CAICT 中国信通院

中国人工智能产业发展联盟
Artificial Intelligence Industry Alliance

集智 专题报告

大模型一体机应用研究报告

（2025 年）

中国信息通信研究院人工智能研究所
中国人工智能产业发展联盟

2025年10月

版权声明

前　言

在“人工智能+”的政策背景下，大模型技术快速发展，成为推动产业智能化升级的核心引擎。大模型一体机作为一种集成化、场景化的产品形态，凭借其行业化落地快速、安全可控、易用性强等优势，正成为促进人工智能与实体经济深度融合的关键基础设施，为千行百业的智能化转型提供高效、便捷的技术支撑。

过去一年，全球大模型技术与产业生态加速演进，大模型一体机作为技术落地的重要载体，在技术、应用与产业协同层面迎来重要进展。**技术层面，**大模型一体机通过软硬协同设计、优化算力调度与资源管理、集成全栈开发工具链与预置行业模型，支持数据处理、私有化部署和全开发流程加速，持续突破高性能与高安全性的技术瓶颈。**产业层面，**服务器厂商、云服务提供商、行业应用开发商及大模型技术供应商等纷纷入局，硬件-软件-服务一体化产业链初步形成，围绕大模型一体机的研发、生产、部署和服务的产业链条正在快速形成并联动发展，产业生态加速完善。**应用层面，**大模型一体机加速向千行百业渗透，覆盖政务、医疗、金融、制造、能源等关键领域，降低行业企业智能化应用门槛，各类创新应用层出不穷，规模化效应逐步凸显。

总体来看，我国大模型一体机技术能力持续突破、产业生态初具规模、应用场景百花齐放，但仍面临技术自主创新能力较为薄弱、应用场景适配难、安全隐私保障机制待完善等挑战。展望未来，随

着大模型技术突破和行业需求爆发，大模型一体机有望成为大模型技术普惠化的重要突破口，为“人工智能+”行动提供坚实支撑。

在此背景下，本报告深入剖析了大模型一体机的技术演进、产业发展动态与应用实践，深入分析其典型场景、选型策略和落地路径，旨在为企业应用大模型一体机提供全面参考。同时，展望我国大模型一体机技术产业化发展的新趋势，助力构建自主创新、安全高效的智能化生态体系。

目　录

图 目 录

一、大模型一体机发展概述

近年来，随着大模型技术加速产业智能化升级，大模型一体机通过集成算力、算法与行业解决方案，显著降低企业部署门槛，推动人工智能技术规模化落地。当前，大模型一体机市场格局初步形成，技术路径趋于多元化，硬件加速与软件优化协同发展，逐步构建起覆盖训练、推理及行业应用的全栈能力。作为人工智能基础设施的重要形态，大模型一体机正加速重构产业生态，为下一代人工智能应用提供核心支撑。

（一）大模型一体机的定义

大模型一体机是一种高度集成的、提供大模型应用能力的系统。它通常采用私有化部署方式，封装人工智能应用所需要的复杂组件，提供一种简化、高效且安全的部署与运行环境。其核心理念在于对硬件资源、软件资源、模型资源及垂直领域应用进行深度整合与协同优化，构建一个易用性高的一站式人工智能解决方案，降低企业或机构部署和应用人工智能技术的门槛。从功能上看，大模型一体机融合了高性能人工智能服务器的计算能力与大模型私有化部署的特性，促进人工智能技术更快地在组织内部转化为实际应用和业务价值。

大模型一体机的核心价值在于其系统性地解决了人工智能大规模应用与落地中的若干关键瓶颈与挑战。**首先，大模型一体机显著简化部署流程与运维复杂度。**通过高度预集成化的系统设计和全面的软硬件优化，大模型一体机大幅降低了用户部署和配置人工智能

基础设施的复杂度，有效规避了传统模式下繁琐的硬件选型、兼容性测试、环境搭建、软件栈配置与调优等环节，实现快速部署应用能力。**其次，大模型一体机促进性能与效率的系统级提升。**大模型一体机核心优势源于深度的硬件和软件协同设计，通过对算法、软件与硬件的整体优化，超越单一组件性能极限，实现系统级的性能峰值与能源效率。**此外，大模型一体机能够强化数据安全与合规保障。**大模型一体机普遍支持私有化、本地化部署模式，使得企业或机构能够将敏感数据和核心模型严格控制在自身物理或逻辑安全边界之内，为对数据主权、隐私保护有严格要求的行业，提供符合其合规标准的、安全可控的运行环境。**最后，大模型一体机能够提供高度的应用定制化与业务融合能力。**大模型一体机强调场景化赋能，提供便捷的工具支持用户利用自有数据对模型进行高效微调，或部署私有模型，确保人工智能技术能紧密贴合具体业务需求，有效解决特定流程中的挑战。

（二）大模型一体机的发展背景

2024 年至 2025 年，大模型一体机市场呈爆发式增长，大模型一体机的兴起源于人工智能技术演进与产业需求变革的双重驱动。

从技术发展来看，底层技术持续突破推动大模型一体机的能力提升和成本重构。随着大模型技术逐步从“科研突破”走向“工程部署”，模型开源、压缩与推理优化等一系列关键技术的成熟成为一体机发展的技术底座。一方面，DeepSeek 系列开源模型的发布，使开源模型能够达到与闭源模型相近的效果，降低了优质模型获取

的成本，在提升大模型一体机业务效果的同时，降低了落地成本。另一方面，随着大模型推理优化技术的快速发展，KV Cache、模型压缩等技术通过减少冗余计算，提升推理速度和吞吐率，提升了大模型一体机的应用效果。同时，开源生态推动企业快速适配，形成“模型-工具链-硬件”的协同创新，同时降低了大模型一体机的部署及使用成本。规模效应与技术创新共同推动大模型一体机成本下降，一体化设计减少了复杂的系统集成和调试环节，从而缩短了部署周期，节约了人力成本。此外，大模型一体机通常提供灵活的订阅或租赁模式，将一次性高额投入转化为可控的运营支出，极大地降低了企业的初期投入门槛。

从产业层面来看，产业链上下游的加速协同与创新融合促进大模型一体机快速发展。随着人工智能产业化进程不断深入，硬件供应商、软件供应商、模型供应商、应用供应商等形成紧密协作的产业生态，共同推动大模型一体机实现从技术研发到商业落地的全链条创新。在硬件生态方面，多元算力协同格局正在形成。硬件供应商如华为昇腾、寒武纪等通过开放指令集架构，与整机厂商深度协同优化。在软件生态方面，平台工具链的生态共建形成协同效应，华为推出全栈国产化训练推理加速套件以提升开发应用效率，百度针对一体机场景推出轻量化训练推理工具，提升易用性。在模型生态方面，模型生态呈现分层协作特征，模型供应商如深度求索、阿里云通过开源模型，与应用供应商共同开发垂直场景解决方案。产业链的完善和生态系统的构建为大模型一体机的发展提供了保障，

不仅促进了技术的共享与合作，还推动了整个行业的快速发展。

从政策层面来看，多项政策推动大模型一体机市场繁荣发展。近年来，政府从研发、应用和基础设施多层次协同发力，推动大模型技术创新和产业化发展，从而促进了大模型一体机的快速发展。**在研发方面，**2022 年科技部发布了《关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见》，强调场景创新驱动技术升级和产业增长，鼓励企业通过强化基础理论与关键技术攻关，提升创新水平，加大应用示范效应。**在应用方面，**2025 年国务院发布了《关于深入实施“人工智能+”行动的意见》，强调以人工智能引领“新质生产力”，引导各行业加快智能化转型。在国家顶层设计的引领之下，地方政府亦加大布局大模型一体机建设。深圳市政府发布《深圳市加快打造人工智能先锋城市行动计划（2025—2026 年）》，明确提出要“培育发展大模型一体机，重点开拓一体机在金融、政务、医疗等本地部署需求旺盛的应用市场。”系列政策推动了大模型一体机的应用需求。**在安全合规方面，**全国人大常委会发布《中华人民共和国数据安全法》，要求数据处理者采取技术和管理措施保障数据安全，推动一体机强化本地化数据存储与加密技术，确保训练和推理数据不出域。同时，《中华人民共和国个人信息保护法》也明确了处理个人信息应当遵循的原则和要求，而大模型一体机通过本地化部署的方式，能够更有效地实现个人信息的严格隔离和精细化访问控制，从而有效规避数据泄露风险，在法规的推动之下得到了更有利的发展。

从应用需求来看，行业场景驱动需求爆发推动大模型一体机的井喷式增长。随着各行各业对智能化升级的需求不断加深，大模型从“泛化工具”逐步走向“场景专家”，企业与机构对于低成本、低延迟、强隐私保护的本地化人工智能能力提出了更高要求，而大模型一体机正好在此背景下应运而生。首先，对私有部署与数据安全的强需求，使大模型一体机在政务、金融、医疗等高敏感领域率先落地。在政务领域，国内多城市已部署多台大模型政务一体机，支持本地公文生成、政策解读、政民互动等任务，有效提升政务服务的智能化水平。在金融领域，多家银行部署 AI 投顾与智能客服一体机，实现对客户需求的实时响应与智能分析，同时满足银行对合规与低延迟响应的双重要求。此外，针对行业用户降低技术门槛、提升管理效率的综合需求，大模型产品形态从平台到一体机逐渐演化。大模型一体机“软硬一体”的形态，封装了模型加载、推理加速、知识集成和接口调度等复杂流程，降低了行业用户的使用门槛，大大缩短了从部署到见效的时间周期。许多垂直行业厂商也开始围绕行业需求定制大模型一体机方案，例如政法一体机、医疗助手一体机、呼叫中心一体机等产品形态不断丰富，形成“场景即服务”的生态格局。

二、大模型一体机技术架构及选型参考

随着大模型一体机从概念验证迈向规模化落地，其技术架构持续演进并形成差异化发展路径。当前主流大模型一体机已突破传统服务器的设计范式，通过“硬件-软件-模型-应用”架构的垂直整合，

构建起覆盖数据处理、模型训练、推理加速及场景落地的全栈技术体系。产业已衍生出多种大模型一体机的产品形态，在选型的过程中需兼顾算力规模与成本效率、算法兼容性与生态成熟度、平台可扩展性与运维简易性等因素，来为各行各业提供稳定、高效、可控的整体解决方案。

（一）大模型一体机技术架构

大模型一体机的技术架构遵循分层设计、软硬协同、深度优化的原则，构建从底层硬件资源、软件资源、模型资源到上层智能应用的全栈式、垂直整合系统，如图 1 所示。这种“硬件-软件-模型-应用”的技术架构能够为大模型的开发、训练微调、推理及全生命周期管理提供一个高效、稳定、安全且易于使用的平台，有效支撑大模型的应用任务。

图 1 大模型一体机技术架构

硬件层作为物理基座，提供核心的计算能力、存储能力以及网

络能力。计算资源方面，大模型一体机的 AI 计算能力通常由一种 AI 加速硬件组成，例如通用图形处理器（GPGPU, General-purpose computing on graphics processing units）、面向 AI 任务优化设计的各类神经网络处理器（NPU, Neural Processing Unit）等。**存储资源方面，**大模型一体机需满足大模型对海量参数和数据的存储需求，达到大容量、高性能和可扩展的要求。对于常规场景的持久性存储需求，通常采用 NVMe SSD 的存储方式。针对性能要求高、可扩展性强的场景，需要采用专用的 AI 存储方式。**网络资源方面，**大模型一体机需要灵活的网络资源配置。针对训练场景，通常采用 200-400G InfiniBand/RoCE 的高带宽配置，以实现参数同步效率最大化；针对推理场景，通常采用 10-100G InfiniBand/RoCE 低时延、高可靠的网络配置。

软件层是连接硬件与应用的关键桥梁，提供资源管理和 AI 开发的能力。资源管理方面，资源池化能力通过资源虚拟化技术，如容器化和 GPU 虚拟化，将物理资源抽象化并进行池化管理，再结合智能的资源调度策略，实现计算、存储、网络资源的按需动态分配和弹性伸缩，优化资源利用率。**AI 开发能力方面，**软件层提供了覆盖模型全生命周期的工具和服务，涵盖数据处理、开发训练和部署推理的整个流程。其中数据处理功能支持数据的采集、清洗、标注和存储，确保高质量的数据输入。模型训练环境集成了主流的深度学习框架和高效通信库，支持分布式训练和模型评估。模型部署与推理环节则集成了模型压缩技术和加速库，以提升模型的运行效率和

稳定性。软件层不仅将离散的硬件资源转化为可编程、可编排的计算服务，还为大模型的开发提供灵活的技术支撑。

模型层聚焦于核心算法能力的封装和业务价值转化，是连接软件层提供的 AI 开发能力与上层应用需求的桥梁。模型层通常预置或支持便捷加载多种主流基础大模型，这些模型覆盖了不同规模和模态，为用户提供了开箱即用的基础智能能力。关键在于提供强大的模型定制与优化能力，特别是支持全参数微调和参数高效微调技术，使用户能用自有数据快速适配模型。为了评估和提升模型性能，模型层通常集成模型评估体系，提供标准的评估指标和方法。此外，模型层还强调模型管理的重要性，提供模型版本控制和管理功能。通过对基础模型的集成、强大的定制优化工具以及完善的评估和管理体系，模型层将软件层提供的开发能力转化为可直接应用的核心智能，为上层应用层提供了坚实的算法基础。

应用层作为技术栈的顶层接口，聚焦于模型层提供的智能化能力与具体业务需求的结合，实现大模型在不同领域的广泛应用与价值落地。应用层通过标准化的 API 接口、SDK 和可视化用户界面对外提供服务，降低了模型调用的门槛，使得各类应用能够便捷地集成大模型的强大能力。许多一体机还封装了针对特定行业的场景化解决方案模板，例如智能客服、文档处理、代码生成等，加速了行业应用的落地。此外，应用层强调对知识库集成与检索增强生成（RAG, Retrieval-Augmented Generation）技术的支持，通过结合外部知识库，显著提升模型回答的准确性和领域专业性。在应用层，

通常会提供智能体构建、工作流编排等工具，满足企业多样化的 AI 应用开发需求。同时，基于预置的大模型，应用层也会提供如问答助手、办公助手、编程助手等成品 SaaS 服务，直接赋能最终用户。应用层通过构建从模型推理到业务反馈的闭环系统，使得大模型一体机能够灵活嵌入各种智能应用场景，最终完成从技术能力到实际业务价值的转化。

综上所述，大模型一体机的技术架构是一个复杂且不断演进的系统，其核心目标是通过软硬件的深度融合和协同优化，为大模型的全生命周期提供强大且易用的基础设施。这种集成化的架构显著降低了用户部署和使用大模型的门槛，加速了人工智能技术在各行业的落地应用。

（二）大模型一体机产品形态分类

为了精准响应不同应用场景、预算规模以及用户群体的需求，大模型一体机的产品形态呈现出多元化的特征。其分类通常围绕硬件资源配置、模型规格以及应用领域等多个维度展开。

从核心处理任务的角度看，大模型一体机可以分为推理一体机和训推一体机。大模型推理一体机主要聚焦于将训练完成的模型部署到生产环境，实现高效、低延迟的实时推理服务。其架构设计优先考虑推理加速能力和能效比，通常采用定制化推理芯片和模型压缩技术以降低计算资源消耗。软件层面配备动态批处理、请求调度引擎和异构计算调度，保障高并发请求下的响应速度和服务质量。推理一体机广泛应用于智能客服、实时语音识别、图像分析、推荐

系统等场景，满足对响应时延敏感且并发量大的业务需求。此外，推理一体机支持多模态融合推理和长文本处理，适应复杂的应用场景。**大模型训推一体机**集成了训练和推理两大功能，其核心能力在于支持微调训练与实时推理的统一资源管理和调度，硬件上配备高性能 AI 加速器、超大带宽内存、高速存储及低延迟互联，以保障训练与推理任务的高效切换和并行执行。软件栈不仅支持主流训练框架和算法，还集成模型开发生命周期管理工具，实现从数据预处理、模型训练、微调到推理部署的全流程协同。训推一体机适用于政企、金融、医疗等对数据安全和本地化部署有严格要求的行业，能够显著缩短模型迭代周期，提升算力利用率，降低总体拥有成本。经中国信息通信研究院调研，目前产业中仅推出推理一体机的企业约占 34.0%、仅推出训推一体机的企业约占 17.0%，同时推出推理一体机和训推一体机的企业占比 48.9%。可以看出目前产业落地大模型一体机还是以推理一体机为主导。这是由于一方面模型推理是当前 AI 落地应用的主战场，许多企业不再自己训练模型，而是直接调用或部署现有模型进行应用开发，这也催生了对高性能、低成本、易部署的推理专用设备的巨大需求。另一方面，与训练相比，推理对算力的绝对要求较低，这使得一些中小型或创业公司能够凭借其在特定硬件或垂直行业优化上的优势切入市场。

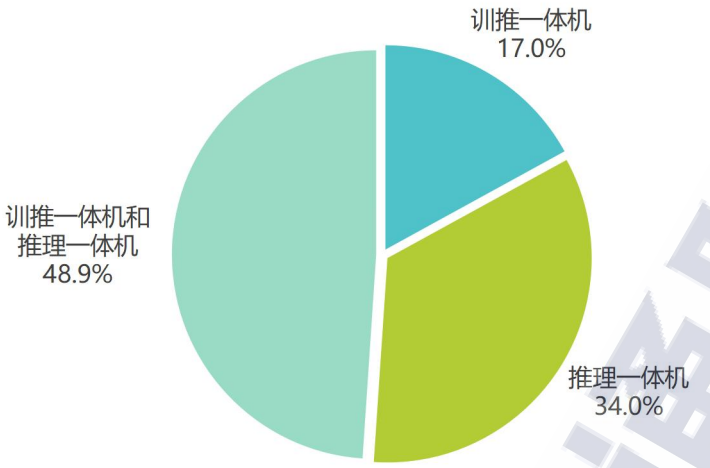


图 2 按功能划分的大模型一体机厂商分布占比

从应用场景来看，大模型一体机可以分为通用型和行业型。通用型大模型一体机适用于多行业、多场景的通用需求，侧重构建开放兼容的底层架构，具备灵活的任务扩展能力。在硬件层面支持高性能计算芯片、大容量内存、高速存储和高带宽网络；软件层面集成主流深度学习框架，提供全栈开发工具链，支持可视化管理工具；模型层面内置多种基座大模型，支持语言、视觉、多模态等基础模型，支持多种通用场景的快速部署。**行业型大模型一体机**则针对金融、医疗、制造、教育、政务等特定垂直领域进行了深度优化和定制，体现在硬件选型上强化领域计算特性适配，以及软件层面上支持行业特定的模型微调和行业知识库，提升模型的行业专业性和适应性。

表 1 大模型一体机应用场景分类

	通用型	行业型
适用场景及特性	适用于多行业、多场景的通用需求，强调模型的广泛适配性和灵活性。	针对特定行业场景进行深度优化，结合行业数据和知识，提供定制化解决方案。

硬件配置	配备高性能计算芯片，支持大模型的高效运行。支持大容量内存、高速存储和高带宽网络。	针对行业需求优化硬件配置，如专用 AI 加速器、行业定制芯片。集成高效存储和网络设备，支持实时数据处理。
软件能力	集成主流深度学习框架，提供全栈开发工具链，支持可视化管理工具，包括硬件组网、资源监控和故障定位等。	深度结合行业应用软件，支持行业特定的算法优化和模型微调，集成行业知识库和检索增强技术。
模型能力	内置多种基座大模型（如 DeepSeek 系列、Qwen 系列等），支持微调和增量训练。预置模型涵盖语言、视觉、多模态等基础模型。	基于行业数据和知识进行微调，提升模型的行业专业性和适应性。支持复杂场景的多模态模型。
应用适配	支持多种通用场景的快速部署，如知识问答、智能客服、文本生成、图像识别等。	针对行业场景进行深度优化，如医疗影像分析、金融风险管理、工业设备故障诊断等。

经中国信息通信研究院调研，目前产业中仅推出通用一体机的企业约占 21.3%、仅推出行业一体机的企业约占 31.9%，同时推出通用一体机和行业一体机的企业占比 46.8%。通过数据可以看出，**目前大模型一体机的发展趋势已趋向于行业化**，纯粹的通用一体机市场相对孤立。这类企业可能主要服务于对 AI 硬件有基础需求、但尚未确定具体应用场景，或者需要进行定制化开发的客户。目前，市场对针对特定行业、特定应用场景的 AI 一体机需求非常高。企业不再满足于一个“万金油”式的通用方案，而是需要能够深度集成行业知识、优化工作流的专业化设备。

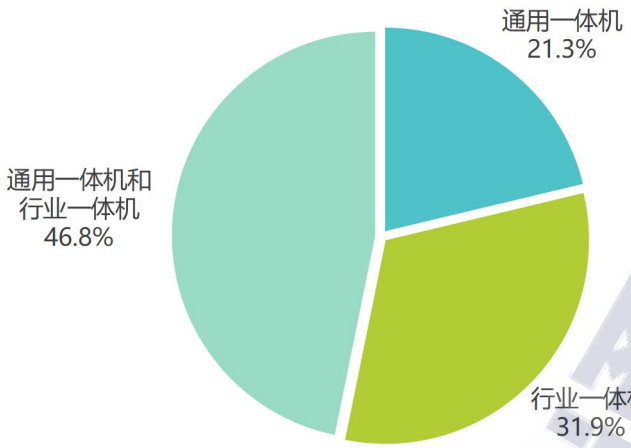


图 3 按应用类型划分的大模型一体机厂商分布占比

（三）大模型一体机选型参考

大模型一体机作为当前人工智能产业化落地的关键基础设施，正迅速渗透到政务、金融、医疗、制造等各个行业领域。然而，面对市场上琳琅满目的一体机产品和解决方案，企业如何做出科学合理的选型决策成为关键挑战。本节从应用方视角出发，系统梳理大模型一体机选型的核心方法论，深入分析不同行业的选型侧重点，提供评估方法帮助企业合理进行大模型一体机选型。

不同行业对大模型一体机的需求重点存在显著差异，深入理解这些行业特性对大模型一体机的选型决策至关重要，下面将针对政务、金融、医疗、制造等典型行业，分析其在大模型一体机选型过程中的考量因素和具体方法。

政务行业在大模型一体机的选型中会重点考量数据安全和自主可控的要求。为保证数据安全，政府部门通常要求数据本地化存储，采用物理隔离的内部部署方式以满足合规需求，避免数据外泄和跨境传输风险。因此，政务选型时格外关注国产化适配、私有化部署

以及安全防护能力。政务工作涉及大量敏感信息，只有将大模型一体机部署在本地并采取国密算法加密等措施，才能防止机密数据泄露。此外，政府应用场景偏向公文生成、政务服务问答等，往往希望定制模型以贴合政策咨询、审批流程等垂直需求，从而提高服务准确性和效率。政务用户也建议选择具备长期支持能力的厂商，以确保系统可维护性和生态适配，满足后续升级迭代需求。

金融行业在大模型一体机选型中着重关注安全性和时效性。由于金融行业设计大量敏感的财务信息和交易数据，金融机构需要采取严格的数据安全措施，倾向于选择私有化部署并进行安全加固的一体机，以满足金融监管机构的各项要求。这类一体机通常会内置数据隔离沙箱、支持国密算法模块，以保证客户敏感数据的隔离保护。此外，为支持海量交易和用户请求的实时处理，金融行业在选型一体机的过程中更加需要高可靠性和低延迟。目前，金融业应用大模型主要聚焦内部业务提效，如智能合同审核、资产对账、市场研报生成等。对于涉及高精度数值计算的场景，许多银行证券还持谨慎态度，在选型时会评估模型输出的准确性和可控性。

医疗行业在大模型一体机选型中更加注重数据隐私、专业准确性和可靠性。为保障患者的敏感医疗信息安全，医疗机构通常要求数据不出域，通常选择私有化部署方式，并采取严格的访问控制和数据加密措施。因此，医疗行业选型时格外关注一体机的数据安全性、本地部署能力以及细粒度的权限管理和审计功能。医疗场景对模型的专业性知识高，医疗机构倾向选择支持加载医疗专用模型和

知识库的大模型一体机，将医院自有数据用于模型训练，使其更贴合诊断决策支持、医疗问答、新药研发等垂直场景。这些应用场景要求模型有高准确性和解释能力。此外，医疗机构需要确保系统的稳定性和可靠性，建议选择具备高可用性、长期技术支持和与现有医疗系统良好兼容性的厂商。

制造业在大模型一体机选型中更加注重实时性能、场景适配能力。制造业应用大模型主要集中在设计研发、生产制造、运营管理和产品服务等场景，这些任务往往涉及高频实时数据处理，对模型推理速度和系统稳定性提出较高要求。因此，制造企业在选型时通常优先考虑具备低延迟推理能力、支持边缘侧部署的一体机产品。由于制造场景对精度和可靠性要求较高，选型时还需评估模型的场景适配能力和对行业专有知识的支持，企业倾向选择支持定制化微调和私有知识嵌入的模型，确保模型输出贴合自身工艺流程和质量标准。

虽然各行业在大模型一体的选型方面侧重有所不同，但用户单位在选型大模型一体机大体遵循以下流程和决策逻辑：

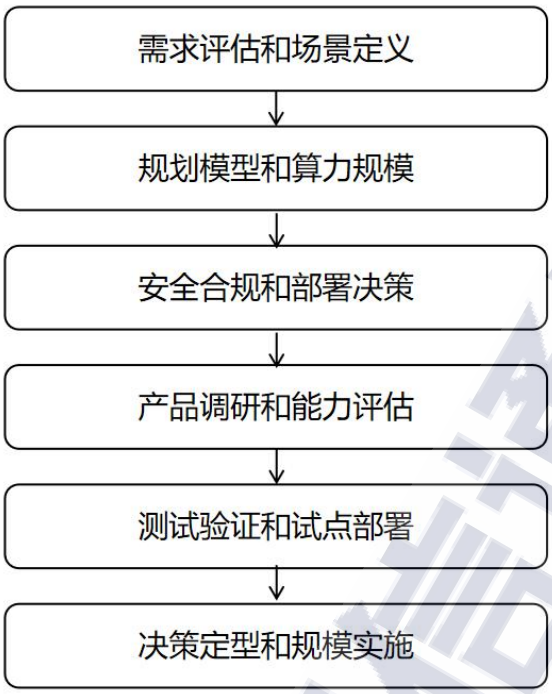


图 4 大模型一体机选型流程

首先是需求评估和场景定义。明确业务需求，评估企业引入大模型应用的目标场景和性能要求。明确任务聚焦训练还是推理、应用的领域是智能客服、公文写作还是图像识别，评估不同场景的并发量、实时性和吞吐量需求，以确定合适的模型类型和规模，并选择相应的数据资源。

其次是规划模型和算力规模。根据业务场景选择合适的大模型类型并确定参数规模。进一步依据大模型的类型和参数确定所需算力硬件配置，包括 GPU 型号与数量、CPU 型号、内存容量、存储型号、机柜规格和网络带宽等。政务、金融、能源等行业客户尤其要关注国产化兼容性。此外，在算力规划时还需考虑峰值并发和扩展性，选定一体机配置应能支撑高峰期负载，并预留接口方便后续增加节点或升级 GPU，以免最初选型过小后期无法平滑扩容。

然后是安全合规与部署模式决策。明确企业在数据安全、合规

性方面的需求，决策部署方案。如果企业涉及敏感数据或有监管要求，通常会选择完全本地化部署的一体机，以确保物理隔离和数据“不出域”。需核对一体机提供的安全特性，如访问控制、加密存储、审计日志、防火墙策略等。同时确认方案对国产软硬件的支持程度，包括操作系统、中间件、GPU 加速卡等兼容性，以避免合规隐患或供应链风险。

之后是产品调研和能力评估。收集市场上不同厂商的大模型一体机产品，重点考察厂商的技术实力和方案成熟度。在性能方面，重点考察吞吐量、并发能力、延迟时间等指标，来保证大模型一体机满足业务场景的任务量级和处理效率需求。在软件生态方面，重点关注软件工具链能力、预置应用能力以及技术支持能力。同时，也需要综合考量成本预算、售后能力等因素，结合自身预算，寻求技术与成本的最佳平衡。

此外是测试验证与试点部署。在比选不同一体机方案后，通常会对候选方案进行概念验证（PoC, Proof of Concept）或小规模试点部署，以实际检验其功能和性能。这一阶段由厂商提供测试机或在用户机房部署一套设备，与用户的业务数据进行对接测试。典型的验证内容包括：功能测试、性能测试和兼容性测试等。通过一段时间的业务模拟运行，应用方可以发现并反馈问题，如模型输出错误案例、系统稳定性不足之处，然后与厂商一起优化调整。

最后是决策定型与规模化实施。综合试点结果和各项指标，最终选定最契合需求的一体机方案，签订采购部署合同。此后进入正

式上线阶段，在部署过程中保持与厂商技术团队紧密合作，做好人员培训和运维交接。用户方企业还需制定评估机制，持续监测一体机在实际业务中的效果和收益，定期审视是否需要扩容硬件或引入新的模型版本，以保障系统始终满足业务发展需求。

通过以上流程，应用单位可以从自身业务痛点和目标出发，有条理地筛选出最合适的大模型一体机方案，实现技术落地与价值转化。

三、大模型一体机产业发展情况

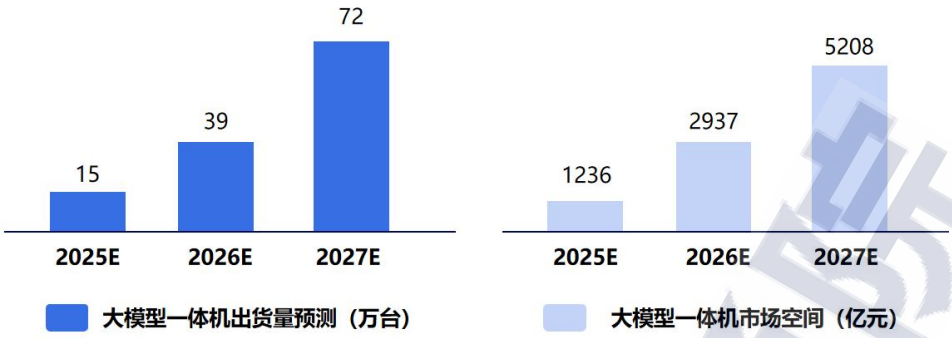
随着人工智能技术的不断突破与应用需求的快速增长，大模型一体机产业迎来了蓬勃发展期，并逐步形成多元化的产业生态格局。当前，大模型一体机正从单一硬件加速向“算力基建+算法优化+场景赋能”的全链条协同方向演进，通过整合芯片厂商、云计算服务商、算法开发商及行业解决方案提供商的核心能力，构建起覆盖技术研发、产品交付与商业落地的完整价值链。

（一）大模型一体机市场定位及规模

大模型一体机的目标市场主要面向将数据主权、安全可控、模型私有化作为核心竞争力的行业与企业。由于大模型一体机具有软硬件协同、本地化部署为主以及多层安全体系的特性，使其更加适用于注重数据安全与隐私保护、对实时性要求高以及对应用易用性需求较高的场景。从开发流程上，主要面向推理或轻量训练为主的场景；从行业应用上，主要面向政务、金融、医疗、能源等数据安全与合规要求高的行业。同时，**大模型一体机也存在其使用场景的**

局限性。由于其软硬一体化、本地部署的特性，大模型一体机在处理超大规模模型训练和高并发推理任务时，容易遇到性能瓶颈，难以支撑大规模参数模型的持续高效训练。此外，本地化部署的方式提高了企业的初始采购成本和后期运维成本，包括硬件购置、机房改造、电力供应及专业运维团队建设等。因此，大模型一体机在超大规模训练场景和海量高并发服务请求中存在明显局限，更适用于对数据隐私要求极高、业务规模相对稳定、且具备一定资金与技术支持能力的行业用户。

市场需求持续增加，大模型一体机步入高速增长通道。在生成式 AI 技术爆发与企业智能化转型需求的双重驱动下，大模型一体机市场渗透率持续提升。浙商证券预测，2025 至 2027 年大模型一体机需求量将分别达到 15 万台、39 万台和 72 万台，对应的市场空间将从千亿级别迅速扩张，至 2027 年有望突破五千亿元人民币。据中国信息通信研究院统计，市场上已迅速有接近百家厂商推出 AI 一体机产品，这些参与者涵盖硬件与服务器企业、云计算厂商、大模型创业企业及行业应用开发商。在基础设施层面，硬件厂商通过整合高性能算力与行业方案积极抢占市场，云厂商则依托弹性算力与生态加速布局私有化部署，在应用市场层面，头部企业凭借技术或生态优势占据主要份额，形成了多元竞争的格局。



来源：浙商证券

图 5 大模型一体机出货量及市场空间

（二）大模型一体机产业链分析

大模型一体机作为集成高性能硬件、先进模型和行业应用的复杂系统，其产业链构成涵盖了从硬件供应商、软件供应商、模型供应商、应用供应商到整机供应商等多个关键环节。各环节不仅技术门槛高，且相互依赖、协同紧密，形成了一个多层次、多维度的生态体系。

来源：中国信息通信研究院

图 6 大模型一体机产业图谱

硬件供应商为大模型一体机提供算力底座，是性能与可靠性的基础保障。硬件供应商处于产业链上游，为大模型一体机提供 GPU、

CPU、存储设备、网络设备和电源等配套部件，其提供的算力、网络和存储能力决定了一体机运行大模型的效率。目前国内已有超过 20 家人工智能芯片厂商成为大模型一体机的供应商，包括华为、新华三、问道以芯等，国产硬件生态呈百花齐放之势。庞大的市场需求和政策扶持，为硬件供应商提供了前所未有的发展机遇。

软件供应商是大模型一体机产业链中的关键环节，为大模型一体机的高效运行和应用落地提供支撑。软件供应商主要负责提供操作系统、AI 平台软件、模型管理工具、推理引擎及开发工具链等软件能力。通过提供操作系统保证整机稳定运行，提供 AI 框架和算法库，为模型训练和推理提供基础算法支持，提供模型推理引擎、中间件和驱动适配等，使硬件算力与大模型算法能够高效协同，通过集成模型仓库、开发工具链、容器化部署等技术，极大降低了用户的使用门槛和开发成本。目前，国内已有超过 30 家大模型平台或工具链厂商成为一体机软件供应商，包括百度、中国移动、新华三等，软件供应商面临着多样化的行业需求和复杂的应用场景，需提供高度定制化和场景适配能力。随着大模型应用从通用向垂直行业深入，软件层面的定制化、优化和服务能力成为竞争焦点，**目前主流的产品形态会在大模型一体机中预置检索增强生成**（RAG, Retrieval-Augmented Generation）**技术和智能体**，来使大模型一体机提供更加符合特定场景或业务需求的能力。经中国信息通信研究院调研，目前 34.0%的企业会在推出的部分大模型一体机中配置检索增强生成能力，61.7%的企业会将所有一体机都配置检索增强生成能力，

44.7%的企业会在推出的部分大模型一体机中配置智能体，51.1%的企业会在推出的所有大模型一体机中配置智能体。

模型供应商是大模型一体机产业链中链接基础算力与应用场景的关键环节，是大模型智能化能力的核心来源。模型供应商通过构建高质量的基座模型和针对特定行业的专用模型，赋能大模型一体机实现高效的训练、推理和管理功能，极大提升一体机的应用价值和行业适配性。当前国内市场涌现出超过 30 家模型供应商，包括百度、华为、浪潮云等，他们通过持续优化模型性能和增强模型的行业适应能力，推动大模型一体机在政务、公安、医疗、教育等多个领域的深度应用。经中国信息通信研究院调研，89.4%的企业会在大模型一体机中预置开源模型，仅 10.6%的企业仅使用闭源模型，开源模型在大模型一体机中的应用占比极高。

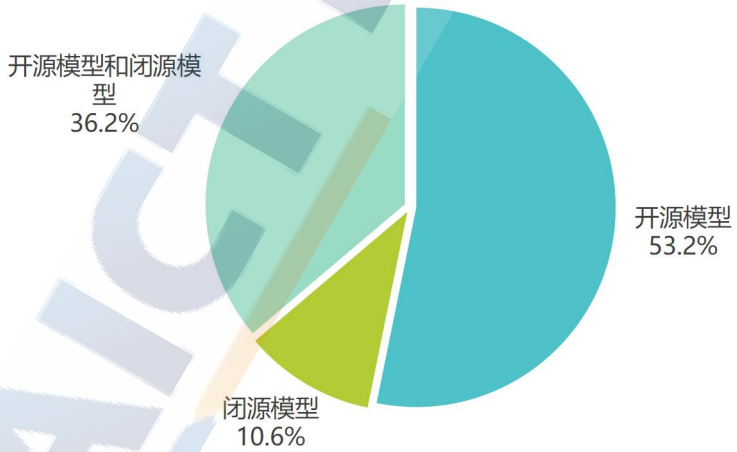


图 7 大模型一体机应用开源、闭源模型占比

应用供应商将大模型一体机的算力、软件和模型能力转化为具体行业解决方案，推动一体机在各行业的深度应用。应用供应商位于产业链下游，通过结合行业业务流程和需求，开发定制化的智能

化应用，提升行业效率和服务质量，成为大模型一体机产业链中实现价值落地的关键力量。目前，国内已有华为、新华三、中国移动等多家领先企业活跃于该领域，形成了丰富的应用生态。应用供应商的核心价值在于将复杂的人工智能技术转化为易用、高效的行业解决方案，降低用户技术门槛，提升业务效率和智能化水平。

整机供应商承担着将硬件、软件、模型等多方面资源进行深度集成和系统设计的关键职责，是一体机性能、稳定性和用户体验的核心保障。整机供应商负责一体机的架构设计、硬件选型与集成、软件系统集成及 AI 算法优化，确保各组件协同高效运行，满足大模型训练与推理的复杂需求。目前国内整机供应商包括华为、百度、新华三等头部企业，这些厂商具备强大的研发能力和系统集成经验，通过软硬件协同优化，实现“开箱即用”、高性能、低功耗和安全合规的一体化解决方案，极大降低了用户的部署和运维难度。整机供应商的核心价值在于整合产业链上下游资源，打造稳定可靠、性能卓越且易用的产品，推动大模型技术的规模化应用。

大模型一体机产业链的健康发展，需全链条协同应对来自技术、供应链与合规性等多维度的挑战。硬件供应商作为基础算力支撑方，面临高端芯片研发制造难度大、短期国产芯片性能存在差距等核心技术壁垒，同时高端制造环节依赖先进代工和特定原材料，在地缘政治影响下供应链稳定性存在不确定性，亟需突破技术封锁和外部限制以支撑产业持续发展。软件供应商面临基础框架迭代快、软硬件协同复杂及开源合规等挑战，需持续优化异构算力支持与分布式

加速，并保障不同部署场景下的兼容性与稳定性。同时，还需满足企业客户对安全性、可解释性和私有化部署的严格要求，构建符合监管要求的全栈软件能力。**模型供应商**则需应对算法迭代迅速、算力成本高企及数据合规监管趋严的挑战，千亿参数模型的训练与推理资源消耗巨大，需持续优化混合精度、MoE 架构和模型蒸馏等技术，同时数据获取与使用也须严格遵循隐私保护和内容安全要求。应用供应商身处落地前沿，需直面行业需求复杂多变、场景适配难度高、技术与业务融合深度不足等挑战，必须在快速响应市场变化、持续优化产品体验和保障数据安全合规等方面实现系统性突破。整机供应商在系统集成方面面临产品迭代快、兼容性复杂和稳定性保障难度大等问题，必须统筹性能、成本与能效的平衡，并灵活响应不同客户的定制化需求。

四、大模型一体机应用实践

近年来，随着大模型技术在各行业的深入应用，大模型一体机凭借其软硬件一体化优势，逐步成为推动智能化转型的重要支撑。大模型一体机显著提升了智能客服、智能编程、智能写作等通用场景的应用效率与体验，降低了企业和机构的技术门槛和运营成本。作为连接前沿技术与实际业务的关键桥梁，大模型一体机正在政务、金融、医疗、制造等多个领域实现了规模化落地，助力构建安全、可靠且高效的智能服务体系。

（一）场景应用实践

目前，大模型一体机凭借其集成化、高性能和易用性的特点，

在智能客服、智能编程、智能写作等各类应用场景中展现出日益广泛的应用潜力，显著提升了工作的效率和效果。经中国信息通信研究院统计，74.5%的企业推出了应用于智能客服场景的一体机，63.8%的企业推出了应用于智能写作领域的一体机，74.5%的企业推出了应用于智能检索领域的一体机，80.9%的企业推出了应用于智能数据分析的一体机。

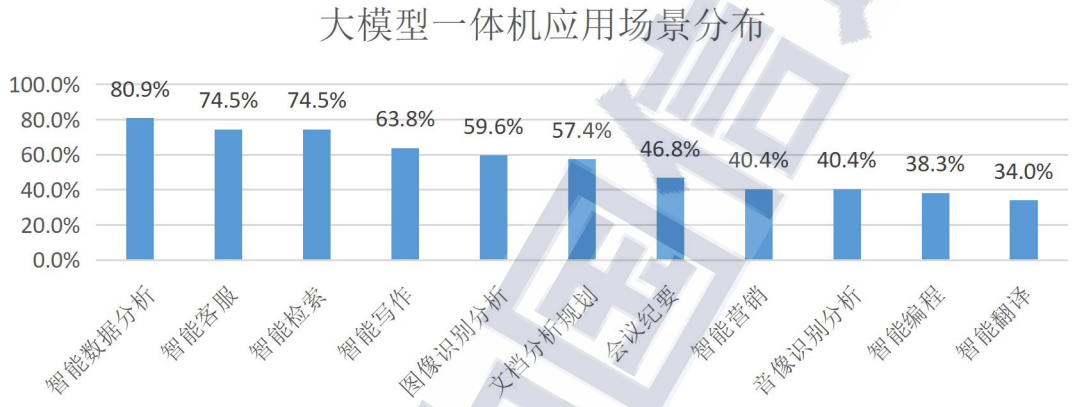


图 8 大模型一体机在各应用场景中的分布占比

大模型一体机赋能智能客服，提升服务应答能力。传统智能客服面临着专业性不足、响应速度慢和知识更新滞后等核心挑战，无法深入理解行业术语和复杂问题逻辑。为解决此类问题，大模型一体机在智能客服领域提供了私有化的大语言模型服务，结合企业数据实现智能交互。通过部署在本地的一体机，能够实现在企业内网快速上线智能客服功能，保障数据不出域、模型可控，符合金融、政务、医疗等行业对数据隐私的严格要求。在技术路径上，一体机依托于大模型自然语言理解与生成能力，使其能够理解口语化提问、支持多轮对话并保留上下文，从而弥补传统系统缺乏交互记忆的缺陷。在此基础上，一体机通常采用检索增强生成（RAG, Retrieval-

Augmented Generation）技术，将企业内部知识库与通用大模型相融合：在生成回答前先检索相关内部文档作为“外挂知识”，极大提升了回复的专业性并减少了“幻觉”错误。此外，由于此类知识检索和客服问答场景通常都在本地进行，也进一步避免了数据泄露风险，确保了业务的安全性。部署形态上，大模型一体机通常软硬件集成，开箱即用融入现有 IT 架构，可连接企业知识库、业务数据库以及客服对话接口，实现与原有系统的平滑集成。**深圳某半导体企业使用华为大模型一体机搭建企业智能 IT 热线极大提升企业运营效率。**通过便捷的硬件安装和模型部署流程，其交付效率提升 200%。在企业 IT 服务场景中，通过智能文本问答、坐席实时辅助、处置建议推荐等功能，最终达成 80%工单自助闭环，显著提高工作效率、降低运维压力。

大模型一体机驱动代码生成，降低开发门槛提升工程效率。大模型出现之前的开发流程存在诸多低效环节，开发人员需要花费大量时间写样板代码、重复性的逻辑，同时，代码质量保障也给研发带来额外负担。这些问题会导致项目周期拉长、人员投入高、错误率高，使企业研发效率受限。为解决此类问题，大模型一体机提供了智能编码助手的解决方案，其部署形态通常是在企业内部以一体机或云旁部署的形式，引入训练有素的代码大模型，并通过插件或 API 集成到开发者常用的 IDE、代码托管平台中。技术路径上，通过自然语言生成代码、智能补全与重构帮助开发者快速编写和重构代码。同时，根据函数逻辑自动生成单元测试用例，并输出代码调

试建议，降低代码测试和排错的成本。最后，根据代码内容，AI 自动撰写注释、接口文档，便于团队理解和维护。**浪潮云睿大模型一体机应用于智能编码场景后**，使开发时间成本显著降低，平均每个项目的开发时间缩短了 55%，原本开发一个中等规模项目需 45 天，应用一体机进行代码生成后仅需 20 天，大幅提升了开发效率。

大模型一体机重构公文写作，大幅提升创作效率。公文写作作为党政机关和大型企业的基础工作，是一项耗时费力且要求严谨的工作，过程周期长、人工投入大，常常拖慢事务办理效率。以往拟写一份公文可能需要数小时甚至数天的反复打磨和审核，增加了行政成本。为解决此类问题，大模型一体机在公文写作场景提供了自动拟稿和辅助校对的整体方案。部署在政企内网环境的公文生成一体机往往内置经过公务文书语料专门训练的大语言模型，能够根据用户的简要指令快速生成符合公文格式要求的初稿，并自动完成格式和措辞的规范化处理。模型结合了机关已有的公文模板和专业知识库，确保生成内容符合公文和政策要求。

（二）行业应用实践

目前，大模型一体机正加速渗透政务、金融、医疗、公共安全、制造、零售等关键行业。一方面，通过安全可控的本地算力与行业知识库深度融合，驱动政务服务提质增效、金融风控精细化、医疗诊疗智能化；另一方面，以“快速交付+智能体”模式缩短了从业务需求识别到价值兑现的周期，显著降低了组织使用大模型的技术门槛与运维成本，为产业智能化升级提供坚实底座与持续动能。经中

国信息通信研究院调研，61.7%的企业推出了应用于金融行业的一体机，74.5%的企业推出了应用于政务行业的一体机，61.7%的企业推出了应用于教育行业的一体机，57.4%的企业推出了应用于医疗行业的一体机。

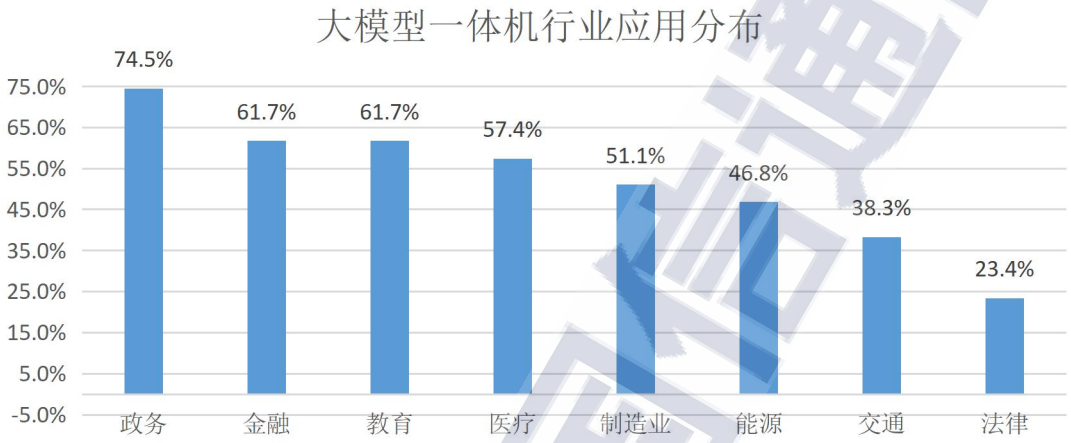


图 9 大模型一体机在各行业中的分布占比

大模型一体机赋能金融，提升风控与合规效率。在数字化和智能化的浪潮下，金融行业面临风险管理压力与日俱增、合规审查成本居高不下、客户服务体验亟待提升等多重挑战。面向上述挑战，当前越来越多的金融机构应用大模型一体机实现模型的快速启动及微调优化，保障高效可靠的业务运营管理。**在投研与风控方面，**券商利用大模型一体机实时抓取财经资讯、公告等数据，自动生成高质量研报，并对投行业务底稿和招股书进行智能校验，提高投研效率和准确性。保险机构借助大模型简化核保流程，自动分析投保人资料评估风险，加速理赔服务。**在信贷审批与合规审查方面，**银行部署大模型一体机作为智能信贷助手，可以自动提取申请材料关键信息，辅助进行贷前信用评估与贷后风险预警。合规部门借助模型

对海量交易和报告进行审查，自动识别可疑行为和异常模式。通过本地大模型对接内部知识库，审批流程从人工经验判断提升为智能审核，既提高审批效率又保证合规要求。**在智能客服与财富管理方面，**大模型一体机内置的智能知识问答助手可以让员工快速检索处理产品和业务知识，并提供多格式文档的对话交互查询，支持内容溯源，显著提升客服响应专业度。智能坐席辅助系统深度理解客户意图，提供流程导航和知识提醒，减轻坐席人员负担。此外，在财富管理领域，大模型助手可辅助理财顾问进行客户需求分析、产品推荐和资产配置方案，提供个性化的投资建议，提升服务品质和精准度。

大模型一体机赋能政务行业，提升公共服务效率和质量。政务领域长期以来存在业务办理效率不高、知识利用不足等问题。一方面，各级政府日常要处理大量公文材料和审批事务，传统模式下公文起草、审改流程繁琐冗长。另一方面，不同部门之间信息孤岛现象突出，“知识孤岛、数据碎片”导致决策支持不充分、跨部门协同低效。同时，政府数据高度敏感，要求人工智能应用必须安全可控、本地部署。通过引入大模型一体机，能够提升行政办公效率和公共服务质量。**在公文起草与审校方面，**大模型一体机正在辅助撰写公文、报告和会议纪要等文稿，根据输入要点快速生成初稿，人工再行修改定稿，大幅缩减写作时间。**在政务服务与咨询方面，**大模型一体机可以针对政务大厅的智能导办和政策咨询场景，提供智能客服功能。群众通过网站或政务 App 提出问题，即时从庞大的政

策法规库中检索并生成回答。**在行政审批与治理决策方面，**大模型一体机可以担任智能辅助审批员的角色，对提交的申请材料自动分类归纳要点，辅助审核人员发现遗漏或风险点，加速审批流转。在政府治理中，能够整合应急管理、城市运行等多源数据，实现智能决策支持。**黑龙江省大数据中心应用浪潮海若大模型一体机提升运营指挥大厅政务数据分析效率。**通过对话交互实现分析决策数据的灵活实时获取，通过构建知识库和问答模型，面向政府数据指标分析展示场景，提供数据分析研判和辅助决策能力，大幅提升运营决策效率。例如通过一体机支持比数专家能力，数据比对耗时从原来的 5 小时，降低到了 30 秒以内；一体机提供数据快速获取能力，帮助业务人员大幅度提升数据获取效率，获取数据的周期从原来仅工作日找数据，需 1-3 天获取所需数据，提升到 7*24 小时随时获取数据，耗时仅需 2-3 分钟，助力高效的运营数据决策流程。

图 10 大模型一体机应用于政务场景的数据分析

大模型一体机赋能制造业，加速生产与运营模式革新。在全球化竞争日益激烈、个性化需求不断增长的背景下，制造业面临着生产效率提升、产品研发周期缩短、质量控制精细化以及供应链协同

优化等多重挑战。为了应对这些挑战，越来越多的制造企业引入大模型一体机，以实现设计、生产、管理全流程的智能化升级。**在智能研发与设计方面，**制造业企业利用大模型一体机对海量研发数据和设计图纸进行深度学习，辅助工程师快速生成产品概念、优化设计方案，并进行虚拟仿真测试，大幅缩短产品上市周期。**在智能生产与质量控制方面，**大模型一体机能够实时分析产线数据，识别设备异常、预测维护需求，并优化生产排程。同时，通过对产品图像、传感器数据进行智能识别与分析，实现缺陷的精准检测和质量的在线控制，提高良品率。**在供应链优化与管理方面，**大模型一体机能够整合供应商、生产、物流等多方数据，进行需求预测、库存优化和风险预警，提升供应链的韧性和响应速度。**在客户服务与售后方面，**大模型一体机内置的智能问答系统可以快速响应客户咨询，提供故障诊断和维修指导，提升客户满意度。通过本地化部署大模型，制造业能够确保核心技术和生产数据的安全，加速向智能制造转型。**日照钢铁集团应用新华三大模型一体机优化炼钢过程。**在生产制造方面，监测并分析炼钢过程中的关键变量，如温度、化学成分和炉料配比等，预测并调整生产参数，使炼钢成本降低了约 15%，成品率提升 10%。同时，应用大模型一体机进行智能调度，综合生产线的生产能力、设备状态和市场订单等信息，实现了生产调度优化，使生产效率提升 10%，设备利用率提升 15%，改善了生产线的综合效益。在供应链管理方面，应用大模型一体机分析市场需求、原材料价格及供应商表现等多维度数据，优化了原材料采购和库存管理。

在原材料采购上实现了 10%的成本节约，并在满足市场需求的同时，降低了库存周转时间。

大模型一体机赋能教育行业，提升教学质量与管理效率。教育行业长期存在优质资源分布不均、个性化教学不足和教师负担过重等挑战。为解决这些挑战，教育机构正积极探索大模型一体机的应用，以构建更智能、高效的教育生态系统。**在个性化教学与学习方面，**大模型一体机能够分析学生的学习数据、知识掌握情况，智能推荐定制化的学习路径和教学资源，实现“因材施教”。它能够根据学生的提问，提供多维度、深层次的知识解答，并进行智能批改作业和提供即时反馈。**在智能教学辅助方面，**大模型一体机可以辅助教师快速备课，自动生成教案、课件和试题，减轻教师的日常工作负担。它还能通过对课堂互动和学生表现的分析，为教师提供教学改进建议，提升课堂效率和趣味性。**在教育管理与决策方面，**大模型一体机能够整合学校的各类管理数据，如学生学籍、成绩、考勤等，进行智能分析和预测，为学校管理者提供数据驱动的决策支持，优化资源配置。

大模型一体机赋能能源行业，提升能源管理和运营效率。在全球能源转型与“双碳”目标的背景下，能源行业面临着数据量爆炸式增长、运维管理复杂、安全风险突出以及能源利用效率亟待提升等一系列挑战。为应对这些挑战，越来越多的能源企业正引入大模型一体机，以实现模型的快速部署、高效训练与优化，从而保障其业务的高效可靠运行。**在电力调度与智能电网方面，**大模型一体机

可以对海量实时数据进行分析，精准预测电力负荷，并优化调度策略，确保电网稳定运行，减少高峰时段的供电压力。**在设备运维与故障预测方面，**通过模型对发电机组、输电线路、变压器等设备的运行数据进行深度学习，可以自动识别异常模式，提前预警潜在故障，将被动维修转变为主动预防，大幅降低停机时间和运维成本。**百度大模型一体机应用于福建电力行业，**通过电力知识记忆理解、多模态融合分析，给电网“问诊把脉”，辅助业务人员实现重过载问题的精准诊断并及时“对症下药”。通过学习 2000 多项故障案例，单次诊断工作可由 20 个工作日缩短至 5 个工作日，全省每年预估可节约诊断工作 1500 人天，大型保供电方案编制时间由原来 10 小时缩短至 10 分钟。

五、大模型一体机发展趋势

近年来，随着大模型技术的不断突破与成熟，大模型一体机作为集成算力、模型与应用的综合解决方案，正加速推动智能计算向更高效、更安全、更便捷的方向发展。凭借硬件深度优化、软件工具链完善及多元化模型生态的协同进步，大模型一体机不断提升算力效能与系统智能化水平。同时，其在行业场景的垂直深化和广泛渗透，促进了政务、金融、医疗、制造等领域的深度融合。伴随着安全防护机制的内生化与便捷化部署能力的提升，大模型一体机正成为构建安全、可靠且高效智能服务体系的核心支撑。

（一）大模型一体机的全栈技术能力持续深化

随着大模型参数的不断扩大，应用场景的逐渐丰富，大模型算

力性能需求不断提高，部署和优化的软件栈变得更加复杂，模型逐渐向多模态、多行业持续扩展。为应对上述挑战，大模型一体机正朝全栈协同优化持续演进。

在硬件层面，大模型一体机的算力、网络、存储及能耗管理能力都将不断提升。传统的硬件架构面临算力瓶颈、延迟长等挑战，大模型一体机将通过从“单机箱”到“机柜级一体化”演进，提升模型的实时推理与多模态推理吞吐能力。为了提升模型参数加载和梯度同步效率，大模型一体机将采用**高效存储与内存体系**，优化存储层级和带宽，减少数据传输瓶颈。同时，为了提升大模型训练的扩展性和效率，大模型一体机将采用**网络互联与分布式协同**的方式，支撑分布式训练和推理的高速低延迟网络架构，保障多节点间高效通信和同步。此外，为了提高系统稳定性和持续运行能力，大模型一体机将采用**能效管理与散热技术**，通过动态功耗管理、智能调度和先进散热技术，降低整体能耗，在降低运营成本的同时，符合绿色计算的发展趋势。硬件深度优化为大模型一体机提供了强大的算力基础和高效的资源利用，是实现大规模模型训练和部署的前提。

在软件层面，大模型一体机的软件工具链正在向自动化、可视化、易用化的方向发展。完善的软件工具链是释放大模型一体机硬件能力的关键。为了提高开发效率和模型质量，大模型一体机将支持**全流程自动化的软件工具**，覆盖数据预处理、模型训练、调优、部署和监控等环节，极大降低人工干预，加速从算法创新到产品落地的周期。为了保障系统稳定运行和持续优化，大模型一体机将提

供**统一的开发和运维平台**，统一接口和管理视图，支持跨硬件、跨模型的统一调度和资源管理，实现模型生命周期管理、版本控制和性能监控。为了降低运维成本，大模型一体机将提供**智能化运维和故障预测**能力，通过引入 AI 驱动的运维系统，实现故障预警、资源调优和性能自适应，提升系统可靠性和用户体验。此外，大模型一体机软件正在向**低代码、无代码**方向发展，通过降低技术门槛，使业务人员也能参与模型构建和应用开发，促进 AI 技术的普及和应用创新。软件工具链的完善不仅提升了开发者的生产力，也为大模型一体机的广泛应用提供了坚实的软件支撑。

在模型层面，大模型一体机的模型生态正在不断扩展，构建多元化的应用场景。为了适配更加丰富的业务场景，大模型一体机将推进**多样化模型库建设**，包括通用大模型和行业专用模型和多模态模型等。模型库的丰富性和开放性促进了模型复用和快速定制，缩短了应用开发周期。为了实现大模型在一体机上的高效部署，大模型一体机将推进**模型适配与优化技术**，通过模型压缩、蒸馏、剪枝等技术，兼顾性能和资源消耗，保证模型在不同硬件平台和业务场景中的表现。为了满足复杂场景下的综合智能需求，大模型一体机正在推进**跨领域的模型融合**，推动语言、视觉、语音等多模态模型的融合，提升模型的理解和生成能力，拓展智能应用的边界。同时，大模型一体机还在推进**生态协同与开放合作**，鼓励产业链上下游协同创新，推动开源社区、企业和科研机构共同构建开放、共享的模型生态，促进技术创新和商业模式创新。模型生态的扩展不仅丰富

了大模型一体机的应用价值，也推动了人工智能技术向更广泛、更深入的行业渗透。

（二）大模型一体机将持续深化行业化场景化能力

随着大模型技术在各垂直领域的快速渗透，传统的大模型往往缺乏行业特定知识，直接应用于专业领域效果有限，而各行业又存在数据隐私、合规要求和专业门槛，需要定制化的解决方案。在此背景之下，大模型一体机正从通用能力平台向深度行业化、场景化解决方案演进，将通过领域知识增强、业务场景专属优化等方式，构建更精准、更高效的行业智能体系。

为了更好地适配垂直场景，各行业将涌现专用的大模型及知识库集成方案。大模型一体机将预置或接入各行业权威的数据和知识图谱。政务领域将推进融合政策法规和政务数据构建政府专用大模型，实现智能政策解读、政务问答等能力；金融领域将集成金融知识库和风控模型，为投研、风控提供专业分析支持；医疗领域将结合医学文献和临床数据库，打造医疗问答和辅助诊断模型等。通过将行业知识与大模型结合，一体机能够提供场景化、专业化的人工智能服务，弥合通用大模型与具体业务需求之间的差距。未来，不同行业的大模型细分化趋势将更明显，每个行业可能拥有自己优化的模型版本，既精通该领域专业知识又具备通用对话能力。这将推动“千模型千面”的行业人工智能生态形成。

在各重点行业中，大模型一体机将催生出一批行业专属型智能体应用，并逐步标准化、规模化部署。政务领域会推广面向公众的

政策咨询智能问答体，提供法规政策解读和政务信息服务；金融领域广泛应用智能客服和投顾分析助手，用于客户服务、投资咨询和反欺诈风控；教育领域发展 AI 教学助手，可根据学生情况提供个性化答疑和学习指导；医疗领域出现临床决策支持 Agent，协助医生检索医学知识、进行诊断参考；工业制造领域部署智能质检与运维助手，通过数据分析实现设备故障预测和生产流程优化。这些行业智能代理的落地将极大提升行业效率。可以预见，行业通用 AI 应用模块化将成为趋势，各行业会沉淀出一批可复用的智能体应用模板，通过一体机快速部署到实际业务中，推动行业智能化升级。

（三）大模型一体机将兼顾安全性与便捷化部署

对于金融、政务、医疗等行业而言，数据安全与治理合规是部署大模型的首要考虑。大模型处理的数据往往涉及个人隐私、金融交易、政务机密等敏感信息，需要实施内容审核和权限管理等模型治理措施。未来，大模型一体机将在提升安全可信的同时实现部署形态多样化和服务模式创新。

大模型一体机将内置全方位的安全防护能力。一方面，在硬件层集成加密芯片、可信执行环境等安全模块，对数据存储和传输进行高强度加密，确保数据即使在本地也受到银行级保护。通过硬件隔离和沙箱技术，将不同用户、不同应用的运行环境彻底隔离，防止数据串扰和未授权访问。另一方面，在软件层面引入内容安全“围栏”机制：实时监控模型的输入输出，自动检测并过滤敏感或违规内容，防止大模型产生有害信息。同时，一体机管理平台将提

供安全审计和访问控制功能：记录所有用户对模型的访问和调用日志，支持分级权限管理和实时监控。这样企业可以对模型的使用实现可追溯、可控管，符合内部风控和外部监管要求。未来随着国内《生成式 AI 服务管理办法》等法规落地，模型的合规性、可解释性要求提高，预计一体机厂商会在模型治理上加强投入，确保其产品满足不断演进的法律和伦理规范。

在保证安全性的同时，大模型一体机的部署形态将更加灵活多样，以满足不同场景需求。一方面，私有化本地部署仍是主流形态，企业可以将一体机部署在自有机房或本地数据中心，实现与现有 IT 环境的无缝集成。这些一体机往往采用标准机架式设计，支持模块化扩展，当业务规模增长时可以通过叠加更多计算模块来线性扩容。另一方面，一体机的边缘化部署将兴起。针对工厂车间、零售门店等对时延敏感或网络受限的场景，小型大模型一体机可部署在边缘侧，就近提供 AI 推理服务。这些边缘一体机具备体积小、功耗低的特点，但仍内置完整的大模型能力，适合分布式部署在各业务现场，实现云边协同。同时，云边协同管理平台将成为一体机的重要组成部分。厂商可能提供统一的管理软件，使企业能够同时管控云端大模型服务与本地一体机设备，统一调度算力和模型资源。在需要更大算力时，一体机可无缝对接公共云或私有云资源，实现混合部署。在数据敏感或网络中断时，则可切换为本地独立运行，保证业务连续性。最后，部署的简易与自动化依旧是一体机演进重点，未来一体机将在即插即用基础上进一步做到自适应配置与智能运维。如：

自动根据负载调整算力分配、自动安装安全更新和补丁、通过 AI 辅助进行故障预测和运维决策等，极大降低企业使用大模型技术的门槛。

（四）大模型一体机产业生态持续协同深化

随着大模型一体机技术的不断成熟与广泛应用，产业生态正进入持续协同、深化发展的阶段。随着大模型一体机在企业端、行业端的快速渗透，其产业生态正在从单一厂商的“产品提供”模式，加速走向多方参与、深度合作的“价值共创”模式。这种协同深化不仅体现在技术层面，更体现在产业链上下游资源的整合和商业模式的创新上。

大模型一体机产业生态格局向多元化和协同化方向发展。大模型一体机的产业生态正在由以整机厂商为核心的单中心结构，演进为涵盖硬件供应商、软件供应商、模型供应商以及应用供应商的多元化、开放式生态。一方面，生态参与者间的合作模式将不断深化，从“技术集成”走向“生态共创”。整机厂商不再仅提供封装式产品，而是联合上下游伙伴共建软硬一体化生态，推动标准接口、资源池化、服务平台化的发展。另一方面，行业用户也将从被动接入者转变为共同创新者，参与模型微调、场景适配和闭环反馈。整体上，大模型一体机生态正从“单点突破”走向“系统融合”，形成技术共研、资源共建、价值共创的良性循环。

在大模型一体机产业生态协同深化的过程中，产业标准化和联盟化趋势日益深入。一方面，随着大模型一体机产业链条的延展与

参与主体的增加，统一的标准体系与协作机制成为推动生态高质量发展的关键基础。中国信息通信研究院人工智能研究所联合众多企业正在推进大模型一体机系列标准与评测基准，以帮助应用方提供选型参考、提升业务效率，帮助技术提供方优化产品技术能力、提升生态合作效率。另一方面，大模型一体机产业联盟化趋势加速显现。中国人工智能产业发展联盟 AI Infra 工作组正在联合大模型一体机应用方和技术提供方开展体系化的产业生态活动，包括研制大模型一体机产业图谱、大模型一体机应用案例征集、大模型一体机供需对接会议等，以推动产业各方协同创新。总体来看，大模型一体机产业生态的持续协同深化，正通过标准体系建设和产业联盟化方式，推动生态体系从碎片化走向体系化，从竞争并存走向协同共生，形成更加开放、规范、高效的产业发展格局。

编制说明

本报告由中国信息通信研究院人工智能研究所编写完成。编写过程中，得到以下单位的大力支持，在此特别感谢：

新华三技术有限公司

华为技术有限公司

北京百度网讯科技有限公司

浪潮云信息技术股份有限公司

浪潮电子信息产业股份有限公司

京东科技信息技术有限公司

深圳市矽赫科技有限公司

中国信息通信研究院 人工智能研究所

地址：北京市海淀区花园北路 52 号

邮编：100191

电话：010-62301618

传真：010-62301618

网址：www.caict.ac.cn