

企业AI应用行动指南2024

易观分析
2024年5月

本产品保密并受到版权法保护

Confidential and Protected by Copyright Laws

更多免费行业报告

并购家

www.ipoipo.cn

并购家 免费行业报告网

IPOIPO.CN 每日更新 不用注册会员 无广告 永久免费

随着人工智能技术的快速发展，社会和经济正逐步迈入万物AI时代，企业积极探索AI在业务中的应用，以提升效率和创造价值。

《企业AI行动指南》旨在为企业提供全面的AI大模型部署指导，涵盖AI大模型的发展概述、战略决策和部署步骤、方法及选型建议。该指南还指出部署过程中的常见误区，并提供策略自检工具。

《企业AI行动指南》由易观AI咨询顾问团队根据实战经验编制，随着客户需求的不断变化，我们将持续更新报告，帮助企业梳理决策流程、提高效率、规避风险，并抓住AI带来的机遇。

指南中包含“AI场景选择三度模型”、“行业AI应用成熟度曲线”、“大模型选型评估矩阵”等实用工具，帮助企业选择合适的AI应用场景和模型。如需详细解读，可通过二维码联系易观分析获取专业指导。



易观分析

01

2023年，大模型元年，
这一次真正撼动了人类

中国大模型的发展呈现出蓬勃生机，形成了多元共进、创新竞发的活跃态势



LLM开启的生成式人工智能到底可以解决什么问题，与之前相比，进化在哪儿？



企业由数字化阶段迈入智能化阶段，AI能力成为企业核心能力

易观分析认为：

- 智能化是互联网化、数字化的延伸而不是颠覆，演进过程中，技术的价值从过程逐步向结果转移；
- 未来企业，组织形态将会被重构，组织边界模糊，组织结构灵活机动，以AI为核心重新定义；
- 企业的AI能力（含“AI”量）成为核心能力。

智能化

“数据+算力+算法”定义世界逻辑，将智能决策与资源优化配置相结合，进一步提高决策科学性与实时性，并加速进入产业数字化深水区

数字化

数字技术向实体经济与企业经营环节渗透，依托网络进行信息高效传递，从而解决资源调配问题

互联网化

互联网快速扩张，跑马圈地，培养用户习惯，经典消费互联网运转逻辑

01

AI+：基于企业现有流程和服务，使用AI技术赋能与增强能力

02

企业拥有AI原生产品和服务

03

围绕着AI原生产品，企业重构组织形态和能力

大模型所加速的生成式人工智能已经渗透到多个场景



文本

- 对话/问答
- 文档/文本/文案生成
- 内容/会议摘要等
- 语言翻译
- 文学/剧本创作等

01



代码

- 自然语言生成代码
- 代码补齐
- 生成SQL
- 生成软件测试用例
- 合成数据等

02



图像

- 图像分类/分割
- 工业设计
- 医学影像标注与解剖结果构建
- 艺术/商业作品创作
- 图像修复
- 天文观测、卫星遥感观测等

03



音视频

- 信息播报
- 语音编辑/翻译
- 影视内容分析编辑
- 视频增强/风格迁移
- 音乐/视频生成

04



3D

- 电影/游戏/动画制作
- 建筑/家居设计
- 工业制造
- 工业/艺术设计
- 医疗健康
- 虚拟现实等

05



分子发现

- 药物设计
- 材料科学
- 食品与农业
- 能源
- 个人护理等

06

大模型能力与AIGC相结合向企业经营关键环节渗透



行业应用全面铺开，实践案例示范价值在于效益比测算与优化

金融行业

- 前端对客的核心价值与场景主要为提升服务体验，包括智能营销、智能客服等方面，同时也在中后台运营管理方面，例如信贷审批、核理赔等流程性任务层面提升工作效率
- 智能投研与投顾、智能财富管理与量化交易等金融特定任务方面则需要金融大模型的训练与微调，探索创新价值

实践案例：中国农业银行ChatABC，应用于多轮对话、内容摘要等场景



- 建设目标：着眼于大模型在金融领域的知识理解能力、内容生成能力以及安全问答能力，实现金融知识理解和智能问答应用
- 建设要点：对于大模型微调、提示工程、知识增强、RLHF等大模型关键技术进行综合应用，结合农行业务支持知识库、内部问答数据以及人工标注数据等金融知识进行融合训练调优
- 参数规模：百亿级
- 当前能力：可初步具备自由闲聊、行内知识问答、内容摘要等多类型任务的服务能力
- 探索场景：已在行内多个渠道以多轮问答助手、工单自动化回复助手等形式面向内部员工开放试用

电商/零售行业

- 核心价值目前主要体现在面向用户运营侧的服务体验提升，以及营销运营过程中的内容生成与提效，主要应用场景包括智能客服、商品海报与文案生成等
- 中期数字人赋能用户营销正在提速，长期来看将从前端向中后台乃至供应链与产品研发延伸

实践案例：阿里通义大模型赋能商家，AIGC生成3D商品与店铺营销物料



娱乐/游戏行业

- 包括影视视频、游戏等在内的娱乐行业是高度依赖于内容资产行业，包括图像、音视频、3D资产等，借助于大模型与AIGC能力首先提高内容生产效率是当务之急
- 无论是影视还是游戏，都在营造世界观与价值观的过程中需要故事线的引导与文本对话的链接，相对应地，短期看文本与对话生成等，中长期看智能NPC乃至游戏策略设计

实践案例：启元世界游戏AI解决方案



实践案例：Runway Gen-2赋能视频创作



教育行业

- 围绕受教育者与C端用户：覆盖K12、高等教育、职业教育等不同类型，主要侧重于提升交互体验以及个性化教学方面展开
- 围绕赋能教育机构与施教者/老师：覆盖备课、教学、考试评测与学生管理多个环节，在原有教育信息化的基础上进行智能化升级

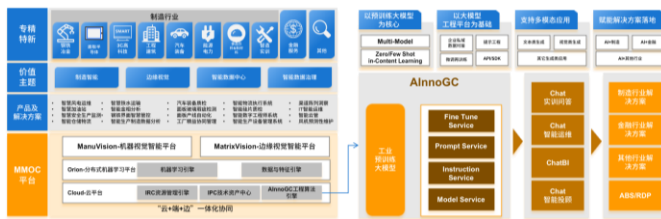
实践案例：科大讯飞星火大模型赋能教育行业



工业/制造

- 工业/制造行业是知识密集型行业，尤其是进入到工业细分行业，知识密度与信息壁垒，包括工艺、成分、流程等，都高度差异化，相应地，工业行业大模型的必要性甚高
- 目前主要在AI辅助研发设计、工业质检（即质量检测与缺陷分析）、生产流程智能化等方面应用，未来则需要探索AI+机器人的智能化升级，用大模型链接物理世界

实践案例：创新奇智人工智能技术栈及“奇智孔明”应用探索



医药/医疗

- “AI+医疗”主要应用场景与价值在于电子病历生成与分析、AI辅助诊疗，包括AI影像识别与临床辅助决策等，赋能医生与医疗机构提高平均专业水平，提高工作质量与效率
- 新药研发利用语言模型等进行靶点发现、化合物筛选、临床实验研究等，尤其是在药物发现环节充分发挥探索性价值

实践案例：华为基于Pangu Drug Model提供AI辅助药物研发平台，覆盖新药研发全流程



易观AI大模型产业全场景图谱

AI大模型应用层

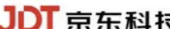
行业应用层

金融 医疗 教育 工业 传媒/影视 零售电商 娱乐游戏 企业服务 政务

用户应用层

内容消费 创作工具 企业级用户 政府机构用户

AI通用基础大模型

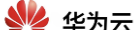
 文心大模型	 日日新大模型	 悟道大模型	 九天·众擎基座	 昆仑天工
 通义千问	 盘古大模型	 智谱·AI 智谱GLM	 星辰语义	 紫东太初
 混元大模型	 言犀大模型	 百川大模型	 鸿鹄图文	 云天书
 星火大模型	 豆包大模型	 ABAB	 360智脑/奇元	 K7大模型

AI大模型工具平台

 千帆大模型	
 火山引擎	
	
	

AI大模型基础层

算力基础

AI芯片	云服务	智能计算平台
 英伟达	 阿里云	 阿里云飞天智算
 AMD	 腾讯云	 青云智算平台
 寒武纪	 百度智能云	 华为昇腾计算
 地平线	 华为云	 百度百舸异构计算

智能服务器

 浪潮

 HUAWEI

 中科曙光

数据基础

数据服务	向量数据库
 海天瑞声	
 星辰数据	 星环科技
	 爱可生

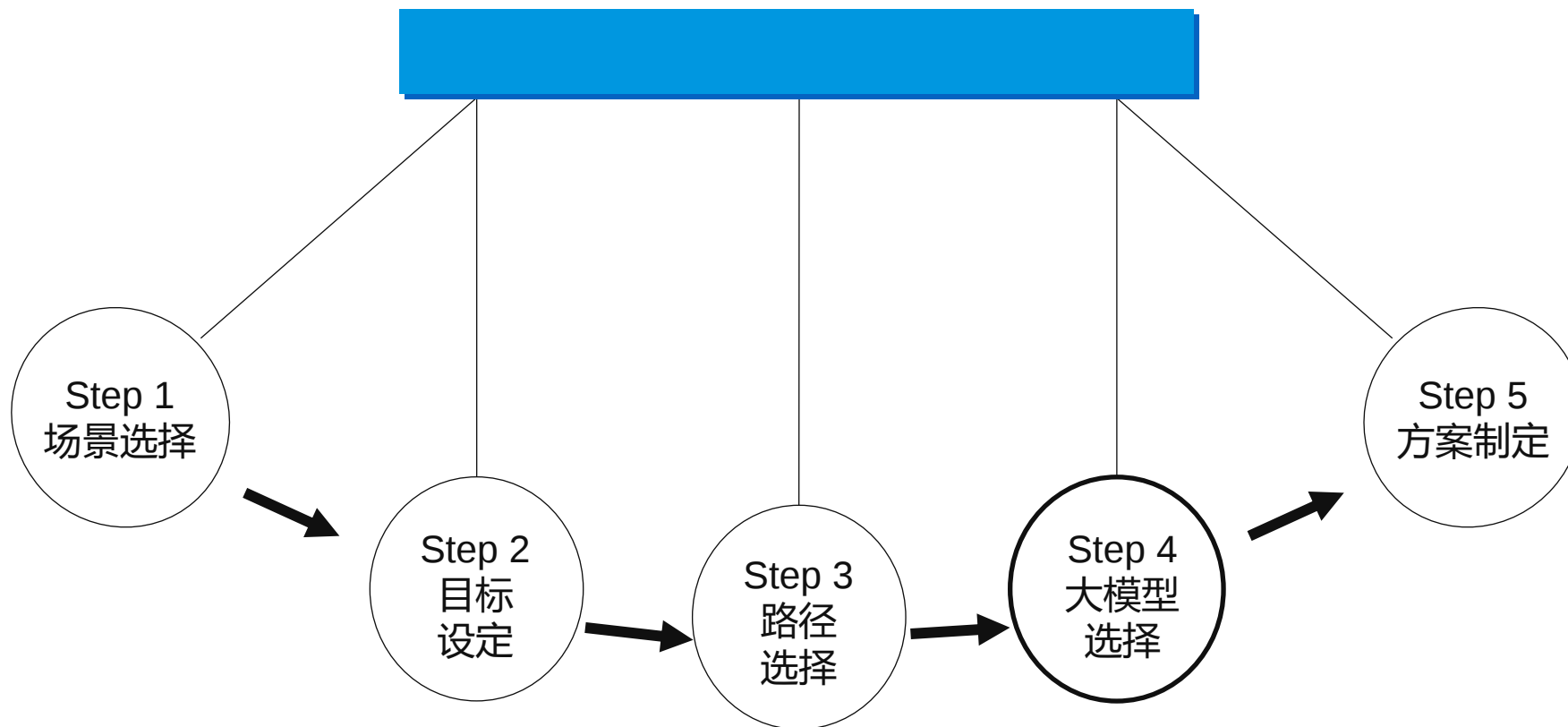
算法基础

02

AI策略的选择与决策

Step 0: 理解AI, 了解AI行业应用最佳实践



Step 1: 场景选择的主要考虑问题

AI应用成熟度

- 技术成熟度（可以参考Gartner的hyper cycle）
- 企业AI通用应用成熟度（可以参考易观企业应用AMC曲线）以及企业AI应用最佳实践
- 行业AI应用成熟度（可以参考易观行业AMC曲线）以及行业AI应用最佳实践
- 本企业的场景成熟度评估（需要考虑业务本身的健康度、标准化程度、AI化意愿等）

企业资源就绪度

- 组织能力与团队（包括战略与文化、组织与团队、技能与培训等）
- 预算
- 数据
- AI工具和基础设施
- 其他

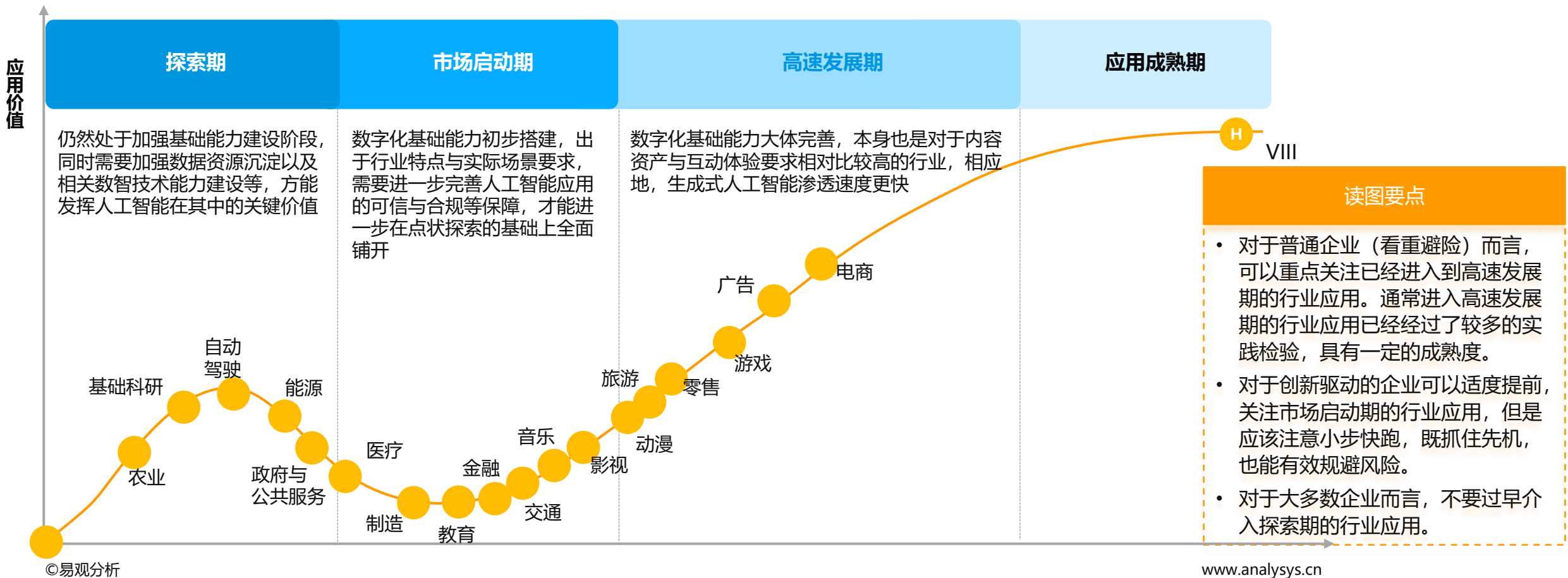
企业部署价值度

- AI投入成本分析
- AI的收益分析
- AI风险评估

示例：各个行业应用专有场景的成熟度

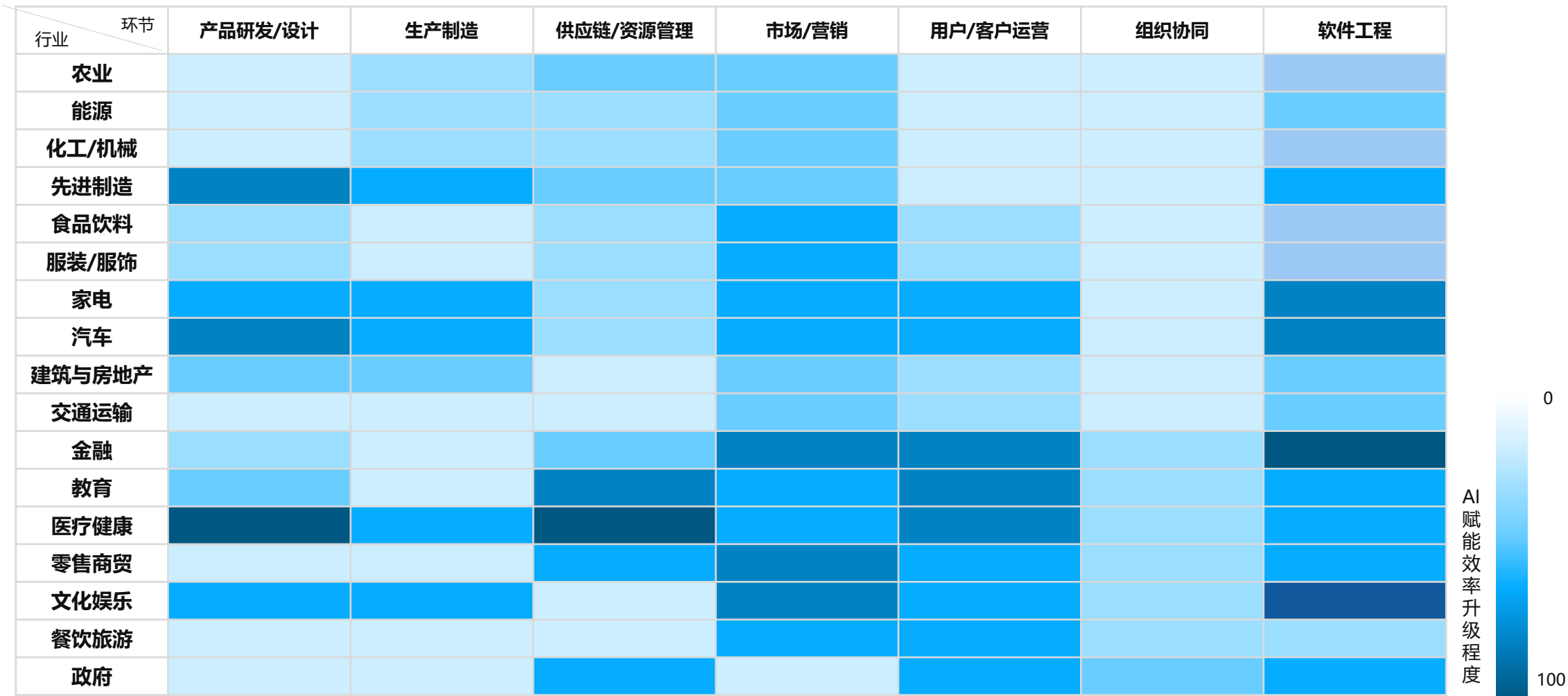
中国生成式人工智能行业AMC应用曲线（2023）

行业发展特点： ← 稳中求进 ————— 效率为先 →



示例：各个行业通用场景AI化的成熟度

深入行业关键环节的AI价值渗透度

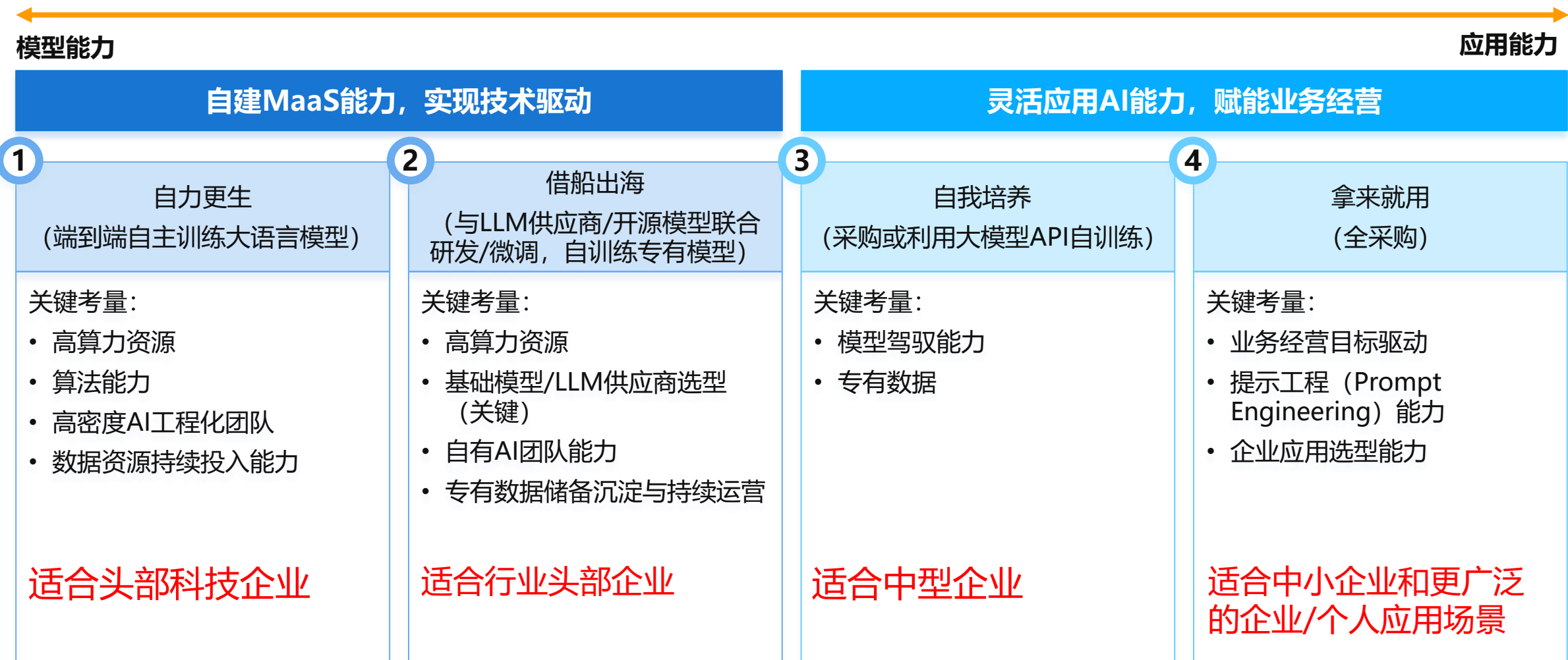


Step 2: 目标制定

AI策略的目标，要有业务负责人把控而不是单纯的技术负责人。



Step 3: 路径选择



企业部署AI大模型路径一：端到端自主训练大语言模型

- 易观分析认为，端到端自主训练大语言模型已经成为国内科技巨头布局的重点之一，这不仅是基于商业考量，更是基于技术、战略和生态系统建设的全方位考量。

端到端自主训练大语言部署关键考量要素

端到端自主训练大语言模型更适用于有一定IT投入与研发能力的企业，有较为完善的云服务及数据生态。**关键考量要素**包括：投入**算力资源**的能力、**AI工程化建设团队能力**以及**数据资源投入能力**。



算力资源



AI工程化团队



数据资源投入能力

自建规划

- 企业需要明确定义自主训练大模型需要解决的问题或实现的目标；
- 收集和清洗相关的数据。

模型训练

- 根据需求和可用数据，选择适合的大语言模型架构和训练算法；
- 训练自己的模型阶段，需要进行模型训练、调参和验证，以达到预期的性能指标。

部署集成

- 部署到生产环境中，并与现有的系统和应用程序集成，需要考虑到安全性、性能和可扩展性等方面的问题。

端到端自主训练大语言模型成为国内科技巨头布局重点

拥有自主研发和掌握的大语言模型将为科技巨头带来技术领先的优势，通过自主训练，科技巨头可以针对本土语言和文化特点进行优化，提供更贴近用户需求的智能服务。此外，大语言模型已经显现出在各行各业都有广泛的商业应用潜力，科技巨头同时对外输出模型服务。

典型实践 百度 阿里巴巴 Tencent 讯飞星火 HUAWEI

百度 文心大模型

百度文心大模型师基于自主研发的端到端训练框架，使用大规模的中文语料进行训练，具有深度学习和自然语言处理技术的先进特性。此外，文心设计之初就考虑了与产业的深度融合，提供了全流程的支持工具和方法，以适应不同行业需求。

阿里 通义千问大模型

通义千问是一款大型中文自然语言处理模型。在模型架构设计上更侧重于实现特定的应用场景或任务。当前，阿里通义千问已经应用于智能客服、文档自动化、内容生成、智能推荐等多个阿里生态场景，为用户提供更智能、更个性化的服务和体验。

腾讯 混元大模型

腾讯混元大模型是一款结合了自注意力机制和卷积神经网络的混合层级架构模型，具有多粒度特征融合、跨模态支持、大规模参数等特点。混元支持多种语言模态的处理，包括了文本、图像、语音等，拓展了模型的应用场景和能力。

企业部署AI大模型路径二：利用开源模型或与LLM供应商进行联合研发与微调，训练专有模型

联合研发与微调部署大模型路径的关键考量要素

利用开源模型或者联合研发与微调部署大模型路径更适用于有较强行业专属场景业务需求，并需要实现技术驱动赋能的企业。

关键考量要素包括：一定的算力资源基础与投入、基础模型选型或LLM供应商选型、自有AI团队实现专有模型可持续运营与迭代、专有数据储备沉淀与持续运营能力。

01

业务需求分析与规划：

企业需要明确他们的需求，包括模型的预期用途、性能指标、数据集需求等

02

选择开源模型或LLM供应商：

根据需求和预算，企业需要选择合适的开源模型或LLM供应商，可能涉及评估不同模型的性能、适用性和许可协议等因素

03

模型微调：

准备大量的训练数据，或使用选定的开源模型或LLM作为基础进行模型微调，通常涉及在企业自有数据集上进行训练

04

部署和集成和评估优化：

一旦微调完成并达到预期性能，就可以将其部署到生产环境中，包括将模型集成到现有的软件系统中，并确保其稳定性和可靠性

联合研发与微调部署路径成为行业头部企业首要选择

联合研发和微调部署路径为企业提供了一种灵活且高效的方式来推动企业AI大模型的部署，并且通过这种路径可以使企业能够更有效地利用现有大模型的技术成果和资源、快速提高自身创新能力，并更快地将AI应用赋能于自身业务优化上。

典型实践 ICBC



BloombergGPT



国家电网
STATE GRID

中国工商银行
X
昇腾AI大模型

中国工商银行发布基于昇腾AI的金融行业大模型；

构建了从规划、开发到产业化的大模型全流程使能体系，这包括大模型开发使能平台和大模型部署套件，旨在简化大模型的开发和部署流程。

BloombergGPT

BloombergGPT是一个包含500亿参数的语言模型，专门为金融领域设计和训练，以支持金融行业内的自然语言处理任务；

训练数据主要来自彭博社的金融数据，针对金融领域的专业术语和行业趋势。

国家电网
X
百度文心

国家电网与百度合作，共同研发了电力行业的专有大型模型；

在通用文心大模型基础上，设计电力领域特定的预训练任务，结合电力行业的数据和知识进行训练，以提升模型在电力行业特定任务中的应用。

企业部署AI大模型路径三：采购或者利用开源大模型API，利用向量化方式优化自身应用

采购API部署大模型路径的关键考量要素

采购或利用开源大模型API的企业AI部署路径适用于大多数企业，IT能力建设并非重点，但是需要利用AI能力提效降本，赋能业务经营。

关键考量要素包括：企业自身应用对于AI能力的抽象以及对模型的驾驭能力，专有行业领域数据以及企业业务数据沉淀以及向量化处理能力。

	需求分析和规划： 企业首先需要明确他们的AI需求，确定他们希望利用大模型API解决的问题或任务
	选择合适的大模型API供应商： 评估大模型API供应商，并选择最适合企业需求的供应商，涉及考虑供应商的技术能力、性能、价格、服务支持等方面
	API接入和集成： 包括获取API访问权限、了解API的使用方法和限制，并将API集成到企业的软件系统中
	模型调优和性能测试： 企业需要对模型进行调优并性能测试。包括使用验证数据集对模型进行测试，并根据性能指标进行调整和优化

采购API部署路径成为中型企业借力AI升级服务的选择

采购API并将其部署是中型企业在借力AI能力升级服务水平与平衡AI投入成本后普遍采取的方式，API部署通常更快速，企业无需从头开始构建AI模型或基础设施。此外，API提供了灵活性和可扩展性，中型企业可以根据实际需求选择适合的API，避免过度投资和资源浪费，还可以根据业务增长的需要轻松扩展或调整API的使用。

典型实践



企查查通过使用向量数据库来增强其搜索引擎的功能，并在端侧部署模型，可以在多个层面上提升产品的AI功能和整体品质

可以在保持较低成本的同时，提供更加智能和高效的服务，满足用户对于高性能AI应用的需求



多邻国采购GPT模型来改进其语言学习工具，使其能够更自然地生成语言示例和提供更准确的语言练习

多邻国通过集成最新的AI技术，尝试教育领域的创新，以提高用户的学习体验和效率

企业部署AI大模型路径四：调整企业应用的选型策略，提升AI能力评价权重

提升AI能力评价权重的方式部署路径的关键考量要素

以调整企业应用的选型策略，通过提升AI能力评价权重的方式来部署AI的路径适用于大多数企业拥抱数智化的场景，是企业软件AI能力的整体升级。

关键考量要素包括：以业务经营为目标适配驱动AI部署，提示工程能力，更新企业软件选型标准与合作伙伴等。

AI能力评估

企业首先需要对其现有的AI能力进行评估，涉及对现有AI技术、人才、数据和基础设施等方面进行调查和分析，以确定企业目前的AI成熟度和能力水平，在此基础上确定AI可以提供价值的领域和场景。

AI技术选型和供应商评估

在选型过程中，企业可以通过提升AI能力评价权重的方式，将更多的重要性放在AI能力方面，以确保选择的技术能够满足其业务需求并实现预期的效果。

AI能力部署和集成

企业可以开始部署和集成AI技术到其现有的业务应用中。这可能涉及定制化开发、数据集成、系统集成等工作，以确保AI技术能够顺利地与现有系统和流程整合。部署后，企业需要进行持续的监控和优化，以确保AI技术的稳定性、性能和效果。

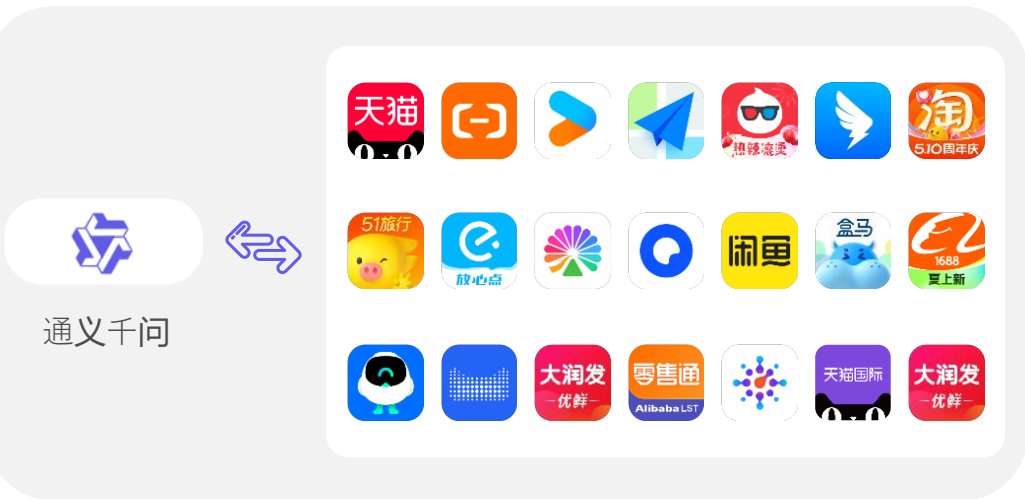
升级企业软件集成AI能力适用于大多数企业构建数智化场景

通过这种路径可以有效提升企业业务效率、增强用户体验、开拓新业务机会、降低成本和风险，并符合数字化转型的趋势，有助于企业提升竞争力并实现持续发展。

典型实践



阿里旗下生态APP接入大模型升级AI能力



Step 4: 大模型企业与合作伙伴选择

应用层



中间层

中间层生态是大模型能力的重要体现



模型层

备注：大模型产品仅为示例，未包括海外大模型企业



基础设施层



模型能力

★★★★★

需要关注基础大模型的关键能力，包括语言能力（简单理解、知识运用、推理能力、特殊生成等）、安全和价值观以及通用任务能力，可以参考FlagEval（天秤）语言大模型评测体系



产品能力

★★★★★

需要关注基础大模型的产品化封装与解决方案能力，包括大模型能力抽象与API化易用性、大模型训练与微调环节支撑与服务保障能力、任务场景实践能力，以及运营维护保障能力等



生态能力

★★★

侧重大模型生态发展策略以及布局情况，包括中间层生态是否完整丰富以便于模型的精调与持续运营，关键环节国产供应链布局与适配度，行业伙伴以及最佳实践案例，开源策略与协议等



可持续发展能力

★★

基础大模型非一蹴而就就需要持续投入与迭代，需要关注大模型企业战略路线与资源投入程度和专注度、人才梯队建设与核心人才情况等

中国主流大模型选型评估——模型能力x产品能力

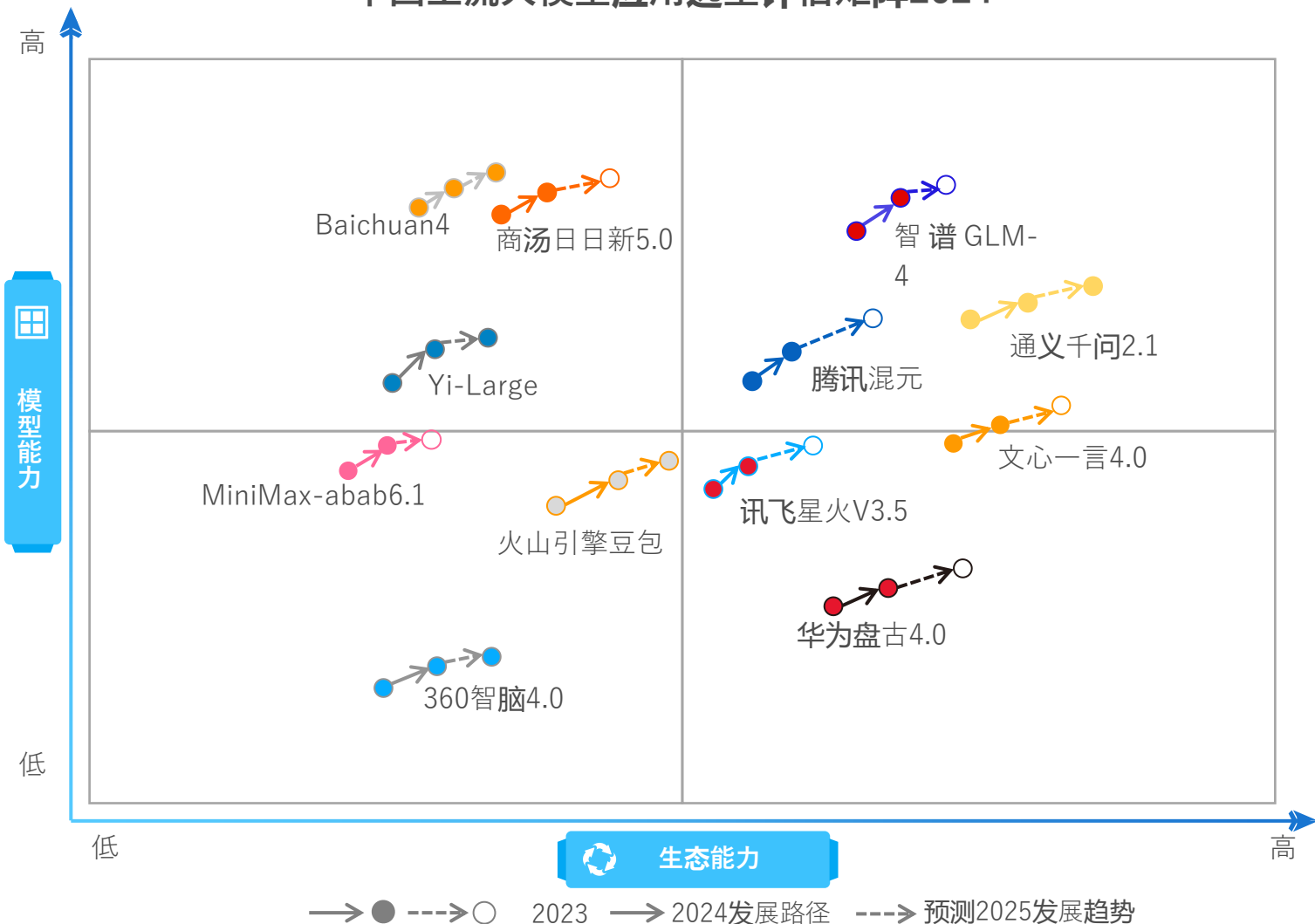
中国主流大模型应用选型评估矩阵2024



- 选取中国大模型12家代表厂商进行评估。本应用选型实力矩阵评估主要聚焦于产品能力的模型应用选型，从模型能力和产品能力两个维度综合反映大模型应用选型市场上主流大模型的市场表现和竞争地位的研究模型。
- 易观分析认为，随着大模型的发展受到多方面利好支持以及AI大模型赛道的不断发展，各大模型将继续在产品能力上进行优化和提升。总体来看，中国主流大模型在产品能力方面呈现出百花齐放的局面，各具特色的产品为企业在不同领域的智能化转型提供了多样化的选择。随着技术的不断进步和市场需求的不断变化，各大模型的产品能力必将在未来持续提升，为企业带来更多的创新机会和商业价值。

中国主流大模型选型评估——模型能力x生态能力

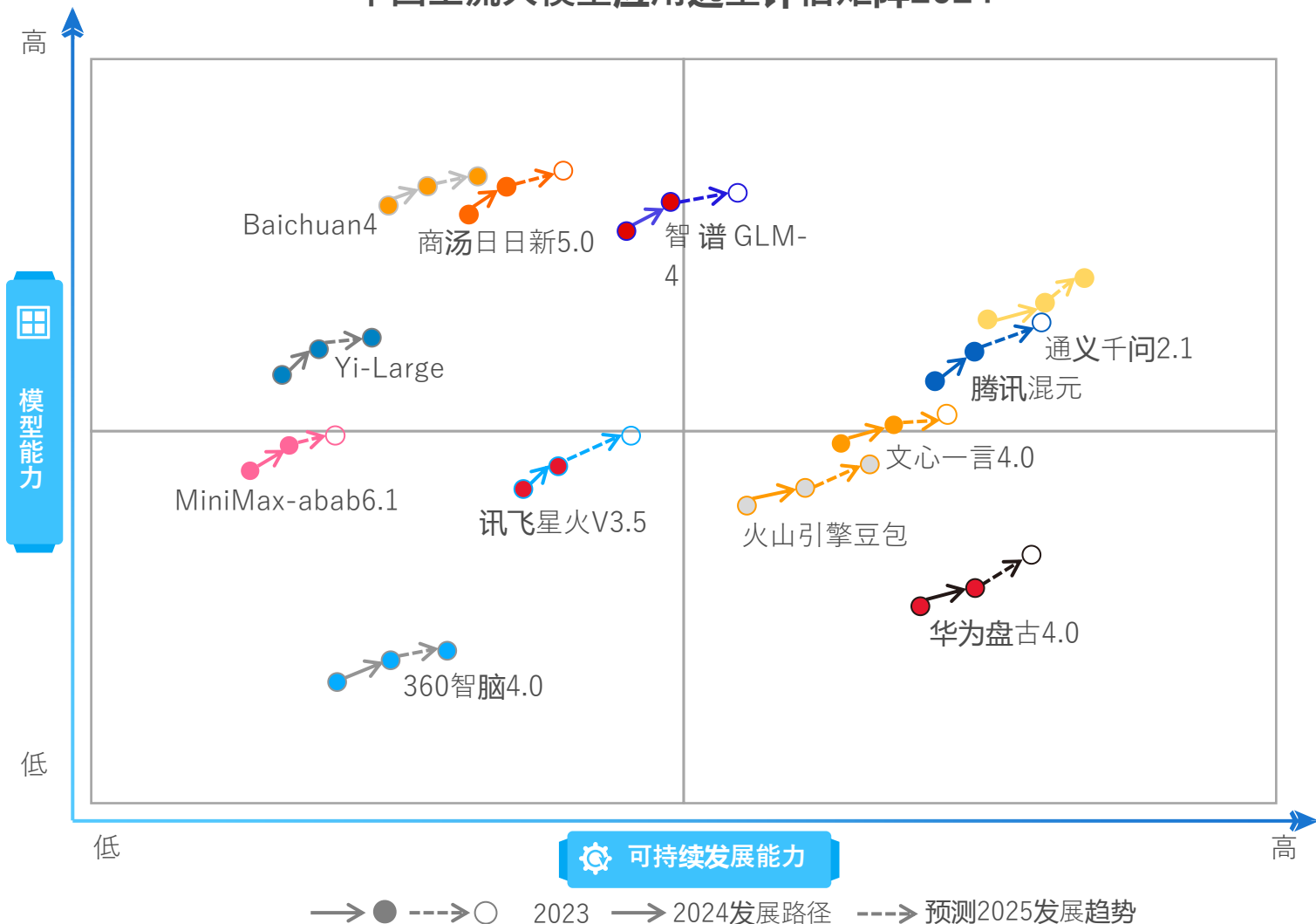
中国主流大模型应用选型评估矩阵2024



- 选取中国大模型12家代表厂商进行评估。本应用选型实力矩阵评估主要聚焦于生态能力的模型应用选型，从模型能力和生态能力两个维度综合反映大模型应用选型市场上主流大模型的市场表现和竞争地位的研究模型。
- 易观分析认为，大模型的生态发展策略和布局日益受到重视，各大模型企业通过构建开放、协同的生态系统，增强其生态发展能力。例如，大模型企业积极与上下游合作伙伴、科研机构 and 行业用户紧密合作，共同开发和推广大模型应用。这样的生态系统不仅提供丰富的应用场景，还促进了技术的快速迭代和优化。

中国主流大模型选型评估——模型能力x可持续发展能力

中国主流大模型应用选型评估矩阵2024



- 选取中国大模型12家代表厂商进行评估。本应用选型实力矩阵评估主要聚焦于可持续发展能力的模型应用选型，从模型能力和可持续发展能力两个维度综合反映大模型应用选型市场上主流大模型的市场表现和竞争地位的研究模型。
- 大模型企业均加大对可持续发展计划的战略和投入。首先，各企业不断增加研发投入，推进技术创新，确保在技术前沿保持竞争力。其次，人才是可持续发展的核心。通过与知名高校和科研机构合作，企业不仅培养大量专业人才，还吸引顶尖专家加盟。最后，各企业制定了详尽的长期战略规划，明确未来的发展方向和目标，以确保企业在激烈的市场竞争中保持持续的创新动力和发展活力。

主流大模型选型评估版本及相关简介

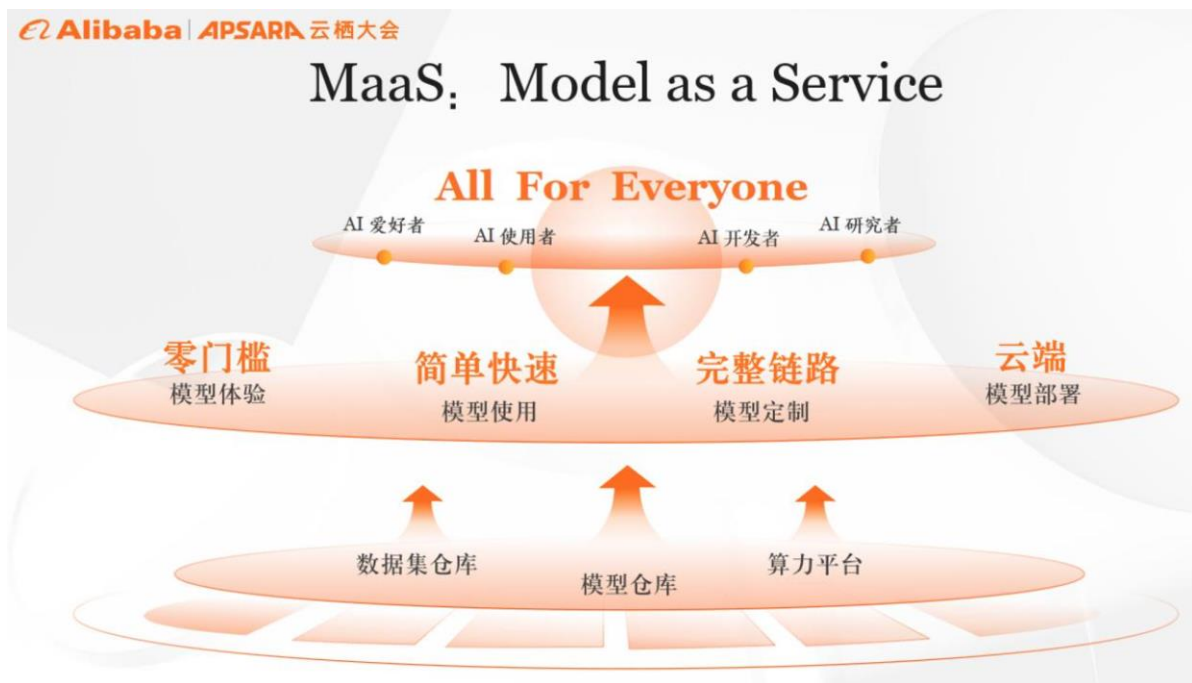
模型	所属企业	发布/迭代时间	应用选型评估模型版本及相关简介
Baichuan4	百川智能	2024年5月22日	百川智能发布最新一代基座大模型 Baichuan 4，并推出旗下首款 AI 助手 “百小应”
腾讯Hunyuan-pro	腾讯	2024年5月17日	腾讯混元官方公开发布的API版本，具备语言理解和生成能力的千亿参数规模的大型语言模型，支持多模态交互
豆包大模型	字节跳动	2024年5月15日	官方发布的API版本，豆包大模型原名 “云雀”，是字节跳动发布的大模型
Yi-Large	零一万物	2024年5月13日	零一万物发布千亿参数规模的大模型 Yi-Large
商汤日日新5.0	商汤科技	2024年4月23日	官方公开发布的SenseChat V5的API版本，是商汤科技推出的新一代大型语言模型，它在多模态理解和生成、知识问答、逻辑推理等方面具有显著的性能提升
讯飞星火V3.5	科大讯飞	2024年1月30日	官方公开发布的星火V3.5的API版本，是科大讯飞推出的一款具有高性能的语言模型
智谱GLM-4	智谱AI	2024年1月16日	官方公开发布的GLM-4的API版本，智谱GLM-4是智谱AI公司开发的千亿参数规模的大型多模态认知智能语言模型
MiniMax-abab6.1	稀宇科技	2024年1月16日	官方小范围内测API版本，版本号为abab6.1，是MiniMax发布的MoE（Mixture of Experts）大语言模型
通义千问2.1	阿里云	2023年12月1日	官方公开发布的API版本，通义千问2.1是阿里巴巴达摩院推出的一款具有多轮对话、知识问答、文本生成等能力的迭代升级版大型语言模型
文心一言4.0	百度	2023年10月17日	官方发布的文心一言V4.0的API版本，文心一言4.0是百度推出的最新版本大型语言模型，在多轮对话、知识问答、文本生成等方面进行了优化和提升
盘古大模型4.0	华为	2023年7月7日	官方发布的API版本，是华为诺亚方舟实验室推出的大型语言模型
360智脑4.0	360	2023年6月13日	官方公开发布API版本，是360公司推出的最新版大型语言模型，具备跨模态生成能力，支持文字、图像、语音、视频等多种数据类型的处理和生成

数据说明：本表格只为本报告分析内容应用选型评估所涉及的大模型做相关说明，表格内排名不分前后。表格范围中的大模型企业为不完全统计，政务、运营商、企业工具、能源、工业等细分行业大模型暂无参加本次应用选型评估。本主流大模型应用选型评估体系中的模型能力参考于多家权威大模型综合评测结果，产品能力、生态能力、可持续发展能力的资料和数据来源于对行业公开信息的分析、对业内资深人士和相关企业深度访谈，以及易观分析师综合以上内容作出的专业性判断和评价。分析内容中运用Analysys易观产业的产业分析模型，并结合市场分析、行业分析和厂商分析，能够反映当前市场现状，趋势和规律，以及厂商的发展现状。

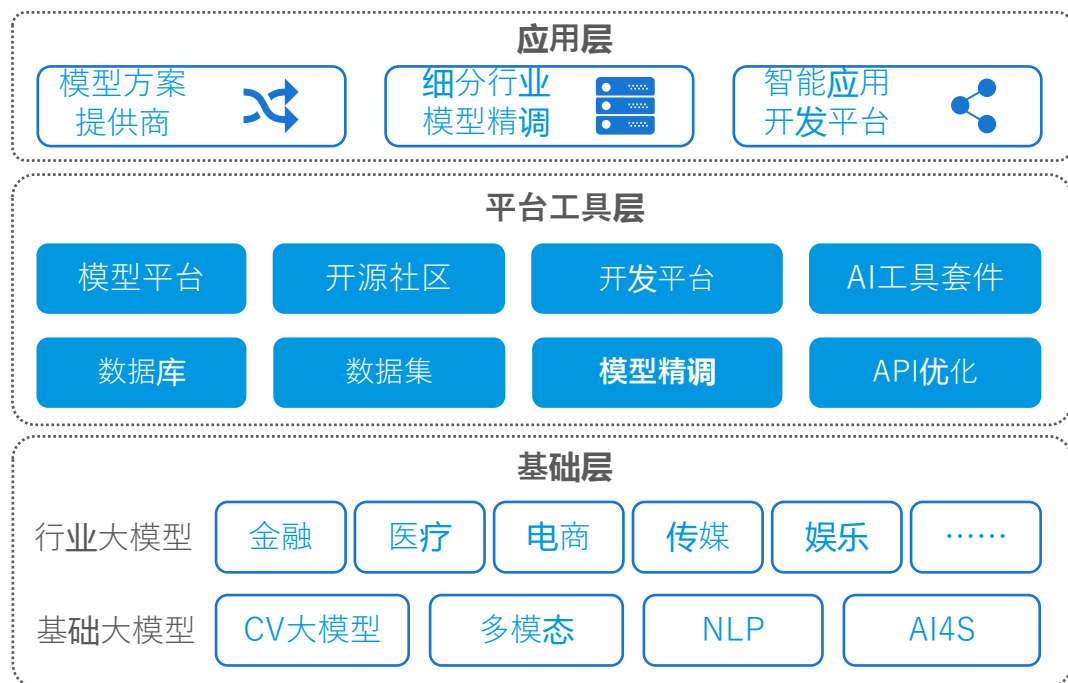
MaaS（Model as a Service）成为主流商业模式，为企业和开发者提供快速构建和部署AI大模型及应用的途径

- 易观分析认为，MaaS业务形态的出现将加速AI大模型的创新和应用开发，从而更便捷地实现行业定制化和个性化应用。
- MaaS为企业和开发者提供了快速构建和部署AI大模型及应用的途径。通过利用云端的弹性计算资源和预训练的模型，开发者能够更迅速地将创意转化为实际的产品和服务。通过MaaS，企业能够规避高昂的建设和维护成本，而只需根据实际使用情况支付费用，这使得小型企业和创业公司也能够充分利用先进的机器学习技术，从而推动更多的创新发生。此外，MaaS不仅能够提供预训练的通用模型，还能够为特定行业或应用场景定制模型，企业和开发者可以根据自身需求对模型进行定制和优化，从而实现更精准和个性化的应用。

阿里巴巴：大模型MaaS解决方案



大模型Maas服务应用的基本架构

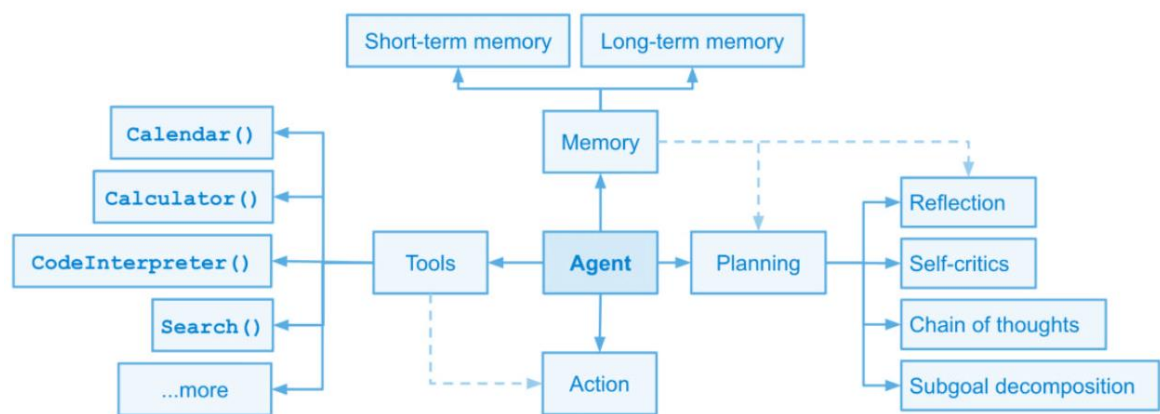


AaaS（Agent as a Service）将帮助大模型最终落地于多元应用场景中

- AaaS（Agent as a Service）通过简化大模型的部署和管理、灵活应对不同应用场景、提高资源利用效率、提供持续更新和维护、加强数据安全和隐私保护以及支持多种交互方式，极大降低了技术门槛，使得大模型能够广泛应用于各种行业和场景，推动其最终落地。

大模型驱动的自主智能体系统架构

- 一个基于大模型的 AI Agent 系统可以拆分为大模型、规划、记忆与工具使用四个组件部分。
- Agent = LLM + 规划技能 + 记忆 + 工具使用的基础架构，其中 LLM 扮演了 Agent 的“大脑”，在这个系统中提供推理、规划等能力。



AI Agent的多元应用场景

应用场景	具体应用实践
个人助理	查找或回答问题，预定票务或管理日历等活动
智能客服	处理客户问题，协助解答问题
推荐系统	Netflix和Amazon等平台使用AI Agent推荐个性化内容
智能监控	安全领域中，AI Agent监控视频流并识别异常
自动驾驶	集成多种AI Agent用语导航、避障和决策
医疗诊断	分析医疗数据辅助诊断，或药物研发、生物蛋白质合成
虚拟助手	如Siri、Alexa和Google Assistant等语音交互服务
金融管理	提供金融建议、风险评估、合规评估或欺诈检测等
智能文档	文档分类、信息分析提取和摘要、翻译等
软件开发	支持应用程序的变成、测试和调试等
任务管理	生成高效的任务并执行

数据来源：Lilian Weng《LLM Powered Autonomous Agents》，易观分析整理

© 易观分析

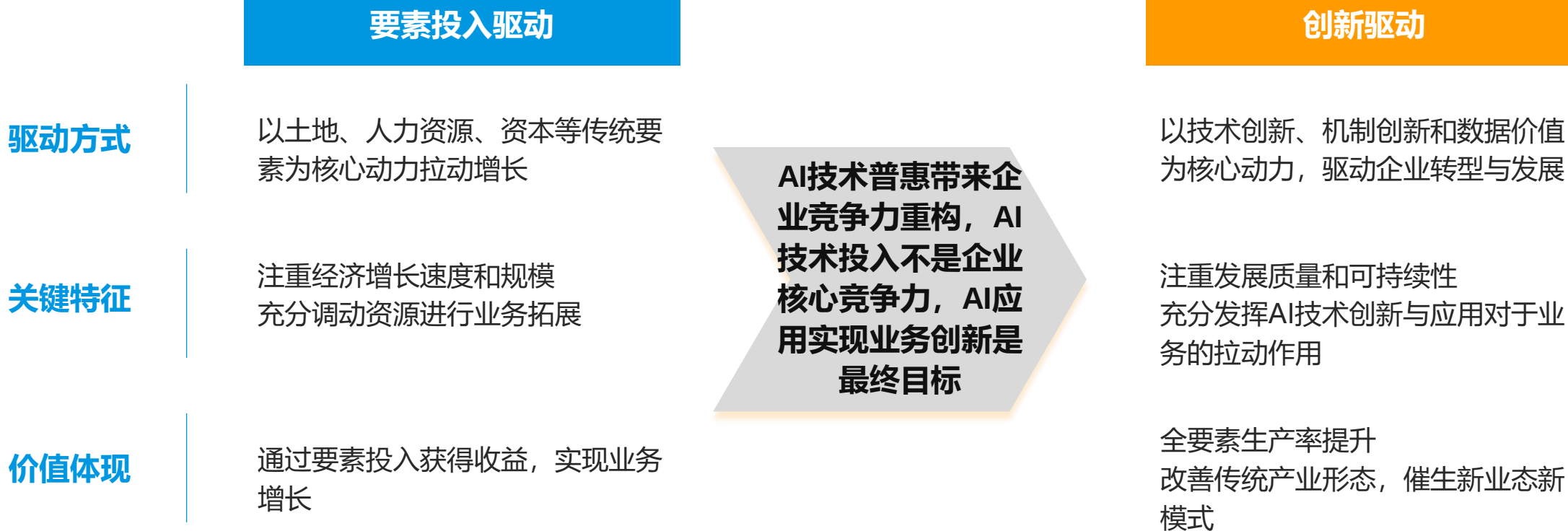
www.analysys.cn

03

企业部署AI的常见误区

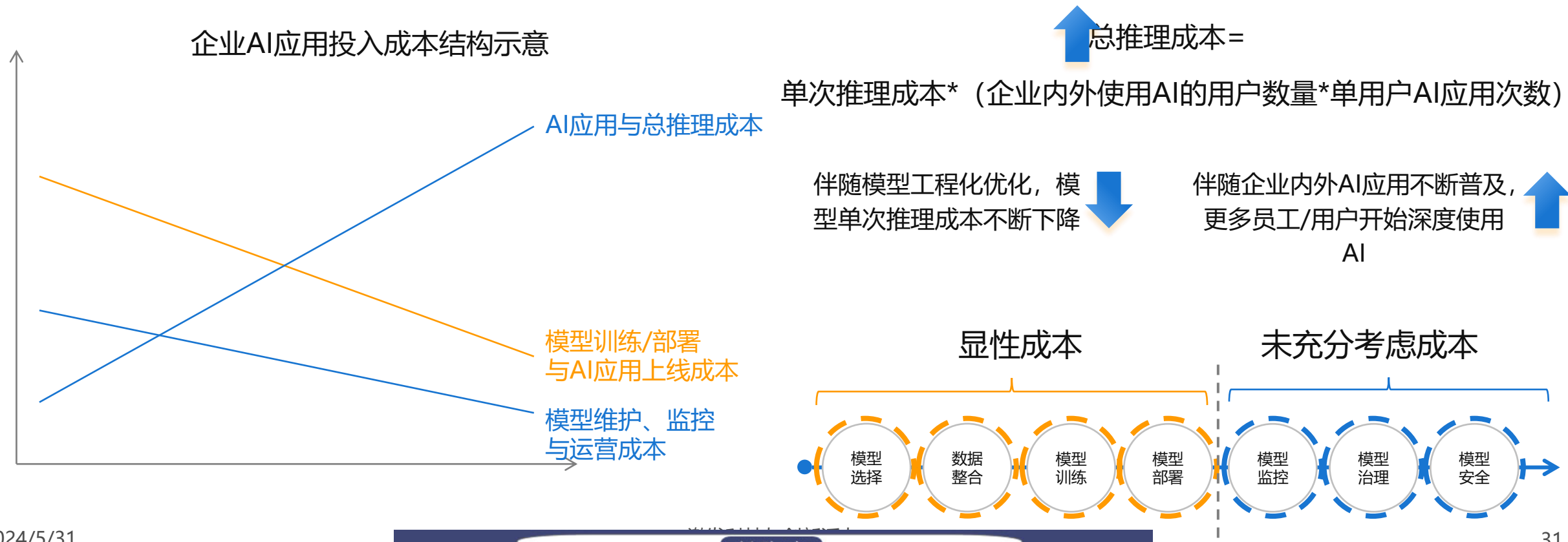
误区一：把手段当做目的，认为AI技术投入直接决定了企业核心竞争力

企业加大AI投入时，应避免仅关注短期效率提升和成本下降。AI的真正价值在于推动业务模式创新和持续发展。企业需从战略层面考虑AI应用，确保其与核心业务和长期目标紧密结合。通过利用AI开发新产品、优化客户体验、创新营销策略，企业不仅能提高效率，还能增强创新能力和市场竞争力，实现创新驱动的发展。企业应将AI视为战略资产，结合商业模式和核心竞争力，确保AI投入带来长期商业价值。



误区二：只关注显性成本，忽视AI应用规模化普及带来的成本投入

企业普遍压缩成本支出的背景之下，AI投入成为少数仍然增长的板块之一，这对于AI技术的快速渗透和价值落地是具有重要意义的。同时，也出现了部分企业在进行AI投入财务测算的过程中，存在着结构性的误区，如下图所示，企业普遍能够充分认知模型的训练与部署，以及生成式人工智能应用上线的部分需要成本投入，但是有可能忽视后续模型与AI应用的过程中发生的成本支出。其一，模型的训练是一个持续的过程，监控模型性能变化进行优化、保护数据集以及由此带来模型安全等问题也是一个持续投入过程；其二，表面上看，模型推理成本不断下降，但是企业需要考虑的是总推理成本与AI应用受益之间的平衡等。



误区三：对AI投入收益预期过快，未充分考量收益的阶段性变化

在当前的市场环境上，企业都希望通过AI带来新的业务增长。但是AI部署后短期内还是先带来效率的提升以及人员成本的节省（如果考虑训练成本可能不能给企业真正节约成本）。真正可以让企业感受到业务变化的收入增加、满意度提升等最快都要2-3年。企业部署AI要有合理的收益预期。



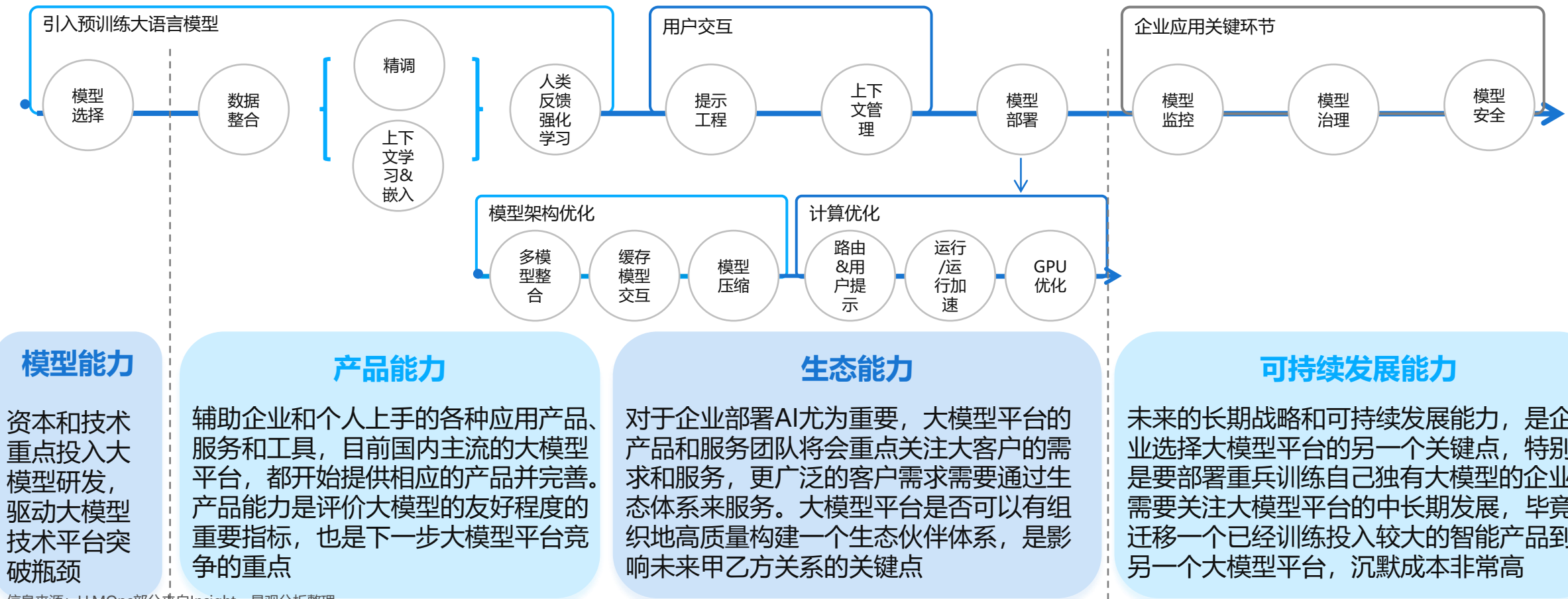
- 提高效率：AI可以自动化许多任务，从而提高企业运营的效率。
- 成本节约：通过自动化和优化流程，企业可以减少人力成本和其他相关成本

- 收入增长
- 客户满意度提升
- 竞争优势：AI技术的应用可以为企业带来竞争优势，尤其是在快速变化的市场中。
- 增强决策：AI可以提供数据分析和洞察，帮助企业做出更明智的决策。

- 创新业务和新产品开发：AI可以促进新产品和服务的开发，推动企业创新。
- 市场领导地位：长期而言，AI可以帮助企业建立或巩固其在市场中的领导地位。
- 可持续发展：AI有助于企业实现可持续发展目标，通过优化资源使用和减少环境影响。
- 重塑行业

误区四：将模型能力作为唯一标准，忽略AI应用落地是系统工程

大语言模型与生成式人工智能应用落地是系统工程，相应地，在如下各个环节都需要进行提前规划和设计，才能在引入AI应用部署到生产环境之后也能够持续稳定安全输出价值。但是，企业普遍目前在AI规划与合作伙伴选择的过程中，过于聚焦模型能力本身，而可能忽视合作伙伴在产品封装和体验、生态能力建设等方面的考虑，需要一并纳入合作伙伴选择和评估的范畴。



信息来源：LLMOps部分来自Insight，易观分析整理

误区五：在AI应用的风险控制方面走极端，或忽视或过于谨慎

企业在部署人工智能（AI）时，若忽视潜在风险，可能会急于将未经充分测试的系统投入实际应用。这样做不仅可能导致数据泄露、产生偏见结果，还可能触犯行业合规性要求。相反，如果企业过于关注风险，可能会陷入分析瘫痪，导致犹豫不决，错失AI带来的巨大机遇。因此，成功部署AI的关键在于找到一个平衡点，既要充分识别和管理潜在风险，又要敢于把握AI带来的机遇。这包括进行彻底的风险评估、确保数据安全和隐私保护、建立透明的AI治理框架，以及在部署前进行充分的测试和验证。通过这样的方法，企业可以确保AI系统的可靠性和合规性，同时充分利用AI的潜力来推动业务增长和创新发展。

模型安全与可控制问题

- 模型的安全性和可控制性，关键在于确保其在遵守安全规范和法律法规的同时，满足不同行业风险管理和用户安全需求。这是AI领域的主要挑战，目前通过安全训练部署流程、实时监控干预机制和增强技术来提升。

大模型的对齐问题

- 对齐问题，既包括人工智能与人类社会价值观保持一致，也包括与不同国家价值观，不同类型企业经营以及商业法则相匹配等，前者最为关键，这也是目前最为关注的AI不受控制的风险之一，目前在通过RLHF与RLAIF不同方式来实现。

模型的可解释问题

- 模型的可解释性，不仅包括使人工智能模型的理解和决策过程对人类用户透明，还涉及与不同领域的专业知识、不同类型用户的需求以及应用场景的具体要求相匹配。这也是目前人工智能领域面临的主要挑战之一。目前通过直观模型结构、决策可视化技术和增强可解释性来实现。

信息的知识产权、数据隐私和数据安全问题

- 模型应用中的数据隐私和数据安全，关键在于确保AI模型在处理数据时遵守隐私保护和安全规范，同时满足不同行业的数据管理需求和用户隐私期望。这是AI领域的主要挑战，目前通过加密技术、数据访问控制和隐私增强技术等措施来实现。

AI部署建议：业务为纲，机制兜底，关注与AI驱动相适应的组织能力建设

01 业务为纲规划人工智能上线与推广计划

“所有行业都值得用大模型重新做一遍”，前提是围绕用户与客户价值的体验升级与业务发展，相应地，人工智能与企业数智化转型一致，业务驱动是核心原则，围绕业务发展扫描数智化洼地，规划人工智能应用用例，获取业务价值，并形成迭代优化扩大AI应用范围的闭环

02 专有数据资产沉淀与管理，应对模型训练与应用

无论是自主训练大语言模型，还是围绕自身行业与业务场景进行模型精调，都需要依赖于企业过往积累的专业领域知识沉淀，进行专有数据语料的准备，从而能够让基础模型的“通才”能力发展成为具备行业属性的“专才”，专有数据是未来模型能力平民化后的重要差异化

03 设定人工智能应用合规与风险管理防范机制

生成式人工智能仍然可能出现“幻觉”、数据泄露等各种风险；同时，关于人工智能与大模型应用方面的立法与规范仍然处于意见征集与调研的过程中，企业有必要主动制定人工智能应用合规与风险管控机制，包括AI开发、应用与审核规范，数据安全规范，员工应用权限规范等等，从而“安全合规”地应用人工智能提升企业竞争力

未来组织能力围绕人工智能发展，既包括创造人工智能，也包括应用人工智能，前者是少数，主要在于提升人工智能专业与高级人才密度，并通过组织协同设计，系统化地提升人工智能科研与工程化水平。

后者是多数，也就是大部分人工智能企业更需要考虑的是，人工智能应用，尤其是生成式人工智能全面铺开过程中，对于组织架构、组织中关键角色与职能、以及对于员工的潜在影响等。

对于个体而言，普遍关注“人工智能会/不会替代什么职业”，甚至引发了对于某些职业发展方向的焦虑，一方面，积极的员工与个体正在迅速拥抱人工智能，在工作的过程中应用生成式人工智能提升工作效率，可能正在出现人工智能赋能于员工优先于组织的情况；另一方面，大众员工可能出现观望、等待甚至无所适从，“无用”内卷的状况，。

这就需求企业对于组织角色与员工发展进行整体规划，不同职能与角色人与人工智能的协同边界如何确定，组织中关键角色与员工的技能与“AI商”升级如何系统化通过培训等手段推动，都是企业系统化提升组织能力与竞争力的关键举措

组织系统协同员工能力协同进化与升级 04

	考虑角度	自测问题
战略与目标	AI的战略目标？	- 企业部署AI的目标是什么？（行业示范？市场竞争力？财务表现？产品创新？成本最优？技术领先？……） - 企业计划的投入是怎样的？ - 企业的收益预期？（一年？三年？五年？……）
场景/产品/技术成熟度	场景成熟度/产品成熟度 技术与应用成熟度	- 客户需求是否明确且稳定？产品/场景是否稳定运营？产品的收入规模有多少？处于生命周期的什么阶段？ - 流程是否标准化？是否有数据积累？人员的水平？… - 技术成熟度和最佳实践案例 - 行业应用成熟度和最佳实践案例 - 企业应用成熟度和最佳实践案例
资源就绪度	数据 技术与基础设施	- 数据流程评估（预处理保证了数据去重、标准化、格式化等；数据提炼避免模型泛化等） - 数据类型评估（数据类型决定了数据的可用化程度/语义化程度） - 知识管理工程（数据分类、知识图谱、工作流程等） - 数据安全与合规等自检（重点） - 计算资源、存储与网络是否准备好 - 推理系统（包括推理优化软件、自动化部署系统等）是否部署 - 监控与运维、技术支持与培训是否安排 - MaaS的成本管理和预测是否完成？
组织与团队	文化与愿景 组织能力 单兵能力	- 企业的文化宣传是否到位？相关的预热工作是否开展？ - 相关团队的人员组成、架构设计？团队的技能水平？ - 企业的工作流程？考核激励机制？ - 企业配备的基础资源和工具是否到位？ - 企业相关的人员培训是否开展？ - 员工对于AI战略的接受度是否达到一定水平？

激发科技与创新活力

易观



数字经济全景洞察

易观千帆



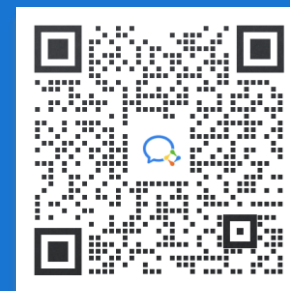
金融行业数字用户
同业对标

易观方舟



全场景私域用户运营

易观社群



加入易观社群
与行业伙伴一同交流