的开发训练等全过程提供相关的技术参考和建议。图书馆领域大语言模型的开发步骤和主要工作可能包括以下几个阶段：

1) 模型选型：根据图书馆领域的需求和数据特点，选择合适的大语言模型。常见的大语言模型有 BERT、GPT、XLNet 等。
2) 模型调优：根据图书馆领域的数据，对选定的大语言模型进行调优，以提高模型在图书馆领域的表现。
3) 效果评估：对调优后的大语言模型进行模型效果评估，以确定模型是否达到预期效果。
4) 模型部署：将调优后的大语言模型部署到云计算环境中，以便在实际应用中使用。
5) 模型使用：在图书馆大模型使用过程中通过上下文提示、思维链提示，以及向量知识库嵌入等方式激发模型性能、提升模型使用效果，为用户提供更好的服务。

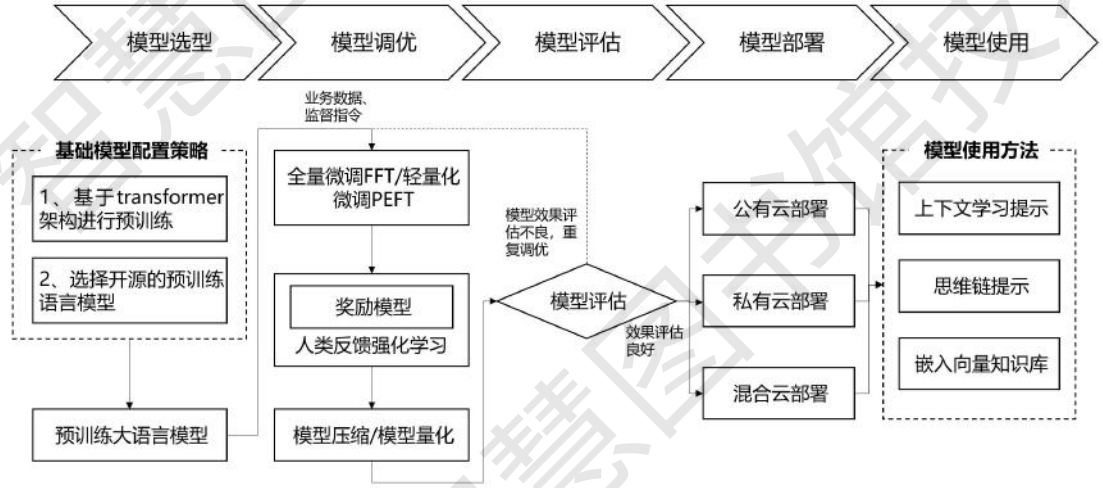

图 7：智慧图书馆大模型开发实施路径

1、模型选型

（1）大语言模型技术比较

下表列出近年来主要的大语言模型的模型大小、模型微调、性能测评、预训练数据规模、硬件资源成本、训练时间等信息。

表：主要的大语言模型技术比较[2]

TABLE 1: Statistics of large language models (having a size larger than 10B in this survey) in recent years, including the capacity evaluation, pre-training data scale (either in the number of tokens or storage size) and hardware resource costs. In this table, we only include LLMs with a public paper about the technical details. Here, "Release Time" indicates the date when the corresponding paper was officially released. "Publicly Available" means that the model checkpoints can be publicly accessible while "Closed Source" means the opposite. "Adaptation" indicates whether the model has been with subsequent fine-tuning: IT denotes instruction tuning and RLHF denotes reinforcement learning with human feedback. "Evaluation" indicates whether the model has been evaluated with corresponding abilities in their original paper: ICL denotes in-context learning and CoT denotes chain-of-thought. "*" denotes the largest publicly available version.

	Model	Release Time	Size (B)	Base Model	Adaptation IT	Adaptation RLHF	Pre-train Data Scale	Latest Data Timestamp	Hardware (GPUs / TPUs)	Training Time	Evaluation ICL	Evaluation CoT
Publicly Available	T5 [73]	Oct-2019	11	-	-	-	1T tokens	Apr-2019	1024 TPU v3	-	✓	-
	mT5 [74]	Oct-2020	13	-	-	-	1T tokens	-	-	-	✓	-
	PanGu-α [75]	Apr-2021	13*	-	-	-	1.1TB	-	2048 Ascend 910	-	✓	-
	CPM-2 [76]	Jun-2021	198	-	-	-	2.6TB	-	-	-	-	-
	T0 [28]	Oct-2021	11	T5	✓	-	-	-	512 TPU v3	27 h	✓	-
	CodeGen [77]	Mar-2022	16	-	-	-	577B tokens	-	-	-	✓	-
	GPT-NeoX-20B [78]	Apr-2022	20	-	-	-	825GB	-	96 40G A100	-	✓	-
	Tk-Instruct [79]	Apr-2022	11	T5	✓	-	-	-	256 TPU v3	4 h	✓	-
	UL2 [80]	May-2022	20	-	-	-	1T tokens	Apr-2019	512 TPU v4	-	✓	✓
	OPT [81]	May-2022	175	-	-	-	180B tokens	-	992 80G A100	-	✓	-
	NLLB [82]	Jul-2022	54.5	-	-	-	-	-	-	-	✓	-
	GLM [83]	Oct-2022	130	-	-	-	400B tokens	-	768 40G A100	60 d	✓	-
	Flan-T5 [64]	Oct-2022	11	T5	✓	-	-	-	-	-	✓	✓
	BLOOM [69]	Nov-2022	176	-	-	-	366B tokens	-	384 80G A100	105 d	✓	-
	mT0 [84]	Nov-2022	13	mT5	✓	-	-	-	-	-	✓	-
	Galactica [35]	Nov-2022	120	-	-	-	106B tokens	-	-	-	✓	✓
	BLOOMZ [84]	Nov-2022	176	BLOOM	✓	-	-	-	-	-	✓	-
	OPT-IML [85]	Dec-2022	175	OPT	✓	-	-	-	128 40G A100	-	✓	✓
	LLaMA [57]	Feb-2023	65	-	-	-	1.4T tokens	-	2048 80G A100	21 d	✓	-
	CodeGeeX [86]	Sep-2022	13	-	-	-	850B tokens	-	1536 Ascend 910	60 d	✓	-
	Pythia [87]	Apr-2023	12	-	-	-	300B tokens	-	256 40G A100	-	✓	-
Closed Source	GPT-3 [55]	May-2020	175	-	-	-	300B tokens	-	-	-	✓	-
	GShard [88]	Jun-2020	600	-	-	-	1T tokens	-	2048 TPU v3	4 d	-	-
	Codex [89]	Jul-2021	12	GPT-3	-	-	100B tokens	May-2020	-	-	✓	-
	ERNIE 3.0 [90]	Jul-2021	10	-	-	-	375B tokens	-	384 V100	-	✓	-
	Jurassic-1 [91]	Aug-2021	178	-	-	-	300B tokens	-	800 GPU	-	✓	-
	HyperCLOVA [92]	Sep-2021	82	-	-	-	300B tokens	-	1024 A100	13.4 d	✓	-
	FLAN [62]	Sep-2021	137	LaMDA-PT	✓	-	-	-	128 TPU v3	60 h	✓	-
	Yuan 1.0 [93]	Oct-2021	245	-	-	-	180B tokens	-	2128 GPU	-	✓	-
	Anthropic [94]	Dec-2021	52	-	-	-	400B tokens	-	-	-	✓	-
	WebGPT [72]	Dec-2021	175	GPT-3	-	✓	-	-	-	-	✓	-
	Gopher [59]	Dec-2021	280	-	-	-	300B tokens	-	4096 TPU v3	920 h	✓	-
	ERNIE 3.0 Titan [95]	Dec-2021	260	-	-	-	-	-	-	-	✓	-
	GLaM [96]	Dec-2021	1200	-	-	-	280B tokens	-	1024 TPU v4	574 h	✓	-
	LaMDA [63]	Jan-2022	137	-	-	-	768B tokens	-	1024 TPU v3	57.7 d	-	-
	MT-NLG [97]	Jan-2022	530	-	-	-	270B tokens	-	4480 80G A100	-	✓	-
	AlphaCode [98]	Feb-2022	41	-	-	-	967B tokens	Jul-2021	-	-	-	-
	InstructGPT [61]	Mar-2022	175	GPT-3	✓	✓	-	-	-	-	✓	-
	Chinchilla [34]	Mar-2022	70	-	-	-	1.4T tokens	-	-	-	✓	-
	PaLM [56]	Apr-2022	540	-	-	-	780B tokens	-	6144 TPU v4	-	✓	✓
	AlexaTM [99]	Aug-2022	20	-	-	-	1.3T tokens	-	128 A100	120 d	✓	✓
	Sparrow [100]	Sep-2022	70	-	-	✓	-	-	64 TPU v3	-	✓	-
	WeLM [101]	Sep-2022	10	-	-	-	300B tokens	-	128 A100 40G	24 d	✓	-
	U-PaLM [102]	Oct-2022	540	PaLM	-	-	-	-	512 TPU v4	5 d	✓	✓
	Flan-PaLM [64]	Oct-2022	540	PaLM	✓	-	-	-	512 TPU v4	37 h	✓	✓
	Flan-U-PaLM [64]	Oct-2022	540	U-PaLM	✓	-	-	-	-	-	✓	✓
	GPT-4 [46]	Mar-2023	-	-	✓	✓	-	-	-	-	✓	✓
	PanGu-Σ [103]	Mar-2023	1085	PanGu-α	-	-	329B tokens	-	512 Ascend 910	100 d	✓	-

（2）大语言模型选型思路

[2] 《A Survey of Large Language Models》

对于图书馆领域的大语言模型，不同的图书馆基于不同的用户群体和资源规模等可能有不同的需求，想要选出适合图书馆领域的最佳模型，需要从多个角度进行评估和对比。**从技术角度来看**，首先，大语言模型的数据处理能力是一个关键的考虑因素。一个好的图书馆模型应该能够灵活地存储、展示和处理文本和多媒体资源，帮助用户从大量文献中提炼整理出能帮助到用户的知识。用户感兴趣的文献资源可能是用户关心的某个专业领域的学术论文也可能是关于某个新闻事件的相关报导等，需要根据不同的数据类型使用领域自适应、本地知识库辅助的方法来定制一个垂直领域的大模型。**从用户体验来看，**系统的用户界面和易用性也非常重要，需要提供友好的操作体验和多用户高并发的支持。此外，图书馆专业咨询领域的模型需要具有较好的平衡性，即平衡大语言模型的事实性和创造性，不能造成用户的事实性误导，必要时可以引入搜索引擎的结果。**从成本角度来看**，要选择一个适合图书馆的模型还需考虑实施时间和费用方面。对于许多图书馆，选择一款开源的模型，并进行一定的领域自适应、指令微调、对接本地知识库等个性化改良可能是一个可行的策略。

具体而言，模型选型主要有以下考虑因素：

1) **功能**：包括多轮对话的能力、生成高质量文本的能力、可能的多模态能力，逻辑推理能力以及足够广泛的通用知识基础，比较好的泛化能力等功能，将是模型的应用和定制的关键要素；
2) **性能**：需考虑模型的推理速度，以及开源模型在同等硬件条件下的微调效率，收敛效率等性能指标；
3) **许可证**：指的是针对开源模型的许可证类型，比如是宽松自由软件许可协议，还是著作权许可证；他人修改源码后，是否要求还是开源，等等的考虑。许可证类型将影响到行业大

模型的深度研发和商用的模式；

4) **安全/伦理考量和可解释性**：安全/伦理考量指模型在受到攻击和恶意输入时是否能保持数据和应用安全稳定，并维持预定的道德伦理标准；可解释性是指模型能对其输出结果提供可预测，可逻辑推演，可归因的结果；

5) **工具链支持**：基于基础大模型定制行业模型，或开发 AI 应用，都会涉及到诸多工程化的问题，比如数据整理、数据标注、各种微调和插件支持、工作流集成、模型部署等等，工具链完备程度将对模型的定制和开发效率具有巨大影响；

6) **（技术和产业）生态系统**：生态系统会对模型与其他系统和服务的集成、协作产生巨大的推动作用。广泛健康的生态合作系统提供并支持标准化接口和协议，使得生态合作方可以方便快捷低成本地交互和合作，形成良性规模效益。

图书馆大模型涉及各类信息查询、数据编目、知识摘要、内容索引、情报分析、咨询交互等事务，其选型需要结合自身业务特点，综合加权考虑上述要素：**比如在学术研究、情报分析等方面如果涉及敏感数据较多，则在安全和可解释性上需要加以更多关注；如果考虑未来进行大规模业务定制、延伸和拓展，对是否开源/许可证类型的选择就需要非常谨慎；如在线用户的并发数较多，对体验要求也比较高，则需要重点分析模型的性能和功能。**

（调用接口、大模型调优这些不同方式的安全性考虑）

2、模型调优

（1）模型调优的技术路线和主要的调优方法

从参数规模的角度，大模型的微调分成两条技术路线：

其一是对全量的参数进行调优的训练，被称作全量微调 FFT(Full Fine Tuning)。FFT 最大的优点就是特定数据领域的训练效果较好。但 FFT 也会带来一些问题，主要有两个：一是训练的成本

比较高，算力需求较大，往往一块 GPU 是不够用的，比如微调 Llama2 7B 就需要更多显存；二是灾难性遗忘(Catastrophic Forgetting)，用特定领域的训练数据去微调可能会使模型在这个领域的表现变好，但也可能会把原来表现好的其他领域的能力变差。

其二是只对部分参数进行调优的训练，被称作轻量化微调 PEFT(Parameter-Efficient Fine Tuning)。这种微调策略的优点是：微调的代价低，降低计算成本，缩短训练时间；降低训练的硬件要求，较小显存的 GPU 和内存上也可以完成训练任务；部署的代价低，可以重复利用已经部署的大模型，不需要重复训练和部署多个模型；防止灾难性遗忘，使得大模型不至于因为学习了新任务而丧失了对先前已经训练好的任务的处理能力。**因此，PEFT 是目前比较主流的微调方案。**

下面按照以上两种技术路线梳理常见的模型微调方法，具体如下：

1）全量微调路线

a) **Supervised fine-tuning 有监督微调**

Supervised fine-tuning 是一种将预训练的语言模型适应于特定任务或领域的训练技术。其基本思想是采用已经在大量文本上进行训练的预训练语言模型，然后在小规模的任务特定文本上继续训练。在这个过程中，预训练模型的权重被更新，以更好地适应任务。所需的 fine-tuning 量取决于预训练语料库和任务特定语料库之间的相似性。如果两者相似，可能只需要少量的 fine-tuning。如果两者不相似，则可能需要更多的 fine-tuning。在 NLP 中，Supervised fine-tuning 最著名的例子之一是由 OpenAI 开发的 GPT 模型。GPT 模型在大量文本上进行了预训练，然后在各种任务上进行了微调，例如语言建模，问答和摘要。经过微调的模型在这些任务上取得了最先进的性能。

有监督微调能够利用预训练模型的参数和结构，避免从头开始训练模型，从而加速模型的训练过程，并且能够提高模型在目标任务上的表现。有监督微调在计算机视觉、自然语言处理等领域中得到了广泛应用。然而有监督微调也存在一些缺点。首先，需要大量的标注数据用于目标任务的微调，如果标注数据不足，可能会导致微调后的模型表现不佳。其次，由于预训练模型的参数和结构对微调后的模型性能有很大影响，因此选择合适的预训练模型也很重要。

b) instruction tuning 指令调优

指令调优是用自然语言形式的格式化实例集合微调预训练大语言模型的方法。指令通常是一种更详细的文本，用于指导模型执行特定操作或完成任务。指令可以是计算机程序或脚本，也可以是人类编写的指导性文本，目的是告诉模型如何处理数据或执行某个操作，而不是简单地提供上下文或任务相关信息。指令调优需要收集或构建指令格式的实例，然后以监督学习方式微调预训练大语言模型。在微调后，大语言模型可以展示出卓越的能力，泛化出能解决未见任务的能力。

以 InstructGPT 为例，指令调优基本流程如下：

- 准备自然语言指令集：针对特定任务，准备一组自然语言指令，描述任务类型和任务目标，例如情感分类任务的指令可以是“该文本的情感是正面的还是负面的？”。
- 准备训练数据集：针对特定任务，准备一个标记化的数据集，其中每个数据样本都包含输入文本和标签，例如情感分类任务的标签可以是“正面”或“负面”。
- 将自然语言指令和数据集转换为模型输入：将自然语言指令和数据集转换为模型输入，例如对于情感分类任务，将自然语言指令和文本拼接作为输入，例如：“该文本的情感是正面的还是负面的？这家餐厅的食物很好吃。”

- 在指令上进行微调：在指令上进行微调，以适应特定任务的需求，提高模型在任务上的性能。

2）PEFT 轻量化微调路线

a) Prefix-tuning

Prefix-tuning 是一种自监督学习方法，通过预测句子前缀的方式进行训练。该方法在生成文本任务中表现出色，它通过针对不同任务设计的前缀生成任务和损失函数来学习表示。Prefix-tuning 在输入 token 之前构造一段任务相关的 virtual tokens 作为 Prefix，然后在接入下游任务进行训练时，预训练语言模型中的其他部分参数固定，只更新 Prefix 部分的参数。

这种微调方式的优点是：通过添加特定的前缀来微调模型，不需要修改底层权重或重新训练，减少了计算资源的使用；可以根据任务要求定义不同的前缀，实现对不同任务的个性化微调。缺点是：对于不同的任务，需要手动设计合适的前缀；前缀的长度和内容可能会影响性能，需要进行调整和实验。

b) Prompt-tuning

Prompt-tuning 是一种学习文本生成模型的指导性提示的方法。通过设计和优化生成任务的提示，可以引导生成模型生成更准确和合理的输出。Prompt-tuning 在生成式对话系统和其他文本生成任务中有广泛的应用。Prompt-tuning 将 prompt 扩展到连续空间，仅在输入层添加 prompt 连续向量，通过反向传播更新参数来学习 prompts，而不是人工设计 prompts。Prompt-tuning 冻结模型原始权重，只训练 prompts 参数，训练完成后，只用同一个模型可以做多任务推理。Prompt-tuning 使用 LSTM 建模 prompt 向量间关联性。

这种训练方式的优点是：通过在输入中添加特定的提示（prompt），可以将模型引导到所需的输出；可以根据任务需要，设计不同的提示来实现个性化微调。其缺点是：提示的设计和选择可

能需要一些经验和尝试；提示的效果可能受限于特定的任务和数据集；不太好训练，省了显存但不一定省时间；多个 prompt token 之间相互独立，可能会影响效果。

c) LoRA

LoRA 是一种轻量级的自监督学习方法，旨在通过对原始数据进行特征提取来学习有用的表示。它不需要标注数据，而是通过设计不同的任务和损失函数来生成伪标签，并使用这些伪标签进行训练，从而学习到适用于多个任务的通用表示。LoRA 的基本原理是冻结预训练好的模型权重参数，在冻结原模型参数的情况下，通过往模型中加入额外的网络层，并只训练这些新增的网络层参数。

这种训练方式的优点是：具备迁移学习和多任务学习的能力，可以共享底层表示并在不同任务之间进行迁移；新增参数数量较少，减少了对整个模型的重新训练的开销，微调成本显著下降，可以获得和全量微调类似的效果。其缺点是：LoRA 中参与训练的参数量较少，解空间较小，效果相比全量微调有一定的差距；对于上百亿参数量的模型，LoRA 微调的成本还是较高。

d) Adapter

Adapter 是一种轻量级的模型扩展方法，它在现有预训练模型的中间层添加额外的适配器层，用于特定任务的学习。适配器层只学习与特定任务相关的参数，其他层保持不变，从而减少了模型参数量和计算成本。

这种训练方式的优点是：实现简单，只需在原预训练模型的特定位置添加额外的适配器层；可以在不同任务之间共享底层表示，减少参数训练数量，降低计算资源的使用；支持快速微调，可以快速将预训练模型应用于新任务。其缺点是：适配器仍然需要微调，可能需要更多的数据和计算资源来达到较好的性能；适配器的设计可能需要一定的经验和调整，以获得最佳的性能。

此外，**模型在经过微调后一般需要进行对齐调优，以增强模型对于人类价值观或偏好的对齐，主流技术是基于人类反馈的强化学习（Reinforcement Learning with Human Feedback）**。

人类反馈强化学习是一种结合了强化学习和人机交互（HCI）的方法，通过接收来自人类用户或专家的反馈来调整自己的行为。这种方法减轻了传统强化学习中需要大量试错的问题，使得大语言模型更加高效、快速地学习任务。在 RLHF 中，通过让人类对不同的模型输出进行排序或评分来收集人类反馈，从而提供奖励信号（reward signal）。收集到的奖励标签（reward labels）可以用来训练奖励模型（reward model），进而反过来指导 LLM（Language Model）适应人类的喜好。这一方法被 ChatGPT 使用而得到大力推广。RLHF 的特点在于它能够直接利用人类反馈来指导智能系统的学习过程。这种方法不仅能够提高大语言模型的学习效率，还能够使大语言模型更好地理解人类用户的需求和偏好，从而更好地完成特定任务。

RLHF 的训练步骤包括三个部分：

- 预训练语言模型：使用经典的预训练目标训练一个语言模型，收集示范数据，确定监督策略；
- 聚合问答数据并训练一个奖励模型：基于预训练语言模型来生成训练奖励模型的数据，并在这一步引入人类的偏好信息，训练奖励模型；
- 用强化学习方式微调：利用奖励模型输出的奖励，用强化学习方式微调优化模型。

另外，由于大语言模型有大量的模型参数，需要占用大量的内存来进行推理，使得实际中部署大语言模型的成本非常高，**模型压缩技术因此成为一种重要的调优方式以减少内存占用和推理延迟，使得大模型可以在资源有限的环境中使用。**

模型量化是一种常见的模型压缩方法，将浮点数参数和激活值映射为整数，从而减小模型的内存占用。模型量化的优点是可以减小模型的内存占用和推理延迟，使得大型语言模型可以在资源有限的环境中使用。然而，模型量化也存在一些缺点，如量化误差可能导致模型性能下降，特别是对激活值的量化更加困难。模型量化主要有两种方法：量化感知训练（QAT）和后训练量化（PTQ）。QAT 方法需要重新训练整个模型，而 PTQ 方法则无需重新训练。对于 LLMs 来说，由于参数数量巨大，**PTQ 方法更受青睐，**因为它的计算成本较低。

（2）图书馆领域大语言模型微调的训练方案

图书馆大语言模型通常是基于开源的预训练语言模型结合指令、专业数据集等进行微调、调优，可采用 LoRA、prompt-tuning、prefix-tuning 等微调方法进行轻量化微调，以降低训练工作量，在最大限度保留预训练语言模型原有能力的同时提高模型在图书馆业务领域的性能表现。

图书馆大语言模型微调需要根据不同的应用场景、不同的调优方法准备对应的数据集，用于图书馆大语言模型微调训练的数据来源可能包括读者问答、书目数据、值班文档、法规文件等数据资源。

同样根据不同的微调方法，在训练数据的预处理方面会有不同的格式和要求，训练数据的预处理过程可能包括数据汇聚、质量清洗和过滤、重复数据删除、减少隐私、标记处理、准备预训练。

3、模型评估

大语言模型（LLMs）的性能评估是评估其能力和效果的重要方面。下面从评估方法、评估标准和测试任务三个角度对大语言模型的性能评估进行介绍。

（1）评估方法

针对大语言模型的性能评估，主要包括定量评估和定性评估的

评估方法。定量评估主要通过计算各种指标来衡量模型的性能，如准确率、困惑度、BLEU 分数等。这些指标可以用于评估模型在语言建模、条件文本生成、代码合成等任务上的性能。定性评估则通过人工分析和判断模型生成的文本的质量、流畅度、一致性等方面来评估模型的性能。

（2）评估标准

大语言模型的性能评估需要制定相应的评估标准。根据不同的任务和应用场景，评估标准有所不同。例如，在语言建模任务中，常用的评估标准是困惑度，即模型对给定上下文的下一个词的预测的准确性。在条件文本生成任务中，可以使用 BLEU 分数来评估生成文本与参考文本之间的相似度。此外，还可以根据任务的特点制定相应的评估标准，如在代码合成任务中，可以使用代码的正确性和可读性来评估模型的性能。

表 2：大模型评估内容及评估指标

评估内容	评估指标
语言建模能力	困惑度（perplexity）
长程依赖建模能力	LAMBADA 准确率、困惑度
条件文本生成能力	BLEU、ROUGE、METEOR 等自动评估指标，人工评估
代码生成能力	代码正确性、代码质量评估指标
对话系统能力	对话质量、流畅度、相关性评估指标
知识问答能力	精确率、召回率、F1 分数
文本摘要能力	ROUGE 评估指标、人工评估
机器翻译能力	BLEU、TER、METEOR 等自动评估指标，人工评估
情感分析能力	情感分类准确率、情感强度评估指标
逻辑推理能力	推理准确率、推理时间
多模态理解能力	图像标注准确率、图像生成质量评估指标
多语言理解能力	多语言翻译准确率、多语言对话质量评估指标
多领域适应能力	领域特定任务准确率、领域适应性评估指标
多样性和一致性评估	N-gram 覆盖率、人工评估
对比实验	准确率、效率、可扩展性等性能指标
难度级别评估	难度级别测试集准确率、难度级别任务完成情况

根据图书馆大语言模型的应用场景，需要根据具体任务和应用场景，选择合适的评估指标来衡量模型的性能。例如，在文献服务

中，可以使用准确率、召回率和F1分数来评估模型对用户查询的回答的准确性和完整性。在读者咨询中，可以使用用户满意度调查或人工评估来评估模型生成的回答的质量和可理解性。

（3）测试任务和测试数据集

为了评估大语言模型的性能，需要使用合适的测试集。测试集应具有代表性，能够涵盖模型在不同领域和不同难度下的表现。常用的测试集包括Penn Treebank、WikiText-103等。此外，还可以设计特定的测试任务来评估模型在特定能力上的表现，如语言理解、文本生成、推理等。这些测试任务可以涵盖不同的领域和难度级别，以全面评估模型的性能。根据图书馆大语言模型的应用场景，可以选择包含图书馆领域相关文本的数据集作为测试集数据。这些数据可以包括图书馆目录、文献摘要、读者咨询记录等，并确保数据集的多样性和代表性，涵盖不同主题、不同文体和不同难度级别的文本。

4、模型部署

（1）主要的部署方式

1）公有云

公有云由第三方服务提供商通过互联网向用户提供虚拟机(VM)、开发平台或软件应用等多元化服务。这些服务可能免费提供，也可能会根据基于订阅或按使用量付费的定价模式收取费用。公有云提供商拥有并管理客户用于运行工作负载的数据中心，负责维护所有硬件和基础架构，并提供高带宽网络连接，确保可以快速访问应用和数据。公有云架构是多租户环境，用户共享虚拟资源池，多租户的工作负载可同时运行在共享的物理服务器上，而每个云租户的数据在逻辑上都与其他租户的数据相隔离。公有云的优势如下：

- 成本更低，无需购买硬件或软件，仅对使用的服务付费；
- 无需维护，维护由服务提供商提供；

- 近乎无限制的缩放性，提供按需资源，可满足业务需求；
- 高可靠性，具备众多服务器，确保免受故障影响。

2）私有云

私有云是专为某家企业运营的云基础架构。通常，私有云托管在本地，位于客户自己的防火墙内，但也可以托管在专用云提供商，此种情况下，客户对基础设施具有专属的单独访问权限。私有云能够提升数据安全性和合规性，避免与其他云客户共享资源可能带来的性能和安全影响。与公有云相比，私有云的优势包括以下方面：

- 控制能力和安全性更高、更高的隐私级别；
- 灵活性更强，应用和基础架构的定制能力更强，组织可自定义云环境以满足特定业务需求。

3）混合云

混合云融合私有云和公有云，使用技术和管理工具，让企业根据需要在两者之间无缝迁移工作负载，进而实现最佳性能、安全性、合规性和成本效益。借助混合云，企业可将无法轻松上云的敏感数据和原有任务关键型应用保存在本地，同时利用公有云部署 SaaS 应用，利用 PaaS 快速部署新应用，以及利用 IaaS 按需提供额外的存储或计算容量。大多数企业云采用者都选择混合云架构，这便于他们灵活地为每个工作负载选择最佳云环境（公有云或私有云），或者随着需求的变化在不同云之间迁移工作负载。混合云具有以下优势：

- 控制力，组织可以针对需要低延迟的敏感资产或工作负载维护私有基础结构；
- 灵活性，需要时可利用公有云中的其他资源；
- 成本效益，具备扩展至公有云的能力，可仅在需要时支付额外的计算能力；
- 轻松使用，无需费时费力即可过渡到云，可根据时间按工作负载逐步迁移。

（2）智慧图书馆大模型应用部署建议

表 3：智慧图书馆大模型应用部署建议

智慧图书馆大模型应用	架构	部署建议
智慧图书馆大语言模型	服务器	私有云
面向读者的服务应用	Web 端、小程序、APP	公有云或混合云
大模型接入的图书馆业务系统	PC 端、web 端	私有云

（3）算力需求及硬件配置

1）算力需求估算

训练端算力需求，与模型参数、训练数据集规模正相关。

训练算力需求=2×# of connections×3×# of training examples×# of epochs[3]（其中，# of connections 是指神经网络中，相互依赖的神经元数量，例如在一个完全链接的神经网络中，N 层输入与 M 层输出，# of connections=N*M，通常 parameters 可以近似于# of connections，# of training examples 指数据集数量；# of epoch 指训练数据集上的完全通过次数）

推理端算力需求，与模型参数数量、平均序列长度、并发需求量正相关。随着大语言模型从简单文字交流，向多模态发展，对于推理算力需求将大幅提升。对于响应速度而言，相比训练，推理的响应速度要求更高（通常用户能接受的响应时间，在几秒之内），因此所需要的并发 GPU 算力相应提升。

推理算力需求=模型大小*推演批次大小*平均序列长度*推演速度[4]（注：理论数值，其中并发请求数量、模型架构、输入数据等，均可能为影响因素）

本文对于类 ChatGPT 应用的推理端算力需求给出估算如下（假设单片 A100 用于 AI 大模型每秒生成 1757 个单词，与单次客户需要

[3] 资料来源：Estimating Training Compute of Deep Learning Models(epochai.org)

[4] 资料来源：华安证券《详解大模型训练与推理对算力产业链的需求影响》

生成的内容数量相当）：如果仅是给企业内部使用，则假设每天访问量为 5000 万人次，每人与 ChatGPT 对话 5 次，由此测算，由于 AI 大模型推理需要新增的 AI 加速卡需求空间为 4.3 万个，新增的 AI 服务器需求空间为 5425 台。如果面向个人用户开放使用，则分别假设每天访问量为 1 亿或 3 亿人次，每人与 ChatGPT 对话 5 次，由此测算下来，由于 AI 大模型推理需要新增的 AI 加速卡需求空间为 8.7 万或 26.0 万个，新增的 AI 服务器需求空间为 1.1 万或 3.3 万台。

2）硬件配置

按照计算芯片的组合方式，一般可以分为："CPU+GPGPU"，"CPU+DSA"，和"CPU+DSA+GPGPU"三种类型。这三种类型目前都已在云计算场景广泛应用和部署。

DSA 即领域专用加速器，是用于一些特定场景或算法族计算的芯片级加速。最早的 GPU 也属于 DSA，也就是图形加速的 DSA。随着 GPU 逐渐演化，将非常小的 CPU 核心加入 GPU 形成 GPGPU 架构后，才具备了通用化的计算能力。

a) CPU+GPGPU

CPU+GPGPU 是较早且部署众多的一种。由于这种架构的计算灵活度高，也可用于模型训练和非 AI 类计算，适合任务种类繁多且差异化大的云计算场景。

b) CPU+DSA

CPU+DSA 是目前 Google 云计算应用较多的方式。例如 Google 去年发布的 Pathways 计算系统（包含 6144 块 TPU）就是这类架构的典型代表。这类架构计算灵活性稍低，但是计算性能和成本都非常明显优于 CPU+GPGPU 模式，广泛用于 GPT-4 或其他算法部署场景。

c) CPU+DSA+GPGPU

CPU+DSA+GPGPU 介于前两者之间，充分提高了灵活性又明显降低了计算成本。这类架构需要算法设计/部署人员有丰富的异构架构

部署经验。

图 8：计算架构对比

5、模型使用

图书馆大语言模型在使用过程中，可以应用上下文学习、思维链等提示策略，帮助模型更准确地理解任务要求和上下文信息，以指导模型生成符合要求的输出，遵循特定的规则和逻辑，充分激发大语言模型的语义理解、内容生成、推理等能力，并使其更好适应图书馆业务领域的任务要求和场景需求，提高图书馆大语言模型在特定任务上的性能。同时，也可使用嵌入向量知识库等方式在不改变模型参数的情况下实现专业数据、知识的存储和记忆，提高大模型的事实回忆能力，并获得更加精准、可信的检索查询和回答效果。

（1）基于上下文学习的提示词

ICL 是一种利用大语言模型的特殊提示形式，通过提供任务描述和示例演示来引导模型识别和执行新任务。通过这些提示，模型可以在没有参数更新的情况下，根据演示的示例来识别和执行新任务。使用上下文学习提示，需要从任务数据集中选择几个示例作为演示，并与任务描述一起组合成特定顺序的自然语言提示。然后，将测试实例（自然语言提示）作为输入附加到演示中，作为模型生成输出的输入。通过任务演示，模型可以识别和执行新任务，而无需显式的梯度更新。另外，对于提示词的生成，可以使用预定义的模板将输入-输出对转化为自然语言提示,也可使用零样本提示“Let’s think step by step”来生成中间推理步骤，或者通过问题分解和顺序求解子问题来生成提示词。

（2）思维链提示

思维链提示是一种改进的提示策略，通过在提示中包含中间推理步骤，使模型能够逐步推理和生成结果，提高大语言模型在复杂推理任务上的性能。与传统的提示策略不同，CoT Prompting 在提示中引入了推理路径，使模型能够进行多步推理，从而更好地解决复杂任务。这种提示策略在解决算术推理、常识推理和符号推理等复杂推理任务上表现出色。构建思维链提示词需要选择合适的中间推理步骤来引导模型进行逐步推理，需包含任务描述、输入数据、上下文信息和提示风格等关键要素，同时也可使用标记（如引号和换行符）来突出重要部分，或分隔任务描述和上下文信息，帮助模型更好地理解和执行任务。

（3）向量知识库 embedding

一般来说，大语言模型学习知识的方式主要有两种：一是通过模型微调训练调整模型参数；二是通过嵌入向量知识库。在模型微调训练成本较高、难度较大的情况下，嵌入向量知识库也不失为一种给大语言模型灌输知识并获得较好的检索问答效果的不错的选择。大语言模型嵌入向量知识库即为用户利用向量知识库管理自有知识资产并结合大模型构建垂直领域的智能服务。向量数据库存储和处理向量数据，提供高效的相似度搜索和检索功能，通过向量嵌入，将用户知识库文档和数据转化为向量表示，并与大模型交互，实现专有的、面向垂直行业的智能化应用。一种基于大语言模型和向量知识库构建的典型应用案例是 GPT4 PDF Chatbot Langchain 项目，其原理是 GPT4 PDF Chatbot Langchain 项目的底层逻辑是将用户提供的各种 PDF 文档转化成文本文件并将其融合分解，通过智能化技术将其向量化到向量数据库中，用户使用对话模型和 ChatGPT 发起对话，ChatGPT 根据提示语，找到最相关的文档并通过 GPT 生成与需求最匹配的答案。

大模型微调和嵌入向量知识库的比较如下：

表 4：大模型微调和嵌入向量知识库比较

大模型学习知识的方式	优势	劣势
微调 Fine tuning	1) 模型可以学习新的风格并实现较好的生成效果，比如新的句式、文风、语气特点等； 2) 模型输出或回复更加自然、流畅。	1) 模型输出可能不够可靠，可能生成虚假、错误信息或胡编乱造 2) 模型需重新训练、训练成本较高、难度较大； 3) 模型缺少对自由数据和知识资产的长期记忆机制。
向量数据库 Embedding	1) 模型在向量知识库覆盖领域的检索、问答的输出更加准确、可靠； 2) 学习知识的效率更高，实现成本更低。	1) 模型针对嵌入的数据只能做检索查询、问答，难以将这些数据融入到模型的语言和知识体系。

大语言模型嵌入向量知识库主要包括三个步骤：向量化、提问和回答。

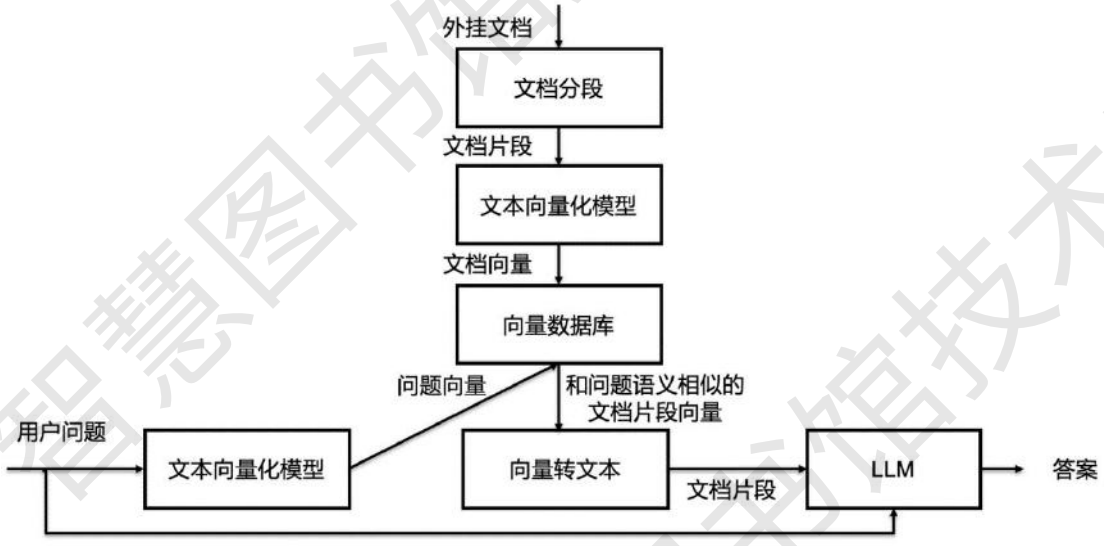

图 9：大模型嵌入向量知识库

1) 构建向量知识库：使用嵌入模型为想要索引的内容创建向量首先，收集目标索引的内容，可能包括书籍、文章、网站、图

片等信息，例如 pdf、txt、md、docx 等格式的文本。其次，将收集到的内容进行切块处理，将原始数据分解成可以独立处理的小块。文本拆分有不同的方式，可根据规则进行拆分，即根据中文文章的常见终止符号进行文本分割，如单字符断句符、中英文省略号、双引号等，或者根据语义进行分割，将文本拆分为语义不同的小块（通常是句子）。通常，由于语义的不确定性，使用规则分割方式可能会取得更好的效果，文本分句长度为 800。接下来，使用嵌入模型为每部分内容获得向量结果。嵌入模型是一种能够将高维数据映射到低维空间的技术，使得数据可以在计算机上进行处理。例如，可以使用 Word2Vec 模型将每个单词转化为一个向量。常见的向量化模型有 OpenAIEmbedding、text2vec-base-chinese、instructor-large。最后，存储向量数据形成向量数据库，这个数据库可以用于各种应用，如搜索引擎、推荐系统等。

2) 用户提问及向量化：用户进行提问并对提问进行向量化处理

首先，用户通过用户界面或命令行输入提出问题，用户可以自由地提出任何与向量知识库相关的问题。接下来，使用嵌入模型对问题进行向量化处理，将用户问题转换为向量形式，以便计算机能够更好地理解和处理，用户提问的向量可以与向量知识库的特征向量进行比较和匹配，从而提供更准确的答案或相关信息。

3) 检索向量并回答：在向量数据库中检索相似向量，返回回答

首先，在向量数据库中检索相似向量，通过计算向量之间的相似度，例如余弦相似度、欧氏距离等，找出与用户提问向量最相似的向量。之后，根据向量数据库将相似向量转换成文本或内容，这个过程可能会用到词嵌入、文本生成等自然语言处理技术，转换后的文本或内容可以更直观地表达含义。最后，将转换后的文本或内容发送给大模型，大模型根据接收到的信息进行语言组织和处理，然后生成回答。

五、智慧图书馆大模型创新生态建设路径

充分整合各方资源优势，推动资源互补、协同创新，构建政产学研用多方参与、互利共赢的智慧图书馆大模型创新生态，为智慧图书馆大模型创新和应用提供丰厚土壤，营造开放包容、融合创新的发展环境。

（一）加强数据开放共享和分析挖掘

开放图书馆行业业务数据，为图书馆领域大模型开发和应用提供高质量的数据支撑。图书馆可以公开部分图书馆行业的业务数据，例如读者问答、书目数据、值班文档等，依托技术公司整理大模型训练微调所需要的指令数据集，率先构建图书馆大语言模型训练微调的数据集并促进数据开放共享，为预训练大语言模型的微调优化提供高质量的数据支撑，促进大模型在图书馆领域的落地应用，完善图书馆领域大模型创新开发的基础资源支撑。

基于大模型技术等加强图书馆数据分析利用，充分挖掘数字人文研究价值。图书馆可以通过设立数字人文业务，针对馆藏资源进行再次开发，将馆藏资源（如MARC元数据、全文影像图片数据、音视频数据和文本化的资源内容）进行重新知识化的组织，在数据基础上以辅助人文研究，为人文研究提供新方法、新手段和新视角。基于大模型的应用加强图书馆馆藏资源的深度挖掘和分析，从而发掘出新的知识和信息，助力人文研究者探索更多的研究领域和方法。例如，利用自然语言技术对大量的文本数据进行快速分析，识别出文中的关键信息和关系，为研究者提供更加深入、精准的研究材料。支持图书馆员利用信息技术的最新发展辅助馆藏数字化和数据基础设施的建设，提高馆藏数字化的效率和数字馆藏利用率。

举办数据创新大赛，开展数据创新应用的交流合作。图书馆可以举办开放数据竞赛，通过开放图书馆历史文献资源，凝聚起一批又一批来自社会各界的数据创客团队，收获大量优秀创意作品。图

书馆还可与国内外的研究机构、高校和企业进行合作，共同开发一系列基于高质量数据的应用，如知识图谱、智能检索和推荐系统等，以期为用户提供更为便捷、智能的服务。

（二）搭建多元服务集聚的开放平台

基于微服务架构搭建开源服务平台，赋能智慧图书馆应用创新。随着图书馆发展进入以智慧图书馆为特征的第三代时期，当前的技术挑战不仅仅是满足传统的集成管理系统的需求，更在于搭建一个能够为各类用户和开发者提供稳固、高效的技术支持，推动创新与协同工作的无缝融合的赋能平台。基于 FOLIO 的开源架构搭建本土化的下一代图书馆服务平台，以微服务架构为基础，支持微服务组件的灵活调用以及其它系统和应用的无缝集成，实现智慧图书馆各类应用和服务的开发部署、集成对接及接入整合，从而为智慧图书馆业务应用提供统一的平台支撑，高效赋能智慧图书馆大模型应用场景建设。依托云瀚平台，图书馆可结合实际业务需求选择、搭配和部署适合的功能模块，实现服务和应用的定制化和个性化。

推动云瀚平台智慧化升级，赋能智慧图书馆大模型创新应用。充分把握大模型技术发展机遇，推动云瀚平台融合大模型技术进行智慧化升级，打造集成开发 AI 模型和应用的工具链及整套环境的一体化平台，对外提供包括人工智能模型和应用开发、训练以及全生命周期管理的一站式、工程化 MaaS 服务，将云瀚平台升级为融合人工智能算力服务、模型开发训练和部署等多元化服务的开放平台，同时也可在云瀚平台上应用基于大模型技术的编程助手、研发助手，进一步提升云瀚平台创新效能，打造大模型赋能的智慧图书馆应用开发新范式，为智慧图书馆大模型应用场景建设提供有力支撑。

（三）完善大模型创新应用标准规范

建立完善图书馆领域大模型创新应用的标准规范体系，促进智慧图书馆大模型创新应用互联互通、开放共享。围绕智慧图书馆大

模型开发、训练、测试评估等，制定技术、业务、数据、服务、产品等方面的标准规范，使智慧图书馆领域大模型技术应用在技术、接口、界面、交互等方面保持一定的一致性、兼容性，为图书馆领域大模型创新和技术应用提供标准规范依据。构建图书馆大模型的科学评价机制，从模型的准确性、稳定性、响应速度到对复杂请求的处理能力等多方面进行细致评估，为大模型应用效能评估提供科学依据。完善图书馆大模型可信标准及评测规范，涵盖基础技术、系统应用、业务场景等各方面，将可信要素全面融合大模型研发、生成、经营等全流程。制定智慧图书馆大模型训练数据质量管理标准规范，对数据的公平性、来源、处理和整合方式进行严格的质量定义，确保数据的真实性、完整性和有效性。

积极探索大模型合规实践和价值对齐，确保智慧图书馆大模型应用的安全与合规性。探索制定智慧图书馆大模型的合规操作指引，在技术、数据处理、用户交互和应用等多个层面制定合规准则，确保智慧图书馆大模型在各种应用场景中的合规性和安全性。如加强对数据的保护措施，采用先进的加密和匿名化技术，确保个人数据和隐私权益不受侵犯，确保数据来源的透明性和可追溯性。完善图书馆领域大模型与伦理、价值观对齐的方法论体系，倡导大模型价值观校准、深入开展伦理审查，利用人工标注来校正和消减模型中的固有偏见，为大模型在智慧图书馆的应用确立人文与技术的均衡，使图书馆大模型在服务读者时既高效又具有道德伦理的考量，促进图书馆领域大模型与人类价值观、伦理和谐融合。

（四）依托联盟营造开放的创新氛围

依托云瀚联盟营造开放包容的创新氛围，促进图书馆行业资源共享、开放创新。依托云瀚联盟和社区对云瀚平台的开源套件进行定期维护和更新迭代，研制相关的评测标准开展智慧图书馆大模型评测并发布评测报告，同时也可针对图书馆行业大模型创新开展分

析研究、发布趋势报告，对智慧图书馆大模型创新应用的成功实践进行推广培训，鼓励图书馆行业积极开展大模型创新应用，营造开放包容的创新氛围。持续扩大联盟合作范围，率先开创元宇宙和web3.0时代的新型组织关系，并推动资源共享、开放合作，推动各类图书馆、开发商、服务商及相关机构互相协作，共享知识，开发出满足特定需求的应用和服务，国内图书馆可积极参与FOLIO社区，与国际同行开展合作，促进技术交流和资源整合，积极探索适合中国的图书馆行业联盟模式。

（五）开展行业人才培养与交流互动

加强图书馆行业复合型人才培养与专业培训。为了充分利用大模型技术加快推进智慧图书馆建设，需要培养一批既懂得图书馆学、信息科学，又了解大模型技术的复合型人才。这要求图书馆工作人员不仅需要有扎实的图书馆学和信息科学知识，还需要对人工智能、数据科学和计算机科学有所了解。因此，需要在传统的图书馆学教育中融入人工智能、数据科学和计算机科学的元素，培养学生的跨学科思维和技能。同时，也可以鼓励图书馆员通过进修、研讨会和工作坊等方式，学习和掌握新的技术和方法。针对图书馆工作人员开展专业知识的普及培训也是培养专业人才的重要支撑。通过定期的培训、研讨会和在线课程，使图书馆员对大模型有一个基本的认识，了解其在图书馆领域的应用和价值。同时，也可鼓励图书馆员进行实践和创新，探索大模型在实际工作中的应用方法和技巧。在这个过程中，依托图书馆联盟可以实现专业培训的多样化。一方面联盟内的不同角色，可以提供多样化的教育和培训资源；另一方面，联盟内也可以组织培训项目，分享最佳实践和经验，提高培训的效果和范围，使图书馆员接触丰富的知识和技能。

围绕智慧图书馆大模型创新应用积极开展交流互动。在大模型创新实践中，图书馆还需要与其他图书馆、学术机构和技术企业进

行合作和交流，使图书馆获得更加广阔和深入的视角，帮助图书馆更好地满足用户的需求。可通过问卷调查、访谈和用户研究等方式，收集图书馆内部和外部的需求和建议，指导大模型的研发和优化。同时，也应密切关注图书馆领域的发展趋势和变化，确保大模型技术应用的与时俱进，满足图书馆的长远发展需求。图书馆联盟通过促进交流互动可以了解图书馆彼此间不同的服务模式、用户群体和需求，这种多样性有助于确保大模型创新应用的通用性和适应性。一旦某个图书馆在大模型的应用中取得了成功，其经验和技术可以通过联盟快速推广到其他图书馆。

参考文献

[1]吴建中. 建设智慧图书馆，我们准备好了吗？[EB/OL].(2021-11-28)[2023-09-12].
https://mp.weixin.qq.com/s/iJQELXIPodIUkK9dbcemgg
[2]刘炜.智慧图书馆十问[J].图书馆理论与实践,2022(03):1-6.DOI:10.14064/j.cnki.issn1005-8214.20220215.001.
[3]蔡丹丹,张智敏,贺晨芝等.图书馆IT应用十大趋势[J].图书馆建设,2023(01):76-83+94.DOI:10.19764/j.cnki.tsgjs.20230045.
[4]中国人工智能大模型地图研究报告[R].北京:中国科学技术信息研究所,2023.
[5] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Yang Chen, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Wen Ji-Rong. "A Survey of Large Language Models." ArXiv.org (2023): ArXiv.org, 2023. Web.
[6] Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. "Emergent Abilities of Large Language Models." ArXiv.org (2022): ArXiv.org, 2022. Web.
[7]叶鹰,朱秀珠,魏雪迎等.从ChatGPT爆发到GPT技术革命的启示[J].情报理论与实践,2023,46(06):33-37.DOI:10.16353/j.cnki.1000-7490.2023.06.005.
[8]尹沿技,王奇珏,张旭光.详解大模型训练与推理对算力产业链的需求影响[R].合肥:华安证券,2023.