CAICT 中国信通院

集智 蓝皮书 No.202409

人工智能发展报告

（2024 年）

中国信息通信研究院

2024年12月

版权声明

更名声明

原“集智”白皮书更名为“集智”蓝皮书。“集智”蓝皮书将继续秉承原有的编撰理念和高质量标准，致力于提供有价值的信息和洞见。

前　言

人工智能浪潮席卷全球，正以前所未有的速度、广度和深度改变生产生活方式，对全球经济社会发展和人类文明进步产生深远影响。近年来，语言大模型、多模态模型、智能体和具身智能等领域不断出现突破性创新，推动人工智能迈向通用智能初始阶段。与此同时，人工智能的工程化持续加速推进，新产品新模式层出不穷，行业应用走深向实。

特别是过去一年，全球大模型井喷式发展。**技术层面**，缩放定律（Scaling Law）依然有效，语言大模型技术多维度能力持续进化，视觉大模型和多模态模型加速迭代，探索交叉模态融合处理。计算平台与模型创新紧密耦合，大规模分布式训练成为框架的新发力点，分布式训练支持、混合精度计算支持、高速互联通信等新要求驱动计算底座迭代升级。软件工具链全面优化升级，加速模型生产质效变革、提升模型部署推理效能、助力智能应用快速部署。高质量多模态数据集成为推动模型能力提升的关键，高水平数据标注和合成数据等新技术取得快速发展和突破。**应用层面**，专用智能应用逐步成熟，通用智能落地前景广阔。重点行业人工智能应用走深向实，贯穿产品研发设计、生产制造、营销服务、运营管理全流程，在提质增效的同时，逐步渗透并引导产业变革。从产业链各环节应用来看，大模型落地呈现“两端快、中间慢”的阶段特征。“选、建、用、管”体系化推动落地应用成为加速人工智能走向实用化、普惠化的行业共识。**安全方面**，人

工智能技术应用带来自身安全、衍生安全两大类风险挑战，各国治理进程不断提速，全球人工智能治理正处于“从原则走向实践”的关键阶段。**展望未来**，引入强化学习等技术来增强大模型能力仍是近期技术演进的重点方向，多模态模型、智能体有望加速突破，具身智能成为迈向通用人工智能的重要一步。面向中远期，类脑智能等颠覆性技术的成熟，有可能为人工智能发展带来更广阔的想象空间。随着人工智能赋能新型工业化向纵深发展，人工智能在实体经济中的应用场景将进一步拓展，加速向生产制造环节渗透，加速迈向全方位、深层次智能化转型升级新阶段。

在此背景下，我院发布《人工智能发展报告（2024年）》，旨在总结梳理人工智能技术创新方向、产业升级重点、行业落地趋势和安全治理进展，展望人工智能发展机遇，以期与业界分享，共同推动人工智能产业蓬勃发展。

目 录

图目录

表目录

一、总体态势

人工智能浪潮席卷全球，正以前所未有的速度、广度和深度改变生产生活方式。世界主要国家纷纷将推进人工智能技术创新与应用作为国家战略的重要方向，我国高度重视人工智能在培育新质生产力、塑造新动能方面的重要作用。习近平总书记指出，人工智能是新一轮科技革命和产业变革的重要驱动力量，将对全球经济社会发展和人类文明进步产生深远影响。2024 年 1 月，国务院常务会议研究部署推动人工智能赋能新型工业化有关工作，强调以人工智能和制造业深度融合为主线，加快重点行业智能升级，大力发展智能产品，高水平赋能工业制造体系。当前，人工智能正处于迈向通用智能的初始阶段，并成为推动经济社会持续发展的关键动力。

（一）人工智能技术演进走向新范式

以 Transformer 架构为基础的大模型不断取得新突破，在大数据、大算力加持下，逐渐实现从单任务智能到可扩展、多任务智能的跨越。这一关键突破，标志着人工智能技术发展**走向新范式**。以大模型为代表的人工智能技术展现出了类人智能的“涌现”能力，呈现规模可扩展、多任务适应及能力可塑三大特征。**一是**规模可扩展。模型的规模可扩展性不仅体现在参数的扩大，更依赖高质量数据集的供给以及大规模算力集群能力的增强。当前在模型参数保持不变的情况下，提高数据质量、扩大数据集规模或提升算力规模水平，

都能够显著增强模型的复杂性和处理能力。**二是**多任务适应。大模型支持多任务多模态能力持续增强，可执行任务已经从文本对话拓展到多模态理解、多模态生成等场景。**三是**能力可塑。通用大模型在训练阶段通过结合增量预训练、有监督微调、知识图谱等方法，实现将专业数据和知识注入模型中，提升大模型在专业领域的应用能力；在推理阶段，通过引入检索增强生成、提示词工程和智能体等技术，将更丰富的上下文信息和专业知识引入模型推理过程，解决更复杂的推理任务，优化模型表现。

具体从大模型算法演进态势看，深挖现有体系架构潜力，以实现理解推理能力和训练效率倍增仍是当前发展主线。模型研发主体纷纷围绕算法理论融合（如 Transformer 架构与其他路线结合）和模型改造（如扩大上下文窗口、思维链复杂推理、优化注意力模块、网络架构稀疏化、多模态特征对齐与统一理解等）展开创新升级，从而提高模型性能表现。近期 OpenAI o1 模型通过模仿人脑思考的思维过程，显著提升数学、物理、编程等复杂任务的性能水平。与此同时，非 Transformer 模型的底层算法也在不断创新。例如，基于图神经网络的 GraphCast、GNoME 在气象和材料领域已取得重大突破，基于物理约束的 PINN 网络、基于算子学习的 DeepONet 和基于傅里叶变换的 FNO 网络已成为求解偏微分方程（PDEs）的重要手段。

（二）人工智能工程化迈向新阶段

工程化技术是推动人工智能从实验室走向生产环境的关键桥梁，

也是人工智能在垂直行业应用落地的必经之路。在此过程中，人工智能工具链发挥着核心作用，其覆盖数据处理、模型训练微调、部署推理、应用开发、监控运维和安全可信全流程，是实现智能化转型的基础设施和加速器。当前，人工智能工程化的重点逐渐从大模型的训练微调向应用开发和落地转变，构建起围绕大模型及其应用的工具链，标志着人工智能工程化进入了新的产业化阶段。

开发工具链加速大模型技术迭代速度。开发工具链作为连接算法、数据与应用场景的关键纽带，对大模型的训练和推理至关重要。**在训练方面，**开发工具围绕分布式训练持续优化，显著提升了大模型的训练效率，如 DeepSpeed、Megatron-LM 等分布式训练框架通过支持更丰富的并行策略，以及更丰富的计算加速策略，有效支持产业界超大规模模型的预训练。同时，训练框架围绕参数高效微调等方面的技术创新，可以有效降低计算和存储成本。**在推理方面，**开发工具链聚焦优化量化、剪枝等压缩技术持续突破，加速推理过程并降低部署成本。同时，开发工具通过完善并行推理、混合精度推理、推理缓存等技术，可以有效降低计算资源消耗，提升推理服务速度。

应用工具链拓展大模型应用广度。大模型应用工具主要围绕 Agent（智能体）、多模型编排、大小模型协同、知识库集成、检索增强生成（RAG）及多组件融合等核心要素持续创新。Agent 的引入，实现了复杂任务的自动化执行与智能决策；多模型编排则有效

解决了单一模型局限性问题，通过灵活组合大小模型提升系统性能；大小模型协同机制，在确保精度的同时优化了计算资源利用；知识库与 RAG 技术的结合，极大增强了模型的知识推理与生成能力，确保结果的精确性；多组件的融合应用，则进一步丰富应用场景，提升了系统的灵活性与可扩展性。应用工具链不仅极大降低了大模型应用的开发门槛，还显著提升了智能应用的性能与用户体验。

（三）人工智能安全治理工作紧密推进

在人工智能飞速发展的浪潮下，全球人工智能治理合作持续升温，各国政府、国际组织、私营部门及社会各界携手并进，各主要经济体治理体系日渐明晰，产研合作愈发紧密。各类安全治理框架、安全治理工具日新月异，标志着全球人工智能治理跨入更加成熟的新阶段，以应对人工智能技术快速发展带来的机遇与挑战。

全球人工智能安全治理合作愈发紧密，各主要经济体治理体系渐趋明晰。国际合作方面，交流合作更加频繁，强调“负责任”、普惠发展理念。联合国在全球人工智能治理中发挥主渠道作用，二十国集团、七国集团等密集推出人工智能治理举措，人工智能安全峰会聚焦安全议题提供全球对话平台。与此同时，全球积极推动人工智能普惠发展。联合国大会通过了关于人工智能的里程碑式决议加快实现可持续发展目标，我国发起“一带一路”倡议、搭建“数据丝路”、成立中国-金砖国家人工智能发展与合作中心，并在“人工智能能力建设国际合作”高级别会议上提出《人工智能能力建设

普惠计划》，都致力于让人工智能为全人类带来“惠益”。**治理体系方面，**各主要经济体治理体系渐趋明晰，旨在维护本土产业发展需要。我国兼顾人工智能发展与安全，提出建立人工智能安全监管制度，发布《全球人工智能治理倡议》《人工智能安全治理框架（1.0 版）》。欧盟出台《人工智能法案》构建统一治理格局，美国发布拜登行政令推行行业自律的治理架构，英国、新加坡、日本等国加速构建立足本土产业发展需求的安全治理方案。

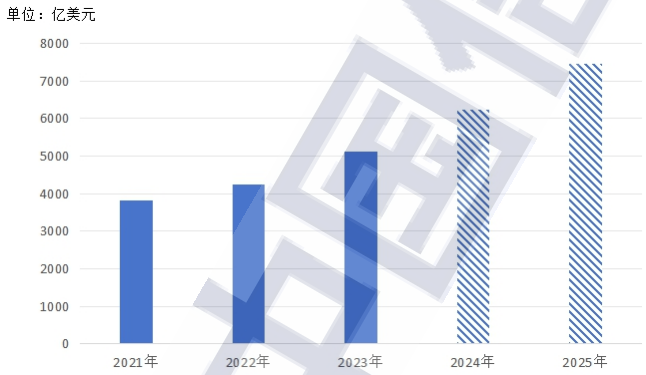
人工智能安全前瞻研究和产业实践深度结合，安全技术应用能力明显提升。前瞻研究方面，麻省理工学院、伯克利大学和南洋理工大学等研究机构提出模型间对抗新范式，深入探索人工智能模型自身安全边界。清华大学、北京大学和腾讯等机构积极开发新型模型水印算法，增强人工智能应用的安全可追溯性。**产业实践方面，**各国推动安全治理走深向实。美国打造 Dioptra 测试平台评估人工智能安全可靠性，英国人工智能安全研究所推出 Inspect 工具集评估模型能力和整体模型安全，新加坡迭代 AI Verify 工具箱以期提升人工智能可信度，中国推出大模型公共服务平台，集成数字内容水印、生成内容检测、不良信息识别等技术工具，提升大模型的安全合规能力。

（四）人工智能产业稳中有进迎来新动能

全球人工智能产业保持高速增长。据 IDC 预测，2024 年全球人工智能产业规模将达到 6233 亿美元，同比增长 21.5%[1]。具体来看，

[1] IDC

有两个方面的重要原因。**一是**大模型涌现式发展，为人工智能产业高速增长提供了核心动力。自 2023 年起，全球基础模型数量快速增加，相较于 2022 年增长数量翻倍；2024 年以来全球基础模型新增或迭代近百个，保持了较强的创新态势[2]。**二是**生成式人工智能技术加速产业化进程，促进全球人工智能规模化发展。据 Gartner 预测，到 2026 年，超过 80%的企业将使用生成式人工智能 API，或部署生成式人工智能的应用程序。



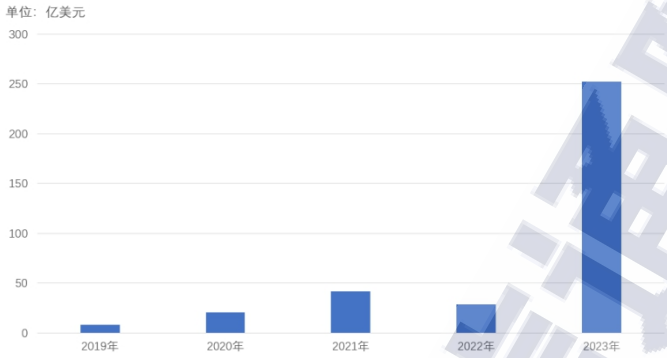
来源：IDC

图 1 全球人工智能产业规模（单位：亿美元）

大模型领域拉动全球人工智能投融资金额上扬。2024 年上半年，全球人工智能投融资金额达 316 亿美元，同比上升 84%。在全球融资紧缩的背景下，受益于大模型发展和企业融资带动，人工智能领域融资占全行业融资比例持续上升，从 2022 年的 4.5%上升至 2024 年上半年的 12.1%。2023 年，生成式人工智能投融资规模达 252 亿美元，约为 2022 年的 9 倍，占 2023 年所有人工智能相关投资的约

[2] 基于斯坦福大学 Ecosystem Graphs 数据及头部企业发布事件统计

四分之一[3]。2024 年上半年，全球金额最大的 10 笔融资事件中有 6 笔为大模型企业融资，金额总计达 135 亿美元。



来源：The AI Index 2024 Annual Report

图 2 全球生成式人工智能投融资规模（单位：亿美元）

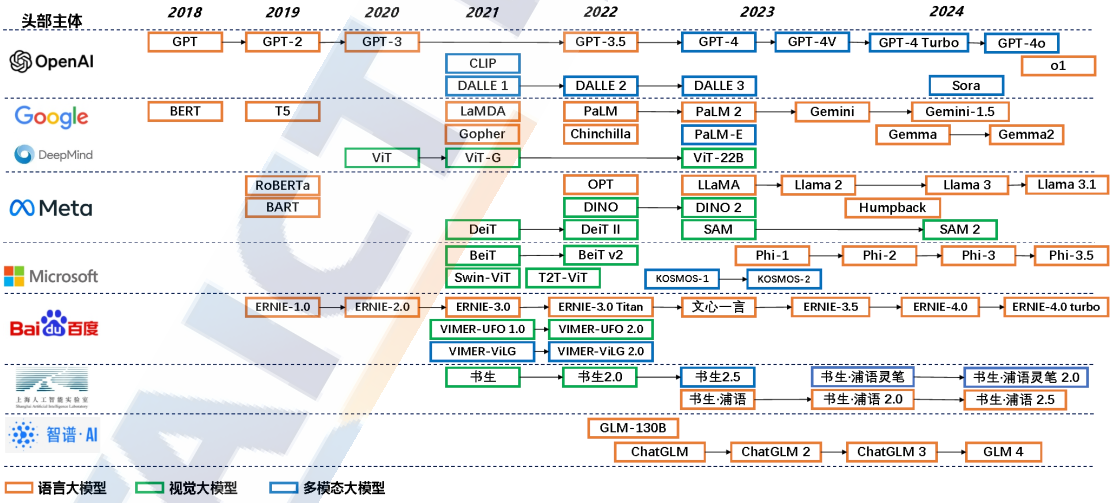
人工智能创业企业发展持续创新高。人工智能创业企业是智能化时代影响技术产业格局的生力军，也是推动全球人工智能产业生态繁荣的重要力量。截至 2024 年第二季度末，全球共有人工智能独角兽企业 242 家，其中 2024 年上半年新增 15 家，占所有新增独角兽企业的 40%。独角兽企业业务分布广泛，2024 年新增独角兽企业业务领域涵盖生物制药（如 Xaira Therapeutics）、软件开发（如 Cognition AI）、数据平台（如 Weka）、搜索问答（如 Perplexity）、科研写作（如 Sakana AI）等。部分独角兽企业商业模式逐渐清晰，技术实力和发展前景已获得市场认可，其中 Astera Labs、出门问问等已于 2024 年上半年成功上市。随着大模型应用落地门槛的持续降低，各行业也将涌现出更多人工智能创业企业，垂直行业赛道成为未来创新涌现的重要场景。

[3] The AI Index 2024 Annual Report

二、技术创新

（一）基础模型仍在快速演进迭代

缩放定律驱动下的模型能力持续提升，基础大模型的语言、视觉和多模态能力快速迭代。从时间维度来看，2022 年 ChatGPT 的出现引领大模型浪潮兴起；2023 年国内大模型呈井喷式爆发态势，能力快速迭代，模态持续拓展；2024 年大模型推理理解能力跃迁，并开始探索垂类领域应用落地。OpenAI 等基于大量工程实验和反复验证提出缩放定律，揭示了模型能力与计算能力、参数量和数据量间的定量关系，业界也遵循该定律指导资源要素投入、推动模型创新发展，近年来在模型技术能力、通用泛化水平等方面取得一系列突破性进展。目前，大模型支持模态已逐步从自然语言处理拓展到多模态理解和生成等场景。



来源：中国信息通信研究院

图 3 语言、视觉和多模态三类基础模型布局

1. 语言大模型：技术能力演进加速，在幻觉、成本等

问题上仍面临挑战

从 2023 年至今的基准测试结果来看，全球大模型能力已经出现阶跃式提升。**语言大模型能力提升主要体现为以下四方面。一是上下文窗口长度扩展，提升全局能力。**大模型的上下文窗口长度是指模型在执行文本生成任务时，能够处理的前置文本的数量或长度，决定了模型对信息的理解深度和广度，对于理解和生成连贯、一致且准确的文本具有重要意义。当前，国内外主流大语言模型均具备 128k 以上的上下文长度处理能力，可一次性处理数十万单词或汉字。**二是知识密度增强，储存更多知识。**随着数据、算力、算法协同发展，大模型知识密度持续增强，平均每 8 个月翻一番。2020 年 6 月发布的 GPT-3 大模型有 1750 亿个参数，2024 年 2 月面壁智能发布 MiniCPM-2.4B 模型在实现同等性能的同时，参数规模降至 24 亿，相当于知识密度提高了约 86 倍。**三是 MoE 混合专家架构能够容纳更多知识，精准刻画任务。**MoE 稀疏激活多个专家子模型支路，加权融合多个子模型结果，实现更加准确的输出，提高推理计算效率。目前，谷歌的 Gemini-1.5 Pro、Mistral AI 的 8x7B 与 8x22B、阿里云 Qwen-1.5 MoE、阶跃星辰 Step-2 等头部大模型均采用 MoE 架构，已成为当前大模型的重要演进趋势。**四是通过强化学习（Reinforcement Learning，RL）将思维链（Chain of Thought，CoT）内化进模型，提升复杂推理能力。**2024 年 9 月 OpenAI 发布的 o1 系列模型在后训练（Post-Training）阶段采用强化学习和思维链的技术

方案，不仅在“慢思考”后回答复杂问题的表现优异（尤其是在 STEM 领域的推理能力显著增强），还具有了自我反思与错误修正能力，使自博弈强化学习有望成为提升语言大模型逻辑推理能力的技术新范式。

表 1 语言大模型演进迭代情况

公司	模型	上下文长度
Meta AI	Llama 2	8k
	Llama 3.1	128k
OpenAI	GPT-4	32k
	GPT-4 Turbo	128k
	o1/o1 mini	128k
Anthropic	Claude 3.5	200k
阿里云	Qwen	8k
	Qwen-1.5	32k
	Qwen-2.5	128k
百度	ERINE 4.0	8k
	ERINE 4.0 Turbo	128k
上海 AI 实验室	书生浦语	8k
	书生浦语 2.0	200k
谷歌	Gemini-1.0	32k
	Gemini-1.5	1000k（100 万）

语言大模型虽然在文本理解与生成、复杂逻辑推理任务上取得了突破，但在幻觉问题、训练成本方面仍然面临挑战。一是复杂逻辑推理和泛化能力仍需强化。以 OpenAI o1 系列模型为例，虽然 OpenAI 通过强化学习和思维链等技术方案使得 o1 系列模型在 STEM 领域的推理能力得到大幅提升，但在开放性、复杂度更高的

问题和场景中泛化能力仍然不强。一方面，可以通过在特定领域的数据集上进行微调，提高模型在该领域的逻辑推理能力。另一方面，可以将知识图谱与大模型结合使用，提供额外的结构化知识，帮助模型实现更准确地理解和推理。**二是幻觉问题无法彻底消除。**当前语言大模型在生成文本或理解信息时，可能会产生不准确或完全虚构的内容，影响生成内容的可靠性与安全性。当前业界在以下三个方面试图降低幻觉：提升训练数据质量，对训练数据进行清洗和验证，以去除错误、不一致或误导性信息，确保训练数据的准确性和可靠性是预防幻觉问题的关键；使用检索增强（RAG）技术，RAG 模型结合了检索机制和生成机制，能够从大量外部数据库中检索相关信息，并结合这些信息进行生成，从而提高内容的准确性；增强长上下文处理能力，大模型通过处理长文本信息，更好地理解上下文及复杂的逻辑关系和情境，减少生成幻觉的风险。**三是训练成本仍然偏高。**当前大模型的训练成本仍然偏高，这主要体现在数据需求、算力消耗及基础架构上。在数据层面，大模型需要大量的数据来进行有效的预训练，数据的收集、标注、清洗和预处理都需要大量的时间和资源。此外，高质量数据往往需要人工标注，不仅耗时而且成本高昂，尤其是在需要专业知识的领域。在算力层面，训练大模型需要巨大的计算资源，通常需要使用 GPU 集群高性能计算设备，目前这些设备的使用成本仍然较高。在算法层面，当前基础单元架构的改进关注特定任务或条件下单一维度性能的提升，通常会

以牺牲其他方面性能为代价，如计算效率、内存占用等，目前还没有出现一种能够全面超越现有 Transformer 架构的基础单元。

表 2 语言大模型调整及解决方案

挑战方向	主要解决思路
复杂推理问题	合成推理数据、类推提示法、过程监督、拒绝采样微调等
幻觉问题	信息检索增强技术、调整解码策略以及完善预训练和对齐策略等方式等
训练成本问题	更加高效的优化器和提升训练框架的稳定性，合成数据和自动标注提升训练数据的构建效率

2. 视觉大模型：Transformer 赋能图像理解，扩散模型实现图像生成

视觉 Transformer 模型（Vision Transformer，ViT）将 Transformer 架构从自然语言扩展到视觉领域，成为判别式视觉任务的主流架构。2020 年，谷歌发布的 ViT 模型首次将图像适配到 Transformer 架构，在 ImageNet 数据集上取得了超过 CNN 的表现，从此奠定了 ViT 在视觉领域基础架构的地位。目前，业界主要聚焦模型结构和下游任务两方面对 ViT 模型进行改进。在模型架构改进方面，微软的 Swin Transformer、Meta 的 MAE、DeiT、SAM、DINO2、苏黎世联邦理工学院的 PVT、McGill 大学的 CvT 等模型从多尺度、知识蒸馏、自编码等方向改进 ViT 网络结构，在图像分类、目标检测与分割、图像检索、深度估计等传统视觉任务上取得突破。**与此同时，以 ViT 为代表的判别式视觉大模型仍面临以下几个方面的挑战。一是计算**

资源需求高，基于 Transformer 架构的视觉大模型在计算上具有二次复杂度，对计算资源有较高要求。**二是训练数据依赖性强，**ViT 等视觉大模型需要大规模数据集进行预训练以获得更好的性能，在小数据集上的表现可能不佳。**三是自监督学习挑战，**自监督学习是视觉大模型训练的关键环节，但如何有效地设计自监督任务以充分挖掘数据特性与模型潜力仍然是一个开放性问题。**四是模型部署与推理加速，**为了在实际应用中部署视觉大模型，需要有效的模型压缩和加速技术，以减少模型大小并提高推理速度。

扩散模型成为图像生成领域的主流方案，展现巨大应用潜力。扩散模型（Diffusion Models）基于马尔科夫链的扩散过程逐步从噪声中重构出所需的数据，广泛应用于高质量图像与视频的生成、编辑与修复等场景。扩散模型相较传统生成模型在以下三方面展现优势：**一是高质量样本生成，**扩散模型能够生成高分辨率、高保真度的图像，视觉效果逼真。**二是训练稳定性，**与生成对抗网络（GAN）等传统生成模型相比，扩散模型的训练过程更为稳定，减少了模式崩溃的风险。**三是灵活性与可控性，**扩散模型允许通过条件输入（如文本描述、草图等）来引导图像的生成方向和风格，支持生成多样化的图像样本，包括艺术创作、风格迁移等多种创新应用。目前头部人工智能厂商聚焦通过扩散模型持续提升图像生成能力，包括 Stability AI 的 Stable Diffusion、OpenAI 的 DALL·E、谷歌的 Imagen 等。**与此同时，以扩散模型为代表的生成式视觉大模型仍面临以下**

三个方面的挑战：一是推理速度，扩散模型的推理过程需要多个步骤迭代生成，导致推理时间较长。**二是幻觉问题，**当前模型存在生成图像内容与客观事实不符的情况。**三是评估指标，**当前扩散模型生成样本的评估主要基于 FID 分数，这一指标反映图像全局的表征能力，无法全面反映样本的细节恢复效果和多样性。

3. 多模态模型：四种实现方式探索交叉模态处理

多模态大模型融合了多种感知途径与表达形态，能够同时处理文本、图像、语音等多种数据，并进行深度的语义理解和交叉模态处理，具备深度人机交互和全面智能应用的潜力，是通用智能的重要实现路径。多模态大模型主要有四种实现方式，按模型实现功能可以分为理解类与生成类两条主要路径。

一是多模态理解模型。多模态理解模型对齐视觉特征与文本特征实现跨模态的统一理解，分为以下两类技术路线。**一方面，基于语言大模型底座，配合多类外部专家模型共同实现多模态处理。**如微软的 Visual ChatGPT 模型将 OpenAI 的 ChatGPT 与 22 种不同的视觉基础模型（VFM）相结合，使用户能够超越语言限制，实现多模态交互，如聊天交互、图像编辑等任务。谷歌的 PaLM-E 模型利用现有 LLM 和语言嵌入方法解决多模态问题，将连续具体的多模态输入转化为 LLM 可识别的向量特征，用于大模型训练，实现语言问答、视觉问答等任务。**另一方面，通过跨模态特征对齐学习，实现多模态输入的统一和融合。**OpenAI 的 CLIP 模型通过对比学习，将图像

与文本通过各自的预训练模型获得的编码向量在向量空间上对齐，从而理解和推理图像和文本之间的关系，被广泛用于图像检索、视觉问答、图像生成等领域。

二是多模态生成模型。多模态生成模型基于对不同模态信息的理解，具备文本、图像、视频、语音信息的生成能力，能够根据输入指令创造新的数据内容或增强现有数据的表达能力，分为以下两类技术路线。**一方面，**DiT（Diffusion Transformer）**结合扩散模型与 Transformer 优势，成为视频生成模型主流架构。**DiT 架构用 Transformer 代替了传统扩散模型中基于卷积网络的 U-Net。Transformer 具有更强的上下文处理能力和理解生成能力，扩展能力更强、收敛效率更好。OpenAI 的 Sora、谷歌的 Veo、快手的可灵大模型均具备超一分钟长时长、1080P 高清视频生成能力。**另一方面，端到端统一多模态架构，实现跨模态生成与实时交互响应。**OpenAI 的 GPT-4o 和谷歌的 Gemini 均采用了端到端单体模型的方式学习文本、视觉、语音等不同模态的统一表征，实现跨模态实时交互相应。GPT-4o 能够根据手机拍摄视觉信息与用户对话交互实现多模态统一理解，具备“听、看、说”的多模态交互能力，平均响应速度 320ms，达人类对话水平。

表 3 多模态模型技术路线表

类型	路线	典型
多模态理解	语言大模型调度	微软 Visual ChatGPT
		谷歌 PaLM-E
	跨模态特征对齐	OpenAI CLIP

		微软 KOSMOS
		DeepMind Flamingo
		Salesforce BLIP
多模态生成	扩散模型	Stability AI Stable Diffusion
		OpenAI DALL · E
		OpenAI Sora
		快手 KLING
		Runway Gen-3
	端到端理解与生成架构	谷歌 Gemini
		OpenAI GPT-4o

（二）计算平台与模型创新紧密耦合

1. 模型创新依赖计算平台，协同价值凸显

以大模型为代表的通用智能范式正在驱动人工智能计算平台升级。当前，“大模型+大算力+大数据”成为可能实现通用智能的主要路线之一，基础大模型底座的智能水平与迭代速度成为各国科技竞争的战略焦点。然而，大模型目前仍是一种实验科学装置，升级迭代需反复实验，是复杂的系统工程。以 Llama 3.1-405B 为例，使用 1.6 万张 NVIDIA H100 GPU，在 15.6T token 数据上预训练需 54 天[4]，期间经历 466 次工作中断，这种工程实验科学装置更加依赖先进、高效、可靠的软硬件系统支持。相比专用算法，大模型的创新与基础软硬件体系正加速耦合。一味追求算力规模扩张无法满足大模型创新需求，而是要更加强调应用、算法、关键软件栈、底层硬件全方位协同发展，实现系统收益最大化。但这也对软硬件协同水平提出了新挑战，中国信通院 AISHPerf[5]基准测试表明，软硬件系统

[4] The Llama 3 Herd of Models，Meta Llama Team.

[5] AISHPerf：Performance Benchmarks of Artificial Intelligence Software and Hardware

正面向大模型训练、推理需求加速迭代升级，模型轻量化部署、混合精度计算、分布式训练策略优化等新特性是近期软硬件产品升级迭代重点，厂商积极推动不同模型网络架构与硬件的深度适配。

大模型技术的原始创新和应用迭代落地高度依赖先进的软硬件协同技术生态体系。一方面，模型原始创新与底层硬件协同显著加强，构建新的模型结构与组件往往需考虑底层硬件的支持程度，如针对模型架构优化的 Flash Attention、Flash Decoding 等创新技术。**另一方面**，面向差异化的赋能场景，需要软硬件系统结合场景需求特点在训练、推理等环节高度协同，从算力集群调度、框架分布式训练能力、端侧推理加速等方面系统提升模型调优速度、任务精度，降低训练推理成本，提升人工智能赋能传统行业质量和效率。

2. 框架关键技术能力持续提升，大模型加速框架成为新发力点

PyTorch 引领框架发展，国产框架快速崛起。从全球来看，目前，基础通用框架两强并立局面被打破，PyTorch 已主导学术界，使用占比持续提升。PapersWithCode 数据显示，PyTorch 框架在论文中使用比例从 2020 年 9 月的 51%稳步提升至 2023 年 9 月的 60%，而同期 TensorFlow 则从 20%骤降至 3%左右。从技术能力来看，2022 年底发布的 PyTorch2.0 将计算图捕获正确率从 50%提升至 99%，解决其上一版本动态图编译困难的致命缺陷，同时将 2000 余算子整合优化至 250 个左右，仅需一行代码即可实现 1.5 到 2 倍的 Transformers

模型训练加速，大幅提升大模型支持能力，编译效率大幅提升，受到业界广泛欢迎，逐渐扩大与 TensorFlow 的竞争优势，先前持续数年的框架两强并立局面被打破。**从我国来看，国产框架技术能力不断完善，基于国产框架的行业解决方案正在向垂直领域快速渗透。**近年来国内涌现了一批如百度飞桨、华为昇思、一流 OneFlow、之江天枢等开发框架，支撑构建一批更加符合本土产业特色和场景需求的解决方案。随着人工智能进入大规模赋能新型工业化阶段，国产深度学习框架迎来新一轮发展机遇，向行业融合渗透不断加强。如百度飞桨已凝聚 1070 万开发者，基于飞桨创建了 86 万个模型，服务 23.5 万家企事业单位；华为 Mindspore 社区用户达到 780 万，总 PR 数达到 97.7k，已在互联网、医疗、安防、政府、科学计算领域广泛落地应用。

大规模分布式训练成为框架的新发力点，一批大模型加速框架显现。当前，开发框架主要面向大模型分布式训练异构资源管理调度、多节点任务调度等方面完成优化，呈现两种发展路线，**一是基于原有框架实现分布式训练功能，**例如，微软 DeepSpeed、英伟达 Megatron 基于 PyTorch 强化大模型分布式支持能力、提升训练效率。其中，微软 DeepSpeed 针对分布式训练中计算资源稀缺问题，提升异构硬件统筹调度能力，丰富计算资源供给。DeepSpeed 在多 GPU 系统上展现出较好的分布式扩展性，相较于 Megatron，其应用更为广泛，包括计算机视觉、自然语言处理、推荐系统等，旨在提高大

模型的训练速度和效率。**二是集成分布式能力的一体化通用发展路线，**例如，百度飞桨框架原生支持超大规模分布式训练能力，推出端到端自适应分布式训练技术，实现了低成本自动并行开发、最优并行策略自动选择和异步流水调度，突破了模型结构和硬件环境多样导致的分布式训练策略开发复杂、训练性能调优难的技术瓶颈。

3. 大模型训推需求推动芯片加速迭代，各类市场主体差异化创新

大模型热潮进一步推动计算底座迭代升级。大模型计算特性对硬件要求极高，带来分布式训练支持、混合精度计算支持、高速互联通信等新要求新挑战，驱动计算底座迭代升级，呈现三大趋势特点：**一是芯片架构向定制化演进，迎合 Transformer 计算特性。**如英伟达自 Hooper 架构引入 Transformer 引擎提升算法计算性能，并利用启发式算法实现数据精度动态切换（Blackwell 架构二代 Transformer 引擎已支持 FP8、FP6、FP4 等多种低精数据），在保证性能的前提下降低计算总量；芯片创业公司 Etched 推出仅支持 Transformer 架构的 Sohu 芯片，牺牲编程能力提升计算速度，推理吞吐量达到 H100 的 20 倍。**二是存储与互联重要性日益提升。**随着大模型参数持续增长、输入输出数据长度快速提升，模型参数和计算缓存 kv 值消耗的内存空间呈指数级增长，存储和互联成为主要瓶颈，在芯片单位面积算力接近天花板且性能相对过剩的背景下，头部硬件厂商创新升级重点从卷算力向卷内存、卷互联转变，如 AMD

MI300X 宣传时已淡化算力色彩，重点突出显存和互联指标，英伟达 B200 显存容量和显存带宽提升幅度（240% × H100），均超过算力提升幅度（220% × H100 @FP16）。**三是强调软硬协同升级释放硬件计算潜力。**如 AMD ROCm 6.2 更新扩展了专为语言大模型所设计的 vLLM 库支持，提升了 Instinct 系列加速器的 AI 推理能力；英伟达参与 FlashAttention 3 注意力算法设计，充分利用 H100 芯片动态 warp 寄存器分配、FP8 精度支持等特性，相比 FlashAttention 2 速度提升 1.5-2 倍。

多方试图破局，出现三类挑战者。尽管目前英伟达垄断人工智能计算生态，但面对高昂的采购成本和庞大的市场空间，各方持续寻找替代英伟达的解决方案，出现三类挑战者。**一是以 AMD、英特尔为代表的半导体巨头，**凭借深厚技术积累、庞大资金支持和市场渠道优势，推出面向大模型和人工智能的高性能计算产品，如 AMD Instinct MI325X 芯片、英特尔 Gaudi2 芯片等，在内存容量、存储带宽、性价比等方面形成差异化竞争优势。**二是以** Cerebras、Groq、d-Matrix、Graphcore **等为代表的芯片初创企业，**尝试通过超大尺寸芯片、存内计算、近存计算等非常规技术路线取得突破，已获得 OpenAI、微软、三星等行业巨头投资。**三是以微软、Meta 等为代表的互联网巨头加快自研芯片进程，**试图摆脱对英伟达依赖，提升议价能力，如谷歌 TPU 已更新至第五代（TPU v5p），支持多模态大模型 Gemini 训练；微软推出 MAIA 100，采用 5nm 工艺，服务微软

云大模型训推；Meta 发布首款自研推理芯片 MTIA v1，基于 7nm 工艺 ASIC 芯片，与自身 PyTorch 框架高度适配。此外，量子、类脑、光计算等前沿颠覆式路线也加紧与大模型应用结合，规模商用虽有差距，但为复杂高效计算系统实现开辟新路径，如清华大学光计算芯片“太极”实现 160TOPS/W 的超高能效，能够以更低的资源消耗和更小的边际成本支撑大模型训练推理。

（三）工具链不断完善加速大模型研发应用

大模型工具链是指一系列集成化的软件工具和平台，旨在支持大模型开发构建、训练优化、应用开发、部署推理和运维管理全流程。工具链的持续升级对于大模型开发和应用至关重要，是构建模型服务体系（Model as a Service，MaaS）的平台能力支撑，目标是灵活便捷供给大模型服务。首先，工具链不断升级，能够有效应对大规模模型训练的复杂性挑战，提高训练效率和推理效能。其次，工具链的集成和易用性降低了模型开发部署门槛，缩短了开发周期，加速实现一站式模型部署。最后，工具链不仅可以集成 Agent 框架、检索增强生成（RAG）等新技术，还能支持灵活的模型与组件调用，促进智能应用的快速构建和部署。

来源：中国信息通信研究院

图 4 大模型工具链架构图

1. 模型训练工具：加速模型生产质效变革

训练工具能力全面升级，有效支撑大规模训练任务。一方面，为了更加高效地完成大模型训练任务，涌现出多种训练加速技术。一是计算资源优化技术，如 PyTorch 支持的混合精度训练、Adafactor、Flash Attention 等技术，能够通过减少计算和存储需求，提升模型效率。目前，谷歌、微软、腾讯、蚂蚁等头部企业广泛采用混合精度训练等技术减少显存占用并提升训练速度。**二是应用计算优化策略来提升模型的执行效率，**如 DeepSpeed 支持的算子融合、梯度积累技术等技术，能够在资源有限的情况下，通过优化计算策略，加速计算过程。此外，收敛性优化技术通过提高模型的收敛速度，提升模型训练效率，并提高模型的泛化能力。目前，主流的深度学习训练框架均支持收敛性优化技术，如 DeepSpeed、PyTorch、JAX 等均支持 Adam、Adagrad 等自适应学习率优化器，能够在训练过程中动态调整学习率，使模型能够更快地收敛。**另一方面，为提升大模型**

在特定场景的适应性，业界推出多种微调技术以提升训练效率。目前主流微调技术分为全量微调和参数高效微调（PEFT）技术。**全量微调**精度高、泛化能力强，但计算成本较高，一般适用于精度需求较高的复杂任务场景。**参数高效**微调能够显著节省训练时间和计算资源，适用于资源受限或者需要快速部署迭代的场景，已经成为产业实践的主流选择。主流的参数微调方法包括低秩适应（LoRA）、前缀调优（Prefix Tuning）、提示调优（Prompt-Tuning）等，当前产业主流的训练工具套件，如百度千帆大模型平台、阿里百炼大模型平台均支持上述微调技术。

2. 模型推理工具：提升模型部署推理效能

推理工具能力不断升级优化，为大模型落地提供高效支撑。一是模型压缩工具持续整合更多压缩技术。大模型通常需经过模型压缩以适应更广泛更多样化的部署环境，如何兼顾压缩比例与性能损耗是关键。以量化、剪枝为代表的压缩技术持续演进，通过低比特量化、稀疏化、模型结构搜索、参数自动寻优等方式实现模型训练中、训练后的低损与高效压缩。如百度的模型自动化压缩工具 ACT（Auto Compression Toolkit）可实现压缩流程自动化[6]，商汤的神经网络量化工具 PPQ 通过图优化等技术实现高效的压缩能力[7]。**二是国内外机构专注于大模型推理引擎的研发创新。**推理引擎针对大模型

[6] PaddleSlim/example/auto_compression at develop · PaddlePaddle/PaddleSlim (github.com)
[7] openppl-public/ppq: PPL Quantization Tool (PPQ) is a powerful offline neural network quantization tool. (github.com)

推理场景的低时延、高吞吐要求，从显存优化、高性能算子、服务调度等多个维度进行优化设计，已成为当前大模型部署推理的主要工具，如伯克利大学 LMSYS ORG（Language Model Systems Organization）vLLM、英伟达 TensorRT-LLM、HuggingFaceTGI（Text Generation Inference）、微软 DeepSpeedDeepSpeed-MII 等。我国科技企业也纷纷布局该领域，如腾讯一念 LLM 同时支持英伟达 GPU 和华为 NPU 卡[8]，阿里魔搭 DashInfer 支持 CPU 卡上的高效推理[9]，蚂蚁 GLake 通过对键值对缓存实现透明管理和存算解耦，进一步提升了推理性能和兼容性[10]。**三是模型与推理工具之间呈现高度融合与协同优化的趋势。**如百度文心一言大模型结合飞桨框架协同优化，使得推理性能提升百余倍；腾讯太极平台通过显存+主存统一存储管理技术，并通过模型算子之间的显存共享和优化，使端到端推理性能提升至业界平均水平的 2.3 倍。

3. 应用开发工具：助力高效打造 AI 应用

大模型服务化供给加速了智能技术的普惠化落地，激发了基于大小模型的智能应用开发需求。与传统应用软件开发相比，智能应用开发在应用模式方面转变为以模型为核心，**在开发方式方面**演变成零代码、低代码等多种形式，以满足不同技术能力的用户群体需求。**以上变化对模型应用的开发提出了新的要求，一方面，**开发所

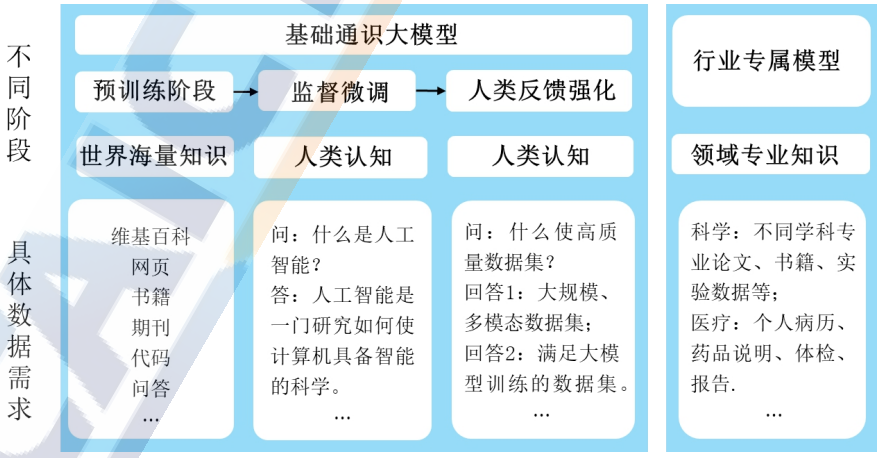
[8] https://github.com/pcg-mlp/KsanaLLM/blob/main/README_cn.md
[9] https://www.modelscope.cn/headlines/article/497
[10] https://github.com/intelligent-machine-learning/glake

依赖的能力框架、工具链条等需适应大模型应用的开发需求。Agent、RAG 等框架成为常用 AI 应用开发框架，大小模型组合成为应用落地的主要技术方式。同时，搜索、格式转化等工具插件将模型的能力进行延伸，在应用中扮演着越来越重要的作用。**另一方面，**快速变化的市场需求对应用开发效率的提高更为迫切。面对新需求与新变化，模型应用开发工具逐步走向市场，并帮助用户快速搭建丰富的模型应用。

大模型应用开发工具趋于平台化，不仅整合了必要的开发工具、框架与服务，还极大地简化了从模型调用到应用部署的全过程，为开发者提供了高效、便捷的创新环境。一方面，国内模型应用开发平台整合了软件开发的全流程工具，并提供了大量 AI 能力组件。例如，百度智能云千帆 AppBuilder 面向不同开发能力的用户和开发场景，分别以零代码态、低代码态、代码态的产品形态，帮助开发者构建 AI 原生应用。字节跳动扣子应用开发平台集成了超过 60 款各类型的插件，可以极大地拓展 AI Bot 的能力边界，并且提供了简单易用的知识库能力和数据库长期记忆功能。**另一方面，**国外模型应用开发平台集成了很多开源框架，并提供了丰富的 API 和工具，满足各种应用场景的需求。例如，谷歌的 Vertex AI 中包括 TensorFlow、Keras、Colab 等在内的开源框架，提供了模型训练和部署的 API，能够快速实现基于大模型的应用解决方案。

（四）高质量多模态数据集成为模型能力提升关键

大模型发展已经进入多模态融合阶段，作为人工智能学习、训练和验证的“燃料”基础，**大规模、高质量、多模态数据集对于多模态大模型能力提升愈加重要，以数据为中心的人工智能时代正在加速到来。**近期，由华盛顿大学、Salesforce Research 和斯坦福大学等机构联合团队推出的包含万亿 token 的史上最大多模态开源数据集 MINT-1T（Multimodal INTerleaved）正式发布，经验证以该数据集为基础预训练的 XGen-MM 模型在视觉描述、视觉问答、多图像推理等基准性能方面取得了显著提升。为加速构建人工智能高质量数据集，**面向大模型的新一代数据工程成为核心技术手段。**大模型的数据工程涵盖训练数据集的数据采集、数据预处理、数据标注、质量评估、数据合成、开放共享等全生命周期，不仅需要保证数据的数量和多样性，更要强调数据的质量和有效性，并通过严格的数据治理和管理，确保数据的安全性和合规性，降低数据使用中的风险。



来源：中国信息通信研究院

图 5 不同阶段的具体数据需求情况

1. 数据预处理：多模态词元融合和实时处理成为主要发展方向

数据预处理技术正朝向多模态融合、智能化、实时性全面进化的方向发展。一是多模态词元化序列向量有效融合。随着大模型向多模态方向发展，预处理技术逐渐整合文本、图像、音频、视频等多种类型的数据，探索建立模型识别的多模态统一词元序列空间方法，实现高效、一致、标准的预处理流程，以支撑模型对复杂多源信息的理解和生成能力。比如，OpenAI 的 GPT-4o 模型实现了图像、文本和音频等不同多模态的词元向量统一对齐，平均反应时间仅有 320 毫秒，与人类的对话反应速度已经不相上下。**二是自动化与智能化程度持续提高。**当前数据预处理过程更加依赖自动化工具和算法，未来亟需利用 AI 技术自我优化预处理步骤，减少人工处理过程干预，提升效率和精确度，比如自动识别数据模式、智能选择预处理策略。**三是实时处理与流式数据处理能力不断增强。**面对大规模实时数据流，预处理技术创新侧重于低延迟处理技术，比如流式计算、实时分析和即时反馈机制，确保模型能够及时响应最新数据。**四是利用边缘计算加速处理效率的趋势逐步显现。**为了应对大模型数据量的指数级增长，预处理技术更加倾向于利用边缘计算和分布式处理架构，减少数据传输成本，提高处理效率和响应速度。比如使用 Apache Spark 等分布式计算框架，在集群中并行处理数据，可有效提升数据预处理效率。

2. 数据标注：新一代高水平数据标注提升高质量数据集供给能力

大模型发展需要新一代高水平数据标注。当前，随着深度学习和人工智能模型的复杂度提升，对高质量、精细化标注数据的需求愈发迫切，这不仅要求数据标注技术能够高效处理大规模数据集，还需要具备对多模态数据（如图像、语音、视频及文本）和跨领域数据综合处理的能力，**数据标注逐渐向专业化、智能化、多模态方向发展。一是自动化与智能化标注工具创新成为焦点。**当前，基于计算机视觉、自然语言处理等技术的自动标注工具快速涌现，这些工具利用算法初步完成标注，再由人工进行校验和修正，可大幅提高标注效率，降低成本。比如国内数据标注企业海天瑞声已建成一体化智能数据处理服务平台,可实现语音、图像、视频以及文本等全领域数据的自动标注处理。**二是多模态数据标注技术的融合逐渐成为趋势。**随着 AI 应用向更复杂的场景拓展，单模态数据已无法满足需求，跨领域的多模态标注技术，结合图像、声音、文本和视频的多模态联合标注，正在成为数据标注的新趋势。例如，由 Human Signal 开发的 Label Studio 开源数据标注工具，可支持文本、图像、语音等多模态数据标注，广泛应用于 NLP、CV、语音识别等领域，显著提高了 AI 模型训练效率。**三是持续学习与反馈机制引入促进数据标注质量和效率双重提升。**通过将标注后的数据反馈给 AI 模型，不断训练和优化模型性能，形成标注-训练-反馈的闭环，不仅能提升

模型精度，还能指导标注策略的动态调整，确保标注工作更加高效和具有针对性。**四是跨学科融合深度和广度进一步拓展。**随着人工智能技术的不断发展，越来越多的重点行业领域开始应用数据标注技术，不同行业领域的数据标注需求呈现多样化和专业化的特点，需要跨领域的专业知识和技术支持。

3. 质量评估：数据质量评估和模型反馈机制共同推动数据质量不断提升

当前，人工智能数据集质量评估需求体现在完整性、准确性、一致性、时效性和可解释性等多个方面，评估技术发展趋势主要聚焦以下几个关键方向：**一是质量评估与反馈机制深度融合。**数据质量评估引入客观的数据质量评估指标和模型反馈机制，使得数据使用者可以评价数据集的实际综合表现，并反馈给数据提供者以改进数据采集和处理流程。2024 年 6 月，OpenAI 推出了 CriticGPT，旨在帮助人类评估和检测大型语言模型（LLM）生成的代码输出中的错误，CriticGPT 通过训练生成自然语言反馈，可以评估出代码中的质量问题，并且在检测自然发生的 LLM 错误时，其生成的评审比人类评审更受欢迎，准确率高达 63%。**二是多模态数据质量评估框架快速发展。**针对图像、语音、文本等多种类型数据，设计发展了综合评估模型，确保跨模态数据的一致性和互补性。通过融合计算机视觉、自然语言处理和语音识别技术，实现多维度数据质量的全面评估。**三是偏差与公平性评估成为数据质量评估重要组成部分。**鉴

于 AI 系统易受偏见数据影响，数据质量评估技术致力于检测并量化数据集中存在的偏差，确保训练数据的均衡性和代表性，减少模型输出的不公平性。通过算法审计和统计测试，系统性地识别并纠偏，保障 AI 应用的公正性。**四是动态数据质量监控体系逐步完善。**利用实时分析和流处理技术连续评估数据质量，即时反馈数据问题，支持快速响应。这不仅有助于维护数据的时效性和准确性，也确保了 AI 模型在数据变化时的稳定表现。

4. 数据合成：合成数据有望解决大模型潜在数据瓶颈

当前，大模型的训练数据严重依赖现有的互联网公开数据。有研究预测，到 2026 年大型语言模型的训练就将耗尽互联网上的可用文本数据，**未来需要借助合成数据解决大模型的数据瓶颈。**目前，合成数据正迅速向金融、医疗、零售、工业等诸多产业领域拓展应用。根据 Gartner 预测，到 2024 年，60%用于 AI 开发和分析的数据将会是合成数据，到 2030 年，合成数据将成为 AI 模型所使用数据的主要来源[11]。2024 年 6 月，英伟达正式发布全新开源模型 Nemotron-4 340B，具体包括基础模型 Base、指令模型 Instruct 和奖励模型 Reward 共三个模型。其中，指令模型 Instruct 的训练仅依赖大约 2 万条人工标注数据，其余用于监督微调和偏好微调的 98%以上训练数据都是通过 Nemotron-4 340B SDG Pipeline 专用数据管道合成。

[11] Gartner, "Maverick Research：Forget About Your Real Data - Synthetic Data ls the Future of Al," Leinar Ramos, JitendraSubramanyam,24 June 2021

当前，合成数据技术创新主要呈现以下几大趋势：**一是合成数据模型走向深度进化。**传统的数据合成方法多依赖统计学和机器学习的基本原理，当前数据合成技术聚焦于深度学习算法模型，特别是生成对抗网络(GANs)的广泛应用。GANs 通过一对竞争性神经网络—生成器和判别器的博弈过程，实现了前所未有的数据真实度与多样性，诸如 StyleGAN、BigGAN 等高级变种网络技术，极大拓宽了数据合成的应用边界。**二是多模态合成能力不断突破。**多模态合成技术通过整合不同模态的特征表示，能够同时生成声音、视频、3D 模型等多种类型的数据，不仅丰富了合成数据的维度，也促进了多模态理解和生成任务的进步，为复杂场景应用（如自动驾驶、虚拟现实等）提供了重要的技术支持。**三是强化学习与合成数据逐渐融合发展。**近期数据合成技术开始与强化学习算法深度融合，用于模拟复杂环境下的交互数据，帮助智能体在安全、成本效益高的虚拟环境中学习策略。这种结合不仅解决了现实世界数据获取难、风险高等问题，还极大地提升了智能体的学习效率与适应能力，尤其是在自动驾驶、机器人导航等领域展现出巨大潜力。**四是隐私保护与合规性技术不断增强。**面对日益严格的个人数据保护法规，数据合成技术创新性地提供了隐私保护解决方案—差分隐私、联邦学习与合成数据的结合，使得在不暴露原始敏感信息的前提下，也能生成可用于训练的高质量数据集，这不仅保障了用户隐私，也为金融机构、医疗保健等行业利用 AI 技术创造了条件。

三、应用赋能

随着大模型时代到来，人工智能技术能力快速迭代，持续推动各行各业的发展路径变革，全面带动大规模产业升级。在传统专用智能应用基础上，大模型通过进一步提供智能对话、文本创作、图像生成和视频生成等通用能力，提升赋能经济发展、民生服务、科学发现等各领域的深度和广度，将对全球经济社会发展和人类文明进步产生深远影响。

（一）人工智能赋能阶段性特征显现

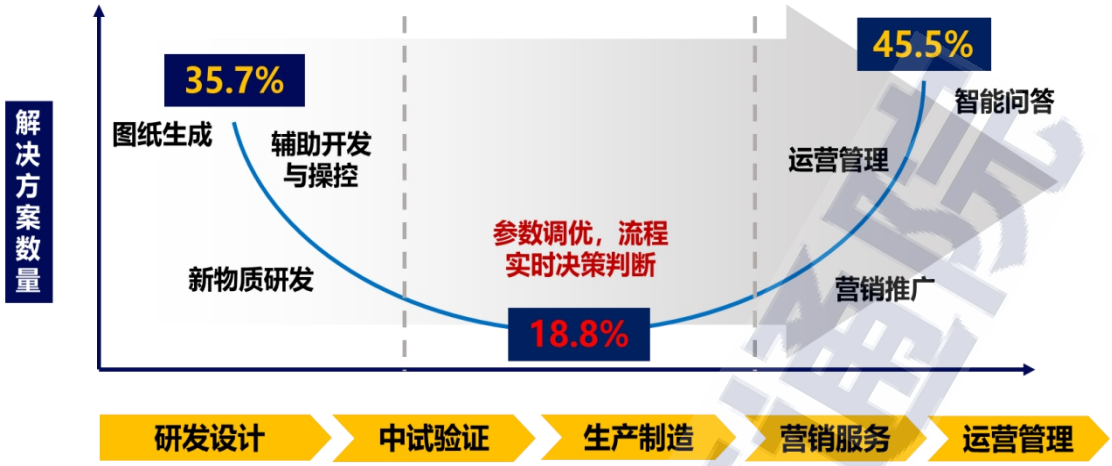
专用智能应用逐步成熟，通用智能落地前景广阔。一方面，专用小模型与行业场景融合深入。通常来看，小模型包括传统结构小模型及小参数预训练模型两类。其中，传统结构小模型网络结构以卷积神经网络、循环神经网络等为核心，在图像识别、语音识别等任务中接近到人类水平，由于其规模较小，训练及推理成本低，目前已在实际生产中广泛部署应用。随着人工智能技术需求增长，长尾场景增多，小规模预训练模型也可一定范围内解决多个下游任务。**另一方面**，随着全球掀起大模型应用探索热潮，大模型凭借更强大的分析、预测和交互能力，以及对场景任务的适应性，将有望带动工业技术产业实现创新性变革。短期看，大模型将重点赋能通用性较强或具备充足数据的场景，以提升执行效率为主。长期看，随着大模型不断向研发、生产管控等核心环节深度赋能，将显现变革效应，影响底层逻辑、产品形态甚至整个产业体系。值得探讨的是，

大模型的价值并非“替代”传统小模型，未来人工智能应用将呈现“大小模型协同”发展态势。

面向企业侧和消费侧的应用展现出不同的发展态势。面向企业侧，大模型应用更注重专业定制和效益反馈。我国提倡在社会生产领域和大众消费领域共同推进人工智能发展，这与我国产业结构特征密切相关。随着“人工智能+”行动等政策深入推进，大模型在多个行业成效显现，尽管各行业成熟度高低有别，但整体表现出较高的探索积极性。在以软件开发为代表的生产性服务业领域，部分行业具备相对良好的数据基础，实用效益的显现吸引了更多参与者加入，部分应用能够实现全流程优化，大模型应用成熟度较高。在交通、医疗等社会民生关键领域，大模型信控方案、无人驾驶、智慧诊疗等，已经实现点状场景的示范应用。在制造业领域，研发生产提质增效成为当前大模型落地的主要诉求。目前，在药物/材料研发和工业设计等研发设计环节、在安全巡检和质量监测等生产环节、以及在统计图表生成等经营管理环节，大模型均有应用落地，展现出巨大变革潜力。然而，由于线下生产流程的复杂性、严谨性和专业化要求，大模型在实时生产环节应用进展相对较慢。**面向消费侧，大模型应用更强调普惠适用和创意生成。**消费侧大模型应用百花齐放，对话助手类产品热度不减，创新应用不断涌现并逐渐成为用户使用的重点。根据相关统计数据，截至 2024 年 7 月，聊天助手类产品以 24 亿的流量位居首位，但流量环比降低超 15%；相比之下，图

像设计类产品流量环比上升了 16.83%。与此同时，用户需求不再局限于文字生成，创新生成式工具成为当前最具吸引力的功能。据知名投资机构 8 月发布的《Top100 消费级生成式 AI 应用》榜单显示，有 52%的公司专注于图像、视频、音乐、语音等多种模式内容的生成或编辑，且新上榜的 12 家公司中有 58%属于创意工具领域。

大模型应用在产业链各环节分布呈现“两端快、中间慢”特征，即产业链两端的研发设计和运营服务等知识密集型、服务密集型环节落地相对较快，生产制造等中间环节相对较慢。**从两端环节看，一方面**，科学研究、研发设计等知识密集型场景理论基础坚实，且普遍拥有高质量数据集，大模型赋能科研（科研智能，AI for Research and Development）的作用得以充分发挥。例如，某药物分子大模型可以减少新药研发中对小分子化合物的人工筛选计算量，使先导药的研发周期从数年缩短至数月，研发成本降低约 70%。**另一方面，**营销、运营等服务密集型场景跨行业通用性较强，成为大部分行业企业首选的大模型“试点”场景。以某医疗大模型为例，在诊前阶段，数字人就医助手能提供 7×24 小时的智能客服及专业科普服务，在分诊、导诊环节辅助人工服务，极大提升诊疗效率。**从中间环节看**，大模型在生产、制造等低附加值场景的落地存在一定局限性，面临场景选择难、低时效性、低可信度等问题。但在视觉等领域也出现了成熟的应用模式，如 TCL 通过视觉技术实现液晶面板缺陷检测，准确率超 90%、生产周期缩短了 60%。



来源：中国信息通信研究院

图 6 基于百个优秀案例统计的 AI 应用产业链分布

总体而言，当前大模型技术条件下，落地应用并非适用所有场景。目前大模型适用的场景侧重于对话交互、创意生成、知识管理类，而对于可解释要求高、确定性要求高、实时性要求高、场景动态性高、样本数据不易获取的场景，大模型如何有效应用还需要进一步探索。因此，**大模型赋能需要针对具体应用场景合理选择。**如，在产品设计、技术研发、知识管理、仿真验证等语料丰富、问题边界清晰的领域，大模型能发挥强大的自然语言处理能力能，极大提升劳动者的生产效率和创造力，从而推动工业技术产业实现创新性变革。而在实时生产中，由于对质量管理和流程精的高度要求，以及高质量专业数据获取的现实困难，尚不能采用大模型生成的“弱解释性”结果直接指导生产现场。综上所述，数据规模与质量、场景核心业务逻辑与大模型的创意生成和交互能力之间的匹配度，是选择的重要考虑因素。

（二）重点行业人工智能应用走深向实

装备行业重点关注研发与制造流程优化、产品智能化升级等方向，逐步渗透并重塑生产模式。**一是优化智能制造流程。**人工智能技术通过与工业软件、工业控制系统等关键工业要素的深度融合，能够结合市场需求、物料储备和设备状态实现工业生产的智能调度，形成生产过程的高效协同机制，支撑工业企业实现生产的智能决策。例如，AnyLogic 仿真平台通过强化学习降低重型机械运动，实现制造产线优化，协调对象的数量增加 66%，运动次数减少 11%。**二是提升智能产品与服务价值。**汽车、轨道交通、工程机械等装备逐步向智能化产品演进，基于视觉的环境识别成为目前主要探索方向。航空和交通领域成为开展增值服务的重点行业，如国外某航天公司飞行器座舱内的 AI 驱动系统可以通过评估和通知燃油水平、系统状态、天气状况和其他基本参数来帮助优化实时飞行路径。**三是产品设计与仿真优化。**人工智能辅助设计软件能够根据市场需求快速迭代产品设计、大幅提升仿真效率，有利于增强市场竞争力。例如，北汽福田应用 AI 找到最佳的设计路径，消除原结构太重和产品质量缺陷带来的问题，零部件从最初的 4 个零件变为 1 个，重量减轻 70%，强度增强 18.8%。**四是助力环保与可持续发展。**人工智能在节能减排、废弃物管理等方面发挥重要作用，通过优化生产流程、提高资源利用效率，减少环境污染。此外，人工智能还能助力绿色产品研发，推动装备制造业向低碳、循环方向发展。

消费品行业聚焦产品创新与智能化营销管理，正逐步改变消费者的购物习惯与体验，推动行业向更加个性化、智能化的方向发展。**一是新产品研发快速响应市场变化。**人工智能技术助力消费品企业快速响应市场变化，开发具有创新性和差异化优势的新产品。以智能家居为例，小米生态链中的产品，如智能灯泡、智能门锁等，通过集成人工智能技术实现了语音控制、场景联动等多种功能，提供了更加便捷、舒适的生活体验。**二是精准营销与智能客服。**一方面，人工智能通过分析消费者的购买历史、浏览行为等数据，预测消费者偏好、购买趋势，帮助企业定义目标用户并生成专属营销方案，实现精准营销和个性化产品推荐。例如，衔远科技的 ProductGPT 营销大模型，结合商品特性与热门趋势，可在几分钟内生成富有创意的营销内容，利用率高达 80%。另一方面，基于自然语言处理技术提供面向顾客的智能聊天与推荐等服务，替代传统的人工客服 24 小时不间断地为用户提供咨询解答、订单跟踪等服务，提升服务效率与用户体验。**三是供应链管理智能化。**人工智能的应用使消费品供应链变得更加智能与透明，从原材料采购到生产、物流、销售等各环节的信息都可以被追踪，并基于 AI 分析预测市场需求以制定合理的生产计划。例如，京东物流利用大模型的数智化供应链技术，聚焦从智能规划到智能仓储与运配，再到智能客服与营销的全链路降本增效，实现采购自动化率超过 85%，平均现货率超过 95%，库存周转天数降至近 30 天。

原材料行业聚焦生产过程管控优化，利用人工智能技术逐步改变传统的资源开发、加工和利用方式。**一是资源勘探与开发智能化。**人工智能结合地质学、遥感技术等多领域知识，提高矿产资源勘探的精度和效率。例如，加拿大的 GoldSpot Discoveries 公司使用机器学习算法分析地质数据，成功预测了多个金矿的位置，提高了勘探的成功率。**二是生产流程优化与节能减排。**人工智能在工业生产过程中实现精准控制，优化工艺流程，降低能耗和排放。例如，在钢铁行业中，宝钢股份利用人工智能技术实时监控高炉运行状态，通过调整操作参数，提高了冶炼效率和产品质量，减少了能源消耗和环境污染。**三是废弃物管理与资源回收。**人工智能通过图像识别、物联网等技术手段，实现废弃物的精准分类与回收。例如，瑞典的 Recycleye 公司开发了一套基于人工智能的废物分类系统，能够准确识别不同类型的废弃物，并进行分类处理。同时，还可以利用人工智能分析废弃物成分，探索其再生利用途径，促进循环经济发展。

（三）体系化推动人工智能落地应用成为共识

当前，人工智能应用持续走深向实，行业大模型已在金融、医疗、教育、零售、能源等多个行业领域实现了初步应用，并产生了明显的经济效益和社会效益。通过总结多方案例，大模型在落地应用通常涵盖需求分析、模型选型、中台建设、模型应用、运维管理、风险管理等重点环节，体系化推动落地应用成为引导人工智能技术实用化、普惠化发展的行业共识。

1. 开展战略需求分析是企业布局大模型的前提

大模型作为引领时代发展的战略性技术，已成为各行各业竞相发展的焦点。企业希望通过布局大模型对传统的业务流程、组织架构和经营模式进行全面升级和改造，以提升运营效率、降低成本、增强市场竞争力，并更好地满足客户需求。在布局大模型之前，企业通常全方位开展战略需求分析，统筹规划大模型所需各类资源，进而为大模型落地应用提供有利支撑。比如，思必驰科技股份有限公司在深入分析轨交领域的智慧乘客服务、智慧运营运维需求后，将自研 DFM-2 行业大模型与智慧轨交方案相结合，协助苏州轨交打造轨道交通多线路中心级语音平台，为乘客提供智慧出行新体验。多方案例表明，需求分析可以有效地帮助企业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，为后续的决策制定、资源配置和研发测试提供坚实的基础，助力企业在智能化转型中行稳致远。

2. 明确选型方案是企业研发大模型的关键一步

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，根据自身切实需求明确大模型技术选型可以为企业后续模型研发和应用夯实基础。通过分析百度“文心一言”、阿里“通义千问”等通用大模型和度小满“轩辕”、中石油“昆仑”等行业大模型的构建研发过程，**大模型技术选型通常包括模型生态、模型部署、模型协同、算力推算**等方面。企业在**选择开源或闭源两类模型生态**时，通常综合评估自身开发成本、开发周期、性能、安全性等要求，

基于所选模型生态通过搭配标准化的接口和丰富的工具包可以进一步提高模型开发的质量。**合理的模型部署策略是模型稳定可靠运行的基础。**公有云或私有云部署策略的选择涉及企业自身数据敏感性、数据规模、算力规模等要素。在模型部署过程中，企业通常考虑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。**模型协同是实现模型高效应用的重要手段。**企业通过整合自身大小模型资源，构建大小模型协同链路，充分融合大小型模型的优势，进而提高模型性能以适应不同的应用场景。比如，部分企业将大模型部署在云端以处理大规模计算任务，将小模型部署在终端设备，如智能手机上，以实现快速响应和低延迟。**开展模型推理算力评估可以提高企业资源配置效率。**企业进行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，并结合 AI 芯片显存大小推算所需芯片数量。一般情况下，如果任务对生成、理解、推理、决策的要求较高，推理算力充足，可以选择较大的模型，如 Qwen-72B、Llama-3 70B；对于准确度要求不高的较简单的任务，推理算力有限，或在边缘设备及端侧部署，可选择十亿级别的模型，如 Qwen-7B、Llama-3 8B。

3. 建设企业人工智能能力平台是工程落地的核心

人工智能能力平台正在成为企业全面智能化转型的基础底座。AI 能力平台通过统一管理算法、数据、计算资源和模型，以标准化的方式优化这些关键资源，帮助企业降低技术门槛，缩短开发周期，

并提升智能应用的构建、部署和维护效率，从而推动企业全面智能化。通常来看，AI 能力平台的核心功能包括统一纳管计算资源、高效开发流水线、统一管理运营 AI 资产等。**对于初创企业，AI 能力平台的构建应聚焦核心业务，利用云服务和开源工具快速搭建基础设施。**通过敏捷开发模式，企业能够在短期内推出最小可行产品（MVP），并使用容器化技术如 Docker 和 Kubernetes 简化部署，确保 AI 服务灵活支持业务需求。在有限资源条件下，初创企业可以通过这种方式快速构建可复用和扩展的 AI 能力平台。**对于大中型企业，AI 能力平台建设需要深度整合已有的大模型平台、数据能力平台和业务系统。**企业应通过梳理业务逻辑，使 AI 能力最大化赋能业务流程；技术上，则需构建高效的数据整合和模型管理平台，通过微服务架构提供低延迟、高可用的模型服务，提升业务运营效率。中国工商银行通过大小模型协同，构建了统一的 AI 能力平台，广泛应用于智慧服务、产品创新、风险防控等多个领域，显著提升了业务创新效率，实现了 AI 能力平台通过标准化开发和管理，加速智能应用的落地，并为金融行业提供了成功路径。

4. 构建智能体应用进一步释放大模型应用潜能

智能体作为将大模型转变为生产力的主要应用形态，通过智能体工具调用、智能体工作流、智能体人机交互等方式，能够快速理解和响应产业需求，拓宽大模型应用场景，为企业的数字化转型和智能化升级提供强大助力。**智能体工具调用有效解决大模型“有脑**

无手”的问题。大模型在感知、认知、推理等方面表现出色，但仍缺乏将决策转化为实际行动的能力。智能体可以实现意图理解、任务分解、任务规划，可通过调用小模型、实用工具或检索数据库等完成具体任务。**智能体工作流进一步推动模型高质量输出。**智能体工作流在任务执行过程中可以将任务拆分为不同步骤，通过合理规划和多次迭代，实现更高质量的模型输出，确保任务顺利完成。比如，在评测大模型代码生成能力的 HumanEval 数据集上，GPT-3.5（零样本）的正确率为 48.0%，GPT-4（零样本）的正确率为 67.0%，远远高于 GPT-3.5。如果在 GPT-3.5 上搭配智能体工作流，GPT-3.5 的推理表现将超过 GPT-4。**智能体人机交互实现人类和智能体的优势互补。**人类在模糊概念理解、创造性思维、情感判断等方面具有特定优势，智能体在数据处理、任务规划、推理决策等方面更具优势。通过交互式学习，智能体可以逐步积累更多的人类经验，实现更高的鲁棒性和适应性。

5. 打造运维管理体系助力 AI 生产过程规范化

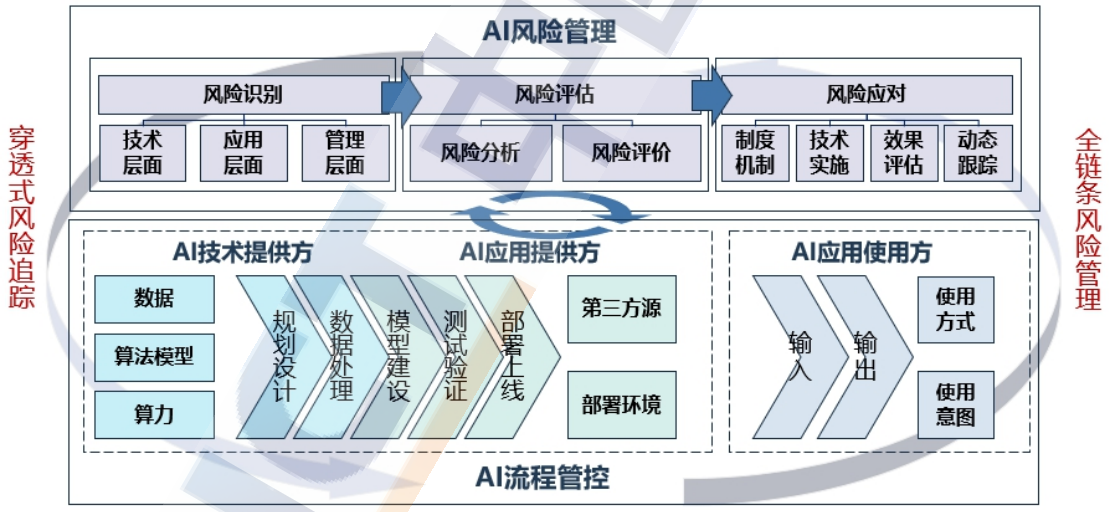
机器学习研发运营体系（MLOps）作为一种系统性方法论，在传统机器学习领域引起了广泛关注。市场上涌现了众多 MLOps 工具和平台，如 TensorFlow Extended (TFX)、MLflow、Kubeflow 等，这些工具提供了数据工程、模型训练、模型部署与交付、模型监控与运营等模型生产全流程的支持。通过将 MLOps 纳入机器学习工作流中，打造模型研发运营管理体系，实现模型的快速迭代、持续交付

和持续运营，提高业务响应速度，从而系统性解决**模型烟囱式生产周期长、生产过程和资产管理欠缺、跨团队协作难**等问题。而大模型时代的 MLOps，**基于流程化、自动化、持续迭代、可管理等原则，提升大模型的可修正和可运营能力，加速大模型规模化落地步伐**，提升运营管理效率，打造价值闭环。例如，某金融机构的反欺诈场景，通过 MLOps 体系管理和运营 AI 模型，可从每笔交易中动态学习，从而提高检测可疑活动的能力，显著增强欺诈检测能力。某企业物流场景，通过 MLOps 管理网络路线规划的模型，实现交通和天气数据不断实时更新，从而实现更高效的路线交付和更优的客户满意度。

6. 注重风险管理为大模型落地保驾护航

人工智能技术的应用场景持续拓宽的同时，新型人工智能技术应用风险持续涌现，全球将人工智能安全治理列为“优先议题”。对于产业界而言，亟需从风险管理和流程管控的角度出发，构建一套精准识别、全面防范、有效管控人工智能风险的治理落地方案。**在风险管理方面**，构建“风险识别-风险评估-风险应对”的风险管理链路，结合 ISO/IEC 42001《人工智能管理体系》等国际标准架构，立足我国产业实践提出风险管理方案。目前，已有百度、阿里云等企业推动构建基于风险的全生命周期合规化管理体系。首先，从技术、应用和管理等多维度进行风险识别，找准风险点；其次，对风险进行分析和评价，确认风险的危害程度；最后，根据特定风险类别实施对应的风险应对方案。例如，通过对人工智能模型的关键核

心资产进行全流程加密，对模型生成内容进行评估测试以及对恶意内容进行过滤等方式应对人工智能安全风险。**在流程管控方面**，打通“技术提供方、应用提供方、应用使用方”全链条风险管理链路，将风险治理理念及人工智能治理要求贯穿于人工智能规划设计、数据处理、模型建设、测试验证和部署上线等各类活动的全流程，满足不同技术领域、不同业务规模的产业实践需要。产业界也有头部企业积极开展面向流程的实践，如中国移动编制《人工智能安全风险防控工作指引》，依托“技术提供方、应用提供方、应用使用方”三类角色并围绕平台、数据、模型算法、业务服务、人员组织等治理要素，推动落实相应安全措施。



来源：中国信息通信研究院

图 7 人工智能风险管理体系

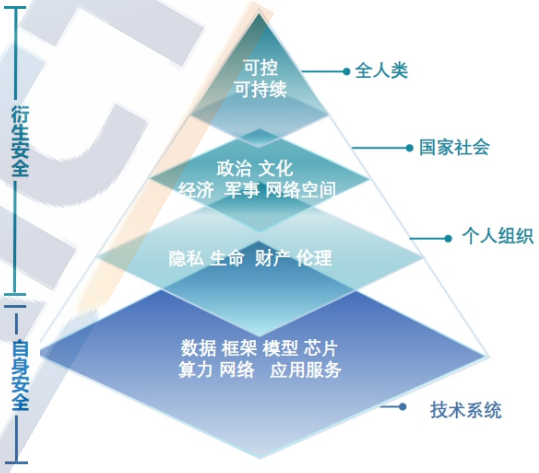
四、安全治理

人工智能在服务经济社会发展的同时，也诱发出数据安全、隐

私保护、虚假信息传播、劳动力取代、科技伦理挑战等诸多风险。为寻求应对策略，全球纷纷调整人工智能安全治理布局。纵览全球举措，国际合作愈发紧密，各国治理进程不断提速，产业组织发挥重要作用，安全技术体系逐步完善，全球人工智能安全治理正处于“从原则走向实践”的关键阶段。各国积极构建人工智能治理体系，建设标准体系，提升技术安全能力，加强社会参与和提升公众素养，**旨在实现安全与性能平衡发展，让人工智能驶向更加有序、普惠全人类的未来。**

（一）人工智能技术应用带来多重挑战

得益于算法不断突破、预训练大模型迅猛发展、多模态技术融合应用，人工智能技术产业发展进入快车道。作为未来颠覆性技术，人工智能在催生新产业、新模式、新动能，成为新质生产力的同时，也引发多维度的安全挑战，包括自身安全和衍生安全风险。



来源：中国信息通信研究院

图 8 人工智能风险示例

人工智能的自身安全主要是人工智能技术系统的安全问题，涉及技术系统部署所依赖的传统基础设施，以及模型、框架等人工智能系统特有的部分。因此，人工智能技术系统一方面面临传统的信息安全问题，如网络钓鱼、DDoS 攻击及网站篡改等网络威胁；另一方面面临一些新问题，如算法模型可解释性不足、框架安全漏洞、数据标注不规范等挑战。算法模型方面，由于深度神经网络存在非线性、大规模特点，导致在理论上难以证明其行为，同时因“模型幻觉”造成生成内容不可信。人工智能框架方面，安全漏洞可能遭到恶意利用。据报道黑客利用开源人工智能框架 Ray 的安全漏洞，数千家网络服务器遭受攻击，超过 10 亿美元算力遭到“劫持”[12]。数据标注方面，标注数据的不准确和不一致将导致模型在训练过程中学习到错误知识，生成错误、歧义或不符合语境的内容。

人工智能的衍生安全问题主要是由于技术系统风险管控不当，以及技术系统被滥用、误用或遭到外部攻击，对个人组织、国家社会，乃至全人类造成的安全问题。**一是对个人组织产生隐私、生命、财产、伦理等风险。**随着人工智能技术的发展和应用，人工智能已经潜移默化地融入了个人和组织的生产生活中，技术被滥用或失控后，个人和组织的切身利益会遭到侵害。2023 年以来，全国已经发生多起不法分子利用人工智能换脸、换声技术实施诈骗的案例，涉

[12] https://www.oligo.security/blog/shadowray-attack-ai-workloads-actively-exploited-in-the-wild

案金额高达百万元人民币[13]。**二是对国家社会带来政治、文化、经济、军事、网络空间等方面的风险。**近年来，人工智能技术已经被用于军事等领域，用于降低军事投入和开辟数字战场。如以色列军队利用人工智能技术打击了加沙地带 1.1 万多个目标，一天内发现并摧毁了 150 个隧道[14]。**三是对全人类造成不可持续、不可控的风险。**人工智能模型性能突破需大量、高效的算力支持，但模型训练导致大量资源浪费，抬高碳排放水平。以 GPT 训练为例，1750 亿个参数的 GPT-3 模型能耗相当于 1287 兆瓦时的电力，还产生了 552 吨二氧化碳[15]。在未来，前沿人工智能模型在化学、生物、放射性物质、核武器（CBRN）领域还可能造成不可控的极端风险。联合国秘书长古特雷斯多次警告人工智能的快速发展可能会导致严重的意外后果，并将其与核武器比肩[16]。

（二）全球人工智能安全治理正处于“从原则走向实践”的关键阶段

1. 国际层面推动形成治理共识，围绕安全议题合作愈发紧密

国际组织推动形成共识文件。联合国成立“人工智能高级别咨询机构”，负责分析人工智能国际治理并提出政策建议。2024 年，联合国大会先后通过《抓住安全、可靠和值得信赖的人工智能系统

[13] https://www.thepaper.cn/newsDetail_forward_23245389
[14] https://www.thepaper.cn/newsDetail_forward_25306308
[15] https://www.thepaper.cn/newsDetail_forward_27032901
[16] https://www.thepaper.cn/newsDetail_forward_23722410

带来的机遇，促进可持续发展》，以及中国主提的《加强人工智能能力建设国际合作》决议，为全球提供制度蓝图。同时，不断推动将治理落在实处，2024 年 9 月联合国人工智能高级咨询机构发布最终报告《为人类治理人工智能》，为应对人工智能相关风险和治理差距提出具体建议。**经济合作与发展组织**（OECD）于 2024 年 5 月更新人工智能治理原则，推动人工智能重要定义达成共识。**金砖国家**于 2023 年 8 月，已经同意尽快启动人工智能研究组，推动建立有共识的治理框架和标准规范。

全球围绕安全议题开展紧密合作。2023 年 11 月，英国举办全球人工智能安全峰会，28 国及欧盟签署《布莱切利宣言》。2024 年 5 月，韩国和英国共同主办人工智能首尔峰会，27 国及欧盟签署《首尔宣言》，将“安全、创新、包容”列为三项原则。此外，英、美、欧盟等 11 个国家和地区签署《首尔人工智能安全科学国际合作意向声明》，各国将依托人工智能安全研究所，强化前沿人工智能系统研究合作，促进技术资源、大模型信息、测评数据共享，推进人工智能安全科学研究。2024 年 11 月，美国于旧金山召开国际人工智能安全研究所网络成立会议，以推进人工智能安全方面的全球合作和知识共享。

2. 全球主要经济体完善治理体系，安全落地实践成为重要抓手

欧盟采取统一立法治理架构，布局风险管理等标准配套举措。体系上，采取风险分级的基本治理框架。2024 年 5 月，欧盟理事会正式批准《人工智能法案》，按照人工智能潜在的不同风险级别，划分为不可接受风险、高风险、有限风险、低风险四种风险等级，并引入不同的规则来解决这些风险。**举措上，通过加强管理、产业标准化等手段寻求人工智能安全配套方案。**欧盟成立专门的人工智能办公室，以监管人工智能的发展，支持可信人工智能的使用，并防范人工智能风险。2023 年 5 月，欧盟委员会发布支持欧盟人工智能政策的标准化请求，包含人工智能系统风险管理、符合性评估和质量管理在内的 10 余项标准已正式成为下一步工作计划。2024 年 9 月，100 多家人工智能相关方签署《人工智能公约》，通过自愿承诺、互助合作，如分享信息和资源等方式，共同应对未来可能出现的挑战。

美国沿袭行业自律治理方案，开展安全测试夯实技术监管。架构上，统筹监管资源促进行业自律。2023 年 10 月，美国总统拜登签署关于《安全、可靠和值得信赖地开发和使用人工智能》的第 14110 号行政令，发布全面的人工智能治理方法，明确了各监管机构的行动目标与期限。截至 2024 年 11 月，促使 16 家领先公司自愿承诺推动安全、可靠和可信的人工智能发展的工作。**行动上，政府积极推动安全测试。**第 14110 号行政令明确提出由美国国家标准与技术研究院（NIST）负责标准研制工作。同时，美国商务部将通过 NIST

制定指导方针和最佳实践，以促进开发和部署安全、可靠和值得信赖的人工智能系统的行业共识标准，包括创建评估和审计人工智能能力的基准、开展“两用基础模型”的红队测试等。2024 年 8 月，美国加利福尼亚州发布第 1047 号《前沿人工智能模型安全可靠创新法案》，明确各类测试、安全和执法标准。

英国治理方案注重“促进创新”，开发开源平台推进安全实践。方案上，营造有利于创新的政策环境。2023 年 3 月，英国发布《促进创新的人工智能监管方法》提出优先考虑指导、资源措施等较轻的干预手段，强调监管的合比例性，释放人工智能创新活力。**手段上，推出开源安全评估平台掌握模型安全性。**2024 年 5 月，英国人工智能安全研究所推出开源的“Inspect” 人工智能模型安全评估平台，帮助产业评估人工智能模型的核心知识储备量、推理能力与自主能力等性能，提高人工智能模型透明度及可重复性。

新加坡实施温和干预方案，打造可验证的安全测试机制。体系上，不断完善人工智能治理框架。2019 年 1 月，新加坡发布《人工智能治理模型框架》，为新加坡系统化探索可信人工智能生态系统奠定重要基础。2024 年 5 月，新加坡迭代发布《生成式人工智能治理模型框架》，聚焦于生成式人工智能特点，细化相关原则，为企业提供详细的指导和建议。**实践上，推动人工智能测试工具集建设。**2022 年 5 月，新加坡发布人工智能治理工具包“人工智能验证”（AI

Verify），结合技术测试和基于流程的检查方法，帮助企业对自身的人工智能系统进行评估。2024 年 5 月，新加坡在 AI Verify 中再添“登月计划”（Moonshot），包含基准测试、红队测试、测试基线，帮助开发人员根据风险基线测试人工智能模型，推动人工智能安全应用。

我国展现新型举国体制治理优势，着重保障人工智能应用安全。整体上，从框架规范到精准治理的体系建设。我国遵循《新一代人工智能治理原则——发展负责任的人工智能》《新一代人工智能伦理规范》勾勒出的基本框架和行动指南，秉持科技向善的基本理念，出台《关于加强科技伦理治理的意见》《科技伦理审查办法（试行）》等文件，加强科技伦理审查和监管。与此同时，延续人工智能治理领域精细化的立法特征，出台《互联网信息服务深度合成管理规定》《生成式人工智能服务管理暂行办法》，发布《人工智能生成合成内容标识办法（征求意见稿）》，聚焦互联网信息服务等重点领域的监管。**措施上，通过安全评估、备案等举措，保障信息服务领域安全。**2024 年 4 月，国家网信办发布首批生成式人工智能服务已备案信息的公告，截至 8 月，国内已有近 1919 个深度合成算法、190 个生成式人工智能服务在国家网信办完成备案，形成良好示范效应。

3. 产业组织发挥技术研究和治理协同优势

产业组织积极发挥技术研究和治理协同优势，通过发布治理框架，制定标准规范等多种形式促进人工智能治理。

治理框架方面，2023 年 1 月，NIST 发布《人工智能风险管理框架》（AIRMF1.0），提供了一套组织流程和活动来评估和管理风险。2023 年 8 月，美国 AI Now 研究所等机构联合发布“零信任人工智能治理”框架，为政策制定者提供一套治理路线图。2023 年 12 月，中国信息通信研究院依托中国人工智能产业发展联盟（AIIA）筹建安全治理委员会，发布“人工智能风险管理体系”，旨在持续推动人工智能安全治理技术能力提升和安全应用落地。

标准规范方面，ISO、IEC、IEEE、CEN-CENELEC、SAC-TC 28-SC42 等国内外标准组织在 2022-2023 年就人工智能安全已开展大量标准化工作。2023 年 2 月，ISO 和 IEC 联合发布《人工智能风险管理指南》，指导企业有效实施和整合人工智能风险管理的流程。5 月，欧盟委员会要求 CEN 和 CENELEC 起草新的欧洲标准，以支持《人工智能法案》实施。近年，我国人工智能标准化工作不断提速， 2024 年 6 月，工业和信息化部等四部委联合印发《国家人工智能产业综合标准化体系建设指南（2024 版）》，加快构建满足人工智能产业高质量发展和“人工智能+”高水平赋能需求的标准体系。

安全技术方面，2024 年 1 月，NIST 发布了关于对抗性机器学习攻击的报告，总结提出了包括数据投毒、模型窃取、成员推断和属性推断攻击等多种攻击方法。2024 年 4 月，MLCommons 制定基准测试 v0.5 验证评估 AI 模型的安全性，使用超 43,000 个测试提示词、

Meta 的 Llama Guard 来评估大模型对危险提示的响应。7 月，MLCommons AI Safety 工作组发布 v1.0 测试版本，构建相对全面的方法来衡量大型语言模型的安全性，其中包括部分未公开状态的“隐藏测试”，以确保测试的安全可信性。2024 年 4 月，中国信通院依托中国人工智能产业发展联盟发起大模型安全基准测试 AI Safety Benchmark。在内容维度，整理了 50 余万条测试输入，涵盖了底线红线、信息泄露和社会伦理等风险类型。在模型安全维度，结合了 16 种新型的模型攻击方法，设计了 80 余种攻击模板。截至 8 月，已经对国内外 25 家开闭源大模型进行了安全测试，初步摸清当前大模型的安全边界。

4. 企业积极开展负责任的技术研发与应用

为应对人工智能在当下遇到的可信问题与挑战，企业积极探索应对人工智能风险的新方案，从管理和技术两个维度落实自律。

管理体系更加重视全流程管控。一方面，企业设置内部治理组织，统筹人工智能治理工作。IBM、微软、谷歌、Lucid AI、百度、商汤、阿里、蚂蚁、旷视等多家科技公司设立了伦理委员会，并统筹推动对相关产品实施伦理审查。另一方面，发布伦理准则指引，指导企业具体实践。IBM、微软、谷歌、百度、腾讯、阿里、商汤、360 等国内外多家企业推出企业人工智能原则，包括对社会有益、安全、保护隐私、公平、透明、可解释、可控、负责任等方面。

安全技术方案向一体化、定制化发展。一方面，安全技术方案整体涵盖了风险识别、评估、防御等多个环节。**风险识别和评估方面，**微软推出风险识别工具 PyRIT，可评估大模型生成内容的安全性，推动了风险识别的自动化和智能化。谷歌与 Jigsaw 发布的 Perspective API，可实时识别、评估和过滤网络有害言论。**风险评估和防御方面，**奇安信集团发布 AI 安全整体应对方案，包括 AI 安全框架、解决方案、评估服务和测试工具等，可对大模型内容安全、科技伦理等风险进行综合防范。另一方面，安全技术方案开始支持个性化的配置。以微软 Azure AI Content Safety 为代表的安全解决方案不仅实现了对有害内容的自动识别与干预，还赋予开发者自定义过滤规则的能力，与用户的定制化需求深度融合，进一步增强了内容的合规性。

五、发展展望

当前，人工智能正处于迈向通用智能时代的初始阶段，技术体系和产业生态正在加速构建。近期来看，人工智能重要发展方向包括：**一是增强语言大模型能力仍是技术升级的重点方向之一，**推理或将获得更多关注和资源投入。通过探索精细化的 Scaling Law，提升数据的数量和质量，不断提升模型的复杂逻辑推理能力，降低成本、减少幻觉，推动语言大模型成为未来通用人工智能的中枢。“慢思考”能力更强的 o1 系列大模型发布，预示着自博弈强化学习有望成为提升语言大模型逻辑推理能力的技术新范式；在预训练阶段边

际效益递减、后训练缩放定律（Post-Training Scaling Law）显现的背景下，未来可能有更多算力被投入到后训练和推理阶段。**二是多模态模型有望加速突破**，从以语言大模型为骨干的多模态模型，向原生多模态模型演变，提升图文理解和跨模态交互能力。**三是智能体**凭借其强大的环境交互、任务执行、自我优化等能力，将进一步拓宽人工智能的应用场景，大幅提升用户体验和工作效率，为人工智能赋能千行百业夯实基础。**四是具身智能**为智能体赋予“身体”，使其能够与物理世界交互、探索、获取经验并改进自身行为，实现思维智能与行为智能的有机融合，成为迈向通用人工智能的重要一步。**中远期来看**，与当前数字芯片不同的模拟计算、量子计算芯片等硬件或将逐步成熟，在此基础上发展的类脑智能等颠覆性技术，将为人工智能发展带来更广阔的想象空间。

人工智能的技术浪潮将推动更大范围的行业转型升级，助力行业迈向智能化新阶段。大模型展现出的巨大潜力不仅将促使人工智能产业迎来重要的拐点，还将进一步推动我国生产力和生产关系的深刻变革。**在行业大模型发展方面**，其演进将聚焦于三个核心维度：**一是增强行业通用性**，随着行业数据集的完善，针对特定行业的通用型大模型将逐步涌现。企业可在此基础上通过定制化开发满足不同需求，降低成本，促进智能化普及。**二是提升模型的专业稳定性与准确性**，确保经过行业专家优化的模型得到持续关注和应用，形成稳定可靠的应用模式，通过多模态数据处理，提升人工智能的决

策效率和准确性。**三是更多元化的人机交互方式**，人与大模型的交互方式将从文本向语音、视频甚至脑机接口等方式转变。新兴的交互方式将使大模型更易于使用，降低技术门槛，促进人工智能技术的广泛应用。**在赋能应用场景方面**，行业大模型应用在提升文档检索、操作指导、设计图生成和智能客服等基础功能之上，还将深入到生产流程优化等核心环节。通过综合分析业务和生产环节中的多模态数据，以数据驱动的方式优化决策过程，推动行业向更高层次的智能化发展。未来，大模型行业赋能的趋势将从当前提高交互能力的阶段，逐步向提高业务创新和集成发展的阶段迈进，最终实现与产业深度融合，推动行业变革，迈向通用智能时代。

人工智能安全治理工作迈向深水区，探索切实有效、多方共治、敏捷应对的落实方案成为全球共同议题。面向人工智能的下一个时代，人工智能产业生态正在加速形成和发展，全球人工智能安全治理不仅是抢占战略制高点和发展机遇的“关键点”，也是全球和人类需要共同应对的“必答题”。未来，人工智能安全治理的深化和落实更需多元协同共治。**一是**加速完善人工智能安全风险识别方法论，随着人工智能技术日益融入经济社会发展各领域全过程，其安全风险面不断扩大，更加敏捷、精准的安全风险识别机制仍需持续探索。**二是**不断强化风险评估与防范策略，重点从人工智能基础设施、算法模型、上层应用以及产业链等方面进行评估，形成切实有

效、动态迭代的风险识别与应对策略。**三是**持续加强人工智能安全技术治理，加强对算法模型毒性、鲁棒性、公平性等方面的评测技术工具研究，强化技术治理能力。**四是**推动开放协同的国际合作，中国需要与全球共同努力，开展人工智能安全治理的基础理论研究和共性技术研发，推动技术标准和指引的深入实践与应用，加强安全治理国际交流与合作，达成更广泛的共识，共同释放人工智能潜力，有效防范和应对治理风险。

中国信息通信研究院

地址：北京市海淀区花园北路 52 号

邮编：100191

电话：010-62302914

传真：010-62304980

网址：www.caict.ac.cn