



大模型应用落地白皮书

企业 AI 转型行动指南



目录

CONTENTS

核心观点	01
第一章：祛魅务实，大模型加速从探索走向落地	02
1.1 业务驱动，大模型助力效率飞跃，实现融合的体验创新	03
1.2 百舸争流，大模型服务商竞逐AI浪潮新时代	06
第二章：知易行难，企业落地面临的挑战与机遇	08
2.1 大模型落地面临多重挑战	09
2.2 领先企业已从大模型成功落地中率先获益	13
第三章：加快AI转型，构建全方位的大模型业务落地能力	17
3.1 大模型业务落地能建设三阶段	19
3.2 破除落地大模型的思维误区	23
第四章：头雁效应，大模型深入众多应用场景	25
4.1 大模型应用场景不断扩宽，应用日渐成熟	26
4.2 众多行业企业深入大模型落地实践	30
第五章：攻克有径，跨越大模型落地技术难题	50
5.1 大模型落地部署技术步骤	51
5.2 精准选模、高效落地、持续挖掘——落地三要素	55

第六章：信赖之选，火山引擎大模型服务助力省心AI转型..... 58

6.1 豆包大模型 59

6.2 火山方舟大模型服务平台 61

6.3 扣子 64

6.4 HiAgent 66

结语及未来展望 69

核心观点

● 大模型技术已经迈入与业务深度融合的阶段

如同云与机器学习、深度学习等AI小模型一样，当前大模型技术也进入了与业务深度整合的关键时期。64%的中国企业预计其对AI的投资将增长10-30%，各企业正基于具体的业务场景，积极探索大模型技术的实际部署与应用潜力，以促进业务发展。

● 大模型在企业落地的周期和应用速度超出预期

大模型能够大幅提升企业生产力、驱动业务创新和增长，是具有突破性和颠覆性的技术。尽管将大模型技术与业务深度结合是一个复杂的专业过程，但在专业技术服务商的持续支持下，企业实现大模型部署的周期已缩短至平均6-12个月，尤其在数字化领先的企业中，这一过程更为迅速。

● 企业正致力于深化业务场景探索，以期释放大模型落地带来的价值

为了充分利用大模型带来的商业机遇，企业正不懈探索技术与业务结合的新模式，逐步打造与自身发展战略相契合的大模型应用实践，以期通过大模型技术的创新应用为企业带来更大的商业价值。

● 大模型产品和服务的选择和技术伙伴能让企业落地大模型事半功倍

超过47%的企业认为，与领先的大模型厂商建立可靠的合作关系是项目成功的关键。为提升大模型在企业侧的落地效率，企业应根据自身业务需求和发展规划，重点考虑具备AI大模型全栈开发能力、模型及插件工具丰富、内嵌垂直场景经验模板的厂商，并选择可以提供事前、事中、事后全周期咨询和切实可行实践方案的服务商通过制定全局周密的策略，并结合个性化的业务实践经验，企业将能够更有效地推动大模型的成功落地，加快AI转型。

01

祛魅务实

大模型加速从探索走向落地

1.1 业务驱动，大模型助力效率飞跃，实现融合的体验创新

大模型技术已经成为人工智能领域的焦点，驱动了AI应用的升级和创新。基于大规模数据集的预训练，让模型具备了高效处理复杂语言结构、视觉模式以及多模态信息的能力，促使人工智能在认知理解与决策支持方面迈向更高的阶段。在产业应用方面，大模型的实施正在改变企业业务开展、产品服务、运营管理的传统模式，为各行业的智能化升级提供了强有力的支撑。企业端对大模型的应用也已从初步的技术探索与创新尝试，逐渐步入以实现商业价值为核心的新时代。众多企业正加快步伐，寻找能够迅速构建商业闭环、验证价值的应用场景，期望借助大模型实现运营效率的飞跃、用户体验的创新，以及生产力的提升。

● 1.1.1 加码投资，企业扩大试点

加大投入力度，AI大模型的关注度持续攀升。AI大模型在提升流程效率、增强创新能力方面发挥着关键作用，并为企业在市场竞争中赢得优势，例如在个性化营销内容创作、产品设计创新、用户体验优化、员工助手打造以及知识专家系统开发等方面。IDC全球调研数据显示，37.7%的受访企业正在重点投资AI大模型，并预计在未来三年内引入AI软件及相关培训和服务。此外，64%的中国企业预计其对AI的投资将增长10-30%。这一显著的投资增长反映了业界对AI大模型巨大潜力的日益认可。

拓展试点范围，企业正积极探索AI的应用潜力。企业在AI大模型的应用上展现出极高的兴趣和参与度，他们通过内部研发、联合创新、跨界合作等多种模式，积极探索AI技术的实际应用场景和潜在爆发力。IDC调研数据显示，在过去一年里，全球范围内平均对AI大模型项目进行了34次概念验证（POC）测试，这一数字远超其他IT项目，且企业对AI大模型测试的满意度高达70%。这一数据表明，AI大模型在解决企业需求、优化运营流程、辅助业务拓展方面具有显著效

果。此外，大多数CXO级别的高管表示，他们将继续增加对AI大模型的各项投入，以加强企业在AI转型方面的能力建设。

● 1.1.2 多维价值，坚定企业探索大模型潜力的决心

IDC研究认为，大模型技术对于企业价值的贡献可从对内赋能与对外服务两大维度进行阐释。具体而言，该价值可细化为针对企业员工、用户群体、营业收入及市场拓展四个方面，进而构建出一个全面的“AI大模型价值图谱”。此图谱详细描绘了大模型技术在不同维度上对企业产生的深远影响和价值增值，包括工作效率提升、业务引导助理、知识数据洞察、用户体验创新、生产工具赋能、产品服务升级、管理方式变革、市场营销决策8大方向。

图1 AI大模型价值图谱



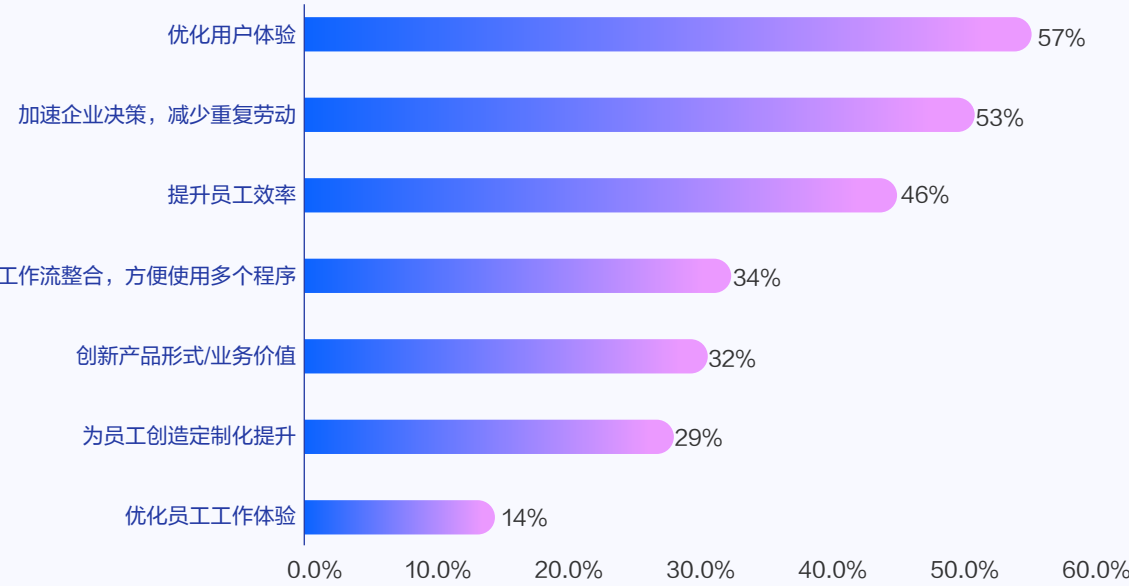
来源：IDC大模型价值研究，2024年

- ◆ **面向员工：**一是利用AI大模型提供的办公软件、流程管理软件、开发设计软件来提高员工工作效率、缩短重复和复杂工作时间；二是利用企业内部知识、经验等无形资产和员工培训、企业规范、服务规则等文本材料，打造具备企业记忆的专属AI智能体。
- ◆ **面向用户：**一是对用户信息、消费习惯、兴趣爱好等信息进行综合分析，更好地总结、对比、预测用户数据变化和趋势；二是提供全新的使用交互体验，如借助智能客服、数字人等，更好地理解用户意图并进行产品推荐。
- ◆ **面向营收：**一是为设计、编程、制造等工作人员的生产工具赋能，自动生成并创新内容，加快产品的迭代周期和创新速度；二是创新产品形式，从产品管理、价值创新、上市计划、价格管理等方面更好地赋能业务。
- ◆ **面向市场：**一是主动分析外部市场变化和内部经营数据，进行合规管理、风险预测、潜在效益分析和建模；二是预测整体市场表现、各地区竞争分析、供应商能力和相关风险，帮助制定策略，并自动生成相关宣传物料。

从AI大模型价值图谱可见，大模型以其强大的能力和广阔的应用前景，有望深入到众多领域，释放出巨大的企业应用空间和潜力。

此外，根据IDC针对企业用户的大模型调研数据可以看出，无论是优化用户体验、加速企业决策，减少重复劳动、提升员工生产效率，或是推动产品创新与服务个性化，大部分的受访企业都普遍对大模型带来的价值抱有高度期待。这一预期进一步提升了企业将大模型技术转化为实际应用的动力。企业普遍认为大模型在推动业务增长、构建核心竞争力、提升用户体验等方面扮演着关键角色。因此，也有越来越多的用户更加积极地投入到这场技术革命中，深挖其背后的商业潜力。

图2 采用AI大模型给企业带来的价值



数据来源：IDC大模型企业落地进展调研，N=100，2024年8月

1.2 百舸争流，大模型服务商竞逐AI浪潮新时代

在企业用户持续增加的投资和日益增长的业务需求推动下，大模型服务商们纷纷加快步伐，竞相推动这一技术从创新突破走向企业应用的实际转化，致力于将大模型从服务商自身的探索阶段推向企业落地，以满足各行各业对智能化转型的迫切需求。这一趋势不仅促进了大模型技术的快速成熟，也为企业用户带来了更加丰富和高效的应用解决方案，助推了大模型落地需求端和大模型技术服务供给端的双向循环发展。

一方面，大模型技术服务在产品服务和技術层面实现了诸多的创新突破。为降低企业的使用门槛，领先的大模型厂商已经成功构建了从零到一的端到端解决方案，极大地减少了重复开发的必要性。企业可以借助这些既有模型和服务，避免“重新造轮子”的投入，从而将更多资源和精力集中在业务创新上，加速智能化转型的步伐。例如火山引擎持续丰富豆包大模型家族，基于基础底座大模型最新

发布视频生成、文生图、图生图、语音合成、声音复刻、音乐、同声传译等更贴合实际场景的模型，将模型能力进一步专业细化。此外，为加速企业搭建内部智能体的进度，火山引擎还打造了HiAgent——一个企业专属的AI创新应用平台；它允许业务人员利用提示词、知识库、插件等工具，以低代码方式实现AI落地，集成内部数据，降低AI开发的难度、积累AI中台的厚度。

另一方面，大模型技术服务商也在不断地进行能力全面升级，深度服务业务场景。为了帮助企业更有效地落地大模型技术，技术服务商们不遗余力地采用了多种手段来升级自身的服务能力，深入挖掘各种契合企业业务场景的需求。除了持续的技术研发、优化算法、提升数据处理能力之外，他们还积极加强对行业知识的积累，确保为企业能提供落地、高效率的解决方案。一些大模型技术服务商还通过构建行业大模型落地联盟、提供专业咨询和培训服务、加强落地场景探索等，全方位地提升自身在帮助企业应用大模型过程中的专业性和实用性，从而确保AI大模型能够真正融入企业的核心业务，发挥出最大的价值。如：**火山引擎围绕行业需求场景，构建大模型应用生态**，连续成立汽车大模型生态联盟、智能终端大模型联盟、零售大模型生态联盟，联合业界头部企业、技术厂商和合作伙伴，帮助企业能够以极低的试错成本将大模型技术应用落地到业务场景，共同探索AI驱动未来发展，提升运营和开发效率，全面创新用户体验。

02

知易行难

企业落地面临的挑战与机遇

在评估大模型的商业潜力时，企业和技术服务商都普遍持乐观态度，并认识到这一技术有望开启新的商业机遇。然而，将大模型的潜力转化为实际业务效益的过程，无疑充满了诸多挑战。在此次革命性的大模型转型旅程中，企业不仅要拥抱创新带来的效率和效能提升，还需精心应对成本控制、人才短缺、技术复杂性以及信息安全等一系列考验。

2.1 大模型落地面临多重挑战

● 挑战1：高成本、复杂投入下的投资收益挑战

IDC调研显示，算力成本、隐形的机会成本、投资回报的长周期以及不足的人才储备是企业落地大模型遇到的第一道难题，具体包括：

▶ **算力成本：**算力资源的消耗是当下阻碍AI大模型落地的最主要因素，这一成本对不少企业的财务状况构成了较大的压力。调研数据显示，92%的企业认为在大模型工程化落地阶段，缺少算力资源是最大的挑战；细分来看，89%的高管认为模型训练成本高，81%的高管认为模型推理成本高。对比而言，模型的调优成本已经相对低价，仅有35%的企业认为模型调优（Fine-tuning/Prompt）成本高。这些成本无疑是企业财务投入的重要考量点，使得企业在做出是否引入大模型的相关决策时显得犹豫不决，不得不权衡预算和投资回报之间的比重。

图3 算力资源是企业落地大模型最大的挑战



数据来源：IDC大模型企业落地进展调研，N=100，2024年8月

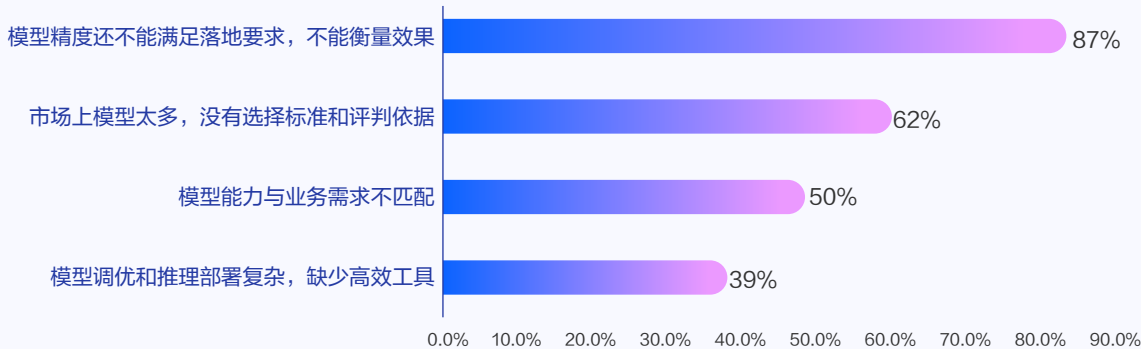
- ◆ **选择错误下的机会成本：**机会成本，虽然并非直接的经济支出，但它同样是企业在是否引入大模型决策时不可忽视的成本之一。它代表了企业为了选择某一方案而放弃的、可能是更佳替代方案的价值。尤其在瞬息万变的市场环境中，企业抓住机遇的时间窗口极为有限；一旦错失或在大模型初始的选择时对其性能、适配度、应用能力等方面欠考虑，企业可能面临巨大的机会成本，甚至可能要承受数倍的额外负担，如后期不得不进行的模型切换、基础设施的改造升级等。
- ◆ **实际与预期回报仍有差距：**根据IDC的调研，虽然企业对AI大模型项目抱有极高的期望，普遍期待能够实现1-3倍的投资回报率（ROI），但现实情况却与这些美好愿景有所偏差。目前，大多数企业观察到的投资回报实际上低于50%，这一数据与其对大模型的高度关注和前期投资预期形成了较大的落差。面对这样的现状，企业在投入AI大模型项目时，普遍较为审慎，且在内部投资决策时受到较大的阻力。
- ◆ **多方面人才积累不足：**目前多数企业在AI人才方面的储备尚显不足。企业落地大模型不仅亟需专业的大模型开发人才，还需要具备算法优化、数据处理、计算能力、测试验证等技能的实用型人才。由于AI大模型的全方位开发人才难以在短期内迅速积累，企业在探索和实施大模型应用的过程中，内部人才不足，而外部招聘难觅懂自身业务又了解大模型落地的专业人才，使得落地受阻。

● 挑战2：模型选配难题精准匹配难，适配挑战加剧

- ◆ **部分场景下模型效果接受度低：**87%的企业认为模型精度还不能满足落地要求，无法衡量具体效果；具体表现在涉及用户信息、面向生产和决策的任务中，对模型的逻辑推理、任务执行要求更高，而当前大模型的泛化性使得企业在模型优化上面临更大的挑战。

- ▶ **大模型选型困难：**在模型的选择方面，IDC调研显示，62%的企业认为市场上模型选择太多，缺乏选择的标准和评判依据。企业通常参考模型准确率排行榜、社区口碑推荐来选择模型，并自行搭建内部业务数据测试集来简单评测效果，而这一过程又缺少完整、科学、丰富的评估体系和标准。
- ▶ **通用模型能力与专业需求不匹配：**IDC调研显示，50%的企业认为模型能力与业务需求不匹配，原因是通用大模型无法满足专有场景的需求。
- ▶ **模型上线性能难以保证：**正式上线后的服务响应时间、稳定性、并发数、高吞吐、可扩展性往往不可预测，而这些正是企业最关心的问题。

图4 公司落地大模型面临的挑战-模型层



数据来源：IDC大模型企业落地进展调研，N=100，2024年8月

● 挑战3：模型部署落地：细节挑战遍布，每一步都是考验

- ▶ **效果调优路径多、执行难：**59%的企业认为模型调优（Prompt Engineering/Fine-tuning）是大模型开发中投入最多且挑战更大的工作之一。由于模型优化方式、路径、调整程度选择多样且企业缺少足够经验和技術支撑，导致该过程复杂且耗时，使得经验欠缺的企业在执行过程中面临众多困难。
- ▶ **分裂的开发环节造成事倍功半：**企业在实施大模型的落地过程中，涉及从开发到部署的多个复杂步骤，包括但不限于二次训练、数据管理、参数优化、

效果精细调整、Prompt工程、RAG（检索增强生成）、生态插件集成、模型性能评估、模型剪枝与蒸馏、模型维护管理以及算力资源调度等十余个关键环节，每个环节都涵盖了众多细致的工程化任务，对技术团队和开发人员来说是一项复杂而繁琐的挑战。

- ▶ **没有稳定的数据底座则无法发挥最大价值：**在推进大模型应用的过程中，数据的作用至关重要。多模态数据的实时采集、处理、分析，以及确保数据的质量、安全和有效管理，是企业在实施大模型时必须优先考虑的核心问题。IDC调研显示，企业在数据资产管理、构建数据架构以及数据维护等方面遇到了不少挑战。特别是在开发AI应用时，有68%的企业高管认为有必要对内部数据资产进行整理，66%的高管期望建立如数据湖的支撑性架构，还有62%的高管认为需要投入资源构建知识管理体系，并对数据进行清洗和标注。

● **挑战4：潜在安全风险与可解释性的双重挑战**

- ▶ **缺少全周期安全可信方案：**大模型服务拥有更长的链条，涉及全周期的数据和模型管理、模型调优、使用交互、查询调用等，因此需要考虑为大模型搭建专门的安全模块。
- ▶ **模型生成内容准确性和可解释性不足：**8%的企业表示希望提高模型和数据的可解释性，这需要大模型学习、链接关键业务信息，并在模型微调提高精度的同时，确保关键信息不泄露、不被模型学习了解，以及数据不可随意被查看和问询。

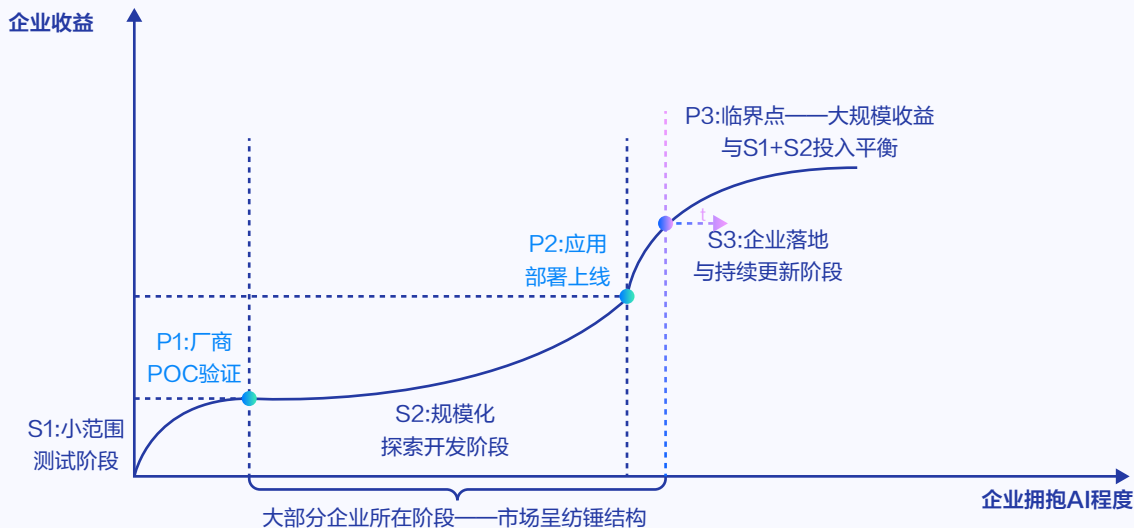
2.2 领先企业已从大模型成功落地中率先获益

在积极探索和落地大模型的过程中，企业虽然面临不同程度的挑战，但整体来看，先行拥抱和落地AI大模型的企业已经展现了清晰的收益曲线，不少率先解锁大模型红利的先锋实践案例为期待大模型为自身业务带来变革的观望者注入了落地的动力。

2.2.1 大模型落地收益曲线

IDC以如下收益曲线以描述市场情况，横纵坐标分别表示企业拥抱AI的程度和企业获得的收益。在企业跨过前期探索和正式投资建设后，AI大模型的部署和落地可以为企业带来明显可见的收益，包括工作效率的提升、用户体验的升级和市场竞争力的增强。但投资回报与前期投入持平的时间点并不是固定的；企业对大模型拥抱程度越高、资源投入和落地范围越大、应用深度越深，这个时间点也会越早到来，企业收益也会更明显。

图5 大模型收益曲线



来源：IDC，2024

- ◆ S1: 小范围测试阶段——即开始尝试探索AI大模型应用阶段。该阶段资源投入较少，可明显为企业带来新的体验和价值增长；但由于仅处在小范围验证阶段，所以增益空间有限。
- ◆ S2: 规模化开发阶段——即进入大模型服务规模化投入阶段。由于需要整体架构的开发和服务解决方案的落地和时间验证，企业所获得的收益并不明显；但随着大模型应用的部署上线，带来的收益也随之增长。
- ◆ S3: 企业落地与持续更新阶段——即不断完善更新大模型服务来增加收益阶段。该收益得益于内部成本降低、人员效率提高，以及外部产品服务创新升级带来的市场竞争优势。
- ◆ P1: 厂商POC验证——该点表示为企业在经历大模型小范围探索后进行大规模资源投入前的厂商服务验证POC时刻。
- ◆ P2: 应用部署上线——该点表示为AI应用大规模部署上线的时刻。
- ◆ P3: 盈亏平衡临界点——该点表示大模型带来的收益与投入支出平衡的时刻。

● 2.2.2 抢跑AI时代：先锋企业率先解锁大模型红利

实例揭示，AI大模型在企业应用中的成效显现。大模型具备可以快速落地、不用重复进行基础工作建设的优势，因此对企业而言落地更快、更易见效。IDC调研显示，大模型的落地周期多在6-12个月（48.5%）和12-18个月之间（30.3%）。相比于其他IT解决方案，AI大模型应用可以更快地上线；尤其是对于互联网企业或是已有AI应用基础的企业，新的AI大模型能力、垂直场景应用的上线时间更短，最快可1个月内完成落地。因此，率先落地的先锋企业已经收获大模型落地的红利。

上汽乘用车利用豆包大模型更快聆听反馈“用户之声”。上汽乘用车通过引入豆包大模型，可以快速处理来自公域和私域的“用户之声”，并精确理解每条内容，在服务反馈、热点事件跟踪和质量改进等领域取得显著成效，为用户带来更加优质的体验和服务。

火山引擎助力中国飞鹤落地AI技术。火山引擎提供HiAgent平台，针对企业级市场进行深度定制，通过段位划分策略，帮助企业员工逐步掌握HiAgent使用，同时通过原厂咨询、内置最佳实践应用效果评测确保达到生产级标准。平台将AI技术融入各个业务流程中，从而提升整体运营效率和市场竞争力，为数字化转型提供了坚实基础；在某智能问答项目中，能够处理广泛的消费者咨询，实现100%的问答响应率，同时保持了超过95%的高准确率。

火山引擎携手中手游为《仙剑世界》打造AI NPC。利用火山引擎RAG方案，中手游打造游戏精灵“圆满”，使其具备更智能的辅助功能，其可以根据玩家的行为数据和游戏内的进程，提供最适合玩家角色当前状态的支持，如任务推荐、玩法说明等，从而减少玩家在游戏时的阻碍感，提升游戏乐趣。火山方舟平台充分满足中手游对游戏的高RPM/TPM要求，通过按需调整配额，保障游戏在超大流量下的服务稳定性，并打造极致性价比，助力降低技术、资金与人力成本。

企业CXO普遍预计，未来一年大模型服务将会带来更多的效益改善。面向未来，随着AI的进一步能力升级和服务范围扩大，企业内高管预期大模型会从多个维度带来实际效益。IDC数据显示，预计未来一年大模型可帮助企业降低18%成本、增加2%收入、缩短24%流程时间、提高17%员工工作效率、提高7%资产利用率、提高19%产品创新水平、提升6%业务洞察力、提升2%高管决策速度。

图6 未来一年AI大模型给企业带来的效益



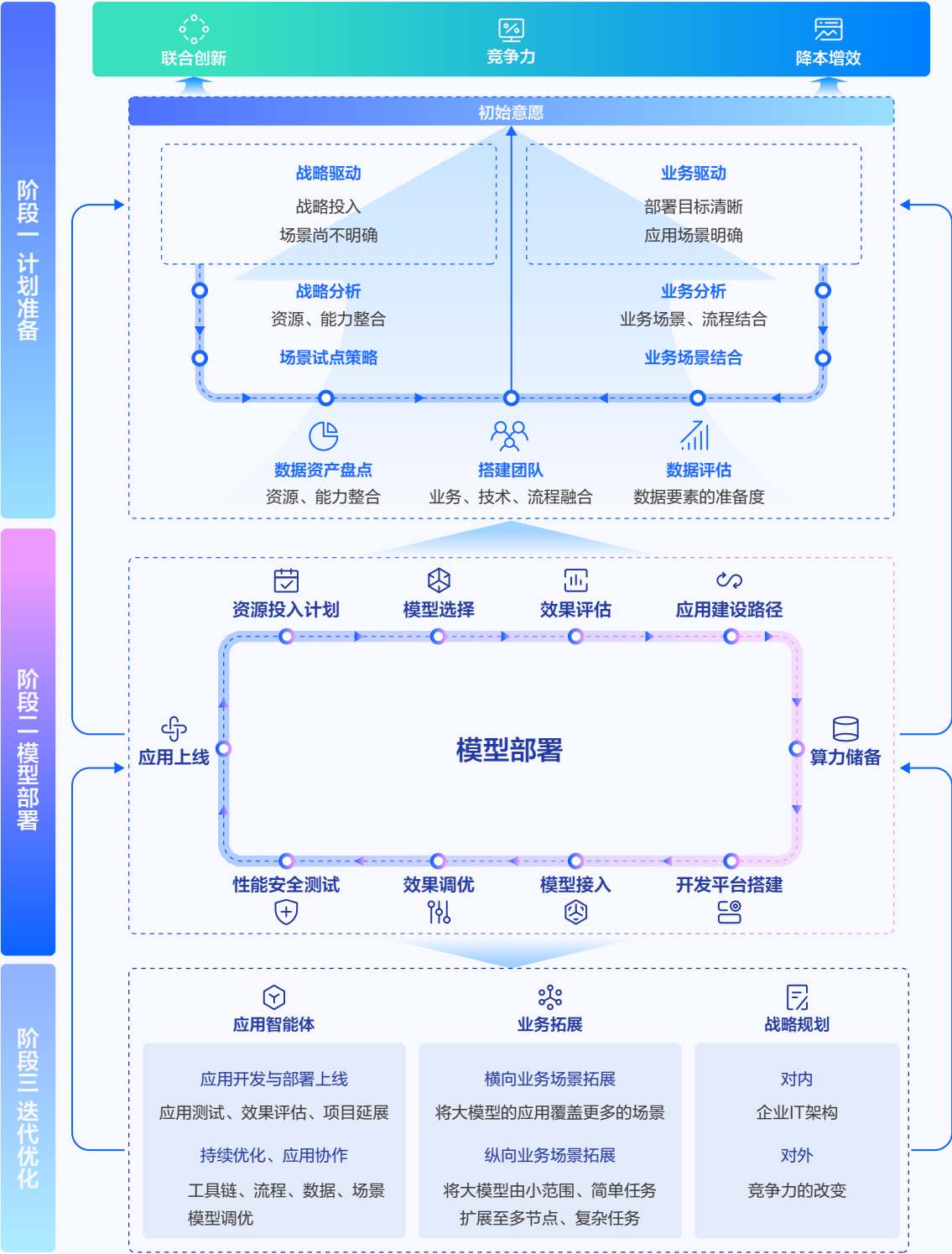
数据来源：IDC大模型企业落地进展调研，N=100，2024年8月

03

加快AI转型

构建全方位的大模型业务落地能力

在当前大模型技术迅猛发展的背景下，企业适时、科学地引入大模型对其长远发展至关重要。无论企业目标如何，处于哪个发展阶段，把握大模型机遇都是必要的。企业管理者需从战略高度规划大模型的落地，这不仅是一次技术革新，也是企业管理理念的升级，关乎企业未来市场竞争力和可持续的经营发展。



来源：IDC，2024

3.1 大模型业务落地能力建设三阶段

提前做好计划准备、扎实推进模型部署、持续优化迭代三大阶段对企业建设全方位的大模型业务落地能力、实现AI转型至关重要。具体来看：

● 阶段一：计划准备阶段

首先明确初始意愿。企业高管在了解AI大模型解决方案时切忌盲目跟从技术潮流，而是需要从企业整体的层面明确落地AI大模型的原始驱动力，分析与AI结合的可能性、行动路线并做好预期收益的管理：

- ◆ **战略驱动：**即企业可能没有明确的AI+场景落地计划，但考虑企业的发展环境、行业趋势、竞争态势，有必要将大模型落地作为战略性规划的一环。从战略布局出发，分析内外部发展环境，盘点企业内部资源和能力，制定场景试点策略，制定筛选场景的标准和路径。
- ◆ **业务驱动：**企业已经有了非常明确的业务场景，迫切地需要引入大模型的能力来进行应用场景的改造，以实现降本增效、提升效率、优化体验等目标；执行切实可行的业务分析策略，了解场景现状、预期目标、数据的就绪度、梳理业务流程；深入探索业务场景结合点和需求程度，将大模型解决方案与特定的业务环境、业务流程或业务需求紧密联系起来。

其次剖析目标任务，了解就绪度。在确定需求和目标后，从资源盘点、流程梳理、能力自查三个方面来考虑并拆解目标，充分参考了解大模型落地所需的企业资源投入，制定详细周密的时间和建设计划：

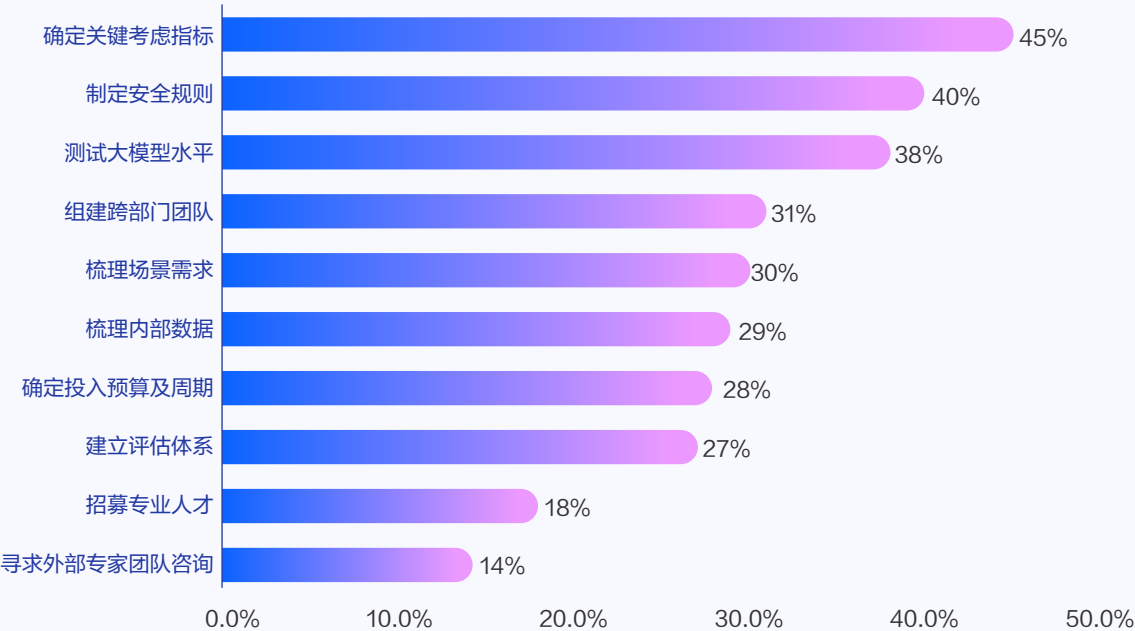
- ◆ **一是资源盘点，**在最短时间内梳理内部资源，如数据资源、算力资源，并分析资源整合的可能性和优先级，确定可为AI大模型的落地提供完备的后续支撑；了解是否数据充足、治理水平如何、是否有大量未清洗数据、是否需要外部购买数据、有无数据获取渠道、算力资源如何购买以及计费方式；

❖ **二是流程梳理**，AI的搭建、应用并不是点对点的数据传输，在模型调优、模型管理、查询调用、内容生成的过程中，均涉及企业、部门、团队内外的多个组织方和参与方，只有明晰AI执行的过程方式，才可以更好地进行开发建设和运营管理；

❖ **三是能力自查**，大模型的建设需要大量专业人才和业务人才共同合作，因此企业应重点考虑评估人才资源基础和经验知识储备，制定研发创新的人才投入形式，是否需要低代码、轻量化的开发平台，以及是否需要外部合作来进行定制化开发。

最后，搭建跨部门团队确保大模型落地的推进。大模型是对企业人员、市场、产品、业务场景的赋能，不仅仅是“底座”“中台”等概念；其需求的提出和最终实现均来源于使用方，而非仅靠数据部门、科技部门的采购开发就可以落地。企业应充分获得高级管理者的支持，搭建跨部门协作团队，制定人、财、物资源的管理和监管制度，确保资源的即时响应和快速供给，实现资源的有效整合和优化配置，从而推动大模型在企业的深度融合和高效运行。

图7 您认为开发大模型时前期准备工作的重要性



数据来源：IDC大模型企业落地进展调研，N=100，2024年8月

● 阶段二：模型部署阶段

模型部署是最核心的落地环节，将直接影响企业应用AI大模型的实现效果，因此需要按照科学的路径进行开发建设。此过程包括如下重点步骤：

- ◆ **资源投入计划：**明确开发AI大模型落地过程中所需要的专业人才、资金投入、数据准备和管理，以及业务领域所需的知识经验；
- ◆ **模型选择：**结合具体的需求从模型精度、效果、参数规模、场景匹配度来筛选适合自身的模型；
- ◆ **效果预测：**预测AI大模型上线后的实际效果，包括准确率、用户体验、响应时间、TPM、RPM（TPM：Tokens Per Minute每分钟Tokens数量；RPM：Requests Per Minute每分钟请求数）；
- ◆ **应用建设路径：**规划AI大模型的建设周期、落地具体场景、覆盖使用广度、垂直场景应用深度；
- ◆ **算力准备：**考虑推理与计算成本和使用方式，以及购买的规模、服务的选择；
- ◆ **开发平台搭建：**为AI大模型的应用率先搭建好底层开发平台，覆盖LLMOps全流程，接入丰富的Agent工具链、API管理服务和数据引擎平台；可以重点考虑提供低代码、全流程大模型开发能力的厂商；
- ◆ **模型接入方式：**即考虑模型接入和部署应用的方式，包括云端接入、API管理、线下部署；面向简易快速上线以及企业级、部门级应用定制化开发，挑选可以提供多产品选择的服务；
- ◆ **效果调优：**即提升模型效果使其更满足自身业务需求，主要可通过Prompt工程、RAG增强检索生成、有监督调优、基于人类反馈优化等；
- ◆ **性能安全测试：**即在模型上线前确保性能稳定和安全可信；

- ◆ **应用上线：**即考虑应用上线范围和管理，包括跨平台、系统部署、跨数据源生成、多版本管理。

● 阶段三：迭代优化阶段

AI大模型在应用测试和效果评估后可正式上线，但这并不是整个流程的结束，企业还应重点考虑模型的可成长性。随着使用量和使用方向的增多，应进一步开展调优训练和知识管理，达到“越用越好”的效果。因此，模型的部署完成并非终点，企业仍需持续关注并致力于大模型在三个关键层面的迭代与优化：智能体应用效能的提升、跨业务领域的拓展应用，以及对企业整体战略规划的对内和对外的影响：

- ◆ **智能体优化，**即针对基于大模型的智能体的能力和水平的持续提升。

对于开发上线的AI智能体应用，企业可以持续提升模型的精度、改善用户体验，例如通过AB测试评估调优后的效果，并针对在应用服务过程中遇到的工具缺失、流程混乱、数据管理困难、场景需求升级等问题进行优化完善，以保证其发挥最大的价值和效果。

- ◆ **业务扩展，**即将已成功的基于大模型的应用扩展至相关或更复杂的业务领域以挖掘更多的价值。

- **横向业务拓展：**在实际应用AI大模型中，企业往往会首先选择1-2个易于落地、急需改善的场景进行项目试点，当完成项目上线并评估可持续投资后，可结合积累的技术、开发、流程、管理经验，共享同一套平台、工具和资源，进一步将模型应用到其他场景中，服务不同的人员，打通、管理更多的数据和实践经验。
- **纵向场景延伸：**知识管理、对话问答、查询分析是AI大模型最先落地的场景，随着模型的升级和应用需求的加深，企业应考虑将小范围、简单任务扩展至多层次、复杂任务，打造可以同时满足多个任务的智能体，探索、扩大大模型的服务深度和边界。

◆ **战略规划**，即从企业整体的视角评估大模型部署落地后对企业整体竞争力的改变，同时关注其对IT基础设施的影响。

- **持续关注企业对外的差异化竞争优势**：当AI大模型应用上线后，企业应重点考虑竞争环境，了解其他企业与AI的结合深度和布局计划，评估、判断自身企业在当前和未来的竞争力和优势对比，从而制定、明确AI大模型下一轮的投资和开发计划，以促进应用和产品的迭代，形成长期差异化竞争优势。
- **对内关注企业整体数字化基础设施的能力**：大模型上线后，企业IT基础设施面临一系列的变化，包括计算资源需求、服务器负载、存储空间、数据管理、安全性和隐私保护等。因而企业需要关注现有IT系统的升级和优化，以确保大模型的高效运行。

3.2 破除落地大模型的思维误区

对大模型落地这一新的企业课题，在准备、实施和优化三大阶段中，企业容易产生认知误区。这些误区如同迷雾，笼罩在企业前进的道路上。为帮助犹豫观望的企业梳理发展路径，避免在大模型落地过程中走不必要的弯路，我们归纳了先行者在探索过程中常见的思维陷阱：

技术指标	VS	商业应用	把技术指标（参数量、数据训练规模）等当作商业应用成功与否的指标，而忽略了大模型的实际效果。
成本	VS	需负担的成本	大模型的训练确实需要大量的计算资源和资金，但并不意味着企业用户需要承担大模型全部的费用和成本投入。目前，大模型技术服务商已经通过预训练、模型微调、算法优化等技术降本的方式，使企业以相对合理的成本享受大模型的技术优势。企业用户可以直接选择成熟的预训练模型，不需要从0开始训练。
开发流程	VS	全生命周期的开发工具	开发大模型不是需要用户独立建立开发流程、实现端到端开发，技术供应商能提供全周期的开发平台和流程推荐，并提供插件、测试、数据管理等“开箱即用”和低代码的工具组件，帮助企业更加敏捷、快速地应用AI。
短期产出	VS	长期投入	关注短期内大模型带来的价值产出，而忽略长效投入下的价值收益。模型上线不等于AI应用落地的结束，还需要长期的投入。仅关注短期产出，可能将忽略数据飞轮效应的长远收益。
安全隐私	VS	大模型商业价值	担心引入大模型会伴随着产生数据隐私和数据安全的风险，而忽略了大模型落地对现有数据资产和数据价值发挥的推动作用。实际上，大模型技术供应商围绕大模型开发和使用提供了全栈安全可信解决方案，可保证全流程的模型和数据安全。

来源：IDC，2024

04

头雁效应

大模型深入众多应用场景

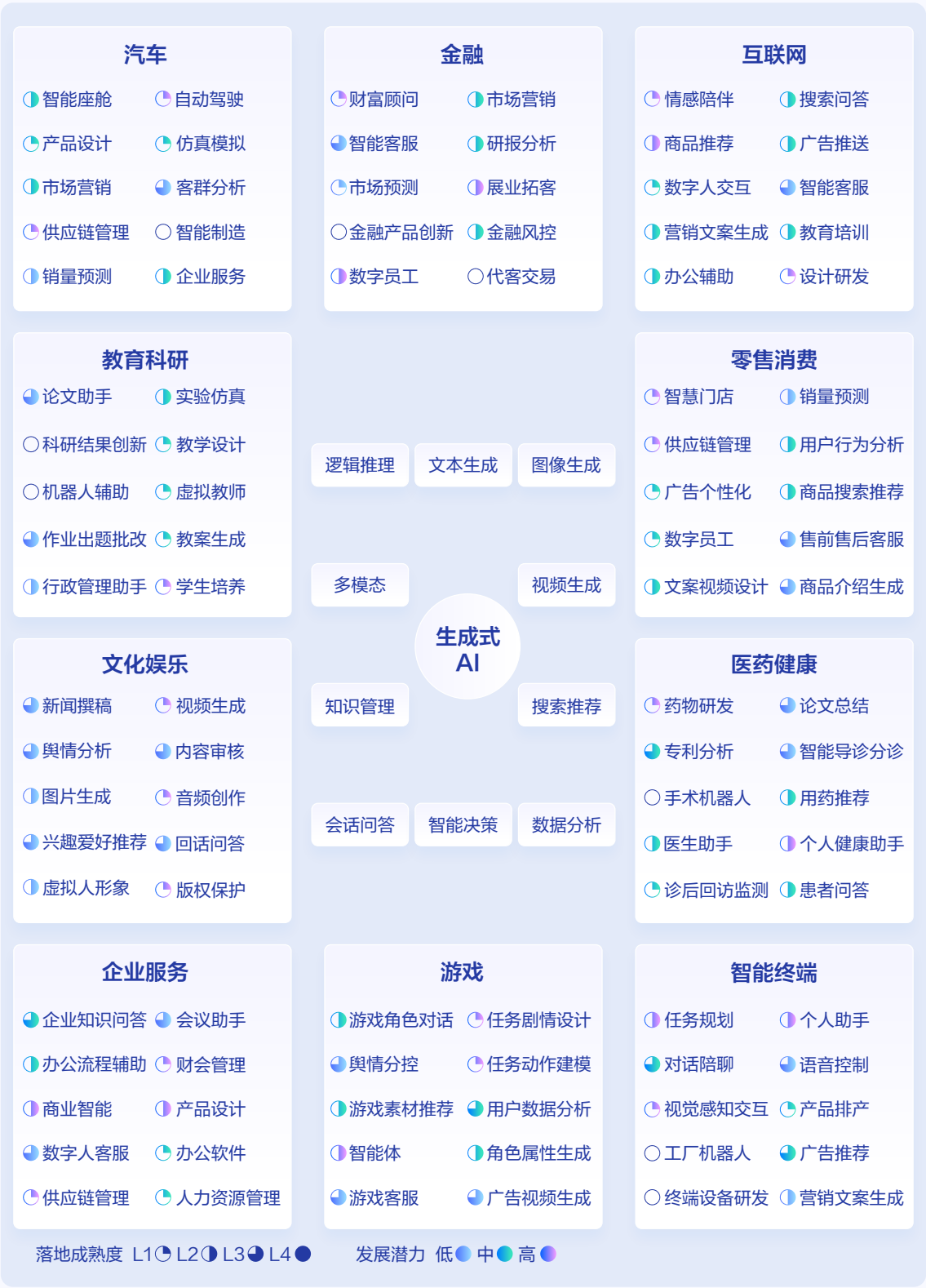
随着大模型在众多企业中的落地实践不断取得成功，其应用场景正逐步拓宽。这些企业在不同业务场景中的积极探索和验证，使得大模型的能力得到了广泛认可。从最初的业务试点到如今的深度融入，大模型的应用范围不断扩展，涵盖了生产、管理、营销等众多领域。

4.1 大模型应用场景不断扩宽，应用日渐成熟

IDC研究总结了十大核心能力和十大应用领域，每个应用领域下梳理十个细分场景，梳理了AI大模型落地应用场景全景图，并分析其落地成熟度和发展潜力，以期为市场和企业提供参考，帮助企业选择最合适落地的场景和服务。

- ◆ **十大核心能力：**包括文本生成、图像生成、视频生成、推荐搜索、数据分析、智能决策、会话问答、知识管理、多模态、逻辑推理；
- ◆ **十大应用领域：**包括金融、互联网、零售消费、医药健康、智能终端、游戏、企业服务、文化娱乐、教育科研、汽车；

图8 AI大模型落地应用场景全景图



来源：IDC，2024



金融

当前AI大模型落地场景重点是在知识管理以及智能服务相关领域，例如智能客服、内部知识管理、研报分析、市场营销。从部署方式上看，也更倾向于私有化部署。但AI大模型依然存在生成内容不准确、逻辑判断差等技术不足，且其核心业务是为客户提供理财和财富管理服务，数据安全保密性要求高。



互联网

互联网对AI大模型的接受程度最高，企业希望实现运营成本的降低和产品设计创新，如搜索问答、广告推送、商品推荐、营销文案生成、教育培训、办公辅助等已开始规模化落地应用。此外，情感陪伴也表现出了更大的用户潜力。



零售消费

图文生成能力已广泛应用于商品介绍生成、售前售后客服场景，大幅降低人力成本并保证内容准确性。此外，销量分析和预测、用户行为分析等数据分析功能，也通过BI升级的形式来服务更多人员。



医药健康

当前AI大模型主要用于辅助场景，如药物研发中靶点发现、分子合成，以及智能问诊助手、导诊分诊等。另外，养老场景中也有较大的应用潜力，例如，利用AI实现视频安全监控、文化娱乐、情感陪伴、出行辅助等功能，可以为用户提供更智能更有温度的服务。



智能终端

以手机、电脑、智能家居为主的终端设备在积极融合AI能力，开发智能对话、任务控制等基础能力。面向未来，手机有着明显的优势——可移动，APP功能丰富，应用生态底座全面，SoC芯片持续升级且具备摄像头拍摄能力，可作为大模型的个性化开发底座来进行问答、执行、控制。笔记本电脑则更多围绕系统工具的升级来实现辅助办公、代码开发、视觉感知和控制等功能。智能家居在对话理解和语音控制的过程中，可以给人更自然的体验，而不仅是关键词识别和触发。另外，同时具备多模态感知和可移动能力的智能眼镜、AR设备也是大模型落地的热点领域。



游戏

优质游戏的开发成本巨大，成功后也会带来明显回报。对于优质内容和游戏玩法的设计、数值策划、闯关类型、人机匹配等，大模型尚不能完全替代；如当前大部分设计策划仍由人来设计测试，不需要复杂的算法介入，人机匹配则大多是机器学习领域的强化学习和模仿学习。目前AI大模型多用于2D图像生成、语音生成、智能客服、违规词监测、3D人物和动作生成，以及智能客服、办公辅助等场景。



企业服务

这一赛道的企业已经将“AI+”列为2024年重点目标，投资动力明显，更新路径也从基础的内容生成、数据分析、知识问答，延伸至个人助理、跨模态生成、跨数据检索、多轮对话、长文生成等能力。



文化娱乐

新闻生成、图片素材生成、音乐生成等单模态能力已在现实生活中广泛应用。随着GPT-4o的发布，视频生成领域、图片生成视频、文字生成视频、视频自动剪辑生成等方向，也开始在互娱领域快速落地。



教育科研

头部教育机构拥有足够的题库积累，覆盖幼儿教育、学前教育、小学教育、初中教育、高中教育、职业培训、大学和考公考研全年龄段，依靠AI生成能力来提供试题生成、批卷打分、口语分析、虚拟教师等功能。大模型已在英语培训市场展现出巨大的商业机会。



汽车

汽车与AI大模型的结合主要分成四个模块，包括智能座舱、企业服务、自动驾驶和智能制造。目前多家车企已上线智能座舱和企业服务，如车内控制、导航娱乐、知识管理、市场营销等，通过语音对话、意图识别、RAG检索、指令生成、Function Call、API调用、插件调用、内容整理和生成这一流程完成驾舱控制。

4.2 众多行业企业深入大模型落地实践

随着大模型应用场景的不断拓展和应用的日益成熟，我们看到了越来越多的大模型成功落地案例。这些案例如同路标，为众多企业指明了落地的方向，提供了实实在在的参考。企业可以从中了解到先进的技术应用经验，感受到大模型带来的实际效益，从而更加坚定地在自身业务中推进大模型的应用。

赛力斯：携手豆包大模型提升用户满意度

● 企业介绍

赛力斯始创于1986年，是一家以新能源汽车为核心业务的技术科技型汽车企业，旗下主要产品包括AITO问界系列高端智慧新能源汽车、蓝电新能源汽车、瑞驰电动商用车等，秉承着“推动汽车能源变革，创享智慧移动生活”的使命，专注技术自研，在三电技术、增程技术、电子电气架构和超级电驱智能技术平台方面处于行业领先地位。

● 需求驱动力

主要需求：提升用户反馈闭环效率，优化业务部门决策支持

- ▶ 提升用户反馈闭环效率：通过优化从收集、分析到响应用户反馈的整个流程，加快用户问题解决速度，提高用户满意度；
- ▶ 优化业务部门决策支持：基于对用户反馈的及时处理与准确判断，提升相关业务部门的业务决策效率。

● 赛力斯与火山引擎合作重点

赛力斯工作重点：

- ▶ 厘清用户反馈处理流程中各项工作环节，借助豆包大模型建设用户之声管理平台（VOC：Voice of Customers）；
- ▶ 梳理公私域平台有关用户反馈的数据源头，打通企业工单系统，同时对接内部协同办公软件，如飞书、钉钉、企微等，依托火山引擎提供的数据采集与分析能力，为业务部门提供第一手高质量用户反馈信息。

火山引擎解决方案：

火山引擎基于豆包大模型的理解、分类与总结能力，携手赛力斯共创用户之声管理平台，实现了用户反馈分析、处理工作由人工承接转为模型承接，达成用户反馈闭环的智能化升级。

- ▶ 情绪正负向判定：豆包大模型可对反馈内容做上下文理解，输出“正向”、“中立”、“负向”标签，避免人工主观漏判/误判，该类标签可用于报表聚类展示；

- ◆ 内容标签提取：在火山引擎600+“开箱即用”的汽车行业标签支持下，豆包大模型可基于客户预设的标签体系对内容打标，包含品牌、车系、车型、维度、观点等，输出结果定位对应业务部门；
- ◆ 内容观点总结：豆包大模型能提取用户反馈内容中的主要观点，例如其可将一篇3,000字的反馈精简至200-300字，并结构化地提炼核心观点，从而帮助运营大幅提升对于内容的理解处理效率；
- ◆ 实时数据采集与分析：火山引擎数据产品助力用户之声管理平台，可实时采集公私域数据，包含资讯、评价、投诉、建议等多个维度的图文、视频信息，并支持品牌、车系、车型的下钻分析，维度涵盖用户声量监测、来源渠道分布等。

● 落地效果

- ◆ 在豆包大模型的帮助下，VOC管理平台有效提升了赛力斯的用户反馈闭环效率，缩短了处理时长，实现了用户满意度的提升；
- ◆ 基于火山引擎提供的数据分析能力，赛力斯的风险监控水准得以进一步提升，并推动各个业务部门基于用户反馈及时制定决策与实施，从而优化企业运营。

上汽乘用车：有效提升“用户之声”的反馈处理效率

● 企业介绍

上海汽车集团股份有限公司（简称“上汽集团”）作为国内规模领先的汽车上市公司，努力把握产业发展趋势，加快创新转型，正在从传统的制造型企业，转变为面向消费者提供移动出行服务与产品的综合供应商。

● 需求驱动力

主要需求：高效处理海量数据，精准分析用户反馈，快速响应市场变化

- ◆ 高效处理海量数据：需要快速处理和分析大量用户的反馈信息，以确保不会错失任何重要的用户声音；
- ◆ 精准分析用户反馈：需要能够准确识别和理解用户反馈中的关键点和情感倾向，以便更好地满足用户需求；
- ◆ 快速响应市场变化：需要通过高效的数据处理和精准的反馈分析，获得及时且有价值的洞察，以支持快速的业务决策，从而在竞争激烈的市场中保持领先。

● 上汽乘用车与火山引擎合作重点

上汽乘用车工作重点：

- ◆ 整合来自各类垂直媒体、论坛、微博、企业APP等渠道的用户反馈数据，为豆包大模型的分析理解提供素材；
- ◆ 根据豆包大模型对各个来源用户反馈信息的分析结果，进行下一步的产品改进、服务优化和市场策略调整。

火山引擎解决方案：

- ◆ 针对社交媒体、论坛、新闻网站等公域平台上大量的用户评论与反馈，豆包大模型提供优秀的自然语言理解能力，可以快速筛选、分类和分析这些数据，提取出用户的主要关注点和情感倾向；

- ◆ 豆包大模型在理解能力方面不断升级，可通过对用户反馈的深入分析，更准确地识别出产品和服务中的不足之处，还可针对长篇文章评论，实现文章深层次表达主题的识别，精炼提取用户的反馈和诉求；
- ◆ 豆包大模型以更强模型优势、丰富的模型能力，助力上汽乘用车全面提升服务反馈、热点事件跟踪和质量改进等。

● 落地效果

- ◆ 引入豆包大模型技术之后，上汽乘用车在处理来自公域和私域的“用户之声”时，既实现了对大量信息的快速处理，同时又做到了对每一条内容的精确理解；
- ◆ 在豆包大模型强大能力的帮助下，上汽乘用车在服务反馈、热点事件跟踪和质量改进等领域取得显著成效，为用户带来了更加优质的体验和服务。

海尔消金：豆包大模型为信贷资产管理提质增效

● 企业介绍

海尔消费金融成立于2014年，是由海尔集团发起设立的持牌金融机构，经银保监会批准设立的全国性持牌消费金融公司，旗下拥有“海尔消费金融”“够花”两款APP,通过科技金融为用户提供消费信贷服务。

● 需求驱动力

- ▶ 场景智能化升级：大模型可以应用到金融业的贷前、贷中、贷后等各个环节，通过多模态数据识别客户信贷风险、客服小结标准化、降低客诉意愿、提升客户满意度等方面，提升效率。
- ▶ 提高员工效率：提升客服坐席效率、减少人工录入时间、降低错误率，以提高效率降低成本。当前坐席摘录人均每天200余次、记录耗时2+小时，解决描述不准确、表达口语化等问题。
- ▶ 改善服务体验：在客户服务的全生命周期中，积极引入大模型智能化支持手段。借助音色复刻、RTC（Real-time Communication）等先进技术，全力打造智能化服务体系，显著提升客户的体验感与满意度。

● 海尔消金与火山引擎合作重点

海尔消金工作重点：

- ▶ 整理贷款知识、客户信息、员工手册等内容，做好数据安全管控后，与豆包大模型连接，提升关键词查询和模糊匹配生成的准确率；
- ▶ 组织员工培训，学习大模型使用方式和建议，建立生成Prompt模板。通过扣子搭建适合不同业务人员的智能体。

火山引擎解决方案：

- ▶ 基于字节跳动豆包大模型精调的大模型，能够满足海尔消金90%以上的智能化场景需求，包括风控、客户服务、贷后资产管理、意愿度识别等；
- ▶ 火山引擎提供了火山方舟大模型服务平台、扣子AI原生应用开发服务平台、算力管理平台等全栈式大模型工具；

- 提供完善的模型应用数据治理体系，提供丰富多样的独享技术方案，严格遵循安全合规的要求，确保安全、合理地使用大模型。

落地效果

- 在资产管理过程中，通过与火山引擎的合作，引入豆包大模型能力，在初始应用阶段，效果提升达到15-20%；
- 海尔消费金融坐席助手摘录过程中可达到超过95%的准确率。每天节约坐席专员1-3小时，同时提升智能问答系统的准确率。

后续支持

- 海尔消金与火山引擎共同成立金融大模型联合创新实验室，致力于构建消费金融垂直大模型，并在精准营销、智能客服、贷后管理、风控等业务场景中持续探索大模型的创新应用，加速海尔消金的智能化转型进程；
- 未来将联合研发智能客服机器人，利用多模态智能服务、自动质检、情感分析等，自动摘录，提升客户整体的服务效率和体验。

海尔消金CIO 梁树峰

“在引入大模型后，现在实际的落地效果非常显著，最典型的要数海尔消金的贷后管理工作。此外在客服领域里边，也对客户意图进行识别，可以更精准地理解客户的意图，还便于事后对客户进行管理。”

中国飞鹤：利用AI实现企业全面的AI转型升级

● 企业介绍

作为全国乳品行业龙头企业，中国飞鹤1962年创立于赵光农场，迄今已有60多年历史，是中国最早的奶粉企业之一。根据第三方调研机构全新数据，2019年-2023年度，飞鹤连续5年销量位居全国首位。同时，飞鹤奶粉连续两年全球销量第一，旗下星飞帆系列已连续3年成为全球第一大单品。截至2023年1月，中国飞鹤在婴幼儿奶粉中市占率达21.5%，稳居市场第一。

● 需求驱动力

- ◆ 全面智能化：将AI技术深度整合到企业的数字化转型中，从数字人、智能推荐、内容分析等试点项目逐步扩展至全面的能力建设和最终的全局AI应用实施；
- ◆ 优化用户体验：销售员工水平参差不齐，业务效率和客户交互提升潜力巨大，需要打造统一的AI助手帮助解答用户问题，优化购物体验。

● 飞鹤与火山引擎合作重点

飞鹤工作重点：

- ◆ 重塑基础设施能力，携手火山引擎搭建先进的云架构、业务中台和数据中台，确保数据处理的高效性和业务流程的灵活性，支撑上层AI应用运行；
- ◆ 选择AI能力中台核心组成部分，面向实际业务场景和需求，重点建设大模型能力、数字人、数字孪生、智能推荐、音视频、VR/AR等；
- ◆ 将应用场景分成用户运营触达、渠道销售和企业管理三大类，每个类别下重点建设具体应用场景，如追溯回答、消费者咨询、优选客服等，以满足不同环节的需求。

火山引擎解决方案：

- ◆ 基于飞鹤信息化建设“3+3+2”的战略蓝图，飞鹤聚焦搭建集成基础设施层、能力层、接入层、场景层全覆盖的AI能力中台。通过明晰业务需求、愿景及资源状况，火山引擎为飞鹤AI能力中台做整体规划，制定分步实施策略，促使中台成功落地；

- ▶ 围绕HiAgent平台，火山引擎针对企业级市场进行深度定制，提供成熟解决方案，通过段位划分策略，帮助企业员工逐步掌握HiAgent使用，同时通过原厂轻咨询、内置最佳实践和应用效果评测，确保达到生产级标准。

● 落地效果

- ▶ HiAgent平台发挥了核心作用，将AI技术融入各个业务流程中，从而提升整体运营效率和市场竞争力，为数字化转型提供了坚实基础；
- ▶ 某智能问答项目能够处理广泛的消费者咨询，实现100%的问答响应率，同时保持了超过95%的高准确率。

飞鹤副总裁 冯海龙

“AI大模型已在消费者服务、线下活动检核、企业办公、导购培训、数据分析和智慧农牧等多个业务场景中广泛应用，为飞鹤的创新发展注入新动力。”

中手游：为《仙剑世界》打造丰富的AI NPC生态

● 企业介绍

中手游是领先的全球化IP游戏运营商，以IP为核心，通过自主研发和联合研发，为全球玩家提供精品IP游戏。围绕自有IP《仙剑奇侠传》，持续为粉丝创造精品内容和互动体验，打造世界级IP，并将《仙剑世界》打造成全球首个国风仙侠虚拟世界。

● 需求驱动力

主要需求：为《仙剑世界》打造AI NPC，落地RAG技术，保障游戏服务稳定

- ▶ 为《仙剑世界》打造AI NPC：基于大模型搭建丰富的AI玩法，打造具有长期记忆和成长能力的AI NPC；
- ▶ 落地RAG技术：通过RAG技术进一步提升模型的推理效果，优化AI游戏内容的表现；
- ▶ 保障游戏服务稳定：借助模型及系统稳定强大的推理QPS保障能力，保障游戏服务稳定运行。

● 中手游与火山引擎合作重点

中手游工作重点：

- ▶ 针对《仙剑世界》游戏中可落地大模型的场景进行了深入探索，如意图识别、情绪识别；
- ▶ 对基座大模型、RAG解决方案中的向量数据库等关键组件，进行了大量评估，并搭建开发框架。

火山引擎解决方案：

- ▶ 提供豆包·角色扮演模型，助力中手游打造行为独立、可对玩家的行为做出个性化的反应游戏氛围NPC；
- ▶ 基于豆包大模型能力帮助中手游打造玩家专属剑灵“圆满”，其可为玩家提供问题解答、任务引导、功能唤起、智能传送、战斗BUFF、时辰播报等功能；
- ▶ 通过召回能力优秀的豆包·向量化模型、P90延时仅18.2ms的向量数据库VikingDB等组件，搭建一站式RAG解决方案，为中手游建设游戏RAG知识库提供技术支撑；

- ▶ 提供依托火山引擎充沛公有云GPU资源池打造，底层算力充足，且支持分钟级完成千卡扩容的大模型服务平台火山方舟。

● 落地效果

- ▶ 基于豆包大模型打造的AI NPC生态在《仙剑世界》游戏中被全面应用，塑造了一个对于玩家而言陪伴感更强的游戏世界；
- ▶ 火山引擎RAG解决方案，帮助游戏精灵“圆满”具备了更智能的辅助功能，例如其可以根据玩家的行为数据和游戏内的进程，提供最适合玩家角色当前状态的支持，如任务推荐、玩法说明等，从而减少玩家在游戏时的阻碍感，提升游戏体验；
- ▶ 火山方舟平台充分满足了中手游对于游戏的高RPM/TPM要求，通过按需调整配额，保障游戏在超大流量下的服务稳定性，同时提供全面的算法与工程优化能力，支撑豆包大模型在更加细分场景中的快速落地，并打造极致性价比，助力中手游降低模型使用时的技术、资金与人力成本。

中手游技术中心总监

“目前大模型已经在中手游项目开发流程中的多个环节投入应用。火山引擎这样一位高效、值得信赖的云服务商伙伴，在这中间扮演着至关重要的角色。”

浙江大学信息技术中心：落地全方位大模型应用体系，构造智能化教育环境

● 浙江大学介绍

浙江大学是一所特色鲜明、在海内外有较大影响的综合型、研究型、创新型大学，设有7个学部、40个专业学院（系）、1个工程师学院、2个中外合作办学学院、7家直属附属医院。截至2023年底，学校有全日制学生67656人、国际学生5514人、教职工9557人。2022年，浙江大学入选第二轮“双一流”建设高校。

● 需求驱动与创新驱动

主要需求：融合大模型能力，获得跨越式发展

- 融合模型能力：大模型时代技术背景下，校园教学、科研、管理、服务等系统，亟待融入大模型能力，使得各系统更加智能化，各系统组成的校园系统更加智能化；
- 融合模型能力：在政策背景下，全校师生亟待平台支持，以实践人工智能的学习；
- 获得跨越式发展：浙大西湖之光算力联盟已经建成了良好的基础设施，由火山引擎大模型方案构建的AI应用，将继续推动自身的发展和基础设施的发展。

● 浙江大学信息技术中心与火山引擎合作重点

浙江大学工作重点：

- 围绕教学、科研、教务、校园服务、本地生活等需求，建设体系化解决方案；
- 解决场景多、多方协作、时间紧、效果要求高的全方位挑战。

火山引擎解决方案：

火山引擎HiAgent平台为建设“浙大先生”门户及AI科学家等系列智能体场景应用，提供了四大能力支持：

- 多模型接入和多维数据整合：实施多模型接入策略，根据不同场景应用需求精选最合适的模型，以多维数据整合能力有效处理不同数据库的多样化业务数据，为各场景应用提供可靠的数据支持；

- ◆ 集成化工作空间协同开发：高效构建并利用集成化工作空间，实现多供应商之间无缝的数据流转与深度协同开发，减少“数据孤岛”信息摩擦，确保整个开发链路每个环节都能基于最新、最准确的数据决策行动，提升应用开发的迭代效率和质量；
- ◆ 灵活的发布与集成机制：通过灵活地发布和集成Chatbot、Copilot、Agent多种形式的AI应用，实现“浙大先生”大模型应用体系的多元化，帮助师生按需自动化处理日常任务；
- ◆ 模型应用自动化调测优化：提供效果评测、应用Tracing、Debugging等能力，支持工作流与智能体的快速调试，支持通过评测集、评测任务方式自动化验证应用效果与问题发现，加速应用的迭代优化，确保应用在实际场景中的稳定性和可靠性。

● 落地效果

- ◆ 仅花7天时间，就实现了“浙大先生”大模型应用体系的完整、高效构建，让AI应用走进课堂、校园、实验室等场景，为超过6万名在校师生等打造了更智能化的教学教务、科研创新、校园生活等全新体验；
- ◆ 在大模型能力支持下，“浙大先生”应用体系为用户提供高度定制化的AI服务与支持，全面促进浙江大学教学、教务、科研智能化升级与效率提升：
 - 门户：无缝对接浙大统一身份认证系统，师生可凭校园账号轻松登录并享受相关服务
 - AI科学家：提供全学科数据库、知识库、通用AI科研助手等工具，快速链接算力平台、模型库、算法库，大幅提升科研效率
 - 课堂智能问答：深度融合大模型与向量数据库，精准处理教案、考试要求等课堂资料，即时解答学生疑问、提升学习效率与质量
 - 教务咨询：以智能交互提供专业教育咨询，强化政策文件检索与理解，助力学生高效掌握教务信息
 - 百事通助手：整合超600项网上服务事项，覆盖学习、生活、行政等多个场景，为师生提供全方位办事指引
 - 本地生活助手：精准捕捉意图，为师生筛选、推荐贴合需求的餐饮，并提供页面地址及推荐理由，满足多样化用餐生活需求
 - 慧学外语：提供丰富多样、高效使用的课程信息与学习对话体验，助力外语学习

苏泊尔：构建AI驱动的“居家食养健康”物联场景生态

● 企业介绍

苏泊尔成立于1994年，2004年在深交所上市，是中国炊具行业首家上市公司。苏泊尔拥有6大研发制造基地，分布在杭州、武汉、绍兴、玉环、柯桥和越南，通过持续创新和对品质的始终坚持，在小家电行业保持领先地位。产品涵盖烹饪用具、厨房电器、家居与个人护理电器等多个领域。云饌平台是苏泊尔旗下的智慧生活共创平台，为消费者居家健康烹饪全场景提供了领先的解决方案。

● 需求驱动力

主要需求：提供个性化健康食谱，增强苏泊尔APP使用体验，建设智能炒菜机生态

- ◆ 提供个性化见刊食谱：提升厨房烹饪品类电器的内容生产效率，同时满足消费者对于个性化专属谱的需求；
- ◆ 增强APP体验：提供多场景中食谱智能检索和推荐体验；
- ◆ 建设智能炒菜机生态：提升智能炒菜机生态服务能力，为炒菜机用户提供一日三餐“居家食养健康”主题的配餐服务。

● 苏泊尔与火山引擎合作重点

苏泊尔工作重点：

- ◆ 基于扣子专业版，打造云撰食谱创作、云饌AI寻味、云饌居家食养健康等智能体；
- ◆ 优化以往PaaS-SaaS-APP的软件架构模式，借助“Agent as API”方式实现智能体与苏泊尔APP、苏泊尔IoT产品直连，提升开发效率和用户体验。

火山引擎解决方案：

火山引擎扣子专业版提供了灵活编排搭建AI智能体的能力，加速苏泊尔云饌系列智能体的构建；扣子平台灵活的API支持，令苏泊尔以“Agent as API”的方式应用智能体成为可能：

- ◆ 云撰食谱创作：结合用户绑定的厨电产品及家庭饮食情况，智能体可借助豆包大模型和文生图能力，为用户生成专属的个性化健康食谱；

- ▶ 云饌AI寻味：可在扣子专业版的智能体编排技术支持下，于苏泊尔APP寻味栏目中，实现多场景食谱智能检索和推荐等功能；
- ▶ 云饌居家食养健康：借助扣子平台智能体语音交互OpenAPI、音色克隆、音色合成API等能力，提升苏泊尔智能炒菜机的服务水平，为消费者提供每日三餐“配餐式”的健康美食日历服务。

● 落地效果

- ▶ 云饌食谱创作智能体针对苏泊尔旗下电饭煲、电压力锅、烤箱、空气炸锅等重点品类，月均服务超15万苏泊尔APP会员，好评度逐步提升。
- ▶ 云饌AI寻味智能体日均服务苏泊尔APP会员一日三餐2万多次。

*数据取自苏泊尔云饌平台，2024年10月至今。

苏泊尔AIOT中心总监 张作强

“苏泊尔联合豆包大模型与扣子AI智能体平台，深入居家健康烹饪等多领域，以'Agent as API'直连产品，优化了IoT平台架构，提效研发且用户体验得以优化。凭技术创新构建'居家食养健康'物联生态，赋予烹饪电器人格化体验，给消费者带来智能便捷的健康烹饪新生活。”

海底捞：基于智能化客户评价分析，打造更好餐饮服务

● 企业介绍

海底捞品牌创建于1994年，历经二十多年的发展，海底捞已经成长为国际知名的餐饮企业。截至2022年12月底，海底捞拥有上千家直营餐厅。海底捞多年来历经市场和顾客的检验，成功地打造出信誉度高、融汇各地火锅特色于一体的优质火锅品牌。2018年9月，海底捞于港交所挂牌上市。

● 需求驱动力

主要需求：高效完成客户评价分析

高效完成客户评价分析：海底捞自有及三方平台门店已累计超千万条用户评价，以人工方式对评价进行分类分析成本高、效率低，海底捞希望提效用户评价分析过程，加速沉淀产出总结报告。

● 海底捞与火山引擎合作重点

海底捞工作重点：

- ◆ 建设数据中台，整合顾客餐后评价、投诉工单、第三方餐饮平台点评与评价等反馈；
- ◆ 打造后端环境进行数据治理，过滤无效内容，将有效评价推送至大语言模型。

火山引擎解决方案：

基于豆包大模型，火山引擎为海底捞搭建了用户评价分析模型，通过对用户评价进行情绪、观点的高效智能抽取，形成服务质量、菜品质量洞察。

- ◆ 豆包大模型具有更强性能，可通过语言处理能力，对顾客评价中有关环境、菜品、服务等不同维度内容进行抽取，同时判断顾客评价时正向/负向情绪：
 - 从全局角度，对顾客评论给出“好中差”整体评价结论；
 - 提供进一步精细化分析能力，从服务、产品、卫生、性价比等多个角度判断顾客情绪。
- ◆ 提供不同颗粒度总结报告，以直观、数字可视化形式呈现在海底捞数据中台，方便查阅。

● 落地效果

在火山引擎豆包大模型能力帮助下，海底捞实现了高效、精准的用户评价分析，为门店绩效管理及后续经营管理的持续迭代与优化提供了指导。

海底捞信息科技部产品负责人 杨炫之

“AI在餐饮行业的应用，不能一味地考虑边际效用，也需要结合服务体验，海底捞是以服务立足的企业，在AI方案的选择上，也一直在追求“以人为本，体验第一”的原则。”

更多企业落地案例实践

招商银行落地智能体，形成更全面AI能力

客户需求：

- 加速智能化转型：推动AI创新从高性能基础设施向智能体多元场景拓展；
- 提供更智能化、个性化客户服务：以自然流畅语言交互方式，满足金融行业用户多样化个性需求。

解决方案：

火山引擎为招商银行的智能化创新提供了拥有低门槛智能体构建能力的扣子平台，以丰富的企业级插件、灵活的集成机制、多模型适配能力，加速招商银行AI能力落地应用：

- 以扣子平台为依托，招商银行举办了为期三个月、范围覆盖全行的“大模型应用创新大赛”，加速大模型及智能体技术在金融行业的融合应用，推动更多金融智能体项目涌现；
- 依托扣子平台打造的“掌上生活优惠”及“财富看点”等智能体，为用户提供生活优惠查询以及市场行情分析等一系列服务。

落地效果：

打造的智能助手以更自然的交互方式，以及更智能化、个性化的服务，提升用户满意度及粘性；招商银行内部大模型及智能体技术的普及，为后续更多贴合自身业务需求的AI智能体构建奠定了基础。

领克汽车将AI打造成销售顾问的得力助手

客户需求：

- 产品知识培训：优化内训流程，强化销售顾问知识体系，理解客户需求；
- 增强销售技巧：提升服务技能，增强销售顾问服务、沟通技巧和跟进策略；
- 智能化工具助理：智能化建档，匹配现代消费者购买行为，丰富销售工具，建立健全现代化工具。

解决方案：

火山引擎与领克汽车构建了同SalesCopilot技术深度融合的豆包大模型销售助理，以及实时对练和评级系统、用车知识工具等：

- 豆包大模型销售助理：通过豆包大模型，销售顾问可以获取关于本品和竞品汽车产品的深度知识，以及市场趋势的即时更新，更好地理解客户需求和优化销售策略；
- 实时对练和评级系统：SalesCopilot提供的对练系统能够模拟各种销售场景，让销售顾问在实际接待和跟进客户前，能够充分练习并提升应对各种问题的能力；
- 用车知识工具：销售顾问可以通过这一工具快速访问到关于汽车维护、操作和功能的详细信息，并可以通过客户手册，查阅以视频等方式输出的丰富内容，从而获得准确而详细的产品知识。

落地效果：

更精准地满足客户需求，提供个性化、更高效的服务，助力销售业绩成功达成，打造业务增长新空间；灵活的底层算力支持，使得领克汽车能够以极低成本快速落地大模型，降低端侧推理成本。

想法流基于模型多模态能力保障更强用户互动

客户需求：

- 提升用户互动性：打造严格遵循人设要求，且具备基本常识对话问答能力的AI角色，以主动开启新话题等方式推进聊天进展，增加用户粘性；
- 多场景支持：希望获得对多个细分专业场景中、用户不同交互行为的支持；
- 服务稳定：为保证更好的对话体验，提升用户留存率，需要延时稳定在40-50ms，并保障RPM&TPM请求。

解决方案：

火山引擎为想法流提供了效果更强的豆包大模型，以多模态AI能力，帮助平台为用户提供丰富奇妙的内容体验：

- 以提供PE提示词调优方式协助AI游戏的制作，并将制作指南开放提供给PGC创作者参考；
- 通过豆包大模型语音合成、图片人脸合成等多模态能力，支持多场景中不同交互行为需求；
- 通过MoE架构模型和充足资源保障，满足客户对延时和RPM/TPM的要求。

落地效果：

通过接入豆包大模型，想法流获得了用户每日互动次数1.5-2.5倍、人均互动轮次1.5-3.5倍于其他模型的增长，部分内容场景中人均对话高达150轮以上。同时，在火山引擎千万级TPM资源保障下，平台线上请求成功率达99.95%，token间时延平均控制在40-50ms内。

深维智信借大模型获营销能力、培训效率双提升

客户需求：

- 营销机会洞察：基于语义进行智能化客户画像提取，对销售会话数据进行准实时的AI分析，实现更精准企业营销；
- 员工培训引导：基于大模型能力，为企业培训提供贴近实战的高效对练服务。

解决方案：

火山引擎为深维智信提供了豆包大模型，以模型能力对运营数据分析过程进行提效，同时以模型能力加持、实现更智能的对话模拟，帮助提升客服能力：

- 火山引擎提供豆包多种窗口尺寸模型及角色扮演模型，满足多场景需求；
- 分析销售会话数据，提供基于语义的智能质检配置和分析，识别客户异议、卡点分析、关键矛盾、交易复盘，实现销售过程透视，服务销售管理；
- 通过AI提炼高频问题、真实需求、客户画像，从客户声音出发，科学制定销售策略，借助精准营销提升转化效率。

落地效果：

语义分析的召回率相对于常规质检工具提升55%以上，帮助客户将新人入职培训时间缩短50%，平均成单时长减少22%，平均人效提升31%。

和府捞面以模型能力提效用户评论分析

客户需求：

- 用户评论分析：分析用户评论数据，识别情感倾向并对其进行分类；
- 经营策略调整：基于用户对菜品口味、门店服务质量的评价，调整经营策略。

解决方案：

扣子专业版提供低代码开发环境，使得没有编程基础的业务人员也能轻松完成智能体的开发。和府的业务人员借助扣子产品的强大编排能力和丰富的插件功能，结合豆包大模型的技术支持，实现了更高效、更精准的用户点评分析：

- 通过豆包大模型的强大文本分析能力，对用户评论数据进行情感分析，精准识别正面、负面、中性等情感倾向；
- 基于模型的多层次理解，结合环境、服务、菜品等关键词，对评论进行智能分类，并提取相关标签，综合准确率超过95%。分析结果以JSON格式输出，便于开发团队进行后续数据的处理和应用。

落地效果：

和府捞面打造的“顾客点评分析”智能体，能够替代人工、高效完成顾客点评工作，为门店建立对顾客的高效洞察，并为其在菜品、服务等方面进行策略调整提供了有力依据。

中和农信借助扣子完成抖音生态内容质检

客户需求：

- 将人工抽检变成AI全检：高效全量地提取抖音视频内容；
- 实现自动化质检分析：高效分析内容效果并及时更新调整。

解决方案：

借助扣子专业版强大的编排能力和丰富的插件能力，中和农信在没有耗费研发人力资源的情况下，实现了质检智能体的开发上线，完成了过去单靠人力无法做到的工作。

- 借助扣子单Agent（LLM模式）搭建，并将智能体发布为API，实现了每天都通过API来自动获取员工自媒体账号的视频内容并进行效果分析；
- 智能体主要依靠视频理解插件和豆包大模型的能力进行信息获取和处理等批量、自动化场景，免除了意图识别、工具调用的时间延迟。

落地效果：

扣子提供的低代码搭建环境、丰富的插件和强大的workflow组件，使得中和农信智能体的开发过程中几乎没有耗费研发资源，2人经过3天的开发和调试就达到了满意效果，成功上线。

05

攻克有径

跨越大模型落地技术难题

基于对众多企业应用实践的总结分析和对技术应用的研究，IDC总结了大模型部署落地的技术路径。这套路径融合了实际操作中的宝贵经验，为企业引入和应用大模型提供了清晰的指南，确保了大模型技术落地的高效性和稳定性，助力企业的智能化转型。

5.1 大模型落地部署技术步骤

企业在落地大模型应用时，建议重点考虑可提供全栈一体、低代码、垂直赋能能力的开发平台和服务。这些平台和服务商可以将完整、科学的AI部署步骤联系起来，同时其内置的丰富的经验模板和插件，能使企业快速、高效、准确地搭建落地目标场景并达到预期效果。此外，在面对不同投入周期、资源、场景、环境时应提供多样化的选择空间，允许企业以更灵活、类似插件搭积木的方式来创新AI。



来源：IDC，2024

● 资源投入计划：

- ◆ 确定团队建设路径，整合大模型开发人才、调优人才、数据人才、业务人才，重点考虑是否需要引入外部人才或展开外部合作，保证大模型开发应用各环节的充足人才储备；
- ◆ 保证资金阶段性的持续投入，并制定里程碑，让每项成果都能被看到；
- ◆ 盘点内部数据，制定数据采集、数据清洗和数据治理计划，使其更好服务于AI大模型的训练和推理；
- ◆ 考虑企业内部是否有可以直接学习和引用的模板和经验沉淀，打造差异化壁垒。

● 模型选择：

- ◆ 了解大模型市场提供产品的功能特性和应用效果，面向需求定制，使用内部和外部测试集来测试模型效果，并充分体现评估集对业务指标的考量，选择评测指标达到业务需求的模型；
- ◆ 从模型实际落地的效果、可参考的实践案例和demo来挑选模型；
- ◆ 从训练和推理效率、延迟以及调用成本方面挑选最合适规格的大模型，可参考模型大小、硬件支持水平、服务SLA水平；
- ◆ 重点考虑该模型与企业目标业务场景的匹配度，是否需要调优或经过简单调优即可上线应用来解决问题。

● 效果评估：

- ◆ 在选择模型后，企业应考虑从具体落地的场景和使用效果出发，来预测模型上线后的业务指标达成率、实际体验效果、各功能平均时延、最大吞吐等工程指标；

- ▶ 预估需要调整优化的投入和开发难度，确保使用者拥有最好的AI应用体验，实际效果和响应时间的长短会极大影响用户的感受和留存率。

● 应用建设路径：

- ▶ 明确建设周期，制定详细的时间计划和关键时间节点，做到多线程统一管理；
- ▶ 了解大模型的应用边界，从赋能方向、使用人群等确定大模型的落地场景；
- ▶ 在明确落地场景后，需要进一步考虑大模型的覆盖广度和创新深度，划分大模型可以服务的业务、市场、开发等部门人员和外部客户群体，并考虑如何在现有大模型的基础上进行内容和服务创新。

● 算力准备：

- ▶ 随着大模型应用广度和深度的增加，大模型的覆盖范围和使用频率会规模化提升，推理时算力资源的消耗额度也将同步增加；企业应重点考虑算力购买规模，制定阶段化、梯度上升的算力准备策略；
- ▶ 明确算力资源的使用消耗方式，倾向选择可以实现瞬时响应和扩缩容的服务；
- ▶ 算力有公有云、私有云等多种方式，应根据自身业务体量和实际需求挑选成本最优、服务最合适的方案。

● 开发平台搭建：

- ▶ 搭建LLMOps平台，保证AI大模型的开发建设和管理拥有完整的工具链条，具备全栈平台能力；
- ▶ 利用外部AI厂商成熟的Agent工具实现AI大模型的快速开发和应用，直接使用预置的丰富垂直场景化模板和工具；

- ▶ 随着模型版本、上线应用场景的增多，做好统一API接口的管理；
- ▶ 为AI大模型采集、产生和使用的多模态数据搭建可靠的数据平台底座，实现文本、图像、视频等多样化数据的一体化管理。

● 模型接入：

- ▶ 从场景安全需求、数据隐私保护需求、管理便捷程度、成本等方面来选择模型的接入方式，如云端接入、API管理或线下部署；
- ▶ 重点考虑后续调优的成本以及模型更新的代价；目前来看，云端接入是多数企业的选择，且在后续模型版本更新后也能通过更低的成本实现无缝衔接。

● 效果调优：

- ▶ 可选择Prompt工程、RAG知识库等简单、低代码、无代码方式进行效果调优，需重点考虑AI厂商是否具备丰富的经验沉淀、算法模型和模板预置，如RAG过程中提供成熟的嵌入、检索和重排序模型；
- ▶ 选择SFT、RLHF等复杂方式进行调优；这一过程需要更高水平的开发人员对大模型的架构、参数和训练数据进行调整。

● 性能安全测试：

- ▶ 在模型上线前最关键的步骤是进行性能安全测试，包括性能、安全、风控测试，对内容生成效率、数据内容泄露风险、敏感词等进行测试；例如TPM（Tokens Per Minute每分钟Tokens数量）、RPM（Requests Per Minute每分钟请求数），保证大流量、高并发场景下运行稳定；
- ▶ 重点考虑互联网场景应用更深、服务人数更广的AI厂商，其拥有更多的数据和经验，在测试时水平更高，可以有效避免安全和敏感问题。

● 应用上线：

- ◆ 根据业务场景重点考虑AI大模型的跨平台、跨系统上线，统计应用上线出口，明确是否需要开发额外的用户界面、客户端，还是与已有服务界面融合即可；
- ◆ 做好底层数据的管理，保证模型上线后不同使用群体、不同问题和任务可以在既定的规则和使用范围下应用数据，同时保护用户的数据隐私；
- ◆ 做好版本管理，以应对模型应用的增多、测试和更新。

5.2 精准选模、高效落地、持续挖掘— 落地三要素

在大模型落地的众多技术步骤与细节之中，精准选模、开发平台、伙伴与同行这三个方面显得尤为关键。精准选模确保了所选模型与业务需求的高度匹配，为后续应用打下坚实基础；开发平台则是模型训练与优化的关键场所，直接影响到模型的性能发挥与落地效率；而伙伴与同行则关乎技术支持、资源共享与协同创新，对于模型的成功落地同样至关重要。

● 5.2.1 模型的选择：企业实现AI转型的关键

- ◆ 充分的业务场景论证将保证商业应用的成功。实验室环境下的模型水平与实际业务落地存在一定认知差距，企业需要重点考虑该模型背后是否有更大的使用量、有充分的业务场景论证；更多的模型使用可以保证更快的迭代升级，包括通用基础模型和多场景的模型家族，能带来更好的、更贴合业务场景的、可以带来真正商业价值的AI服务。

- ◆ **多模型家族给企业更多的选择空间和更紧密的匹配度。**面向不同业务场景，企业需要文生文、文生图、声音复刻、语音合成、语音识别、视频生成等多样化、多模态模型，并实现多源数据的同时接收、判断、思考、处理、检索、生成需要的内容，以选择与业务高度适配的大模型服务体系。
- ◆ **模型应具备广泛塑造、即时可用的灵活性。**大模型作为企业创新提效工具，需要具备可塑造性。企业应关注该模型在利用不同方式调优时的难易程度，以及模型厂商是否提供了更简单、自动化的工具模板，将模型不可变的基础部分进行封装，针对适合企业进行调整的架构和参数提供高代码、低代码工具。同时，即时可用、便捷插拔、弹性扩缩容也是实际业务中需关注的因素。

● 5.2.2 一站式大模型服务平台：解决模型部署的复杂难题

- ◆ **选择提供全栈化、自动化、智能化使用体验的大模型服务平台。**AI大模型不同于传统的计算机视觉、自然语言处理等小模型，大参数模型所使用的开发、调优工具更多、更复杂，对平台功能的一体化要求也更高，导致企业很难依赖过去自建的开发流程和平台。平台不仅需要覆盖从场景分析、模型选型、数据集管理、RAG、工具插件、数据评估、效果调优、测试评估、部署上线、监控优化的全开发链路，还需要提供自动化的流程和工具，减少人员重复、不必要的工作负担，让其专注于模型的效果优化。例如，火山方舟在提供多样化且易于使用的调优框架的基础上，准备了丰富的实践案例和落地指南，并围绕RLHF开发了奖励函数更易实现的低成本方案。
- ◆ **保证从POC到上线运营的全生命周期数据和安全。**企业CXO高管需意识到，大模型的安全可信不同于传统的安全概念，涉及数据输入、知识管理、问答处理、记忆检索、输出监控的全链条周期，这些都会影响大模型本身和生成内容的安全和可信。例如，火山引擎的数据加密服务可实现对话过程的全链路加密、纯内存、零日志，加密密钥掌握在客户自己手中。

◆ **考虑AI厂商团队是否能提供专业的技术服务支持和咨询保障。**调研显示，有28%的企业认为内部缺乏AI大模型相关技能和人才，26%的企业担心生成结果不准确或内容幻觉问题。而企业自身的人才团队和经验无法保证平台以最优路径搭建，需要外部AI厂商的专业支持，辅助、引导其搭建适合自身业务的、定制化的、满足实际个性化需求的大模型服务平台。因此，可以从团队服务规模、案例数量、目标场景经验积累、技术相应程度、客户评价等方面来挑选合适的厂商。

● 5.2.3 合作伙伴：持续挖掘大模型应用潜力的关键

◆ **选择好的技术服务商可以为企业带来更大的直接价值和潜在收益。**IDC调研数据显示，47%的企业认为与AI大模型伙伴建立良好的合作关系来推动项目落地是能否取得成功的最重要因素。服务领先的AI厂商可以给项目带来效率提升、产品创新、收益增加和成本降低等价值，大幅提升企业生产力、驱动业务创新和增长。同时，选择好的技术服务商也可以享受后续模型更新、持续专业服务、业务咨询指导、生态圈交流、社区知识共享等更大收益。

◆ **头部AI厂商已经先行探索并积累大量经验。**大模型的落地需要专业的伙伴和长时间检验，头部AI厂商已经在汽车、零售、教育、泛互联网等多个行业服务并有了丰富的实践指导和案例参考；计划采购AI服务的企业则可借此对大模型有更充分的了解并知悉如何发展自身业务，避免重复试错带来的成本，避免重复建设、反复调优、多轮验证带来的时间消耗和市场机会的错失。

06

信赖之选

火山引擎大模型服务加速企业AI转型

火山引擎作为国内大模型领域的重要技术服务商之一，凭借其豆包大模型，火山方舟大模型服务平台，如扣子、HiAgent等多样的企业级智能体开发平台，及其在众多行业的成功落地经验和高效的服务响应，成为企业AI转型中值得选择的优质服务商。

6.1 豆包大模型

● 业务驱动模型效果持续增强

自火山引擎2024年5月发布豆包大模型以来，依托企业内部真实业务场景数据、外部案例合作以及实际场景中的工程化问题，豆包大模型架构和参数不断优化和增加，其模型效果已获得众多企业认可。截至9月，豆包大模型的日均tokens使用量已经超过1.3万亿，4个月的时间里tokens整体增长超过10倍。

● 模型家族适配多场景需求

豆包持续强化大语言模型、语音大模型、视觉大模型的能力水平，并在通用模型的基础上，面向行业领域不断丰富模型家族，多样化的垂直场景细分模型可适配客户的专业需求，更加专业和深入的案例实践可以加速垂直场景模型优化迭代。豆包通用模型pro为自研大语言模型专业版本，支持256k长文本，具备更强的理解、生成、逻辑等综合能力，适配问答、总结、创作、分类等场景。与pro版本对应的是lite轻量化版本，可为企业提供更低token成本和更低延迟。豆包模型家族的视频生成、角色扮演、语音合成、语音复刻、文生图、图生图模型已广泛在内外场景应用，例如电商营销、动画教育、城市文旅、音乐MV、微电影、短剧、沉浸式听书、陪伴式AI交互、视频配音、电商产品配图等。

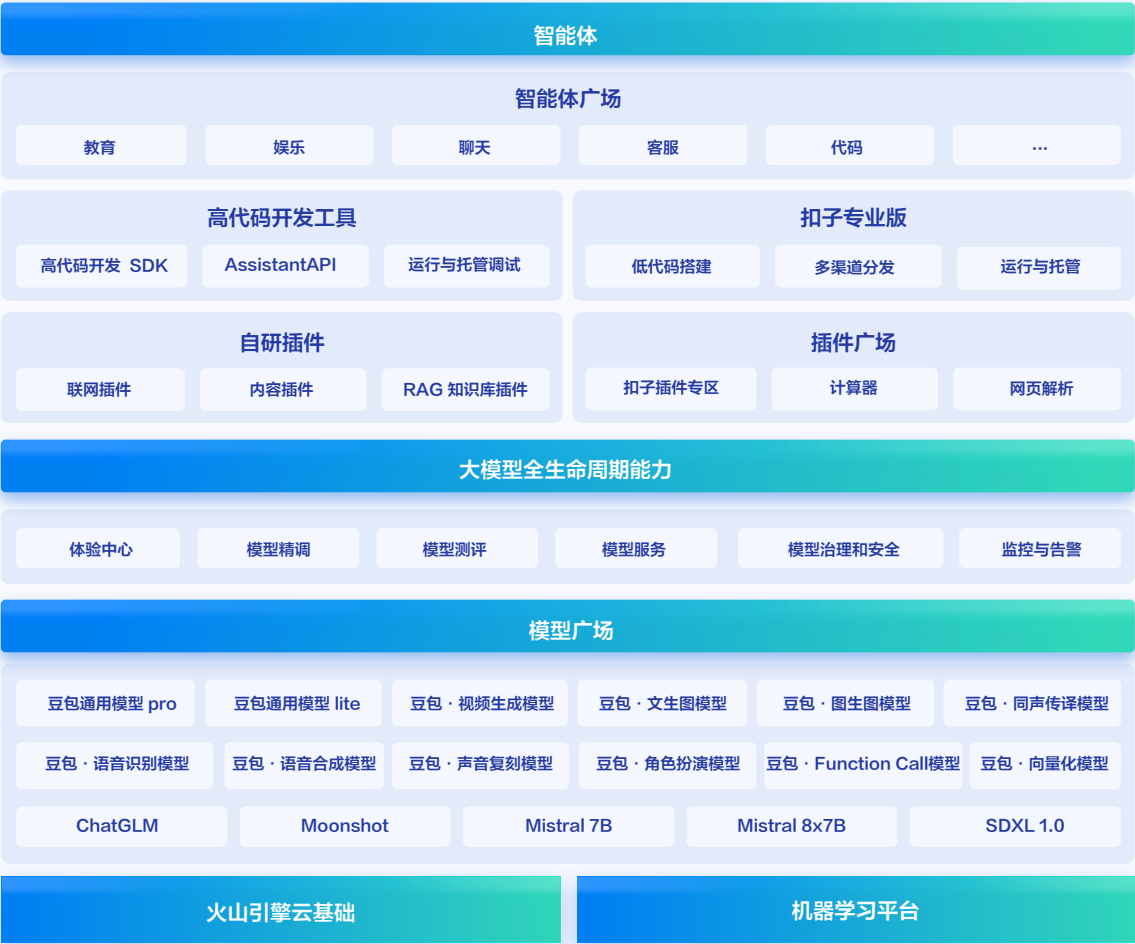
适配多种业务场景，驱动业务增长



来源：火山引擎，2024

6.2 火山方舟大模型服务平台

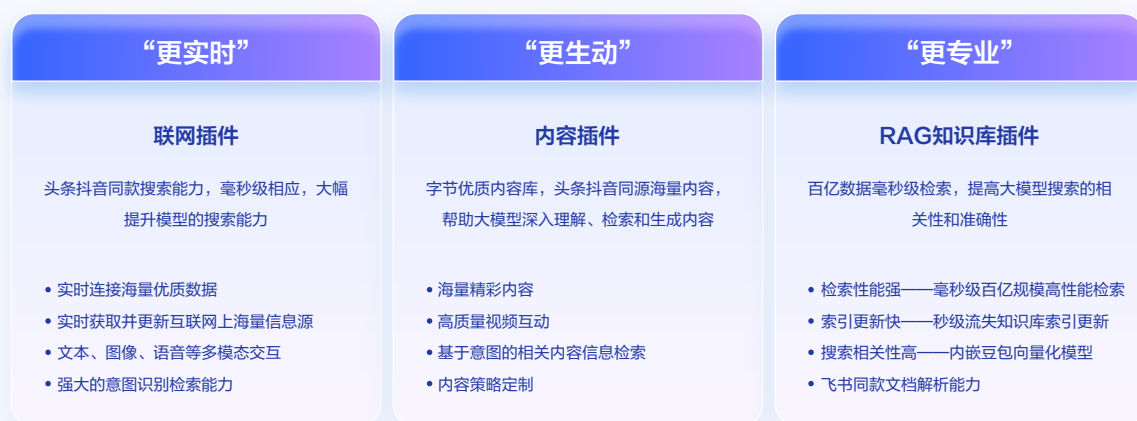
功能全景图



来源：火山引擎，2024

◆ **强劲的系统承载力保证大模型应用发挥乘数效应。**火山引擎提供领先的算力充沛供给、推理层优化策略和高水平系统调度能力，保证企业面临大模型需求涌现时可以拥有瞬时可用、极致弹性的海量资源。在面对科研、汽车、教育等不同场景下的高并发需求时，豆包通用模型pro可最高支持10k RPM和800k TPM，远超同行水平，并且可以根据实际需求进一步扩容，帮助大型企业在生产环境中落地大模型。

- ❖ **全周期安全可信方案帮助企业构建大模型安全“堡垒”。**火山引擎持续升级安全可信解决方案，面向企业关心的身份认证、数据保密、环境隔离、信息无痕、操作可审计等问题进行能力专项持续优化，如在任务结束后，沙箱会自动销毁数据，保证用户会话全程无痕；通过领先的可信容器沙箱技术，杜绝外部风险入侵和内部数据泄露。同时，面向推理、精调安全审计，火山引擎可全周期、全覆盖监控接入访问、KMS访问、沙箱登录、沙箱连接、沙箱容器逃逸、vArmor拦截日志，保证数据可保护、可审计、可追溯。
- ❖ **丰富的插件生态使得企业更好对接外部内容。**火山方舟自研联网插件、内容插件和RAG知识库插件，将运营内容和技术积累共享给企业，以增强模型落地效果。例如，联网插件提供头条、抖音同款搜索能力，可实时连接获取互联网海量优质可信数据，大幅提升用户的意图识别检索水平；内容插件可对接更多新闻、视频和垂类内容信息；RAG知识库插件基于飞书文档解析技术，可自动将输入文档内容切片、归类、嵌入，并达到百亿数据的毫秒级检索，大幅提高企业内部知识搜索的相关性和准确性。



来源：火山引擎，2024

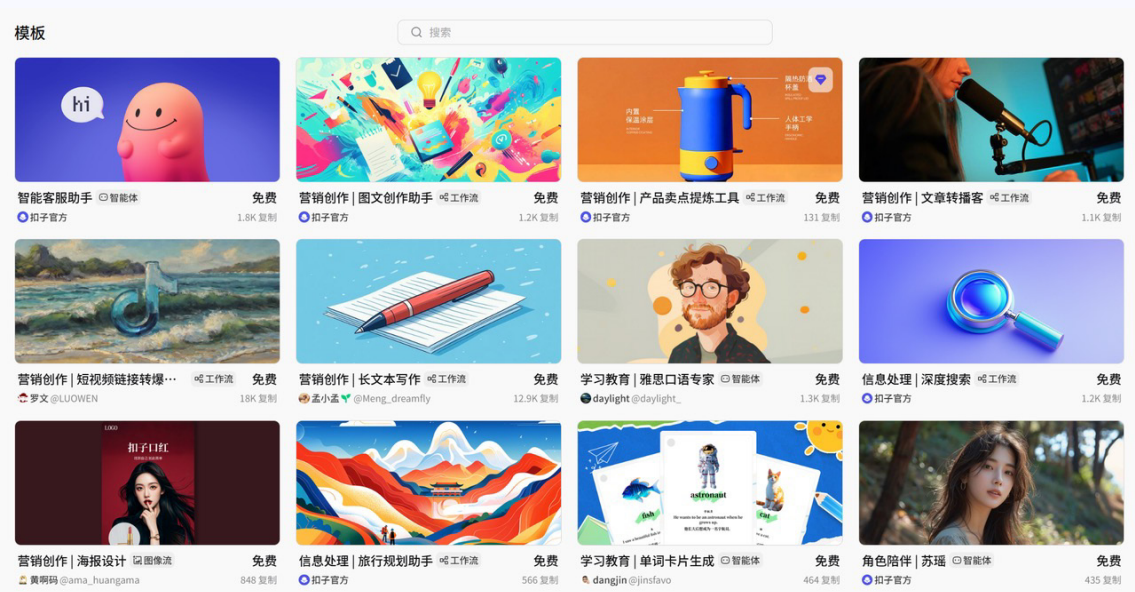
▶ **长期持续释放的服务价值。**不同于一次性交付服务，AI大模型作为企业的动态资产，只有不断地学习新知识和快速迭代，其作用和价值才可持续发挥。为此，火山方舟打造客户全周期服务链条，总结实践经验形成方法论，帮助客户理解大模型应用边界和落地预期，提供场景咨询、评测体系建立、目标规划、ROI计算、模型选择、应用落地、效果达标、业务分析、服务优化、新能力测试、增购扩容支持。IDC观察到，企业在初步落地大模型应用后，在持续更新模型能力的同时，还会加大对IT和业务数字化转型的投入，包括云服务、数据管理、开发工具、流程软件等，这些也是火山引擎所具备的。



来源：火山引擎，2024

6.3 扣子

- 降低开发难度，更好地服务企业用户。面向AI应用开发人员不足、服务稳定性要求高的企业，为降低大模型应用的开发门槛，支持用户快速搭建AI应用，进一步缩短从构思到落地实践的时间，火山引擎发布扣子专业版，以专业级的稳定支撑、火山方舟模型资源快速接入、更好的团队资源配额为核心价值，满足企业对高并发、高稳定性的要求，将应用开发周期缩短95%以上，同时加强数据安全和资产安全保护。
- 丰富的插件和流程智能化管理，快速打造企业级超级应用。扣子专业版定位为企业级应用打造平台，链接插件、工作流、图像流、知识库、数据库等一系列工具，无缝衔接拥有更强模型效果的豆包大模型家族，以及企业精调模型与开源模型，加强团队紧密合作和开发流程化管理，实现逻辑串联和多技能并行，面向企业用户提供自身积累沉淀的、丰富的应用模板，以及调测、发布、集成、观测的全周期服务。例如使用新闻插件，可以快速打造上线一个能够播报最新时事新闻的AI播音员。



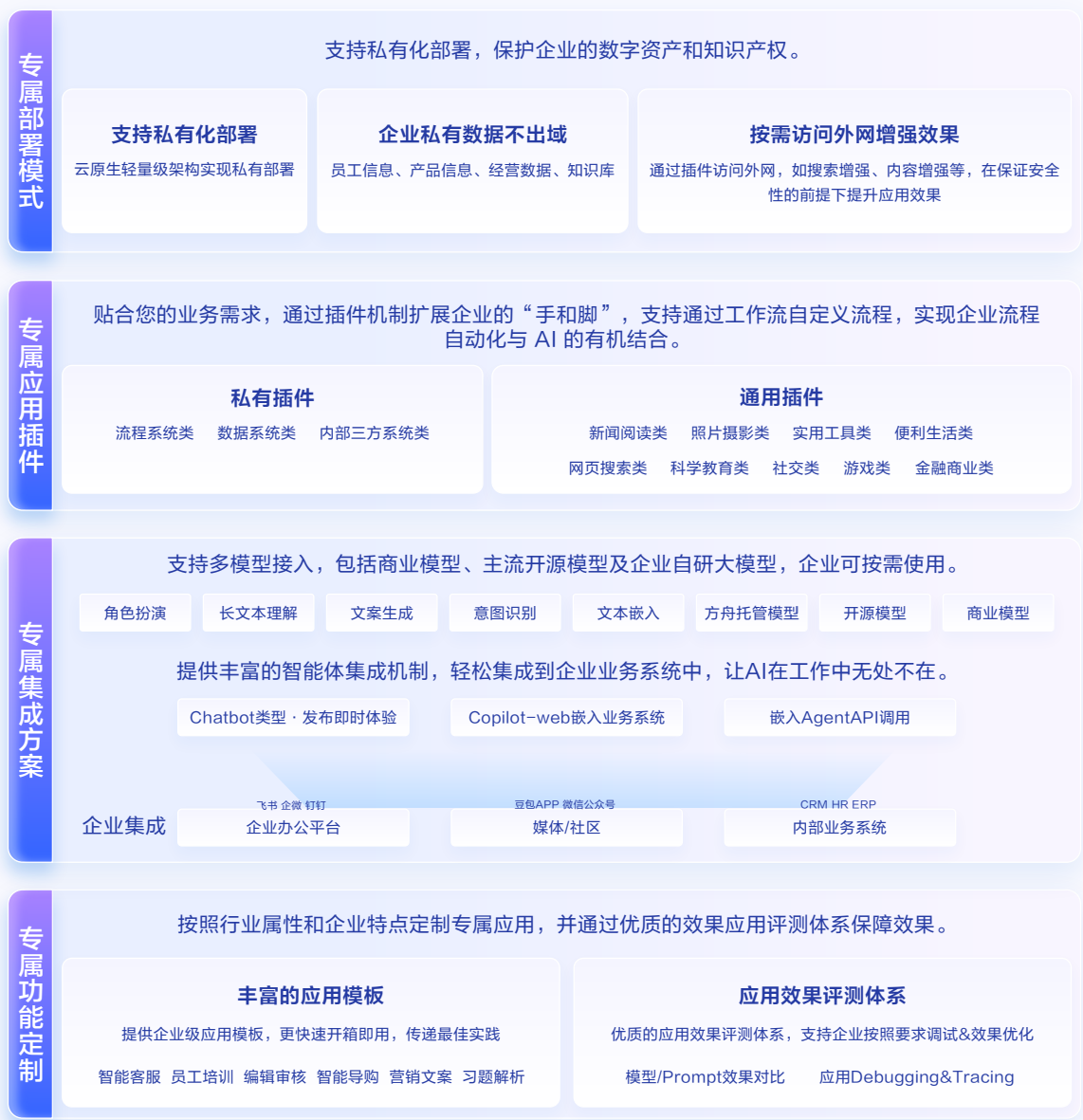
来源：火山引擎，2024

◆ **上线扣子专业版模板商城，快速复用AI优秀实践模板。**为解决市场上无参考、用户寻找模板困难、个性化修改过程复杂的问题，扣子专业版新推出了模板商城。通过有效的分类和搜索功能，以及展示热门数据、官方推荐和专题，用户可以精准地了解 and 接触目标领域和其他领域的模板，从而激发创作灵感、跨界开发应用。目前商城已经上线了智能客服助手、图文创作助手、雅思口语专家、名画写真、作业批改等优秀模板，在火山引擎内部和外部客户中被大量使用和认可。火山引擎仍在不断优化模板商城生态，帮助用户在0-1阶段，以低成本、低代码方式更好地搭建出实用的POC方案并快速上线。

6.4 HiAgent

◆ **更懂AI转型的专属应用创新平台，助力更多企业快速创新落地。** HiAgent定位为企业专属的AI应用创新平台，目的是帮助企业能够用零代码或低代码的方式走完从模型到应用的“最后十公里”。HiAgent可以向下兼容多种模型，原生整合豆包大模型的同时还能兼容第三方的闭源或者开源模型，并通过提示词工程、知识库、插件、 workflow编排四要素来辅助轻松构建智能体，预置丰富的模板和插件，并给予自定义的自由空间。目前HiAgent已经为招商银行、海尔集团、中国国际航空等500强企业和浙江大学等顶尖985高校提供服务，落地200+场景，打造Agent 500+款，惠及数百万终端用户，持续沉淀丰富的行业模板为更多企业提供服务。同时，提供端到端的AI咨询服务以及切实可行的AI转型实践方案，助力企业轻松落地AI创新业务。

四大特性助力企业构建专属AI应用创新平台



来源：火山引擎，2024

◆ **更灵活适配企业业务，将丰富的模型选择、行业插件与业务系统深入集成。**

HiAgent提供了更加丰富、面向垂直领域的模型家族和一站式服务，联合金融、教育、消费、医疗健康等行业头部ISV打造了100+行业插件，并支持插件可定制，帮助企业无缝衔接现有业务系统。企业可以利用AI专属平台提升创新速度、加深知识密度、增加能力厚度。HiAgent支持专属部署模式、专属应用插件、专属集成方案和专属功能定制，帮助企业搭建具备自身特色的、定制化的应用创新平台，为企业提供一个系列具备专业流程的模板，引导用户通过简单操作即可完成知识切片、提示词编写、多Agent调用管理。此外，在丰富的“开箱即用”功能集市基础上，沉淀积累融合企业内部数据、经验、流程的能力，将杂乱的数据和信息变成有用的知识，打通各部门系统流程，做厚AI中台能力、做到举一反三，发挥飞轮效应。

◆ **更强的安全保障让企业大模型安心落地。** HiAgent支持RAG知识库和大模型全栈私有化部署，为知识库提供数据安全防护，并提供覆盖Agent全生命周期的安全保障，同时面向安全隐私，集成HiAgent、知识库、大模型的全栈私有化部署解决方案，可以实现事前、事中、事后的全周期保护，消除恶意代码注入、幻觉输出，实现知识库数据保护、细粒度访问控制。

结语及未来展望

当下，大模型及其落地处在一个快速变化和不断演进的过程中。大模型技术服务商既要紧密贴合企业实际业务需求，又要不断突破技术革新的边界。在这个动态发展的过程中，只有将市场需求与技术发展紧密融合，相互激发，才能确保大模型的发展在未来既顺应市场潮流，又保持技术的前瞻性。

● 从企业侧来看，企业对AI大模型充满期待，且更加依赖AI头部厂商

IDC调研显示，整体市场与经济环境稳中向好，有22.2%的受访用户认为未来1年不会减少对AI和自动化的投资预算，且受访企业一致同意AI大模型需要更多的投入和伙伴。大模型能力的快速迭代以及低代码工具生态的丰富，让企业会更加倾向于选择模型直接应用，并不进行复杂的调优。对此，拥有更多数据和行业实践经验的AI头部厂商将更具优势，也具备更多的人才和资源与企业陪跑，助其完成持续迭代和业务创新。

● 从技术本身来看，模型效果持续增强，将解决更复杂的开放式任务

Scaling Law（规模定律）将长期存在；随着数据量的持续上涨和任务复杂度的提升，基座模型的效果将会随着训练和调优持续增强，大模型的能力和专业性也将进一步提升，例如数学逻辑推理、高维计算、视频任务、多模态生成。小参数模型的能力因任务和应用领域不同而差异化明显，大型模型则更加通用，在复杂、开放式任务中的表现水平更高，在面向特定应用和服务需求时的可迁移性、可塑性也更加优越。

● 垂类模型、多模态、智能体是AI大模型突破应用边界的关键

当前，大模型的模型应用正在从图片、文本生成，走向视频、多模态交互方向。随着越来越多的企业应用大模型并创新个性化的模型服务，行业领域垂直、细分业务的模型家族将会更加丰富，并出现高水平的模型和高质量的生成内容，提升使用满意度。同时值得更多关注的是，Agent智能体的开发将是增加大模型服务广度和深度的关键，使其可以同时处理多项实际任务和问题，执行复杂的操作并给出更加满意的结果。虽然目前大模型在数学、逻辑推理和优质内容生成上仍有较大发展空间，但随着头部AI厂商融合更多行业落地实践，创新思维链、逻辑链，大模型也将会更好地服务企业快速实现AI转型。

关于 IDC

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

IDC China

IDC中国（北京）：中国北京市东城区北三环东路36号环球贸易中心E座901室

邮编：100013

+86.10.5889.1666

Twitter: @IDC

blogs.idc.com

www.idc.com

版权声明

凡是在广告、新闻发布稿或促销材料中使用IDC信息或提及IDC都需要预先获得IDC的书面许可。如需获取许可，请致信 gms@idc.com。翻译或本地化本文档需要IDC额外的许可。

获取更多信息请访问www.idc.com，更多有关IDC GMS信息，请访问<https://www.idc.com/prodserv/custom-solutions>。

版权所有2024 IDC。未经许可，不得复制。保留所有权利。