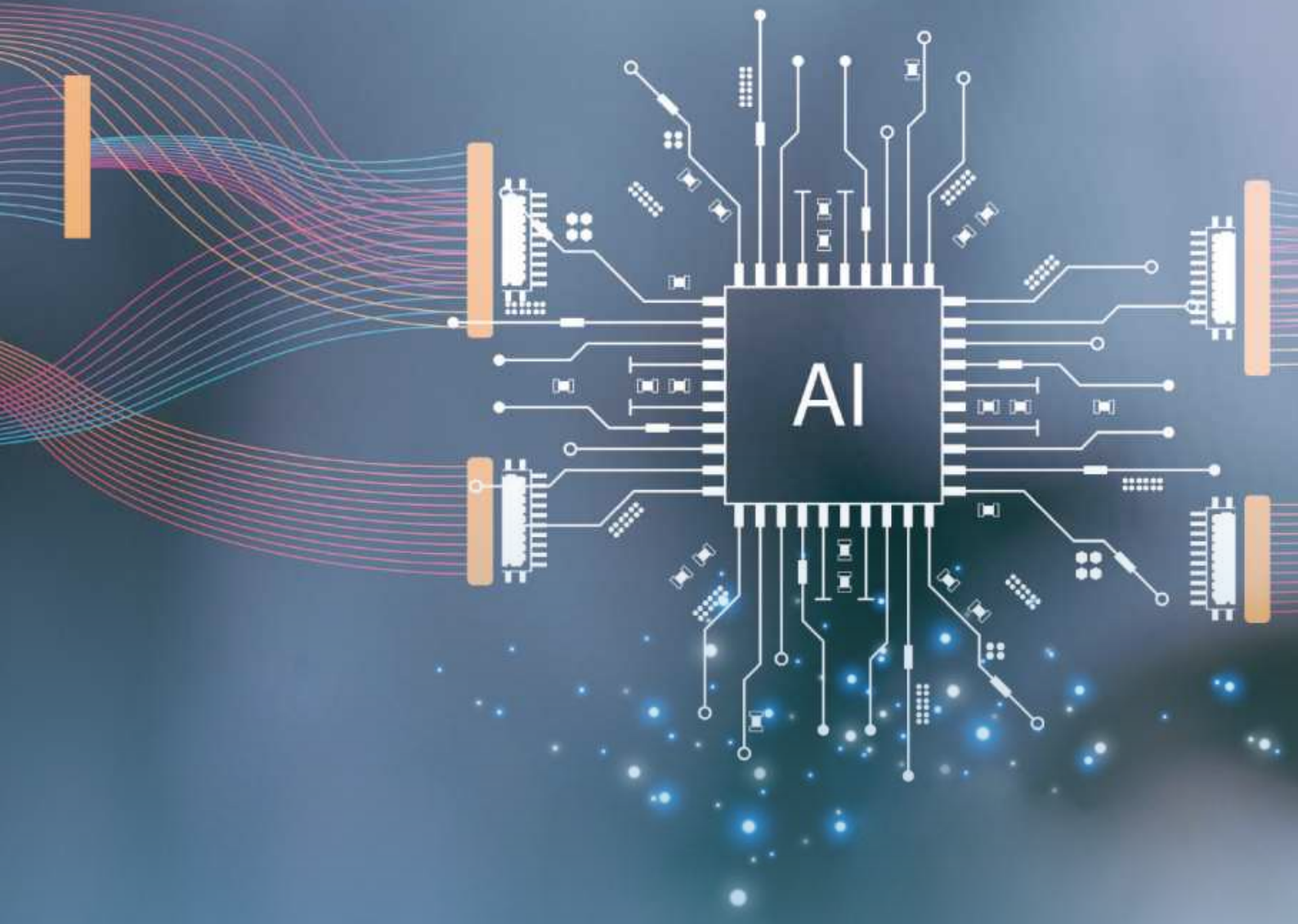
IDC

赞助商：NVIDIA



技术革新引领未来——生成式 AI 塑造核心发展引擎

人工智能产业链联盟

星主： AI产业链盟主

知识星球

微信扫描预览星球详情

目录 Table of Contents

IDC 观点

观点一：技术协同发展推动生态完善

在当今快速演变的技术生态系统中，多技术协同升级已成为推动新兴技术发展的核心动力。这一过程涉及人工智能（AI）、大数据、云计算等关键技术的深度融合，也关系到各个行业之间的相互渗透，技术和行业互相交织形成了一个创新生态。例如，金融场景已可以将产品与大模型进行结合、生成交易数据，从而弥补真实数据的不足，并优化欺诈识别模型的训练。时至今日，行业融合多种新技术的成功案例层出不穷，显示出协同效应对技术创新周期的加速作用。

观点二：数据持续积累推动新的处理范式

IDC 将 AI，尤其是生成式 AI，视为下一个重大变革性和有影响力的技术转变。我们正在进入一个 AI 无处不在的时代。此次变革与过往计算机革命及云计算转型相比，不同之处在于其驱动力为数据而非硬件设备；这也标志着智能化进程首次以数据为核心导向的重大飞跃。而生成式 AI 的核心价值之一在于它解锁了非结构化数据中的价值。数十年来，从非结构化数据中提炼有用信息一直是一项艰巨挑战。据 IDC 统计，2023 年非结构化数据将占所有存储数据的 77%[1]，这表示一个数据密集型创新周期已经来临。在此背景下，能够高效处理、解析并转化这些非结构化数据为可行性洞察的 AI 技术，将成为推动各行各业转型升级的关键力量。

1: Worldwide Global DataSphere and Global StorageSphere Structured and Unstructured Data Forecast, 2023–2027

观点三：算力是产业数字化转型的重要推动力

IDC 预测，到 2027 年，70% 的经济价值将以高信息密度的商品和服务形式呈现；为保持经济增长速度，强大的算力是信息产品发展的支撑基础[2]。数据分析、大模型训练以及推理等关键步骤都需要算力作为底层资源。而算力配置是否有弹性并且可扩展，直接影响到企业业务的运行稳定性以及响应市场变化的速度。因此，构建并优化算力体系，成为企业把握大模型时代发展机遇和提升核心竞争力的战略重点。

观点四：大模型技术发展关注点逐渐转向高效、经济和环保方向

在生成式 AI 的演进中，高效性、经济性和环境友好性的考量日益受到重视。大模型预训练对计算资源的需求极高，相关能耗问题也逐渐凸显。如何在不牺牲模型精度与训练效率的前提下，通过优化计算架构、节能技术等方法，实现降低运行成本、减少能源消耗，并最终达到低碳环保的长远目标，确保大模型技术的可持续发展，以更优地促进社会经济向绿色转型并迈入高质量发展阶段，成为当前研究与实践的核心议题。

2: IDC FutureScape: Worldwide Data and Content Technologies 2023 Predictions

第一章 生成式 AI：推动科技进步与产业变革的强大驱动力

1.1 生成式 AI 引领产业智能化落地，开启经济发展新篇章

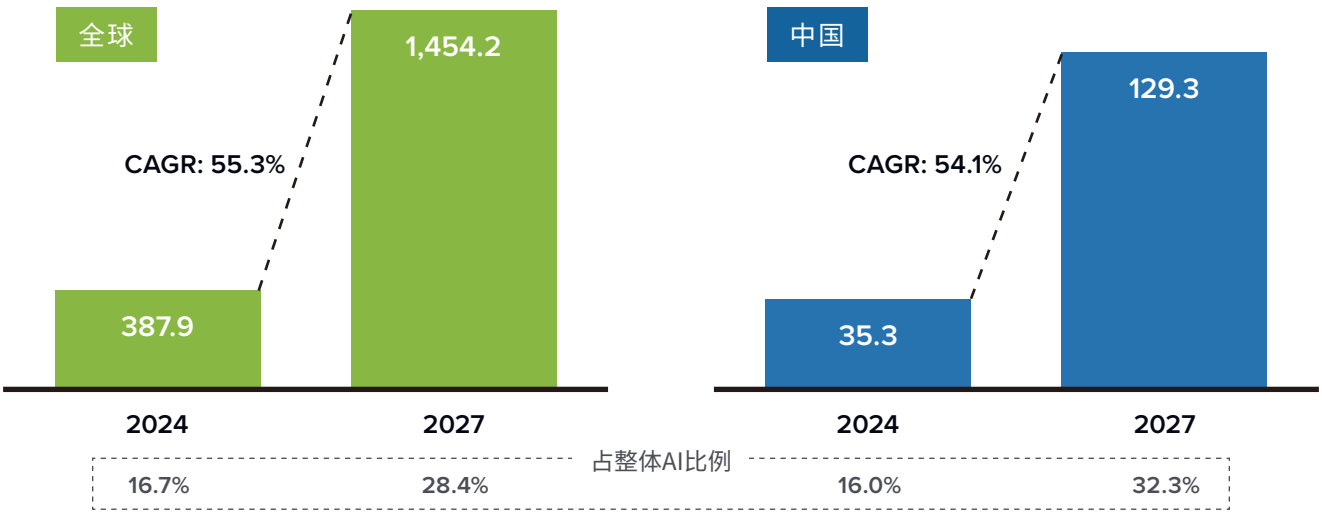
随着 AI 技术的推进，生成式 AI 已跃升为数字时代的前沿领域。从最初的基于规则的简单创作，发展至今日由深度学习驱动的创造性产出，生成式 AI 技术实现了由量变到质变的深刻转型。这一历程，包含了计算能力的几何级跃升，数据资源的持续累积，以及机器学习、深度学习算法的不断精炼与革新。尤其在近十年间，生成式对抗网络（GANs）与 Transformer 模型的诞生，为文本、图像乃至视频内容的自动生成开辟了创新级可能性，极大地拓展了创意表达的边界。

IDC 预测到 2027 年，全球生成式 AI 市场规模将攀升至 1454 亿美元，中国市场的投资亦将达到 129 亿美元；这一发展趋势的动力源自技术迭代的加速、

应用领域的拓宽，以及企业对 AI 创新驱动的不懈投入。除了大模型 AI 厂商外，NVIDIA 作为加速计算技术的领航者，在此进程中也发挥着核心作用，NVIDIA AI Enterprise 平台通过加速计算能力、优化的软件栈和容器化服务，降低了企业部署和运用复杂 AI 模型的门槛，加速了从研究到生产的转化过程。值得注意的是，该平台能够支持训练千亿乃至万亿参数量级的大模型，给生成式 AI 技术落地带来可能性。

图 1

生成式 AI 支出规模（单位：亿美元）

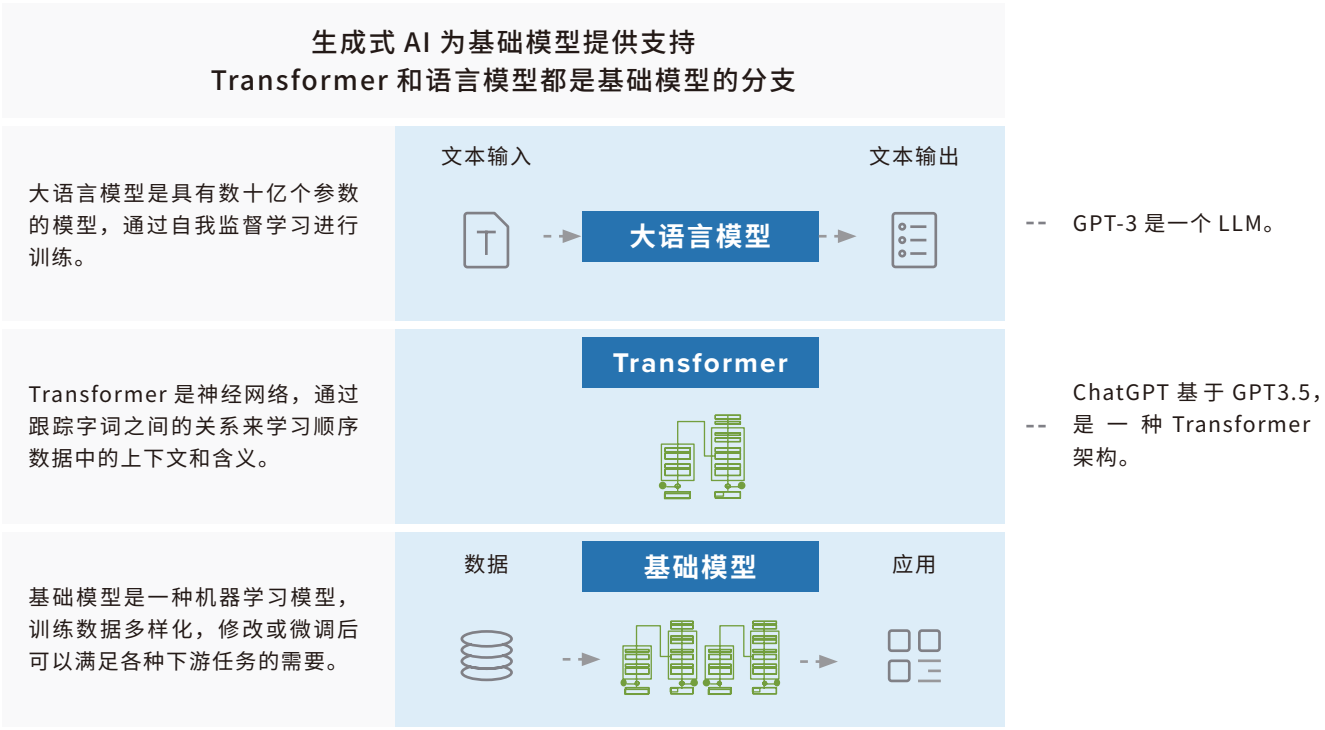


来源：IDC, 2024

生成式 AI 技术的商业化与技术进步并驾齐驱，不仅在国际舞台上催生了一系列科技创新，也见证了本土企业的迅速崛起与差异化战略的实施。大模型目前主要分为文本、图像以及视频三种模态，在不同的数据和场景中发挥作用。

图 2

Transformer 架构革新生成式 AI ，开启自然语言处理的新篇章



来源：IDC, 2024

文本对话技术的迭代升级，率先为大模型开辟了应用前景：文本生成技术的飞速发展，归功于文本数据资源的多样化和易获取；这些数据在互联网的每一个角落，包括但不限于社交媒体、新闻文章、学术论文、历史档案等，其多样性、广度与深度为模型提供了丰富的学习材料。另外，Transformer 架构的问世，成功解决了循环神经网络（RNN）在处理长序列信息时的局限性。ChatGPT 作为语言生成领域的先锋，凭借其卓越的对话创造、代码生成及跨领域知识解析能力，彰显了生成式 AI 技术的高水平成熟度及广泛的应用潜力。GPT-3.5 模型拥有 1750 亿参数及先进的自注意力机制，采用多层 Transformer 解码器堆叠架构，使模型具备了上下文感知的对话、代码合成及跨学科知识解析能力。在国内，ChatGLM 与 Baichuan 等大模型亦展现出色表现。ChatGLM 采用的双流自注意力机制增强了对复杂语言结构的解析力，其灵活性和较低的

资源消耗，特别是通过模型量化技术实现的 ChatGLM-6B 模型在边缘端的低门槛部署，极大地推动了高级语言模型的普及。Baichuan 则整合了意图理解、信息检索、强化学习等关键技术，并借助有监督微调与人类意图对齐策略，在知识问答、文本创作等多领域取得了卓越成效。

图像生成技术的革新，进一步拓展了大模型的创意边界：图像创作需要融合计算机视觉与深度学习技术。在早期发展的过程中，生成新图像在真实度与细节还原度上存在一定的局限性，导致图片失真；而新一代技术则凭借大量的训练数据集和复杂的算法架构设计，使生成图像的真实性显著提升，Stable Diffusion 和 DALL-E 2 是图像模型的代表。从技术方面来看，Stable Diffusion 利用扩散模型架构，从随机噪声中解析出清晰图像，其核心优势在于其可以在低计算资源的基础上保持生成高分辨率图像；同时，其开源特性更是激发了社区用户的积极性，形成模型从使用到迭代的正向循环。DALL-E 2 则是运用 Transformer 架构实现的文本到图像的直接映射，通过多模态数据的预训练，使模型能够推理出不同的图像特点，从而有效转化文本中的抽象概念和细节，并通过分层构建图像的方式确保生成内容的结构合理性和细节饱满度。

视频创作技术的飞跃，补全了大模型在动态场景中的不足：视频生成技术的发展得益于多模态技术升级已取得的重要进展，从最初的动画合成到处理复杂动态场景和非线性叙事结构。在技术快速迭代的背景下，以 VideoGAN 和 Sora 为代表的视频生成模型，极大提升了视频创作的效率。VideoGAN 利用深度学习技术可生成连贯的视频片段，通过时间相关损失函数和循环一致性约束确保帧间连贯，结合时空注意力机制和 LSTM 等复杂网络结构，以捕捉和保留视频序列的时空特征，实现视频的自然流畅。Sora 凭借其时空一致性与动态适应性脱颖而出，其集成的 LSTM 与 3D CNNs 协同工作，确保视频序列在时间维度上的平滑过渡和逻辑连贯，同时引入条件生成机制，赋予用户高度定制化和交互式的视频创作体验，进一步模糊了现实与虚拟的界限，开创了内容创作的新境界。

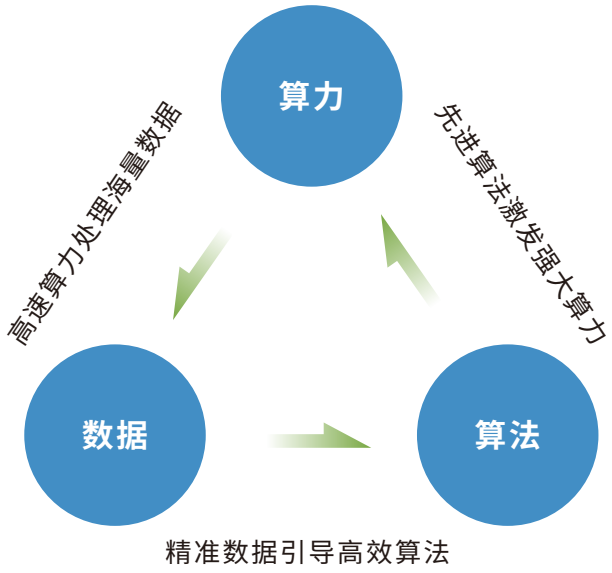
技术层面上，生成式 AI 正不断向高精度、低延迟和多模态方向发展。模型架构上依然是以 Transformer 及其变体为主，但目前也在探索更高效的注意力机制和模型压缩技术等方式，以降低模型的计算成本并提高部署效率。另外，针对特定领域的细粒度优化也是新的技术热点，如医疗、金融等行业模型。厂商方面，包括 NVIDIA、谷歌、微软在内的国际企业，以及国内的阿里、百度等公司，都在布局通过提供高性能计算资源及上层生态来共同支撑生成式 AI 的持续发展。随着技术的不断迭代与应用场景的不断开拓，生成式 AI 推动的智能化转型正稳步前行，其带来的效益将会逐渐体现在社会经济的各个层面。

1.2 硬件迭代、算法突破、数据改善共促生成式 AI 发展

在生成式 AI 的快速演进中，算力的强化、算法的演进以及数据的积累是三大核心要素，在新技术浪潮中共同发挥作用，持续拓展新技术边界，并将生成式 AI 推向新的高度，确保其在多种应用场景中表现更卓越。

图 3

生成式 AI 发展三大要素



来源：IDC, 2024

算力支撑：硬件革新与技术协同发展

算力是生成式 AI 发展的物理基础，高性能计算硬件的持续进步为模型训练提供了强大的支撑。GPU 因其高度并行的计算能力，成为训练大模型的理想选择。近年来，专门针对 AI 计算优化的 TPU（Tensor Processing Unit）、

ASIC（Application-Specific Integrated Circuit）等加速器的出现，也提升了计算效率，降低了能耗。这些硬件创新，结合高速互连技术，如 NVLink、InfiniBand 等，为大规模并行计算提供了必要的基础设施。

除了硬件基础设施外，多集群并行计算技术如 Horovod 和 PyTorch Distributed 等框架是协调硬件资源的关键软件组件。这些框架在通信协议上做了一定优化，以获得高效的模型参数同步与负载均衡，从而有效地解决了多 GPU 的协同问题。具体来说，Horovod 在 Ring-AllReduce 算法下减少了模型更新的通信时间；而 PyTorch Distributed 提供了灵活的分布式训练，在支持多种并行模式的情况下，使训练过程得到明显的加速，同时降低了资源消耗。此外，高效率的数据传输与同步，随着模型规模的扩大变得格外重要。远程直接内存访问（RDMA）技术与高速网络通过减少数据传输的复制步骤并缩短延迟，确保大规模集群间数据的高效交换，增强模型训练的稳定性和效率。这些技术与智能的数据放置策略相配合，使大规模并行计算的效率进一步优化。

算法层面：深度与广度的双重跃升

Transformer 架构带来了一场自然语言处理领域的革命，该架构通过引入自注意力机制显著提升了对长序列数据的理解和生成能力，它能够使模型并行考虑输入序列的所有位置，彻底改变了传统的序列数据处理方法。

生成式 AI 技术迭代的核心推手是算法创新。Transformer 架构带来了一场自然语言处理领域的革命，该架构通过引入自注意力机制显著提升了对长序列数据的理解和生成能力，使模型并行考虑输入序列的所有位置，彻底改变了传统的序列数据处理方法。自注意力机制的精髓在于，它能够赋予模型学习输入序列中任意两部分之间关系的能力，这种全局视角对于理解和生成自然语言至关重要，因此基于 Transformer 的 BERT、GPT 系列等迅速成为主流的自然语言处理模型。此外，我们看到，模型规模也在随着算法的不断演进而迎来增长，除了得益于 Transformer 架构高效的并行处理能力外，分布式训练技术的成熟也不可或缺，如模型并行、数据并行和混合并行等这些技术在大规模模型训练中有效地解决了内存限制和通讯瓶颈等问题。

在面对大模型的训练推理效率方面的挑战时，Mixture of Experts（MOE）架构被提出。该架构通过将模型分解为多个专注于处理输入数据特定领域的专家子网络，并采用门控机制来挑选最适合的专家执行任务，实现了计算资源的动态优化配置与高效利用。这一设计不仅增强了模型处理复杂任务的能力，也为处理极为庞大的数据集开辟了道路，同时还确保了模型的可扩展性与灵活性，是大模型设计的一个重要发展趋势。

数据优化：数据量与多样性并重

数据质量的高低是生成式 AI 模型精确性和泛化能力的根本所在，因此多数企业目前正致力于数据治理流程的优化，通过采用存算一体架构及数据湖解决方案来提高数据的存储和处理能力。存算一体架构通过紧耦合设计减少了数据移动能耗与延迟，显著提高了能效比和处理速率，降低了数据传输中的损耗。而数据湖解决方案则为企业提供了一个集中管理平台，该平台能够支持结构化、半结构化以及非结构化数据的高效存储与分析，为模型训练提供了丰富多样的数据源。

硬件迭代、算法突破与数据改善构成了生成式 AI 发展的铁三角，三者相互促进，不仅共同推动着新技术的快速前行，也催化了从理论到实践的跨越。

在数据模态方面，IDC 调研显示，生成式 AI 创建的数据中有 36% 是文本，远高于其他数据类型，但是到 2028 年，生成式 AI 创建的 75% 的数据将均匀分布在文本、图像和视频之间，其余为代码、音频和科学数据，形成多模态数据的局面[3]。因此，多模态数据的融合分析是未来发展的重点，即通过结合多种模态的数据，使模型对环境中的隐含信息进行更准确的捕捉与分析，从而提高模型对复杂场景的理解能力。在多模态数据集的基础上，还可以通过数据增强技术，如图像旋转、平移、添加噪声以及文字资料的同义替换、句子结构调整等，使训练数据的丰富程度得到进一步的提升，从而增强模型的鲁棒性和泛化能力，为多元化和复杂的 AI 应用场景奠定基础。硬件迭代、算法突破与数据改善构成了生成式 AI 发展的铁三角，三者相互促进，不仅共同推动着新技术的快速前行，也催化了从理论到实践的跨越。

3: The Potential Impact of Generative AI on the Global DataSphere and Global StorageSphere

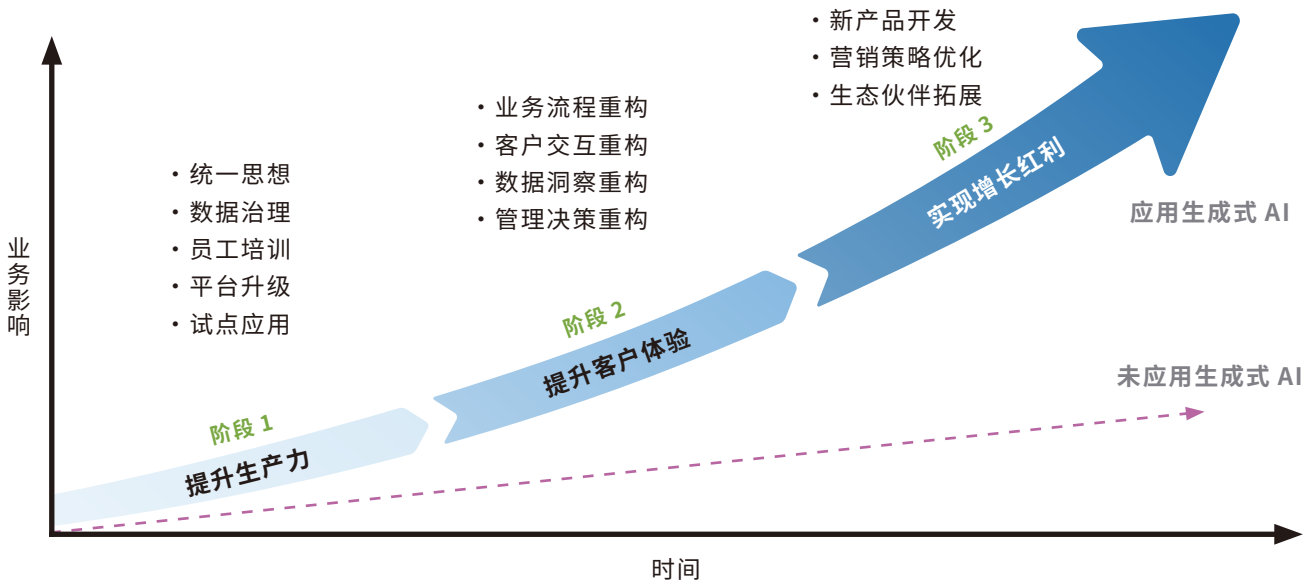
第二章 驾驭生成式 AI：企业智能化转型的实施流程与核心影响

2.1 生成式 AI 在企业端的应用：明确目标，实现价值

生成式 AI 的应用逐渐深入到企业运营的各个流程，即从基础设施的完善到业务流程的自动优化，再到内容生成的价值创造。企业在开始阶段通常是通过数据治理、技能培训等手段，为 AI 的应用打下坚实的基础；中期则在业务流程中嵌入 AI 技术，在客户体验和经营效率上实现改善；最后进入创新驱动阶段，生成式 AI 将会成为企业创新的催化剂，不仅在产品研发、市场策略等方面能够加快步伐，更在安全合规、生态拓展等核心领域构筑防线，使企业的核心竞争力在各个方面得到全面的提升。

图 4

生成式 AI 在不同发展阶段的核心价值与准备工作



来源：IDC, 2024

基础构建阶段：数据治理与技术融合，激发内部效率

在生成式 AI 融入企业的初期，应抓好技术在企业中融合的基础工作，对企业已有 IT 架构进行评估和优化，使之能够有效支撑 AI 系统的高效运转。数据治理是这一阶段的核心工作，企业需要通过建立完善数据预处理以及数据收集、清洗、存储和分析体系来保证数据质量，从而夯实 AI 模型训练或推理的基础。据 IDC 调研数据显示，95% 的受访用户预期生成式 AI 将促使组织存储更多的数据，直接导致数据容量需求激增 [4]。2023 年，IDC Global StorageSphere 数据显示，全球数据中心已累积存储超过 3800EB 的数据量，其中约 67% 的数据被保存在云端数据中心 [5]。鉴于此趋势，技术供应商应持续评估生成式 AI 对其产品矩阵的影响，以应对 AI 应用

4: The Potential Impact of Generative AI on the Global DataSphere and Global StorageSphere
5: Worldwide Global DataSphere and Global StorageSphere Structured and Unstructured Data Forecast, 2023–2027

的日益普及所带来的数据量级的快速增长。同时，企业内部的准备工作也不容忽视，包括对员工进行 AI 相关知识与技能的培训，为企业的全面智能化转型预先布局，确保技术升级与人员能力提升并驾齐驱，共同推动企业向智能化时代稳步迈进。

体验重塑阶段：用户为中心，深化个性化体验

中期阶段，由于基础架构的逐渐稳定，企业开始将目光转向用户体验的提升上。在这一阶段，生成式 AI 将从单点试用逐渐转向广泛采用，旨在打造无缝的、个性化的用户产品。以 Chatbot 为例，它可以为客户提供 7*24 小时的服务，通过生成式 AI 技术的整合，智能客服不仅像以前一样可以回答常规预设的问题，还可以提供基于用户对话历史、行为习惯甚至是情绪感知的动态响应，为用户提供个性化的互动体验。此外，在中期阶段，生成式 AI 对于企业业务流程、数据洞察分析和管理决策的重构也开始产生不同程度的影响。对于企业内部来说，生成式 AI 不局限于创作工作，更是深度融入到业务逻辑中，利用先进的机器学习算法对海量数据进行深度挖掘与识别分析，达到精准模拟复杂系统行为的效果，从而动态调整业务流程或产品供应链，达到智能自动化的目的。

创新驱动阶段：技术前沿探索，塑造未来竞争力

当企业内部流程和产品都不同程度地融入了生成式 AI 后，企业的智能化转型也步入深水区。在这一阶段，企业将会积极投入到多模态生成、强化学习等前沿技术的应用探索中，生成式 AI 的应用和落地将显著提高企业竞争力。在产品设计与研发领域，生成式 AI 辅助工具能够明显加速从概念到产品的转化过程，通过产品设计原型的快速迭代，及时响应市场变化与消费者需求，缩短产品上市周期。同时，在深度融合生成式 AI 后，企业通过模拟市场情景与预测分析，为管理层提供科学决策依据，在降低创新风险的情况下，提高管理层决策的科学性和有效性。整体来说，进入第三阶段，企业产品不论是从概念到营销，从研发到销售，还是从效率到效果，方方面面都将会有显著的提升。除此

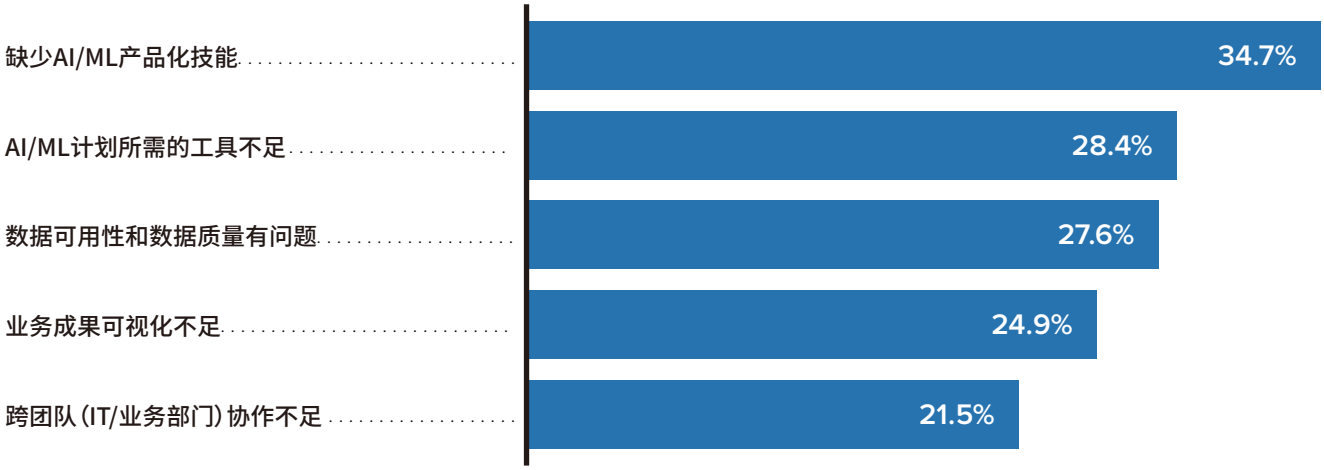
之外，生成式 AI 还有机会成为企业间跨界合作的桥梁，企业可以通过 API 开放与平台共享，联合产业内外的相关方来共同挖掘 AI 技术在行业领域的应用潜力，实现生成式 AI 在塑造未来商业格局中的无限可能。

2.2 企业需求在新技术时代下的演变：迎接挑战，拥抱变化

新时代技术浪潮促使生成式 AI 成为企业战略转型的核心，深刻影响组织结构、战略、文化及风险管理等多方面。企业需求逐渐向运营效能与人力资本优化、用户体验极致化、商业模式重塑与价值创新等方面转变。然而，在拥抱这一技术革新的旅程中，企业也不可避免地将应对多种挑战：

图 5

企业实施 AI/ML 计划面临的主要挑战



来源：IDC, 2024

AI/ML 技术有效产品化的难题

在将 AI/ML 技术转化为可市场化的产品时，企业经常面临技术与市场需求脱节、产品化路径不明确等问题。这要求决策者不仅要具备深厚的行业知识，还需有

预见技术趋势和产品创新点的能力，以便将抽象的技术概念转化为实际应用价值。此外，高昂的研发成本、长周期的产品迭代，以及如何平衡技术先进性与用户接受度也是主要障碍。

缺乏高效支持 AI 开发和部署的工具链

为了完善涵盖数据准备、模型训练、测试、部署和监控的完整工具链，企业在扩展 AI 项目时往往面临着重大挑战。此时，整合 RAPIDS（可在 GPU 上加速数据科学工作流的开源库）变得至关重要。这不仅需要一个强大且集中的技术栈，并要求将 RAPIDS 无缝融入其中，同时还需要高效的团队协作和持续的维护能力，以确保其符合严格的数据安全协议。这些全面的需求往往超出了单个团队或部门的能力范围。

数据孤岛与处理低效

数据是 AI 项目的命脉，但数据往往分布在不同系统中，存在清洗困难、缺少连续性、噪声大等问题，严重影响模型的训练效果和最终的业务应用。企业需要投入大量资源进行整合和治理，需具备对数据的高度敏感性，同时也需要大量场景数据作为底层支持。

业务成果验证困难

尽管 AI 技术有潜力带来显著的业务价值，但其成效往往难以准确预测和衡量。如何设定合理的 KPI、计算技术投资回报率，以及如何建立直接关联 AI 干预与业务增长的因果关系模型，是企业需要直面的挑战；尤其在技术发展初期，很难直观评估技术投入所带来的直接收益，需要更长远的战略规划。

跨部门协同与知识共享障碍

信息孤岛和部门间的沟通壁垒通常是阻碍技术在企业中快速应用和创新拓展的主要问题，在组织架构复杂的大型企业中更是如此。同时，在知识体系环节，目前大部分厂商缺乏有效的知识管理系统和协作工具，使得内部经验和学习成果难以在组织内部进行传承。

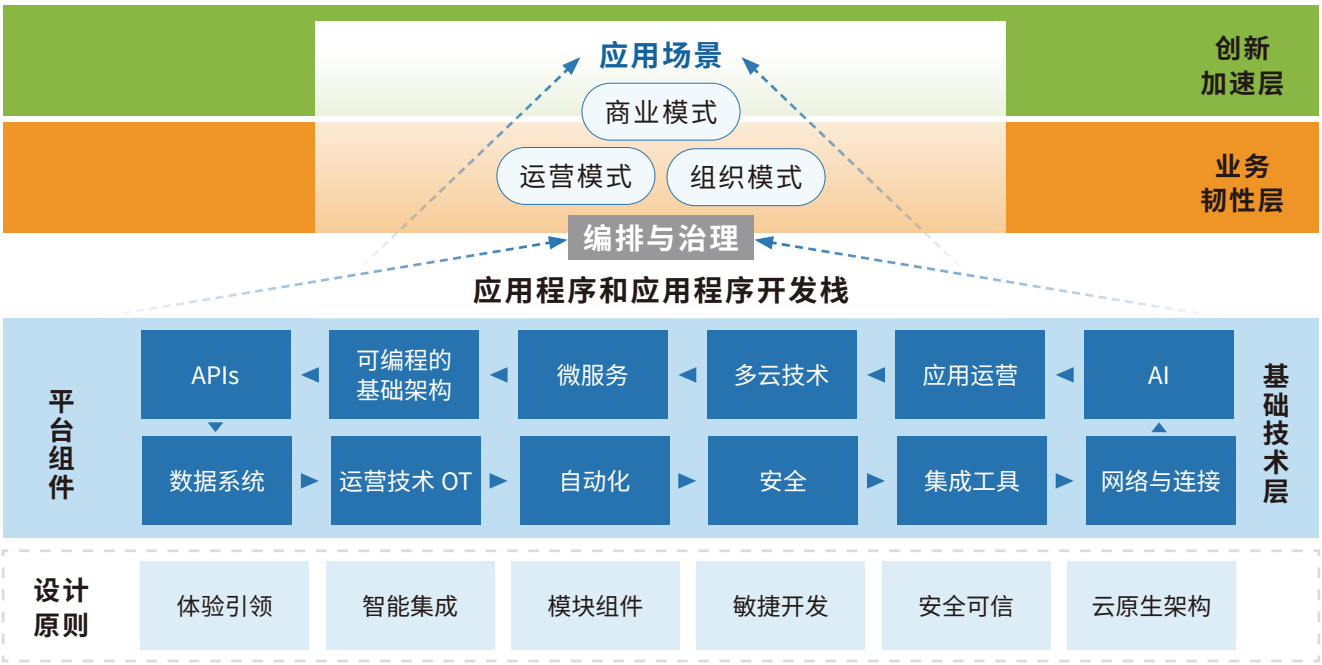
面对众多挑战及复杂多变的市场环境，企业日益倾向于采纳一种综合技术或深度整合的解决方案，因为完整的解决方案能够将新技术彻底嵌入企业日常运营的每一个环节，包括数据的采集、处理、深入分析，直至指导决策，形成一个闭环的管理体系。

2.3 端到端的生成式 AI 解决方案：满足企业真实需求的关键

在技术采纳的高级战略方向与企业深层次结构转型的影响下，端到端生成式 AI 解决方案作为关键性进展应运而生，它桥接了用户和技术，直接满足了企业对智能化与运营效率提升的全面需求。

图 6

端到端生成式 AI 解决方案：企业数字化转型的关键架构图



来源：IDC, 2024

精准对接业务需求，量身定制智能策略

端到端的 AI 解决方案在面对各行各业独特的业务场景和个性化需求时，能够展现出其不同于传统 AI 的灵活性，通过模块化设计，如模型选择、训练优化、

API 封装、UI/UX 设计等，为企业提供一站式技术产品化路径，企业也可以自由组合不同的 AI 模块，从而达到从基础的数据处理功能到复杂的分析预测应用的平滑过渡和灵活扩展。这种“可插拔”的模块化构建方式，不仅使企业能够快速应对市场的变化，同时也降低了技术门槛，确保 AI 技术能够精准对接业务需求，以最快速度实现技术从科研场景到落地的价值最大化。

集成工具链，持续优化实现动态适应

端到端解决方案能够提供集成开发环境（IDE）、自动机器学习（AutoML）平台、云原生部署服务、持续集成 / 持续部署（CI/CD）等一整套完整的工具链，简化了企业从原型到生产的过程。同时，与传统静态技术部署不同的是，端到端 AI 解决方案一般会内置基于反馈循环的学习与优化机制，使 AI 模型在实际应用过程中能够不断地接收业务反馈并进行自我学习与调整策略，从而持续进化升级，更紧密地贴合业务变化与市场需求。AutoML 就是一个很好的案例，能够自动根据用户交互数据优化模型，使得 AI 在预测销售趋势或客户服务响应上随着时间推移而逐渐精准，实施结果显示，使用自适应技术的企业在六个月内平均服务效率提高 20% 左右 。

深挖数据价值，提升处理效能

端到端的 AI 解决方案通过先进的算法与高度集成的平台，能够有效破除数据碎片化的问题，为企业运营分析创建一条从采集到落地的无缝“数据流”。在这一过程中，即使是少量或零散的业务数据，也能在解决方案精密的算法加工下转化为指导业务战略的洞察，换句话说，端到端的 AI 解决方案能够将有限的数据激发出无限的价值。据统计，得益于端到端 AI 解决方案能够整合分散在各个系统中的数据，企业在采用处理平台后，数据处理速度平均可提升 30%[6]。

据统计，得益于端到端 AI 解决方案能够整合分散在各个系统中的数据，企业在采用了处理平台后数据处理速度上平均可提升 30%[6]。

6: Future Enterprise Resiliency & Spending Survey Wave 4, IDC, April 2024, N=889

量化技术影响力，确保决策有理有据

端到端解决方案可以通过内置的业务影响分析模块和 A/B 测试框架，帮助企业设计实验，量化 AI 应用的业务影响。这些解决方案可以通过高级分析工具，如预测模型的拟合度报告、业务流程改进的量化指标等，使企业能够直观监控 AI 项目对关键绩效指标的提升。此外，基于生成式 AI 的预测分析能提供更精确的业务预测，从而为企业高管的决策提供支持。

强化跨部门协作，激发技术无限潜能

通过采用统一标准的 API 接口与行业协议，技术落地过程中的集成复杂度有所降低，减少了系统间的兼容障碍。

端到端的 AI 解决方案着重于构建一个无缝链接且高度协调的数字化生态环境，旨在促进传统与新兴技术的无缝融合、跨职能团队间的数据流通以及企业资源的智能优化配置。通过采用统一标准的 API 接口与行业协议，技术落地过程中的集成复杂度有所降低，减少了系统间的兼容障碍。与此同时，解决方案会集成项目管理和沟通工具，如看板、即时消息系统等，能够有效加强团队间的实时协作，确保技术实施过程中的透明度和效率。

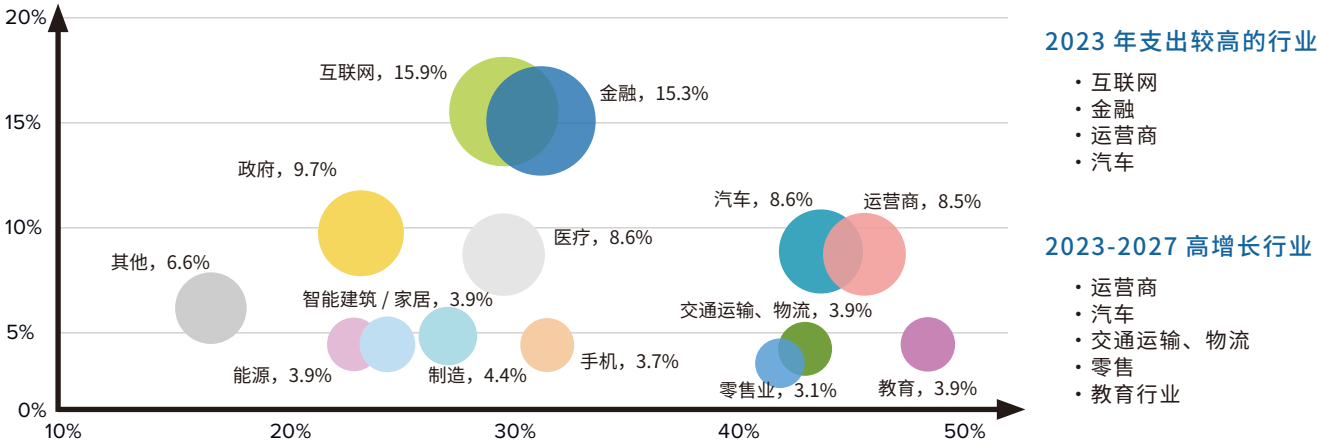
第三章 迈向 AI 智能体，生成式 AI 重塑千行百业

2023 年生成式 AI 在多行业的试点应用，带动了一场业务智能化的浪潮。据 IDC 统计，互联网、金融、医疗三大领域以 15.9%、15.3%、8.6% 的年增长率引领智能化进程，显示了行业对新技术的迫切需求和高度接纳能力。尽管运营商与汽车行业当前年增长率略低，但长远来看，这两大行业未来五年将分别以 43% 和 46% 的复合年均增长率迅速攀升，这代表着 AI 在通信基础建设和出行变革中的广泛应用前景 [7]。

7：IDC 中国，大模型应用场景与市场规模预测研究，2024

图 7

生成式 AI 不同行业支出规模与未来增长

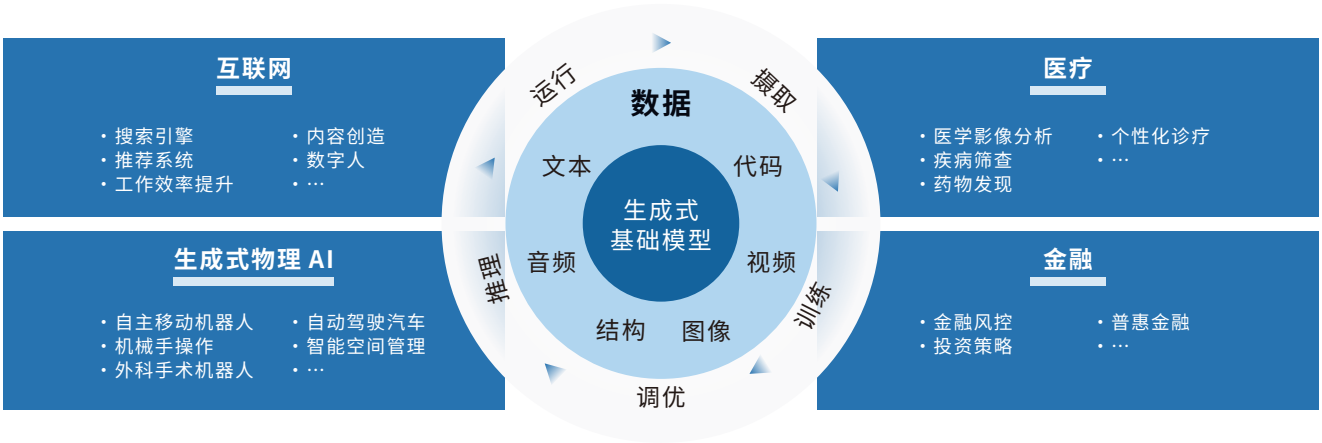


来源：IDC, 2024

生成式 AI 擅长基于数据资产和知识沉淀进行创作与输出，在各行业已经开发了一批会话类、知识管理类的共性应用；然而，由于行业属性各有不同，合规要求有异，生成式 AI 驱动各行各业创新与变革的侧重点亦各有不同。

图 8

生成式 AI 应用场景



来源：IDC, 2024

3.1 互联网行业：虚拟角色与内容生成

互联网多为数字原生企业，具有良好的数据基础和创新基因。互联网行业的挑战在于持续吸引用户注意力、增强品牌互动性，同时保持高效、创新的内容生产。于是，以虚拟角色和内容生成为代表的生成式 AI 技术正成为推动互联网行业进化的关键力量。生成式 AI 对互联网核心业务场景的赋能主要如下：

- **搜索引擎：**生成式 AI 使搜索引擎能够更好地理解复杂的查询意图，并汇集多方信源内容。通过自然语言理解和检索增强生成（RAG）等技术，搜索引擎可以直接提供总结性答案、建议或执行指令。伴随多模态大模型的发展，搜索结果还能以表单、思维导图以及图像、语音、视频等更加多样的方式呈现，使用户获取更优质的知识体验。

- **推荐系统：**生成式 AI 通过深度学习模型，根据用户的偏好和上下文，可以动态生成个性化推荐内容，甚至还可以深度个性化定制详情页面，这不仅能提升推荐的准确率，还能创造惊喜元素，提高用户粘性和满意度。

- **工作效率提升：**在企业内部，生成式 AI 与办公软件结合能够自动化处理大量重复性工作，如客户服务、报告生成、数据分析等，释放员工时间，让他们专注于更具价值的创造性工作，从而提升组织效率和创新能力。

- **内容创作：**面对互联网行业日新月异的大量内容创新及迭代的需求，生成式 AI 可以辅助创意、文案、脚本、编辑等多项内容工作，覆盖包括编曲、视频制作以及游戏角色设计等多模态领域，极大地提高创作者内容的产出效率。同时，生成式 AI 还可以结合时下热点和需求趋势，进行内容的创作和评审，确保内容的商业价值。

- **数字人：**数字人在互联网行业已广泛应用于直播、培训以及客服等场景中，生成式 AI 可以使数字人更好地理解用户的复杂指令甚至感知用户的情绪变化，让数字人与人之间的交互更加真实、灵动。此外，生成式 AI 还可以生成更丰富的虚拟人形象，更加贴合场景需求。

生成式 AI 在产品设计、应用开发与测试、流量分发、营销推广、用户运营等诸多业务场景都开始了有益的尝试。互联网拥抱生成式 AI 不仅在于利用 AI 技术优化现有业务与运营模式，更在于启发互联网新一轮的商业模式革命，打造基于生成式 AI 的新增长点。

案例：生成式 AI 促使定制化品牌内容的自动生成

图 9

百度营销平台“擎舵”

来源：百度 , 2024

百度营销平台推出的“擎舵”项目，利用生成式 AI 技术，根据用户特定需求和品牌特性，自动生成定制化虚拟角色与内容。这项技术覆盖视觉内容、交互对话、故事剧本等多个维度，旨在为用户提供独特的品牌互动体验。通过分析历史数据和用户行为，“擎舵”能生成高度相关的个性化内容，支持一键混剪视频制作，显著提升品牌内容生产的效率与个性化水平。实践表明，该技术已实现制作成本降低 85%、效率提高 100 倍、产品上市周期缩短 75% 的显著成效。此外，百度在构建基于检索和推荐的生成式大模型索引学习平台方面，通过融合判别与生成技术，使定向关键词生成的有效性从 30% 跃升至 100%，开辟了商业内容生成的新模式。

技术进步方面，百度与 NVIDIA 合作开发的 PaddleBox 项目，成功将稠密模型参数规模扩展至百亿级别，实现 100 倍的规模提升，并优化了多机加速性能，使训练效率提升 50%。同时，GPU 技术支持的 PGLBox 引擎，通过异步聚合通信库，实现了大规模图数据的高效多机训练，促进了百亿级语义模型与万亿级离散模型的联合学习，推动了大模型训练技术的实质性进展。（以上案例所展示数据截至 2024 年 GTC 大会）

3.2 医疗领域：药物研发的智能计算平台

医疗行业亟需解决资源分配不均、诊疗效率低下及个性化治疗方案稀缺的问题，尤其是在精准医疗日益增长的需求面前。通过分析大量医疗健康数据，生成式 AI 可以提供更高效、便捷和个性化的医疗服务。在生命科学方面，生成式 AI 大幅降低了医药发现的资金与时间成本。生成式 AI 对医疗行业的助益在短时间内，已得到行业的认证。根据 IDC 调研数据显示，2023 年 6 月全球医疗行业中只有 13%[8] 的企业在生成式 AI 方面进行了大量投资；仅仅四个月后，这一比例就上升到 46%[9]。生成式 AI 在医疗行业的主要应用场景有：

- **医学影像分析：**通过深度学习模型，AI 能精确识别影像中的异常结构，辅助医生进行早期癌症筛查、疾病诊断，如肺癌、皮肤癌的影像识别技术，以及心脏病、脑部疾病的影像辅助分析；还可以通过结合医学影像和病理信息等多模态数据，为医生提供更全面的诊断支持。

- **疾病筛查：**疾病类数据往往受到严格的合规性要求，一直以来训练数据获取困难都是 AI 技术在医疗业落地的一个制约项。生成式 AI 可以通过合成数据使 AI 模型更好地学习疾病诊疗案例，优化临床诊疗的表现。

- **药物发现：**利用生成式模型，AI 可以模拟数百万种化合物的结构和活性，加速新药候选分子的筛选过程。据报道，AI 可为公司降低高达 70% 的药物发现成本 [10]。

- **个性化诊疗：**基于患者的遗传信息、临床数据和疾病模型，生成式 AI 能够预测患者对不同治疗方案的响应，为制定个体化治疗计划提供参考。对话类应用的引入还能更加有针对性地回答广大来自病患的问题，改善患者的医疗服务体验。

8: IDC's June 2023 Future Enterprise Resiliency and Spending Survey, Wave 5
9: IDC's November 2023 Life Sciences Generative AI Survey
10: Insider Intelligence's AI in Drug Discovery and Development report

案例：智能化计算平台助力药物发现流程

唯信计算于 2020 年开始开发分子智能化计算平台 WeMol，旨在填补国产药物研发智能计算平台的空白。

图 10

分子智能化计算平台 WeMol

来源：唯信计算，2024

2021 年，唯信计算加入了 NVIDIA 初创企业加速计划。在 NVIDIA 技术和硬件的加持下，WeMol 以自主研发的 APLHA 系列独特算法为核心，完成了对从小分子、mRNA 到蛋白设计领域的药物发现全流程赋能，将大、小分子药物的生成、设计和计算模拟效率提升数百倍，累计服务各类生物医药企业与科研机构 500 余家。

借助与 NVIDIA 的合作，WeMol 更好地集成与对接了多种大模型以及 GPU 加速算法。例如通过 NVIDIA NIM 微服务解决方案，实现 AI 推理模型的快速部署；利用专注于药物发现的 AI 模型微服务 NVIDIA BioNeMo NIMs，WeMol 能够直接部署计算机辅助药物设计（CADD）AI 模型、DiffDock 分子对接工具、OpenFold 蛋白质结构预测模型和 ESM 蛋白质语言模型，以及针对抗体研究和其他药物发现流程的多种模型。WeMol 支持多种形式的抗体设计、免疫原性预测、LNP 递送系统设计、可开发性优化、mRNA 序列设计及超高通量虚拟筛选等计算，可搭建定制化的分子数字化及智能计算平台。其中，人源化和免疫原性的模型预测准确度能达到 90% 以上；在抗体可开发性和抗体亲和力改造方面，模型计算结果与实验反馈也高度吻合，得到了客户的高度认可。

未来，唯信计算计划将 WeMol 平台拓展至基因组学和医疗影像领域，利用 NVIDIA 的 Parabricks 和 MONAI 等平台工具，为中国医疗医药行业带来更全面的智能化研发平台，助力行业向更加高效、精准的药物研发迈进。

3.3 金融行业：风险管理、投资决策与反欺诈

金融行业向来是实践行业转型的引领者，是最有望诞生第一批成熟落地场景的行业。金融行业有严格的合规和监管要求，必须严格控制风险。生成式 AI 以其强大的数据处理、模式识别和内容生成能力在金融的风险管理、投资决策以及普惠金融等领域有着巨大的应用潜力。通过深度学习和自然语言处理技术，生成式 AI 能够自动化处理复杂的金融数据，实现对风险的快速识别与响应，增强金融机构的风险抵御能力，同时为投资、信贷决策提供更加精准的分析支持。金融行业生成式 AI 的主要应用场景如下：

- **金融风控：**对于风险管理，生成式 AI 能够自动生成关于贷款申请人或投资项目的详细调查报告，涵盖财务状况、信用历史、合规性检查等。通过自然语言处理技术，AI 系统能够理解有关金融规章制度的提问，为客户提供准确的信息支持。生成式 AI 还可以辅助生成风控相关算法代码，例如结合 NVIDIA 全栈技术，如 RAPIDS、Spark 和 Deep Graph Library（DGL），助力银行实现自动化反金融犯罪、改善信用风险建模、更好地完成风险管理和欺诈检测并降低成本。对于欺诈检测领域，生成式 AI 通过数据增强、模拟欺诈场景等方式提升金融欺诈检测的准确性。生成式 AI 也可用来生成额外的合成数据，解决真实数据不足的问题，进而优化反欺诈系统。

- **投资策略：**生成式 AI 可以分析财务报告、市场研究报告，提取关键财务指标和市场趋势，为投资者提供有价值的洞见。在算法交易领域，生成式 AI 通过情感分析社交媒体上的讨论，预测市场情绪和趋势；同时，它能够将投资者的口头描述转化为交易算法的代码，实现策略自动化。例如，NVIDIA NeMo Curator 能够简化数据整理任务，如数据下载、清理、质量过滤、精确或模糊数据去重等；NVIDIA RAPIDS 可在算法交易的因子计算与挖掘和算法开发等环节完成 GPU 加速，提升性能；在生产环节，NVIDIA

Triton 可实现算法推理加速，助力金融机构完成算法交易的部署。**乐天证券的 AI 虚拟投资助手，运用 NVIDIA RIVA+LLM 技术，根据客户数据提供个性化投资建议，实现了高度定制化的客户虚拟投资助手服务。**

- **普惠金融：**农村金融与小微企业融资面临的一个共同挑战是客户信用数据不足，导致传统金融服务难以评估其金融风险。生成式 AI 在优化信贷决策方面展现出巨大潜力，例如在农村金融领域，商业银行借助卫星遥感图像可识别农作物的生长情况与种植面积，以形成信用资产，再通过生成式 AI 解决遥感成像清晰度不高的问题，并且结合地理、天气、宏观政策以及市场供需预测等信息，更精准、智能地推荐授信额度；在供应链金融领域，生成式 AI 与知识图谱相结合，能够完整绘制产业链图谱，进而定位小微企业所在产链位置，并综合上下链信息，实现对企业经营行为的全面洞察，以便准确地评估小微企业的信用风险。此外，基于生成式 AI 的 7x24 小时在线聊天机器人能够为普惠金融、农村金融和小微企业用户提供即时咨询服务，解答贷款、投资相关问题。生成式 AI 的应用不仅能够显著提升金融服务的便捷性和可负担性，而且有助于缩小城乡、大小企业之间的金融服务差距，推动经济均衡发展。

3.4 生成式物理 AI：机器人与自动驾驶

要实现在现实环境中高效运作的机器人与自动驾驶系统，关键能力在于系统需能区分并理解物体特性，同时将高级决策策略转化为精确的动作指令。尤其对于自动驾驶汽车而言，由于其普遍采用电力驱动，能效成为了设计与功能实现中的核心考量因素。历史上，自动驾驶技术面临的一大挑战在于如何准确感知并理解复杂多变的周围环境。生成式物理 AI（Generative Physical AI）的兴起，为构建能够灵活应对现实世界不确定性的机器人提供了全新路径。

相较于传统生成式 AI，生成式物理 AI 进一步整合了对三维空间关系及物体物理特性的深刻认知，显著提升了系统的智能水平。

生成式 AI 技术赋予了自动驾驶汽车感知、理解并执行复杂任务的能力。该技术通常被嵌入到机器人或自动驾驶汽车中，通过集成传感器与执行器的运动技能，实现对现实世界的深度交互与理解。相较于传统生成式 AI，生成式物理 AI 进一步整合了对三维空间关系及物体物理特性的深刻认知，显著提升了系统的智能水平。在开发过程中，开发者利用强化学习在模拟环境中对自动驾驶机器进行训练，这一方法允许 AI 通过无数次试错，在安全、高效的数字环境中快速掌握技能。更为重要的是，这些系统还具备从人类示范中学习的能力，从而不断增强其执行效率与环境适应能力。

生成式物理 AI 可以帮助机器高精度地适配各种环境，为机器人提供动力，使其能够包装纸箱、帮助制造车辆、提高物流和库存管理的运营效率，甚至在手术室为医生提供帮助。

- **机器人：**借助生成式物理 AI，机器在多样化环境中的高精度适应能力显著提升。
  - 自主移动机器人（AMRs）：在仓库场景下，AMRs 凭借集成传感器实时提供的数据，能够在复杂空间内导航，有效规避包括人类在内的各类障碍物，显著提升作业效率与安全性。
  - 机械手操作：通过分析传送带上物品的朝向，机械手能精细调整抓取策略，展现出针对不同物品类型的精准操控技能，提高了包装、装配等任务的自动化水平。
  - 外科手术机器人：在医疗领域，生成式物理 AI 使手术机器人能够掌握缝合、穿针等高精度手术技巧，展现了其在辅助完成复杂医疗程序中的精确性与灵活性，减轻了外科医生的工作负担。

- **自动驾驶汽车（AV）**：配备先进传感器的自动驾驶汽车，在生成式物理 AI 的加持下，能够准确感知并解析周围环境，无论是在高速公路还是城市街道，都能做出决策。该技术增强了 AV 识别行人、应对交通与天气变化、自主执行车道变换的能力，使其能够灵活处理多种不可预见情况，有效提升行驶的安全性与舒适度。

- **智能空间管理：**在工厂、仓库等大型室内区域，生成式物理 AI 通过固定摄像头与视觉模型，可实现对各类实体与行动的全面监控，进而优化动态路由与运营效率。同时，这些系统能准确识别并解读广阔复杂的环境，确保人员安全，提升整体管理水平。

案例一：生成物理 AI 用于外科手术

ORBIT-Surgicalis 项目是 NVIDIA 与大学研究人员合作开发的一个模拟框架，用于训练手术机器人。这些机器人旨在增强外科团队的能力，减轻外科医生在微创手术中的认知负担。该框架包括腹腔镜培训课程中的十几种操作，如抓取和精确放置针头等小物体。

ORBIT-Surgicalis 基于机器人仿真平台 NVIDIA Isaac Sim 开发，采用了由 NVIDIA GPU 驱动的强化学习和模仿学习算法。同时利用 NVIDIA Omniverse 和通用场景描述（OpenUSD）技术进行逼真渲染，增强了模拟的真实感。这种设置可以生成高保真合成数据，帮助训练 AI 模型，如，在真实世界中分配手术工具等任务的视频。

与现有的手术框架相比，通过利用 GPU 加速和并行化优势的手术模拟器，医疗团队能够将机器人的学习速度提高一个数量级。

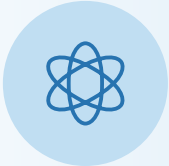
案例二：简化制造工厂机器人开发流程

全球约有 1,000 万家工厂，总价值达 46 万亿美元，制造业为使用生成式物理 AI 提供了大量机会。

电子产品制造商富士康正在利用数字孪生来训练 AI 机器人，从而提高工厂的自动化水平。通过利用包括 Teamcenter 在内的西门子 Xcelerator 产品组合和 NVIDIA Omniverse，富士康正在创建一个虚拟环境，以简化 NVIDIA Blackwell HGX 系统的生产流程。

在数字孪生系统中，富士康的工程师使用 Omniverse 将 3D CAD 元素集成到一个单一的虚拟工厂中，并在其中使用 NVIDIA Isaac Sim 对机器人进行训练。如爱普生的机器人，通过 NVIDIA Isaac Manipulator（机械臂轨迹规划功能）学习物体操作等复杂任务。此外，由台湾 FARobot 公司开发的富士康自主移动机器人（AMR）可利用 NVIDIA Isaac Perceptor（3D 环视感知）在工厂车间内导航，创建实时 3D 地图并避开障碍物，NVIDIA cuOpt 则提供路线优化功能。

这种先进的培训和模拟环境有望显著提高生产效率，每年可降低成本和能源消耗 30% 以上。



生成式 AI 的发展和其在各行业业务场景的融合密不可分，互相促进；换言之，生成式 AI 的发展离不开算力的支持和业务场景的需求，行业合作伙伴也同样需要考虑开发工具的易用性、调取模型的开放性以及行业解决方案能力。

第四章 NVIDIA 的生成式 AI 技术：重新定义计算与智能的边界

对大多数企业而言，大规模的算力资源投入不会成为其发展生成式 AI 的核心，端到端的全栈解决方案才是其落地生成式 AI 的关键。NVIDIA 作为全球领先的加速计算公司，无论是在底层硬件资源，还是软件平台、加速框架、开发工具、行业应用方案等方面，均有着丰富的技术栈和经验，是行业拥抱生成式AI战略的理想合作伙伴。

4.1 硬件支撑：为生成式 AI 提供卓越计算能力

底层算力支持对于 AI 发展的重要性不言而喻。GPU 为深度学习模型提供高效的并行计算能力和数据处理能力，有效加速生成式 AI 的训练和推理过程。然而，随着模型参数量的飞速增长以及模型模态的复杂化，单节点计算能力日渐面临瓶颈。实现计算能力可扩展的关键，在于组建计算集群，满足日益增长的大模型训练、推理，以及 AI 应用部署的算力需求。因此，除了单卡算力外，

运算架构对于算力的智能调度，网络通信对于单卡之间交流损耗的弥补，也都至关重要。

具体来说，NVIDIA 的 GPU 架构从 Fermi 到 Hopper，每次架构升级都带来性能和能效上的显著提升。新一代的 Hopper 架构中，NVIDIA 更新了 Tensor Core，以专用的硬件单元加速模型训练和推理等 AI 工作负载，并引入 Transformer 引擎，实现 FP8/FP16 混合精度计算，动态调整算力，在保持准确性和提供更强安全性的同时，提高吞吐量，加速生成式 AI 的所有工作负载，从而实现效率与性能同时增长。

在网络方面，NVLink 与 NVSwitch 提供高带宽和低延迟的数据传输，以保证计算集群运行的高效性。NVLink 技术可用于 GPU 之间的高速点对点互连，提供高带宽和低延迟的数据传输，并通过 Peer to Peer 技术完成 GPU 显存之间的直接数据交换，进一步降低数据传输的复杂性。这对于分布式环境下运行的复杂 AI 模型尤为重要；更快的纵向互联有助于服务器集群内每个 GPU 性能的充分释放，从而提升整体计算性能。在此基础上，NVSwitch 实现了服务器中多 GPU 之间的高带宽、任意连接，完成多 GPU 通信任务。

作为 AI 应用的核心基础设施，NVIDIA 数据中心可以为生成式 AI 提供强大的计算和存储能力，确保 AI 应用能够稳定运行并处理大量数据。针对不同的 AI 部署需求和算力需求规模，NVIDIA 提供适应多等级数据中心的部署方案。

计算集群规模受互联带宽的限制，会导致 GPU 的利用率随集群规模扩大而降低。对此，NVIDIA 提供 InfiniBand 高速网络，提升 GPU 集群的扩展性。InfiniBand 支持可编程拥塞控制和动态路由，在训练过程中能够同步优化数据整合流程，从而实现所有端口均以全线速进行数据传输，并极大地减轻了交换机对计算性能的制约。此外，InfiniBand 还结合 SHARP（Scalable Hierarchical Aggregation and Reduction Protocol，可扩展分层次聚合和归约协议）技术，减少了网络传输的数据量，缩短消息传递接口（MPI）操作的时间，并提高数据中心效率。

作为 AI 应用的核心基础设施，NVIDIA 数据中心可以为生成式 AI 提供强大的计算和存储能力，确保 AI 应用能够稳定运行并处理大量数据。针对不同的 AI

部署需求和算力需求规模，NVIDIA 提供适应多等级数据中心的部署方案。与此同时，云原生技术利用容器化、微服务、持续集成和持续部署等技术为 AI 应用的开发和部署提供全新的方式。这些都为更快速地构建、测试和部署 AI 应用提供了良好的基础。

4.2 软件与工具：构建全面的 AI 开发生态

单独的加速算力设施难以成为生成式 AI 生产化的有效工具，为此，NVIDIA AI Enterprise 为 AI 技术的加速落地提供全栈化支持。该平台是一个云原生的软件平台，可以简化生产级 AI 解决方案开发和部署的工作，涵盖生成式 AI、计算机视觉和智能语音等诸多方向。其微服务架构不仅易用，还能够在企业级安全、服务支持和稳定性等方面提升模型表现，确保 AI 解决方案平稳过渡到生产环境，为企业提供 AI 支持。

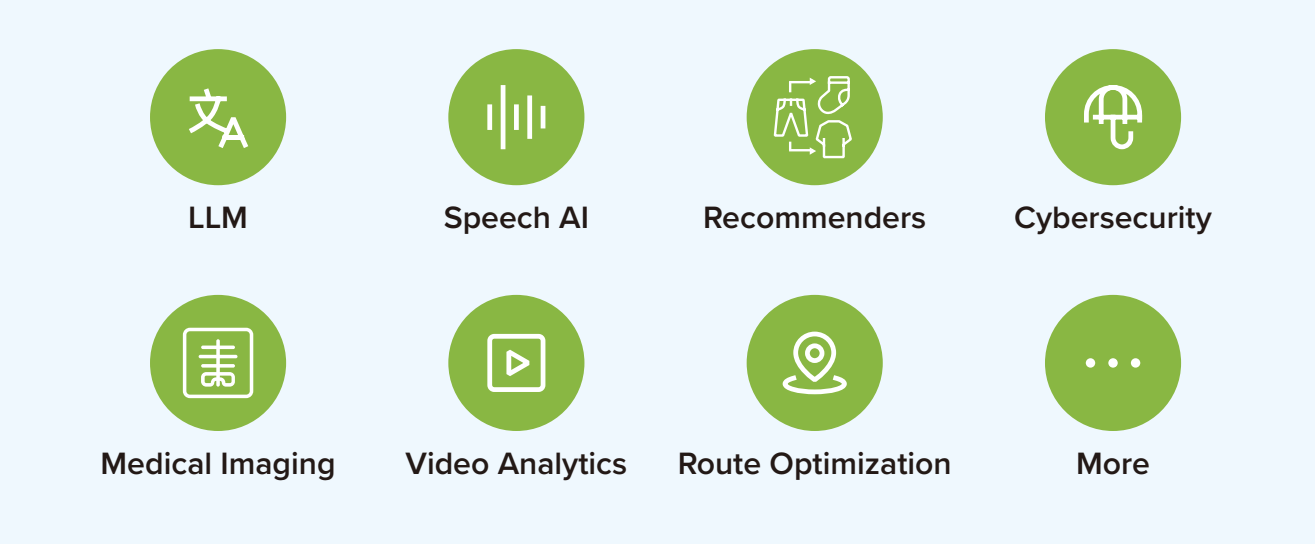
为实现生成式 AI 的快速部署，NVIDIA AI Enterprise 包含了大量为特定应用场景服务的软件开发工具集和模型，针对不同行业和场景需求做了适配。例如，NVIDIA NeMo™ 是一个用于开发定制生成式 AI 的端到端平台，其中包括用于训练、定制和检索增强生成、防护和工具包、数据整理工具以及模型预训练的工具；Clara 用于医疗行业，辅助医疗影像识别和医药研发等场景；Picasso 则支持开发人员开发和部署用于视觉内容创建的模型。这些预备制化的工具集使用户可以快捷、方便地按需部署运行。这些工具集相互协同，使企业能够顺利迁移到最新版本的开源工具，而不会引发连锁反应。通过 NVIDIA AI Enterprise 预配置的工具，用户能够快速、无阻地进行相关应用的部署升级。企业可直接试用先进的基础模型（如 Llama2，Stable Diffusion，Nemotron-3 等），也可通过 NVIDIA NeMo 利用专有数据对基础模型进行调

NVIDIA AI Enterprise 平台上目前已经发布了超过 4,500 个软件包（涵盖开源与第三方选项）以及 60 多个 NVDIA CUDA libraries。这些工具集相互协同，使企业能够顺利迁移到最新版本的开源工具。

优和测试。这些基础模型使用了负责任来源的数据集，企业可将应用连接到 API 端点，在任意位置部署和运营模型。

图 11

AI 用例与工作流



来源：NVIDIA, 2024

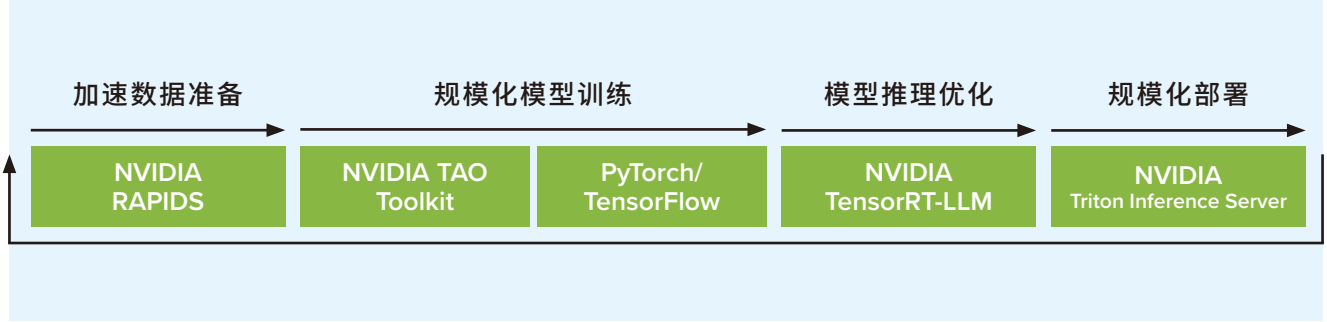
NVIDIA AI Enterprise 软件套件还支持 AI 开发，并包含针对底层计算资源的管理优化解决方案。

- **用于加速数据准备：**用于加速数据准备：用于 AI 和数据开发与部署的工具集集成了 RAPIDS 等先进工具，可通过 RAPIDS cuDF（加速 pandas 数据帧）和 RAPIDS accelerator for Spark（加速 Spark）等零代码更改工具加速数据预处理，使开发人员能够加速其现有代码。

- **用于大模型训练：**NVIDIA AI Enterprise 可与 TensorFlow 和 PyTorch 等主流 DL 框架协同工作，结合 NVIDIA 基于迁移学习技术的低代码框架 TAO，为开发人员创建一个多样化、高效的开发生态系统。
- **用于模型推理优化：**TensorRT-LLM 专注于模型推理阶段的性能优化，确保算法应用达到最佳运行效率。
- **用于大规模部署：**Triton 专为大规模部署而设计，可确保复杂企业环境中模型服务的稳定性和可扩展性。

图 12

NVIDIA AI 模型开发与部署工具

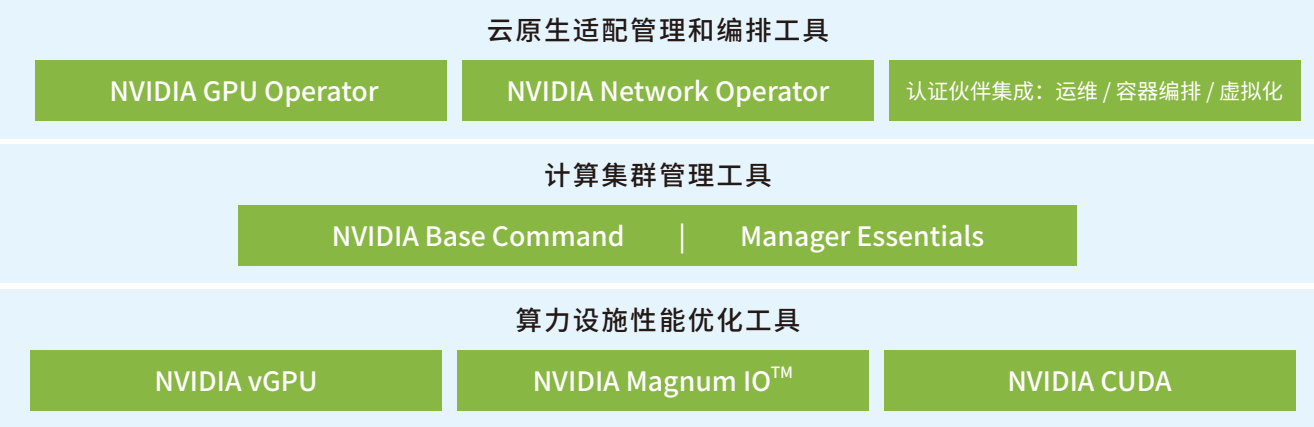


来源：NVIDIA, 2024

NVIDIA 还提供了丰富的工具栈帮助企业实现更便捷的模型部署和应用开发。**云原生管理和编排工具，**通过虚拟化、模块化、容器化手段来高效率和灵活地调用云资源，在优化模型性能的同时，提高模型可迁移性；**计算集群管理工具，**实现对集群的自动调配和管理，支持与 Kubernetes 进行编排，为多云和混合云环境中的异构 AI 和高性能计算集群提供快速部署和端到端管理体验；最后，**算力基础设施加速工具，**完成对 GPU 算力的优化配置。

图 13

NVIDIA 算力的基础设施管理工具



来源：NVIDIA, 2024

微服务架构在 AI 开发中发挥着重要作用，它能够将复杂的 AI 应用拆分为一系列小型、独立的服务，从而提高系统的可扩展性和灵活性。NVIDIA NIM 作为 NVIDIA AI Enterprise 的重要组成部分，专为加速企业级生成式 AI 的推理部署而设计。

NVIDIA NIM 是 NVIDIA AI Enterprise 的一部分，是一套易于使用的预构建容器工具，目的是帮助企业加速生成式 AI 的部署。这些预构建的容器支持多种 AI 模型。只需一个命令，NIM 微服务即可帮助企业客户部署 AI 模型，以便使用标准 API 和几行代码轻松集成到企业级 AI 应用程序中。NIM 基于可靠的基础设施（包括 Triton 推理服务器、TensorRT、TensorRT-LLM 和 PyTorch 等推理引擎）构建，旨在促进企业客户根据其自身需求和选择大规模无缝进行 AI 推理，从而确保企业可以满怀信心地在任何地方部署 AI 应用程序。无论是在本地还是在云端，NIM 都能高效实现大规模加速生成式 AI 推理。

NIM 提供了两种灵活的试用方式：一是通过 preview API，该 API 涵盖了全面的 AI 模型库，用户可以试用并构建 AI 工作流的原型；二是预构建容器，可供用户直接下载并在自有基础设施上部署，可在 5 分钟内完成从下载到运行的全过程。无论是选择在 NVIDIA NGC 云平台上使用 preview API，还是在本地基础设施上部署预构建容器，NIM 均统一提供了一套标准化的 API 接口，最终用户可以轻松地将 AI 功能作为关键组件嵌入到其应用程序中，享受无缝的集成体验。

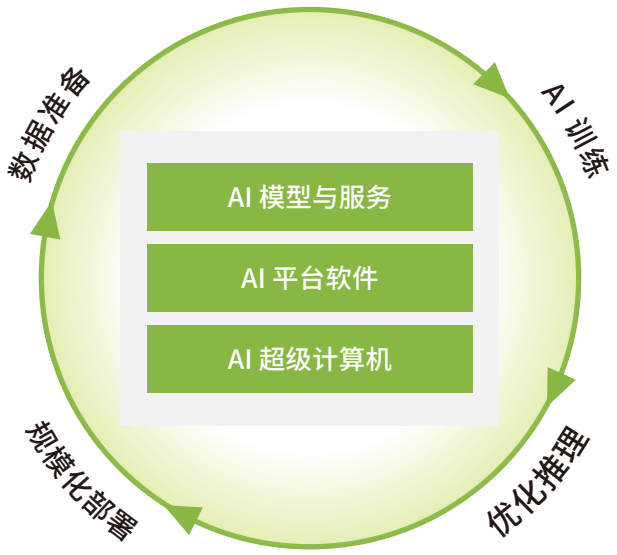
NVIDIA AI Enterprise 不仅提供适合 AI 从业者的出色开发工具、框架和预训练模型，而且能充分满足 IT 专业人员在管理和编排方面的要求。这意味着，企业可以在性能、高可用性和安全性上获得全面保障。性能上，NVIDIA NIM 微服务提供了优化的运行时间，简化了生成式 AI 的开发；安全上，NVIDIA AI Enterprise 通过持续监控安全漏洞和模型定制所有权来保护公司数据和知识产权；部署上，基于标准的容器化微服务可以运行在云端、数据中心和工作站上，确保运行位置不受限；稳定和可靠性上，通过 API 稳定性、软件管理和 NVIDIA Enterprise Support 预测软件生命周期，帮助确保项目保持平稳进行。

4.3 端到端的解决方案：加速 AI 应用的部署与运行

伴随着技术的进步，用户的需求不再集中于某个单独的场景问题，尤其是大模型的发展为企业提供了 AI 应用规模化落地的可能。在此背景下，走向端到端解决方案是必然趋势。此转型不仅超越了软硬件的简单叠加，更实现了数据流程的全程优化与协同，是提升效率、强化安全、优化体验的关键。端到端策略不仅体现了深度整合能力，更是协调优化数据全流程、提升体验感的关键所在。

图 14

NVIDIA AI 应用解决方案



来源：NVIDIA, 2024

NVIDIA 注意到全栈解决方案对于大语言模型（LLM）的重要性，并为此构建了一个全面的技术生态系统。这一生态系统覆盖了高端硬件、专业软件框架，直至模型的训练与部署全流程。硬件层面，NVIDIA 依托其多样化 GPU

软件方面，NVIDIA 推出了 NeMo 框架可以端到端的满足整个 LLM 工作流的需求，其中涵盖数据处理、生成式 AI 模型训练和推理等方面的需求，简化并加速了开发流程。

产品，加上高速 GPU 间通信技术如 NVLink 与 NVSwitch，为模型运行提供了坚实基础。此外，NVIDIA 设计的高性能服务器，配置多块 GPU 并可扩展为集群，极大提升了处理能力，满足大规模运算需求。软件方面，NVIDIA 推出了 NeMo 框架可以端到端地满足整个 LLM 工作流的需求，其中涵盖数据处理、生成式 AI 模型训练和推理等方面的需求，简化并加速了开发流程。在模型训练和推理环节，NVIDIA 推出 HGX 和 MGX 服务器，实现大规模运算能力的扩展，并凭借 GraceHopper 等超级芯片为模型运行提供最大化的性能支持。针对不同数据模态的大模型，NVIDIA 亦有布局。例如，在文本生成图像领域，基础模型 Picasso 服务于文生图应用，能够基于文本数据进行高效训练与微调。而对于融合多种模态数据的大模型，NVIDIA Edify 则表现更佳，不仅能够生成高质量的 3D 模型、物理材质及图像，还在艺术与创意产业中推动了产品的快速创新与迭代，显著提升了工作效率。

在药物研发领域，NVIDIA Clara Discovery 集成了 GPU 加速及优化的框架、工具、应用和预训练模型。BioNeMo 是一种特定领域的框架，用于在超级计算规模下训练和部署基于 NeMo Megatron 的生物分子 LLM，框架包含 Transformer 架构的多种模型。该框架有助于医药企业预测蛋白质结构，探索化学反应、扫描候选药物和分子模拟，可帮助科学家和研究员更快地将药物投向市场。在基因组学领域，NVIDIA GPU 的计算资源和 NVIDIA Clara Parabricks 工具包有助于研究者更快地完成基因组测序。NVIDIA Clara Parabricks 作为一整套现成的基因组学分析解决方案组合，旨在提高速度、优化准确性和可扩展性，支持从 DNA 到 RNA 的分析，以及用于开展初级分析、二级分析和三级分析的应用流程。研究员可利用 NVIDIA 解决方案分析细胞突变的分子特征，识别病毒的突变体变异情况，助力攻克癌症和病毒。

在自动驾驶领域，NVIDIA 生成式 AI 可协助车企构建软件定义智能驾驶汽车，其解决方案覆盖了从感知层到决策执行的各个环节，其 NVIDIA DRIVE 平台集成了高性能计算能力，包括 DRIVE Thor SoC，能够为车辆提供托管自动驾驶功能和车载 AI 应用的基础，再配合 DRIVE OS 安全操作系统确保智驾的

安全性。决策与路径规划方面，DRIVE Chauffeur 平台依托强大的 SoC 资源，能够很好地应对复杂路面情况，制定安全高效的行驶策略。NVIDIA 还通过 DRIVE Sim 和 Constellation 仿真系统，借助生成式 AI 创造多样化的测试场景，加速算法验证与迭代。此外，DRIVE Concierge 作为智能数字助手，融入自然语言处理能力，提升了用户在智能座舱中的体验。这一系列方案不仅紧密协同，还严格遵循行业安全标准，共同推进了自动驾驶技术的成熟与广泛应用。

企业在推进生成式 AI 应用的过程中，需细腻打磨每一个核心环节，从数据到技术，再到成本与安全，每一环都需精心布局，以实现技术的高效运用和业务的稳健发展。

第五章 前景与战略：生成式 AI 将会持续落地，引领产业全面迈向数字化时代

5.1 生成式 AI 未来趋势：应用边界不断拓展，持续发挥智能化价值

多模态 AI 崛起

生成式 AI 的竞争焦点正逐渐从单模态领域扩展到多模态战场。IDC 数据显示，至 2023 年底，中国的数据量将攀升至 30.962 EB，并有望在 2028 年达到 97,057 EB，期间年复合增长率预计为 25.7%[11]。随着数据量的激增，海量多模态数据的标注和训练也将带动数据采集、存储、标注、治理产品的新一轮升级。IDC 认为，未来多模态 AI 拥有良好的发展前景，同时其还将促进自然语言处理（NLP）和计算机视觉（CV）等技术领域的瓶颈突破，形成协同效应。

11: Worldwide IDC Global DataSphere Forecast, 2024–2028: AI Everywhere, But Upsurge in Data Will Take Time

生态系统逐步构建

生成式 AI 的长远发展依赖于一个健康、开放的生态系统。这一生态系统的核心在于促进技术的标准化与模块化，确保不同组件间无缝对接，加速技术从实验室走向实际应用。具体来说，通过数据平台、AI 平台和计算资源的开放，企业不仅能够吸引众多开发者和合作伙伴，更能够促进自身产品深化。除此以外，开源大模型也是丰富生态的路径之一。目前，LLaMA、ChatGLM 都是开源的通用模型。对于用户来说，开源模型可以进行本地部署，提升数据安全性；对于技术厂商来说，通过开源大模型，企业能够共享研究成果、加速技术创新。通过生态合作模式，能够有效缓解单一企业或研究机构所遭遇的资源局限，加速技术突破的进程。

算力的重要地位依旧

据 IDC 测算，2022 年我国智能算力规模已达到 260EFLOPS（每秒百亿亿次浮点运算次数），到 2027 年这一数字将跃升至 1117EFLOPS，期间的年复合增长率预计将高达 33.9%。

智能算力作为驱动 AI 大模型发展的核心动力，其需求随着模型规模的不断扩大而上升。据 IDC 测算，2022 年我国智能算力规模已达到 260EFLOPS（每秒百亿亿次浮点运算次数），到 2027 年这一数字将跃升至 1117EFLOPS，期间的年复合增长率预计将高达 33.9%[12]。伴随算力需求的激增，服务器需求量亦同步攀升。为了满足需求，不少企业开始聚焦数据中心建设和运营。除此之外，特定的 AI 芯片也是重要发力方向，如 ASIC 和 TPU，它们能够针对特定的算法和工作负载进行优化，降低能耗，提升算力效率。展望未来，更多的定制化芯片将会被推出，以适应多元化的 AI 应用场景。

可持续数据中心战略持续发展

为了应对新技术所带来的能耗挑战，低碳、绿色可持续的数据中心战略对企业乃至国家都极为重要。数据中心的战略不再是传统硬件和基础设施的管理升级，基于加速计算平台构建高效易用的数据中心将成为新趋势。在这个过程中，还需要对软件和算法进行关注，通过对网络架构进行优化，从而提高数据的传输

12: IDC 2023-2024 年中国人工智能计算力发展评估报告

效率并预防系统潜在风险，最终实现单位功耗（比如生成每 100 个 token 所消耗能源）的大幅降低。除此之外，在数据中心中采用风能、太阳能和水能等可再生资源，提高每个服务器的 AI 应用密度，以及通过灵活选址策略等方式，均能有效降低电消耗和碳足迹，也有助于达到控制运营成本的效果。例如 Project Natick 的海底数据中心，利用海水进行冷却，大幅降低了冷却系统的能耗；DeepMind 的新算法，也是通过优化数据中心的冷却系统来减少能源的消耗；硬件创新方面，Blackwell GPU 有效解决了大数据处理中的通信瓶颈问题，进而提升了能源使用的综合效率。这些绿色数据中心的实践为行业带来了实际的价值，也体现出了企业的社会责任感，在经济、环境以及品牌形象等方面都具有正向价值。

5.2 IDC 建议：面向企业：技术为本，效益为先，与时俱进

兼顾上下级的从属关系建立完善的合作体系

在中国有 60% 的企业在构建和运用生成式 AI 模型时，选择融入由第三方提供的且经过合法授权的数据资源，这一调研结果不仅凸显了企业对高质量数据的依赖，也反映了在 AI 技术推进的浪潮中对外部数据合作的重视与接纳。

在数据方面，据 IDC 统计，无论是中国企业还是全球企业，在生成式 AI 模型中都在使用多种数据源。值得注意的是，在中国有 60%[13] 的企业在构建和运用生成式 AI 模型时，选择融入由第三方提供的且经过合法授权的数据资源，这一调研结果不仅凸显了企业对高质量数据的依赖，也反映了在 AI 技术推进的浪潮中对外部数据合作的重视与接纳。市场需求不断增长，国家对于数据权限的确定也在不断明晰，企业在拓展市场的过程中应尽快与数据所有者建立合作机制，需要建立一套完善的体系，避免地理区域所带来的冲突关系。总的来说，构建一个结合自主研发、开放合作、高效管理与前瞻视野的技术创新生态系统，对于技术厂商而言至关重要，这不仅能够促进技术的快速成熟与应用，也是在复杂市场环境中稳固地位、把握未来机遇的关键所在。

13: IDC's Future Enterprise Resiliency & Spending Survey, Wave 2, 2024 年 2 月 , n = 896 （北美 : 371, 欧洲与其它地区 : 225, 亚太 : 300 [中国 : 100] ）

加快技术创新凸显端到端解决方案的优势

IDC 数据显示，生成式 AI 带动的基础设施、平台 / 模型、应用程序和服务的支出预计在 2023 年至 2027 年将以 73% 的复合年增长率增长[14]。技术厂商想要在蓝海市场中保持领先地位，就需要聚焦技术创新与协作策略的不断融合。目前，基础大模型的同质性较强，企业应着眼于行业、场景领域，将重心放在解决方案的完整性，以及产品的易用性、系统兼容性等方面上。未来，随着生成式 AI 的法律法规不断完善，企业的竞争优势将会逐渐转移到实际应用上，具体包括 AI Agent、安全可信、数据服务等。基于此构建的一体化解决方案能够降低企业对新技术的应用门槛，满足用户的多元化需求。

通过优化架构打造灵活的成本控制

根据 IDC 调研结果显示，三分之一的企业已将 AI 融入其生产流程，另有半数企业正处于与 AI 技术融合的探索阶段或自主研发符合特定需求的 AI 模型。

随着市场的变化速度加快，企业的不确定性也逐渐增多，在成本控制和企业运营方面需要更加灵活的策略来适应多变的业务需求。从部署方式来看，边缘计算正日益成为主要的 AI 训练和推理环境之一，云部署也在快速增长；AI 服务器的本地部署增长虽相对缓慢，但仍呈现健康的增长趋势。根据 IDC 调研结果显示，三分之一的企业已将 AI 融入其生产流程，另有半数企业正处于与 AI 技术融合的探索阶段或自主研发符合其特定需求的 AI 模型[15]。除了部署方式外，发力微服务架构、优化成本结构也应是企业重要布局方向，微服务的应用不仅允许企业根据实时需求动态调整资源分配，避免资源闲置造成的浪费，还在硬件与软件层面推动了更高层次的虚拟化与选择自由度，进一步降低了成本并提升了系统灵活性。

通过技术升级确保用户的信息安全

伴随着技术发展，安全和个人隐私已经逐渐成为用户关注的重中之重。企业应在多层次防御方面进行发力，在硬件中内建加密加速器和安全处理器，为生成式 AI 后续应用提供基础的安全保障。在软件层面，则需要重点关注数据的加

14: The Potential Impact of Generative AI on the Global DataSphere and Global StorageSphere
15: IDC's June 2023 Future Enterprise Resiliency and Spending Survey, Wave 5

密策略和具体到人的访问和控制权，利用差分隐私技术结合生成式 AI 带来的无间断安全检测系统，持续监督潜在危险，为用户带来安全性的持续升级。在风险应对方面，构建完善及时的响应机制，保证能为用户提供最快速的解决方案，最大限度减少客户损失，则是企业需要提供的服务之一。

IDC Custom Solutions

国际数据公司（IDC）是在信息技术、电信行业和消费科技领域，全球领先的专业的市场调查、咨询服务及会展活动提供商。IDC 帮助 IT 专业人士、业务主管和投资机构制定以事实为基础的技术采购决策和业务发展战略。IDC 在全球拥有超过 1100 名分析师，他们针对 110 多个国家的技术和行业发展机遇和趋势，提供全球化、区域性和本地化的专业意见。在 IDC 超过 50 年的发展历史中，众多企业客户借助 IDC 的战略分析实现了其关键业务目标。IDC 是 IDG 旗下子公司，IDG 是全球领先的媒体出版，会展服务及研究咨询公司。

IDC

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

@idc　@idc　idc.com

AI人工智能产业链联盟

#每日为你摘取最重要的商业新闻#

更新 · 更快 · 更精彩

Zero

AI音乐创作人
水墨动漫联盟创始人
百脑共创联合创始人
人工智能产业链联盟创始人
中关村人才协会秘书长助理
河北北大企业家分会秘书长
墨玫星辰智能科技有限公司CEO
河北清华发展研究院智能机器人中心线上负责人
中关村人才协会数字体育与电子竞技专委会秘书长助理

主要业务:AI商业化答疑及课程应用场景探索，各类AI产品学习手册，答疑及课程

欢迎扫码交流

提供：学习手册/工具/资源链接/商业化案例/行业报告/行业最新资讯及动态

人工智能产业链联盟创始人

邀请你加入星球，一起学习

人工智能产业链联盟报告库

星主：人工智能产业链联盟创始人

每天仅需0.5元，即可拥有以下福利！
每周更新各类机构的最新研究成果。立志将人工智能产业链联盟打造成市面上最全的AI研究资料库，覆盖券商、产业公司、研究院所等...

知识星球

微信扫码加入星球