IDC

2022中国大模型发展白皮书

——元能力引擎筑基智能底座

本研究由百度支持

CONTENTS目录

IDC观点

随着数字化转型需求增长，AI在企业中的应用也越来越多，AI开发门槛高、应用场景复杂多样、对场景标注数据依赖等问题成为AI规模化落地的挑战，而预训练大模型的出现则为人工智能带来了新的机遇与希望。大模型作为政府和企业推进人工智能产业发展的重要抓手，在识别、理解、决策、生成等AI任务的泛化性、通用性、迁移性方面都表现出显著优势和巨大潜力。

IDC预测未来大模型将带动新的产业和服务应用范式，在深度学习平台的支撑下将成为产业智能化基座，企业需加快建设人工智能统一底座，融合专家知识图谱，打造可面向跨场景或行业服务的“元能力引擎”。

具体来看：

大模型具有良好的通用性、泛化性，显著降低人工智能应用门槛。预训练大模型在海量数据的学习训练后具有良好的通用性和泛化性，用户基于大模型通过零样本、小样本学习即可获得领先的效果，同时“预训练+精调”等开发范式，让研发过程更加标准化，显著降低了人工智能应用门槛，成为AI走向工程化应用落地的重要手段。

深度学习平台为预训练大模型的发展保驾护航，两者结合夯实了产业智能化基座。深度学习平台是推动产业智能化转型升级的核心载体，为大模型的算法开发、训练、部署保驾护航。大模型加上深度学习平台，贯通了从硬件适配、模型训练、推理部署到场景应用的AI全产业链，夯实产业智能化基座，将加速产业智能化升级。

大模型在推进产业智能化升级中已表现出巨大潜力，企业应该尽早关注。大模型目前的产业应用包括面向企业提供AI中台基座、深度定制支持产品或生产的优化与创新、开放模型服务等。大模型已经在搜索、推荐、智能交互、AIGC、生产流程变革、产业提效等场景表现出巨大的潜力，企业应该尽早关注，在业务中布局。

未来还需加强大模型与真实场景需求匹配，推动大模型大规模落地。目前中国大模型厂商在模型布局方面较为完善，应进一步围绕行业赋能的广度和深度持续探索，不断夯实基于大模型的产品建设，推动大模型技术从实验室走向实际大规模落地。

01

前言

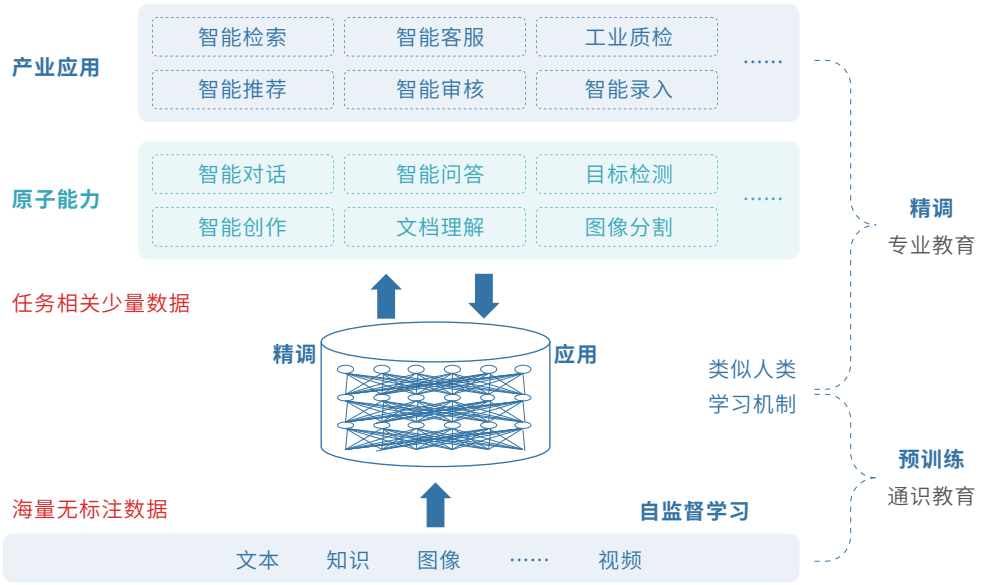
人工智能显著的溢出效应加快推进了新一轮科技革命，也带动了其他技术的进步。随着产业应用的深入、场景复杂度提升，随之而来的是数据的爆发式增长、算法的飞速更新迭代、算力的消耗指数上升，这些都对人工智能的发展提出新的要求。

1.1大模型发端及内涵 »

随着人工智能赋能实体经济进入深水区，企业通常面临数据资源有限、算力投资难度大、高水平人才稀缺的发展瓶颈。大模型作为解决上述问题的最优路径之一，可极大降低企业的技术门槛和开发成本。

IDC定义下的AI大模型是基于海量多源数据打造的预训练模型，是对原有算法模型的技术升级和产品迭代，用户可通过开源或开放API/工具等形式进行模型零样本/小样本数据学习，以实现更优的识别、理解、决策、生成效果和更低成本的开发部署方案。大模型的核心作用是突破数据标注的困境，通过学习海量无标注的数据来做预训练，拓展整体模型前期学习的广度和深度，以此提升大模型的知识水平，从而低成本、高适应性地赋能大模型在后续下游任务中的应用。在实践中，预训练大模型在基于海量数据的自监督学习阶段完成了“通识”教育，再借助“预训练+精调”等模式，在共享参数的情况下，根据具体应用场景的特性，用少量数据进行相应微调，即可高水平完成任务。

图1 训练大模型“预训练+精调”模式



来源:IDC&百度

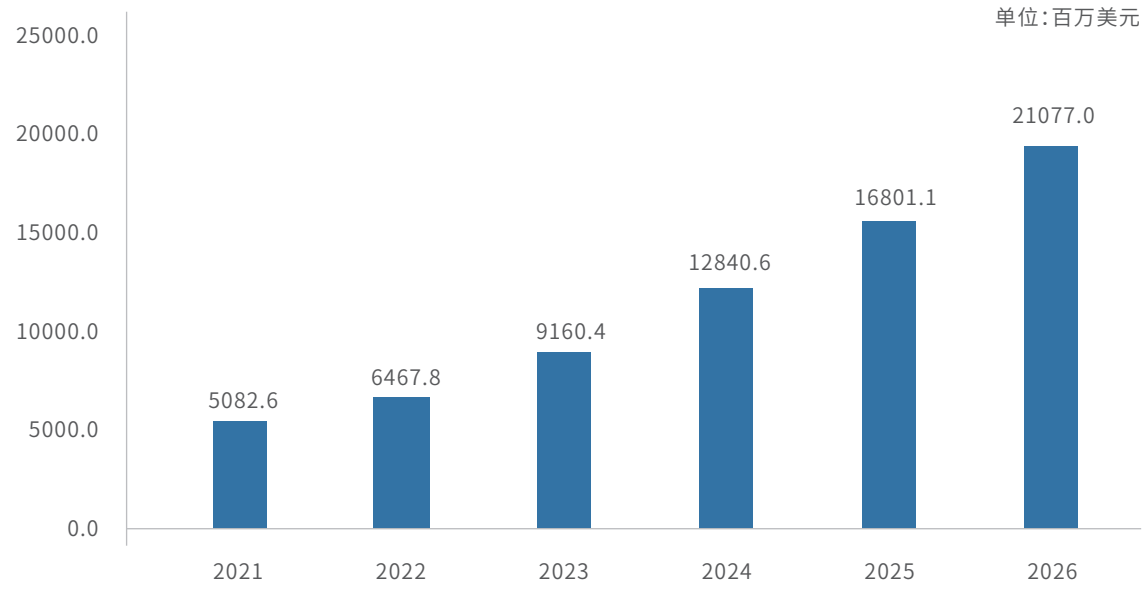
从技术的角度来看，大模型发端于自然语言处理领域，以谷歌的BERT、OpenAI的GPT和百度文心大模型为代表，参数规模逐步提升至千亿、万亿，同时用于训练的数据量级也显著提升，带来了模型能力的提高。此外，继语言模态之后，如视觉大模型等其他模态的大模型研究，也开始逐步受到重视。进一步地，单模态的大模型被统一整合起来，模拟人脑多模态感知的大模型出现，推动了AI从感知到认知的发展。

1.2 国家政策推动中国大模型加速发展 »

AI软件及应用市场快速增长，AI大规模落地成主要关注点

2021年中国人工智能软件及应用市场规模为51亿美元，预计2026年将会达到211亿美元，各行业的需求正大力推进AI的发展，将推动市场的持续增长。

图2 中国人工智能软件及应用市场规模预测，2021-2026



来源：IDC AI Cloud tracker

随着数字经济、元宇宙等概念的逐渐兴起，人工智能进入大规模落地应用的关键时期，但其开发门槛高、应用场景复杂多样、对场景标注数据依赖等问题开始显露，阻碍了规模化落地。AI大模型凭借其优越的泛化性、通用性、迁移性，为人工智能大规模落地带来新的希望。

国家政策对AI产业应用的关注与引导将推动预训练大模型加速发展

在国家层面，各国都在强调人工智能在发展中的重要性，并相继出台相关政策，希望在新一轮产业变革中占据上风。中国在“十四五”期间，针对人工智能的未来发展陆续出台了相关指导方案和激励支撑政策，对人工智能的整体发展方向和技术发展重点做出重要规划，同时提出加强算法创新与应用、推动算力基础设施建设、完善数据基础支撑体系等关键建议，倡导未来不断夯实产业发展新基础。

具体来看，上海市发布《上海市人工智能产业发展“十四五”规划》，《规划》中提到“十三五”时期上海人工智能发展面临的瓶颈：规模化应用深度不足，人工智能的应用以单个场景使用为主，深入传统行业核心业务流程、完整解决行业痛点、实现商业价值的应用较少；而大模型凭借其特性，直击痛点，将会

是未来突破发展瓶颈的关键技术。在基础理论研究中，《规划》还提到，“十四五”人工智能发展的主要任务是深化人工智能通用技术突破，面向自然语言处理、计算机视觉、语音识别等通用技术，支持相关科研机构和企业加快研发；建设先进算法模型，相关测试性能达到国际领先水平；支持对各类算法模型进行深度优化，适配实际应用需求。此外，北京市发布《北京市“十四五”时期高精尖产业发展规划》，《规划》重点关注：全面突破智能芯片、开源框架等核心技术，构建自主可控的产业链体系；建设国家级人工智能前沿研究中心、超大规模人工智能模型训练平台；融合人工智能和产业应用。同时，广州市也发布《广州市人工智能产业链高质量发展三年行动计划》，《规划》提到对大模型及其上下游产业生态链的布局要求和对相关技术平台在落地应用时的可靠性把握：针对昇腾、云从、讯飞等开放平台，未来将重点关注产业技术生态的塑造，促进AI精准赋能，提升人工智能应用的安全性与可信性。

大模型的技术特点、实现方式以及场景应用能力均与“十四五”时期政策期望相符，能够有效解决人工智能所面临的部分挑战。在场景驱动下，大模型技术将不断迭代发展，数据的增长和算力的发展也赋能模型训练和平台优化，形成技术供给和场景需求互通演进的持续创新力。

02

大模型成为AI开发新范式

2.1 人工智能发展的挑战与阻碍 »

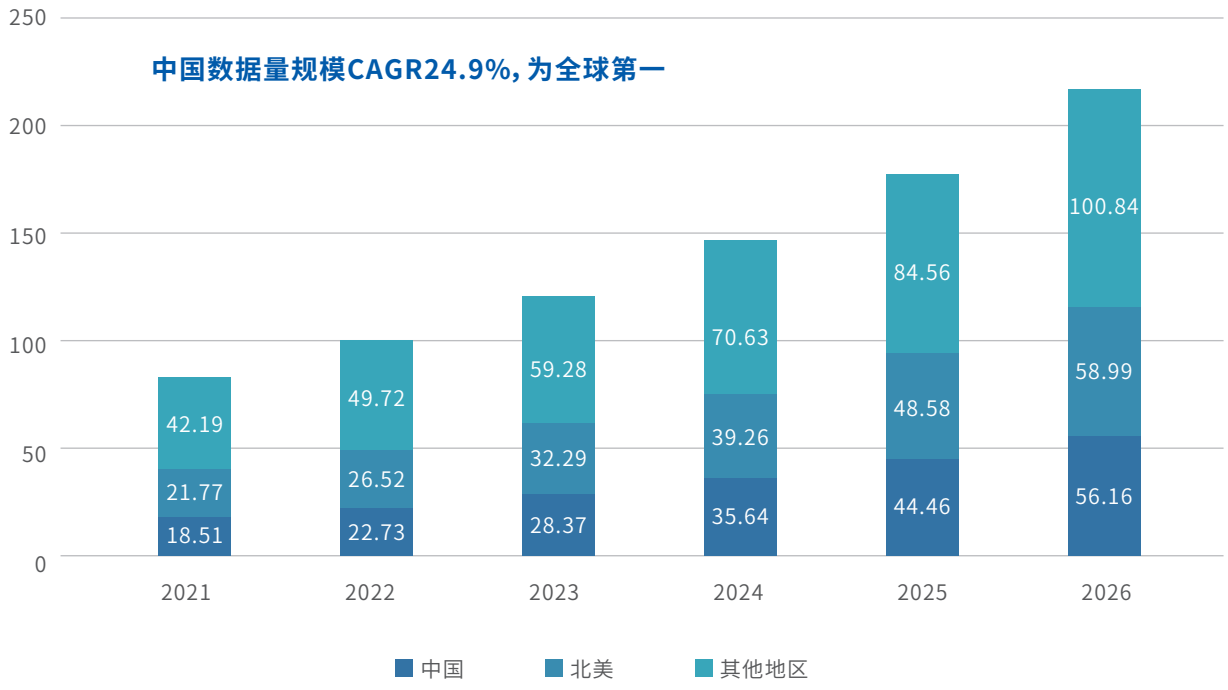
2.1.1数据、算法、算力是AI发展的驱动力也是瓶颈所在

数据、算法和算力是人工智能的三大要素，在核心技术发展以及应用落地中起着至关重要的作用，三者互相作用形成对人工智能的正向推动力。人工智能企业多数都是使用开源框架、自建数据库、购买算力的方式进行研究，但是随着企业信息化和数字化的发展，带来AI场景多样化和数据的井喷式增长，随之也使得算法的复杂度急剧提升、算力的消耗成倍增加，导致不少企业发展受限，亟需技术与产品的突破来解决AI所面临的问题。

信息化的发展使得数据量爆发式增长，带来数据“宝藏”的同时也对技术提出更多挑战

数据是AI发展的基石，随着人工智能、区块链、IoT等新兴技术的发展，中国乃至全球的数据规模都将高速增长。据IDC统计，中国数据规模将从2021年的18.51ZB增长至2026年的56.16ZB，年均增长速度CAGR为24.9%，增速位居全球第一。

图3 全球数据圈：按地区划分，2021-2026（单位：ZB）



来源：IDC Global DataSphere，2022

注：IDC将每年被创建、采集或复制的数据集合定义为数据圈（DataSphere）

随着数据量的高速增长，数据特征高维、模态格式多样的趋势也逐渐明显，对数据的AI建模也相应地更加复杂，涉及到研究对象的多变量维度，如时间、空间维度，计算复杂度会随之呈指数增加，数据标注难度也会增加。同时，海量的数据将不可避免带来更大的数据噪声问题、数据偏见风险，为模型如何有效利用好数据、学习其中的知识带来更大挑战。

数据是产业智能化发展中最宝贵的资源。海量的数据，为人工智能自监督学习带来巨大助力。利用好爆发增长的海量数据，将会是企业充分挖掘数据红利、构建数字经济下竞争壁垒的重要抓手。

应用场景多元化和复杂化，增加了模型生产的难度

随着AI技术的发展、产业应用的深入，应用场景变得更多元、更复杂。例如：工业场景下，有工业质检、安全巡检等应用，质检中不同产线生产的零部件千差万别；智能办公场景下，文档的分类、文档OCR识别、文档知识抽取、文档审校也都是不同的任务类型。解决一个场景的问题，往往需要多个任务的深度融合，涉及多任务统一建模等问题，因而对算法提出了更高的挑战。按照现在主流的算法应用，这意味着技术厂商需要针对不同场景、不同任务生产大量的算法或模型。一方面这将会导致重复性工作量加

大，另一方面也对开发人员的算法能力和业务理解有更高的要求。随着智能化转型的需求增加，AI开发门槛和研发效率问题凸显。

应用复杂度攀升，算力承压持续增加

算力是AI发展的基础设施，是通过对信息数据进行处理，实现目标结果输出的计算能力。除了要求提升计算能力，技术的发展对于软硬件也提出了新的要求。目前整体市场发展还不及预期，具体来说，硬件方面需要针对不同的场景和高性能计算能力进行拓展融合，满足研发企业的多芯部署、分布式优化、高性能计算的需求。目前人工智能芯片主要有GPU、FPGA和ASIC等类型，从英伟达GPU的发展可以看出，算力、内存、网络传输等都在提升，计算能力逐步增强，但在产业落地应用中的成本还相对较高。随着分布式训练的发展，数据存储和网络传输问题成为大模型训练的瓶颈。目前InfiniBand，已经可以支持节点内以及节点之间高吞吐低延迟的数据互联，缓解网络传输的问题，但数据存储仍存在挑战，需要新技术的出现来解决。在软件方面，厂商需要打造完整的开发软件栈，支持计算密集型算子和访存密集型算子协同编译优化，增强通用性编程能力，满足企业针对不同训练推理数据格式和量级进行底层编译以及融合调度和统一运营管理的需求。在整体软件栈中深度学习平台尤为重要，可以提供覆盖AI能力生产、运用、管理等全流程的工程化实践方法，推动产业链上下游协同创新，联动底层算力、数据和上层应用服务，打破企业在数字化转型升级中面临的多种瓶颈，解决数据成本高、模型开发难、算力分配不合理等问题。

2.1.2中国人工智能的其他挑战与阻碍

2022年是实现产业数字化的元年，人工智能加快赋能千行百业，与实体经济深度绑定，在医疗、城市、工业、能源、金融等领域进一步落地应用，给企业带来了新的发展方向， 除了底层数据、算法和算力对人工智能发展所带来的瓶颈以外，IDC认为人工智能领域还面临三大维度下的挑战。

- **数据互通壁垒明显，共建生态存在阻碍。**新一代信息技术与产业的深度融合扩大了网络空间的边界，数据作为一种新的生产要素，已成为推动企业智能化升级的重要资源。但同时，流转无序、区域性限制大、定价机制不完善、监管机制不完备等问题，导致数据不流通，数据对数字经济的放大和叠加作用没有得到充分发挥。数据的流通和共享是释放数字红利的前提，提高数据流通性可以推动社会资源配置的优化，节约社会成本。为此，需要明确数据的权责，完善政策，规范数据的使用，推动数据共享流通，支撑人工智能技术的高速发展。

- **技术门槛高，平台层挑战不断。**AI算法的开发与模型训练、调优有着一定的技术门槛，需要进一步依托在算法框架上的产品与工具套件，降低AI开发门槛。因此，深度学习与大模型平台需要向下衔

接硬件、向上承接应用。未来不管是训练还是推理，硬件的种类会变得非常繁杂，向平台层提出了更高的衔接要求；同时随着AI规模化落地的需求增强，平台面向实际应用需要持续降低模型开发的学习门槛、降低模型优化难度。当前，开发平台发展重点在于提供专业且丰富的技术组件，向下驱动算子和数据管理工具的高性能延展，向上带动产品线研发并推动门槛的降低。

- **人才储备不足，技术发展受限。**除了技术、战略部署、资金投入等问题以外，智能化发展所遇到的最大挑战便是人才短缺。IDC预计到2025年，全球500强中有一半的企业将自己开发软件，这将加剧企业软件人才不足的问题。随着人工智能落地场景的复杂度增加，需要更多既懂业务又能运用AI技术的综合型创新人才。目前企业和高校的合作可促进人才的优化配置，高校为企业提供技术的理论学习，企业为学校提供有效的实践基地。经济全球化的发展不断促进社会资源流动，资源配置方式根本性变革极大提高了资源的利用率，但目前来看人才缺口仍然存在。未来，需要进一步建立人才合作培养生态，接受人才及技术在机构间的循环流动，同时推动降低技术接触年龄，提前布局储备年轻人才力量。

2.2 大模型带来AI开发新范式 »

面对人工智能的各种挑战，预训练大模型的出现提供了通用化解决方案，从无标注数据中通过自监督学习获取大量“知识”，实现用更统一的方式推动人工智能产业落地。

2.2.1 大模型增强人工智能泛化性、通用性

在过去每一次关键技术的通用性得到解决后，生产方式都有巨大改变，生产水平也产生质的飞跃。人工智能是第四次工业革命的重要驱动力，所以，提升人工智能的通用性是加速产业智能化升级的关键。

“大模型”是打通人工智能技术通用性“任督二脉”的关键

过去在分散化的模型研发模式下，单一的AI应用场景下多个任务需要由多个模型共同支撑完成，每一个模型建设都需要算法开发、数据处理、模型训练与调优过程。预训练大模型增强了人工智能的通用性、泛化性，基于大模型通过零样本或小样本精调，就可实现在多种任务上的较好效果。大模型“预训练+精调”等模式带来了新的标准化AI研发范式，实现AI模型在更统一、简单的方式下规模化生产。

2.2.2 大模型降低人工智能应用门槛

大模型基于“预训练+精调”等新范式有效降低AI开发门槛

具体来说，大模型的通用性、泛化性以及基于“预训练+精调”等新开发范式，让AI场景应用的模型定制流程变得更标准化、效果优化更简单，有效降低对标注数据、算法人员能力的要求。围绕大模型布局相关的AI开发工具组件与平台，将大幅加速人工智能大规模产业化进程。例如百度文心大模型在模型层构建了基础（包括NLP、CV、跨模态等）、任务（对话、搜索、OCR等）、行业（能源、金融、制造、传媒等）三层大模型体系，深入考虑各大应用场景特性；向上打造工具与平台层，将大模型能力在开发平台与套件中输出；封装模型训练与精调、模型压缩与部署各环节等。这些都极大降低了AI开发门槛，让更多企业或开发者可以低成本、高效率地获得AI能力，应用到自己的业务中。

2.2.3 深度学习平台为大模型发展与应用护航

深度学习平台的发展已相对成熟，大模型的出现对深度学习平台来说是“如虎添翼”

深度学习平台面向多样的产业需求，基于开源框架提供算法模型以及工作组件和平台能力，向下协调调度硬件算力，向上支持各项任务，包含开发框架、算法模型以及工具平台三大核心层级，呈现出标准化、自动化、模块化特性。大模型则进一步增强模型通用性和泛化性，带来新的模型开发范式。深度学习平台与大模型合力，将进一步降低模型开发门槛、提升研发效率，贯通了从硬件适配、模型训练、研发部署，到场景应用的AI全产业链。

深度学习平台底层开发框架成为大模型与算力之间的桥梁

ASIC等芯片，通过简化底层硬件技术，在大模型与算力之间建立沟通。针对不同的模型和硬件，将资源抽象成统一的分布式资源视图，通过底层硬件感知和映射功能，找到软硬之间的最优组合，并将模型的运算步骤分配到相应的计算卡上，达到负载均衡、提升大模型训推性能的目的。

深度学习平台助力大模型解决训练、推理部署困难问题

超大模型训练、推理需要消耗密集和昂贵的算力等资源，对算法本身提出了极高的要求。在海量数据上训练百亿、千亿、万亿的参数，对模型训练速度、模型精度以及训练资源成本都是极大的挑战，深度学习平台通过超大规模并行方案，支撑大模型高效、高性价比训练。超大规模的模型参数，也让模型预测单次的成本与耗时都大幅提升，成为规模化的产业应用瓶颈。深度学习平台通过提供量化、稀疏、蒸馏、剪枝等能力帮助大模型在精度无损的情况下进行压缩，推动大模型轻量化和模型推理加速，为产业大规模应用做好保障。

大模型与深度学习平台相辅相成，将会持续释放红利，并渗透到各行各业的场景中。

未来，以大模型为生态基座的产业链将成为智能化升级过程中可大规模复用的基础设施。在大模型通用性、泛化性以及降低人工智能应用门槛的优势推动下，人工智能也将会加快落地，形成新的机遇。

03

大模型加速产业智能化变革

3.1 大模型带来AI技术与应用变革潜能被广泛验证 »

3.1.1 NLP大模型

自然语言处理（Natural Language Processing，NLP）是用计算机来模拟、延伸及拓展人类语言能力的理论、技术及方法，是融合语言学、计算机科学、数学等于一体的综合性学科。自然语言处理目前面临的关键问题是人类语言的复杂性和多样性，例如同样的词汇在不同的语境之下意思不完全一致、日常用语中的反讽等反向情感表达、句式结构的多变和缺失所引发的歧义以及方言和“行话”等语言个性化特点。

近十年来，深度学习成为NLP模型研发的主流技术框架，带来了巨大的进步，但仍然受限于对大量有标注数据的依赖，模型泛化性、通用性仍有不足。近几年，随着预训练技术的发展、算力提升以及NLP领域的海量数据和任务特性，大规模预训练模型首先在该领域取得突破。2018年，随着BERT的诞生，大规模预训练语言模型，利用海量的无标注文本自监督学习，即可深入掌握大量语言知识，刷新多个AI权威榜单记录。3亿参数的BERT模型在权威通用语言理解类评测榜单GLUE上的11个任务刷新纪录，将基准值推至80.4%，绝对提升了7.6个点，在机器阅读理解顶级水平测试SQuAD1.1的全部两个衡量指标上超越人类平均水平。由OpenAI推出的GPT系列模型，不仅在效果上刷新了多项记录，更是表现出高水平的生成能力，开放的API服务催生孵化了系列创新产品。国内文心ERNIE系列大模型在GLUE上实现9个任务突破90分，ERNIE3.0系列在问答、分类、情感分析、抽取、识别等93个典型NLP任务上刷新业界纪录。百度文心系列大模型已应用于百度搜索、信息流、小度等重要产品，服务数亿用户，也被广泛应用于百度智能云的智能文档、审校、客服等产品中。

3.1.2 CV大模型

计算机视觉（Computer Vision，CV）是指使用计算机及相关设备来模拟生物视觉的技术，即基于传统

或深度学习算法，赋能计算机理解数字图像和视频，并从各种模态的数据之中提取目标信息。其主要目标是开发“机器之眼”，不仅让计算机具备视觉能力，更让计算机识别、理解“看”到的多模态数据。

计算机视觉作为人工智能和深度学习的子领域，目前主要以深度卷积神经网络（CNN）和Transformer为支撑，针对各个应用场景开发优化类人视觉功能，例如厂商利用图像识别、图像和视频搜索、视频合成等技术应用于汽车交通、媒体标签等常用场景。当前技术上的瓶颈包括杂物遮挡、识别角度等问题。

计算机视觉大模型发展迅速。比如，2021年150亿参数的V-MoE被推出，该模型表现出大模型在缩放视觉模型方面的潜力，并在ImageNet上准确率达到了90.35%。此外，V-MoE具有可扩展性，其表示能力和迁移能力表现为SOTA。国内厂商也逐渐开始在计算机视觉方面深入探索，盘古CV大模型在ImageNet数据集的线性分类评估上，达到了与全监督相比拟的结果，在应用方面可提供OCR文字识别服务，支持通用类、证件类、行业类以及自定义模板识别等多个场景落地应用，目前已经在TFDS图像自动识别精度上超过人类检测员水平。另外，通用视觉模型“书生”（INTERN）在任务上也有优异的表现，在目标检测任务上平均错误率降低了47.3%。据了解，“书生”只需要少量的下游数据，就能超过CLIP基于完整下游数据的准确度。

表1 NLP&CV发展现状与挑战对比

	NLP	CV
现状	分别在语言理解与生成、智能创作、机器翻译、智能对话、知识图谱和定制化语言解决方案落地应用，整体算法发展顺利，数据源可获得性较强	2D数据工业质检、智慧城市落地完善，应用场景多、可商业化市场大，拥有最佳实践；人脸、OCR识别发展较为成熟
挑战	语言的歧义、文化差异及多样化、情感分析困难	3D/4D数据识别面临变形、光照、遮挡等可以依靠大规模预训练模型解决部分痛点的问题；数字人、数字孪生的数据获取困难，算法处理复杂
预期未来发展	以多个数据信息维度约束来验证情感分析及文本分析的准确性	打通数据融合以突破3D/4D获取瓶颈

来源：IDC

3.1.3 多模态大模型

多模态大模型的发展从OpenAI的CLIP(文本图像匹配)，以及Dall·E（文生图）拉开帷幕，目前跨多个模态的数据融合问题开始变成行业探究的重点。多模态是指多个模态感知与认知的融合。对于人类来说，所有感知交互方式的融合形成了社会交流；对于计算机来说，是通过对文本、图片、视频和音频等不同储存信息载体的认知和理解，结合环境因素来模拟人与人之间的交互方式。多模态技术的重要性不言而喻，让人工智能理解人类世界的最优办法就是让AI成功理解多模态信息并能够对此类信息形成分析、推理的逻辑和生成新信息的能力。

近年来，大模型技术发展推动多模态模型不断升级迭代。首先，预训练大模型赋能多模态机器学习的广度和深度，例如通用性AI大模型M6，十万亿级的参数持续提高模型上限，赋能模型应用的通用性，进而拓宽大模型应用广度，覆盖电商、智能交互等业务场景。同时，多模态预训练模型mPLUG荣登全球权威“机器视觉问答榜单”（VQA Challenge 2021）榜首，并超越了人类平均水平。此外，多模态大模型能够实现图像、文本、语音等模态之间的统一表示和相互生成。例如，百度文心ERNIE-ViLG 2.0文生图大模型在公开权威评测集 MS-COCO 和人工盲评中效果位于前列，在语义可控性、图像清晰度、中国文化理解等方面均展现出优势，初步实现在多个场景的商业应用。

我们看到，头部厂商在多模态大模型领域持续布局，注重模型整体通用性的同时不断提升子领域的优化体验和技术升级。未来，基于技术的不断突破，多模态将持续拓展各行业场景下的信息融合应用。

3.1.4科学计算大模型

科学计算领域近年来发展态势向好，持续推进技术突破。科学计算指的是通过计算机高效率完成再现、预测和发现客观世界运动规律及演化特征的全过程，即出于解决科学和工程中的复杂数学问题的目标，优化计算机性能以完成数值计算。

近年来“AI+科学计算”（科学智能）也在引发科研方式的大变革，如生物制药、气象预报、地震探测、材料研发等科研领域，大模型技术同样也在这些领域带来巨大的突破。科学计算的子领域生物计算（Bio-computing），即基于生物学固有理论信息和大量的生物学实验结果及研究分析开发的解决生物学问题的计算模型，正是走在前列的科研方向。2021年以来，生物计算领域持续突破。例如，DeepMind推出的AlphaFold2能够覆盖98.5%的人类蛋白质组，并对20种其他生物蛋白质的结构进行预测；同时，该公司与EMBL-EBI（欧洲分子生物学实验室）合作，推出蛋白质结构数据库以储备和匹配蛋白质3D结构图像。各大企业自此之后纷纷提出AI for Science的概念，着手利用人工智能技术加快重点科学技术研发与突破。目前，国内市场活跃产品有头部厂商打造的通用大模型，融合自监督和多任

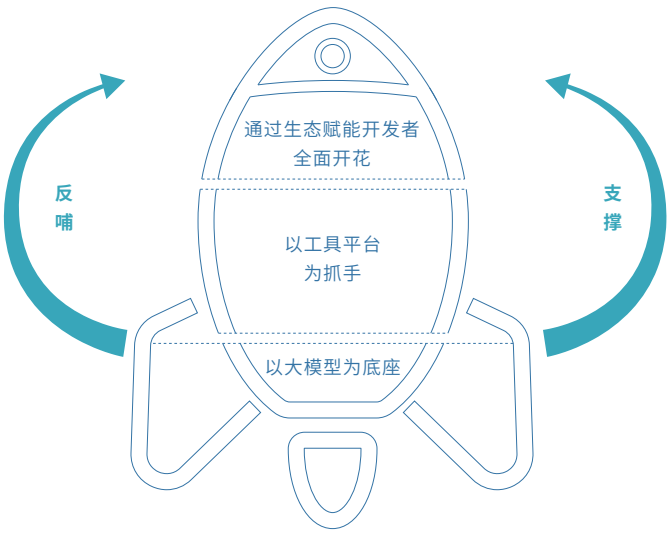
务学习以赋能生物医药行业，例如百度文心的蛋白质结构预测大模型、化合物表征学习大模型等；此外，也有专注于生物计算大模型以发现靶点、研发新药的百图生科以及医药知识图谱平台德睿智药等。

3.2“模型+工具平台+生态”三级协同加速产业智能化 »

新一代信息技术与产业的深度融合是抢占数字经济发展领先地位的必然选择，人工智能是新一轮科技革命中和产业交叉最密集的技术。对AI核心驱动要素的潜能激发，正在从生产规则、落地方式、商业模式等多维度重塑产业智能化。

大模型增强了AI技术的通用性，让开发者以更低成本、更低门槛，面向场景研发更好的AI模型，助力普惠AI的实现。但目前，基础大模型距离大规模产业应用并成为产业基座还有很长的一段路要走，不仅需要有与场景深度融合的大模型体系，也需要有支持全流程应用落地的专业工具和平台，还需要开放的生态来激发创新；三层之间交互赋能，才能形成良性循环的产业智能化生态共同体。我们看到OpenAI在开发GPT大模型的过程中具有相似的思路，在不断加强大模型本身性能的同时，将GPT打包成产品，对外提供API服务，相关开发者、企业、学术机构以及高校都可以申请使用。开放后，大量开发者利用 API 开发出了各种各样的功能，例如翻译机、网站生成器等；OpenAI则通过用户获取了更多的行为数据，形成了对GPT的反哺。由此可见，“模型+工具平台+生态”三层共建有助于业务的良性循环，也更容易借助长期积累形成竞争壁垒。

图4 模型+工具平台+生态"三级协同加速产业智能化



来源：IDC

- **模型层是大模型能力的核心引擎。**模型层从技术发展与产业应用出发，主要包含基础、任务、行业大模型，模型的训练要求高，算力消耗大，建设人员主要为高级算法人员。基础大模型主要优势在于其通用性，可以让该技术方向的相关应用任务都得到进一步提升，但也正是这样的优势导致基础大模型在行业、任务中表现尚未最优。任务大模型是在基础大模型上，面向典型的任务，如对话、搜索、文档智能、人脸识别、OCR等，进一步结合任务特性，优化模型算法，学习任务相关数据与知识，从而使得大模型在任务上表现出更优异的效果，很多任务甚至可以零样本直接应用。行业大模型是在基础或任务大模型上，进一步融合行业数据、知识以及专家经验，提升大模型对行业应用的适配性，目前在金融、能源、制造、传媒、城市等已经有头部企业或机构与科技公司或科研单位联合发布了行业大模型。基础+任务+行业三层大模型相互促进，共同支撑起产业转化。

- **工具平台层将大模型落地研发标准化，推动AI广泛落地。**大模型在深度学习平台的有力支撑下，实现了高效生产并真正为产业所用，深度学习平台为大模型解决硬件适配，提供蒸馏、剪枝、压缩等技术并向外输出部署方案，支撑自然语言处理、计算机视觉、跨模态等各类大模型的应用。同时，基于深度学习平台进一步推出基于大模型的AI开发平台、工具套件、大模型API等，将基于大模型的精调、大模型能力调用产品化，让更多AI应用型开发者或业务专家，可以零门槛或低门槛地将大模型应用于自己的业务中，以此全面释放大模型效能，助力开发者效率提升。

- **生态层是基于大模型能力打造共创、共享社区。**大模型“预训练+精调”的新研发范式，让AI在识别、理解、生成等方面的能力实现突破，带来深度语义理解、智能交互、内容生成的技术与产品变革动能。打造基于大模型能力的生态，提供能力、工具、服务，连接供需，吸引更多的开发者和企业共创、共享，是释放大模型潜力的重要路径。

“模型+工具平台+生态”的模式需要协同优化，拓宽人工智能技术落地的场景覆盖广度，加深产业实际应用的深度，共同加速产业智能化，推动人工智能技术赋能千行百业，惠及千家万户。

3.3大模型加深度学习平台正在成为产业智能化基座 »

3.3.1 大模型推动人工智能向着通用化、工业化、集约化发展

大模型是人工智能走向工程化应用落地的重要手段

当前，人工智能已经从安防识别、智能推荐、语音对话等多种应用场景进入社会生产生活。随着产业发展需求层级的不断深入以及数字经济、智能经济、数字化转型、新基建等政策的加快推进，人工智能赋能千行百业以实现大规模工程化应用落地迎来高速发展的窗口期。如何缩短人工智能研发周期、降低

开发应用成本、提升实际工作效率，已成为各行业关注的核心问题。大模型具备场景通用性和泛化性、工程标准化、大模型建设集约化的特性，可以满足实际产业发展中的应用需求，提供高水平能力的工程化实践案例，打造智能化升级的基础底座。

一是场景应用中的通用性与泛化性

通过学习海量数据，大模型可以不断丰富模型参数和模型结构。此外，通过引入相关知识，将数据与知识相结合，大模型能拥有更高的识别水平和模型迁移性，在广泛的基础任务和特定行业任务上均表现出较好的效果，仅需少量特定标注数据训练就可实现快速落地。换句话说，大模型作为基础设施，在上面进行简单的微调优化，将能够建造不同的建筑。

二是工程标准化

深度学习平台本身已具备标准化特性，AI模型开发包括模型选择、数据处理、模型优化、模型迭代等一系列环节。大模型的“预训练+精调”的范式，令AI模型开发变成基于预训练大模型+少量样本数据精调参数的通用流程，进一步增强AI模型的开发标准化、简化流程。

三是大模型建设集约化

当人工智能技术从实验室环境走向企业生产环境，由于研发环境和开发目的不同，企业更关注技术投入的高能效和低成本，但传统面向单点任务的模型反复开发和训练将不可避免地导致成本的增加，这成为AI赋能千行百业的关键阻碍。大模型恰逢其时，带动整体产业结构以“倒金字塔”形式更健康发展，但大模型本身的研发有数据、算法、算力的高门槛。因此，通过一部分具有领先的数据、算力资源供给能力和算法人才的企业打造大模型基础底座，可以帮助上层各行业领域的服务企业以更低的投入成本和更高的效率建设繁荣的产品应用生态。

3.3.2 深度学习平台解决大模型落地关键挑战，释放大模型潜能

深度学习平台是推动产业数字化转型升级的核心载体

随着人工智能技术的逐渐成熟和应用落地，AI技术平台、基础软件等产品进入大众视野，通过提供基础算法库和全周期开发组件，可以帮助开发者实现更高层级的创新突破和技术更迭。

一是实现算法模型创新开发

从编程范式来说，当前主流深度学习平台均支持动静统一的编程范式，即同时支持动态图和静态图两种类型，在实现动态图高效开发训练的同时，也支持开发后一行代码转静态图的训练加速和部署，可以大幅度提升开发者算法研发准确率和生产部署效果。从算法模型库来说，深度学习平台作为模型开发后资产沉淀的主要承载，可提供业界领先的算子和模型结构，帮助开发者基于大模型进行下一步调优和创新，提升研发和应用效率。从工具组件来说，深度学习平台不断融合强化学习、联邦学习、图学习、量子计算、生物计算等前沿技术，提供所需的专业化框架套件和解决方案，满足大模型在广泛应用场景的落地需求，例如飞桨打造图学习框架PGL，提供异构图数据采样和存储能力，以及图卷积神经网络、图注意力网络、基于图卷积的无监督学习网络等模型，并结合分布式嵌入存储能力实现大规模分布式训练。

二是支持企业全流程协同管理

大模型是人工智能算法的先进性成果，还需配合深度学习平台提供覆盖数据管理、模型开发、训练调优、推理部署、协同监管为一体的全流程研发套件，来满足企业用户侧的应用服务系统化开发管理需求。深度学习平台通过端到端的开发部署能力，纵深打造专业化能力套件，配合通用大模型、任务大模型以及行业大模型，赋能政府和企业不同场景下的自主AI模型开发。例如百度全功能AI开发平台BML整合底层开源框架以及上层数据处理、模型开发建模、模型训练管理以及端侧部署能力，支持多种方式建模和调参选择，辅助企业实现一站式模型定制能力。

三是加快大模型训练部署

在模型训练方面，数亿参数的大模型读写、存储和训练成本巨大，因此需要深度学习平台提供高效的大规模分布式训练技术，根据模型参数以及训练数据量的不同，实现底层资源弹性调度管理，全自动选择最优并行策略技术、高效计算及通信技术。另外，面对多样化的底层硬件，深度学习平台也需要加快GPU、CPU、ASIC、FPGA等多类型芯片的软硬适配，从而在自定义优化、统一硬件接口、自动化编译等方面提升模型训练速度。在模型部署方面，深度学习平台需要满足大模型的云边端部署推理需求，覆盖服务器端、移动端、边缘端、网页端等不同硬件场景的推理引擎，并提供模型压缩工具对大模型进行蒸馏、剪枝来适配端侧设备的性能要求和存储条件。同时，平台也需提供统一的API接口，助力开发者进行无代码调用部署。

04

大模型的评估与典型市场参与者

4.1 产业生态图谱 »

大模型生态涉及底层服务支持、算法平台以及行业应用，厂商主要包括百度、阿里、商汤、华为等人工智能企业，也有智源研究院、中科院自动化所等研究机构，同时英伟达等芯片厂商也纷纷入局。

图5 中国大模型生态

来源：IDC

大模型底层服务支撑基本完善，各厂商围绕核心算法与模型库、上层软件平台深入布局优化。在底层服务支撑层英伟达单卡芯片可完成百亿参数模型训练，将有效支撑大模型训练和应用推广。在基础算法平台层，科技巨头企业以及研究机构积极布局训练框架、模型库和工具平台，大幅降低大模型快速训练部署的算力依赖。最后行业应用层，大模型在搜索、对话、推荐等基础功能应用领域已建立行业标杆，企业更多需关注医疗、遥感、城市、基础科学、元宇宙等复杂场景。

国内的科技巨头都在预训练大模型领域投入研发力量。以百度为代表的自研全栈技术企业生态加速了我国技术普惠与产业赋能，是构建国产化生态体系的重要一环，为驱动千行百业大规模智能化升级、提升产业独立性和抗风险能力奠定基础。

4.2大模型评估框架及评估结果 »

4.2.1 大模型评估框架

为充分评估大模型技术能力、功能丰富度与底层深度学习平台开发能力，以及对各行业赋能的实际效果，并考虑到大模型的未来商业化前景，IDC搭建大模型评估框架V1.0，框架构成为“1-3-6-11”架构，即1个整体评估框架、3个评估维度、6个一级指标和11个二级指标：

分类	一级	二级
产品能力	模型能力	模型丰富度
		模型性能
	工具平台能力	功能丰富度
		平台成熟度
		易上手程度
	开放性	开放可体验的能力数
		对用户数据隐私保护、数据安全措施
应用能力	应用广度	已覆盖的行业数
	应用深度	客户业务流程关键环节渗透度
生态能力	应用生态	基于大模型进行产品开发的开发者数
		基于大模型工具与平台开发者创建的模型或应用数

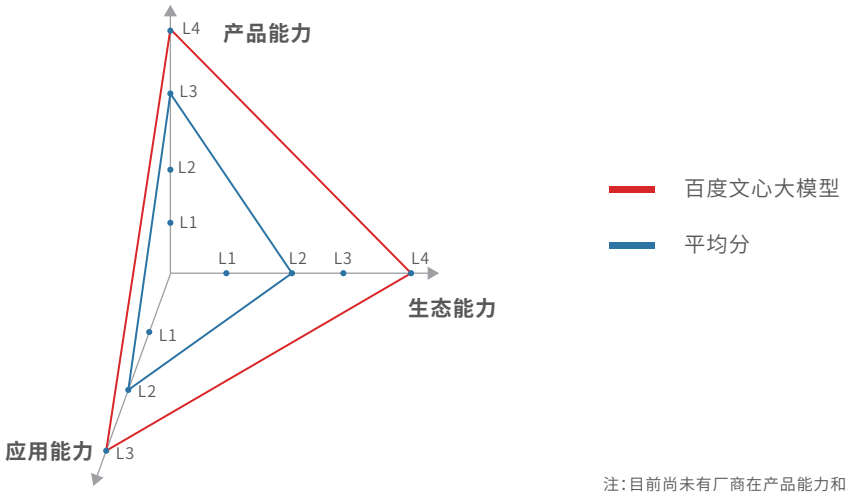
- **产品能力**主要考虑到大模型的技术能力和功能丰富度，以及底层深度学习平台的核心支撑能力，具体包括3个一级指标，分别是模型能力、工具平台能力和开放性。其中，模型能力包括模型丰富度和模型性能，工具平台能力包括功能丰富度、平台成熟度和易上手程度，开放性包括开发可体验的功能数和对用户隐私保护及数据安全措施。
- **应用能力**主要考虑到“大模型+深度学习平台”的实际应用广度和深度、商业化前景，具体包括2个一级指标，分别是应用广度和应用深度，其中应用广度为考察覆盖行业数，应用深度为考察客户业务流程关键环节渗透度。
- **生态能力**主要考虑到大模型市场生态布局情况，考察基于大模型进行产品开发的开发者数量、基于大模型工具与平台开发者创建的模型或应用数。

基于以上框架，IDC将对中国市场主流大模型厂商进行评估，明确在当前的行业局面下大模型的发展进程，帮助企业在开展相关业务时提供选型参考，并帮助平台厂商更好地制定竞争战略。

4.2.2 百度文心大模型评估结果

基于IDC搭建的大模型评估框架V1.0，并选取国内主流厂商（N=9），从模型能力、工具平台能力、开放性、应用广度、应用深度、应用生态共6大维度的11项指标，通过定性和定量两个方式进行打分评估，计算出各厂商在不同维度的得分和平均分情况。IDC发布中国大模型市场2022年百度文心大模型评估结果：

图6 中国大模型市场2022年评估结果—百度文心



来源：[1]IDC

注：目前尚未有厂商在产品能力和生态能力方面达到L5，应用能力尚未有厂商达到L4，因此不在本次评估中凸显

注：IDC中国将大模型市场服务划分为L1-L5五个层级，来衡量大模型产品能力、应用能力和生态能力，层级越高，厂商在大模型市场梯队越靠前，当前大部分厂商能力处于L2-L3层级。

评估结果显示，百度文心大模型在市场格局中处于第一梯队，产品能力、生态能力达到L4水平，应用能力达到L3水平。

具体来看：

百度文心大模型在**产品能力**呈现出较强技术实力和平台积累，“文心大模型+深度学习平台”创新了人工智能研发应用范式，满足市场大规模落地需求，达到行业前端水平；**应用能力**方面，百度已在金融、能源、制造、城市、传媒、互联网等行业拥有实际落地的标杆案例，截止目前文心已累计发布11个行业大模型，且积极布局，致力于解决用户实际痛点，并参与到客户业务流程关键环节中，其应用深度与广度方面在评估厂商中位列前沿；在**生态能力**方面，百度文心大模型在社区用户的基础上，可以实现与开发者、行业用户、上下游产业的正向互动，在评估厂商中处于行业领先位置。

面向未来，不论是百度还是其他大模型厂商，都应围绕整体平台化能力、行业赋能进行持续攻坚突破，解决大模型开发落地难、生态基础薄弱等问题，将大模型变成驱动人工智能产业进一步高速发展的元能力引擎。

4.3 百度文心大模型 »

百度率先在2019年3月发布预训练模型ERNIE1.0，持续投入大模型的技术创新与产业应用，布局了NLP、CV、跨模态等大模型，率先提出行业大模型，构建大模型工具与平台，探索产品与社区，在企业端和用户端均有不同程度的突破。基于以上背景，我们将百度作为典型的市场参与者进行着重梳理和分析。

图7 百度文心大模型全景图

来源：百度

百度凭借海量的知识积淀和丰富的应用场景推出的文心大模型，具备知识增强、产业级两大特色。百度自研的多源异构知识图谱，拥有超过5500亿条知识，被融入到文心大模型的预训练中。百度文心大模型同时从海量数据和大规模知识中融合学习，在知识的指导下，以语义单元为单位进行学习，效率更高、效果更好，可解释性更强。文心大模型已应用于百度搜索、信息流、智能驾驶、百度地图、小度等重要产品，服务数亿用户；在行业落地中，文心率先提出行业大模型概念，通过百度智能云在制造、能源、金融、城市、传媒等行业广泛应用；通过大模型工具平台、开源开放的模型与服务，已有近百万开发者使用文心大模型。

在近年的大模型技术探索与产业实践中，**百度文心形成了支撑大模型产业落地的关键路径，构建文心大模型层、工具平台层、产品与社区三层体系：**建设更适配场景需求的基础、任务、行业三层大模型体系，提供全流程支持应用落地的工具和方法，孵化基于大模型的任务系统与创新产品，营造激发创新的开放生态。

4.3.1 文心大模型的模型布局

文心「基础+任务+行业」三级模型体系：

文心大模型层，结合技术发展趋势、产业实践，构建基础、任务、行业三级模型体系。基础大模型聚焦技术方向的技术挑战、通用性、泛化性探索；任务大模型深入理解任务特性，构建预训练算法、训练数据集，打造紧贴任务的模型能力；行业大模型深度融合行业数据与知识特性，构建更适配行业的模型底座。基础大模型支撑任务与行业大模型的建设，任务和行业大模型结合真实场景与数据反哺基础大模型优化。

目前，文心大模型已经建设了36个大模型，其中基础大模型包含：NLP（自然语言处理）大模型、CV（计算机视觉）大模型、跨模态大模型，任务大模型包含对话、搜索、信息抽取、生物计算等多个典型任务，行业大模型包含与来自8个行业的头部企业或机构共建的11个行业大模型。

基础大模型：文心基础大模型覆盖了NLP、CV、跨模态三大方向

- **文心NLP大模型：**百度发布了文心ERNIE系列NLP大模型，ERNIE3.0基于知识增强的多范式统一预训练框架，深入融合的千亿级知识，具备强大的语言理解能力与小说、摘要、文案创意、歌词、诗歌等文学创作能力。其中与鹏城实验室合作发布了知识增强千亿大模型“鹏城-百度·文心“。目前文心ERNIE已经刷新93个中文NLP任务基准，并多次登顶SuperGLUE全球榜，已在机器阅读理

解、文本分类、语义相似度计算等60多项任务中实际应用。

- **文心CV大模型：**百度文心发布了VIMER系列的CV大模型，视觉自监督预训练大模型VIMER-CAE创新性地提出 “在隐含的编码表征空间完成掩码预测任务”的预训练框架，在图像分类、目标检测、语义分割等经典下游任务上刷新SOTA结果。在此之上，多任务学习模型VIMER-UFO 2.0可抽取轻量级小模型，兼顾大模型效果和小模型推理性能，单模型覆盖20多个CV基础任务，在28个公开测试集上效果刷新SOTA。端到端文档 OCR 表征学习预训练模型VIMER-StrucTexT 2.0解决了训练数据匮乏和传统 OCR + NLP 链路过长导致的模型表达能力不足、优化效率偏低等问题，能够广泛应用于各行各业行的文档、卡证、票据等图像文字识别和结构化理解。

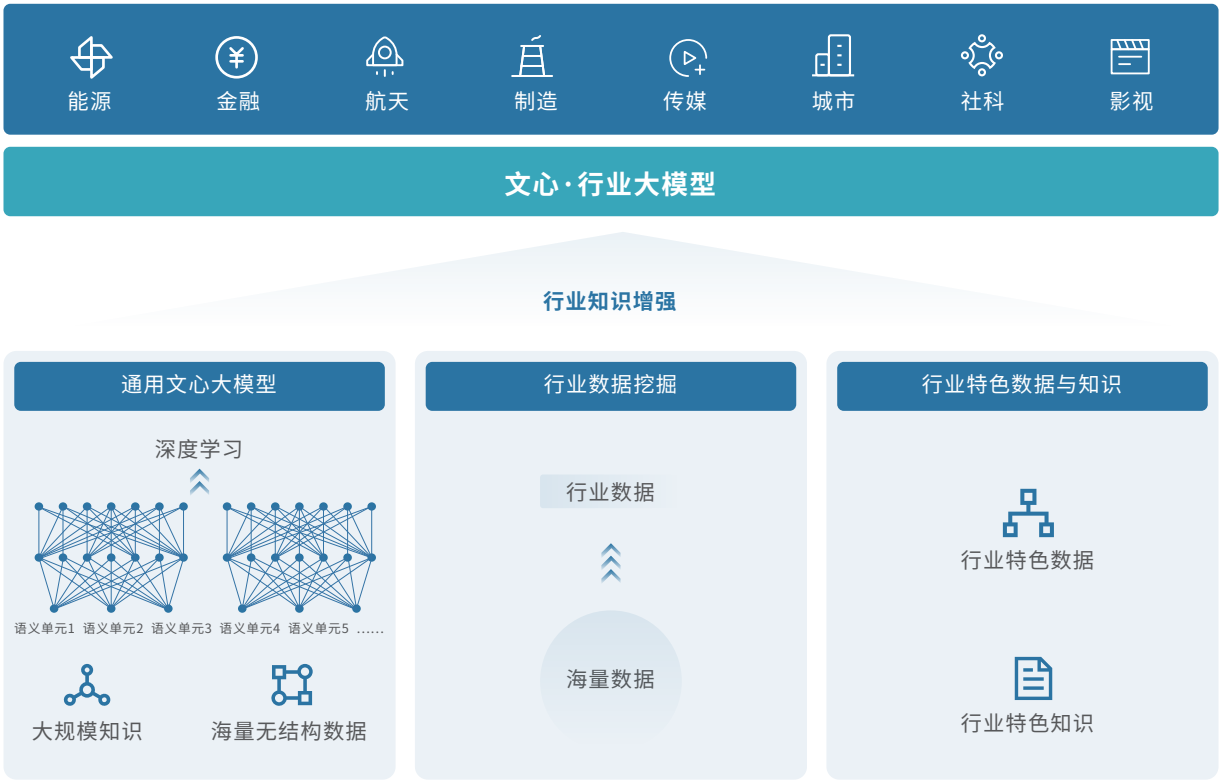
- **文心跨模态大模型：**文心跨模态大模型包括：ERNIE-ViLG2.0文生图大模型、ERNIE-ViL视觉-语言大模型、ERNIE-Layout文档智能大模型等。ERNIE-ViLG2.0是知识增强的 AI 作画大模型，在公开权威评测集MS-COCO上取得了当前该领域的领先效果，在语义可控性、图像清晰度、中国文化理解等方面均展现出了显著优势。跨模态文档智能大模型ERNIE-Layout，基于布局知识增强技术，融合文本、图像、布局等信息进行联合建模，在文档抽取、布局理解、表格理解、文档问答、网页问答等5类11项任务刷新业界SOTA。

任务大模型：

百度文心面向典型任务推出对话大模型PLATO、搜索大模型ERNIE-Search、信息抽取大模型ERNIE-UIE、代码生成大模型RENIE-Code、生物计算大模型等。对话大模型PLATO是基于隐变量的生成式开放域对话大模型，具备接近真人水平的多轮流畅对话能力，开放域对话效果达到世界领先水平。信息抽取ERNIE-UIE是专门基于 ERNIE 通用模型在开放域信息抽取领域进行优化的模型，利用单一模型支持多种类型的开放抽取任务，用户可以使用自然语言自定义抽取目标，无需训练即可抽取输入文本中的对应信息。ERNIE-Code基于海量代码和文本数据进行预训练，引入联合学习，具备跨多种自然语言和编程语言的语义理解和生成能力，已经在代码翻译、代码提取任务上取得不错的效果。文心生物计算大模型构建面向化合物分子、蛋白分子的生物计算领域预训练模型，赋能生物医药行业。HelixFold借鉴AlphaFold2的组合多轨模型结构，完整实现从蛋白序列-蛋白结构-蛋白功能的预测。HelixFold-Single是开源的基于单序列语言模型的蛋白质结构预测大模型，并在抗体结构预测场景下效果超越AlphaFold2。HelixGEM-2主要面向小分子药物研发，融合量子力学第一性原理，创新性地提出多轨机制，每个轨道对化合物不同阶的多体集合进行长程建模，在量子化学属性预测和虚拟筛选双场景上达到领先效果。

行业大模型：

图8 百度文心行业大模型



来源：百度

行业大模型是百度推进文心大模型深入产业落地的一项重要举措，是百度与行业头部企业、机构联合研发的融合行业数据、知识以及专家经验的大模型。目前百度文心在能源、金融、航天、制造、传媒、城市、社科以及影视等领域与国网、浦发、吉利、TCL、人民网、哈尔滨、上海辞书出版社等均有案例应用的行业大模型。这些行业大模型作为重要AI底座，在各行业的技术效果突破、产品创新、生产流程变革、降本增效等维度产生价值。

图9 百度文心行业大模型全景

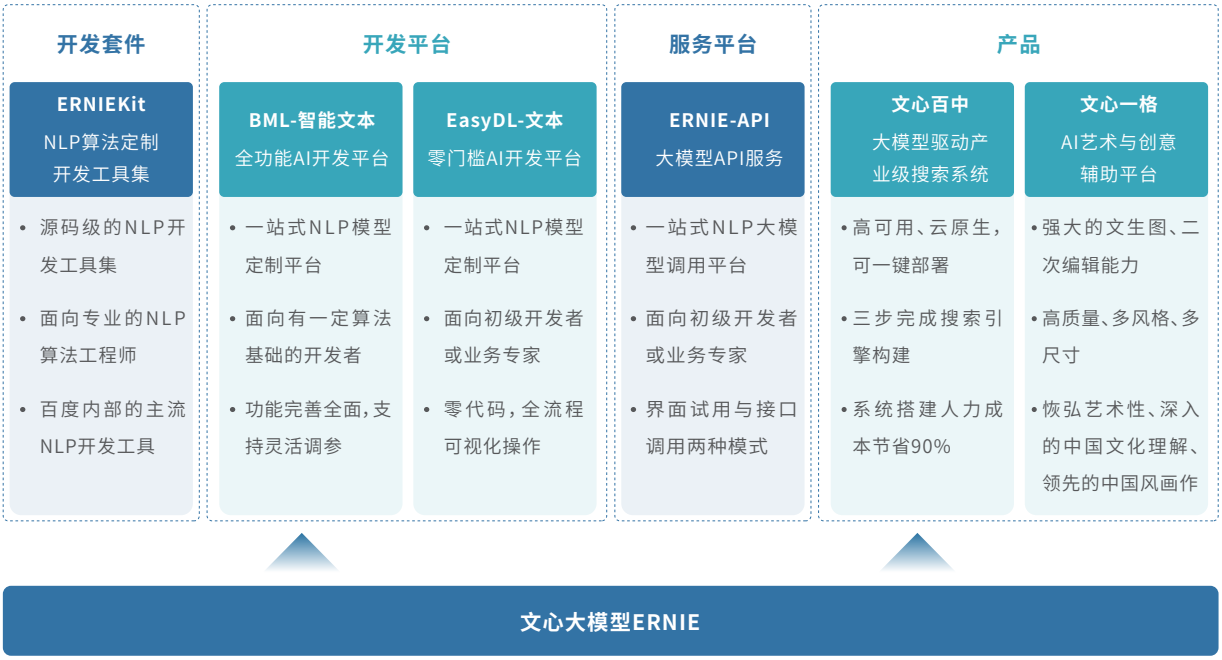
来源：百度

例如，百度与国网合作的NLP大模型，共同打造行业级人工智能基础设施，探索研发电力人工智能联合大模型，不仅提升了传统电力专用模型的精度，而且大幅降低了研发门槛，实现了算力、数据、技术等资源的统筹优化。百度与人民网的合作的NLP大模型，引入舆情数据中心积淀的行业知识来更好训练知识增强的传媒行业大模型，实现更少的标注数据下大幅提升传媒行业自然语言处理任务效果，如新闻内容审核分类、舆情分析、摘要生成等行业任务相对于通用模型提升显著。百度与TCL合作的CV大模型，面向多个产线多个环节的工业质检提供AI基座能力，在TCL几个产线检测mAP指标平均提升10%+，训练样本减少到原有训练样本30%～40%，产线指标即可达到原有产线效果，新产线冷启动效率可提升3倍，产线上线开发周期降低30%。

4.3.2文心大模型产品矩阵

百度文心围绕大模型产业应用的不同研发环节，面向各阶段不同技术、业务背景的开发者或用户，打造系列工具平台与场景化产品。

图10 文心大模型产品矩阵



来源：百度

- **大模型套件：**百度文心推出新一代预训练范式的NLP算法定制开发工具集ERNIEKit，面向NLP工程师，提供全流程大模型开发与部署工具集，端到端、全方位发挥大模型效能。包括数据标注与处理、大模型精调、大模型压缩、高性能部署、场景化工具五大模块能力。

- **AI开发平台：**百度AI开发以双平台模式驱动，面向应用开发者或业务专家提供零门槛AI开放平台EasyDL，面向AI算法开发者提供全功能AI开发平台BML。EasyDL使用百度文心NLP、CV、跨模态大模型作为训练基座，利用少量数据即可获得理想的模型效果，具有零门槛、高精度、低成本数据、超灵活部署四大核心优势。BML深度融合文心大模型，提供Notebook建模、预置模型调参、可视化建模、模型产线建模、Pipeline建模、实验管理等功能，兼具性能和性价比。

- **大模型API：**文心开放了NLP大模型ERNIE3.0、跨模态大模型ERNIE-ViLG、对话大模型PLATO。ERNIE 3.0 提供文案改写、开放问答、摘要、文案创作、小说创作、文本补全等文本理解与创作能力。ERNIE-ViLG提供基于文本描述的AI作画能力，图文相关性强、图片质量高，在中国文化理解、中国风、二次元等方面表现优异。PLATO提供生成式开放域对话服务，逻辑清晰、知识多元、情感丰富，闲聊能力接近真人水平。

- **场景化产品：**在搜索和文生图两个典型的应用场景上，百度文心推出基于大模型驱动的新一代产业级搜索系统文心百中，以及AI艺术与创意辅助平台文心一格。文心百中实现了系统极简化，通过搜索配置、数据导入、搜索体验三步完成搜索引擎构建，具备优秀的语义理解能力，构建一个搜索引擎可节省90%的人力，预置多个常用搜索场景。文心一格，让用户实现一语成画，只需输入一段自己的创想文字，并选择期望的画作风格，即可瞬间生成创意精美的画作；既能生产恢弘绚丽的艺术画，也能生产创意脑洞的超写实图，支持国风、动漫、插画、油画等十余种绘画风格和不同画幅的选择，让每个人都能展现个性化格调，享受艺术创作的乐趣。

4.3.3 文心大模型应用举例

文档智能场景下，赋能文档智能化识别、抽取、录入、审核，助力智能化办公

文档智能化在OCR和智能解析环节，面临格式繁多、布局形式多样、训练数据稀缺、业务定制成本高等问题。百度打造了文心ERNIE-Layout文档智能大模型，针对文档场景融合了文本、图像、布局等信息，引入了布局语义与视觉语义理解能力进行联合建模，能够对文档图片、PDF　文件、扫描件等多模态文档进行深度理解与分析，为各类上层应用提供多语言的模型底座，助力文档内容解析、语义理解、审核分析等全链路的智能化方案升级。

目前，基于文心大模型的智能文档分析平台TextMind，可提供包括文档信息抽取、文本内容审查、企业文档管理、文档格式解析、文档内容比对等全方位一站式的文档智能服务，已形成一套完整的企业文档场景化解决方案，满足银行、券商、法律、能源、传媒、通信、物流等不同行业和场景的文档处理需求。植根市场需求推出的合同智能处理解决方案，则可全流程赋能企业合同管理、法务信息服务，提高合同审查效率及准确性，助力企业办公的数字化转型和智能化升级。

人机对话场景下，赋能开放域拟人对话，助力交互创新

在人机对话场景下，主要面临对话逻辑差、知识准确性不高和缺乏长期记忆等挑战。百度提出了基于隐变量的生成式开放域对话大模型PLATO，同时结合知识内化和知识外用的全面知识增强策略：一方面，模型从海量公开网页与社交数据中学习，将大量的知识记忆到内部参数中；另一方面，模型进一步模仿人类对外部信息的查询和利用，学习在回复生成中融合外部知识。此外，PLATO还实现了在交互过程中实时识别、记忆和使用对话历史。文心对话大模型生成的对话回复逻辑清晰、知识多元、情感丰富，开放域多轮对话能力接近真人水平。

文心对话大模型已经广泛应用于百度搜索、信息流、智能音箱等互联网产品，累计服务超过10亿用户、

超过5亿个智能家居设备。同时通过百度智能对话平台及产品矩阵，广泛赋能通信、传媒、能源、汽车、金融等20多个行业，覆盖行业头部媒体、运营商、航空公司、车企、银行等客户，并催生了数字人客服、AI训练师等新业态新模式。在电话客服场景，百度建立了面向对话理解问题的专用预训练模型，该模型对数据标注量的需求比以往降低45%以上，支持头部电信运营商实现了覆盖全国的智能客服改造，显著减少了人工服务时长与用户等待时长。在数字人客服场景，百度联合头部商业银行发布了客服数字人，2021年累计服务客户超千万人次，销售额上百亿元。通过百度智能对话平台的公有云服务，文心对话大模型还广泛支持了近4万智能对话开发者，累计创建对话应用超过17万个。

无人驾驶场景下，赋能感知场景，提升感知智能

在无人驾驶场景下，面临大规模自动驾驶数据上云、大算力AI芯片性能突围、城市场景下通用自动驾驶产品服务的规模化等问题。过去十年间，百度在自动驾驶领域长期持续投入。目前，百度自动驾驶专利总申请量达到3477件，自动驾驶测试总里程超过4000万公里。背靠百度自研的AI芯片、文心大模型等，自动驾驶实现了全链条关键技术的自主创新。

百度自动驾驶依托文心大模型，从数据和感知模型的角度率先实现智能感知闭环迭代。基于数十亿图文对训练得到的文心-图文弱监督预训练大模型，百度已具备近百万类常用概念（物体、颜色、形状、动作、状态等）的泛化识别区分能力，实现了基于语义概念描述的低成本数据挖掘方案，从而大幅扩充了自动驾驶语义识别数据，如特殊车辆（消防车、救护车）识别、塑料袋误检等，使得自动驾驶长尾问题解决的效率实现了指数级提升。在感知模型上，基于10亿级参数规模的文心-自动驾驶感知模型，实现了感知大模型小型化闭环迭代，自动驾驶感知泛化能力显著增强，有效完成模型域适应迁移并解决远距离和小目标、长尾目标定位不准等问题。目前大模型已经成为自动驾驶感知能力提升的核心驱动力。

此外，随着自动驾驶走向规模化落地，高精度地图成为其发展的瓶颈，百度依托文心大模型将高精地图自动化生成能力大幅度提升至96%，解决了应用成本高的问题。百度提出的“高提纯、高消化”的数据闭环设计理念，全面强化自动驾驶的数据利用能力。据了解，该方案的数据提纯路径是利用车端小模型和云端大模型，实现高效率数据挖掘和自动化标注；数据消化架构实现自动化训练，具备联合优化和数据分布理解的能力，有效地利用高纯度数据进一步提升自动驾驶系统的整体智能水平。

工业质检场景下，助力降本增效、提升产能

在工业质检场景下，伴随工业4.0的到来，传统生产方式转型升级成为工业制造企业亟需思考的问题。传统的工业质检，以人工质检为主，面临着质检效率低的挑战。同时，工业质检产线往往工艺复杂，但对检测精度要求又非常高。

从行业现状出发，文心工业质检大模型，借助行业数据，突破了少样本和强标注限制，在多个工业质检

场景（3C、钢铁、纺织等）上，使得任务训练样本节省30%～40%，开发周期降低了30%，冷启动效率提升3倍，指标提升10%。文心工业质检大模型大幅优化了质检流程，提升了模型研发与运营效率，实现了成本的显著降低。基于文心大模型的工业质检场景方案，可应用于钢筋计数、安全帽智能识别、工业园区电力负荷检测、金属零部件质检、厂区吸烟检测、立体库智能盘点等具体场景，实现高精度性能识别，提升企业生产及交付效率。

AIGC场景下，赋能内容生产，助力数字经济

在AIGC场景下，主要有来自技术和商业两方面的挑战。技术上面临生成内容质量，技术伦理等问题；商业上面临创造力归属、创作伦理、知识产权等问题。百度利用知识增强的大模型基座赋能内容生产，在智能对话、小说等文本生成，文本到图片、视频等跨模态生成方面均有深入的产业应用。

百度推出ERNIE 3.0文本理解与创作API，覆盖写作文、摘要、文案、小说、对联等多项生成能力，每天服务上万用户；也已应用于百度自有产品的创新，比如在小、初学生“写作文”的场景解决学生 “无写作灵感”的痛点需求，形成产品的差异化优势。基于文生图大模型能力，百度开放AI作画API，同时推出AI艺术与创意辅助平台文心一格，用户只需输入一段自己的创想文字，并选择期望的画作风格，即可生成创意精美的画作，目前已服务上百万用户。在视频生成场景，百度以文心大模型技术作为基座，打造支撑通用型、大规模生产的智能视频合成平台，用户仅需一键输入新闻图文内容链接，就可以自动化完成视频制作，整套制作流程只需数分钟，日产智能视频达到万级别、日分发量达亿级别，同时结合知识问答、动态新闻、数字人、活动专题等推出差异化的视频自动生产能力。

4.3.4 文心大模型生态

预训练大模型市场正处于高速发展阶段，需要解决差异化水平下开发者和企业的应用需求。百度飞桨深度学习平台向下适配各种硬件，支持文心大模型的开发、高性能训练、模型压缩、服务部署的各种能力，贯通AI全产业链，串联起全栈化的产业生态体系。

文心大模型+飞桨深度学习平台生态共享，在市场生态方面持续发力

以百度飞桨为代表的国产开发框架已经逐步与产业融合，在社区生态建设上持续发力。飞桨（Paddle-Paddle）以百度多年的深度学习技术研究和业务应用为基础，集深度学习核心训练和推理框架、基础模型库、端到端开发套件、丰富的工具组件于一体，是国内功能丰富、开源开放的产业级深度学习平台。文心大模型是飞桨模型库的重要组成部分，与飞桨共享生态，包含产业级知识增强大模型体系，以及工具平台、API和创意社区助力大模型的高效应用。飞桨深度学习平台解决大模型研发和部署的各类问题，大模型使得AI模型的研发门槛更低、效果更好、流程更加标准化，硬件厂商、开发者以及模型应用企业在文心+飞桨生态中，紧密链接、相互促进，形成共聚、共研、共创的健康生态。

目前生态已凝聚535万开发者，服务20万家企事业单位，与12家硬件伙伴联合发布飞桨生态发行版，推动深度学习平台与更多硬件适配；还与国内科研院所、实验室以及高校强强联手，一同攻克AI技术难关，目前已赋能389所高校，服务747名教师，学分课培养10万余名AI学子 。

05

大模型未来发展趋势

5.1 大模型的发展是大势所趋 »

大模型未来将会助推数字经济，为智能化升级带来新范式：

- **大小模型协同进化，推动端侧化发展。**未来几年，大模型和小模型将会协同推动人工智能的发展，实现明确分工，高效率低成本地解决业务问题。大模型负责向小模型输出模型能力，小模型更精确地处理自己“擅长”的任务，再将应用中的数据与结果反哺给大模型，让大模型持续迭代更新，形成大小模型协同应用模式，达到降低能耗、提高整体模型精度的效果。大规模参数并不是产业所追求的重点，更少的标注数据依赖、更优的模型效果、更高的模型性能以及便捷的部署方式将是未来研究的重点。

- **大模型通用性持续加强，实现AI开发“大一统”模式。**大模型由于其泛化性、通用性，为人工智能带来了新机遇。通过无标注数据进行自监督学习，从而降低标注数据的人力要求，GPT-3、文心大模型均展示出了在未标记数据中的学习成果，并展示了在不同任务与行业领域上的通用性。同时，多模态大模型也逐渐兴起，数据形态差异化问题也将得到解决，未来大模型将进一步致力于构建通用的人工智能底层算法框架，融合多领域的模型能力，在不同场景中 “自我学习”，通过一个大模型解决产业中各种问题。目前，在通用模型的基础上，各行业正利用精调或prompt的方式加入任务间的差异化内容，从而极大地提高了模型的利用率，推动AI开发走向“统一”。

- **大模型从科研创新走向产业落地，通过开放的生态持续释放红利。**大模型最重要的优势，是推动AI进入大规模可复制的产业落地阶段，仅需零样本、小样本的学习就可以达到很好的效果，以此大大降低AI开发成本。目前，我们看到大模型已经开始与领域、行业深度融合，例如，工业质检、蛋白质结构预测等领域的大模型，验证了大模型已不仅在科技企业中应用，也迈出了走向各行各业的步伐。我们看到ChatGPT专注于强化内容创造能力，将生成式AI应用到实际业务中，为大模型带来新的产业落地机遇。具体来看，微软已经宣布将会全线整合ChatGPT，将大模型嵌入搜索引擎和办公软件，进一步推动AI能力的全面赋能和产业落地。

开放、开源是技术逐渐成熟和规模化输出的象征，随着大模型的落地，头部企业将开放技术，赋能中小企业，打造以大模型为底座的生态。目前大模型的开放、开源还主要在算法、API服务、开发工具的使用上，未来需要打造标准算法集、大模型平台、大模型数据集等全栈化的开放生态，将大模型的红利释放给每个开发者，并促进大模型创新应用的出现。

5.2对行业用户的建议 »

- **各行业技术买家都应该尽早拥抱大模型。**大模型是AI方向的必然趋势，所以行业用户应该尽早布局，将大模型纳入业务发展规划，利用大模型进行降本增效、产品革新。越早开始将大模型与自身行业、任务相结合，大模型与技术之间才有更多的时间进行打磨优化，在未来大模型深入各行各业的时候才能拥有话语权，走在行业的前端；同时在大模型和业务融合的过程中，行业用户也将沉淀更多经验，最终实现差异化优势。

- **在合作方面，主要关注大模型与自身业务的适配性。**在选型上行业用户应该关注开放性、兼容性以及可控性，并考察技术厂商大模型产品的技术安全性与稳定性。同时，技术路径与自身业务的契合度也应在考量范畴内，从整体关注技术厂商提供的技术产品是否与自身业务互通。此外，在技术部署的早期就关注产品的兼容互通性也是必要的。当前各类技术产品在技术路线选择上各有差异，即使在相同的技术路线下，也有不同的架构设计，行业用户如能尽早关注到互通能力，就能在未来缓解因产品不同而带来的业务阻碍或重复性工作。

- **与头部厂商联手打造行业标杆。**由于大模型本身的开发门槛高，训练需要海量数据并且要消耗大规模算力等原因，能够发布大模型的厂商均是行业头部。行业用户能够与头部厂商进行联合，不论是在品牌效应还是技术发展上都可能会看到“1+1>2”的效应，带动自身业务发展。但也应当选择适合自身业务的厂商进行合作，切勿盲目合作，可以更多关注厂商大模型技术栈完备性，产业应用经验积累，尤其是自己所在行业的产业应用经验。有了大模型支持，通过自身行业数据打造细分领域或针对性任务的精准模型，有助于成为行业标杆，在大模型成为主流的趋势下占据上风。除此之外，数据的融合也将反哺头部企业大模型精准度、效率的提升，实现双赢。

5.3对大模型供应商的建议 »

- **在技术方面，需要持续探究大模型的生成可控性。**大模型的生成能力有目共睹，目前大模型已经可以用于文生图、文章续写等，但准确度以及如何将用户的想法加入到生成能力中，是大模型的研究方向。通过prompt learning的方式让生成的过程接受使用者提供的条件是解决问题的途径之一，但截至目前仍不能完全达到预期效果。未来，如何建立一个统一可控的生成框架，生成内容的自我检测，以及生成内容的二次修改等内容，仍是值得探索的方向。目前的模型基本上是以文本和图像为主，3D模型少之又少，以数字人为例，最难攻克的便是数字人的面部，即使现实中有很多可参照的原型，但图片所带来的数据，无法支撑3D场景的需求。近期，Magic3D模型的出现打破了大模型在3D方面的僵局，但受到数据的影响，目前还无法将3D模型做成通用型的。IDC认为，未来视频生成一定是具有很大潜力和商业价值的领域，AIGC也将迎来爆发式增长。

- **在安全性方面，大模型的技术安全，以及伴随着大模型落地所带来的伦理问题仍是关注重点。**现在的大模型还处于发展初期阶段，关于安全伦理的业界讨论比较少。但是从以往的经验来看，大模型很容易被攻击，或被有意识地植入一些后门，从而令其在某些特定场景下做出特定响应，无法判断结果的准确性。同时在研究中我们也发现，模型变得越来越大之后，偏见问题也逐步暴露，对于算法的可解释性也越来越难，相关的问题所带来的对于模型的信任度亟待解决。

- **在商业化方面，大模型的路径仍不明确，海外市场发展较早，国内厂商可以重点借鉴。**大模型GPT-3目前在商业化方面已经有探索，应用包括文本合成、语音合成、代码Bug改写等，辅助日常工作的同时亦可满足用户娱乐需求，并提供详细的demo、文档和付费服务。近期，ChatGPT与云计算、搜索引擎等全线产品进行整合，提供了更多商业化的可能性，为国内大模型发展提供了参考范例。未来大模型可以通过开放、开源等的方式将技术产品以及服务开放，构建生态社区，通过调用计数等方式进行大模型的商业化，利用当下的技术和数据积累，产出丰富的服务和价值，继而获取更多的需求和数据，形成良性循环。其次，还可以将大模型作为产业智能化基座，对外进行售卖，企业在基座的基础上直接做上层开发，实现降本增效。面向大企业，也可以支持为其训练自己的大模型，打造更贴合自己企业的智能化基座。大模型本身具有明显的商业化价值，不论是其具有良好的泛化性、可复制性，还是其对于行业的赋能，都显现了人工智能驱动新一轮科技革命和产业变革的巨大力量。厂商应提前应对，占据有利地位，形成和客户的稳定合作，才是长久布局的良策。

关于IDC

国际数据公司（IDC）是全球著名的信息技术、电信行业和消费科技咨询、顾问和活动服务专业提供商。成立于1964年，IDC在全球拥有超过1100名分析师，为110多个国家的技术和行业发展机遇提供全球化、区域化和本地化的专业视角及服务。IDC的分析和洞察助力IT专业人士、业务主管和投资机构制定基于事实的技术决策，以实现关键业务目标。IDC于1982年正式在中国设立分支机构，是最早进入中国市场的全球著名的科技市场研究机构。在中国，IDC分析师专注于本地ICT市场研究，与本地市场 结合度非常高，研究领域覆盖硬件、软件、服务、互联网、各类新兴技术以及企业数字化转型等方面。欲了解更多信息，请登录 www.idc.com。

IDC China

IDC中国（北京）：中国北京市东城区北三环东路36号环球贸易中心E座901室

邮编：100013

+86.10.5889.1666

Twitter: @IDC

idc-community.com

www.idc.com