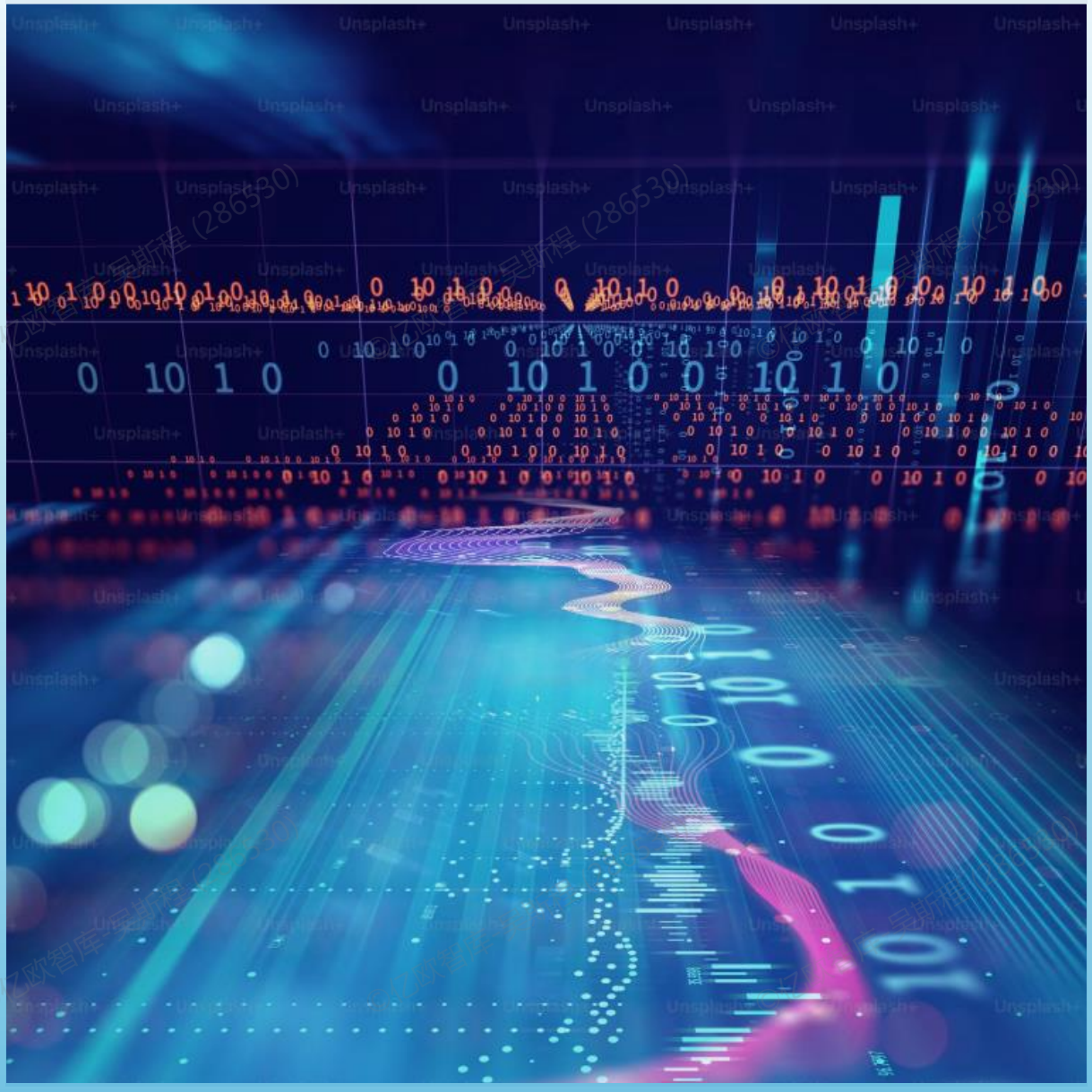


亿欧智库 https://www.iyiou.com/research

2024年企业AI大模型应用落地白皮书

亿欧智库 EO Intelligence

研究报告

目录

CONTENTS

AI大模型应用落地痛点分析

在政策支持、技术变革以及企业数字化转型需求的驱动下，中国企业纷纷开始探索并实践AI对于业务的赋能，积极推进AI大模型的深度应用与落地，与此同时对大模型应用的精度、效果、开发和部署效率等都产生了更高的需求。

但另一方面，企业在落地大模型应用的过程中仍面临诸多挑战。

本章节将重点聚焦企业落地AI大模型应用过程中的现状和需求，并对其面临的痛点和挑战进行梳理。

多重因素驱动下，企业对AI大模型的应用需求不断提高

◆ 在快速发展的数字化时代背景下，AI大模型正在成为众多企业转型升级的关键，同时，政策牵引、技术突破和转型需求等因素也驱动B端企业逐步推进了对于AI大模型的深度应用。

三大因素驱动AI大模型B端应用加速

政策牵引，环境利好

2021年以来，国家和地方层面加速出台AI应用和大模型相关政策，聚焦数据安全、技术创新、应用落地等方面，旨在降低AI应用门槛，加速大模型落地

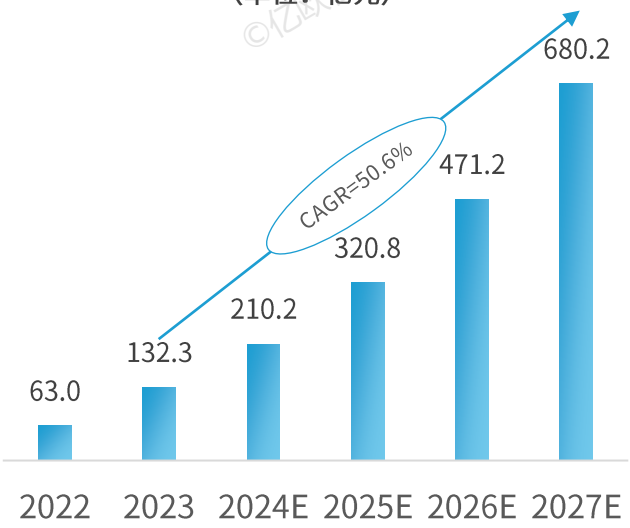
技术突破，基础支撑

深度学习、自然语言处理、多模态等创新技术的不断突破，为AI大模型的开发和应用提供了强大的基础支撑，驱动大模型落地场景的多样性和业务赋能效率的提高

转型需求，市场拉动

在日益激烈的市场竞争和多变的需求下，企业亟需AI赋能运营效率和创新能力提升，AI大模型作为数字化转型重要工具，将帮助企业更好地洞察市场、优化决策

中国大语言模型市场规模及增速
（单位：亿元）



◆ AIGC时代的第一波浪潮是大模型的预训练和训练集群规模的不断扩大，紧随其后，第二波浪潮接踵而至，当前和未来将更加聚焦AI大模型的应用落地。算力、网络等基础设施构筑起高效的计算和存储能力，并基于自然语言处理、算法与模型优化等底层技术保障大模型稳定运行，在此基础上，通用大模型能力逐步完善，并基于专业领域数据涌现出垂直行业和细分场景大模型。

◆ AI大模型在B端企业的应用落地离不开数据、算力和算法的协同支撑。其中，数据作为大模型应用的基石，主要提供丰富且高质量的训练和推理素材；算力作为基础的硬件保障，保障大模型的高效训练、优化及实时推理任务；算法作为大模型应用的核心引擎，可以定义并优化大模型的计算逻辑。

数据、算力、算法构成企业AI大模型应用落地的基础支撑

算法是AI大模型发展的核心

- 智能涌现来自于算法，有了高水平的算法才能有广泛的、高水平的应用
- 通过不断改进和创新算法，企业可以优化AI大模型的性能，提高大模型的准确性和效率

算力是AI大模型的落地保障

- 算力是算法创新的工具，算力不仅源自芯片，更源自系统性的技术创新
- 虽然现阶段大模型训练对算力要求极高，但在业务场景不断复杂的背景下，未来企业AI大模型的推理应用将对算力要求更高

数据是AI大模型应用的原料

- 企业端AI大模型的应用落地需要大规模高质量的数据支撑，并构成模型优化的基础
- 基于多模态的数据训练，AI大模型能学习更多知识，具备更专业的能力，精准高效赋能具体的业务场景优化及决策

◆ 企业落地AI大模型应用的过程就是基于数据、算力和算法的支撑，将大模型能力赋能到业务的过程，但并不意味着拥有了数据、算力和算法，就具备了大模型应用落地的能力，企业还面临从数据到应用、从开发到上线的全面挑战。

数据来源：工信部，赛迪研究院，专家访谈，亿欧智库

企业在落地AI大模型应用过程中面临诸多挑战，难以一己之力实现全流程开发落地

- 对于大部分企业来说，AI大模型应用的实际开发落地面临较高的门槛，从数据的处理到模型的微调，再到算力迁移匹配等各个环节都可能伴随着不同的挑战与痛点。
- 企业在大模型应用过程中面临数据处理工具不足、端到端解决方案缺乏以及数据隐私与安全难题，对企业大模型落地产生影响；而算力多元化和模型多样化，也给很多企业带来了算力迁移和适配以及模型选择的痛点；此外，大模型应用从开发到部署上线的全流程十分复杂，门槛较高，各环节间的协同不足。

企业落地AI大模型应用过程中面临的痛点

工具/解决方案不足

数据处理工具不足

- 缺少针对性强、高质量、易用的数据处理工具/平台。
- 开源数据处理工具的不足以及企业数据的多样性导致难以抽象出通用工具。
- 数据处理工具功能边界不明确，影响数据集质量。

缺少端到端解决方案

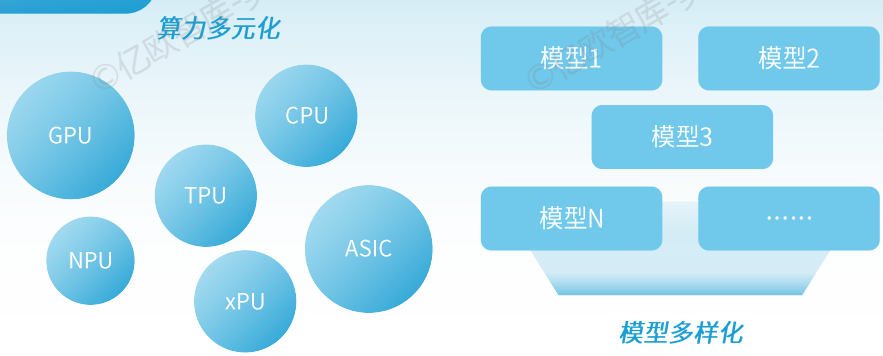
- 大模型当前以中间件或插件形式接入企业业务，涉及多种工具和智能体的集成。
- 需要与不同数据库、应用程序和业务系统兼容，当前还没有形成E2E的解决方案。

数据隐私与安全难题

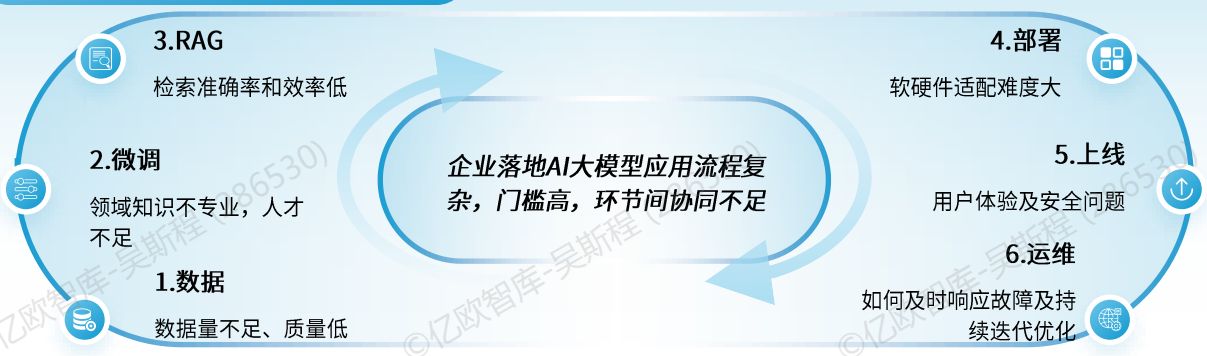
- 大模型在企业端落地面临数据安全与隐私难题。
- 企业需要私有化部署，但云平台锁定效应强，迁移成本高，隐私与安全要求严格。

算力多元化和模型多样化导致适配难

现阶段，AI大模型产业链呈现算力多元化、模型多样化的发展趋势，多元算力融合及多种模型的应用将为企业带来适配难度大和成本控制等问题



大模型应用部署全流程复杂，门槛高



数据来源：专家访谈，公开资料，亿欧智库

面向大模型应用痛点的方法论及实践路径提炼

基于企业对解决大模型应用落地痛点的诉求，中国市场上已涌现出众多服务商，但大多专注解决大模型开发部署过程中单一痛点，难以覆盖全流程，且不具备全方面的能力。进而带来企业后续更高的成本支出以及系统兼容性等问题。

全球生成式 AI 应用尚处早期，企业因数据隐私、安全、成本与性能（延迟等）因素，多采用私有云或本地环境。具有全流程解决方案能力的平台可打通大模型开发、部署与运维全流程，集成丰富授权软件，保障应用性与高效性，有效推动大模型在中国乃至全球的应用落地。 。

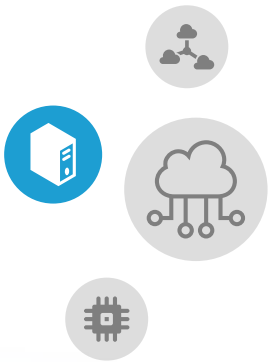
本章节基于企业AI大模型开发部署全流程的痛点与挑战，提出解决思路并提炼核心方法论，以帮助企业高效地落地AI大模型应用。

针对企业落地AI大模型应用的痛点，市场已有初步实践与尝试

◆ 为助力企业解决AI大模型在应用开发落地过程中面临的痛点，当前市场上已涌现出各类服务厂商和多样化的解决方案，例如云厂商基于自身技术积累和算力资源支持，可为企业提供标准化的产品和服务；传统AI应用开发企业基于自身在应用方面的经验积累，可为企业提供更流畅的用户体验；独立的大模型应用开发工具链企业重点聚焦算力多元化和模型多样化适配问题的解决。

市面上针对大模型应用落地的实践总结

云厂商为代表的大厂



- **强大的基础能力：**云厂商通常具备强大的技术研发和训练能力，能够提供解决大模型幻觉、多元算力调度等复杂问题的基础技术支持。
- **通用能力支撑：**云厂商在研发和训练侧构建了通用化能力支持，以满足不同企业间的共性需求。
- **充足的算力资源：**头部云厂商拥有自研大模型和充足的算力资源储备，能够为企业特别是中小企业提供必要的算力支持。
- **全面的服务能力：**云厂商不仅提供大模型技术，还具备应用服务和交付能力，能够提供从研发到部署的全方位服务。

AI应用企业为代表的服务商

- **更顺畅的用户体验：**AI应用企业在开发大模型应用平台时，更注重用户体验的流畅性，使得用户在使用过程中能够获得更加直观和便捷的操作体验。
- **深入的行业理解：**AI应用企业通常对特定行业有更深入的理解和更丰富的经验，能够提供更加贴合行业特点和需求的解决方案。
- **丰富的场景经验：**由于专注于特定的应用场景，AI应用企业能够提供更符合实际业务需求的服务和产品。

新兴大模型应用开发服务商

- **通常有完善的算力运营和调度解决方案。**独立的大模型应用开发工具链企业通过具备较强的软硬件优化技术和多元异构算力适配技术，可以赋能AI应用性能的提升。

◆ 从现阶段市场对于企业AI大模型应用落地服务的实践来看，各类产品和解决方案各有优势也各有需要补足和提升的方面，企业需基于自身实际业务需求选择合适的解决方案。但对于企业来说，需要的更多是聚焦全流程且能力全面的解决方案，真正帮助其解决AI大模型应用开发落地过程中各环节各方面的问题。

数据来源：专家访谈，公开资料，亿欧智库

解决企业AI大模型应用落地面临的痛点，需服务商具备全流程解决方案和全面的专业能力

◆ 大模型技术正逐渐成为推动企业创新和提升竞争力的关键因素。然而，企业在应用大模型时面临着诸多挑战，包括数据处理工具不足、数据隐私与安全挑战、多元算力的适配、多模型的匹配与精调，以及全流程的服务能力等。为了帮助企业克服这些挑战，亿欧智库对企业AI大模型应用落地所需的核心能力进行了梳理，提出了一系列关键能力要求，以确保企业能够高效、经济地利用大模型技术，实现业务的优化和创新。

企业AI大模型应用落地所需的核心能力梳理

破解数据瓶颈的能力

- 在大模型应用开发落地中，面临数据处理工具不足、数据隐私与安全的挑战。
- 针对数据层面挑战，供应商应提供1）灵活高效的数据处理工具，支持多种数据类型和格式的全流程处理；2）高质量数据集生成能力，确保模型训练效果；3）强大的隐私保护与安全支持，满足本地化部署和数据合规需求；以及与企业系统的无缝集成和扩展能力。此外，从数据预处理到模型部署，供应商还应提供端到端的全生命周期服务支持，以实现数据安全方面的管理闭环与有效监控。

针对多元算力的适配能力

- 硬件算力支撑多元化可能是未来的一个发展趋势，除了传统的英伟达芯片外，未来AMD以及国产GPU、ASIC市占率也将得到提升。因此面对未来的需求，大模型服务商需要提供灵活的多元算力迁移、调度方案，以确保企业能够高效、经济地利用各种计算资源，并快速适应不断变化的业务需求。
- 需要关注服务商是否具备强大的硬件适配层以实现跨平台兼容性，同时能基于现有硬件提供优化方案以扩展对复杂模型的支持能力；此外，服务商还需提供灵活的算力资源分配策略，根据企业实际业务需求灵活调配算力资源，例如针对实时性要求高的场景可优先分配高性能AI算力

多模型匹配和精调能力

- 针对企业实际业务场景的差异化，通常需要多种大模型能力的支持，因此要求服务商能够根据企业的业务情况和要求提供不同的基础大模型以及对模型进行适配与精调的能力
- 不同业务场景对大模型的能力要求各不相同，例如，智能客服场景需要大模型具备较强的自然语言理解和生成能力，而智能制造场景则更侧重于数据分析和预测能力。因此，服务商提供多个大模型或能够基于不同大模型进行定制优化，可以确保企业能够根据不同的业务场景选择或调整最适合的模型，从而提高业务效率和效果

全流程打通及服务能力

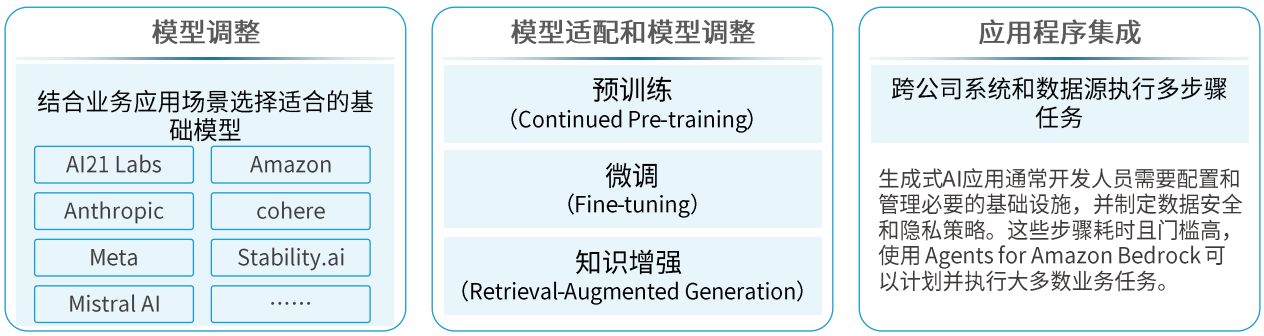
- 企业要想实现AI大模型应用的顺利落地，需要服务商具备全流程解决能力，从数据处理，到模型微调，以及RAG技术引入和知识库构建，再到最后的应用部署和上线运维等，以确保大模型应用的性能优化与精度提升
- 同时，服务商还具备提供定制化方案的能力，灵活应对企业需求，帮助企业梳理大模型应用落地的完整流程。大模型应用各阶段都涉及大量的专业知识和技术，很多企业特别是中小企业不具备专业化能力，需要服务商在各环节都能提供专业服务，并基于企业自身实际应用场景不断优化，帮助企业持续优化模型应用性能和效果

数据来源：专家访谈，公开资料，亿欧智库

企业平台最佳实践：为企业提供AI大模型落地平台，全流程支撑企业各阶段需求

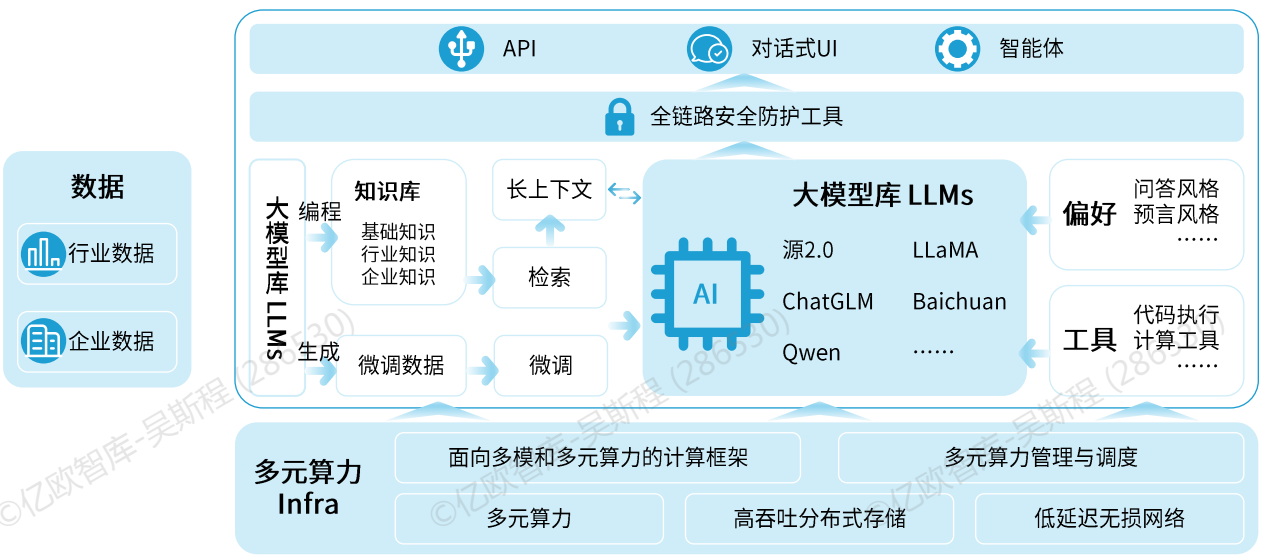
- Amazon Bedrock：是一项完全托管的服务，通过单个API提供来自人工智能公司的高性能基础模型，以及通过安全性、隐私性和负责任的人工智能构建生成式人工智能应用程序所需的一系列广泛功能。使用Amazon Bedrock，开发者可以试验和评估适合业务的基础模型，通过微调和检索增强生成等技术利用企业数据对其进行私人定制，并构建使用企业系统和数据来源执行任务的代理。

AWS Bedrock使用基础模型构建 AI 应用



- 浪潮信息元脑企智EPAI为企业AI大模型落地应用提供端到端开发平台，囊括了全链路、高可用的系列能力工具，能够帮助企业有效降低大模型应用门槛，帮助伙伴提升模型开发效能，打造智能生产力。
- 元脑企智EPAI支持包括CPU、GPU和各类加速卡在内的20+多元计算芯片，通过上层模型算法和下层基础设施的逻辑解耦，降低企业跨算力平台迁移、多元模型部署适配的试错成本，助力企业轻松跨越AI应用开发与部署门槛，加速智能应用创新。提供数据准备、模型训练、知识检索、应用框架等系列工具，支持调度多元算力和多模算法，帮助企业高效开发部署生成式AI应用。

浪潮信息元脑企智EPAI



数据来源：公开资料，亿欧智库整理

基于标杆案例梳理提炼AI大模型应用的路径

- 结合国内外优秀案例，AI大模型落地服务厂商可以提供一站式解决方案，覆盖数据准备、模型选择、模型训练、模型定制、模型部署、应用集成、测试验证以及上线运维等各个环节。但在此之前，企业需要明确落地应用场景以及未来效果预期。
- AI大模型的真正价值和投资回报率取决于企业本身如何推动AI大模型的落地，也取决于AI大模型技术的应用如何为企业带来业务层面的改变。企业应该尽快评估准备情况，制定人工智能战略与落地路线图，为生成式 AI 的应用奠定必要的基础，从而在中长期内通过差异化和重点战略来建立竞争优势。

企业落地AI大模型应用的路径梳理

1.需求梳理及规划

首先，企业需详细梳理业务场景及需求，并快速构建应用原型，通过用户测试或内部评审来验证需求的可行性和用户接受度。并提前对可能用到的技术栈进行预研，包括模型能力、数据处理技术、前后端框架等，确保技术选型与业务需求匹配。

其次，企业需基于业务场景应用需求，选择具备破解数据瓶颈、多元算力适配、多模型匹配和精调、全流程打通能力的供应商。

2.基础大模型部署

基于企业业务场景实际应用需求，选择合适的基础大模型和形式进行部署，如采用API模式部署，需确保接口稳定、安全，并考虑网络延迟对应用性能的影响；如采用本地化部署方式，需确保部署环境的兼容性，提高部署效率。

3.模型微调与定制

根据业务场景的实际应用需求，需对基础大模型进行微调，以提升大模型在特定场景下的性能。复杂且专业度高的场景还需引入RAG技术，内部构建或接入知识库，增强模型理解和生成能力。

在此过程中，还会涉及到数据处理流程的优化，包括数据清洗、标注、增强等，提升数据质量，并实施严格的数据访问控制和加密措施，确保数据处理、存储、传输过程中的安全性。

4.应用测试及验证

AI大模型应用正式上线前需做测试，包括单元测试、集成测试、性能测试、安全测试等，确保应用各模块功能正常、性能达标，并保证安全可靠。

同时，企业还需部署性能监测工具，实时监控应用运行状况，及时发现并解决性能瓶颈。

5.应用上线及运维

通过测试后，需制定详细的应用上线计划，包括回滚策略、应急响应预案等，确保上线过程平稳，并将AI应用与企业现有业务系统做集成，实现数据共享等

在企业内部搭建运维团队，负责应用的日常维护、性能监控、故障排查及持续优化，并建立用户反馈机制，定期收集用户意见，不断优化应用功能和用户体验

数据来源：公开资料，亿欧智库整理

未来趋势预判及策略建议

随着数字化转型的快速发展，AI大模型应用落地将成为每个企业不得不面临的重要趋势，且随着应用的不断深入，未来企业将越来越重视AI大模型对于实际业务场景的有效赋能，同时将对数据、算力和大模型性能等产生更高的需求。

而低门槛、高效地落地AI大模型应用，让其能够真正赋能自身业务效率的提高和成本的降低，不仅需要选择和自身更匹配的大模型，更需要高效的算力基础设施支撑，同时，高效的数据的采集和处理方式、大模型应用的部署模式以及相关团队的有效支持等都必不可少。

本章节将重点分析企业未来在大模型应用落地方面的趋势，并为企业更高效的大模型应用赋能提供策略建议。

企业未来落地AI大模型应用部署的四大趋势研判

1、企业已经感知到大模型的价值，未来将逐步关注ROI

- 全球75%的CIO增加了2024年的人工智能预算。然而，当谈到生成式 AI时，许多组织并没有以正确的方式部署和利用它来释放其潜力。过往在小模型时代，从应用场景到赋能效果都存在清晰可参考的落地路线，然而大模型在这个方面来还没有形成标准案例。
- 目前，多数企业处在大模型的探索阶段，已经在场景应用上感知到大模型的应用价值。未来企业逐步关注大模型投入的ROI上，即大模型是否能真正帮助企业业务实现降本提效。根据Gartner调研，企业领导者期望到2024年大模型可以替代企业3.8%的岗位，到2026年可以替代8.2%的岗位。

2、多模态大模型应用解决多维度业务问题

- 在当前的应用中，大语言模型仍是主流，但世界是多模态的，多模态协同更符合人类感知与表达方式。在实际业务场景中，通过引入图片、语音、视频等数据形态，大模型可帮助企业解决更多维度的问题，多模态也是当前业界的重点发展趋势。由于多模态模型可以捕获跨模态的复杂数据关系，将融合不同信息产生更多样化的结果，参与到更深层次的任务中， 因此相比单模态模型具有更广阔的应用场景，如医疗健康、交通(交通指挥，自动驾驶等)、安防监控等复杂环境。

对比维度	单一模态大模型	多模态大模型
信息的丰富性和完整性	具有局限性，表现为信息不全面和上下文缺失： • 信息不全面：例如仅依赖文本描述可能无法准确理解一个场景；仅依赖图像可能无法准备获取文字内容和背后的含义 • 上下文缺失：单一模态缺乏上下文；如仅有图像信息无法理解其内容	具有丰富性，表现为信息互补和上下文增强： • 信息互补：不同模态的信息可以互补，例如图像提供视觉信息，文本提供详细描述，两者结合效果更好 • 上下文增强：多模态信息可以提供更丰富的上下文，有助于更准确的理解和决策
增强任务表现	单一模态的数据可能会导致部分歧义，如在网络上看到一张图片，描述和人的想象可能完全不同	能对任务进行增强，比如提升准确性和扩展任务范围 其次，多模态可以执行跨模态任务和复杂任务，如自动驾驶领域需要大模型能同时处理视觉、文本、雷达等多类型数据
人机交互的自然化和智能化	交互僵化	交互更自然，可以通过自然的表达，来让大模型理解人类的喜怒哀乐，以及工作和生活习惯

3、将RAG与知识图谱相结合，进一步提升AI在复杂查询处理中的性能

- 当前面向文档类数据检索增强的方法以基于向量数据库通过文本向量化的方式为主，7月微软开源了GraphRAG成为下一个行业热门研究方向。传统的RAG存在一些局限性，如缺乏对实体间复杂关系的理解、固定数量的文本块限制等，将知识图谱引入RAG可以解决这些限制，因为知识图谱提供了一种结构化的方式来表示实体及其关系，使得系统能够：

捕捉实体间的复杂关系

知识图谱通过节点和边的结构清晰地表示实体及其相互关系，使得RAG能够理解和利用这些关系来生成更准确的回答

提供全局上下文理解

知识图谱的全局性视角使得RAG能够跨越文档界限，整合分散在不同文本块中的信息，提供更全面的上下文理解

支持个性化和推荐系统

知识图谱能够根据用户的行为和偏好提供个性化的内容推荐和服务匹配，提升用户体验

数据来源：专家访谈，亿欧智库

企业未来落地AI大模型应用部署的四大趋势研判

4、智能体朝单一智能体能力扩展与多智能体协作方向发展

- 智能体的应用场景广泛，包括但不限于机器人、自动驾驶、智能家居等，现在各类应用中或多或少都在构建让用户去使用的智能体，未来会形成更加复杂的智能体使用情况。扩展单一智能体边界使其能够兼顾多类任务，或者构建多智能体协作机制可能是未来两大落地方向。
- 其中，多智能体框架开始利用层次结构，使一些智能体专注于高级目标，而其他智能体则负责特定于任务的工作，然后向上报告，从效率提升角度看，多智体系统通过智能调度、自动化流程显著提高工作效率；在创新赋能方面，通过跨领域知识融合、创新模式探索，激发行业创新活力；在生态构建层面，多智体AI Agent能构建开放、协作的智能生态系统，推动产业链深度融合与价值共创。

策略建议：基于全流程开发平台底座，实现高效低门槛落地AI大模型应用

◆ 在当今数字化转型的浪潮中，AI大模型的应用已成为企业提升竞争力、优化业务流程的重要手段。面对AI大模型在企业端落地的困难需要基于全流程开发底座来实现以下价值：

01 聚焦业务场景需求并合理选择模型

企业首要任务是深入剖析业务需求，定位核心难题。鉴于大语言模型在文本处理、自然语言理解上的高成熟度和实用性，常被企业视为首选，能显著提升信息处理和生成效率。
若业务涉及多模态处理（如图文生成、视频处理等），则需权衡技术可行性与成本效益。此时，模块化或微服务架构成为优选，将多模态功能作为独立服务融入系统，确保稳定性和扩展性。同时，通过优化系统架构，可有效降低技术集成难度和成本。

02 明确任务性质结合业务逻辑优化流程

确定业务需求与模型后，企业需明确任务性质：辅助生成与决策性任务。辅助生成如文档、代码补全，应利用大模型提升内容质量与效率，优化输入输出流程。决策性任务如风险评估、市场预测，则采取混合智能，结合大模型预测与企业决策逻辑，确保大模型输出为决策辅助，而非替代人类判断，以维持企业控制力并提升决策科学性。

03 评估自身数据丰富度及质量

在梳理业务场景、选择模型和优化流程之后，企业需要开展数据评估工作。评估的重点包括数据的丰富度、多样性、时效性和隐私保护要求等方面。企业需要确保可获取的数据量足以支持模型的训练和微调，并评估数据的多样性以确保模型能够泛化到不同的应用场景。此外，企业还应关注数据的时效性和隐私保护要求，确保数据的合法合规使用。

04 明确技术选型与适配性

企业需要考虑框架的成熟度、易用性、可扩展性以及与企业现有系统的兼容性等因素。通过综合评估这些因素，企业可以选择最适合自身需求的AI大模型框架（如Transformer、BERT、GPT等）。同时，企业还需要确保所选技术能够与企业现有系统高效集成，减少迁移和改造成本。

05 建立持续学习与迭代机制

AI大模型的应用不是一次性项目，而是一个持续优化的过程。因此，企业需要建立大模型应用性能监控和反馈机制，定期评估模型效果并根据业务需求变化和用户反馈进行调优和迭代。通过持续优化大模型的性能和应用效果，企业可以确保AI大模型始终能够满足企业的业务需求并带来实际的价值。

06 培养独立的AI人才与团队

为了支撑AI大模型的应用和发展，企业需要引进并培养具备AI技术能力的专业人才。这些人才将负责大模型的微调、RAG技术的引入以及与其他系统的集成等工作。通过构建独立的AI团队，企业可以确保AI大模型应用的持续性和创新性，为企业的数字化转型和智能化升级提供有力支持。

07 探索大模型应用与业务的深度融合

目前，大模型应用更多用于企业的辅助生成场景。然而，随着技术的不断发展和应用场景的不断拓展，企业应积极探索AI对企业决策的赋能作用。通过数据分析和预测模型等手段，企业可以为管理层提供科学决策依据，推动企业的数字化转型和智能化升级。同时，企业还可以将AI大模型与核心业务深度融合，实现业务流程的自动化和智能化，进一步提升企业的竞争力和创新能力。

数据来源：专家访谈，亿欧智库

亿欧智库

网址：https://www.iyiou.com/research
邮箱：hezuo@iyiou.com
电话：010-57293241

扫码关注亿欧智库
查看更多研究报告

扫码添加小助手
加入行业交流群

北京：北京市朝阳区关庄路2号院中关村科技服务大厦C座4层 | 上海：上海市闵行区申昆路1999号4幢806
深圳：广东省深圳市南山区华润置地大厦 C 座 6 层 | 纽约：4 World Trade Center, 29th Floor-Office 67, 150 Greenwich St, New York, NY 10006