

大模型落地路线图研究报告

(2024 年)

中国信息通信研究院人工智能研究所
人工智能关键技术和应用评测工业和信息化部重点实验室
2024年9月

版权声明

本报告版权属于中国信息通信研究院，并受法律保护。转载、摘编或利用其它方式使用本报告文字或者观点的，应注明“来源：中国信息通信研究院”。违反上述声明者，编者将追究其相关法律责任。

更名声明

原“集智”蓝皮报告更名为“集智”专题报告。“集智”专题报告将聚焦于特定领域或主题的深入探讨，提供更为专业和集中的内容分析。

前 言

近年来，大模型技术能力不断创出新高，产业应用持续走深向实，显著加快人工智能通用化、实用化、普惠化发展进程。大模型兼具理论研究和产业应用双重价值，已成为当前科技前瞻布局和创新研发的焦点，受到国际社会高度关注。系统梳理大模型应用落地的共性需求和关键要素，给出现状诊断、能力建设、应用部署、运营管理等落地路线作为参考，将进一步促进大模型赋能千行百业，为打造新质生产力和推进新型工业化开辟新路径。

本报告重点梳理了大模型应用落地遵循的诊断、建设、应用、管理**四个重要阶段**，归纳总结出能力分析、需求挖掘、方案设计、研发测试、应用开发、效能评估、运维监测、运营管理**八个关键步骤**。基于需求拉动、问题驱动、创新推动原则，研究分析在每一步骤中基础设施层、数据资源层、算法模型层、应用服务层、安全可信层应重点关注的发展要素和亟待解决的核心问题。**现状诊断阶段**，通过分析大模型技术能力、梳理大模型应用场景、评估大模型发展基础，帮助应用方明晰业务发展和转型需求。**能力建设阶段**，设计契合应用方战略规划和业务需求的大模型建设方案，通过系统性研发和测试手段筑牢大模型技术底座。**应用部署阶段**，给出专用大模型优化部署和智能应用定制开发等参考模式，并构建出覆盖大模型设计开发、定制优化、部署运营等全生命周期的应用效能评估体系。**运营管理阶段**，面向大模型平台和服务给出运维监测和运营管理的参考模式，给出实时监测、动态追踪和预警机制等具体方法，指出

建立健全大模型运营管理体系的原则和要点。

未来，大模型有望持续为新一轮人工智能阶跃式发展注入强大动能，进一步助力人工智能实现技术能力不断融合创新、应用场景持续纵深拓展、产业生态加速转型升级、治理体系趋于完善稳健。本研究报告对大模型应用落地的研究和理解还有待加强，报告中如有不足之处，还请各方专家读者不吝指正。

目 录

一、大模型发展情况概述.....	1
（一）全面提升多模态感知能力，推进认知智能走深向实.....	2
（二）大模型场景落地百花齐放，擘画智能应用生态蓝图.....	5
（三）技术选型与工程实践掣肘，亟需开展落地路线研究.....	7
二、诊断大模型能力基础.....	9
（一）评估大模型能力发展现状，深入挖掘业务转型需求.....	9
（二）梳理人财物要素就绪情况，系统评估人工智能基础.....	10
（三）挖掘各类软硬件资源需求，统筹规划业务发展目标.....	12
三、筑牢大模型技术底座.....	15
（一）剖析大模型关键落地路线，科学确立技术选型原则.....	15
（二）设计大模型智能系统架构，制定科学有效解决方案.....	16
（三）系统研发大模型技术底座，全面开展功能性能测试.....	20
四、革新大模型应用范式.....	24
（一）参照业务场景个性化需求，定制化调优专用大模型.....	24
（二）开发大模型原生智能应用，实现大小模型协同赋能.....	25
（三）构建全链路效能评估体系，形成诊建用评有效闭环.....	30
五、构建大模型管理体系.....	33
（一）梳理国内外模型治理要点，明晰运营管理体系价值.....	33
（二）实时监测大模型运行过程，确保高效稳定提供服务.....	34
（三）建立健全大模型管理体系，保障业务高效稳定开展.....	37
六、大模型发展趋势展望.....	40
（一）探索大模型架构优化方案，带动技术应用双重涌现.....	40
（二）紧抓行业数字化转型机遇，全方位打造新质生产力.....	42
（三）加强引导大模型可信发展，对齐人类偏好及价值观.....	43

图 目 录

图 1 大模型底层技术支撑人工智能发展.....	2
图 2 大模型的多维感知能力和认知能力.....	4
图 3 大模型行业应用图谱.....	6
图 4 大模型应用场景.....	7
图 5 应用方能力基础就绪度等级表.....	10
图 6 大模型发展能力就绪度评估指标体系.....	12
图 7 需求分析的关键要素.....	15
图 8 大模型建设方案设计的关键要素.....	20
图 9 百度智能云千帆大模型平台.....	24
图 10 讯飞星火认知大模型云服务平台.....	30
图 11 大模型应用成熟度评估体系.....	33

表 目 录

表 1 模型生态主要选型方案.....	18
表 2 模型优化主要选型方案.....	18
表 3 模型部署主要选型方案.....	18
表 4 模型规模主要选型方案.....	18
表 5 AI Agent 能力及特点.....	27

一、大模型发展情况概述

大模型是指具有大规模参数和复杂结构的深度神经网络模型，通常在训练过程中高度依赖海量的数据资源和强大的计算能力。近年来，算力水平的飞跃、数据规模的激增、算法能力的突破为大模型的长足发展提供了坚实的技术基础。大模型技术能力的持续优化和升级、应用场景的不断丰富和细化，以及各行业实现智改数转的需求日益高涨，共同促使大模型的应用市场呈现出欣欣向荣的发展态势。

大模型技术能力的突破得益于算力设施、数据资源、计算模式、网络架构等方面的源头创新。算力设施上，智能计算集群的快速发展为大模型的算力需求提供了坚实的基础保障。为满足万亿参数大模型的计算需求，NVIDIA 发布了超强人工智能加速卡 Blackwell GB200，通过芯片间的互连技术实现计算能力的代际飞跃。该芯片相较于 H100，在大模型训练性能上提升 3 倍¹，在推理速度上提升 30 倍，显著降低了训练推理的成本和能耗。数据资源上，大规模、高质量、多类型的数据集蕴含着丰富的语义知识，有助于提高大模型的技术能力，充分释放出大模型在不同场景下的应用价值。计算模式上，通过分布式并行、混合精度计算²等方法实现训练加速，通过模型压缩和蒸馏、低精度推理³等方法实现推理加速，可有效提升计算资源的利用率。为加速大模型的训练过程，百度飞桨推出 4D 混合并行策略，通过模型并行、参数切片并行、流水线并行和数据并行等方式可支持训练千

¹

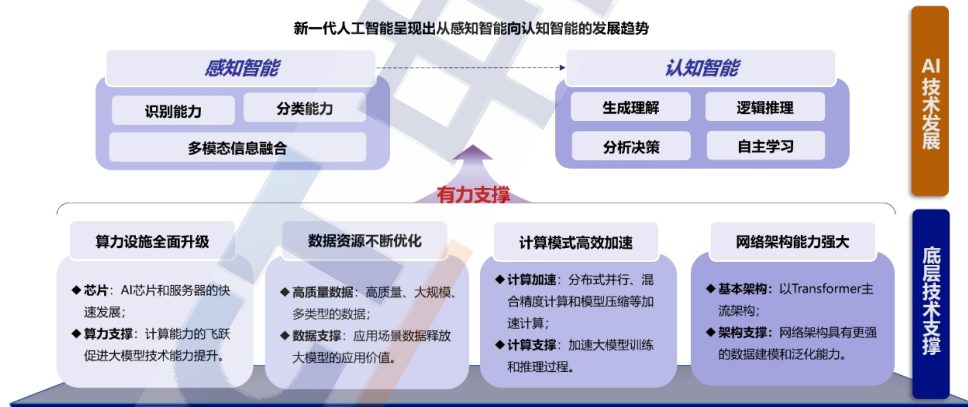
<https://developer.nvidia.com/zh-cn/blog/nvidia-gb200-nv172-delivers-trillion-parameter-llm-training-and-real-time-inference/>

² https://huggingface.co/docs/transformers/perf_train_gpu_one#fp16

³ <https://arxiv.org/abs/2210.17323>

亿级稠密参数模型，相比其他 3D 混合并行策略至少提速 23.7%。网络架构上，以 Transformer 为代表的主流架构不断迭代增强，进一步提升大模型的数据建模和泛化能力，不断突破大模型在长序列任务上的瓶颈。例如，在 Transformer 解码器的基础上融入混合专家模型可以有效提升大模型的性能表现⁴。基于 Mega 架构改进的 Megalodon 能够处理无限上下文，有效提高了大模型在长上下文上的训练推理的质量和效率。

在上述底层技术的有力支撑下，大模型的技术能力不断提升，在“听、说、看”等感知领域已接近或超越了人类水平，在理解、思考和创造等认知领域已取得显著进展。目前，大模型的研究重心逐步向认知智能转变，这是新一代人工智能的发展趋势。



来源：公开资料整理

图 1 大模型底层技术支撑人工智能发展

（一）全面提升多模态感知能力，推进认知智能走深向实

大模型的智慧不断涌现，为感知智能和认知智能的发展注入活力。一方面，大模型的多模态感知能力能够融合文本、图像和语音等不同

⁴ <https://arxiv.org/pdf/2305.14705>

模态信息，实现精准识别和综合分析。另一方面，大模型多维度展现出卓越的认知能力，逐步接近人类水平。

1. 大模型具备多模态感知能力

强大的表征和泛化能力赋予大模型出色的单任务识别分类和多模态融合处理能力。单任务识别分类能力上，大模型具有强大的特征抽取能力，能根据不同类别之间的特征表示准确识别出不同的对象，从而具备更加精准的感知能力，能够在人脸识别、语音识别、图像分类、文本分类等任务上取得更好的效果。多模态融合能力上，大模型凭借强大的表征和泛化能力可以同时处理大量文本、图像、视频、音频等不同模态的信息，通过统一的表示空间实现跨模态的信息交互和知识融合，更加符合人类感知和处理信息的方式。例如，自回归多模态大模型 Unified-IO 2⁵通过将不同模态信息映射到一个共享语义空间，能够在自然语言、图像、音频、视频和具身智能等多种任务中接近或优于单任务模型的性能，展现了多模态大模型的研究前景。

2. 大模型认知能力显著提升

大模型具备较好的理解生成能力，能够深入建模数据间的逻辑关系，依托自主学习机制实现精准的推理决策。理解和生成能力上，大模型结合跨领域知识图谱、上下文学习、持续学习等技术获得理解复杂问题的能力，通过序列到序列模型、扩散模型等技术提升生成创造能力，从而完成对话生成、代码生成、文案创作、视频理解等任务。例如，在智能问答的真实性评价方面，GPT-4 在 TruthfulQA 测试集上

⁵ <https://arxiv.org/abs/2312.17172>

的准确率近 60%⁶，逐步接近人类的表现。推理和决策能力上，大模型智能涌现的重要表现之一是强大的逻辑推理能力，可以通过思维链、提示学习、强化学习等技术挖掘数据之间的关联以实现推理和决策，在逻辑推理、常识推理、决策规划等关键任务上表现突出。针对数学单词问题（Math Word Problem）任务，相较于传统提示学习在数据集 GSM8K 上 25% 的数学推理准确率，基于思维链的大模型 PaLM 能达到近 60%⁷。自主学习和适应能力上，大模型通过迁移学习和自适应学习可实现根据环境和任务的变化自动调整自身参数和结构，并通过持续学习不断接收和处理新的数据，从而加强对外部环境的自主反馈和适应的能力。



来源：公开资料整理

图 2 大模型的多维感知能力和认知能力

大模型的识别、分类、理解、推理、决策、生成等能力显著提升，将加速计算智能、感知智能走向认知智能。大模型强大的认知能力能

⁶ <https://arxiv.org/pdf/2303.08774>

⁷ <https://arxiv.org/abs/2201.11903>

够处理更加复杂的任务，满足定制化的场景需求，提供更加精准的推理、交互、决策等个性化服务，为模型的深入应用提供重要保障。

（二）大模型场景落地百花齐放，擘画智能应用生态蓝图

随着大模型效能不断外溢，感知能力和认知能力显著增强，为上层智能应用和服务夯实了坚实基础，应用路径和商业模式开始清晰，各行业开始积极拥抱新技术，希望借助大模型突破自身发展瓶颈。近年来，大模型已在金融、工业、教育、医疗、政务等行业得到应用，并赋能研发设计、生产制造、经营管理、营销服务等多种应用场景。

1.面向行业加速数字化转型步伐

由大模型引发的新一轮行业变革正向纵深推进，将开启千行百业数字化转型“黄金时代”。目前，各行业已经开始积极评估自身能力现状及应用大模型的可行性，希望借助大模型突破发展瓶颈，达到降本、增效和提质的目的。在金融领域，大模型能够分析海量的金融客户数据，在风险控制、欺诈检测和智能投研等场景下探索应用，帮助金融机构实现高效率、低成本、规模化的人工智能创新应用。例如，中国工商银行将大模型应用于智能客服领域，将平均通话时长缩短 10%，座席服务效率提升 18%，实现座席工作效率的全面升级。在工业领域，大模型的技术能力特点深度契合新型工业化特征，可基于感知预测和决策规划等能力在设计制造、产能优化、知识管理、生产运营、节能环保等场景上全面助力工业领域降本增效。在政务领域，多地政府积极在政府服务过程中引入大模型技术，以提升政务咨询、业

务办理、城市治理、辅助决策等方面的精度和温度，从而有效提升市民体验、政务办公和城市治理效率。



来源：公开资料整理

图 3 大模型行业应用图谱

2.面向场景激发应用创新活力

大模型的技术革新进一步满足了业务场景对智能技术的需求，有望带来更加智能化的用户体验，提升服务效率和质量。在研发设计场景，大模型能够利用研发数据和历史案例解析研发需求，自动生成设计方案、产品原型、开发代码等，进而加速研发设计过程。在生产制造场景，大模型通过分析生产过程中的相关数据支持故障预警、质量控制、流程优化、生产调度等任务，有效提高生产效率和资源利用率。在经营管理场景，大模型通过分析业务数据和市场信息可帮助企业高效实现战略规划、市场预测、业务决策、资源调配，从而提高企业的竞争力和收益率。在营销服务场景，大模型可以基于用户画像提供定制化客服、个性化推荐、预测性营销、品牌形象优化等服务，提高客户满意度和服务效率。在知识管理场景，大模型可帮助企业构建内部

知识库，并辅助开展知识搜索、知识问答和教学辅助等工作，帮助员工快速获取所需知识和信息。在智能助手场景，大模型通过语音或文本与用户进行交互，以智能助手、虚拟个人助理等产品形态提供日程安排、任务管理、信息查询等个性化服务。

研发设计			生产制造		
软件开发	产品研发	方案设计	仓储管理	故障预警	流程优化
通信开发	造型生成	...	质量控制	生产调度	...
经营管理			营销服务		
智慧办公	市场预测	智能供应链	个性化推荐	智能客服	品牌舆情
规划决策	资源调配	...	售后服务	广告生成	...
知识管理			智能助手		
知识库构建	内容生成	知识问答	语音助手	虚拟个人助理	健康助手
知识搜索	教学辅助	...	聊天机器人	专业领域助手	...

来源：公开资料整理

图 4 大模型应用场景

（三）技术选型与工程实践掣肘，亟需开展落地路线研究

大模型广阔的应用前景已打破原有行业竞争格局，各行各业纷纷开始思考如何利用大模型实现革命性提效。然而，当前大模型发展仍存在工程实践复杂、技术选型困难、成功案例缺乏等问题，在应用落地实践过程中面临重重挑战，迫切需要大模型落地路线图作为参考。

1.大模型工程实践复杂，需系统梳理落地路线图作为参考

大模型技术更新迭代快，当前应用实践过程中缺少标准化的落地路径。基础设施方面，在如何构建高性能和高可靠的训练和推理基础

设施，以及在如何根据行业属性或企业性质选择大模型的部署方式等问题上缺乏清晰的技术路径。**数据构建方面**，大规模和高质量的数据集决定了大模型的实际应用效果，因此应用方需考虑如何进行全流程数据治理以及如何构建数据隐私和安全保护体系。**服务能力方面**，面对不同行业 and 不同场景的需求，应用方在如何实现大模型与现有业务数据和信息系统对接，如何开展提示工程，如何开发人工智能原生应用等问题上缺乏成熟的实施方案。

2. 大模型技术选型困难，需明确技术指标和评估方法

当前大模型类型众多，模型应用方需要明确技术指标和评估方法，以对大模型的模态类别、技术能力和应用方案进行选型。**技术指标方面**，大模型开发过程涉及基础设施、数据资源、算法模型、应用模式和风险控制等诸多方面，需要成体系的技术指标选型方案以对大模型进行选择、建设和验收。**评估方法方面**，需要在模型应用的全生命周期开展技术能力先进性和应用场景适用性等评估。从大模型的技术要求出发，在应用前评估现有模型的性能水平，在应用中评估算法模型与实际业务需求的匹配程度，在应用后跟进模型使用效果以制定改进方案。

本报告系统梳理了大模型应用过程，归纳形成大模型落地路线图，期望有效促进行业标准的制定和最佳实践的推广。本报告在研究过程中遵循“需求拉动、问题驱动、创新推动”三大原则，从大模型的现状诊断、能力建设、应用部署、运营管理四个阶段着手，围绕基础设施、数据资源、算法模型、应用服务、安全可信五个层面全方位分析

大模型在落地过程中的问题挑战与解决思路，探索大模型的最佳落地路线。

二、诊断大模型能力基础

（一）评估大模型能力发展现状，深入挖掘业务转型需求

应用方应多维度进行能力诊断，科学客观制定诊断原则。诊断维度上，通过技术能力诊断，有助于应用方深入了解技术方面的现状和问题，明确是否具备足够的技术基础来开发和运维大模型。通过应用场景诊断，帮助应用方更好地把握应用趋势和用户需求，以在实际生产应用中取得更好的效果。通过能力诊断，助力应用方探索潜在的市场机遇和发展空间，明确战略定位和发展方向，建立健全发展机制。诊断原则上，根据应用方在基础设施、算法模型、业务场景、人才团队、战略规划、经费预算等方面的能力现状，分析基础资源的完备度、人才团队配比的平衡度、战略规划与大模型应用的契合度，综合研判应用方能力基础的就绪度，根据就绪情况划分为 L1、L2、L3、L4、L5 不同等级。

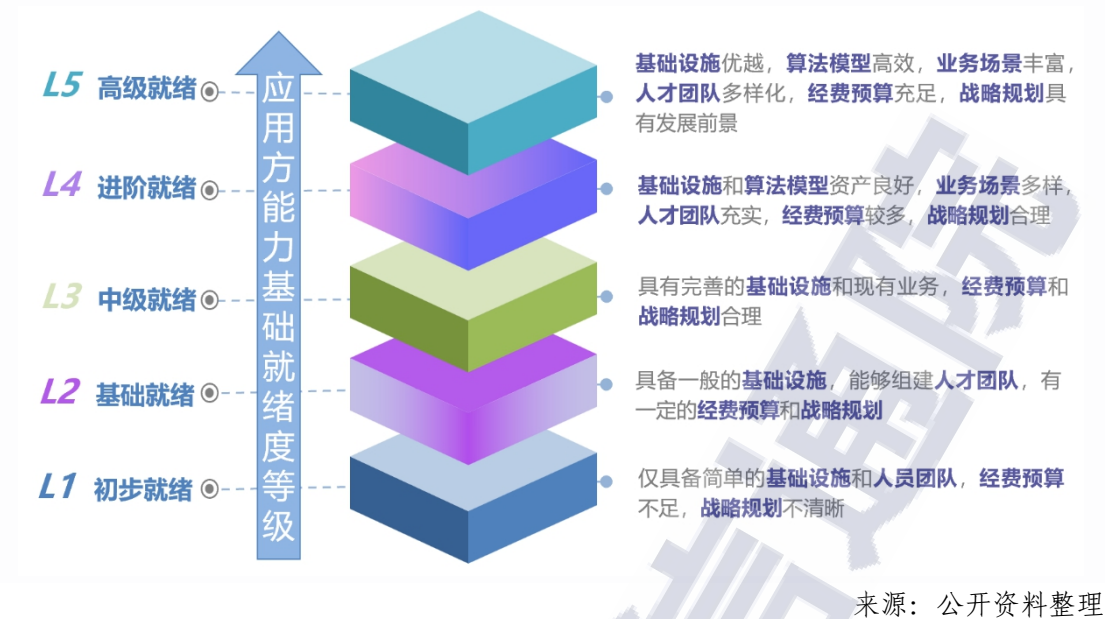


图 5 应用方能力基础就绪度等级表

基于自身能力和业务目标充分挖掘大模型建设需求。需求分析的目标上，多层面的需求分析有助于应用方准确把握自身发展定位，了解现阶段能力与行业平均水平及先进水平的差距，明确自身业务的需求和目标，进一步指导模型的设计、优化和创新，确保大模型高效有序开发。需求分析的范围上，聚焦基础设施、数据资源、算法模型、应用服务、安全可信五个方面，充分分析业务场景需求，全面规划大模型发展要素，确保模型开发的针对性和有效性。

（二）梳理人财物要素就绪情况，系统评估人工智能基础

应用方应综合评估自身能力现状，全面梳理在基础资源、人才团队、战略规划等方面的就绪情况，帮助应用方更清晰地认知自身在大模型领域的能力和潜力，为全面充分的挖掘需求做好准备。

1. 评估基础资源的完备度

应用方应着力评估自身基础软硬件、数据资源、算法模型等方面

的储备现状。**基础软硬件方面**，应梳理已建算力、网络、存储等资源的储备情况，评估软硬件设施的支撑能力。算力上，应评估浮点计算能力、芯片性能、能效比、综合利用率等指标。网络上，评估架构、带宽、延迟、稳定性等指标。存储上，着重评估存储容量、吞吐量和访问时延等指标。软件设施上，评估向量数据库、深度学习框架、操作系统等软件设施的功能、性能及兼容性等情况。**数据资源方面**，应梳理数据规模、类型、分布等情况，评估数据的准确性、完整性、一致性和可用性等质量情况。**算法模型方面**，应梳理现有算法模型资产的种类、数量、部署模式、编程语言、开放程度及兼容性等情况，评估人工智能技术研发基础。

2. 评估人才团队配比的平衡度

应用方应评估人才团队的技术能力和管理能力。技术能力方面，应梳理各类人员所具备的大模型架构设计、算法优化、数据治理、测试验证等专业技能储备情况，通过人工智能领域相关的教育经历、岗位任职、职称职级、工作年限、项目经验、技能专长、论文专著、知识产权、行业影响力等方面综合评估人才团队的技术能力。**管理能力方面**，应采取科学有效的方法评估团队领导、沟通协调、组织规划、分析决策、时间管理、项目管理、环境适应、自我管理学习等管理能力。

3. 评估战略规划与大模型应用的契合度

应用方应评估自身战略规划、经费预算与大模型应用的契合度。战略规划上，应根据自身的使命和愿景、发展定位和未来布局，结合

战略目标、组织结构、内部资源、外部环境、市场需求、创新意识和风险承受能力等因素，综合评估战略规划与大模型应用的契合程度。

经费预算上，应综合评估预算的组成和分配与大模型所需的硬件设备、软件开发、数据采集与处理、人力资源、实验耗材、产品运营等方面的匹配程度。应用方应结合建设需求、资源成本、项目周期、风险与不确定性等因素综合评估现阶段相关的经费预算投入情况能否满足大模型建设和应用的需求。



来源：公开资料整理

图 6 大模型发展能力就绪度评估指标体系

（三）挖掘各类软硬件资源需求，统筹规划业务发展目标

应用方应全面分析基础设施、数据资源、算法模型、应用服务、安全可信等大模型发展要素需求，统筹规划大模型所需各类资源，并评估现有基础设施、信息系统、业务系统对资源需求的满足程度。通过深入挖掘大模型建设、应用和管理需求，并结合企业的发展规划和技术现状，为大模型的选型决策、方案设计、建设实施提供有力参考。

1.根据业务发展需求确定服务场景及目标

应用方应结合战略定位、业务发展、组织管理确定大模型的服务场景及目标。战略定位上，应根据企业属性、发展定位及经费预算等方面，深入分析自身在发展人工智能相关的战略新兴技术和创新新型应用上的需求。业务发展上，应充分分析当前业务发展的优势与不足，全方位探索借助大模型实现产品迭代和服务升级的有效途径，挖掘差异化和个性化服务的价值空间。组织管理上，应加强内外部资源整合，借助大模型实现业务流程的全面优化和高效管理，从而加快数字化转型和智能化改造的进程。

2.根据服务场景及目标确定大模型技术路径

应用方应根据服务场景及目标确定大模型在参数结构、技术能力、部署运维等方面的需求。在模型参数和结构上，应根据业务需求与目标、计算资源、数据特性等因素来确定模型结构和参数规模。在模型技术能力上，应考虑算力资源、技术能力、应用水平以及行业要求等因素，确定大模型的准确性、稳定性、鲁棒性、泛化性、可解释性等指标要求。在模型部署运维上，应考虑业务需求、技术架构、运维资源以及安全防控等多种因素，确定大模型在部署运维过程中的开放性、兼容性、便捷性、灵活性、可扩展性、可观测性、可维护性等指标要求。

3.根据大模型技术路径确定基础设施需求

应用方应根据大模型技术路径，确定大模型训练、推理所需的软硬件设施。硬件设备方面，应根据模型参数规模、模型结构复杂度、

分布式训练策略、数据资源规模等因素确定算力需求。应根据网络架构、带宽、稳定性、存储容量、存储速度等指标确定网络和存储需求。软件设施方面，应根据业务目标、兼容性等因素确定深度学习框架、工具插件、操作系统等软件需求，以构建完善的大模型开发与应用环境。

4.根据大模型技术路径确定数据资源需求

应用方应根据大模型技术路径，确定数据的来源、类型、规模及质量等需求。数据来源和类型方面，应根据业务目标、应用场景、软硬件资源、数据隐私与安全等因素，确定采用公开数据、自建数据或共建共享数据，以及确定文本、语音、图像、视频等数据类型。数据规模方面，应根据计算资源、模型结构、参数规模、业务需求等因素确定数据规模。数据质量方面，应根据大模型的技术指标及业务要求，确定准确性、完整性、一致性、可用性、多样性和时效性等数据质量要求。

5.以风险控制为导向确定安全可信要求

应用方应重视风险控制，根据不同行业和应用场景确定大模型的安全可信要求。安全性方面，应梳理在应用大模型过程中可能存在的数据风险、模型风险、应用风险、服务风险等风险控制需求，制定相应的安全保障机制和策略，确保大模型应用过程和内容产出合法合规。可信性方面，应梳理大模型在稳定性、鲁棒性、公平性、公正性、伦理道德等方面的要求，充分考虑基于大模型的产品服务与人类价值观的对齐需求，通过透明可控的算法、高质量数据集强化可信能力。

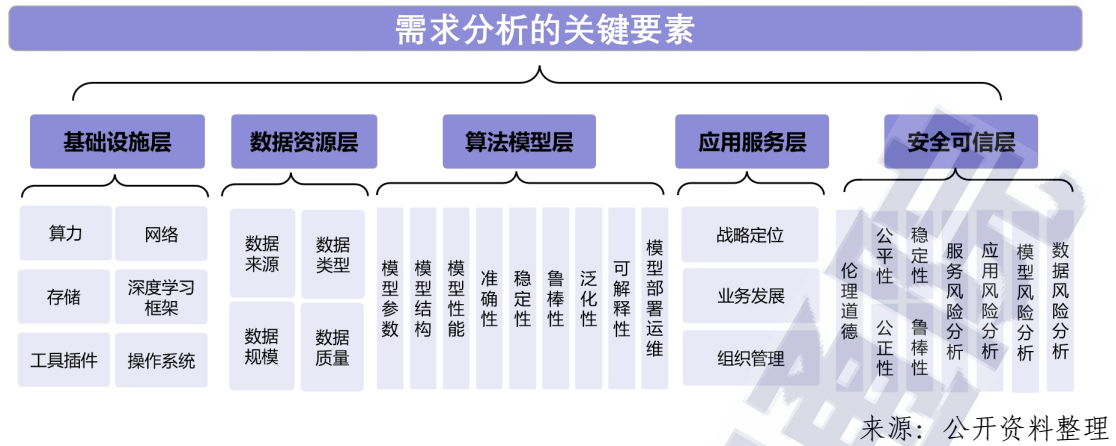


图 7 需求分析的关键要素

三、筑牢大模型技术底座

（一）剖析大模型关键落地路线，科学确立技术选型原则

大模型能力不断提升，应用场景不断丰富，然而在落地应用过程中仍然面临技术、数据、服务、安全等多方面的困难和挑战。因此，应用方在综合评估自身能力和需求后，应根据自身行业属性、业务场景、资源储备等情况选择适合自身战略规划的技术路线，设计科学合理和切实可用的大模型落地方案，为后续大模型的研发和测试夯实基础。

应用方应通过完善的方案设计和系统的研发测试筑牢大模型技术底座。方案设计方面，完善的方案设计可以帮助应用方进一步明晰自身业务需求、优化资源配置、合理管控风险，应用方应综合考虑基础软硬件、数据集构建、模型选型等方案的设计，满足大模型训练和测试的要求。研发测试方面，系统的研发测试可以帮助应用方及时发现大模型存在的问题和不足，为模型的改进和优化提供明确的方向。应用方应构建训练推理平台、应用服务平台，确立可靠的安全风险处

理机制，实现大模型的稳定高效和安全可信。

（二）设计大模型智能系统架构，制定科学有效解决方案

大模型的能力构建是一项复杂的系统性工程，往往牵一发而动全身，因此需要根据应用方的切实需求，设计并验证科学合理、高效可行的大模型建设方案。在方案设计时，应重点考虑基础软硬件选型、数据集构建、模型选型和设计、应用服务、风险控制等因素。

1.面向大模型的基础软硬件方案设计原则

应用方应根据大模型的训练、推理和部署需求，设计基础软硬件方案。硬件选型方面，应用方应根据算力需求、数据规模、网络需求利用现有资源、采购或租用计算核心组件、存储设备、网络设备，为大模型的后续开发打下坚实基础。以大模型推理所需算力为例，其算力需求估算需要综合考虑模型大小、量化方式、访问并发量等，并结合 AI 芯片显存大小推算所需芯片数量。软件设计方面，应用方可选择 PyTorch、TensorFlow、PaddlePaddle、MindSpore 等深度学习框架，借助其丰富的算法库和工具链简化大模型的设计、训练和微调过程。同时，应用方应采用兼容性和扩展性良好的向量数据库等软件系统和操作系统以确保大模型训练和部署的稳定高效。

2.面向大模型的数据集构建方案设计原则

应用方应根据大模型训练、微调、测试等需求，设计数据集构建方案。数据采集方案上，应根据不同的场景和目的，灵活选择收集线上、线下、公有、私有数据，扩展大模型数据规模及多样性，有效支

撑模型训练和微调。**数据标注方案上**，可根据自身数据规模和特点，以及对数据标注效率和准确性的要求，选择采用人工标注、半自动标注、全自动标注等方式。比如，在医疗图像、材料科学、生命科学等数据专业性较强的领域上，可选择人工标注或半自动标注方式。在通用目标识别、检测、跟踪等数据通用性较高的领域，可采用全自动标注方式。**数据管理方案上**，应根据数据的重要性和敏感性设计分类分级的原则，对不同数据采取不同访问权限等措施，以确保数据使用的高效性和安全性。

3.大模型选型和建设方案设计原则

应用方应根据业务目标和资源就绪情况，设计大模型生态、模态、架构、能力、开发、优化和部署方案。**模型生态上**，可根据开发成本、开发周期、安全性等要求选择开源或闭源两种软件生态。**模型模态上**，可根据自身业务需求、产品形态、数据模态等，选择开发语言大模型、视觉大模型或多模态大模型。**模型架构上**，可根据自身数据规模、算力规模、部署方式，选择合适的大模型网络架构、优化算法和损失函数。**模型开发上**，可根据自身开发能力、项目周期、技术要求、验收指标等选择全量开发、微调开发。**模型能力上**，可根据任务的类型和难度，合理选择模型的参数规模和训练策略。**模型优化上**，可以通过提示工程、检索增强生成、微调等方式对大模型进行优化，以满足应用方特定需求。**模型部署上**，可根据自身行业的合规性要求、数据敏感性、数据和算力规模等选择公有云、私有云或混合云三种不同的模型部署策略。

表 1 模型生态主要选型方案

选型方案	特点
开源模型	降低开发成本、加快开发速度，适用于基础研究、个人开发、快速验证、技术共享等场景
闭源模型	可满足定制化、个性化、安全性需求，适用于安全程度要求高、个性化需求大、商业机密性强的场景

表 2 模型优化主要选型方案

选型方案		特点
检索	检索增强生成	可辅助模型进行特定领域知识问答，可在一定程度上缓解模型幻觉问题并增强专业性
微调	全量微调	能较好地拟合数据集、适应任务需求，学习能力强但训练效率较低，适用于微调数据较多的场景
	高效微调	通过减少需要更新的参数量或改变参数更新的方式来提高微调效率，进而减少对计算资源的依赖，并减少训练时间
	指令微调	通过提高大模型的意图理解能力，以提升问答过程的对齐程度
	提示微调	利用特定的输入提示引导模型生成强相关的内容，可激发大模型的补全能力

表 3 模型部署主要选型方案

选型方案	特点
公有云	可根据实际需求动态扩展或缩减计算和存储资源；通常具有较低的开发和维护成本，避免了应用方自建和维护基础设施
私有云	可有效降低敏感数据泄露的风险，能提供更灵活的管理运维方式，可充分利用现有基础软硬件资源
混合云	结合公有云和私有云部署的优势，有效帮助应用方应对突发需求和业务变化，可灵活调整部署方案

表 4 模型规模主要选型方案

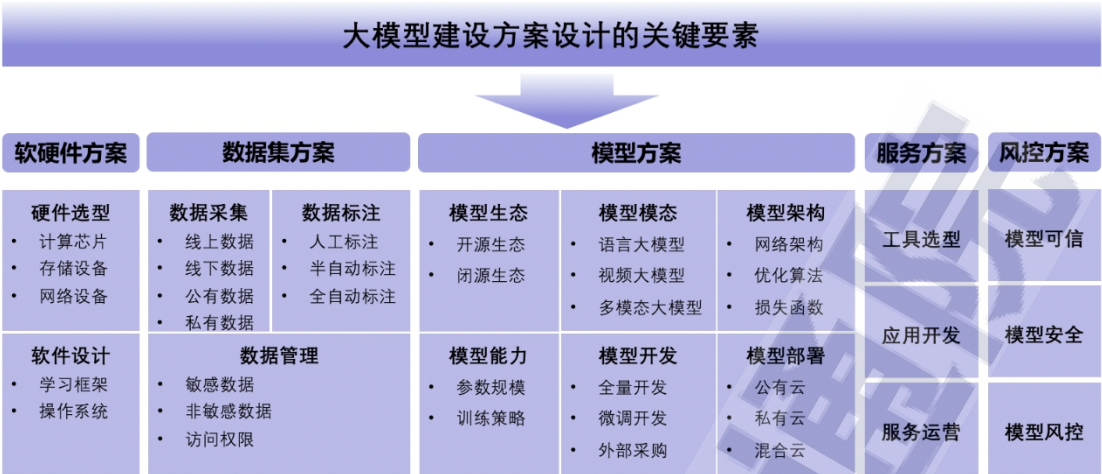
选型方案	特点
百亿及以上模型	适用于对生成、理解、推理、决策的准确率要求较高的复杂任务，训练和推理算力需求较大
十亿及以下模型	适用于简单任务，训练和推理算力需求较小，可在边缘设备及端侧部署

4.面向大模型的应用服务方案设计原则

应用方应根据业务需求，设计大模型的工具选型、应用开发、服务运营方案。多模型路由方案上，可根据业务特点及所属领域，动态选择不同类型、不同领域、不同厂商的大模型，通过集成多个大模型的技术能力以支撑上层应用提供更加精准、稳定的服务。工具选型方案上，可选择 LangChain、智能体等技术或工具与大模型深度融合，使得大模型信息检索更精准、应用构建更高效、行为决策更智能。应用开发方案上，可根据大模型的生成创造、逻辑推理能力进行人工智能原生应用开发，建立用户与大模型之间便捷可靠的交互通路。服务运营方案上，应设计相应的大模型服务运营平台和反馈机制，建立高效的运营流程，以保障模型在各个业务场景的平稳运行。应用方可根据业务和市场的变动，及时调整大模型服务运营方式。

5.面向大模型的风险控制方案设计原则

应用方应根据行业、数据、模型等特点，设计安全可信体系以及风险控制方案。可信方案上，应确定大模型的鲁棒性、透明性、可解释性、安全性、公平性等指标，实现大模型系统的数据安全可信、系统行为可追责、算法模型可解释。安全方案上，应加强数据、模型和应用的安全防护，更好地借助人工智能技术保障业务发展和战略布局。风控方案上，应对大模型应用中的风险进行评估和预测，列出风险清单、进行风险识别、完成风险处置，制定风险管理策略和应对措施。



来源：公开资料整理

图 8 大模型建设方案设计的关键要素

（三）系统研发大模型技术底座，全面开展功能性能测试

应用方应根据前期大模型设计方案进行模型研发，构建大模型训练推理平台和应用服务平台。应用方应开展大模型功能和性能测试，全面了解模型在不同场景下的表现并及时发现和解决问题，进而提高模型的稳定性和可靠性。通过研发和测试的多轮迭代，进一步夯实大模型技术底座，为上层应用服务打下基础。

1.构建高效易用的大模型训练推理平台

应用方应在硬件设备、工具链、软件系统方面实现资源整合，构建大模型训练推理平台。硬件设备协同方面，平台应合理配置训练大模型所依赖的算力芯片和高带宽网络设备等硬件资源，实现显卡之间的数据并行和运算协同。工具链支持方面，应构建全流程开发工具链为大模型提供丰富的通用组件，实现模型的定制化调优和人工智能原生应用的高质量开发。如 LangChain、LlamaIndex 工具链能够简化大模型的应用开发流程，打造更智能、高效、低门槛的人工智能原生应

用开发新范式。**软件系统兼容方面**，应实现大模型在不同操作系统和平台上部署的高度兼容性，从而保证大模型稳定运行。百度智能云千帆大模型平台预置了丰富的模型资源，为用户提供了全流程开发工具链，其推出的训练推理方案简化了开发者的工作流程，提升了模型训练的效率。

2.构建面向大模型的高质量数据集

应用方应通过数据采集、预处理、标注等方式构建高质量数据集。

数据采集方面，可通过采集质量高、规模大、种类全的数据以确保模型能够学习到全面完整的知识，进而提高模型的表征能力和泛化能力。

数据预处理方面，应借助必要的**数据清洗**手段来解决数据中存在的缺失值、异常值、噪声数据、重复数据等质量问题。**数据标注方面**，应将待标注的数据划分为不同的标注任务，根据不同场景的数据需求选择不同标注方式，确保数据标注的准确性、完整性、一致性。

3.构建高质量的算法模型资产

应用方应持续开展模型研发和测试，确保满足准确性和稳定性等指标要求。**模型研发方面**，针对特定需求采用合适的研发方案完成大模型训练，通过微调、RAG、RLHF、提示工程等技术使模型更好地泛化到下游任务，实现既定目标。**模型优化方面**，可利用模型剪枝、模型量化、知识蒸馏、低秩因子分解等模型压缩手段，降低大模型的存储和计算成本。**模型测试方面**，可采用准确率、召回率等指标全面测试模型性能，为模型的优化和改进提供依据。同时可通过对抗攻击测试、隐私保护测试、持续监控和反馈等方式有效测试模型鲁棒性、

安全性和稳定性。

4.开发大模型一体化应用服务平台

应用方应根据平台管理和服务性能需求构建大模型应用服务平台。平台管理方面，应综合利用智能应用管理平台、插件管理平台、低代码管理平台等构建新型应用服务系统，形成智能、高效、自主的企业大脑。服务性能方面，应重点关注服务稳定性和服务响应时间，确保大模型的高并发处理能力，保持高效的性能和稳定的服务。例如，GPT-4o 对音频输入的响应时间可以短至 232 毫秒，与人类在对话中的响应时间相近，提供了更即时的互动体验；GPT-4o 为付费用户提供高达 5 倍的消息限制数量，进一步提高了用户使用体验。

5.构建面向大模型的风险控制体系

应用方应在可信算法、安全机制、风险处理机制方面综合发力，确保大模型应用全流程的安全可信。可信算法方面，可通过逻辑正确验证、人类价值观对齐、计算体系结构和计算模式等方面的技术创新，从根本上提高计算的正确性和安全性。安全机制方面，应构建安全平台帮助大模型提升自身防御性，通过定期对模型和系统进行安全审计和运行监控及时发现安全问题，有效消除安全威胁。风险处理机制方面，应建立风险处理机制实现对模型潜在风险的识别、评估、控制。通过采用基于风险的分类分级规制方法，实现上下游参与者的合作共治。

为实现大模型的集约化开发，降低大模型的训练和优化门槛，应用方可以考虑应用部署大模型开发平台，如百度智能云千帆大模型平

台。百度智能云千帆大模型平台结合大模型的训练、推理和部署需求，聚焦数据集构建、模型选型和设计、应用服务等多个方面，为大模型的开发、优化、管理、部署和集成方案提供支持，全面夯实大模型技术底座，提供有效的解决方案，帮助企业实现降本增效提质。

百度智能云千帆大模型平台落地实践

百度智能云千帆大模型平台作为一站式企业级大模型平台，提供了生成式 AI 生产及应用全流程开发工具链。

在平台方面，平台功能覆盖大模型全生命周期，为企业提供多种大模型，包括百度自研的 ERNIE 系列大模型和第三方开源大模型，此外还提供了从数据管理、模型训练、调优、部署调用和编排集成的全方位支持，包括数据集整理、模型精调、在线测试、模型评估等丰富的系统工具，帮助企业轻松构建和部署行业大模型。在应用场景方面，百度智能云千帆大模型平台面向不同的企业需求提供不同的功能服务。例如，可提供文心一言企业级推理云服务，并进行业务集成；提供在线微调训练功能，快速生成行业场景定制模型服务，满足用户特定需求。



来源：百度

图 9 百度智能云千帆大模型平台

四、革新大模型应用范式

（一）参照业务场景个性化需求，定制化调优专用大模

型

为满足特定领域、行业和任务的需求，应用方应面向不同应用服务定制化开发专用大模型。基础大模型具备强大的通用性和泛化能力，但在特定领域、行业和任务中难以满足细致化和专业化的要求。应用方应深入研究语言、语音、视觉、多模态等特定领域的数据特性和知识体系，细致分析金融、政务、教育、工业、农业、交通、文旅等具体行业的应用场景和业务流程。基于生成、对话、代码、翻译、质检等具体任务的目标，选择合适的大模型并从基础设施、数据资源、算法模型、应用服务、安全可信五个层面定制化开发面向不同应用服务的专用大模型。

应用方应建立多维度大模型评价机制，形成建设应用评估有效闭环。大模型的全方位评估对于开发验证、产品选型和能力提升都至关重要，通过构建覆盖基础设施成熟度、数据资源成熟度、算法模型成熟度、应用服务成熟度、安全可信程度等全链路效能评估体系，有助于全面了解模型的性能、分析模型的优缺点、推动模型的改进与创新，确保模型既能满足实际需求又能不断优化提升。

（二）开发大模型原生智能应用，实现大小模型协同赋能

应用方应将专用大模型转化为应用服务以满足市场的多元化需求，通过面向不同应用服务配备基础设施、构建数据集、实现模型微调、开发模型服务、满足安全可信要求来支撑部署多类智能应用和服务，精确匹配自身发展需求。

1.面向不同应用服务配置基础资源

应用方应采用资源调度、负载均衡、任务管理、容错处理等策略配置基础资源。资源调度方面，可采用动态资源分配、优先级调度、跨集群调度等策略合理分配和管理计算、存储和网络等资源，以满足不同应用开发的需求。负载均衡方面，应考虑计算和网络的负载均衡，通过异构计算资源利用、任务切片与分发等方式实现计算负载均衡，通过优化数据传输、分布式数据存储、网络带宽管理等方式实现网络负载均衡。任务管理方面，应借助任务管理工具灵活地调度、监控和优化任务执行进度，确保应用服务的稳定性和安全性。容错处理方面，应采用数据备份、自动重试、故障隔离等机制来保证应用服务的可靠性，确保在面对硬件故障、网络问题或其他突发情况时，系统能够保持稳定运行。

2.面向不同应用服务构建数据集

应用方应通过数据标注、数据回流等手段来构建面向不同应用服务的数据集。数据标注方面，根据特定领域、行业和任务的需求，收集专业出版物和行业报告等专业数据集，通过构建面向特定领域、行业和任务的提示和回应数据集，训练模型服从并对齐人类意图。数据回流方面，应从日志记录、用户反馈界面、API 调用等自动化工具和程序构建数据采集机制，将反馈数据重新整合到训练和微调过程中，以持续优化和更新模型。

3.面向不同应用场景实现模型优化

应用方应通过专业知识学习、模型压缩方式实现模型精调。专业

知识方面，通过在特定领域、行业 and 任务的数据集上对大模型进行微调或精调，使模型能够学习相应的专业知识，通过调整模型参数、选择合适的学习率和训练迭代次数来确保模型同时掌握通用知识和专业知识。**模型压缩方面**，应通过模型裁剪、知识蒸馏、模型量化等技术降低模型参数规模，减少计算量和存储需求以提高推理速度，通过大小模型协同的方式以更好适应不同的应用场景。

4.面向不同应用场景开发模型服务

应用方应从服务管理、服务运营、能力编排、服务协议等方面开发模型服务。**模型服务管理方面**，应对模型卡片、模型文件、模型调用和模型库进行全生命周期管理，以提供高质量的模型服务，保障模型更新与模型使用有序进行。**模型服务运营方面**，应通过服务质量监控、运营分析以及用户后台管理来帮助模型使用者和提供者实时了解服务动态，并监控模型服务运行稳定性，保障模型服务的可用性和安全性。**模型能力编排方面**，应基于模型服务的基础能力和各类插件进行编排，如 RAG 技术可通过检索增强提升模型的准确性和丰富性，AI Agent 应用可负责具体任务的调度、执行与反馈，知识库可为 AI Agent 提供决策支持。**模型服务协议方面**，应从模型服务质量、服务可靠性、数据保密性、服务计量准确性以及权责明确五部分制定模型服务协议，以统一衡量模型服务水平。

表 5 AI Agent 能力及特点

能力		特点
规划	子目标和任务分解	可将复杂目标分解为更小、可管理的子目标，从而高效处理复杂任务
	反思和完善	可对过去的行为展开自我批评和反思，从错

		误中吸取教训，并对未来步骤进行完善，提高任务完成的质量
记忆	短期记忆	通过学习上下文获得短期记忆的能力
	长期记忆	利用外部向量存储和快速检索以获得长时间保留和回忆信息的能力
工具	小模型	智能体依托大模型实现意图理解，并通过调用小模型、工具或检索数据库完成具体任务
	实用工具	
	检索数据库	
行动	workflow编排	理解意图并将任务拆分为若干个步骤，对工作执行过程进行编排
	反馈迭代	通过多次迭代，实现更高质量的输出，确保任务顺利完成并实现智能体的自我学习
协同	多智能体任务协作	多个智能体扮演不同的角色，协同完成复杂任务，提高工作质量和效率
	资源配置	通过优化资源配置、提高资源利用率，实现更高的适应性和经济性

5.满足不同应用服务的安全可信要求

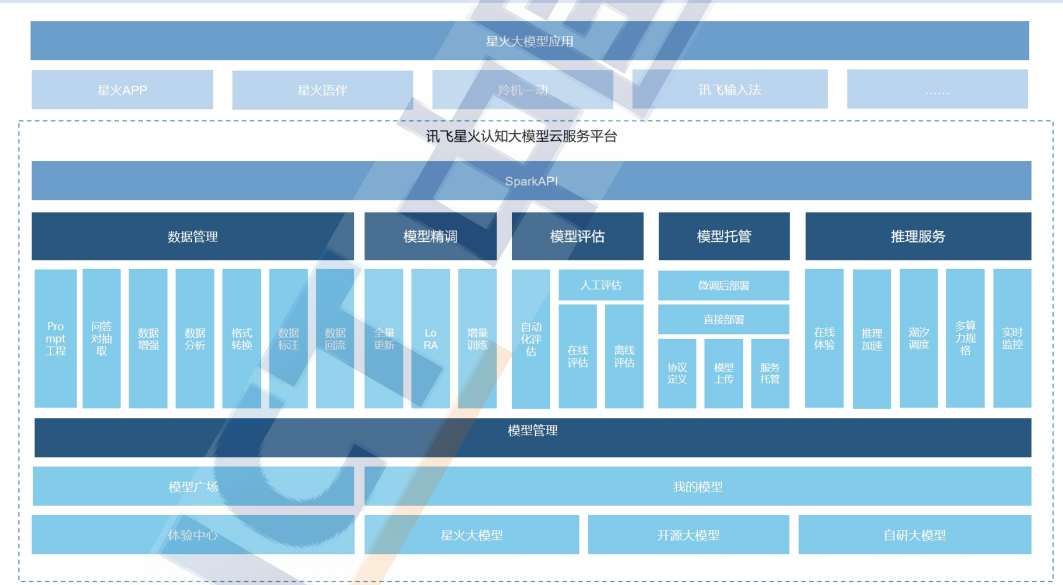
应用方在开发不同应用过程中应满足相应的安全可信要求。隐私保护方面，应采取数据加密、访问控制、技术监控等方式保护应用服务中的用户个人隐私信息。内容合规方面，应通过建立合规指南、内容过滤器和监控机制、定期评估和审查等方式确保应用服务中的信息和内容符合法律法规要求。应用安全方面，从应用设计、实现、运维、管理等多个角度重点关注身份验证、访问授权、进程间通信安全、代码安全、管理与审计安全等方面。服务无偏方面，在提供服务时应同等看待肤色、性别、地域等差异，在应用的设计和服务的提供过程中秉持透明性和包容性原则，持续监控和优化服务流程，以保障服务的公正性和平等性。

为深入用户真实场景解决实际问题，提高模型服务质量和管理能力，实现大模型的实用化、多元化发展，应用方可以考虑采用大模型

应用服务平台，如讯飞星火认知大模型云服务平台。讯飞星火认知大模型云服务平台根据垂直领域的个性化需求，渗透至多种业务场景，提供大模型智能应用的开发、运维、管理的一站式服务，并结合端云协同、边缘计算等能力，涵盖数据管理、模型精调、模型服务等多个角度，为大模型应用落地提供一站式解决思路。

讯飞星火认知大模型云服务应用案例

讯飞星火认知大模型云服务平台为星火认知大模型提供全面的生命周期管理服务。平台不仅涵盖了训练数据的管理，还包括了模型的预训练与微调，以及推理服务的工程化操作和快速部署能力。通过专注于 AI 生产过程的平台化以及 AI 技术的快速实际应用落地，讯飞星火认知大模型云服务平台致力于以 AI 技术为核心，推动系统性的创新。借助于云端、边缘计算、终端设备相结合的模式，实现快速服务交付。本平台支撑着公司教育、医疗、消费者、汽车等多个领域的广泛产品，为支持百万级开发者、亿级终端接入以及千亿级页面浏览量的高效生产平台，展现了较好的性能和广泛的适用性。



来源：科大讯飞

图 10 讯飞星火认知大模型云服务平台

（三）构建全链路效能评估体系，形成诊建用评有效闭环

应用方应建立多维度大模型评价机制，通过构建覆盖基础设施、

数据资源、算法模型、应用服务、安全可信等层面的全流程大模型应用成熟度评估体系，形成诊断、建设、应用和评估的有效闭环，有助于促进模型高效、准确、稳定地执行任务，同时也是衡量投入产出比、数转智改程度的重要依据。

1. 评估基础设施的成熟度

应用方应从资源配备情况与综合效能两方面来评估基础设施的成熟度。资源配备方面，应通过服务器、计算芯片、存储器等方面评估硬件设备的配比情况，通过向量数据库、操作系统、软件框架等方面评估软件设施的配置情况，通过模型开发和数据处理套件、科学计算和代码处理工具、大模型服务平台等评估工具平台的配备情况。综合效能方面，应从计算性能和综合能力两方面评估基础设施的综合效能，通过训练性能和推理性能评估基础设施的计算性能，通过兼容性、可靠性、稳定性、自主性等方面来评估基础设施的综合能力。

2. 评估数据资源的成熟度

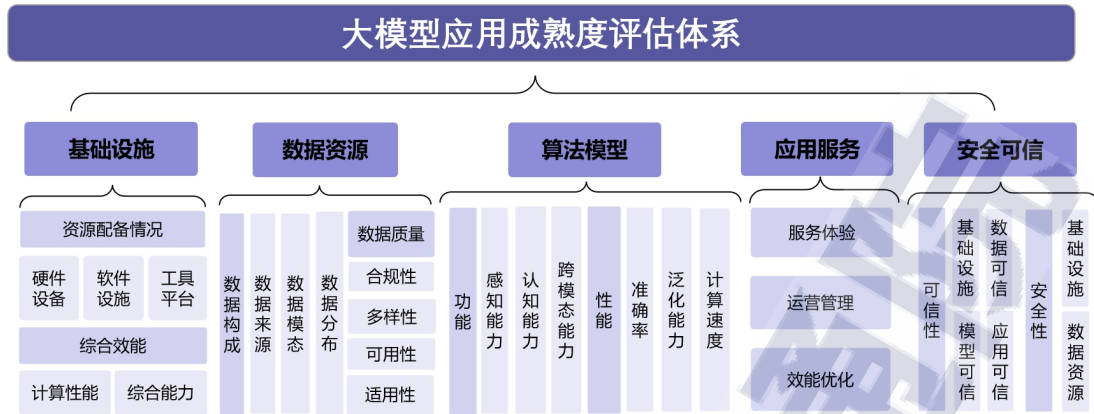
应用方应从数据构成和数据质量等方面来评估数据资源的成熟度。数据构成方面，应从数据来源、数据模态、数据分布等方面来确定数据的构成情况。数据质量方面，应从数据的准确性、完整性、一致性、关联性、冗余度、数量级、更新频率等维度来评估数据质量，从代表性、平衡性、多样性等方面来评估数据的多样性，从可访问性、格式兼容性等方面来评估数据的可用性，从保密性和合规性评估数据安全与隐私保护情况，从扩展性与兼容性评估数据的适用性，从成本效益和投资回报率评估数据效益。

3.评估算法模型的成熟度

应用方应从功能和性能两方面评估算法模型的成熟度。功能方面，应通过识别和分类任务评估算法模型的感知能力，通过生成、理解、推理、决策等任务评估算法模型的认知能力，通过文本、语音、图片、视频等模态数据评估算法模型的跨模态信息融合能力，进一步通过主观和客观方式来评估大模型的自主学习、自我优化、自我适应能力。性能方面，应从准确率、计算效率、并发路数、响应速度等方面评估算法模型的精度，从稳定性、鲁棒性、可扩展性、可重现性、微调方式、部署方式等维度评估算法模型的性能。

4.评估应用服务的成熟度

应用方应从服务体验、运营管理、效能优化等方面评估应用服务的成熟度。服务体验方面，应从稳定性、开放性、扩展性等方面评估大模型的服务质量，通过用户的交互体验、功能体验、情境体验等方面评估大模型的服务体验。运营管理方面，应从数据、调优、交付、流水线等方面评估运营管理的流程化程度，从流水线和工具链评估运营管理的自动化程度，从监控、反馈、迭代等方面评估运营管理的持续闭环情况。效能优化方面，应从场景渗透率、业务优化率、投入产出比等指标来评估大模型的效能优化情况。



来源：公开资料整理

图 11 大模型应用成熟度评估体系

5. 评估安全可靠程度

应用方应从可信性和安全性两方面评估大模型安全可靠程度。可信性方面，从稳定性、兼容性等方面评估基础软硬件的可信程度。从数据采集的可靠性、预处理的有效性、使用的可控性来评估数据的可信程度。从鲁棒性、稳定性、透明性、公平性、可追溯性等方面评估模型的可信程度。从稳健性、可控性、合规性、可反馈性等方面来评估服务的可信程度，从准确性、真实性、可追溯性、合规性、价值观对齐等方面评估内容的可信程度。安全性方面，应从技术能力和管理制度要求来评估基础软硬件的安全性，从完整性、一致性和准确性等指标来评估数据在使用过程中的安全性。

五、构建大模型管理体系

（一）梳理国内外模型治理要点，明晰运营管理体系价值

应用方建立健全大模型治理体系，既是满足外部要求，也符合内部发展需要。当前国内外均高度重视大模型治理，2023 年 6 月，欧

盟发布《人工智能法案》，提出对人工智能模型进行分级分类的治理思路。2023 年 7 月，中央网信办等七部门联合颁发《生成式人工智能服务管理暂行办法》，明确了提供和使用生成式人工智能服务的总体要求，并对生成式人工智能服务提出了分类分级的监管要求，标志着我国迈出了加快人工智能算法模型立法的重要一步。**企业级大模型治理体系建设备受关注**，企业应通过建立完善的自我监管机制、强化企业社会责任、公开透明的责任报告、促进多方利益平衡等措施实现对大模型的有效治理。通过与政府和监管机构建立对话机制、紧密与产学研用各方合作、参与或创建多方协同的行业联盟、关注用户反馈与需求等方式建立健全企业级大模型运营管理体系。

应用方建立健全大模型运营管理体系，是持续实现降本增效提质、提升核心竞争力的重要方式。应用方为达成借助大模型实现战略规划、助推业务发展等目标，可通过规范大模型管理的制度、流程和方法，实现对大模型的技术研发、能力测试、应用开发、能效评估等各项工作进行协调、控制和优化。应用方建立和实施大模型运营管理体系，既是长期发展大模型的基础门槛，也是规范内部流程、提高业务效率和服务质量、降低运营成本，提高公信力和核心竞争力的关键举措。应用方可通过持续构建并完善基础设施运营管理体系、数据治理体系、算法模型治理体系、应用服务运营管理体系、安全可信体系，最终达到建立健全大模型运营管理体系的目的。

（二）实时监测大模型运行过程，确保高效稳定提供服务

大模型部署后依赖收集监测数据和用户反馈以进行持续更新迭代，从而保证良好的运行状态和服务体验。大模型的全面监测包括但不限于监测基础设施的运行状态，监测数据的链路、质量和漂移，监测算法模型的运行情况，监测应用服务的运行状态。通过多方埋点获取实时监测数据，可进一步支撑构建风险预警防范机制，通过对突发事件采取快速反应和有序处理以降低损失，保障大模型应用的高效性、稳定性、透明性和可观测性。

1. 监测基础设施的运行状态

应用方应从硬件和软件两个方面实时监测大模型基础设施的运行状态。硬件方面，应实时跟踪和监测大模型硬件设备的资源利用情况，及时识别资源利用瓶颈和调度优化潜力。通过监测内存、计算单元等使用情况计算得到资源利用率，通过监测响应时间、处理速度得到系统运行效率。软件方面，应实时跟踪监测错误率、异常日志、进程或线程的状态以掌握系统与应用异常情况，通过监测可用性百分比、故障时间得到系统的可用性与连续性指标，及时发现软件系统和服务中的异常状态和潜在风险，保障应用服务的稳定运行。

2. 监测数据的链路、质量和漂移

应用方应从数据链路、数据质量、数据漂移等方面对数据进行实时监测。数据链路方面，应实时监测数据链路的连通性、传输速率和数据流量等指标以诊断数据链路的状态，及时分析网络故障或瓶颈，保证数据顺畅传输。数据质量方面，应监测数据的及时性、有效性、空置率等质量指标，及时发掘和处理数据质量问题。数据漂移方面，

应监测未知数据比例、数据分布、特征关联性数据漂移情况，及时发现数据分布和特征的变化。

3. 监测算法模型的运行情况

应用方应从模型性能、模型服务和模型漂移等方面对模型算法进行实时监测。模型性能方面，应监测模型输入输出的准确率、召回率、F1 值、BLEU、ROUGE、均方误差、平均绝对误差、响应时间等性能指标。模型服务方面，应监测模型服务调用的成功率、并发路数、处理请求的吞吐量和延迟等指标，通过实时监测发现模型服务过程中的异常情况，并及时采取相应的措施进行处理。模型漂移方面，通过在固定时间周期内监测 PSI、CSI 等漂移指标来分析模型漂移情况，通过集成学习、主动学习、特征选择和降维、定期更新模型等方式保证模型长期稳定运行并维持预测的准确性。

4. 监测应用服务的运行状态

应用方应从过程和结果两方面来监测应用服务的运行状态。运行过程方面，应实时跟踪和监测应用服务中数据资源质量、模型性能以及服务稳定性、自动化程度、问题处理速度，及时发掘运行过程中潜在的问题和瓶颈，从而快速响应，确保模型在各种应用服务中高效稳定运行。运行结果方面，应通过监测 token 数量、问题解决率、预测准确率、响应时间、插件贡献度等指标以判断大模型在业务应用中是否达到预期的效果和目标，从而形成服务上线、监测、优化、再上线的螺旋上升。

5. 监测大模型的安全可信情况

应用方应通过监测手段实现安全防控和应急响应，保障大模型的安全可信。在大模型应用部署的全流程进行埋点监测，不仅能实时获取运行数据，还能追加安全防控和应急响应措施可有效预防和减少安全事件的发生，以保障业务连续稳定运营。安全防控方面，应通过持续的安全监测来识别潜在风险，通过加强网络和物理安全防护、进行安全意识培训以及实施数据加密和访问控制等措施来防范安全事件的发生。应急响应方面，应实时监测数据并精确启动应急响应措施，预先制定应急响应计划，建立跨部门应急响应团队，制定大模型响应和恢复明确流程、定期进行应急演练以确保响应效率。

（三）建立健全大模型管理体系，保障业务高效稳定开展

应用方应通过建立健全大模型运营管理体系，提高大模型综合治理水平和应用成熟度，进一步提升自身品牌影响力和核心竞争力。具体而言，应用方可通过持续构建并完善基础设施运营管理体系、数据治理体系、算法模型治理体系、应用服务运营管理体系、安全可信体系，通过建设五位一体的运营管理体系，最终形成健全的大模型运营管理体系。

1. 构建基础设施运营管理体系

应用方应从灾备管理、性能优化、资产管理三个方面构建基础设施运营管理体系。灾备管理方面，应在故障事前预防、故障事中调度、故障事后改进三个关键环节建立灾备管理机制，对专业系统、生产数据、重要资料等进行备份，设计业务系统时应重点考虑冗余等措施，

以确保系统能够持续运行和快速恢复。**性能优化方面**，应采用自动扩缩容技术、资源动态调整技术、架构和算法优化技术、成本效益分析方法、混合云部署策略等实现性能与成本的最佳平衡。**资产管理方面**，应建立统一的基础设施资产管理平台，集成监控、报警、日志分析和资源调度等核心功能，管理内容包括资产录入、资产转移、资产维修、资产借用、资产停用和资产退出等日常管理，包括计提折旧、折旧年报、资产减值准备、资产价值重估等折旧管理，包括分类明细统计报表、部门明细统计报表、新增资产统计报表和退出资产统计报表等报表管理。

2.构建数据治理体系

应用方应从大模型相关的数据架构、质量、应用、安全等方面建立健全数据治理体系。在数据架构方面，使用结构化的方式对大模型训练、微调、测试数据集进行综合分析,组织形成数据架构。根据数据的来源、模态、类型、质量等属性进行数据分布的统计。建立各应用系统、各部门之间的数据集成共享机制 ,促进组织内部数据的互联互通。在数据质量方面，明确数据质量目标,根据业务需求及数据要求制定数据质量的评价准则,包括衡量数据质量的技术指标、业务指标以及校验方法。在数据应用方面，对大模型训练和测试数据集进行统一加工和分析,对应用服务中的监测数据进行分析或建模,对已有数据集进行有选择的对外开放,同时引入外部数据集。在数据安全方面，通过对数据访问的授权、分类分级的控制、监控数据的访问等进行数据安全的管理,满足数据安全的业务需要和监管需求,实现组织内部对

数据全生存周期的数据安全的管理。

3.构建算法模型的治理体系

应用方应在模型更新、模型部署、模型管控等方面构建算法模型治理体系。模型更新方面，应对模型版本的变化开展有效的管理和跟踪，明确记录和审核每次版本更新。应建立质量审查机制，通过引入自动化测试、性能评估、用户反馈等措施及时发现模型问题，持续优化模型性能。模型部署方面，应通过灰度发布、回滚机制等方式进行模型部署和更新，通过模型停用和资源释放确保资源的合理利用。应对多个模型服务进行有效调度、封装和管理，对基于模型的应用服务进行服务编排、服务托管、服务监控、服务告警，并根据需要动态调用相关插件。模型管控方面，应对各类模型的访问和使用权限进行分级分类，应支持模型文件的注册、上传、下载、删除等操作，应支持创建新版本、回滚到历史版本、比较版本差异等操作，应提供操作日志、技术文档、模型超参数等信息以支持模型审计。

4.构建应用服务运营管理体系

应用方应从可观测性、可审计性、可持续性三方面构建应用服务运营管理体系。可观测性方面，应集成日志、指标、组件依赖、自定义监控、告警与通知等工具以展示大模型内部运行过程，以提高大模型的可解释性和可控性。可审计性方面，应通过特征重要性、模型敏感性、模型可视化、模型解释器实现对模型的全面审计，确保大模型的内部结构和行为可以被理解和解释，以提高大模型的透明度和可信性。可持续性方面，应开展成本度量、分析和优化，量化模型的投

入产出比，应综合考虑创新能力、业务战略、市场需求、风险管理等多方面以制定可持续性发展的应用服务管理体系。

5.构建大模型安全可信体系

应用方应从基础软硬件、数据资源、算法模型、应用服务等层面构建大模型安全可信体系。在基础软硬件层面，应构建硬件安全性、可靠性、兼容性等硬件安全可信体系，以及软件安全性、兼容性等软件安全可信体系。在数据资源层面，应构建数据资源管理体系以保证来源透明性、过程可控性、存储可靠性，进而保证隐私保护、过程安全、销毁安全等使用可控性。在算法模型层面，应构建算法模型管理体系保证样本扰动、分布外泛化、对抗鲁棒性等模型鲁棒性、模型可解释性以及模型透明性，保证输出结果的公平性。在应用服务层面，应构建服务的稳健性、安全性、合规性、可反馈性等服务安全可信体系，构建内容准确性、真实性、可追溯性、合规性、价值观对齐等内容安全可信体系。

六、大模型发展趋势展望

（一）探索大模型架构优化方案，带动技术应用双重涌现

当前，大模型的底层架构 Transformer 逐渐显现多种问题，业界正积极探索底层架构的替代方案，新的架构研究思路将带来技术方案的全新升级，形成新一代智能底座。在智能底座的支撑下，各行业积极融入新技术，促使原生应用得以快速涌现，衍生出自主型和辅助型两种原生应用类型，不断渗透到多种应用场景，带来更好的用户体验。

1. 底层架构迭代创新，全力突破模型发展瓶颈

探索大模型底层架构优化方案将推动大模型技术长足发展。现有底层架构问题日益凸显，Transformer 架构稳居核心地位的同时，在计算成本高、可扩展性低、可解释性不足等方面的局限性逐渐显现。新型底层架构正在萌芽，探索新的模型架构、优化算法和训练策略以及混合架构模型已成为未来重点研究方向之一。Mamba、RWKV、InternImage 等新型模型架构开始出现，并出现 MOE 模型、PEG 模型等新的研究方向。新型模型架构不断迭代革新，将激发出全新的技术方案，为大模型的应用落地奠定坚实的技术发展基础。

2. 智能底座推陈出新，带动原生应用涌现

基于大模型的智能底座将驱动新型原生应用涌现。智能底座推陈出新，创新出 Megalodon 新型模型架构及全新神经网络架构 KAN 等。基于 Mega 架构改进的 Megalodon 能够处理无限上下文，有效提高大模型在长上下文上的训练推理的质量和效率。KAN 受到 Kolmogorov-Arnold 表示定理的启发，提出全新架构思路，具有更高的准确性、可解释性和交互性。以 Agent 为代表的自主型原生应用方面，大模型应用的高度自主性、智能决策、个性化服务、高效协作等优势为智能体等自主应用奠定了技术基础，深刻影响人类生产生活方式，将带来更加智能、高效的服务体验。以 Copilot 为代表的辅助型原生应用方面，Copilot 应用能够辅助用户完成多种复杂任务，大幅提高工作效率。辅助型原生应用在准确性、可靠性、高效性和平台兼容性等方面具备潜力和优势。自主型和辅助型原生应用优势互补，通

过将人工智能技术深度融入业务场景，将进一步释放模型价值，加速大模型的应用落地进程，开启智能时代的新篇章。

3.模型技术持续更迭演进，强力助推大模型落地应用

模型技术的持续更新升级，为大模型的落地应用开启新途径，同时也引起了产业界对架构兼容性和平台适用性的广泛关注。**架构兼容性方面**，模型技术的快速发展促使技术方和应用方深入考虑底层架构兼容性和可迁移性问题，用以匹配不同的大模型。当面对底层架构迭代更新或创新时，高效实现与现有设施的兼容和扩展，达到快速应用的目标。**平台适用性方面**，不同的应用开发方式对应用平台的适用性产生不同程度的影响，考虑模块化设计策略，以适应不同应用的集成，大幅提高开发效率。

（二）紧抓行业数字化转型机遇，全方位打造新质生产力

目前大模型正逐步成为金融、教育、政务等多个行业的新型技术基座，能够赋能多个应用场景，加速千行百业的数智化进程。未来大模型将持续提升技术能力，成为打造新质生产力的技术加速器。

1.众多行业数字化转型需求强烈，大模型落地路径需定制化

各行业抢抓数字化转型机遇，将进一步促使大模型应用能力向纵深扩展。一方面，行业特点凸显、场景多元化。在相关政策引导及市场需求的驱动下，金融、政务、教育、传媒等行业正加速应用大模型，行业需求重点和应用场景各具特点。例如，金融业应用大模型既需实

现降本增效，也需注重服务安全。当前金融大模型正深度赋能银行客服、质量检查、坐席运营等场景。而传媒行业应用大模型则注重内容创意与内容质量。传媒大模型可生成数字虚拟主播、创意视频，重塑音视频产品生产流程。另一方面，大模型落地路线应贴合行业实际需要。在大模型建设、管理、应用等不同阶段，技术方应切中行业应用关键点，分门别类梳理大模型落地应用具体路径。例如，部分行业如金融、政务需要模型输出内容保证真实准确，技术方需对训练、微调数据集进行质量把控。而另有部分行业如传媒、家居需要在终端植入大模型，需明确大模型在端侧部署的工程化路径。

2. 夯实人工智能赋能底座，助力新质生产力高质量发展

大模型作为人工智能的前沿技术，能够成为打造新质生产力的中坚力量，成为实体行业的“智能助手”。行业应用上，以大模型为代表的人工智能引发的新一轮科技革命和产业变革正向纵深发展，将为产业带来巨大的机遇和颠覆性的挑战。大模型行业应用掀起新型工业化浪潮，坚持以信息化带动工业化转型，加快建设现代化产业体系，促进社会经济高质量发展。可持续发展上，大模型可持续赋能能源消耗优化、废物管理循环利用、生态评估、生产过程监控等场景，从多个层面加速绿色制造。供给侧、需求侧、基础侧各方联动共同打造面向工业领域的大模型，而大模型也将高效凝练和开放工业应用场景，深化工业数据开发利用。未来，大模型或将进一步提升工业生产力、优化生产关系。

（三）加强引导大模型可信发展，对齐人类偏好及价值

观

随着大模型的应用范围进一步扩展，大模型的安全可信问题日益显著，价值观冲突事件时有发生。面对这一新形势，已有应对措施存在滞后性，仍需采取多种手段提升其可控可信能力。针对大模型可信可控和价值观对齐等问题研制理论方法与实践指南势在必行，面向大模型安全可信、价值对齐的共识需尽早形成。

1.多措并举引导大模型向安全可信发展

安全可信贯穿大模型开发及应用全过程，标准及技术规范将起规范促进作用。当前大模型仍存在数据隐私风险、模型稳定性不足、不当应用的问题，未来需从技术手段、规范标准两方面共同发力，持续推动大模型安全可信。**技术手段方面**，大模型安全、可信、可控的测试验证体系需尽早形成，贯穿大模型研发、训练、推理、应用全过程。同时，除现有的红白测试、攻防测试、偏见检测、鲁棒性验证外，大模型可信安全测试技术需不断迭代，规避黑盒带来的风险。**标准规范方面**，未来需建立健全大模型监管机制，提升大模型安全可信标准及技术要求质量。规范和标准均需引导大模型的研发和应用，确保技术的负面影响最小化，保护利益相关者的正当权益。

2.理论与工程实践共促大模型实现价值观对齐

理论基础与工程实践携手并进，实现大模型与人类价值观的深度对齐。当前价值观对齐存在数据偏见、隐私保护、技术与法律滞后的问题，未来仍需在如下方面做出努力。**理论研究方面**，针对价值观对齐的社会属性，需要明确大模型对齐框架中的价值对齐范围，以及如

何保证大模型的价值观公平性，确保人工智能系统的可持续发展；针对价值观对齐的技术属性，需要通过人类反馈强化学习、可扩展监督学习等技术方案保证大模型的训练目标函数与人类价值目标一致。**工程实践方面**，在大模型落地应用前，需要进行对齐评测，通过安全测评、可解释性分析和人类价值验证等方法充分识别价值对齐风险。在大模型的落地应用过程中，需要接受政策法规的监管，采用用户违规行为监测、第三方审计、内容来源工具等多种治理措施，保障大模型的安全应用和部署。

缩略语

AI	Artificial Intelligence	人工智能
API	Application Programming Interface	应用程序编程接口
GPU	Graphics Processing Unit	图形处理器
KAN	Kolmogorov–Arnold Network	科尔莫戈洛夫-阿诺德网络
MOE	Mixture of Experts	混合专家模型
RAG	Retrieval-Augmented Generation	检索增强生成
RLHF	Reinforcement Learning from Human Feedback	人类反馈强化学习

编制说明

本研究报告自 2024 年 1 月启动编制，分为前期研究、框架设计、文稿起草、征求意见和修改完善五个阶段，面向大模型落地应用的技术供应方、服务集成方、应用需求方开展了问卷调查和深度访谈等工作。

本报告由中国信息通信研究院人工智能研究所撰写，撰写过程得到了人工智能关键技术和应用评测工业和信息化部重点实验室、华为技术有限公司、北京百度网讯科技有限公司、科大讯飞股份有限公司、中国科学院自动化研究所、北京邮电大学、招商局集团有限公司、中国工商银行股份有限公司、海通证券股份有限公司、中泰证券股份有限公司、吉利汽车研究院（宁波）有限公司、中移（杭州）信息技术有限公司、中国电信股份有限公司北京研究院、中移在线服务有限公司、中移物联网有限公司、青岛海尔科技有限公司、国网湖南省电力有限公司电力科学研究院的大力支持。

中国信息通信研究院 人工智能研究所

地址：北京市海淀区花园北路 52 号

邮编：100191

电话：010-62301618

传真：010-62301618

网址：www.caict.ac.cn

