大模型 2.0 产业发展报告

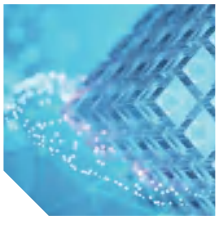
—— 商业落地创涌而现

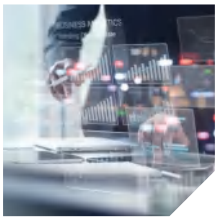
国家工业信息安全发展研究中心标准所
联想集团

2025 年 3 月

基础篇

洞向篇





目 录

【政观经纬】

窦克勤

国家工业信息安全发展研究中心
标准所副所长、研究员

大模型广泛应用将持续推进“人工智能 +”行动

当前，人工智能技术已成为国际经济的新焦点、经济发展的新引擎、社会建设的新机遇。随着算力、算法与数据的协同突破，人工智能技术快速发展，且已从实验室走向产业实践，深刻重塑着人类社会的生产生活方式。我们站在这个新的技术革命起点，抢抓人工智能发展的重大战略机遇，积极推动大模型研制，给人工智能技术发展带来重要突破。大模型凭借强大的泛化能力、多模态理解与持续学习特性，成为推动智能化转型的核心，引领着我们进入一个全新的智能化、高效化的新纪元。

近年来，我国政府相继出台了一系列政策支持人工智能的发展。国务院出台了《新一代人工智能发展规划》，将人工智能上升至国家战略。中央网络安全和信息化委员会出台《“十四五”国家信息化规划》，也提到了人工智能技术。2024 年《政府工作报告》提出开展“人工智能 +”行动。随后我国各地政府也相继出台了相关支持政策，加快推动大模型产业的持

续发展。其中，北京市人民政府办公厅发布《北京市促进通用人工智能创新发展的若干措施》提出“高效推动新增算力基础设施建设，开展大模型创新算法及关键技术研究”；广东省人民政府发布《广东省人民政府关于加快建设通用人工智能产业创新引领地的实施意见》提出“围绕基础架构、训练算法、调优对齐、推理部署等环节，研发千亿级参数的人工智能通用大模型，形成自主可控的大模型完整技术体系”。在国家政策和地方政策的指引下，我国人工智能产业发展迅速，形成了从基础研究到应用落地的完整生态体系。历经多年的技术攻关，国内在大模型、人工智能芯片、开源框架、算法优化等领域涌现出了一系列创新成果，取得了令人瞩目的成就，极大地推动了产业的技术创新，也为实体经济的数字化、智能化、绿色化转型提供了强大动力。

在此背景下，由国家工业信息安全发展研究中心标准所、全国两化融合标委会（TC573）、《数字化转型》期刊和联想集团联合编纂与发布《大模型 2.0 产业发展报告》（以下简称《报告》）具有重要的意义。一是《报告》系统梳理模型架构创新、训练范式革新以及算力基础设施的演进等领域大模型技术发展的最新趋势。通过对比国内外主流大模型的技术参数与应用效能，为从业者提供客观的技术参考。二是提供实践指导，梳理大模型在个人、企业的典型应用案例，揭示其赋能生产生活方式的核心逻辑，推动企业实现转型升级，大模型在企业的智能经营管理、智能设计研发、智能供应链管理、智能生产制造等典型场景的应用有效提升了企业的运营效率和产品质量。三是《报告》对未来发展方向进一步深刻洞察，为产业界和学术界提供指导和参考。希望通过《报告》的发布，能够激发更多业界同仁的思考和讨论，促进产学研各界的交流与合作，在彼此共同努力下，共同创造人工智能产业的辉煌未来。

【产经智见】

戴炜

联想高级副总裁
中国方案服务业务群总经理

大模型 2.0 驱动企业智能化升级破局

近几年，大模型超越了人工智能过去几十年的发展速度。正因如此，大模型也正在引领着以人工智能为代表的新一代信息技术深入到人类的社会生活和经济发展。从技术萌芽到产业应用，是一次巨大的技术跃升，大模型技术也是如此，这也是我们这一产业发展报告的价值所在。

从产业发展现状来看，生成式大模型的技术成熟度已经很高，但产业应用却面临着巨大的鸿沟。与社会个体或创作者日常的大模型应用场景（智能助手、文生图、图生文、翻译及语音设备等）相比，大模型在企业生产经营中的应用要广泛得多，也复杂得多。同样基于目前个人和小组织的应用价值，或许只是大模型价值很小的一部分，而大模型在产业特别是企业中一旦得到大规模深入推广，那将是大模型真正的星辰大海。

不仅如此，对于企业的转型升级而言，大模型技术的成熟，对企业的智能化升级带来更为现实和明确的方向，也对企业智能化升级之后的智能化运营效率和数字化创新能力的提升产生巨大的影响。

联想是一家服务于企业和社会机构数字化转型和智能化升级的企业，在近年的技术探索和产业实践中，我们深切感受到：从基础大模型走向企业级大模型平台，所需要的技术适配和系统平台升级，远比想象的复杂得多。

在近年的探索实践中我们发现，大模型在产业和企业的应用不像之前企业信息技术一样，只是一个独立的应用或系统，而是需要系统性解决企业生产经营平台智能化的问题：需要融合企业原有专用人工智能应用和基础大模型；需要实现大模型的云端和本地的混合部署；需要实现企业商业数据和社会公共数据融合后的大模型训练和调优；更需要实现企业算力平台从通用算力向“通用算力 + 图形算力 + 智能算力”融合的混合算力体系的过渡。我们把这些特征统称为混合人工智能，这也是近几年我们和合作伙伴一起，为大模型应用提供的一种经过产业验证的技术方式。

同时，在企业大模型的发展中，我们依托企业在不同场景智能体的应用，降低大模型应用过高的技术门槛。基于企业应用场景的闭环，依靠适合企业特殊场景的大模型，实现企业独立场景的智能化，同时提供不同智能体的协同，让企业的生产经营和管理运营最终实现智能化。这就是大模型发展过程的 2.0 阶段，从模型研发向模型应用方向发展。

大模型是企业智能化升级的一种高效技术路径，但企业的数字化转型和智能化升级，远比大模型应用复杂得多。在大模型的企业应用中，或者说在企业大模型的发展过程中，我们面临着巨大的机遇和复杂环境，需要全产业的持续创新。

繁花落尽 繁星升起

1956 年 8 月，在美国汉诺斯小镇宁静的达特茅斯学院中，一群科学家正聚在一起，讨论着一个完全不食人间烟火的主题：用机器来模仿人类学习以及其他方面的智能。会议足足持续了两个月的时间，虽然大家没有达成普遍的共识，但是却为会议讨论的内容起了一个名字——人工智能。因此，1956 年也就成为人工智能元年。

在人工智能的漫长征程中，我们见证了技术的波澜壮阔，也感受到了时代的脉动。大模型 1.0 时代，如同一场科技的启蒙，揭开了深度学习与自然语言处理的新篇章。在这个时代，基于大型语言模型（LLM）架构的概率模型，无论是通用模型还是行业模型，都在探索着人工智能的可能性。然而，这些模型在虚拟的非严肃场景中虽然展现出了惊人的潜力，但在商业模式的探索上却显得步履蹒跚。在这个时代，市场上涌现了 200 多家致力于大模型研发的企业，一些企业如繁花般绽放，却最终随着竞争的激烈而消逝，只有几十家能够适应市场并持续运营，其他企业未能在竞争中生存下来。

随着时间的推移，我们迎来了大模型发展的 2.0 时代。这是一个全新的纪元，一个以可商业化为视角的时代。在这个时代，我们不再仅仅关注技术的突破，更开始审视整个大模型发展的产业链、生态和价值体系。我们开始思考，如何将这些强大的模型转化为实际的生产力，如何在现实世界中找到它们的立足之地。

在大模型 2.0 时代，行业 AI 应用正对市场产生深远的影响和改变。随着技术的成熟和应用的深入，人工智能不再局限于虚拟的实验场，而是开始渗透到各行各业，从制造业的自动化到服务业的个性化推荐，从医疗诊断的精准度到金融风险的管理，AI 应用正在重塑传统行业的面貌。

行业 AI 应用对市场的冲击首先体现在效率的提升上。通过引入智能算法和自动化流程，企业能够显著提高生产效率，降低运营成本。其次，AI 技术的应用也为企业带来了新的商业模式和收入来源。例如，基于数据分析的个性化服务能够更好地满足消费者的需求，创造更大的市场价值。

然而，行业 AI 应用也带来了挑战。随着自动化和智能化水平的提高，一些传统岗位可能会被机器取代，引发就业结构的变化。同时，数据安全和隐私保护也成为行业发展必须面对的问题。

在这个时代，我们看到了基于AIPC、AI Phone等新硬件的个人模型的崛起。这些模型不仅仅是技术的展示，更是个人生活和工作的得力助手。它们以更加个性化、智能化的方式，融入我们的日常生活，成为我们生活中不可或缺的一部分。

同时，基于企业化的场景模型也开始崭露头角。这些模型深入到各个行业，从金融到医疗，从教育到制造，它们正在改变着传统的工作方式，提升着生产效率，推动着产业的升级。它们不仅仅是简单的工具，更是企业创新和竞争力的新源泉。

这些新硬件和场景模型的出现，正如繁星般升起，为我们描绘了一个崭新的产业未来的蓝图。在这个蓝图中，大模型不再是高高在上的科技概念，而是实实在在地服务于人类，成为推动社会进步的强大动力。它们成为了产业未来的锚点，引领着我们向着更加智能、高效和人性化的未来迈进。

本书旨在深入探讨大模型 2.0 时代的发展趋势、挑战和机遇。我们将一起见证这个时代的变迁，一起探索人工智能如何更好地服务于人类社会。在这个充满无限可能的新时代，让我们携手前行，创涌未来！

基础篇

第一章
大模型发展进入 2.0 阶段

人工智能旨在让计算机系统模拟人的思维过程和学习过程，其发展和技术演进经历了多个阶段，并产生了自然语言处理、语音处理、计算机视觉数据分析等一系列技术和方法。其中，人工智能大模型，即大型语言模型（Large Language Models，LLMs），是指基于深度学习算法，依托大规模的数据进行训练，利用强大算力资源进行推理和应用，能执行复杂下游任务的神经网络模型，也被简称为大模型。它使用大量的计算资源进行训练和部署，在大规模数据集上完成预训练后，无需或仅需少量数据的微调，即可完成问答、翻译、语音识别、图像识别等复杂任务，直接支撑各类应用。大模型的发展可以追溯到人工智能和机器学习的早期阶段，但真正爆发却是在最近几年，特别是在自然语言处理（NLP）领域。一般来说，大模型的发展经历了以下几个阶段。

• 人工智能探索期

早期探索（1950-2000 年）：人工智能概念提出以后，研究者首先进行了基于简单规则进行推理与基于知识进行推理两个阶段的实践。但实践证明，基于

知识系统和逻辑推理的机器并不智能。随后研究者尝试让计算机自己在数据中学习，基于数据的统计学习出现并成为主流，人们将目光转向了数据和统计理论。在这个阶段，基于神经网络的连接主义学习也得到了一些进展，早期的模型如感知器和反向传播算法为后来深度学习的发展奠定了基础。

- **新一代人工智能起步期**

深度学习的兴起（2006-2012 年）：深度学习的概念在 2006 年被正式提出，它是一种复杂的机器学习算法，主要目的是模拟人脑的神经网络结构，使机器具备类似于人的分析和学习能力。同构建单层模型、需要人工提取特征的机器学习相比，深度学习构建了多层网络，可以自动学习、提取数据特征，在模型复杂程度、能力上都比机器学习有明显的提升，此后深度学习开始广泛应用于自然语言处理任务。

预训练模型的出现（2013-2018 年）：随着深度学习的应用发展，Word2Vec 和 BERT 等预训练模型出现。预训练模型的出现，标志着“通用模型能力 + 特定场景微调”思想的成熟。它的核心思想是将在一个任务上训练得到的模型权重迁移到其他任务上，使得新任务能够利用已经学到的知识和特征。

- **新一代人工智能发展期**

Transformer 架构的突破（2017 年）：2017 年，谷歌团队首先提出 Transformer 架构，该框架基于自注意力机制，解决了自然语言处理任务中循环神经网络（RNN）和卷积神经网络（CNN）在应用中的长距离依赖问题，使长文本理解处理成为可能。这是人工智能领域发展的新突破，Transformer 架构也成为后续大模型的核心技术。

- **进入大模型 1.0 时代**

GPT 系列的崛起（2018 年至今）：GPT 系列模型是 OpenAI 推出的预训练语言模型。GPT-1、GPT-2 的实践已经暗示了扩大数据和参数后预训练模型的潜力。GPT-3 在网络容量上做了更大的提升，最终在不同任务上有了令人震惊的

表现。基于 GPT-3 的对话产品 ChatGPT 在全球范围内也引发了巨大的关注和讨论。GPT-4 在参数上较 GPT-3 又有数量的突破，拥有了处理多模态数据的能力。GPT 系列的崛起证明了在 Transformer 架构下使用海量参数与数据训练的模型的强大能力，“大模型 + 场景微调”成为了行业新的研究范式。

多模态大模型的发展（2021 年至今）：随着 CLIP 和 DALL·E 等模型的出现，大模型开始突破传统的文本处理领域，进入图像、音频等多模态领域，进一步拓宽了人工智能的应用范围。

- **进入大模型 2.0 时代**

大模型发展的 1.0 阶段，是大模型的探索阶段。受到 GPT 的冲击，科技巨头纷纷构建通用大模型，探求大模型在技术上的进一步发展，开始大模型在商业化层面的探索。行业巨头开始筹备构建自己的垂直大模型，在通用大模型基础上利用行业知识进行微调的行业大模型也在金融、医疗、能源等行业出现。

大模型 2.0 阶段，是大模型的应用阶段。随着通用大模型数量爆发与质量快速提升，行业大模型继续发展，在通用大模型基础上基于企业数据和个人预训练数据的企业大模型和个人大模型快速得到应用落地，已经成为大模型最有价值的商业化方向。与此同时，通用多模态大模型产品也得到进一步突破。2024 年 2 月，OpenAI 发布了其文本生成视频的大模型 Sora。Sora 展示了人工智

能在理解和模拟物理世界方面的能力，被认为是通用人工智能的关键一步。2024 年 5 月，OpenAI 推出其新旗舰模型 GPT-4o，能够实时对音频、视觉和文本进行推理，为多模态交互开启无限可能。2025 年 1 月，DeepSeek 正式发布 DeepSeekV3 模型，大幅降低了大模型对算力的要求，同时显著降低了企业应用门槛，加速推动了大模型在企业部署及应用的普及化。

随着大模型从 1.0 阶段到 2.0 阶段的迈进，大模型产业也从初期应用的探索转向应用的规模化与深度化发展，以企业和个人使用者为中心，以大模型应用服务提供商为主导的产业生态正在形成。

01 大模型 2.0 的技术特点和产业生态

（一）大模型 2.0 时代

大模型 1.0 阶段的核心任务是人工智能技术实现快速迭代发展并且逐渐成熟，尤其在文字、语音、图像、视频、代码等领域加速迭代，为大模型技术的商业化应用打下了坚实基础。

大模型发展进入 2.0 阶段的标志，就是在大模型能力快速增强，以及对算力门槛与成本的大幅降低的背景下，大模型技术进入规模商业化。在资本逐渐认可大模型商业价值的背景下，个人和企业用户初步形成付费意愿，越来越多的大模型赋能产品和服务被采购，产业不断探索新的商业模式和业态。主要表现在两个方面：面向个人消费市场，以大模型技术路线的产品与服务越来越丰富，产品与服务在个人端不断探索；面向企业级市场，几乎所有企业与组织都积极探索大模型技术路线为企业带来的创新与变革，大模型开始加速渗透到千行百业的场景中，赋能行业升级。

在这个过程中，智能体 (Agent) 作为大模型在企业落地的重要手段，其种类的丰富和繁荣，是大模型应用规模商业化的一个重要标志。智能体是一种能够感知环境、进行自主决策并执行行动以实现特定目标的智能实体或系统。它基于

人工智能技术，具备学习、规划、记忆、工具使用和行动等能力，能够独立或半自主地完成任务，而不仅仅依赖于用户的明确指令。在大模型 2.0 的发展阶段，大模型在技术、商业化应用、产业等层面均出现了显性变化。

技术上，为了解决基于 Transformer 架构黑箱推理过程、输出可解释可控的问题，以及消除大模型的幻觉，对大模型的训练数据、方法与模型算法持续优化，提高大模型输出的准确性；基于外挂知识库的检索增强生成模型和去概率化大模型逐渐发展，大模型输出的逻辑性、可控性、专业化与行业化进一步加强。随着模型能力的加强，通用大模型的参数也从千亿级向万亿级发展，面向不同应用的不同参数规模的大模型多元化，对算力的要求进一步提高。

商业上，在通用大模型与行业大模型持续发展的基础上，大模型在个人场景与企业场景中找到了可发展的商业模式。基于个人大模型的智能硬件和软件的产品与服务如雨后春笋般出现，基于企业大模型的涉及生产、经营、管理等全域智能化应用正成为企业数字化转型和增强市场竞争力的重要工具。

在产业层面，分别以个人和企业为核心的大模型生态体系正在形成。个人大模型的生态体系主要包括数据、技术基础设施、模型应用、服务与产品、安全性与隐私等多个层面；企业大模型的生态体系主要包括基础层、应用层和战略层。个人通过构建个人大模型推动个人工作方式创新升级，企业通过构建基于大模型的企业智能体进行智能化转型，推动产业结构优化升级，促进社会的生产能力发展。

（二）大模型 2.0 时代的技术特点

在新的阶段，大模型也形成了明显的特点，主要体现在以下几个方面。

更强的理解能力。在大模型 2.0 发展阶段，多模态大模型不仅能处理和理解文本、图像、音频和视频等多模态的数据，还具备了理解跨模态数据的能力。另外，多模态大模型是以思想链展示出类似人类解决问题时的逻辑推理过程，输出具有更高的可解释性。更强的理解能力推动大模型推理的准确性，满足了更多的个人与企业对大模型以及智能体的需求。

更全面的知识储备。在 2.0 发展阶段，数据版权化得到更多关注，数据版权趋势有助于提高私有数据的隐私性和安全性，从而推动更多行业知识能够安全进入大模型进行训练，使大模型拥有更全面的知识储备，进而在回答来自各种领域的复杂问题时，具备更高的泛化能力和更准确的应用范围，推动了大模型在千行百业向深度应用的拓展。

更高效低碳的训练模式。在大模型 2.0 发展阶段，模型压缩技术、RAG 架构等技术创新能够推动训练流程优化升级，TPU 等集成电路的出现能够加速网络的训练和推理过程，异构计算平台结合了多种硬件资源，在保证高性能的同时降低设备功耗，提高了效率，降低了大模型训练成本，让大模型发展更适应绿色环保的理念。

更广泛的产业应用能力。在大模型 2.0 发展阶段，大模型通常提供服务化平台，使得大模型的能力能够以 API 或其他服务的形式对外提供，从而满足不同使用者和应用场景的需求，更广泛地应用于多个产业和领域，为不同企业提供更专业、更精细的技术支持。同时，为深度学习神经网络设计的专用处理器出现并逐渐应用，AI 算力芯片逐步进入个人终端，推动个人生活和工作场景的智能化发展。

（三）大模型 2.0 的产业生态

大模型在现实复杂场景中的应用，主要难点在于打造完整的产业生态体系，进而打通大模型应用的“最后一公里”。现阶段，以个人和企业为核心的大模型产业生态体系正在形成。

个人大模型的产业生态体系主要涉及数据供给、技术基础设施供给、模型应用、服务与产品供给、安全与隐私等多个方面。个人大模型的训练需要个人活动和传感器数据，并对个人数据进行分析与处理；此外，还需要应用程序、操作系统等软件和个人终端产品、个人服务器等硬件提供技术基础设施供给；在模型应用方面，个人大模型还可以提供个性化推荐、健康检测、生产力工具等应用；基于个人大模型，还有定制应用、智能家居、智能摄像头等产品服务；在个人

大模型的开发应用过程中，也要注意个人数据的保护与合规性。利用个人大模型，个人可以在学习、工作、生活、娱乐等多个方面提升体验，实现个人生活的全方位升级和发展。

图 1 个人大模型产业生态图谱

企业大模型的产业生态体系主要包括基础层、应用层和战略层三大层面。在基础层，首先，需要来自企业的内部和外部等多种来源的数据供给；而后利用企业数据训练开发大模型，并将企业大模型应用于决策、自动化、个性化服务等多个方面；在训练企业大模型的过程中，还需要确保算力集群、储存服务器等硬件和大数据平台、云计算工具等软件技术基础资源的供给。在应用层，企业大模型的应用主要涉及市场与竞争、投资与财务、合作伙伴与生态系统、研发与创新、供应链、用户与客户等多个领域，其中通过组织战略合作，可以构建以企业大模型为基础的生态系统。在战略层，主要涉及企业大模型的未来与发展、学习与优化、安全与合规三个方面，通过规划大模型的未来发展、持续改进模型性能并保证过程安全合规，推动企业大模型的可持续发展。在整个企业大模型产业生态体系的构建下，企业能够通过大模型推动自身转型升级，实现长远发展。

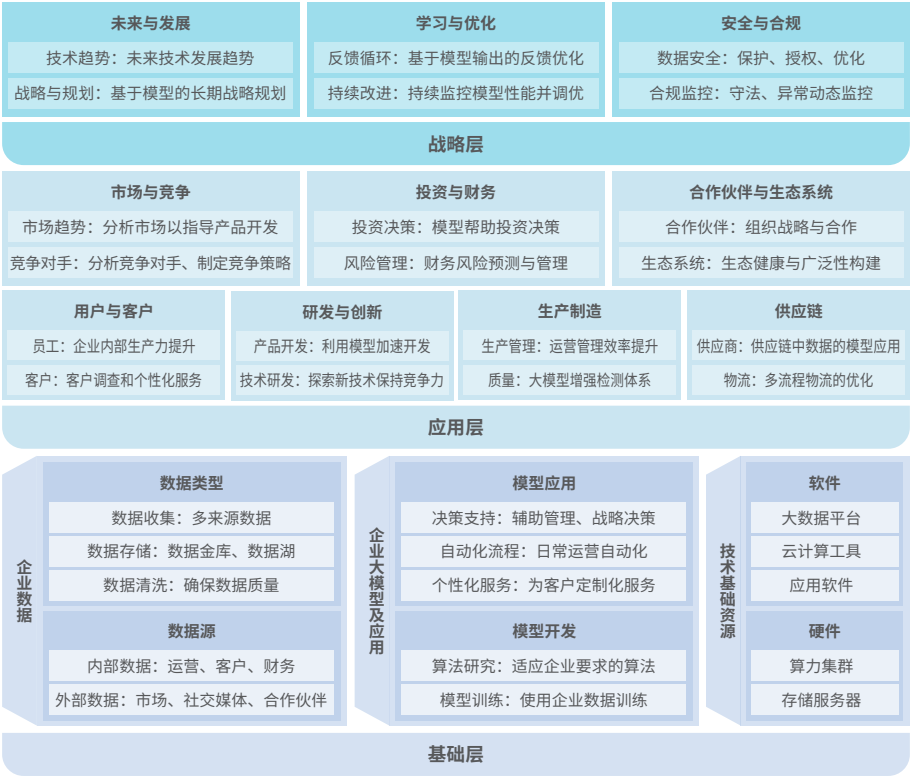


图 2 企业大模型产业全景图

基于个人和企业需求的大模型产业生态体系的构建，是大模型普及的关键。预计未来，个人大模型产业生态体系和企业大模型产业生态体系将逐渐成形，并为个人工作生产效率的升级和企业智能化转型提供关键支撑。

02 大模型 2.0 驱动社会进入智能时代

到大模型 2.0 发展阶段，个人大模型和企业大模型已成为未来大模型发展的高价值发展方向，也成为新质生产力的重要组成部分，不仅促进对个人的工作生活方式变革，而且推动企业的生产管理方式变革，进而极大地推动社会生产力

的发展。

（一）个人生产力明显提高

对个人而言，基于大模型的智能软件、智能终端（如 AIPC、AI Phone）将为个人提供全新的、智能的服务。大模型在内容创作、数据分析等领域的广泛应用，将打破技术壁垒，使人们能够专注于更具创造性的任务，推动个人工作方式的创新升级，使个人的工作生产效率明显提高。随着新终端不断涌现，传统设备的功能将纷纷向各类新终端迁移，个人办公助理将成为最早实现大规模商业化的应用领域。

（二）企业向全栈智能化发展

对企业而言，企业将从基于“大模型 + 场景微调”的局部化场景智能化，向基于“大模型 + 企业私域知识库 + 场景微调”的全栈智能化转变。精准聚焦企业自身生产经营管理的企业大模型，促使企业从生产、经营、管理、决策等多方面进行数字化与智能化重构，实现智能经营管理、智能研发设计、智能供应链管理、智能生产制造，达到提升企业的业务效率、减少重复劳动、降本增效的目的，甚至改变企业组织结构、业务流程、生产形式与产品形态、商业模式等，充分释放大模型带来的新质生产力变革。

（三）促进社会生产力与生产关系变革

对社会而言，大模型 2.0 的发展推动了生产力和生产关系的变革。

大模型可以提升效率，降低成本，能够承接重复工作。在生产领域，大模型驱动了人工智能工程化进程，从定制训练转向预训练，重构产业链，将促进生产力的提升；在家庭生活和工作方式方面，大模型实现了语音和文本交互的高效、自然，提高了工作效率；在产业智能化升级中，大模型在制造、金融、医疗等领域发挥了关键作用，提升了生产效率和服务质量。此外，大模型的快速发展驱动了数字新基建，促进了数字经济的高质量发展，成为推动社会生产力进步的重要力量。

03 各国密集出台人工智能及大模型的支持和监管政策

（一）产业政策

• 全球主要经济体陆续出台支持人工智能发展的相关政策

近年来，全球主要经济体纷纷出台政策以支持人工智能的发展。自 2013 年以来，超过 20 个国家和地区发布了 AI 战略或计划，如欧盟签署的《人工智能合作宣言》和东盟的《东盟数字融合框架行动计划》。美国政府强调人工智能的有效交互，并在 2023 年更新了《国家人工智能研发战略计划》，新增了对高质量数据集和先进算力资源的重点支持。德国在 2023 年推出了《人工智能行动计划》，旨在提升 AI 质量，加强算力和数据质量。日本政府致力于建设国家级 AI 平台，并在 2023 年设立了战略会议讨论国家 AI 战略。英国则发布《人工智能 2020 国家战略》，并宣布建造超级计算机 Isambard-AI，同时鼓励民间资本投入 AI 领域。这些举措共同推动了全球人工智能技术的快速发展和应用。

从全球各国不断出台的产业政策及战略规划分析，未来国际的科技竞争将主要集中在人工智能领域。

• 以大模型为主的人工智能提升至我国国家战略层面

自新一代人工智能开始出现，我国相继出台了一系列政策措施，不断推动人工智能发展步入新阶段。在新一代人工智能起步期，我国逐渐将人工智能提升至国家战略层面，出台相关发展措施；在人工智能发展期，我国出台关于人工智能的发展规划，制定新一代人工智能标准体系；在大模型起步期，随着 GPT-3.5 在全球掀起大模型热潮，我国出台生成式人工智能的相关发展措施，并提出研制大数据和算力相关的标准；在大模型发展期，我国中央和地方出台政策和发展规划，推动人工智能垂类大模型和多模态大模型等持续发展。

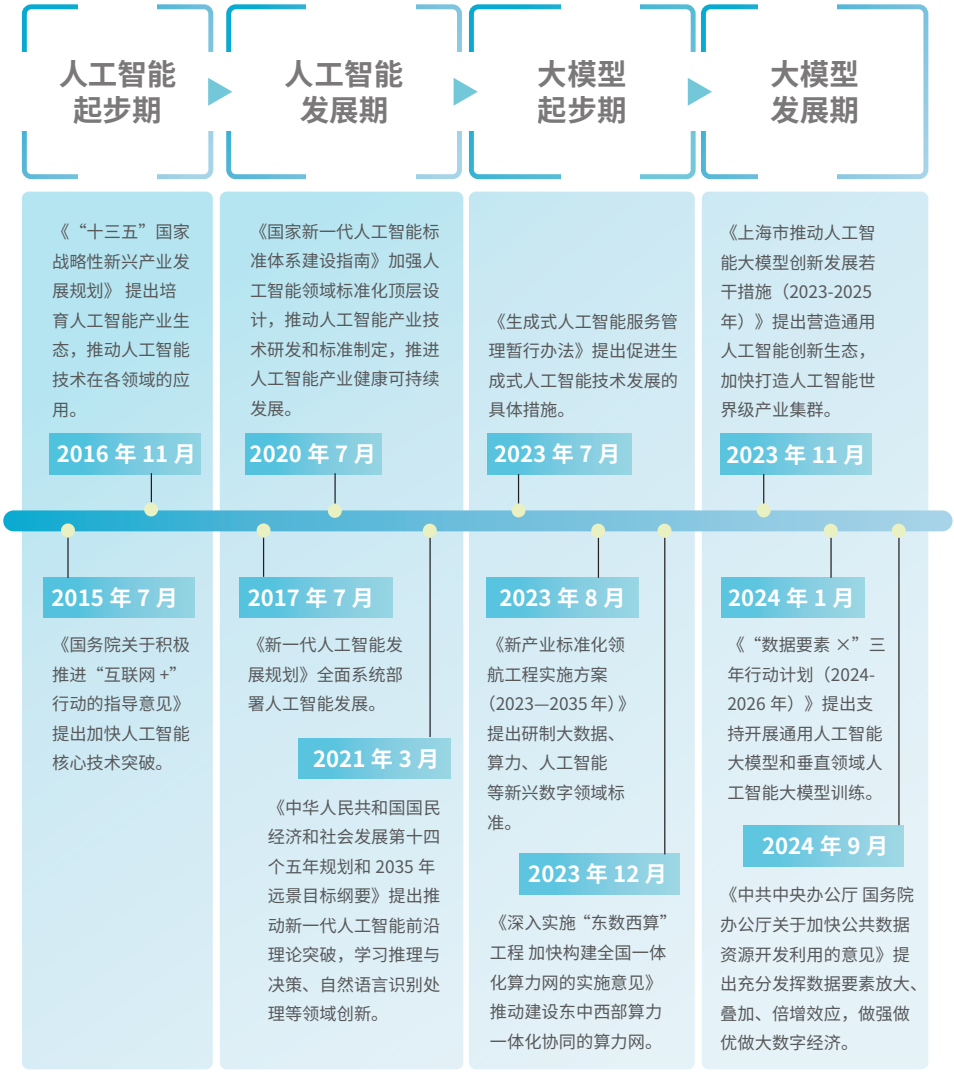


图 3 国家层面人工智能与大模型的相关政策汇总

（二）监管与治理

世界主要经济体如美国、欧盟和英国均对大模型采取了监管措施，在引领 AI 发展和管理风险方面均展现出积极态度，但具体监管策略和方法存在差异。美国只是在现有立法框架内进行规制，没有在联邦层面通过专门针对大模型的综

合性立法。欧盟通过了《人工智能法案》对人工智能进行监管，在全球人工智能治理中领先。英国部分现存的法律法规已经涵盖了对人工智能的规定，早在 2018 年，英国通过了《数据保护法》；2021 年 5 月通过关注 AI 伦理的《自动决策系统的伦理、透明度与责任框架》；2023 年 3 月，英国出台《创新型人工智能监管》白皮书推动人工智能发展；2023 年 10 月，英国议会颁布了《在线安全法案》保护用户安全。

我国对大模型的监管现状体现在多个层面，涉及网络安全、数据安全、个人信息保护等关键领域。2023 年 1 月国家互联网信息办公室等部门联合颁布《互联网信息服务深度合成管理规定》，针对深度合成服务，要求服务提供者和技术支持者依法进行安全评估和算法备案。2023 年 8 月生效的《生成式人工智能服务管理暂行办法》是我国针对生成式人工智能服务领域的首部法律法规，提出了分类分级监管的要求，并明确了服务提供者和技术支持者在内容管理、训练数据、用户权益保护等方面的责任。此外，我国还在积极推进全国层面的人工智能专门立法。《人工智能法（草案）》已列入国务院 2023 年立法工作计划，预示着未来将有更加全面和系统的法律框架来指导人工智能及大模型的发展。

总体来看，全球大模型监管正处于快速发展和完善阶段，各国根据自身的法律体系和社会需求，制定了相应的监管政策和法规。我国也在积极构建符合国情的大模型监管体系，通过立法和行业规范的制定，推动大模型技术的合规发展，并逐步与国际标准接轨。随着技术的进步和社会的发展，大模型监管将继续适应新的挑战，以确保技术的安全、可靠和伦理应用。

（三）我国大模型合规标准的制定

2024 年 6 月，工业和信息化部、中央网信办、国家发展改革委、国家标准委等四部门联合印发《国家人工智能产业综合标准化体系建设指南（2024 版）》（以下简称《指南》）。提出到 2026 年，我国标准与产业科技创新的联动水平持续提升，新制定国家标准和行业标准 50 项以上，引领人工智

能产业高质量发展的标准体系加快形成。《指南》主要内容涵盖了人工智能产业的现状分析、总体要求、建设目标、工作原则、建设思路、重点方向和保障措施等多个方面，并明确了关键技术和标准成果的具体目标，标志着中国在人工智能领域标准化工作的重要进展。《指南》将推动人工智能产业的高质量发展，促进产业科技创新与标准化的联动，加强跨行业、跨领域的协同，有助于提升中国在全球人工智能产业中的竞争力和影响力。

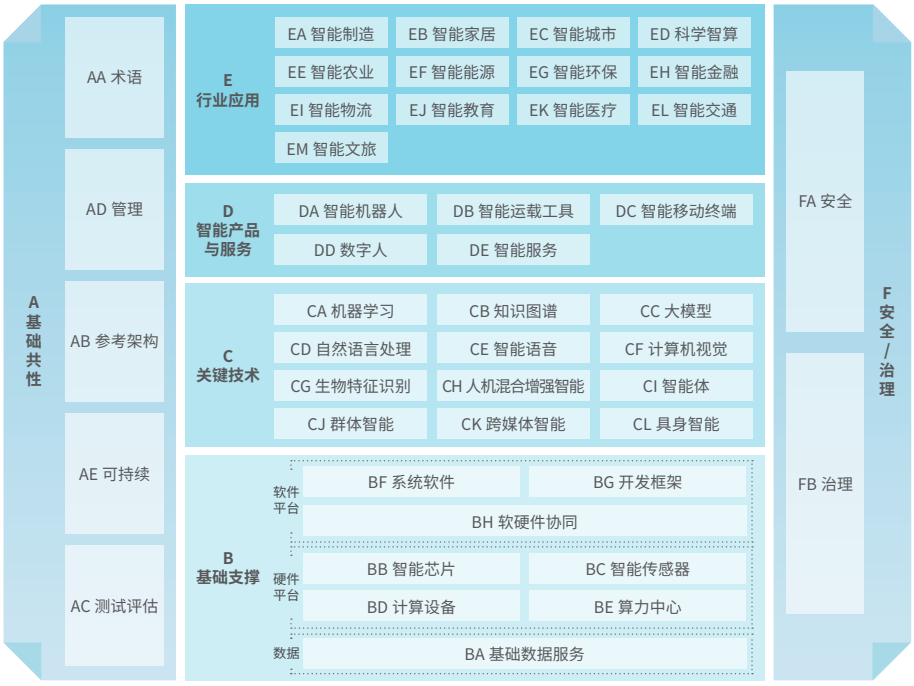


图 4 中国人工智能标准体系结构图

04 大模型引发科技巨头的投资热潮和人才需求的持续增长

（一）科技巨头企业的投资热潮

从自动驾驶到医疗诊断，从智能客服到金融科技，大模型正在逐步改变人们的

生活和工作方式，这也使投资者看到了其中的巨大商机，投资者对于大模型的未来发展充满了信心，纷纷将资金投入到这个领域。

谷歌、亚马逊、微软、阿里巴巴和联想等全球知名的科技巨头都已经在大模型领域进行了布局和投资，通过设立专门的研发机构、招募顶尖的研究团队、投入巨额的研发经费，积极推动大模型技术的研发和应用。

这些科技巨头在大模型领域的投资主要集中在以下几个方面：一是基础技术研发，包括模型结构、算法优化、训练技术等方面的研究；二是应用研发，将大模型技术应用到具体的业务场景中，如智能客服、智能推荐、自动驾驶等；三是生态建设，通过建立开放平台、发布开源工具、提供技术支持等方式，吸引更多的开发者使用大模型技术，共同推动大模型生态的发展。

微软在 2019 年向 OpenAI 投资了 10 亿美元，并在 2023 年年初再次宣布向 OpenAI 投资 100 亿美元。在云计算、办公软件、搜索引擎等方面开展大模型应用均有大手笔投资。谷歌不仅推出了自己的大模型 PaLM2，还向人工智能公司 Anthropic 投资至多 20 亿美元，以加强与微软的竞争力。亚马逊投资 40 亿美元支持 Anthropic 的基础模型开发，Anthropic 的产品 Claude 被认为是 GPT-4 的重要竞品。

国内云计算巨头与大模型创企之间的投资与合作表现活跃。阿里巴巴已经投资了 5 家国产大模型企业（截至 2024 年 3 月），分别是月之暗面、MiniMax、智谱 AI、百川智能和零一万物，尤其是在 MiniMax 和月之暗面这两笔大规模融资中，阿里巴巴是主要的投资方。腾讯投资了智谱 AI、MiniMax 和百川智能 3 家，其中对百川智能、智谱 AI 的投资是与阿里共同参投。联想投资了 relnventAI、X Square 和诺谛智能等大模型企业，其中 relnventAI 由联想创投独家投资，诺谛智能是由联想孵化并投资的行业大模型及应用企业。

联想集团在智能化时代的转型变革过程中，根据自身特点和优势，将大模型投资作为人工智能领域的重要布局。联想基于 AI 原生技术架构的技术突破，打造了针对特定行业或企业场景的联想“智能体”，并形成了端到端的行业解决

方案，提供基于企业智能体的评估、部署、监控和运营等一站式服务，并可以迅速通过 API 实现与外部应用融合，形成了企业智能体系统性解决方案。

（二）人才需求的持续增长

随着大模型技术的快速发展，特别是在自然语言处理、计算机视觉等领域的应用不断拓展，对于具备深度学习、数据科学等专业技能的人才需求呈现出爆炸性增长。

大模型优秀人才缺口大，主要有以下几个原因：一是行业对于能够将大模型技术应用于实际业务场景的高端技术人才有着迫切的需求；二是大模型生态不仅需要单一领域的专家，更需要跨学科知识融合的人才；三是大模型技术在多个行业领域的应用不断拓展，增加了对具备相关领域知识和技能的人才的需求。

为了满足日益增长的人工智能产业需求，近年来国内通过政策鼓励和高校人才培养等措施，加快培养人工智能创新型人才。如 2023 年 11 月，上海市印发《上海市推动人工智能大模型创新发展若干措施（2023-2025 年）》，包括优先推荐大模型创新重点人才纳入国家和本市相关高层次人才计划，重点支持大模型相关紧缺技能人才落户，组织企业、高校、科研机构联合培养跨学科大模型人才等。

基础篇

第二章
大模型 2.0 阶段的关键要素

随着大模型应用逐渐广泛、商业化逐渐落地，大模型发展的关键要素均发生变化，它们相互关联、相互促进，共同推动着大模型的发展。

01 关键要素

大模型具有海量参数和复杂架构，是一种用于深度学习任务的模型，拥有强大的处理能力和表征能力。

大模型关键要素具体包括基础层、模型层、应用层和保障层。大模型以数据、算力、算法和工具为基础支撑，借助数据管理、模型训练、评估优化、服务平台、插件等大模型辅助工具，由大模型供应商开发出基础大模型或行业大模型，再由大模型应用服务商与用户一起将其延伸至制造业、金融、医疗、交通等下游场景应用，切入个人或企业的实际场景。

图 5 大模型生态关键要素

02 基础层

在大模型的发展过程中，数据、算力、算法和工具是大模型发展的基础和支撑。

数据作为大模型能力的来源，数据服务产业将持续发展，数据版权化意识不断增强，推动数据治理、数据安全和隐私保护的法律法规建设不断完善，确保数据的合法合规使用。算力是模型落地的物质基础，国家正在建设算力网络，以智算为主的异构算力结构得到发展，为大模型提供了强大的计算支持，也为人工智能的广泛应用打下了坚实基础。算法是大模型的骨架，它们定义了模型的学习方式和决策过程，算法的创新和优化直接关系到大模型的性能和智能水平。随着深度学习、强化学习等先进算法的不断进步，大模型的学习能力和适应性将得到显著提升。

（一）数据

不同场景、任务对大模型能力的不同需求，需要不同来源的数据支撑。数据作为生产资料，其版权化服务模式流行，将极大推动数据服务产业的发展。

• 对数据数量、质量、样态产生丰富需求

训练集的质量直接影响着大模型训练的成本与结果。随着市场对大模型能力要求的不断增加，对高质量、精细化、定制化的数据需求日益凸显。对于文本类数据集与图像数据集，基于分类、目标检测、语义分割、序列标注等不同的任务也表现出不同样态。AR、自动驾驶等场景的出现，实时图像数据采集和数据自动标注的技术亦在快速发展。

• 数据来源丰富、数据工具发展

大模型训练数据的来源愈发丰富，数据构建的主体由大模型建设商、数据服务商逐渐向个人、企业、行业主体发展。数据构建由通用向私域延伸，个人、企业、行业构建私域数据集的意识加强。这些推动了数据建设平台的发展，数据传输、整理的工具软件愈发丰富、便捷。

• 数据服务产业发展将推动数据版权化

数据交易将推动数据建设平台与数据交易方式逐渐完善，数据版权化意识加强，数据付费成为未来趋势，定制训练数据集的需求激增。根据 Cognilytica 数据，2021 年全球人工智能训练数据市场需求约为 42 亿美元，预计到 2027 年这一需求将增长到 220 亿美元，2021-2027 年复合增长率（CAGR）达 32%。

• 数据治理仍须加强

推动构建高质量数据的同时要加强数据治理。从企业来看，大部分企业的数据治理工作面临着数据量庞大、数据种类繁多、数据管理效率低的挑战，目前尚未出现通用、可靠的数据管理工具，数据治理仅是企业的单兵作战。同时，加强数据治理也是保障国家安全、社会稳定和公民权益的迫切需要。企业数据如

何合理合规，并能保障数据所有者的利益，将是未来行业大模型发展的关键。

- **数据安全需要坚实保障**

数据安全关系到企业的运营稳定、业务发展以及声誉维护。企业不仅要健全数据安全管理系统，更要加强安全技术如加密、访问控制等的研发与应用。数据安全更是涉及国家安全、社会稳定、经济发展以及公民权益的重要议题。随着信息技术的发展，数据安全保障将愈发重要，从立法和执法角度加强对数据安全的监管势在必行。

（二）算力

算力是大模型落地的物质基础，大模型对算力的强需求推动异构算力技术发展。

- **算力需求涌现，智算将成主流**

在大模型落地之前，基于 CPU 芯片的基础算力是日常计算的主要支持，基于超级计算机的超算算力主要应用于科学计算与工程计算等高端领域。随着大模型的广泛应用，支撑人工智能应用和产业发展的智能算力（基于 GPU 和 NPU 芯片）将成为主要算力。

据预测，2022-2027 年中国智能算力规模年复合增长率将达到 33.9%，同期通用算力规模年复合增长率为 16.6%。我国智能算力需求的增长速度远超过通用算力增加速度。

- **异构算力技术得到发展**

大模型面对不同场景的挑战，高效计算成为迫切要求，在此背景下，以 CPU、GPU、FPGA、ASIC 等多种算力协同的异构算力处理体系，因为具有多样性、灵活性、高效性等优点正在崭露头角。

- **算力服务方式走向多元**

大模型落地应用的前提是算力基础设施的搭建。目前，企业应用大模型，如果

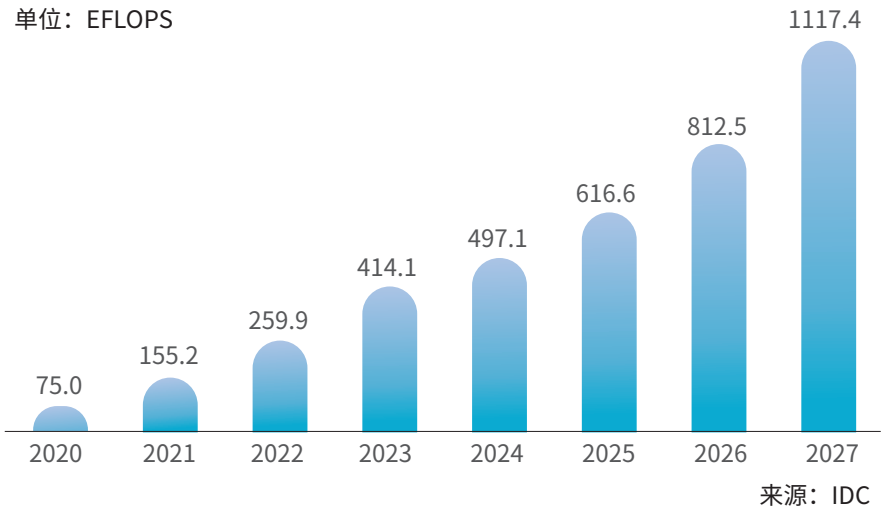


图 6 中国智能算力规模与预测图

采用纯自建算力设施的方式，将面临软硬件采购、机房搭建、系统运维等困难，容易陷入投入高、回报低的困境。购买云算力，将成为企业突围算力壁垒的有效途径。私有云在维护企业数据隐私和业务安全的技术上，能够为企业提供基本的计算服务，公有云可以弥补私有云算力的不足，在通用的场景下进一步满足计算要求，以“私有云 + 公有云”的方式搭建自身混合算力将成为企业搭建算力的通用范式。

与企业算力需求相应，异构算力的交易将从硬件向云端算力拓展，交易方式将从“一次性交付”向“按需订阅、按量计费”转变。算力服务将从算力本身向 AI 服务、中间件服务、数据库服务等配套服务发展，并与它们绑定。随着大模型的广泛运用，特定场景对异构算力的要求将愈发提高，但搭建算力以满足应用的过程将愈发简单。

基于大模型对异构算力的依赖，异构算力提供商将成为大模型及智能体服务生态的重要一员。已经有异构算力提供商，如联想，搭建了可订阅的、一站全包的“臻算服务 2.0”。算力价格与服务能力将成为未来异构算力提供商的主要竞争点。

（三）算法

算法是大模型的骨架。当前大模型的主流架构仍是 Transformer，其推理过程的无法解释性与结果的不可控性无法得到完全解决，未来融合检索增强生成（即 RAG）+ 知识图谱的架构或将成为新潮流。

• Transformer 仍是当前大模型的主流架构

大模型的快速崛起，主要归功于 Transformer 在自然语言处理任务上的突破。从技术架构上看，Transformer 架构是当前大模型领域主流的算法架构基础，虽然有 Megabyte、RetNet 等挑战者的出现，深度学习的大部分算法主要还是在 Transformer 的基础上创新，整个行业的底层算法发展开始放缓。

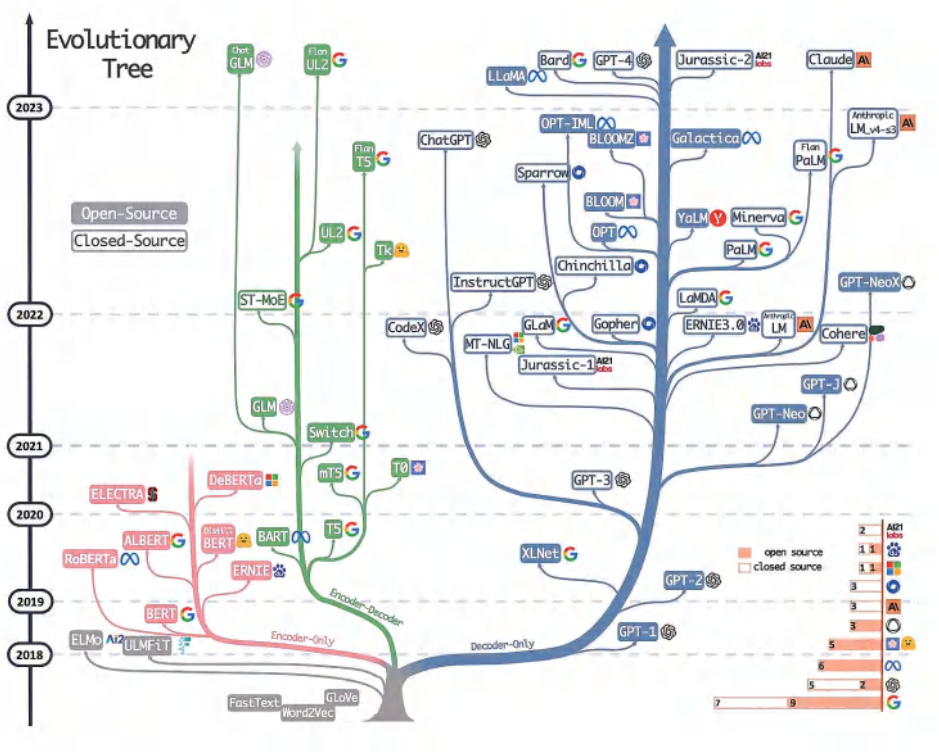


图 7 LLM 进化图

来源：《Harnessing the Power of LLM sin Practice:A Survey on Chat GPT and Beyond》

- **LLM 输出准确性有所提高，RAG 应用广泛**

基于 Transfomer 算法本身的复杂性，大模型算法内部决策过程是黑箱的，其输出的结果具有无法解释性与不可控性，可能会带来算法歧视、信息泄露、模型幻觉等诸多问题。随着大模型训练数据质量的提高、训练方式的优化，目前 LLM 框架下模型输出的正确性已经得到有效提高。

RAG 是当前火热的 LLM 应用方案。RAG 即 Search+LLM，即从外挂知识库中进行信息检索，并将检索到的信息注入 LLM 提示中，以得到最终答案。这种基于外部知识库检索增强的生成模型，其结果有一定依据、更加可控，目前已经有越来越多的应用实例。

（四）工具

为了让更多的企业、开发者可以便捷地将大模型应用于自身业务，全面完善的大模型工具链应运而生。工具链包含大模型训练框架、模型社区、应用开发工具、模型试验与评估、模型试验管理、数据标注、模型部署、模型监控与可监控平台、向量数据库等等，可以大致分为服务于模型开发者的模型训练工具、模型调用工具和服务于模型应用者的一体化模型服务平台。大模型工具不断发

图 8 模型工具分类举例

展，将逐渐降低大模型训练和调优的技术门槛，使得大模型训练、调优流程愈发简单。

03 模型与应用

通用大模型与行业大模型是行业的基座，模型数量激增，模型市场尚未收敛，商业化落地将成为模型生存竞争点。企业大模型服务于企业生产全场景，已经初步形成了以大模型服务商为主导的产业生态。基于个人大模型的产品已经如雨后春笋般出现。

（一）通用大模型与行业大模型继续发展

• 通用大模型数量尚未收敛

受到 GPT 的冲击，不少科技巨头企业均尝试以其自身的优势，构建通用大模型。各家通用大模型在训练数据、参数量、训练框架、任务能力等方面互相比拼，整体向更大参数、更高精度、更强能力方向发展。

虽然通用大模型在广泛的任务上均展现出稳定且强大的性能，但不同通用大模型在不同任务下的能力仍有不同。目前，使用者需要在众多的通用大模型中，同时考察大模型在特定任务中微调后的能力及在通用任务中的能力，最终确认使用的通用大模型。在产业激烈竞争背景下，通用大模型建设商，可以在模型通用能力的基础上打造模型的特殊能力，也可以打造一体化的通用智能体平台，构建模型应用生态。大模型只有在市场中被选择，才能在激烈竞争中生存。

• 行业大模型需要继续锤炼

行业大模型是针对特定行业领域应用的预训练大模型，其训练数据来源涵盖通用公域数据与行业数据。目前，制造、金融、医疗、游戏、法律、交通等行业均凭借各自独特的场景需求，搭建了行业大模型。这些行业大模型的意义在于深入理解和满足行业的特殊场景，为行业智能化、高效化发展提供有力支撑。

未来，随着技术的不断进步和应用场景的拓展，更多行业将构建出符合自身发展需要的行业大模型。

然而，行业大模型的发展仍面临着诸多严峻的挑战。首要问题便是缺乏充足且高质量的行业数据库。对于已经构建行业大模型的行业而言，数据库需要不断得到补充和完善，以适应行业的快速发展和变化。对于尚未构建行业大模型的行业，若无企业牵头并提供丰富的行业经验和资源，构建符合行业需求的数据库将变得尤为困难。

其次，由于行业大模型无法给出可靠、可控的输出，这给那些需要精确、唯一数据的生产场景带来了潜在风险。目前，大模型基于 Transformer 架构尚无法根本解决这一问题，只能通过不断优化数据与训练方法，努力提高模型输出的准确性。

（二）企业大模型商业化加速落地

企业大模型是利用企业自身的知识库并结合企业的使用场景，根据需求对基础大模型进行了充分的优化调整和自有知识库训练，可广泛应用于企业生产经营活动中的定制化、本地化或混合方式部署的大模型。大模型经过企业知识库的训练和优化，可以应用于经营管理、研发设计、生产制造、供应链管理等企业

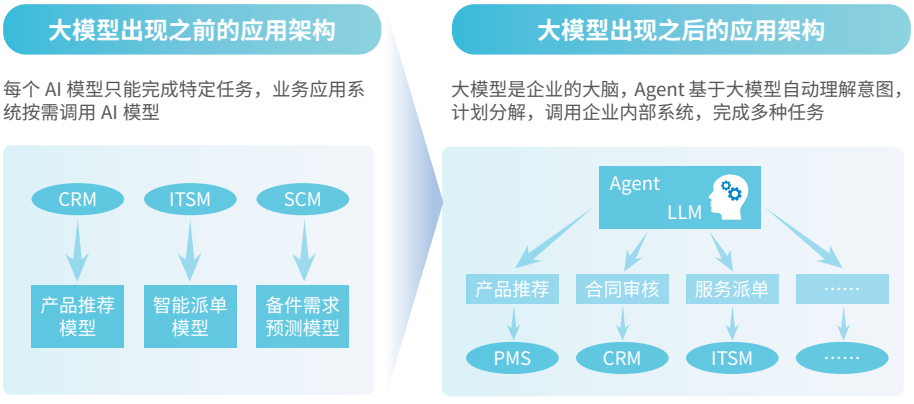


图 9 大模型出现前后的应用架构变化

生产经营的核心流程，为企业数字化重构、智能化运营提供驱动力。

以企业知识库训练的大模型，可以泛化应用于零售、政务、教育、金融和制造等诸多领域，虽然目前仍处在初级探索阶段，但已经初步形成了以智能体服务商为主导的商业生态，将会是大模型最具价值的应用方向之一。

（三）个人大模型产品大量出现

基于个人私域数据训练，并能依据个人使用习惯、偏好和需求提供个性化服务的大模型是个人大模型。基于个人大模型的应用软件已经如雨后春笋般出现，在工作与生活场景中为用户赋能。随着大模型能力的增强和算力芯片的发展，嵌入个人大模型的个人 AI 终端产品也已诞生。

个人大模型落地背后涉及硬件、软件的技术革新，将个人终端产品拉入新的竞争赛道。

04 模型保障层

大模型的发展离不开坚实的保障措施，合规标准的建立是其发展的基础支撑。在推进过程中，数据、模型、应用的安全保障必须全面到位，同时伦理治理亦不容忽视，只有确保价值对齐，大模型才能实现可持续发展。

（一）产业合规标准提供基础框架

人工智能产业标准对于推动技术创新、提升企业竞争力、引导产业升级转型至关重要。它确保技术发展先进可靠，同时通过严格的安全与治理规范，保障技术应用的安全性和伦理性，为产业的可持续、高质量发展奠定基础。

人工智能标准体系的建立主要从以下六个层面展开。一是建立基础共性标准。包括术语定义、参考架构、测试评估、管理和可持续性等方面，为其他标准制定提供基础和框架。二是建立基础支撑标准。涉及数据、算力、算法等技术要

求，为人工智能产业发展提供坚实的技术基础。三是建立关键技术标准。包括文本、语音、图像等人工智能技术规范，推动技术研发与创新应用。四是建立智能产品与服务标准。规范由人工智能技术形成的智能产品和服务模式，确保产品和服务的质量与效能。五是建立行业应用标准。规范人工智能在各行业的应用技术要求，推动产业智能化发展。六是建立安全和治理标准。包括人工智能的安全要求和治理准则，为产业发展提供安全保障。

推动多层面标准的建立，可以形成一个全面、系统、高效的人工智能产业标准化体系，为人工智能产业的健康、快速发展提供坚实的支撑。

（二）数据、模型、应用需要全面安全保障

在模型应用过程中，保障安全至关重要。必须采取一系列有效的安全措施，确保模型应用的全过程都在安全可控的范围内进行，从而为用户提供更加可靠、稳定的服务。

一是保障数据安全。这是大模型落地应用的核心关键之一。训练数据的收集、储存、交易、使用，都涉及隐私与安全问题。大模型不可控的推理回答过程也可能带来无法发现的数据泄露。数据安全是人工智能领域面临的重要风险。维护数据安全需要建立完善的数据管理制度、完整的数据交易链条，匹配各类安全技术与工具。

二是保障模型安全。大模型具有强大的学习和生成能力，但同时也面临着被攻击、被欺骗等风险。大模型的推理过程十分复杂，其决策和判断可能存在一定的偏差和错误。如果这些误判发生在关键领域，如医疗、司法等，可能带来严重的后果。加强模型安全需要加强对模型的安全性检测和评估，及时发现和修复模型存在的安全隐患，确保模型的稳定性和可靠性。

三是保障应用安全。在智能体落地中，涉及的业务场景和应用场景众多，如何确保应用的安全性和稳定性同样重要。保障层需要加强对应用的安全管理和监控，及时发现和处理应用中的安全问题，确保应用的正常运行和用户体验。人

工智能技术的滥用也可能导致不公平和歧视等问题，影响社会公正和稳定。企业和学术界应加强人工智能技术的监管和规范，推动其健康、可持续发展。

（三）伦理治理不容忽视

在 AI 技术的发展和应用中，伦理治理始终是一个不可忽视的重要方面。随着大模型和生成式人工智能的广泛应用，其涉及的伦理问题也日益凸显，社会各界在推进大模型应用的过程中，应坚守伦理原则，致力于构建一套完善的伦理治理体系，以确保 AI 技术的健康发展。

一是保护数据隐私。在大模型训练和应用过程中，应严格遵守数据隐私法规，确保用户数据的合法性和安全性。同时，应积极推广数据匿名化和脱敏技术，进一步降低数据泄露和滥用的风险。

二是关注算法公正性。在大模型的训练过程中，注重算法的公正性和透明度，避免算法偏见和歧视现象的发生。建立算法审查和监管机制，对算法的应用进行严格的监督和评估，确保算法的应用符合社会公正和公共利益。

三是注重 AI 技术的可持续发展。在推动大模型应用的过程中，应充分考虑环境、社会和经济的可持续发展因素，避免技术的滥用和资源的浪费，积极倡导绿色计算和低碳生活，推动 AI 技术与环境保护和可持续发展的深度融合。

四是强调伦理治理的多元化和包容性。在构建伦理治理体系的过程中，应积极吸纳各方意见和建议，推动伦理治理的多元化和包容性，加强与政府、学术界和社会各界的合作与沟通，共同推动 AI 技术的健康发展和社会公众的信任。

（四）价值对齐才能可持续发展

大模型技术的不断发展和深入应用，要与人类的价值观保持一致。无论是在企业环境还是个人应用中，都应当遵循道德准则，尊重人权、保护隐私、促进公平、维护安全，推动大模型技术的可持续发展。

在大模型的研发阶段就应引入伦理和价值的考量，确保模型的设计和应用都符合人类的道德和价值观。需要建立独立的伦理审查机制，对大模型的应用进行事前、事中和事后的审查，确保模型的应用不会偏离人类的价值观。加强公众教育和参与，提高公众对大模型技术的认知和理解，让公众参与到大模型的价值对齐过程中来，共同塑造 AI 技术的未来。

洞向篇

第三章 个人大模型

近年来，个人大模型的概念逐渐崭露头角，其商业应用前景也日益明朗。个人大模型，顾名思义，是一种基于大量个人数据训练得到的大型神经网络模型，它能够根据个人的使用习惯、偏好和需求，提供个性化的服务。

对个人而言，传统的产品逻辑是由开发商通过大数据统计，预测个人的使用需求，提供通用产品服务。在个人大模型的赋能下，在通用的服务基础上，产品可以通过分析用户的私域信息，为用户提供私人的、个性化的服务，服务的精确性将大大提高，用户体验也得到升级。

个人大模型在商业终端的应用已经初露锋芒。在软件方面，基于用户的终端数据，为用户提供个性化的数据服务或管理，如基于用户自然语言描述智能查找文件、智能创作文本等，在工作与生活等场景为用户带来全新的体验。在硬件方面，内嵌个人大模型并支持端侧模型训练的 AIPC 已经出现，内嵌个人智能助手的 AI Phone 已然进入市场。个人大模型在个人终端产品的落地将大大提高个人生产力。

个人大模型的应用还将带动相关产业链的发展。从数据采集、模型训练到应用部署，个人大模型的整个生命周期都需要大量的技术支持和人力投入。这将催生一批专注于个人大模型研发、应用和服务的企业，推动整个产业的繁荣和发展。同时，个人大模型的应用也将对数据安全、隐私保护等方面提出更高的要求，推动相关技术和法规的不断完善。

然而，个人大模型的应用也面临着一些挑战和问题。首先，个人数据的收集和使用需要遵循严格的法律法规和伦理规范，确保用户的隐私权益不受侵犯。其次，个人大模型的训练需要大量的计算资源和时间成本，对企业的技术实力和资金实力提出了较高的要求。此外，个人大模型的应用也需要与其他技术和服务进行融合和协同，以实现最佳的效果和价值。

个人大模型在商业应用领域的崛起已经势不可挡。

01 个人大模型为个人终端产品升级带来新机遇

个人大模型是公共大模型和本地大模型混合应用的产物，它既继承了公共大模型强大的数据处理和学习能力，又兼具本地大模型的个性化和专属性。这种组合使得个人大模型在满足用户多样化需求方面具有得天独厚的优势，能够为用户提供个性化的专属服务，将极大地提升个人终端产品的用户体验。

目前，为个人终端产品部署本地大模型已经成为行业发展趋势。其中，PC 是承载最多场景的个人通用设备，PC 基于硬件支撑，将成为最强个人计算平台；Phone 是承载最丰富场景的设备，基于其包含的最全面个人数据，将成为使用最广泛的个人终端。个人大模型将依托混合人工智能的方式，逐步实现全民普惠。

然而，在个人终端部署本地大模型，对硬件、软件及技术都提出了新的要求。一方面，在个人终端部署个人大模型需要对产品的硬件与软件升级；另一方面，利用个人大模型为用户提供出色的产品和服务成为新的行业竞争点。

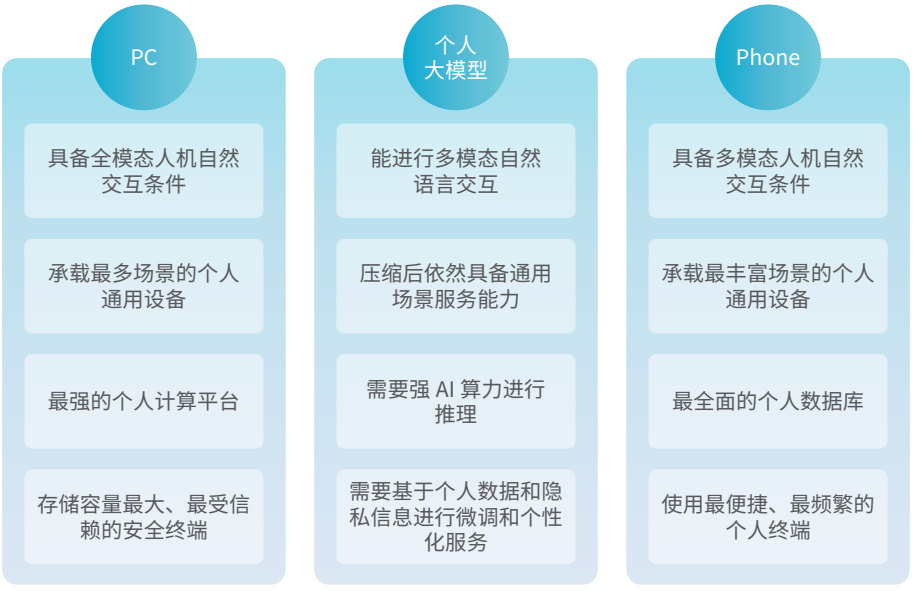


图 10 个人大模型与个人终端产品的结合

个人大模型为个人终端产品升级带来了新机遇，也为企业带来了全新的竞争环境和发展空间。从产业模式和用户功能看，大模型正在改变着终端产业模式。

02 个人大模型对个人终端硬件技术发展提出新要求

个人大模型对异构算力的本地端部署发展提出新要求，这是当前技术发展的必然趋势。随着个人大模型的运用，集中于云端的模式将变得无法满足需求，模型的 AI 算力都需要向端侧和边侧下沉。对于个人大模型的普及应用而言，终端侧算力支持是关键。当端侧内嵌的混合 AI 算力能够达到 10TOPS 时，已经能够在本地完成特定场景的 AI 模型推理，可以在设备智能管理、图像增强、游戏调优等方面提供支撑。当端侧的混合 AI 算力达到 40TOPS 时，使 AIPC 能够支持普通参数规模的本地模型推理，尽管依然需要 GPU 或云端配合才能完成更复杂的任务，但已经能够满足工作、学习、娱乐等场景的大部分 AI 创作类的需求。当端侧的算力进一步提升，AIPC 在端侧能够独立完成模型推理

的能力得到进一步增强，可以完全离线处理大部分复杂任务，终端在功耗控制、影像呈现、复杂运算、游戏体验等方面的表现也能够得到充分的 AI 优化。

在端侧与边侧构建足够的 AI 算力，提升本地模型推理能力，从而形成“端 - 边 - 云”协同的混合算力是行业内不可阻挡的趋势。在搭建本地智能算力上，以 CPU+NPU（神经网络处理单元）+GPU 异构式架构方案提供本地算力是目前最为成熟的方案之一。目前英特尔、高通、AMD 等处理器厂商均已经推出了集成 NPU 单元的处理器产品。

个人大模型的广泛应用，个人端的异构算力将持续发展，不断升级。

03 智能个人助理成为个人大模型应用的重要方式

智能个人助理作为人工智能驱动的代理，能从客户元数据、先前的对话记录、知识库、地理位置等信息中提取相关信息并生成个性化的响应。自苹果 Siri 亮相并引发广泛关注以来，智能个人助理便成为各大科技公司竞相追逐的重要技术领域。

智能个人助理的技术核心在于任务自动化，主要有模板编程、监督学习、强化学习以及大模型等四项技术。生成式大模型的兴起，为智能个人助理领域注入了新的活力，促使各大厂商和创业公司纷纷涉足，推出了一系列基于大模型的智能助理系统或应用。基于智能个人助理，个人用户将通过简单的自然语言与个人大模型进行交互；在个人大模型的赋能下，智能个人助理与用户的交互进一步升级。智能个人助理不仅更加精准地理解用户需求，还能执行更加复杂的任务，如智能查找文件、智能创作等。

然而，从长远的发展视角来看，要实现更高级的智能助理能力，仍需要不断提升智能体对用户和交互环境的感知与记忆能力，以及执行任务时的规划能力和记忆调用能力。此外，模型执行效率、安全性与隐私保护等诸多问题尚未得到完全解决。智能个人助理仍有巨大的发展空间等待探索。

洞向篇

第四章 企业大模型

随着人工智能技术的飞速发展，企业市场迎来一个全新的智能化转型时代，企业大模型也应运而生。

企业大模型是指企业根据自身独特的业务需求和知识体系，对通用大模型进行定制化的训练和优化后，形成的专门服务于企业内部各种业务场景的定制化、本地化或混合方式部署的人工智能模型。企业大模型作为企业智能化转型的关键工具，它能够在企业的经营管理、设计研发、供应链管理和生产制造等活动中，精准地处理和分析庞大的企业数据，不仅能够帮助企业优化现有流程、提升运营效率、降低成本，还能激发企业的创新潜力，帮助企业更好地适应市场变化，增强竞争力，在数字经济的浪潮中抢占先机，实现可持续发展。

01 大模型给企业智能化转型带来的新机遇

大模型的发展给产业链中的各类企业厂商带来了诸多机遇，推动了相关产业优化升级。

对于大模型提供商而言，相关算法和技术的不断升级可以帮助 OpenAI、阿里、百度等通用大模型提供商在激烈的竞争中保持并扩大市场份额，在行业中占据龙头地位；对算力提供商而言，大模型的发展催生了对巨大算力的需求，推动了阿里、百度、联想等企业的云计算服务的发展，也带动了英伟达、联想、浪潮等硬件厂商的 AI 芯片、AI 服务器等硬件升级；对软件提供商而言，大模型的发展推动其开发出更加智能化、个性化的产品，从而保持其在行业中的竞争力；对于集成服务商而言，大模型的发展推动其服务模式创新，从传统的软件集成和部署服务到基于大模型的行业解决方案的转变，帮助服务集成商在市场上占有先机；对应用企业而言，大模型在具体行业和具体场景的智能化应用，为相关企业实现自动化运营和效率优化，同时也为企业创造了新的收入来源，例如，大模型应用在工业质检领域，能够帮助企业更好地提取图像和数据的特征，不仅提升了质检准确性，还提升了工业质检的自动化和智能化水平。

需要强调的是，推动大模型在企业的应用，并不是通用大模型或基础大模型服务提供商，而是行业解决方案服务商。这是因为推动大模型在千行百业的应用，核心还在于行业知识和行业数据的积累，熟悉产业场景与行业 Know How 是关键。

随着大模型给企业升级带来的机遇，将推动企业形成以数据为核心的全业务、全流程和全成长周期的数字化重构，在加速企业技术创新的同时降本增效。这不仅有助于企业在行业中脱颖而出，甚至可能改变整个行业的竞争格局，进而增强企业在全国乃至全球的竞争力。

02 企业智能化转型的价值体系

大模型在企业智能化转型中的应用非常广泛，而且能够帮助企业在竞争激烈的市场中构建自己的优势。但是，大模型在企业的深度应用也带来众多挑战，比如，传统企业各级不同部门的职能会出现交叉和融合，研发将会真正以用户需求牵引为导向，工业品的生产流程将会缩短而且零部件数量将会减少，解决方案将会从分步解决到整体融合解决，供应链将会迁移与再升级等，企业将真正升级为全新的智能敏捷组织，进而形成全新的社会经济模式。

（一）经营管理

在经营管理方面，大模型可以帮助企业进行营销预测、客户管理、客户需求预测和员工业务培训，从而增强企业对市场动态和消费者行为的洞察力以及企业的经营管理效率。

在营销预测环节：大模型可以凭借其多角度的数据分析和处理能力，通过分析历史数据和行业数据，帮助企业预测更精准的销售数量和销售价格等趋势。

在客户管理环节：大模型可以最大化学习人类的思维和表达方式，提供智能客服服务；还可以通过深入分析行业和客户的数据，准确理解客户的言行举止、情感倾向和消费习惯，针对不同的问题提供个性化解决方案，从而提高用户体验，客户的积极反馈也可以反向提升产品和服务价值。

在客户需求预测环节：大模型可以分析客户的购买历史、行为模式和反馈信息等相关数据，预测客户可能感兴趣的新产品或新服务，从而为产品和服务的改进提供数据支持。

在员工业务培训环节：大模型可以进行实时评估和能力分析，自动生成学员能力模型和成长曲线，并生成个性化提升方案，帮助员工持续提升业务能力。

（二）研发设计

在研发设计方面，大模型可以帮助企业进行产品辅助设计、设计草图生成和仿真优化，从而快速迭代产品设计创新，降低企业研发成本，提高企业产品研发效率。

在产品辅助设计环节：大模型能够快速查找具有相似特征的产品模型，生成创新性的产品设计方案来辅助技术人员实现工程制图、模型分析等任务，从而提高产品的设计效率。

在设计草图生成环节：大模型能够帮助设计师根据输入带有产品特征的文本提示来控制图像的生成，从而帮助设计人员快速批量生成草图，显著减少了设计初期阶段所需的时间，同时为设计师的创造力提供了更多可能性。

在仿真优化环节：大模型能够替代传统计算仿真，具备更快的处理速度、更高的精度和适应性，提高了仿真效率。

（三）供应链管理

在供应链管理方面，大模型通过与企业资源计划系统（ERP）进行集成，利用 ERP 系统中的数据进行供行智能化分析，帮助企业进行订单管理、库存管理和物流管理，从而提高供应链的可视化、透明度、协同性和智能性。

在订单管理环节：大模型能够帮助企业实现自动化订单入库、验证和处理过程，减少人工参与，加速订单处理。

在库存管理环节：大模型可以预测库存需求，实现库存的动态管理，避免过剩或缺货情况，保持库存流动性；通过监测仓库环境和库存状况，大模型可以及时预警潜在的安全问题。

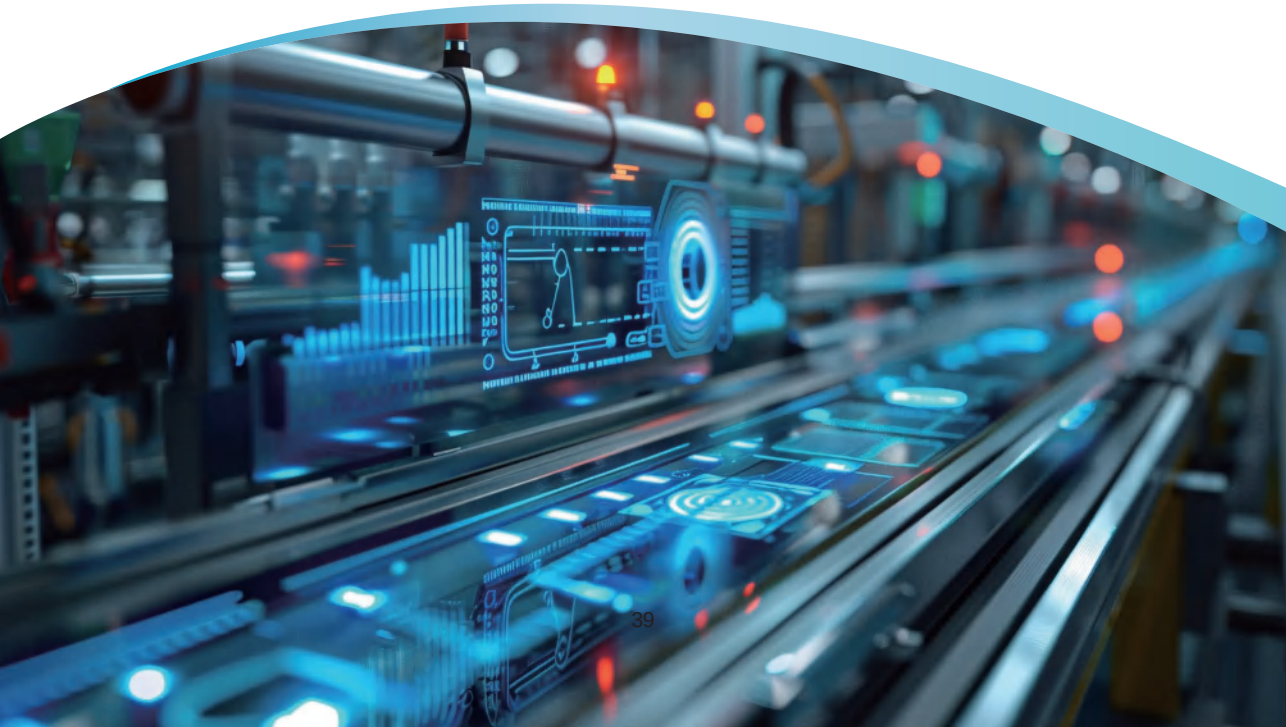
在物流管理环节：大模型可以预测和优化运输路线，减少因交通拥堵或不良天气带来的延误，从而缩短货物到达时间；并且通过分析历史数据和实时条件，提出成本效益最高的运输方案。

（四）生产制造

在生产制造方面，大模型可通过构建企业的生产管理、控制、运维等知识库，推动企业在运营管理、计划调度、质量检测和自动控制等场景的智能化，从而使得企业实现高度自动化的生产过程，显著提高生产效率和产品质量。

在运营管理环节：大模型通过企业知识库的训练，分析操作和运营人员的自然语言指令，提高信息查找和分析的效率，从而提高员工工作效率。

在计划调度环节: 大模型和制造执行系统（MES）融合，分析 MES 系统中的数据，形成智能制造执行体，预测生产需求和资源消耗，帮助企业优化生产排程和物



料管理等，对生产过程进行全面的分析和优化，从而提高生产效率和产品质量。

在质量检测环节：大模型可以将传统的质量管理系统（QMS）进行智能化重构，可以形成产品质量的视觉检测增强体系，使用大模型提供更强的视觉检测能力，用于质检、安全监测等。大模型还可以自动对检测的结果进行分析，并生成检测报告等。另外，在缺陷样本生成方面，大模型可以迅速生成模拟检测图像的缺陷样本，补充小样本的不足，提高模型准确性、缩短训练时间、提高训练效率等。

在自动控制环节：大模型与现有的控制系统集成，实现智能化的控制系统。在工业自动化和机器人技术等领域，控制对象往往具有高度的复杂性和非线性特性，可编程逻辑控制器（PLC）通过与大模型融合，自动生成 PLC 控制代码，能够提高开发效率，降低开发难度。

图 11 企业智能化转型的价值体系

大模型在这些典型场景中的应用，体现了在企业数字化重构以及提升生产力创新方面的关键作用，能够为企业智能化转型提供强有力的技术支持和新的发展动能。

洞向篇

第五章 企业大模型及智能体实践的方法与路径

作为一项新技术，大模型和企业业务融合是一场革新之旅。实践表明，企业在大模型上创新探索的角色不可替代。企业的关注重点不能放在基础大模型或通用大模型的发展上，而要不断积累行业知识与专业数据，熟悉业务场景与行业 Know How，将大模型与产业深度融合。只有将企业自身的数据与业务场景融合，才能真正帮助企业提高经营效率、提升客户服务质量、加速产品创新。因此，在智能化转型的征途中，构建融合企业自身数据与业务场景的企业智能体，才是企业应当着重聚焦并持续深耕的关键方向。特别需要注意的是，在构建企业智能体的过程中，需要根据不同场景来选择使用以大模型为代表的生成式 AI，或基于传统机器学习算法的判别式 AI。一般来说，生成式 AI 适用于自然语言交互、内容生成、知识检索等场景，判别式 AI 适用于生成制造、供应链中对准确率和精确性要求非常高的场景，如智能排产、产能或销量预测、物流路径优化等。

01 企业基于大模型构建智能体的步骤

企业落地大模型，要根据企业自身需求来确定。对于大型企业，当前通常有两种模式：一种是直接采购大模型构建企业的智能应用；另一种是先构建企业自己的大模型，然后在此基础上构建应用。

对于采购大模型直接构建企业智能应用的企业，首先是大模型选型，从市场上已有的开源大模型或大模型服务提供商中挑选一个基础大模型。这些模型已经经过大规模的数据训练，具备良好的语言理解和生成能力。企业可以根据自身业务需求，利用企业内外部数据，对这些基础大模型进行微调，以适应特定的行业知识和应用场景。

对于构建自己大模型的企业，则需要自己构建或从市场中选型大模型开发平台。

其次，选择算力并部署。由于大模型参数量庞大、结构复杂，企业需要提供高性能的计算资源支撑。高性能的算力通常伴随着高昂的成本，尤其当前处于 GPU 与 NPU 算力核心组件快速升级迭代的时期，企业需要选择合适的硬件配置和服务模式。企业可以直接选择第三方服务商的智算服务，也可以自建智算中心。需要注意的是，在第三方提供的智算服务中，除了传统的云计算服务商外，出现了一种新的第三方——它们是行业领先企业，自己建设了智算中心并训练了自己的大模型，拥有丰富的经验，因此它们把智能中心同公有云的智算服务整合，提供给行业或领域内的其他企业使用，这类智算服务更有行业特性，也可以称其为行业智算底座。

然后，选择基于大模型的智能体开发平台及运营和管理等工具平台，采用“五步法”来开发智能应用场景。“五步法”是指企业根据自身智能转型的场景需求，通过定场景、轻量微调、开发插件、知识整理和提示词生成的“五步走”方法落地大模型能力，不断实现 AI 技术的深度应用和业务价值提升。对于构建企业自己大模型的企业，通常需要在建设完成后再进入到该阶段。

在开发上，要基于统一方法进行 AI 应用开发，采用统一的编程接口、数据格式和模型架构，确保不同团队和项目之间的协同性。

当前，国内现有的 AI 原生应用开发平台大多都提供丰富的应用示例、可视化的应用快速编辑器和开源的应用代码框架，支持多种开发框架，提供开发、部署、发布、管理等一站式服务，简化了企业开发流程。例如，联想的智能体开发平台，企业可以结合自身业务需求，利用这些工具进行定制化开发。

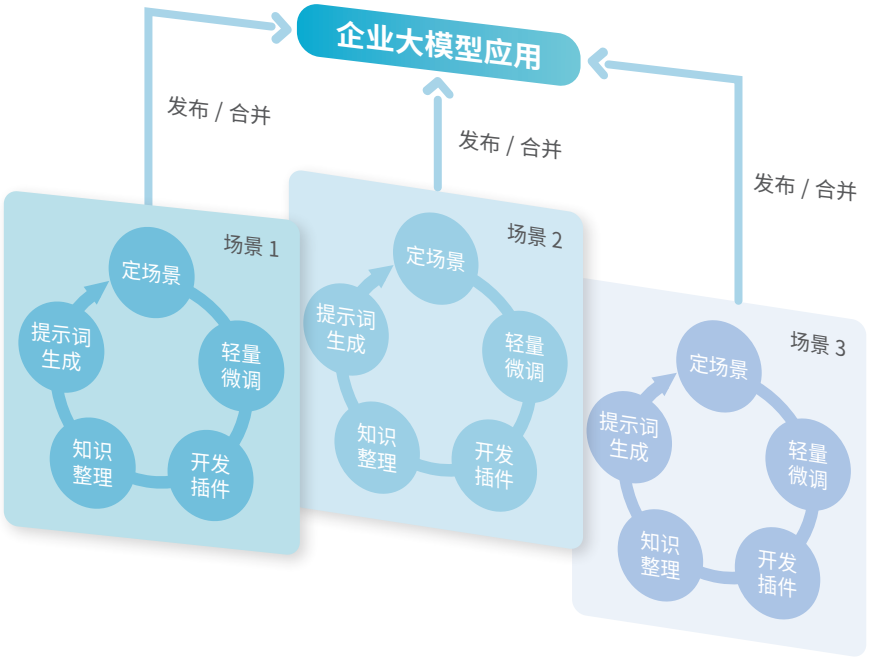


图 12 落地企业智能体的“五步走”方法

第四，企业大模型及智能体的运营与运维。企业需要在性能监控、数据安全等方面进行运维部署。在性能监控方面，可以建立一套完善的性能监控系统，通过 API 或 SDK 等方式对大模型及智能体进行全局监控，及时发现并报告性能异常行为；另外，企业必须确保数据的安全性和合规性，对于难以实施本地部署的企业，可以选择在安全且可隔离的专属数据存储空间中存储和处理数据。通过服务器端的加密技术，这种部署方式能够确保数据的高度安全性和符合法

规要求，有效保护企业数据。

随着产品技术与服务的不断发展，这些流程与步骤也会不断融合、迭代，形成新的服务或模式提供给企业客户。

02 实践案例：联想的智能化转型及联想企业智能体

联想作为中国信息技术领域的领军企业，致力于推动国内大中型企业的智能化转型。联想的客服服务网络已实现全国范围内的广泛覆盖，拥有服务站点逾 4,400 家，管理的企业数据量超过 130PB，管理的算力设备数量超过 5 亿台，专业服务工程师规模达到 24,000 余人。在人工智能在线服务方面，联想集团年均服务次数超过 1,100 万次。

基于 AIPC 的个人智能体和基于 AI 原生应用或实体的企业智能体的结合，体现了联想智能体的核心定位。联想希望通过二者的结合，以智能体为载体成为大模型为个人和企业提供服务的重要方式，帮助企业大模型落地。为此，联想把擎天智能 IT 引擎升级为以企业大模型驱动、智能体为核心的“擎天 3.0”。

联想“擎天”架构的核心是“一擎三箭”。“一擎”是指基于端、边、云、网、智新 IT 架构的擎天智能 IT 引擎；“三箭”是面向三大客群的业务部署，包括

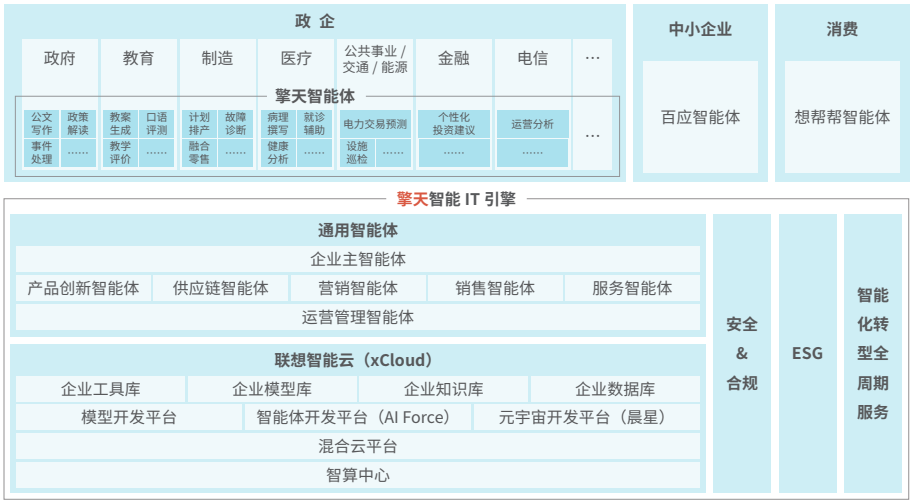


图 13 联想“擎天”架构

为政企客户量身定制的行业智能体解决方案、为中小企业客户打造最懂中小企业的联想百应智能体，以及为消费者提供内嵌主动式智能服务的小天智能体。

擎天智能 IT 引擎的核心架构，是基于联想智能混合云精心构建的智能体引擎，它作为多个智能体的运行平台，承载着企业智能化转型的重任。这个引擎分为上下两层：下层为技术平台，包括 AI 的基础技术平台，如 AI 开发工具平台、智能体运行开发平台和元宇宙开发平台等，并融入了与大模型紧密相关的企业工具库、企业混合大模型和企业知识库平台；上层为业务平台，专注于为不同业务场景量身打造企业通用智能体。

联想在智能化转型的探索过程中，结合多年的行业经验，将一站式订阅的臻算服务和以大模型为基础的端到端全流程解决方案和 AI 原生技术架构等新技术、新产品应用到企业内部的智能经营管理中，又通过定场景、轻量微调、开发插件、知识整理和提示词生成的“五步走”方法落地大模型能力，将大模型真正与企业应用相结合，完成了智能客服、智能营销预测、智能质检和智能供应链等多业务场景的智能体。这一系列的创新实践，不仅显著提升了联想的运营决策效率，更为大模型在企业中的落地应用提供了一套既完整又高效的解决方案，

确保了企业在智能化转型道路上的平稳前行。

在此基础上，联想更进一步，依托自身丰富的实践经验和深厚的行业积淀，为企业提供涵盖智能化转型全周期的专业服务。这一服务旨在全面响应混合 AI 部署对全面性和系统性的高要求，助力企业系统架构顺利向 AI 原生转型，从而在新的智能化时代中抢占先机，实现跨越式发展。

目前联想方案服务已基于擎天智能 IT 引擎形成了行业智能体、通用智能体、“四库三平台”（工具库、知识库、数据库、模型库，智能体平台、元宇宙平台、模型开发平台）等方案体系，为政企用户提供数十种场景智能体的服务产品。

面向中小企业客户群体，联想打造了百应 IT 服务智能体，并基于模态交互、可视化思维链、多方案博弈、多智能体协作、全链路安全等五大技术能力，率先推出 AI 营销、AI 办公与 AI 服务三大核心 AI 应用。这些创新应用旨在破解中小企业面临的 AI 技术引入门槛高、资源获取难度大、实际应用效能提升缓慢等难题，全方位赋能中小企业在 AI 时代的资源拓展、效率提升与成本控制。在 AI 营销领域，借助智能客户追踪、精准用户画像构建及高效客户圈选功能，营销流程得以显著简化，助力企业获客效率实现数倍增长；对于 AI 办公场景，平台支持包括语音、文本、图像在内的多元化交互模式，营造出深度沉浸的人机协作氛围，有效促进员工沟通与协作效率翻倍提升；在 AI 服务层面，平台提供了包括 AI 图像识别诊断、深度逻辑推理、问题精准定位等一系列高级功能，能够迅速响应并解决各类技术难题，大幅提升故障诊断的时效性和准确性，同时实现运维成本的大幅降低，降幅可达 50% 之多。

面向个人消费客户，联想将企业通用智能体中的服务智能体融入小天智能体中，发挥已沉淀的服务知识库和服务大模型的能力，使消费者服务更加智能。在此基础上，联想更进一步，致力于提供具有前瞻性的主动式设备服务，通过深度感知用户场景、精准理解用户意图，并结合多模态的人机交互技术与 AI 智能调度系统，能够在用户尚未主动寻求帮助时便预见问题，主动触达用户解决潜在问题，提升服务的及时性与针对性。

03 大模型在行业智能化转型的典型场景应用

大模型在千行百业智能化转型的应用，将加速局部、单点场景的智能化向全流程智能化方向发展，从而极大地推动制造业、医疗、教育、能源、金融、政府服务等各行各业的智能化转型。

智能经营管理：随着生成式 AI 飞速发展，某国有大型装备制造企业成立了 AI 团队，开始探索 AI 如何加速企业智能化转型步伐。成立伊始，团队坚持“技术服务业务发展”的思路做了深入的业务 AI 需求调研，总结出 23 个 AI 应用场景，快速展开市场调研与技术预研，最终以渐进式的方式选择提升整体 AI 形象的交互式人工智能数字人项目作为试点，拉开了该企业新型 AI 应用的序幕，不仅为企业智能化形象带来了提升，也为了解、掌握 AI 能力提供了练兵场，更通过项目建设发现、培养了更多的 AI 落地实操人才。

智能客户服务：联想应用大模型进行智能客户服务，不仅提升了客户体验、增强了客户忠诚度，也提高了经营管理效率。在智能客服方面，联想将大模型应

用于“魔方”客服系统，为客户提供全时智能客服，精准识别客户的产品使用需求并为座席提供解决方案，在结束单次客服后自动提取关键信息，创建服务工单，不仅提升用户的产品体验和员工的工作效率，更为公司的客户经营管理节约成本。在辅助营销方面，联想 MarTech 平台通过对用户基本属性、商品属性和行为特征等人群标签的提取，使用户画像标签生成的效能获得极大提升，不仅时间成本减少到 1/3，销售转化率也从 0.28% 提高到 1.93%，提升 6 倍以上。

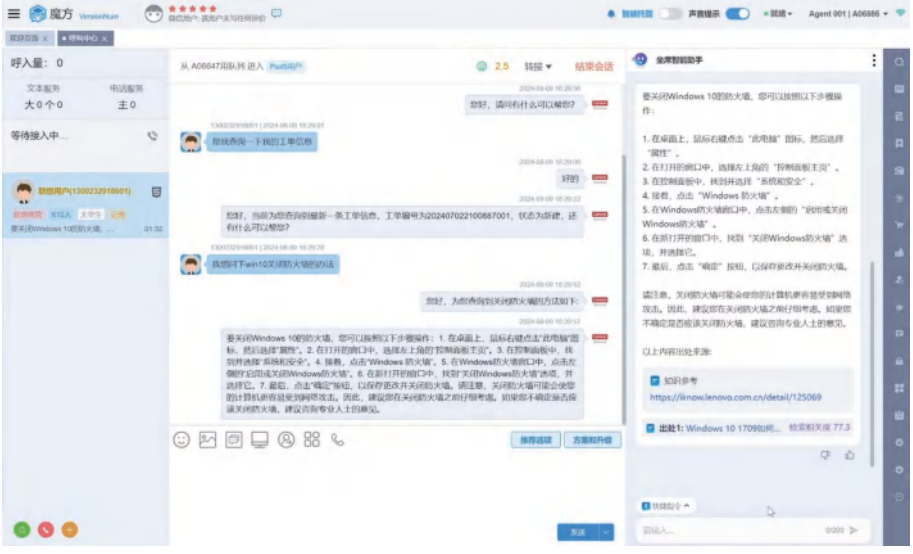


图 14 “魔方”客服大模型自动识别问题，并为座席提供解决方案

智能设计研发：国内航空制造企业与第四范式在 CAD 工业软件上进行共创，引入“式说”大模型，通过语音提问，找到所有与目标设计相类似的三维数模零件，给出多种数模组装方案，使传统的开发形式向数字化转型，提高了研发效率，缩短了研发设计实现周期。

智能供应链管理：顺丰科技通过大模型技术，推出了多项智慧化服务应用。顺丰的小哥服务中心集成了 AI Agent 能力，可以驱动 API 调用和知识库检索，

准确提取收寄标准、运费标准、收派操作等问题中的关键信息，为快递员生成简单易懂的业务解决方式；其智能通系统从各国海关官网不间断地爬取、整理、解读最新的关务规则与收寄标准，并自动转译、分析和提炼，帮助员工快速理解和应用复杂的国际贸易规则，显著提升了物流管理的效率和服务质量。

智能生产制造：联想在产品的质量检测环节，大模型辅助自动化光学检测（AOI）系统进行光学检测，从而高效确保了产品质量的高标准。联想通过大模型对捕捉到的屏幕图像进行处理，识别出屏幕缺陷，如划痕、污渍、色差等，从而使 AOI 系统能够达到每小时超过 300 台笔记本电脑屏幕的检测速度，识别更加准确且大大节省人力成本，实现了极低的误判率和过杀率，从而显著提高生产效率。

工艺优化：某省烟草制造工厂在智能化转型中，巧妙融入智能体技术，实现了从原料加工到成品产出的全流程智能化管理。该工厂基于算法开发平台和智能体开发平台，开发了专属的人工智能算法和智能体，并建立了自己的算法管理平台与智能知识库，使制丝生产线上的设备实现了自我监测与调节，并实现了产线整线水分稳态预测与控制，有效提升了生产精度与效率。同时，构建的工艺质量智能管控系统，结合大数据分析与机器学习算法，为产品质量的持续优化提供了科学依据。这些智能化技术升级改造，不仅加速了生产决策的制定，还增强了企业的创新能力与市场响应速度，为烟草制造业的智能化发展树立了新标杆。

智慧医疗：成都某高端综合三级甲等医疗机构，通过多款智能体和 AI 应用，显著提升了服务品质与管理效率。部署了营销顾问智能体，对销售流程对话的深度语音分析，精准提炼交流精髓，生成阶段性客户画像，为实施标准化、持续性的客户关系管理提供了有力支持。客户服务智能体通过智能分析客户的情绪波动、具体需求及潜在顾虑等信息，定制出更为贴心的诊后服务方案。在医生问诊环节，引入了智能辅助系统，能有效沉淀并提炼大量客户交流信息，生成标准化的综合诊断记录，结合多维度的客户画像，智能辅助医生进行数据分析，不仅提高了问诊效率，还显著增强了诊断的专业性和精确度。此外，该机

构的医疗管理智能服务，智能关联各项指标与维度进行多维度计算，并能自动识别衍生指标的计算逻辑，实现分析结果的即时输出，极大地增强了数据驱动的决策支持能力。

智慧教育：国内某顶尖综合性研究型大学，通过实训智能体，为老师和学生提供了人工智能教学与实践所需的开发和运行环境，推进了产学研一体化教学模式，紧跟技术趋势，增强了学生的动手能力和实践应用能力。河南省某市教体局，通过 AI 智能体以及基础和教育行业大模型，支撑当地数字教育项目，围绕“教、学、评、管、测”五大业务类，实现 18 个大模型实践场景。通过个性化教学和即时反馈，教师调整教学内容，学生获得个性化指导，提高了学习效率和兴趣。通过数据统计和分析支持科学决策，优化了学校的管理流程，提升了整体教学质量和管理水平。

洞向篇

第六章
大模型的未来

大模型作为人工智能领域的一项突破性技术，正在推动生产力的变革，并与新质生产力的发展紧密相连。大模型通过深度学习和海量数据训练，展现出强大的泛化能力和模式识别能力，不仅提升了整个社会生产效率，还促进了新产业、新模式、新业态的发展，推动了传统产业的数字化转型。大模型正在加速发展并同千行百业融合。

另一方面，当前的自回归大型语言模型（AR-LLMs）虽然展现出在文本生成和对话交互等方面的卓越能力，但在知识、记忆、推理、计划等能力上存在明显短板。因此，基于其他架构的大模型也在不断被探索。

01 大模型未来三大发展趋势

当前，随着大模型迭代、算力提升和数据量的增长，大模型性能将快速提升，

部署门槛逐渐降低，应用场景不断丰富。

(一) 通用性进一步提升，全面对齐人类认知

随着模型版本迭代和数据规模与质量的提升，大模型通用性将进一步提升。未来大模型参数规模将不断扩大，泛化能力进一步增强多任务学习能力持续提升，其能力将在更广泛的范围内全面对齐人类的认知水平。语言能力方面，大模型将能够更深入地理解和生成复杂的语言结构，更准确地把握语境和情感，在文学创作、商业文案、法律分析、哲学思辨等高层次认知任务上展现出更加卓越的性能。多模态方面，未来除了现有的文本、代码、图像、视频和音频等能力之外 , 大模型将拓展更多身临其境的模式，包括调动三维视觉、嗅觉和味觉等其他感知方式，通过对不同类型数据的深层理解和综合分析，提供更加全面、准确、细腻的信息处理和决策能力，更好地服务于各个领域。

(二) 模型进一步轻量化，搭载终端更加灵活

Transformer 已成为当前几乎所有大模型的基础算法，但其自带的注意力机制在涉及长序列输入时，计算量呈指数级增长，极大地增加了训练和计算成本，抬高了大模型开发和部署门槛。Transformer 的计算效率不高也导致目前新一轮人工智能热潮下的全球算力短缺。未来，在强大的通用底座型大模型不断发

展的同时，模型向轻量化发展是另一个重要趋势。随着新的基础算法的提出以及模型剪枝、知识蒸馏等算法优化技术的不断进步，大模型在保持高性能的同时，体量不断减小，算力需求大幅降低。轻量化的大模型将能够更加灵活地加载到各类终端设备，直接调用端侧、边缘侧算力进行调优和推理，极大增强了模型灵活性和实时性，同时也更好地保护了用户隐私，拓展了大模型的应用范围和场景。

(三) 融合程度逐渐加深，落地场景不断丰富

大模型技术和原生应用在不断发展的同时，将持续与其他技术深度融合，赋能越来越多的行业，在更多场景深入应用。例如，大模型与机器人技术相结合，为人形机器人提供通用人工智能大脑。大模型与虚拟现实、增强现实技术结合，提供智能化的沉浸式感官体验，开拓新的教育、娱乐等场景。大模型与物联网技术相结合，能够对家庭、社区乃至城市的数据进行处理和决策，更好地服务居民。随着大模型技术的创新发展和融合应用，未来其在金融、医疗、制造、教育、娱乐、交通等领域实现更加丰富的落地应用，为社会带来深刻变革。

总而言之，未来大模型的发展将是通用性与专业性协同、轻量化与性能提升共存、单一技术与跨领域融合并进的过程。在这一过程中，大模型将不断超越当前的局限，展现出更加强大的能力和潜力 , 为人类社会的发展带来更加深远的影响。

02 去概率化大模型成为大模型发展的主要框架

现有的大模型，尤其是自回归大型语言模型，在文本生成与对话交互等领域表现出色，然而亦存在诸多不足与挑战，亟待解决。

常识与理解有限。尽管大模型能够生成流畅的文本，但其对现实世界的理解仍显浅薄，对知识的认知仅限于训练数据中的模式，缺乏现实的世界知识，对特定领域的知识认知十分有限，也很难真正理解人类和动物的一些非语言行为。

缺乏常识导致其在文本生成过程中，出现事实错误、逻辑错误，从而产生不合逻辑或不切实际的回答，降低了其信息输出的质量，影响了模型的可靠性和可信度。

推理与记忆能力受限。大模型在记忆和推理方面远不及人类，无法有效利用先前经验指导当前决策，这影响了模型在复杂场景下的表现。现有的自回归模型，如 Transformer 架构，模型的自注意力机制虽然能捕捉长距离依赖关系，但当其处理过长序列或需要跨越大量文本进行规划和推理时，计算量大增，效率降低。

缺乏目标导向性。自回归模型是通过预测下一个词的概率分布来生成文本的，这种方式非常适合产生流畅的文本，但是模型的训练目标是最小化每一步的预测误差，而不是优化整个文档或对话的结构，因此缺乏对整体结构和目标的考虑。

因此，在大模型现阶段的局限下，去概率化大模型被认为是一种新的发展趋势，联合嵌入预测架构等基于目标驱动的人工智能新架构已崭露头角，这些新模型和架构旨在通过去概率化的方式增强学习、加强记忆与推理能力，以及更好地识别和理解任务目标，从而提升大模型的整体性能和实用性。

去概率化大模型不单纯基于概率进行预测，而是通过规则与学习结合的方式，从而提高模型的记忆、推理和决策能力。因为人类学习不仅依赖具体例子，还依赖通用知识和经验。逻辑规则代表了高水平的认知和结构化知识，对学习过程有显著帮助。比如概率规则学习系统，结合概率论和规则学习的优势，使用概率来量化规则的不确定性，使我们能够从不确定的数据中学习并作出推断。通过这种方式，去概率化大模型将符号知识整合到深度学习模型中，以提高模型的推理能力和可解释性，从而在各种领域中提供有价值的见解，并支持更加精确和可靠的决策过程。

03 目标驱动的人工智能新架构

目标驱动的 AI 系统是一种新型的人工智能架构，其核心理念在于能够识别和理

解给定任务的目标或目的，从而提供融合记忆、学习、推理、能够安全操控的人类智力水平的人工智能助手。通过深入剖析任务需求，AI 系统能够清晰地把握行动方向，并使用这些目标来引导其行为和决策过程。这样，无论面对多么复杂的问题，它都能进行长期的规划，并在遭遇挑战时灵活调整策略，确保每一次行动都与实现预定目标保持一致。

目标驱动的人工智能新架构的实现是一个复杂而精细的过程，它涉及构建包含感知、世界模型和行动选择等多个组件的完整系统。首先，感知组件发挥着至关重要的作用，它负责深入理解当前的环境状态以及任务所设定的目标。通过这一组件，AI 系统能够获取到足够的信息来为后续的行动做出明智的决策。接下来，世界模型组件则负责预测不同行动序列可能产生的结果，并评估这些结果与预定目标的一致性。通过构建这样的模型，系统能够预测未来的状态，从而选择最有可能实现目标的行动路径。最后，行动选择组件基于世界模型的预测结果，选择出最优的行动方案。这一组件能够确保 AI 系统的行为始终与目标保持高度一致，从而实现真正意义上的目标驱动。通过这种实现方式，目标驱动的人工智能新架构能够在实际应用中展现出强大的性能，为解决复杂问题提供了有效的手段。

04 相关研究与实践

图灵奖得主杨立昆教授提出了一种深度学习基于能量的方法——联合嵌入预测架构 (JEPA)。

JEPA 是一种能量模型。与概率模型相比，基于能量的模型 (EBM) 不需要进行概率分布的标准化处理，而是直接通过能量函数来度量变量之间的一致性或不相容性，从而用于预测、分类或决策任务。

由于人类通过被动地观察世界来学习大量关于世界的背景知识，而这种常识性信息是实现智能行为的关键，因此模型必须以自监督的方式学习世界知识表征。JEPA 模型通过自监督学习的方式，使机器通过观察和交互，创建能够理解和预测外部世界变化的内部模型。在涉及图像、视频等模态的数据内容时，通过遮罩策略指导模型产生语义表示，使其在高抽象水平上预测表征，而不是直接预测像素值，从而消除不必要的像素级细节。在提取数据进行训练时，JEPA 架构只从数据中提取相对容易预测的信息，比如，一辆自动驾驶汽车在马路上行驶，马路两侧可能有树木，并且天气可能是大风天气，树木上的叶子会以一种无法预测的随机的方式移动，这种对主场景不会造成重要影响的随机性细节会被编码器视为噪声消除，可以建模和被预测的关键内容被保留下来。

JEPA 不依赖于特定的数据增强或预处理步骤，使得该架构易于适应不同的数据模态和任务，因而更具有灵活性和泛化能力，有望在未来广泛地应用到图像、音频和视频乃至自动驾驶等具体场景，是通过对世界更为深入的理解来推进机器智能的重要一步。

结语

大模型技术作为人工智能发展史上最重要的一次突破，已经引领我们进入了全新的 2.0 时代。而大模型 2.0 时代，最显著的特征就是多领域的发展趋于商业化成熟。

本报告通过深入分析大模型 2.0 的内涵、特点和产业生态，揭示了其在推动社会进入智能时代的演进和变化。我们将见证在大模型推动下，个人生产力的显著提升，企业向全栈智能化的转型，以及社会生产方式的变革。同时，各国政府对于大模型技术的支持和监管政策也日益密集，旨在为这一新兴领域提供健康的发展环境。

在大模型 2.0 时代，数据、算力、算法和工具成为了基础关键要素，它们共同构成了大模型发展的基石。通用大模型和行业大模型的持续发展，企业大模型的商业化加速落地，以及个人大模型产品的涌现，都昭示着一个更加智能化、个性化的未来愿景。而在这个过程中，产业标准、数据安全、伦理治理和价值对齐成为了大模型可持续发展的重要保障。

个人大模型的兴起，为个人用户带来了更加个性化和智能化的服务体验。从智能手机到智能家居，从在线教育到健康监测，个人大模型正在成为我们日常生活中不可或缺的一部分。它们不仅能够提供更加精准的信息推送和决策支持，还能够根据个人偏好和行为习惯，为我们量身打造个性化的产品和服务。在这个过程中，个人大模型不仅提升了我们的生活品质，也推动了商业模式的创新和产业模式变革。

企业大模型的发展，为传统行业带来了前所未有的转型机遇。通过深度学习和大数据分析，企业能够更精准地洞察市场动态，优化决策过程，提升运营效率。在这个过程中，企业不仅能够实现成本的降低和利润的增长，更能够通过流程再造和价值再造，探索全新的商业模式和增长点。这种由内而外的变革，正在

重塑企业的核心竞争力，为企业的可持续发展注入新的活力。

然而，随着大模型技术的不断发展，我们也必须面对概率模型的局限性。传统的大型语言模型（LLM）虽然在过去取得了显著的成就，但在处理复杂任务和确保结果的稳定性方面，已经逐渐暴露出不足。因此，新能量模型的正则化成为了行业发展的新趋势。这种新型模型不仅能够提供更加准确和可靠的输出，还能够在保障数据安全和伦理合规的前提下，推动人工智能技术的健康发展。

在此，我们呼吁全球的政策制定者、研究机构、科技企业和公众，共同关注大模型技术的发展，共同探索和应对伴随而来的挑战。我们需要坚持创新与责任并重，确保技术的发展能够惠及全人类，促进社会的公平与进步。让我们携手合作，共同推动大模型技术的健康发展，为建设一个更加智能、高效、和谐的未来社会砥砺前行。

随着本报告的发布，我们希望能够为大模型 2.0 时代的探索者和实践者提供有价值的参考和启示。让我们在这个繁花落尽、繁星升起的新时代，共同迎接挑战，拥抱机遇，催生人工智能的下 一轮潮涌。