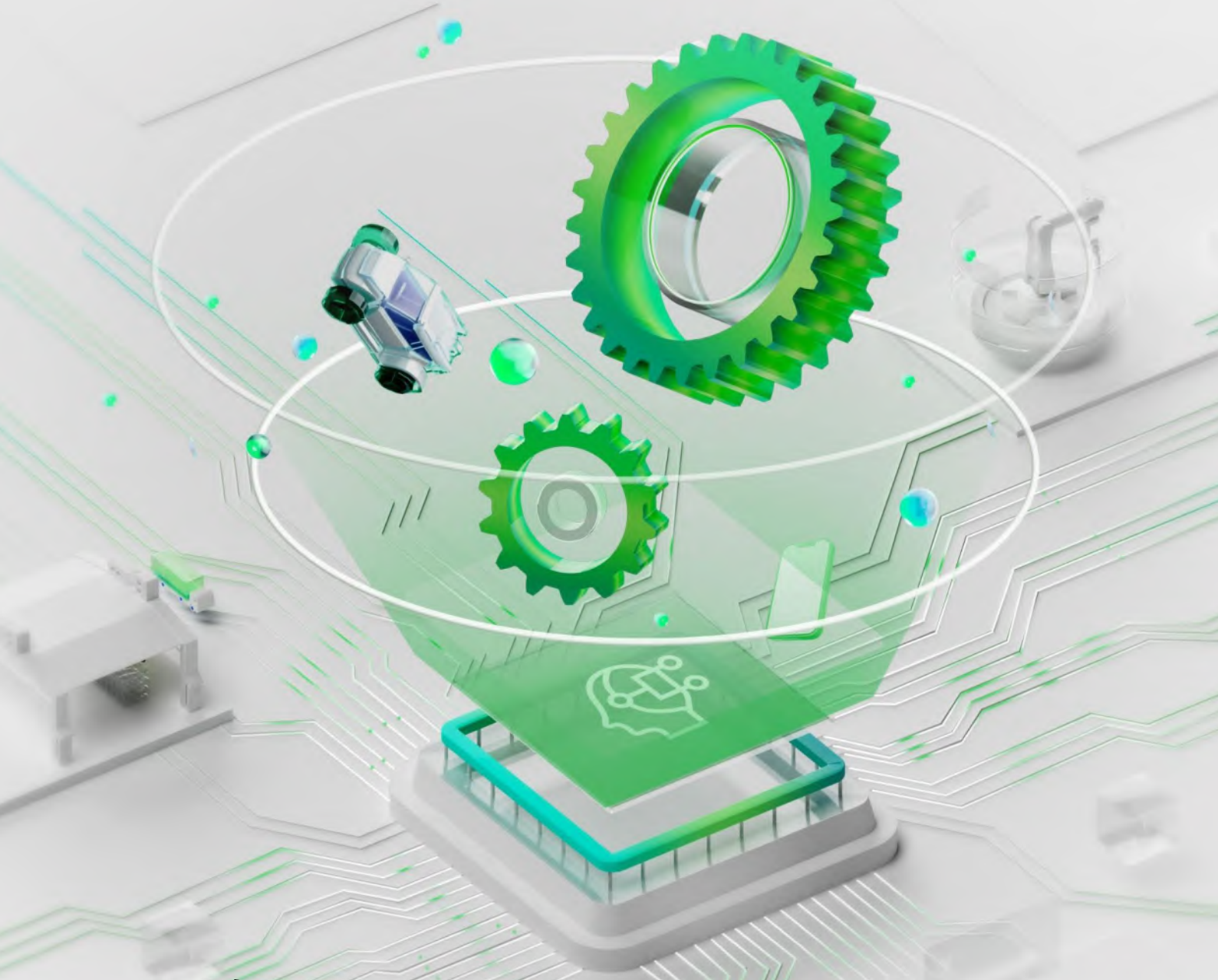
工业大模型应用报告



2024 年 3 月

参与单位

指导单位： 中国通信工业协会

撰写单位： 腾讯研究院

中国通信工业协会物联网应用分会

毕马威企业咨询（中国）有限公司

腾讯云智慧行业五部

目录

1. 大模型为工业智能化发展带来新机遇

1.1. 大模型开启人工智能应用新时代

大模型引领人工智能技术创新和应用。自 1956 年达特茅斯会议（Dartmouth Conference）提出人工智能的概念以来，人工智能技术经历了多个发展高峰和低谷。在这一长期的发展过程中，人工智能技术不断演进，逐步朝着更高的智能水平和适应性发展。2022 年 11 月 30 日，OpenAI 发布了 ChatGPT，引发了行业热潮，直至今日，业界普遍认为，大模型时代已经到来，也象征着人工智能开启了迈向通用人工智能（AGI, Artificial General Intelligence）的新阶段。在大模型出现之前，人工智能通常需要针对特定的任务和场景设计专门的算法，这种方法虽然在特定领域有效，但人们对“智能”的期望是能够适应多种任务和场景的智能系统，单一任务的人工智能系统已经无法满足这些更广泛的需求。大模型能够跨越传统人工智能的局限性，理解和推理的能力有了极大的飞跃，同时也提高了复用的效率，为人工智能技术在更多领域的应用提供了坚实的基础，推动人类社会迈向通用人工智能（AGI）的新阶段。

通用性和复用性是大模型的关键价值。2017 年，Google Brain（谷歌大脑）团队在其论文《Attention Is All You Need》中创造性地提出 Transformer 架构，凭借注意力机制，极大地改善了机器学习模型处理序列数据的能力，尤其是在自然语言处理（NLP）领域。Transformer 架构的出现，为后续的大模型如 ChatGPT 等奠定了技术基础。ChatGPT、Bert 等大模型通过海量数据和庞大的计算资源支持，使得模型具备了强大的通用性和复用性。大模型可以被广泛应用于自然语言处理、计算机视觉、语音识别等领域的各种任务，能够为各种应用和开发人员提供共享的基础架构，并进一步通过微调、零样本学习等方式，直接在一系列下游任务上使用，取得一定的性能表现，支持不同行业、不同场景的应用构建。

大模型展现出三大基础特征。目前大模型并没有明确的定义，狭义上指大语言模型，广义上则指包含了语言、声音、图像等多模态大模型。如李飞飞等人工智能学者所指出，这些模型也可以被称为基础模型（Foundation Model）。我们认为，大模型主要具备以下三大特征：

参数规模大：大模型的参数规模远大于传统深度学习模型。大模型发展呈现“规模定律”（Scaling Law）特征，即：模型的性能与模型的规模、数据集大小和训练用的计算量之间存在幂律关系，通俗而言就是“大力出奇迹”。不过“大”并没有一个绝对的标准，而是一个相对概念。传统模型参数量通常在数万至数亿之间，大模型的参数量则至少在亿级以上，并已发展到过万亿级的规模。如 OpenAI 的 GPT-1 到 GPT-3，参数量从 1.1 亿大幅拉升到 1750 亿，GPT-4 非官方估计约达 1.8 万亿。

泛化能力强：大模型能够有效处理多种未见过的数据或新任务。基于注意力机制（Attention），通过在大规模、多样化的无标注数据集上进行预训练，大模型能够学习掌握丰富的通用知识和方法，从而在广泛的场景和任务中使用，例如文本生成、自然语言理解、翻译、数学推导、逻辑推理和多轮对话等。大模型不需要、或者仅需少量特定任务的数据样本，即可显著提高在新任务上的表现能力。如 Open AI 曾用 GPT-4 参加了多种人类基准考试，结果显示其在多项考试中成绩都超过了大部分人类（80%以上），包括法学、经济学、历史、数学、阅读和写作等。

支持多模态：大模型可以实现多种模态数据的高效处理。传统深度学习模型大多只能处理单一数据类型（文本、语音或图像），大模型则可以通过扩展编/解码器、交叉注意力（Cross-Attention）、迁移学习（Transfer learning）等方式，实现跨模态数据的关联理解、检索和生成。多模态大模型（LMMs，Large Multimodal Models）能够提供更加全面的认知能力和丰富的交互体验，拓宽 AI 处理复杂任务的应用范围，成为业

界探索迈向通用人工智能的重要路径之一。典型如 OpenAI 的 Sora 模型推出，掀起了全球多模态大模型的发展新热潮。

1.2. 大模型有望成为驱动工业智能化的引擎

人工智能推动工业智能化发展进入新阶段。工业发展是一个逐步演进的过程，经历了机械化、电气化、自动化、信息化的阶段后，当前正处于从数字化向智能化迈进的阶段。每个阶段都是工业与各类创新技术的融合，对传统制造业进行升级和改造，提高生产效率、降低成本、提升产品质量。当前阶段，工业领域积累了大量的数据、基础能力和场景需求，为工业场景与人工智能技术的融合提供了基础条件。而人工智能逐渐展现出类似人的理解和分析能力，这些能力与工业场景的融合，将智能化带入到工业生产、运营、管理等领域，不断提升感知、认知和决策等多个环节，有望推动工业发展走向“自适应、自决策、自执行”的智能化阶段。

大模型为工业智能化提供新动力。尽管人工智能在智能制造、工业 4.0、工业互联网等方面有所应用，但这些应用往往受限于特定任务，难以实现跨领域、跨场景的融合创新。过去，人工智能在工业的应用主要聚焦于如质量检测、预测性维护等单一功能，这形成了人工智能应用上“一场景一训练一模型”的局限，尚未形成大规模的应用。然而，大模型的崛起有望带来“基础模型+各类应用”的新范式。大模型凭借其卓越的理解能力、生成能力和泛化能力，能够深度洞察工业领域的复杂问题，不仅可以理解并处理海量的数据，还能从中挖掘出隐藏在数据背后的规律和趋势。此外，区别于传统的人工智能模型只能根据已有数据进行预测和推断，大模型则能够生成新的知识和见解。最后，大模型的泛化能力能够在更广泛的工业场景发挥作用。

大模型为工业智能化拓展新空间。人工智能在工业领域的应用，尽管已经取得了一些显著的成果，但整体来看，其应用的普及率仍然处于相对较低的水平。据凯捷

（Capgemini）统计数据显示，即便是欧洲顶级的制造企业，其 AI 应用的普及率也仅超过 30%，而日本和美国制造企业的 AI 应用率分别达到了 30%和 28%。相较于这些发达国家，中国制造企业 AI 普及率尚不足 11%，显示出这一领域巨大的发展潜力和广阔的空间。相较于以往的小模型，大模型有望挖掘工业领域人工智能应用的新场景，提升人工智能应用的普及率。例如在研发设计领域，大模型能够深度挖掘和分析海量数据，为产品设计提供更为精准和创新的思路。在经营管理领域，大模型能够实现对生产流程、供应链管理等各个环节的监控和智能优化，从而提升企业的运营效率和市场竞争力。

1.3. 大模型应用落地需要深度适配工业场景

大模型的优势在于其强大的泛化能力，可以在不同的领域和任务上进行迁移学习，而无需重新训练。但无法充分捕捉到某个行业或领域的特征和规律，也无法满足某些特定的应用场景和需求，在真正融入行业的过程中，需要适配不同的工业场景，其核心就是要解决以下三个问题。

不懂行业：大模型在处理特定行业任务时，往往表现出对行业知识、术语、规则等的不理解，导致生成的解决方案与实际需求存在偏差，这主要是由于大模型在训练过程中缺乏特定行业的数据和知识，难以覆盖各个行业的专业细节。这种行业知识的匮乏使得大模型在应对工艺流程优化、设备故障预测等专业问题时有所缺陷，难以提供精确、可靠的解决方案，无法满足工业现场的个性化要求。

不熟企业：当大模型接入企业系统时，往往难以全面理解企业的业务流程、数据结构和运营模式，导致生成的解决方案与企业实际需求不匹配。每个企业都有其独特的运营环境和业务需求，而大模型往往缺乏对企业特定环境的深入理解。此外，企业内部的数据孤岛和碎片化的信息系统也增加了大模型理解企业环境的难度。这种不熟

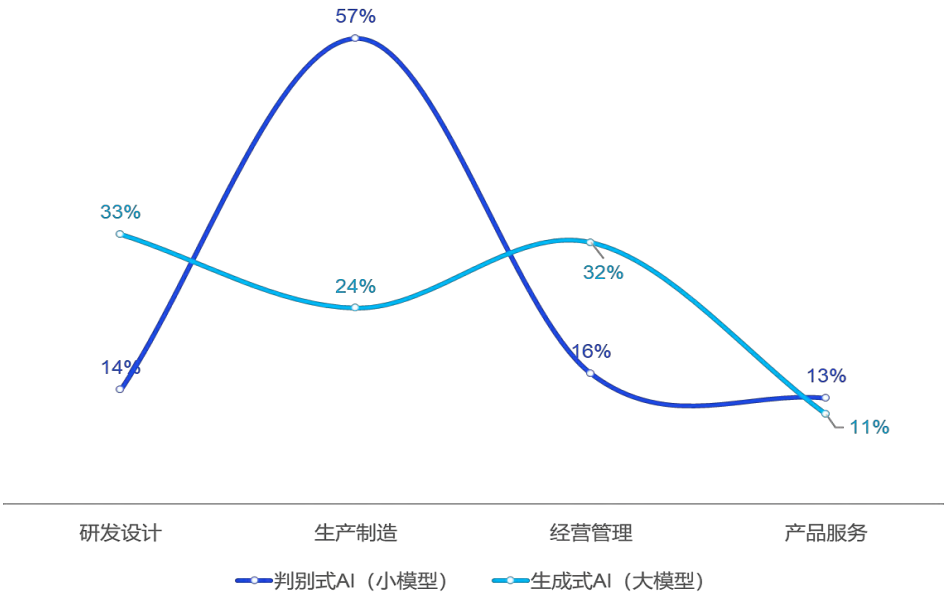
企业的问题使得大模型难以真正融入企业的运营流程，无法平滑地嵌入到现有的各类系统中。

存在幻觉：在某些情况下，大模型会产生与实际情况不符的错误输出，即“幻觉”现象。这主要是由于模型在训练过程中受到了噪声数据、偏差样本等因素的影响，或者由于模型的复杂性和数据维度过高导致过拟合。这种幻觉现象对工业领域的影响是全方位的，无论是生产过程中的质量控制、设备维护，还是供应链管理、市场预测等环节，错误的输出都可能导致严重的决策失误和经济损失。特别是在对安全性、可靠性要求极高的工业场景中，如航空航天、核能等领域，幻觉现象可能带来灾难性的后果。

2. 大模型和小模型在工业领域将长期并存且分别呈现 U 型和倒 U 型分布态势

从工业智能化的发展历程可以看出，在大模型出现之前，人工智能技术在工业领域已有较多应用。在前期阶段，工业人工智能的应用主要是以专用的小模型为主，而大模型开启了工业智能化的新阶段。结合两者不同的技术特点和应用能力，目前在工业领域形成了不同的分布态势。

图表 1 生成式 AI（大模型）和判别式 AI（小模型）在工业主要领域分布情况[1]



2.1. 以判别式 AI 为主的小模型应用呈现倒 U 型分布

根据中国信通院[2]对 507 个 AI 小模型应用案例的统计分析，这些应用主要集中在生产制造领域，占比高达 57%，而在研发设计和经营管理领域的应用则相对较少。这种分布呈现出明显的倒 U 型。

小模型的核心特点是学习输入与输出之间的关系。小模型通过学习数据中的条件概率分布，即一个样本归属于特定类别的概率，再对新的场景进行判断、分析和预测。它的优点是通常比大模型训练速度更快，而且可以产生更准确的预测结果，尤其适用

[1] 507 个小模型应用数据引用自中国信息通信研究院《工业智能白皮书（2022）》，99 个大模型应用数据由本文编写组收集、整理、统计分析所得
[2] 中国信息通信研究院《工业智能白皮书（2022）》

于对特定任务进行快速优化和部署的场景。以工业质检领域为例，小模型能够从海量的工业产品图片数据中，学习到产品的外观特征、质量标准和缺陷模式等关键信息。当面对新的样本时，小模型能够迅速判断样本是否合格，从而实现对产品质量的快速检测。同样在设备预测性维护方面，小模型通过对设备运行数据的分析，能够学习到设备正常运行的模式和潜在的故障特征。一旦监测到异常情况，小模型能够及时发出预警，提醒工作人员进行检修或维护。

小模型的能力更适合工业生产制造领域。首先，小模型能够基于有限数据支撑精准的判别和决策，而生产过程需要针对不同场景进行精准的分析和决策，这两者间的契合使得小模型在生产制造领域具有独特的优势。其次，生产制造过程对准确性和稳定性有着极高的要求，任何微小的误差都可能导致产品质量下降或生产线停工。小模型在训练过程中，能够针对具体场景进行精细化的调整和优化，从而确保模型的准确性和稳定性，这使得小模型在生产制造领域的应用更为可靠和有效。最后，小模型在成本投入方面相对较低，使得其在生产制造现场的应用更具经济性，并在有限的硬件条件下，能够稳定运行。

小模型的定制化需求制约了其进一步渗透。尽管小模型在生产制造领域表现出色，但其应用过程中也面临着一些挑战。以判别式 AI 为代表的小模型通常需要依靠个性化的业务逻辑进行数据采集、模型训练与调优，往往只能处理单一维度的数据。这一过程不仅消耗大量的算力和人力，而且训练后的模型往往无法在多行业通用。例如，工业缺陷检测领域的视觉模型往往需要针对一个产品或者一个设备训练一个模型，产品或设备不同，就要对模型进行重新训练，这种定制化的需求在一定程度上制约了小模型在工业领域的进一步渗透。

2.2. 以生成式 AI 为主的大模型应用呈现 U 型分布

根据对 99 个工业大模型应用案例的统计分析，大模型在研发设计和经营管理领域的应用相对更多，整体上呈现出 U 型分布。这表明大模型当前的能力更适配于研发设计和经营管理，在生产制造环节的能力和性能还需进一步提升。

大模型通过构建庞大的参数体系来深入理解现实世界的复杂关系。大模型的核心在于学习数据中的联合概率分布，即多个变量组成的向量在数据集中出现的概率分布，进而通过使用深度学习和强化学习等技术，能够生成全新的、富有创意的内容。与传统的数据处理方法不同，大模型并不简单地区分自变量与因变量，相反，它致力于在庞大的知识数据库中提炼出更多的特征变量。这些特征变量不仅数量庞大，而且涵盖了多个维度和层面，从而更全面地反映现实世界的复杂关系。以自然语言处理为例，大模型通过学习大量的文本数据，能够掌握语言的规律和模式。当给定一个句子或段落时，大模型能够基于联合概率分布生成与之相关的新句子或段落。这些生成的内容不仅符合语法规则，而且能够保持语义上的连贯性和一致性。此外，大模型还能够根据上下文信息理解并回答复杂的问题，展现出强大的推理和创造能力。

大模型更适合综合型和创造类的工业场景。在综合型工业场景中，由于涉及到多个系统、多个流程的协同工作，需要处理文档、表格、图片等多类数据，变量之间的关系往往错综复杂，难以用传统的分析方法进行精确描述。大模型通过深度学习和复杂的网络结构，可以捕捉并模拟这些关系，从而实现对复杂系统的全面理解和优化。在创造类工业场景中，大模型的优势体现在其强大的内容生成能力上。例如，在产品外观设计方面，传统的设计方法往往依赖于设计师的经验和灵感，设计周期长且难以保证设计的创新性。而大模型通过学习大量的设计数据，能够掌握设计领域的规律和模式，进而生成符合要求的全新设计内容，提升产品设计的效率和质量。

大模型在工业领域的应用潜力仍有待释放。首先，大模型技术本身正处于快速发展的阶段，尽管已取得了显著进步，但在成本、效率和可靠性等方面仍有待进一步提升，以适应工业领域日益复杂的需求。其次，工业场景众多且各具特色，大模型作为新技术，需要逐步与各个工业场景紧密结合，在逐步提升技术渗透率的过程中，挖掘可利用的场景，并根据行业特定需求提供定制化的解决方案。最后，工业领域自身的数据分散且缺少高质量的工业数据集，同时在实际生产中如何确保工业数据的隐私和安全也是企业关注的重点，这些现实问题也限制了大模型的推广应用。

2.3. 大模型与小模型将长期共存并相互融合

目前大模型在工业领域还未呈现出对小模型的替代趋势。尽管以生成式 AI 为代表的大模型被视为当前 AI 的热点，但在工业领域的实际应用中，大模型的能力和成本问题导致其尚不能完全取代以判别式 AI 为代表的小模型。一方面，小模型在工业领域具有深厚的应用基础和经验积累，其算法和模型结构相对简单，易于理解和实现，其稳定性和可靠性得到了验证。另一方面，大模型在成本收益比、稳定性和可靠性等方面存在问题，其在工业领域的探索还处在初级阶段。小模型以其高效、灵活的特点，在特定场景和资源受限的环境中发挥着重要作用；而大模型则以其强大的泛化能力和处理复杂任务的优势，在更广泛的领域展现着巨大潜力，两者将长期共存。

大模型与小模型有望融合推动工业智能化发展。对于小模型而言，利用大模型的生成能力可以助力小模型的训练。小模型训练需要大量的标注数据，但现实工业生产过程可能缺少相关场景的数据，大模型凭借强大的生成能力，可以生成丰富多样的数据、图像等。例如，在质检环节，大模型可以生成各种可能的产品缺陷图片，为小模型提供丰富的训练样本，从而使其能够更准确地识别缺陷和异常。此外，大模型可以通过 Agent 等方式调用小模型，以实现灵活性与效率的结合。例如，在某些场景下，大

模型可以负责全局的调度和决策，而小模型可以负责具体的执行和控制。这样既能保证系统的整体性能，又能提高响应速度和灵活性。

3. 工业大模型应用的三种构建模式

大模型的构建可以分为两个关键阶段，一个是预训练阶段，一个是微调阶段。预训练主要基于大量无标注的数据进行训练，微调是指已经预训练好的模型基础上，使用特定的数据集进行进一步的训练，以使模型适应特定任务或领域。针对工业大模型，一是可以基于大量工业数据和通用数据打造预训练工业大模型，支持各类应用的开发。二是可以在基础大模型上通过工业数据进行微调，适配特定工业任务。三是可以在不改变模型参数的情况下，通过检索增强生成（RAG）为大模型提供额外的数据，支持工业知识的获取和生成。这三种模式并不独立应用，往往会共同发力。

图表 2 工业大模型应用的三种构建模式对比

	预训练工业大模型	微调	检索增强生成
数据需求	无标注及标注的工业数据，静态数据	标注的工业数据为主，静态数据	外挂行业数据库，动态数据
特点	具备部分工业领域的通用理解能力	适用于工业领域的具体任务	不改变模型快速接入行业信息
优点	对工业通用知识的理解	精准执行工业特定任务	快速利用外部信息资源，减少幻觉
缺点	成本较高，缺乏对特定任务的优化能力	泛化能力较弱，可能过拟合	不具备对行业的深度理解能力
适用场景	作为基础模型支持多种工业应用的开发	借助高质量的标注数据实现特定任务	快速结合数据库进行信息检索和输出

3.1. 模式一：预训练工业大模型

无监督预训练主要利用大量无标注数据来训练模型，目的是学习数据的通用特征和知识，包括 GPT-3/GPT-4、LLaMA1/LLaMA2 等，都是通过收集大量无标注的通用数据集，使用 Transformer 等架构进行预训练得到。预训练之后的模型已经足够强大，能够使用在广泛的任务领域。例如，当无监督预训练技术应用于 NLP 领域时，经过良好训练的语言模型可以捕捉到对下游任务有益的丰富知识，如长期依赖关系、层次关系等。然而，另一方面完全基于互联网等通用数据训练的大模型缺乏对行业知识的理解，在应对行业问题上表现出的性能较差，因此在预训练阶段可以使用通用数据加行业数据进行模型训练，使得在基础模型的层面就具备了一定的行业专有能力。

无监督预训练工业大模型的优点是可以具备广泛的工业通用知识，最大程度地满足工业场景的需求，实现模型的最优性能与稳定性。但这一模式的缺点是需要大量的高质量工业数据集，以及庞大的算力资源，对成本和能力的要求较高，面临技术和资源的巨大挑战。在最终应用前，无监督预训练工业大模型与 GPT3 类似，同样需要通过适当的指令微调、奖励学习、强化学习等阶段，形成面向最终场景的应用能力。

SymphonyAI [3]推出了基于无监督预训练的工业大语言模型，该模型的训练数据包含 3 万亿个数据点，12 亿 token，能够支持机器状况诊断，并回答故障状况、测试程序、维护程序、制造工艺和工业标准相关的问题。

制造流程管理平台提供商 **Retrocausal**[4]发布的 LeanGPT™，也采用了无监督预训练的模式，是制造领域的专有基础模型。基于 LeanGPT™这一基础模型，Retrocausal 还推出了 Kaizen Copilot™的应用程序，可以帮助工业工程师设计和持续改进制造装配流程。

3.2. 模式二：微调

微调模式是在一个已经预训练完成的通用或专业大模型基础上，结合工业领域特定的标注数据集进行进一步的调整和优化，从而使模型能够适应具体的工业场景需求，更好地完成工业领域的特定任务。在微调期间，需要使用特定任务或领域量身定制的标记数据集来训练，与模型预训练所需的巨大数据集相比，微调数据集更小，单个任务的微调通常只需要几千条到上万条有标注数据即可。通过微调，大模型可以学习到工业细分领域的知识、语言模式等，有助于大模型在工业的特定任务上取得更好的性能。在当前主流的行业大模型构建路线中，众多行业模型都是使用基础大模型+行业标注数据集来微调得到的。

[3] Industrial LLM - SymphonyAI with Microsoft
[4] Kaizen Copilot – Retrocausal

这一模式的优点是可以充分利用基础大模型的泛化能力，同时通过微调的方式，提升模型的适配性和精度，能够在特定的任务或领域上取得更好的效果，也可以针对具体行业或公司的语气、术语进行定制化。缺点在于需要收集和标注具体工业领域和场景的数据和知识，增加数据准备的成本和时间，若数据不足或嘈杂会降低模型的性能和可靠性，也可能会过度拟合，导致性能下降，或者灾难性遗忘。

Cohere[5]推出全面的微调套件，其中包括生成微调、聊天微调、重新排序微调和多标签分类微调等解决方案，可以满足企业在微调各种 AI 应用时的需求。基于微调，企业可以定制模型，在文本生成、摘要、聊天、分类和企业搜索等目标用例上获得更好的性能。

3.3. 模式三：检索增强生成

检索增强生成模式是指在不改变模型的基础上，结合行业领域的数据、知识库等，为工业场景提供知识问答、内容生成等能力。检索增强生成（Retrieval Augmented Generation，RAG）结合了检索（Retrieval）和生成（Generation）两种方法，基本思路是把私域知识文档进行切片，向量化后续通过向量检索进行召回，再作为上下文输入到基础大模型进行归纳总结。具体而言，首先是将外部数据通过 Embedding 模型存储到向量数据库中。当用户输入查询内容时候，经过 Embedding 模型和向量数据库的内容匹配，得到 Top 排序的结果作为上下文信息一起输入给大模型，大模型再进行分析和回答。检索增强生成在私域知识问答方面可以很好的弥补通用大语言模型的一些短板，解决通用大语言模型在专业领域回答缺乏依据、存在幻觉等问题。

这种模式的优点是可以快速利用现有的基础大模型，无需进行额外的训练，只需要构建和接入行业或企业私有的知识库，就可以实现对工业领域的知识理解和应用，

[5] Introduction (cohere.com)

也可以部分消除大模型的幻觉，减少数据泄露，提高信任度和访问控制。这种模式的缺点是基础大模型可能无法充分适应工业场景的特点和需求，导致效果不佳或不稳定。

Cognite[6]利用检索增强生成（RAG）技术，将大模型与其工业 DataOps 平台 Cognite Data Fusion 结合起来，为工业客户提供基于数据的洞察和解决方案。通过将不同来源和类型的工业数据进行向量化，并存储在一个专门的向量数据库中，可以作为 RAG 的检索源，与用户的自然语言提示一起输入到大模型中，使模型能够提供更加精准的建议或解决方案。

C3.AI[7]推出的 Generative AI 利用检索增强技术，将制造企业知识库与大语言模型分开，从而生成准确、一致的结果，且能够追溯到源文件和数据，以确保信息的准确。另外，Generative AI 还通过嵌入相关性评分机制，在未达到相关性阈值时回答 “我不知道” 。例如在设备运维场景下，操作员可以利用简化的工作流程来诊断设备故障根因。当操作员发现生产问题时，可以直接进入 C3 Generative AI 搜索故障排除指南和教科书，以找出潜在原因。

3.4. 三种模式综合应用推动工业大模型落地

在工业大模型的训练模式中，我们可以看到三种主要的方法，每种方法都有其独特的优势和挑战。无监督预训练模式通过大量无标注数据来提升模型的泛化能力，适用于工业场景的广泛需求，但需要巨大的资源投入。基础大模型加有监督微调模式则在保留通用能力的同时，通过特定领域的数据微调，提高了模型的适配性和精度，但需要高质量的标注数据。基础大模型结合检索增强生成（RAG）模式，通过利用预训练的基础大模型和行业知识库，为工业场景提供即时的知识问答和内容生成服务，这种方法的优势在于快速部署和利用现有资源，但可能在特定工业场景的适应性上存在

[6] RAG is all the RAGe (cognite.com)
[7] C3 Generative AI Now Publicly Available on Google Cloud Marketplace - C3.ai, Inc.

局限。总结来说，这三种训练模式为工业大模型的开发提供了多样化的选择，在实际应用中，这三种模式并非只采取一种方式，往往企业最终发布的应用模型针对不同的应用场景，综合采用多种构建方式。

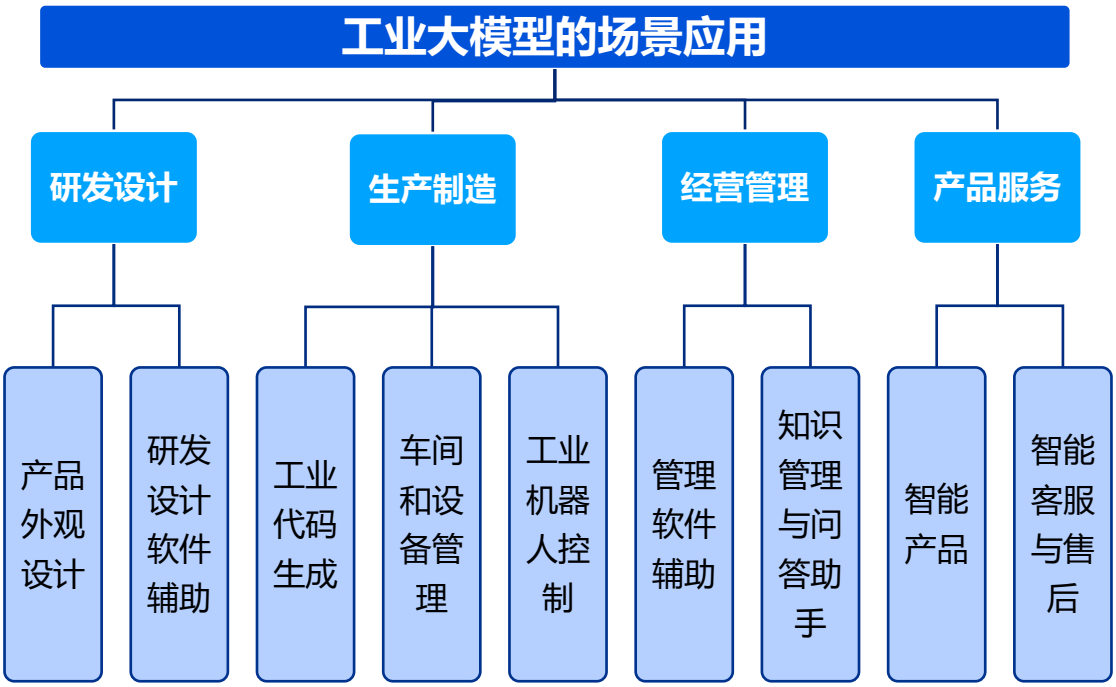
以 **NVIDIA**[8]（英伟达）为例，开发了名为 ChipNeMo 的定制大模型，采用了无监督预训练、微调等多种模式。该模型训练收集了 Bug 总结、设计源（Design Source）、文档以及维基百科等数据，训练的 token 超过 240 亿，在商用开源的 Llama2 基础上，采用领域自适应预训练、带有领域特定指令的监督微调（SFT），以及领域自适应检索等技术对模型进行优化，能够有效的支持芯片设计的一般问题问答、总结 Bug 文档和 EDA 脚本编写等功能。

[8] ChipNeMo: Domain-Adapted LLMs for Chip Design | Research (nvidia.com)

4. 大模型应用探索覆盖工业全链条

从工业产品生命周期的角度，可以将工业场景概括为研发设计、生产制造、经营管理、产品服务等四个主要环节，根据整理的 99 个工业大模型的应用案例，对工业大模型的场景应用总结如下：

图表 3 大模型在工业全链条的应用探索



4.1. 大模型通过优化设计过程提高研发效率

4.1.1 产品外观设计

工业产品设计涵盖了外观设计与结构设计两大关键环节。在这两个环节，大模型都展现出了其独特的价值。在结构设计方面，借助大模型的生成能力可以快速为工业产品或零件提供多种设计方案，缩短产品开发的时间并提供多种创造性的产品选项，让设计师专注于产品设计的核心工作。在外观设计方面，设计师只需提供简短的文字

描述或草图，大模型便能迅速生成多张高保真度的设计效果图。这些效果图不仅满足了设计师的个性化需求，还为他们提供了丰富的选择空间，方便进一步修改与优化。

CALA[9]作为时装设计平台，将 OpenAI 的 DALL·E 生成式设计工具整合到其服务体系中，极大地促进了设计师创意的快速实现。通过输入相关的设计概念关键词，CALA 能够迅速产生一系列的服装设计初稿，显著地缩短了设计周期，提高了工作效率。然而，CALA 并非一个完全自动化的设计工具，其使用过程依然依赖于设计师的专业技能和丰富经验。尽管如此，CALA 显著降低了新设计师的入门难度，并有效提升了资深设计师的工作效率。

海尔设计[10]联合亚马逊云科技以及合作伙伴 Nolibox 共同开发的 AIGC 解决方案，将大模型图像生成技术成功应用于产品设计、用户界面设计、色彩材质设计以及品牌设计等多个领域。该解决方案全面覆盖了新品设计、产品改款升级、以及渠道定制化等工业设计业务场景。其中，概念图的生成得益于 Nolibox 基于开源大模型 Stable Diffusion 的应用开发，有效地助力形成高效、精准的设计成果。

丰田研究所[11]推出的“生成式人工智能工具”是一款专为车辆设计师打造的 AI 助手，旨在提供创新支持。这款工具能够根据文本提示生成精确的设计草图，并允许设计师通过调整定量性能指标来构建原型草图。工具融合了计算机辅助工程的优化理论与生成式 AI 技术，能够将工程约束自然地融入设计流程中。这意味着，在生成满足设计师风格要求的图像的同时，还能综合考虑并优化诸如风阻、底盘高度等关键工程参数。

4.1.2 研发设计软件辅助

[9] CALA · AI-Powered Design & Collaboration
[10] 亚马逊云科技联手 Nolibox，助力海尔创新设计中心打造 AIGC 工业设计解决方案 (amazon.com)
[11] Human-Centered AI | Toyota Research Institute (tri.global)

大模型可以与 CAD、CAE 等工业设计软件结合，通过连接相关数据库，更好地调用相关的设计模块，提升研发设计的效率。以 CAD 为例，现有的海量标准化素材库提供了大量工程制图、布局规划等数据，大模型可以利用这些数据，结合设计者的创意思路和特殊需求，生成多样化的设计方案，供设计者进行参考。另一方面，亦可对设计方案进行快速优化调整，帮助工程师以更快的速度和更少的错误创建布局。

Back2CAD[12] 基于 Elaine CAD Bot、ChatGPT 和 Amazon AWS 等的支持推出 CADGPT™，支持虚拟助手、智能推荐、文档生成、代码生产、CAD 项目辅助等各类功能。以虚拟助手为例，CADGPT 能够基于用户前期的设计和偏好，提出替代性的方案或者现有方案的改进意见，帮助用户短时间内能够获得更好的设计结果。在代码生成方面，CADGPT 可基于用户输入的提示词生成适当的代码片段。

Synopsys[13]（新思科技）推出了一款创新的芯片设计辅助工具——Synopsys.ai Copilot。这款工具融合了先进的生成式人工智能技术，旨在加速芯片设计的整个流程。新思科技与微软合作，整合了 Azure OpenAI 平台的生成式 AI 技术，使得设计工具具备了与工程师进行对话的智能能力。 在日常工作中，工程师可以利用 Synopsys.ai Copilot 来应对芯片设计过程中遇到的各种复杂挑战。通过与工具的智能对话，工程师能够更加高效地解决问题，优化设计方案，从而显著提高设计效率。

Cadence[14]推出了 Cadence.AI LLM，这是业界首个针对芯片设计的大型语言模型（LLM）技术。该工具的核心功能在于加载和处理架构规范、设计规范、集成连接规范以及芯片设计本身，为用户提供了一个强大的交互平台。用户能够通过自然语言与工具进行互动，提出各种指令，如要求列出芯片设计中的不规则网络名称、识别所有

[12] CADGPT | Back2CAD (backtocad.com)
[13] 面向芯片设计和 AI 应用的 AI 驱动型 EDA 套件 | Synopsys.ai
[14] Cadence Creates Industry's First LLM Technology for Chip Design - Cadence Community

潜在的不规则引脚、自动化测试平台的连接设置、以及辅助完成工具脚本和 RTL 代码的编写。

Ansys[15]推出 Ansys SimAI™，一款将 Ansys 仿真软件与生成式人工智能结合的云端应用，可以快速评估新设计的性能。Ansys SimAI 并不依赖于几何参数来定义设计，而是以设计本身的形状作为输入，即使形状的结构在训练数据中的记录不一致，也能有助于更广泛的设计探索。对于需要进行海量计算的项目，该应用可将所有设计阶段的模型性能预测功能提高 10-100 倍。客户可以使用以前生成的 Ansys 或非 Ansys 的数据来训练 AI。雷诺集团利用 Ansys SimAI，加速了汽车零部件的设计和测试过程， Ansys SimAI 可以让雷诺集团的工程师在数分钟内测试一个设计，并迅速分析结果，从而在项目的上游阶段探索更多的技术可能性，并加快产品整体上市进程。

4.2. 大模型拓展生产制造智能化应用的边界

生产制造环节是工业生产的核心场景，对安全性和稳定性的要求较高，目前大模型在该环节的渗透率整体不高，主要集中在代码生成、车间和设备管理和机器人控制等环节。

4.2.1 工业代码生成

大模型在工业代码生成的应用领域广泛，涉及到自动化、机械加工等领域。将大模型应用于工业代码生成的优势在于可以提高工业代码的质量和效率，减少人工编程的时间和成本，提高了研发者的开发效率，特别是重复性高、逻辑简单的任务。同时，自动生成的代码还可以减少人为错误的发生，提高代码的可靠性和可维护性。现有的代码生成方法或工具在处理简单需求的场景时表现较好，如行级代码补全和初级的函

[15] Ansys 宣布正式推出 Ansys SimAI™ | ansys.com

数级代码生成。在复杂的函数级代码生成、深入的问题分析和软件系统设计方面，还需要进一步改进和优化。

Siemens[16]与微软合作推出了 Siemens Industrial Copilot，西门子 Industrial Copilot 允许用户迅速生成、优化自动化代码并加速仿真流程，将原本需要数周的任务缩短至几分钟。该工具整合了西门子 Xcelerator 平台的自动化与仿真信息，并结合微软 Azure OpenAI 服务提升数据处理能力，同时确保客户对数据的完全控制，不用于 AI 模型训练。 Industrial Copilot 旨在提升整个工业生产周期的效率，通过自然语言交互，使维修人员得到精确指导，工程师能迅速使用仿真工具，从而推动工业创新和生产力的提升。

SprutCAM[17]结合 ChatGPT 推出 AI 产品 Éncy。这款 AI 助手通过结合 OpenAI 的 API 接口，能够理解和生成自然语言，帮助 CNC 工程师简化机械加工任务。Éncy 能够执行多种任务，包括生成基于文本描述的代码，以及使用 Python 编写代码来创建.dxf 或.stl 文件。此外，Éncy 还能支持工程师操作机床，回答与 SprutCAM X 软件操作相关的任何问题。例如，当工程师给出指令“在点(100, 25)处钻一个直径 10 毫米的孔”，Éncy 即可生成相应的 CAM 执行代码。

4.2.2 车间和设备管理

在车间管理方面，大模型能够协助监控生产线，确保工艺流程的顺畅与高效；在设备管理领域，大模型通过支持预测性维护减少停机时间，并通过精准的数据分析指导维护决策，有望成为工业智能化转型的关键驱动力。

[16] Unlocking the Power of Generative AI: Siemens Industrial Copilot - Insights - Siemens Global Website
[17] Éncy - virtual AI assistant for SprutCAM X - SprutCAM X

Sight Machine[18]推出 Factory CoPilot，一款集成了 Microsoft Azure OpenAI Service 的工业数据分析工具，它通过智能化分析简化了制造问题的解决和报告流程。Factory CoPilot 提供了一个直观的交互体验，用户可以像询问专家一样轻松获取分析结果。利用自然语言界面，Factory CoPilot 能够自动整理 Sight Machine 平台上的上下文数据，生成易于理解的报告、邮件和图表。它还能引导用户进行根因分析，加快问题诊断。此外，通过持续分析，Factory CoPilot 有助于识别和解决非计划停机、设备效率低下和质量问题，推动制造流程的持续优化。

Vanti[19]推出 Manufacturing COPILOT，目标是解决当前制造业专业人员在数据管理和分析方面面临的挑战。通过融合和整理来自 ERP 系统、制造执行系统（MES）、传感器以及历史记录器等多样化数据，该平台改变了数据处理方式。同时基于大模型能力，允许用户以自然语言询问并与数据互动，将复杂的数据分析过程转换为简单直观的对话。Manufacturing COPILOT 不仅能处理和分析原始数据，还能识别并解释复杂的生产行为，转化为易于理解的、可操作的洞察。借助数据可视化技术，它提供了数据的图形化叙述，增强数据的可解释性，帮助制造业专业人士进行数据驱动的决策。此外，它还简化了测试流程，使用户能够通过自然语言查询快速验证假设，并根据可靠数据进行调整，显著提高了生产效率和操作效率。

ABB[20]与 Microsoft 合作推出 ABB Ability™ Genix，是一个集成了 Microsoft Azure OpenAI 服务的工业分析和人工智能套件。它的核心功能在于提供数据分析、机器学习和用户交互的增强能力，利用生成式 AI 优化工业流程，提高操作效率和可持续性。在实际应用中，通过 Copilot 功能，操作员能够更直观地与工业系统交互，实现预测性维

[18] Factory CoPilot from Sight Machine - Generative AI
[19] GEN AI-Powered Analytics Platform for Manufacturing (vanti.ai)
[20] ABB Ability™ Genix Industrial Analytics and AI Suite

护和实时优化，从而减少停机时间，提升生产效率。根据 ABB 预计，Genix Copilot 提供的数据分析洞见有望将资产生命周期延长 20%，将设备意外停机时间减少 60%。

美国钢铁公司[21]（U. S. Steel）与 Google Cloud 合作，推出的首个基于大模型的应用程序 MineMind™，利用 Google Cloud 的 AI 技术简化设备维护过程。该应用不仅辅助维护团队进行卡车维修、订购零件、提炼复杂信息，还提供全面的参考资料以确保准确性。MineMind™预计将改善维护团队的体验，更有效地节省技术员时间和降低卡车维护的成本，目前该应用已经在 60 多辆运输车上运行。预计完全运行时，MineMind™将帮助维护技术员减少 20%以上的工作时间。

4.2.3 工业机器人控制

大模型的出现可以帮助机器逐渐实现真正像人类一样交流、执行大量任务。工业机器人和自动化工厂作为智能制造的核心载体，将作为大模型和智能制造的中间桥梁。根据微软发布的《ChatGPT for Robotics: Design Principles and Model Abilities》，目前大模型主要通过两个层面对工业机器人进行辅助，第一，作为预训练语言模型，可以被应用于人类与机器的自然语言交互。机器通过 ChatGPT 理解人类的自然语言指令，并根据指令进行相应的动作。第二，大模型可以帮助机器在执行路径规划、物体识别等任务时做出相应的决策。

RoboDK[22]推出了 RoboDK’s Virtual Assistant，一个基于大模型的 AI 应用，专为机器人编程和仿真提供智能化的支持和指导。RoboDK’s Virtual Assistant 通过与 Microsoft Azure OpenAI Service 的集成，实现了机器人数据的高效处理和分析。该应用提供了一个自然语言用户界面，使机器人开发者和使用者可以与 AI 应用进行交互，请求专家的

[21] U. S. Steel Aims to Improve Operational Efficiencies and Employee Experiences with Google Cloud's Generative AI :: United States Steel Corporation (X) (ussteel.com)
[22] Unleashing the Potential of Large Language Models in Robotics: RoboDK's Virtual Assistant - RoboDK blog

建议和指导。同时也可以协助用户完成诸如自动创建和修改机器人程序、优化机器人运动和路径、提高机器人性能和安全性等任务。RoboDK's Virtual Assistant 还可以学习公司特定的信息，如机器人型号和参数、机器人应用和场景、机器人操作和故障排除手册等，为用户提供即时的支持，回答特定的问题。例如，如何选择合适的机器人、如何设置机器人工具和工件、如何解决机器人碰撞或奇异性问题等。

梅卡曼德[23]与汉堡大学张建伟院士实验室达成合作协议，共同致力于机器人多模态大模型的研究与创新。双方正合作开发一款集成视觉、语音和语言处理功能的综合性机器人大模型。该模型旨在赋予机器人对环境中多元信号的感知与理解能力，并通过自然语言实现与人类的有效沟通。预期的研究成果将显著提高机器人的智能程度，促进其与人类更加自然地协作与互动。

斯坦福大学教授李飞飞团队[24]发布了名为"VoxPoser"的项目，该项目用大模型指导机器人如何与环境进行交互，通过将大模型接入机器人，可将复杂指令转化成具体行动规划，人类可以很随意地用自然语言给机器人下达指令，机器人也无需额外数据和训练。

4.3. 大模型基于助手模式提升经营管理水平

4.3.1 管理软件辅助

经营管理环节具备较强的通用性，因而成为大模型较容易应用的工业场景。大模型在管理软件辅助方面的应用，主要是通过自然语言交互等方式，实现对经营管理数据的智能化分析和处理。通过对 CRM、ERP、SCM 等管理软件的赋能，大模型能够提高客户关系、订单管理、供应链管理等方面的效率和质量，为企业提供更精准和个性

[23] 梅卡曼德与汉堡大学张建伟院士实验室携手探索机器人多模态大模型 - 梅卡曼德机器人 (mech-mind.com.cn)
[24] [2307.05973] VoxPoser: Composable 3D Value Maps for Robotic Manipulation with Language Models (arxiv.org)

化的服务。大模型还能够根据用户的需求或描述，自动生成报告、简报、订单等多种形式的内容，为用户提供更便捷和高效的信息服务。

Salesforce[25]推出一款名为“AI Cloud”的客户关系管理（CRM）软件，帮助客户提高生产力和效率。AI Cloud 的服务范围覆盖 Salesforce 的旗舰产品：从 Einstein 服务到工作消息应用程序 Slack 和数据分析软件 Tableau。AI Cloud 总共有九种模型：销售 GPT、服务 GPT、营销 GPT、商业 GPT、Slack GPT、Tableau GPT、Flow GPT 和 Apex GPT。销售 GPT 可以快速自动制作个性化电子邮件；服务 GPT 可以根据案例数据和客户历史创建服务简报、案例摘要和工作订单；营销 GPT 和商业 GPT 可以将受众细分，以便根据每个买家的特点定制产品描述，或者提供如何提高平均订单价值等建议。

鼎捷软件[26]子公司鼎新电脑与微软 Azure OpenAI 合作，发布了由 GPT 大模型赋能的 PaaS 平台 METIS，基于该平台推出个人智能助理（预约会议、汇总信息、催办任务、提示行程等功能）、企业知识大脑（METIS ChatFile，能够解析文件并智能分类，基于 GPT 大模型实现自然语言问答交互）、AI 辅助开发（AI 赋能需求分析、系统设计、程序开发）等多个功能。

4.3.2 知识管理与问答助手

企业知识管理与问答助手类应用已经成为大模型在企业端落地的先行场景，以知识库为代表的问答助手类应用落地广泛，通过对企业的产品、服务、流程、规范、文档等方面的图像和数据进行分析，大模型能够构建和更新企业的知识库，为企业提供全面和准确的知识管理。同时，通过对用户的需求或问题进行理解和回答，大模型能够为用户提供智能的问答和辅助，解决用户在工作中遇到的各种问题，提高用户的工作效率和满意度。

[25] AI Cloud: Trusted GPT Artificial Intelligence for Businesses - Salesforce.com US
[26] 鼎新“数智驱动·未来企业”在线峰会成功举办，重磅发布 METIS 新平台+新应用-鼎捷软件 (digiwin.com)

Andonix[27]推出了 Andi，一个专为工厂工人设计的 AI 驱动的制造聊天机器人。Andi 实现了工厂数据的智能化分析和处理，并提供了一个自然语言用户界面，使工厂工人可以与聊天机器人进行人性化的对话，请求专家的帮助和支持。Andi 可以协助工人完成诸如自动监控机器和流程性能、解决问题、生成行动计划、检查清单和工作指导等任务，还可以学习公司特定的信息，如机器操作和故障排除手册、质量系统、人力资源手册等，为工人提供即时的支持，回答特定的问题，如如何修复特定的机器故障代码、识别导致机器停机最多的三个问题、确定最近一小时的一次合格率（FTQ）或者澄清公司的假期政策等。

Hitachi[28]正在利用生成式人工智能，将维护和制造方面的专业技能传授给新员工，旨在减轻经验丰富员工退休的影响。熟练的工人利用多年经验培养的直觉，来检测可能导致事故或故障的细微异常——如设备的声音、温度或气味的变化，然而这些制造业中的隐形知识存在传递困难。日立已经开发了一个 AI 系统，可以根据工厂的三维数据，生成图像，将可能的故障——如烟雾、塌陷、轨道弯曲——投影到实际的轨道图像上，支持维护工人身临其境的学习如何检查异常。该系统有望通过让他们学习可能导致严重事故的多种问题，来提高维护工人的技能，并允许用户通过虚拟现实设备在远程地点参与培训。

4.4. 大模型基于交互能力推动产品和服务智能化

产品智能交互

在产品服务优化环节，将大模型的能力集成到产品中，也成为消费电子、汽车等领域产品智能化能力提升的探索焦点。

[27] Introducing andi - Andonix
[28] Human-AI collaboration in the industrial sector : Research & Development : Hitachi

国光电器[29]推出的智能音箱 Vifa ChatMini 内置了 ChatGPT 和 文心一言双模型，在保持了专业声学标准的基础上，与传统的智能音箱相比，Vifa ChatMini 在自然语言生成和情感表达方面具有显著的优势，可应用到老年人和儿童等特定用户群体，用于情感支持和智能学习陪伴，也可作为智能助手应用在日常工作和规划中。

BMW[30]（宝马）基于亚马逊 Alexa 大语言模型提供的生成式 AI 技术打造全新一代个人助理。 可以为驾乘人员提供更人性化的帮助，及时解答有关车辆的疑问；通过语音可实现人车智能化交互，为用户带来情感化数字体验。

Mercedes-Benz[31]（奔驰）发布了全新 MBUX 虚拟助理，奔驰表示新款 MBUX 基于 MB.OS 操作系统打造，而 MB.OS 会搭载基于 AI 和大语言模型的全新虚拟助手，能够提供更自然的语音反馈和对话。

腾讯新一代智能座舱解决方案 TAI4.0 从场景和用户体验出发，深度利用汽车的感知能力和大模型的学习理解能力，构建从多模交互到个性化服务的完整智能化闭环体验。基于插件工具、内容生态，为用户在智能交互、效率提升、亲子娱乐等场景下提供丰富的 Agent 能力，比如行程规划，生成式 UI 等。

智能客服与售后

Tana [32]（芬兰固体废料回收设备制造商）与 Solita 合作开发和集成定制的生成式 AI 辅助工具。Tana 的员工的故障排除过程通常是在大量的用户手册和旧的事件报告中寻找解决方案，文档数量庞大且复杂，因此很难找到正确的解决方案。通过使用 Azure OpenAI 服务的大型语言模型，Tana 创建了一个服务于售后团队的智能助手，针对售后

[29] CHATMINI - the world's first smart speaker equipped with ChatGPT - vifa
[30] Next generation BMW voice assistant to be based on Amazon Alexa technology. (bmwgroup.com)
[31] Mercedes-Benz heralds a new era for the user interface with human-like virtual assistant powered by generative AI (mbusa.com)
[32] Teemu Lintula & Solita - Tana's AI-powered assistant - Tana

的相关问题，智能助手将根据相关的文档给出答案，同时还引用了其答案来源的详细信息，售后团队可以自己检查实际的来源文档。

腾讯将大模型客服知识问答的 SaaS 核心能力下沉，升级为智能知识引擎 PaaS 平台，以平台能力赋能各式各样知识问答前端应用的构建。基于腾讯大模型知识引擎，比亚迪开发了 AI 语音助手应用，对其 VDS 设备（Vehicle Diagnostic System，车辆诊断系统）进行了全新升级。比亚迪维修车间的汽车维修工人，双手经常需要佩戴绝缘手套、或者沾有油污，不方便操作点击 VDS 设备。而新员工在查询汽车相关信息、维修专业知识、业务工单等方面也会存在不熟悉、缺乏业务经验等现象。智能问答机器人可以作为 VDS 内置的 AI 语音助手，只需要通过口语化的表达咨询，就可以快速实现维修知识问答，并调取相关的内容进行可视化前端呈现。

5. 工业大模型的挑战与展望

5.1. 工业大模型应用面临数据质量和安全、可靠性、成本三大挑战

挑战一：数据质量和安全是工业大模型构建的首要问题

工业数据质量参差不齐。工业领域涵盖广泛，包括 41 个工业大类、207 个工业中类、666 个工业小类，导致数据结构多样，数据质量参差不齐。此外，由于工业生产过程中的各个环节相互交织，数据之间的关联性和复杂性也较高。数据的来源、采集方式、时间戳等都会影响数据的准确性和完整性。这种数据结构的多样与质量的参差不齐给工业大模型的训练和应用带来了挑战。为了克服这一问题，需要投入大量的时间和资源进行数据清洗、预处理和校验，以确保数据的准确性和一致性。

工业数据安全要求较高。工业数据通常包含企业的核心机密和商业秘密，如工艺参数、配方、客户信息等。这些数据如果泄露或被竞争对手获取，可能会对企业的竞争力和市场地位造成严重威胁。因此，工业企业对于数据的保护和隐私关注度非常高，限制了数据的共享和流通。

挑战二：工业大模型需满足高可靠性和实时性要求

工业大模型应用对可靠性有更严格的要求。工业生产环境往往涉及复杂的工艺流程、高精度的操作控制以及严苛的安全标准。任何模型预测或决策的失误都可能导致生产事故、质量问题或经济损失。因此，在有些领域，工业大模型应用必须具备极高的准确性和稳定性，以确保在各种复杂和动态变化的工业场景中都能提供可靠的支持。

工业大模型应用还受到实时性的制约。工业生产对实时性的要求非常高，很多场景需要模型能够在毫秒级甚至微秒级的时间内做出响应。同时，由于计算资源的限制，模型的大小和计算复杂度也需要得到合理控制。这就需要在保证模型性能的同时，尽可能地降低计算复杂度和内存占用，以实现高效的实时推理。

挑战三：高额成本限制了工业大模型应用的投入产出比

工业大模型的训练和推理成本高昂。大模型通常需要庞大的数据集进行训练，而这些数据的收集、清洗和标注都需要耗费大量的时间和资源。此外，训练过程中所需的计算资源也是巨大的，包括高性能的计算集群、大量的存储空间和高速的网络连接等。

工业大模型的应用需考虑长期运营成本。除了初始的训练成本外，模型的维护和更新也是一个持续的过程。随着工业环境和数据的变化，模型可能需要进行定期的重新训练和调优以保持其性能。这不仅需要投入更多的计算资源和人力资源，还需要建立完善的模型管理体系和监控机制，以确保模型在实际应用中的稳定性和可靠性。

私有化部署的成本投入较大。对于工业应用而言，数据安全是一个重要的考虑因素。许多工业场景需要私有化部署，以确保数据不被泄露。然而，私有化部署通常需要更高的硬件、维护等成本。由于当前工业大模型仍处于初级阶段，投入产出比并不明确。企业对于在工业大模型上的成本投入可能会产生一定的困惑和担忧。

5.2. 工业大模型应用将伴随技术演进持续加速和深化

预判一：基于少量工业基础大模型快速构建大量工业 APP 满足碎片化应用需求

在通用人工智能到来之前，工业大模型的应用模式将是“基础大模型 + 工业 APP”。通过基础大模型和工业 APP 的结合，能够广泛且快速地应对工业领域的挑战，推动各类工业场景的智能化升级。基础大模型将在大量的通用数据和工业数据上进行训练，从而学习到工业领域的通用知识和模式。这些模型将凭借其在特定领域的数据基础，实现对通用大模型的超越，且能够适应不同工业场景和任务的需求。考虑到工业涉及到的领域极其庞杂，并不会以一个工业大模型解决所有问题，更可能的结果是结合行业领域特征形成数十个或数百个基础的工业大模型。工业 APP 是在工业大模型基础上快速构建的各类应用，并且针对特定的工业场景和任务进行优化和定制，这类工业 APP 的数量将达到数万个或更多。由于工业场景复杂并呈现碎片化的模式，在工业大模型基础上，企业可以快速构建符合自身业务和场景的应用，满足个性化的诉求，同时企业可以更加便捷地将各类大模型应用集成到自身原有的业务流程中，实现快速和便捷的智能化应用部署。

预判二：大模型的新突破带来工业应用的新场景

大模型技术本身仍处于发展的早期阶段，各类新的技术和应用模型不断涌现，比如长文本能力的提升、Sora 等视频生成能力的增强，将进一步扩展大模型在工业应用的场景，Agent、具身智能等大模型应用模式的创新也将深化大模型在工业领域的应用。以 Agent 为例，作为一种智能代理系统，具有自主决策和行动能力，可以与环境进行交互并实现自主学习，为工业大模型的应用提供了更加灵活和智能的解决方案。具身智能则将人工智能技术融入到实体中，使得设备和机器具备更加智能化的交互和应用能

力，为工业生产带来更高效、安全的生产方式。Sora 等新应用的出现在图像生成领域带来了新的突破，对工业领域的数字孪生等场景带来新的可能。

预判三：大模型成本的降低将加速工业领域的应用

随着大模型的不断发展和参数规模的增加，所需的计算资源和存储空间也随之增加，给模型训练和部署带来了巨大挑战。在工业领域，对成本较为敏感且应用场景复杂，对大模型的部署及成本提出了较高的要求。但业界也在探索各类模型压缩技术，在保证模型精度的同时，通过剪枝、量化、蒸馏等方式，可以有效地减少模型的参数量、计算复杂度和存储需求，从而降低了训练和推理的成本。这些技术的应用使得即使在资源受限的环境下，也能够训练和部署高效且精确的工业大模型。通过模型压缩，不仅可以降低硬件成本，还能够提高模型在移动设备、边缘计算等资源受限环境下的性能表现。综合来看，这些大模型成本降低技术将加快大模型在工业领域的渗透速度。