第一新声智库
THE FIRST NEW VOICE

2024年中国AI大模型产业发展与应用研究报告

第一新声研究院 | 产业研究报告

THE·FIRST·NEW·VOICE

研究范畴及目的

2018年，OpenAI发布第一个生成式预训练模型GPT-1，模型参数量达到1.17亿， 开启了AI大模型的训练热潮；2022年底，ChatGPT的火爆，吸引了全世界对大模型的广泛关注，也激发了中国**AI大模型发展热潮**。2023年3月中旬，百度发布大语言模型产品文心一言，拉开了国内AI大模型产业 **“百模大战”**的序幕，之后通义千问、盘古大模型、星火认知大模型、360智脑、豆包等纷纷问世。2024年中国AI大模型产业落地明显加速，AI大模型产品化、商业化和产业化的发展脉络基本形成。

AI大模型是指拥有亿级以上参数的**深度学习模型**，从应用场景角度可分为通用大模型和垂直大模型，其中垂直大模型又可以分为行业大模型和垂直场景大模型。通用大模型，聚焦基础层和技术攻关；垂直大模型，聚焦垂直领域解决方案，在通用大模型基础上开发行业和场景专用模型，面向政务、金融、医疗、教育、交通等垂直行业和营销、客服、运营等通用场景。

2024年以来，我国AI大模型的产业应用已经迅速展开，本报告通过调研中国AI大模型产业发展和市场应用，梳理AI大模型的发展历程、产业生态、商业模式、行业应用、市场规模、发展趋势等，并对典型垂类行业和场景应用案例进行剖析，以期达到以下目的：

- 通过研究AI大模型产业的发展历程、产业生态、商业模式、行业和场景应用等，帮助企业了解AI大模型产业的发展脉络和发展方向；
- 重点分析AI大模型产业的市场规模、发展驱动、市场应用、发展趋势，帮助AI大模型企业把握市场需求动向和发展趋势；
- 深入调研AI大模型的典型行业和场景应用，挖掘行业里应用相对成熟、落地价值度高、市场反馈良好的优秀服务商与实践案例，为客户产品选择提供参考。

目录 CONTENTS

01 研究背景

1. AI大模型可按照应用领域和输入数据类型进行分类
2. 至2024年中国AI大模型产业发展加速进入商用阶段

02 发展现状

1. TOP5大模型应用行业：互联网、金融、医疗、教育、政务
2. 国产大模型已大幅降价，为广泛商业化应用奠定基础
3. 中国AI大模型产业图谱
4. AI大模型商业化落地：55%为定制化模式、40%-45%为API及订阅模式
5. 五种部署方式：直接调用、能力嵌入、扩展应用、定制模型、全栈构建
6. 2024年AI大模型应用市场规模约157亿元，2022-2027年复合增长率达148%
7. 未来3年，中国AI大模型产业将逐步进入需求侧驱动阶段

03 应用场景及案例

1. AI大模型在互联网、政务、金融等行业应用场景日益丰富
2. 金融行业痛点及解决方案
3. 医疗健康行业痛点及解决方案
4. 教育行业客服场景痛点及解决方案
5. 政务领域市场监督管理场景痛点及解决方案
6. 零售消费行业痛点及解决方案
7. 制造行业知识管理场景痛点及解决方案

04 发展趋势及挑战

1. AI大模型四个主要技术方向
2. 融合应用软件、智能助手和AI Agent是AI大模型市场应用的三个主要方向
3. 缺乏高质量数据集是大模型商业落地面临的关键挑战

PART ONE

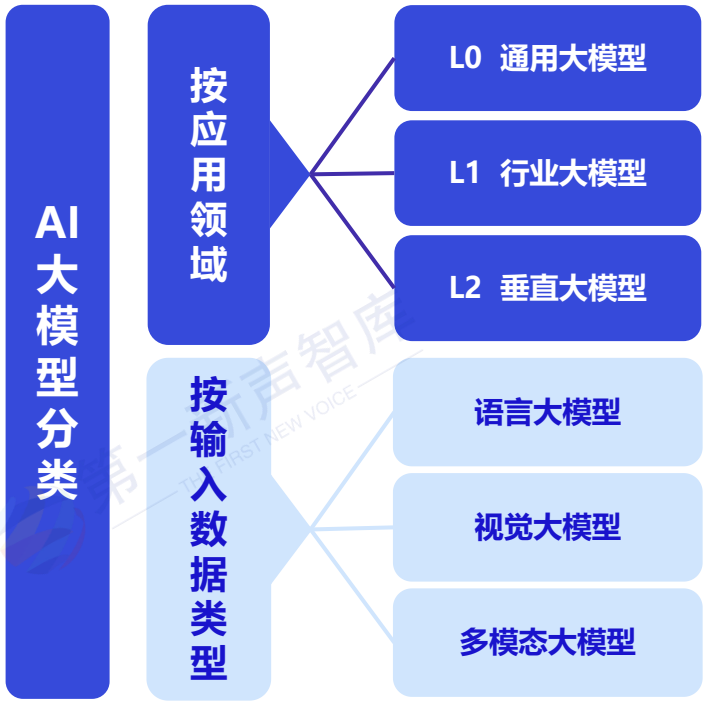
研究背景

01

定义及分类

AI大模型可按照应用领域和输入数据类型进行分类

AI大模型是指拥有**亿级以上参数**的深度学习模型。AI大模型利用深度学习算法和人工神经网络技术等AI技术，通过学习大量的数据提升预测能力，其性能与模型的**参数规模、数据集大小**和训练用的**计算量**之间存在幂律关系；AI大模型基于注意力机制，通过在大规模、多元化的无标注数据集进行训练，具有较强泛化能力，应用在广泛的场景和任务。按输入数据类型，AI大模型分为**语言大模型，视觉大模型、多模态大模型**等；从应用领域角度分类，AI大模型分为**通用大模型、行业大模型、垂直大模型**。



L0 通用大模型	在多个领域和任务上通用，具有基础的认知能力，各学科知识都懂一点；通用大模型就像完成了大学前素质教育阶段的学生。
L1 行业大模型	针对特定的行业或领域的大模型，通常使用行业相关的数据进行预训练或微调，来提高在该领域的性能和准确度；行业大模型就像选择了专业的大学生，对所在专业领域的知识有了更深入的了解。
L2 垂直大模型	针对特定任务或场景的大模型，使用更加专业的知识和数据进行预训练和微调，在专业任务上的性能和效果更优；垂直大模型就像研究生，在所在具体领域有了比较深入的研究。
语言大模型	通过处理大量文本数据来理解和生成人类语言的大模型，执行自然语言处理任务，如文本分类、问答、对话、摘要总结等；常见的语言大模型包括ChatGPT、文心一言、讯飞星火等。
视觉大模型	是一种基于大规模数据和强大计算能力训练的深度学习模型，通过模拟人类视觉系统的工作原理，通过复杂的神经网络结构，实现对图像和视频数据的高度识别、分析和表达的效果；常见的视觉大模型包括智谱*CogVLM*、书生·视觉大模型等。
多模态大模型	多模态大模型指的是将文本、图像、视频、音频等多模态信息联合起来进行训练的模型；典型的多模态大模型包括腾讯的hunyuan-vision、商汤科技的日日新等。

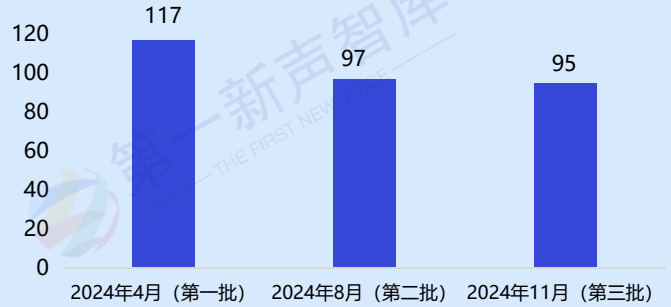
至2024年中国AI大模型产业发展加速进入商用阶段

1956-2006年，深度学习和神经网络技术的提出和发展，为AI大模型的出现奠定了技术基础，**大模型技术萌芽**开始出现；2006年后自然语言处理技术、**Transformer**架构的发展，为**大模型预训练算法技术和架构**奠定了基础；2018年OpenAI和Google分别发布**GPT-1**与**BERT，**预训练大模型成为自然语言处理领域的主流；2022年底，OpenAI推出ChatGPT引发全球大模型发展热潮，2023年中国国内大模型训练开始井喷，出现**“百模大战”**现象；2024年中国政策加大行业落地推动力度，商业发展加速。

- **1956年：**约翰·麦卡锡首次提出“人工智能”概念，标志着AI领域的诞生
- **1998年：**法国学者Yann LeCun等人构建LeNet-5，标志着机器学习从浅层模型向**深度学习模型**的转变，为自然语言处理和计算机视觉等领域的研究奠定了基础
- **2013年：**Word2Vec模型诞生，提出将单词转换为向量的“词向量模型”，推动了自然语言处理的发展
- **2014年：**对抗式生成网络（GAN）诞生，标志着深度学习进入了生成式模型研究的新阶段
- **2017年：**Google提出了基于自注意力机制的Transformer架构，为大模型的预训练算法架构奠定了基础
- **2018年：**OpenAI和Google分别发布了GPT-1与BERT，标志着预训练大模型成为自然语言处理领域的主流
- **2020年：**OpenAI推出了GPT-3模型，参数规模达到了1750亿，在零样本学习任务上实现了巨大性能的提升

- **2023年：**GPT-4发布，具备了多模态理解与多类型内容生成能力，进一步推动了大模型技术的发展
- **2023年：**中国掀起“百模大战”，发布各类大模型数量超过100个，涵盖通用大模型、行业大模型、基于通用大模型或行业大模型的应用服务型大模型等

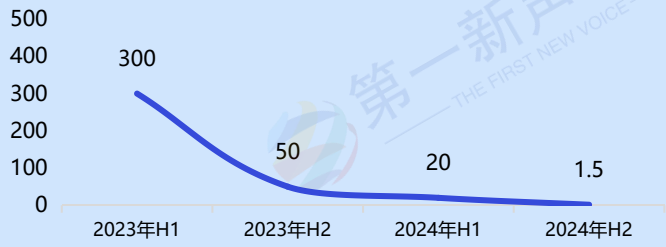
2024年生成式人工智能服务备案数量



数据来源：第一新声研究院整理

- 2024年1-7月，中国央国企采购大模型项目数量已超过950个
- 截至2024年11月，中国获得备案的大模型数量达到309个，中国大模型开始在垂类行业众多场景落地
- 截止2024年底，中国大模型产品使用价格进一步下降，为大模型广泛商用落地提供了基础

2023-2024年国产大模型价格走势（元/百万Tokens）



数据来源：第一新声研究院整理

1956 **技术萌芽与技术探索期** 2023 **AI大模型训练井喷期** 2024 **AI大模型加速商用期**

PART TWO

AI大模型产业发展现状

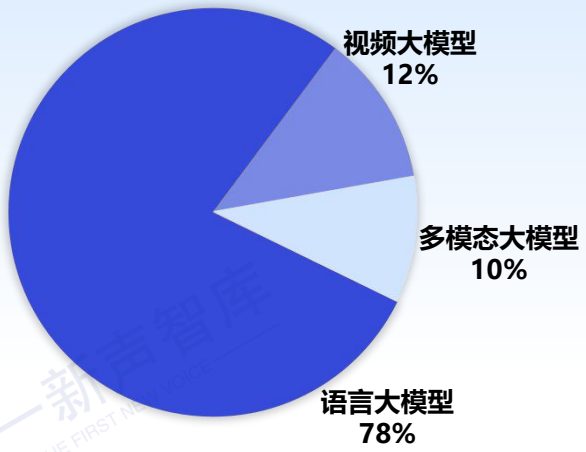
02

大模型发展与应用结构

TOP5大模型应用行业：互联网、金融、医疗、教育、政务

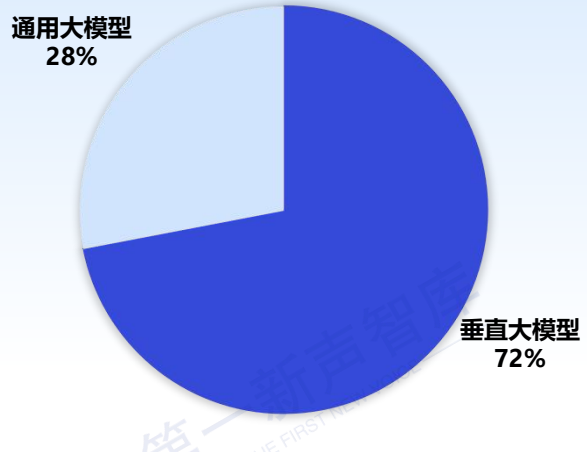
我国AI大模型需要依照《生成式人工智能管理暂行办法》进行备案，截止到2024年11月，已有3批次共计309个大模型通过国家互联网信息办公室备案。从应用类型看，通用大模型占比28%，垂类大模型占比72%；从应用领域看，互联网行业、金融行业、医疗行业、教育行业、工业行业大模型占比均超过10%。

------ 模态结构 ------



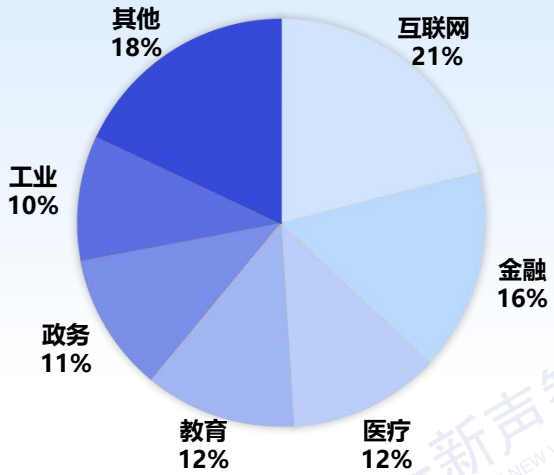
洞察一：得益于语言大模型的能力和丰富的应用场景，语言大模型的数量明显多于其他模态；随着大模型多任务适应性和多模态能力持续增强，对齐视觉特征和文本特征，实现跨模态的统一理解，并根据指令创造新的内容或增强现有数据表达，成为当前大模型创新的焦点。

------ 类型结构 ------



洞察二：闭源大模型的商业化输出和开源大模型的涌现，推动大模型落地门槛持续降低，垂直行业成为创新的重要赛道，吸引了产业资本支持；2024年，我国垂直大模型涌现，快速向垂直行业渗透。

------ 行业结构 ------



洞察三：垂类大模型覆盖互联网、金融、医疗、教育、政务等众多行业；目前垂类大模型行业用户对知识助手、智能客服、智能营销、编码助手等应用接受程度高，同时对数据分析、办公助手等应用抱有较高期待。

数据来源：第一新声研究院整理

国产大模型已大幅降价，为广泛商业化应用奠定基础

当前AI大模型产品进化路线有两条，一是通过增加模型参数量、扩大数据集、提升训练计算量来获得性能更强大的大模型产品；二是通过**优化模型架构适应性**和**计算效率**，获得更具**性价比**的产品，如**70B参数的模型**，通过优化架构和训练策略，可获得**接近或超越更大规模模型**的性能。随着大模型能力和性价比的提升，国产大模型厂商开始大幅降价，截止到2024年底，我国典型AI大模型的价格下降至0.5元/百万Tokens以内，为大模型应用的广泛落地打下了基础。

中国AI大模型代表厂商及最新大模型

AI大模型代表厂商	大模型版本	模型特点
智谱华章	GLM-4-Plus	4050亿参数，具备全面的语言理解能力，更好的智能遵循能力和高质量数据构造能力，结合跨模态能力和实时推理能力，适用于聊天机器人、内容创造、教育辅导、多模态交互等场景
字节跳动	豆包通用模型pro	8000亿参数，基于DIT架构，在图像和语音处理方面性能突出，其视频生成能力可帮助用户提供专业级视频内容
阿里巴巴	Qwen2.5-Turbo	上下文处理的长度极限获得突破，推理速度显著提升，适用于文本理解、代码处理等场景，兼具高效性与低成本优势
百度	文心大模型4.0 Turbo	训练速度相比上代产品提升数倍至数十倍，生成内容更具条理性，输入输出价格下降70%，适用于智能客服、智能助手、内容创作、数据分析等场景
深度求索	DeepSeek-V3	参数规模6710亿,采用了混合专家架构（Mixture-of-Experts, MoE），这种架构使得模型在有限硬件资源下仍能实现高性能
科大讯飞	讯飞星火4.0 Turbo	突出的数学能力和代码能力，新增多模态视觉交互及超拟人虚拟人交互功能，适用于智能写作、多模态交互、医疗、教育、司法和政务服务等场景

中国典型AI大模型产品最新价格

公司	模型名称	价格(元/百万Tokens)	
		输入	输出
智谱华章	GLM-4-Air	0.5	0.5
字节跳动	Doubao-vision-lite-32k	1.5	4.5
阿里巴巴	Qwen-vl-plus	1.5	4.5
百度	ERNIE 3.5	0.8	2
腾讯	hunyuan-standard	0.8	2
科大讯飞	Spark Pro	0.5	0.5

产业图谱

2024年中国AI大模型产业图谱

商业化模式

AI大模型商业化落地：55%为定制化模式、40%-45%为API及订阅模式

AI大模型商业化形式

商业模式	使用/部署方式	适用领域
定制化（55%）	■ 本地化部署：软硬件一体，提供预训练和微调等服务（80%在B端）	■ 适用于党政、金融、能源、工业等行业的大中型企业和组织机构；如中广核引入科大讯飞大模型平台，采用本地化部署方式
	■ 云部署：提供算力服务和大模型预训练及微调服务	
API及订阅（40%-45%）	■ SaaS、PaaS、MaaS等方式调用服务，按流量、Tokens、产出内容、时间等方式计费	■ 适用于电商、医疗、教育等中小型用户；如腾讯云MaaS服务用户已超千家
广告（＜5%）	■ 嵌入智能终端或APP，用户免费使用，向广告主收取广告费	■ 适用于智能终端等面向大规模人群的ToC场景； 如Kimi已开始涉足大模型的广告投放业务

来源：第一新声研究院

AI大模型市场应用的商业化模式逐渐清晰：

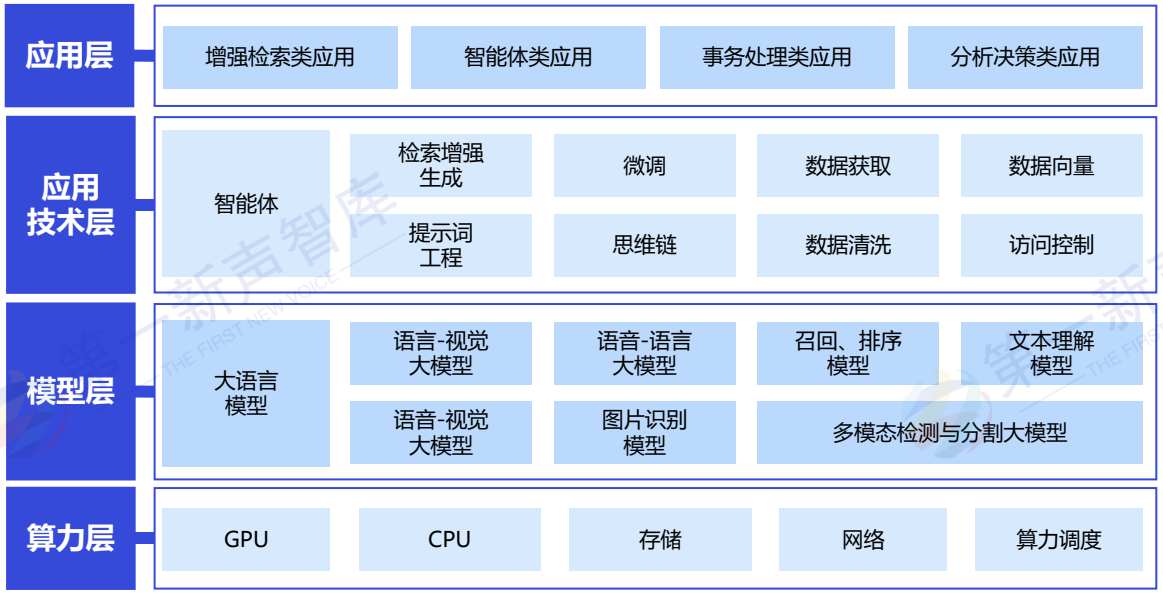
- **定制化模式面向大型政企：**大型企业AI大模型应用时，更倾向于定制化，并采用本地化部署模式；例如浦发银行最新招标的2024年大模型应用体系建设项目，采购定制化算力设备和大模型软件，且要求算力设备和大模型软件满足信创要求。当前，大模型技术和产品迭代迅速，定制化模式下，客户会要求大模型服务商定期迭代更新服务。
- **API及订阅模式适用于中小企业及机构：**采用API及订阅模式采购大模型服务，具有节省资源、快速集成、实时更新和可扩展性强等优势，适用于中小型企业；如豆包大模型等采用API对外提供问答等服务。此外，企业用户调用API需要注意数据安全和隐私保护。
- **嵌入智能终端和APP中收取广告费的模式逐渐落地：**随着大模型应用的普及，将大模型嵌入智能终端和APP中，向广告主收取广告费的模式，将成为大模型产品变现的重要方式，未来大模型应用将成为互联网广告提升和优化流量的重要抓手。2024年11月，苹果与百度达成合作，国行版iPhone接入百度最新的AI大模型Ernie 4.0，iPhone将能够提供更精准、更智能的语音助手、图像识别和数据处理等功能，将拉开大模型规模化嵌入智能终端、收取广告费的序幕。

AI大模型应用落地路径

五种部署方式：直接调用、能力嵌入、扩展应用、定制模型、全栈构建

大模型应用需求落地一般分为四个阶段：（1）**场景需求评估：**评估企业当前的大模型技术、应用场景和能力，做好大模型应用落地的准备，包括技术能力评估、应用场景梳理、能力分析等。（2）**部署能力建设：**设计和构建符合战略规划和业务需求的大模型能力体系，包括大模型建设方案设计、系统研发和功能测试、数据与算法准备等。（3）**大模型应用部署：**将大模型部署到具体的业务场景中，提供定制化的智能解决方案，实现大模型的商业化应用，包括定制化优化与应用开发、效能评估与闭环管理、全生命周期管理等。（4）**大模型运营管理：**建立大模型运营管理体系，保障大模型的长效运行，并通过实时监测和反馈机制提升运营效率，包括实时监测与动态追踪、持续优化与管理体系完善等。

AI大模型架构图

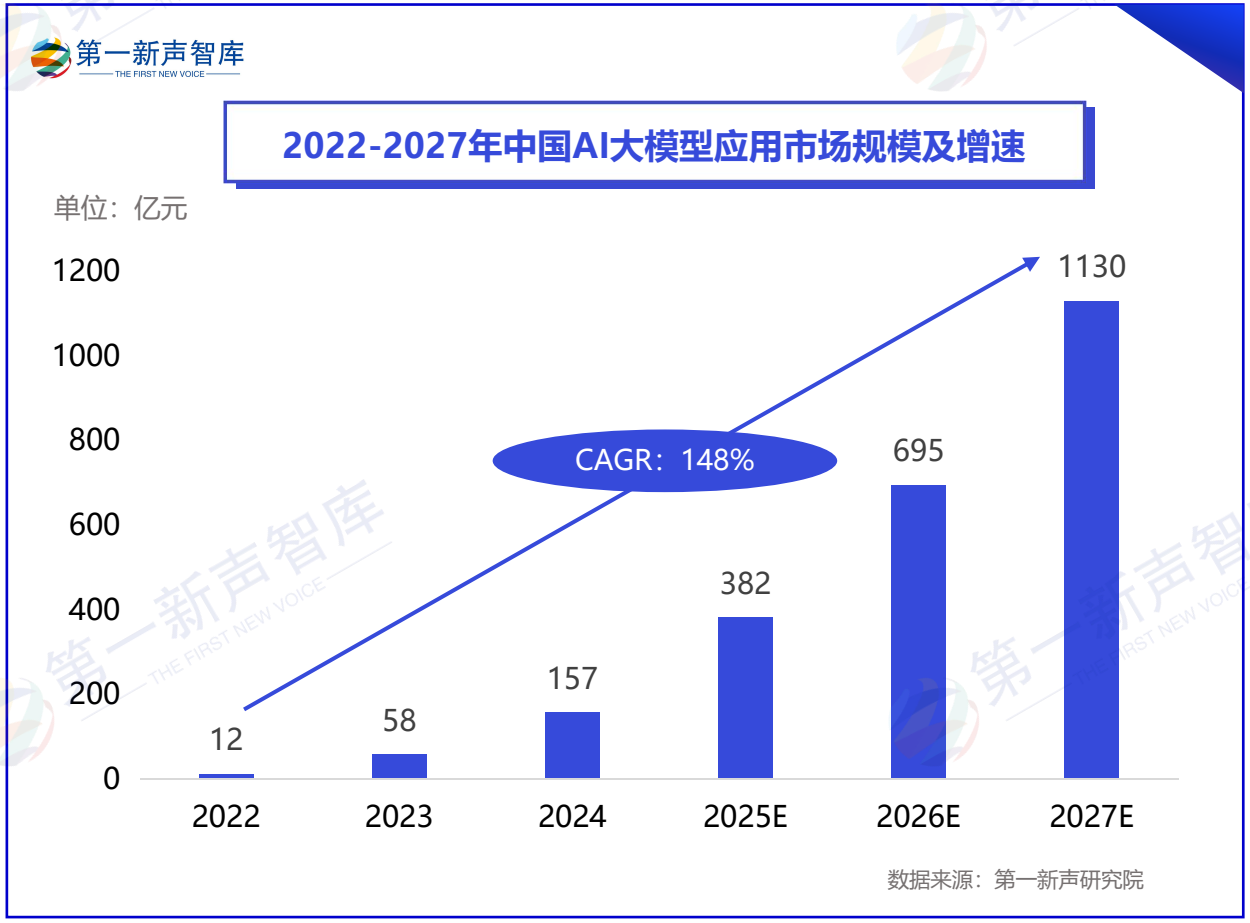


大模型应用部署方式

全栈构建（定制大模型）	定制模型	扩展应用	能力嵌入	直接调用
应用软件	应用软件	应用软件	应用软件	应用软件
大模型工具链（数据检索、提示词工程等）	大模型工具链（数据检索、提示词工程等）	大模型工具链（数据检索、提示词工程等）	大模型工具链（数据检索、提示词工程等）	大模型工具链（数据检索、提示词工程等）
AI大模型（基础模型及微调）	AI大模型（基础模型及微调）	AI大模型（基础模型及微调）	AI大模型（基础模型及微调）	AI大模型（基础模型及微调）
算力基础设施	算力基础设施	算力基础设施	算力基础设施	算力基础设施

重量部署 ⟷ **轻量部署**

2024年AI大模型应用市场规模约为157亿元，2022-2027年复合增长率达148%



根据第一新声智库研究，2022-2027年中国AI大模型应用市场规模复合增长率将达到148%，至2027年，AI大模型市场规模将达到1130亿，AI大模型行业达到盈利临界点。

- **洞察一：** 2024年，中国AI大模型商用加速，根据第一新声研究院不完全统计，2024年公开的大模型中标项目超过1000个，整体应用市场规模将达到157亿，市场用户主要以定制化和API调用模式为主；大模型应用市场规模包括企业用户购买大模型产品、大模型服务、大模型应用服务和软硬一体化大模型应用平台形成的市场总量。
- **洞察二：** 中国AI大模型商用主要通过三种通路，一是大模型厂商直接向终端用户提供模型产品和部署服务；二是大模型服务商通过系统集成商，将大模型产品以行业解决方案的形式提供给用户；三是大模型服务商通过API接口，通过企业级软件服务商或直接向用户提供服务。系统集成商和企业级软件服务商将成为AI大模型提升产业渗透的关键驱动方。
- **洞察三：** 2024年，中国大模型产业新增GPU需求量超过190万张，算力投资达千亿规模，其中80%的新增算力用于头部互联网大模型训练和自有业务支撑，20%用于行业用户大模型能力建设和对外提供MaaS服务，2024年，服务商大模型业务收入，主要用于覆盖算力成本，尚未实现盈利，整体市场规模约157亿元，预计至2025年后将有部分大模型企业开始盈利。

未来3年，中国AI大模型产业将逐步进入需求侧驱动阶段

根据第一新声智库调研，当前大模型市场还处于产品供给驱动为主的阶段，预计未来3年，随着大模型应用被行业用户广泛接受并积极推动应用场景落地，中国大模型市场发展将进入以最终用户需求驱动为主的阶段。

产品供给驱动

通用大模型厂商

- 研发和训练能力更为强大的通用大模型产品
- 例如：字节跳动不断迭代豆包大模型，2024年5月新版本豆包大模型的综合性能提升了20%以上，且还在不断升级

垂类大模型厂商

- 聚焦特定行业或领域，提供更高的准确性和行业属性大模型
- 例如：蜜度-蜜巢大模型，聚焦政务与媒体等领域提供大模型解决方案

企业级应用厂商

- 将大模型能力融入应用软件，或基于大模型开发企业级应用
- 例如：SaaS服务商，将大模型集成到现有的应用程序或服务中，提供更智能的体验，如智能客服、智能营销、知识管理等

最终用户需求驱动

国央企

- 2024年2月，国资委提出中央企业要“开展AI+专项行动”，加快建设一批智能算力中心，进一步深化开放合作，构建一批产业多模态优质数据集，打造从基础设施、算法工具、智能平台到解决方案的大模型赋能产业生态；
- 根据第一新声智库调研，未来3年，每家央企每年在大模型领域的投入普遍超过2亿元。

政府

- 政府主导的大模型采购主要分为两个部分，一是政务大模型平台，用于提升政务效率；二是政府主导的区域智算中心配套大模型平台，用于支持区域企业智能化转型，凝聚新型产业发展形态；
- 预计未来3年，以地市政府为主导的大模型采购规模将超过百亿/年，将极大推动AI大模型产业发展，尤其是满足信创要求的大模型。

其他行业用户

- 金融、工业、教育、医疗、交通等领域大模型的落地场景需求丰富程度高，头部用户具有建设软硬一体的大模型能力平台强烈诉求，中小用户对大模型抱有较高期待，倾向于采购MaaS和SaaS模式赋能业务；
- 未来1-2年，政策和需求牵引下，大模型应用将逐渐从客服、营销、知识管理、聊天等场景延伸到风控、运营分析等对大模型能力和准确度要求更高的场景。

PART THREE

应用场景及案例

03

AI大模型应用成熟度

AI大模型在互联网、政务、金融等行业应用场景日益丰富

AI大模型在ToB领域已广泛应用于互联网、金融、政务、工业、教育、消费等多个领域，行业用户通过引入AI大模型解决方案，优化业务流程、提升决策效率、创新服务模式，积极探索如何利用最新大模型技术推动企业数字化和智能化转型。

第一新声研究院整理绘制

金融行业痛点及解决方案

金融行业普遍存在以下痛点：（1）**营销获客难：**金融机构存在用户需求分散、获客成本高等痛点；（2）**风险管理体系效率低：**金融机构在信用风险控制、贷中监控、反欺诈、贷后管理等领域面临很多难点，如在与黑灰产不断升级演变的攻击对抗中，提升机构全流程业务安全、基础安全、技术安全等；（3）**产品设计精准定位难：**随着外资机构、新的金融公司、股份制银行的快速发展，金融机构对优质客户的竞争也愈发激烈，如何打造直击用户需求的金融产品，成为各家金融机构竞争的关键之一；（4）**全流程数字化转型成本高、周期长、难以快速形成收益：**金融机构数字化转型，需要布局全流程数字化，局部数字化会造成无法和前后流程对接，而全流程数字化也面临成本高、建设周期长的痛点，难以快速形成收益 。

火山引擎金融大模型解决方案服务体系

场景赋能	用户增长（MAU）	资产提升（AUM）	管理提效（ROI）			
智能应用	内容套件 • 热点洞察 • 内容创作 • 内容管理 • 内容推荐	算法套件 • 个性化推荐 • 多方安全计算 • 智能反欺诈 • 联邦学习	数据套件 • A/B测试 • 公域洞察 • 用户画像 • 自动化营销 • 用户增长分析	AI套件 • 智能创作 • 智能双录 • 智能外呼 • 数字人		
AI开发	Coze专业版	Hi Agent				
大模型服务	体验中心	模型精调	模型评测	模型推理	Prompt优化	智能体广场
基础模型	豆包 · 视频生成模型	豆包 · 文生图模型	豆包 · 语音合成模型	豆包通用模型Pro	豆包通用模型lite	
	豆包 · 声音复刻模型	豆包 · 角色扮演模型	豆包 · Function Call 模型	豆包 · 向量化模型	豆包 · 图生图模型	

海尔消金-消费金融垂直大模型应用成果

客户问题：

1.消费金融具有“长尾客户”数量多、单个客户价值低等特征，导致成本控制难度大、客户服务提升难等；

2.目前，海尔消金已经拥有超过300个在线运营系统，全流程智能化、自动化的数字化改造困难。

解决方法：

1.基于豆包大模型进行精调，适配问答、总结、创作、分类等场景，打造消金大模型；

2.利用火山方舟平台丰富的插件和AI原生应用开发服务，结合Coze应用开发平台，结合消金场景数据，形成落地应用。

满足智能化场景需求	每天节约专员时间	问答系统准确率
90%+	1-3小时	88%+

医疗健康行业痛点及解决方案

医疗健康行业痛点及解决方案

我国医疗健康行业普遍存在以下痛点：（1）**就医高峰期患者候诊时间长：**疾病高发期，患者骤增，候诊时间长，患者就医感受差；（2）**检查和治疗效率低：**医生为规避医疗风险，会选择通过检查甚至是多次检查来确认诊断结果，导致效率低下；患者在等待检查结果时间长，患者在拿到检查结果后，难以理解报告内容，依赖医生的解释等导致检测和治疗效率低；（3）**用药处方开具涉及信息众多：**用药处方开具需综合评估关系到疾病表现出来的症状、患者年龄、患者生活的环境等超200个维度，医生很难从200多个维度来周密思考用药；（4）**中医领域医疗资源不足：**中医领域面临名医少、传承断代、医疗资源不足，中医医生依赖经验及阅历，同时由于中医数据资料庞大、典籍丰富，中医培养难度高等挑战。

智谱AI医疗健康行业解决方案

场景赋能

医疗机构	互联网医院	零售药房	健康管理机构	运动健康机构	医疗美容机构	医共体/医联体
疾病预防	疾病筛查	医药销售	医生问诊	疾病治疗	医院运营	患者康复
• AI营养师 • 健康百科 • 保健建议	• 在线问诊 • 报告解读 • 疾病自测	• 导购辅助 • 禁忌查询 • 销售质检	• 智能导诊 • 检查推荐 • 检验单诊断	• 治疗建议 • 用药建议 • 医嘱质检	• 用药知识库 • 制度问答 • 数据分析	• AI回访 • 用药指导 • 康复计划

解决方案能力

医学信息提取器	AI医疗对话助手	医生诊断助手	医生研究助手	药品销售助手	门店经营助手
• 摘要生成 • 关键信息结构化提取 • 自动化标签生成	• 诊前轻问诊 • 健康掰开	• 检验单诊断与解读 • 检验推荐	• 大纲生成 • 文献引用查找 • 研究内容生成与质检	• 病情解读 • 禁忌查询 • 对话质检	• 报告生成 • 对话式经营 • 数据查询

模型基座

GLM-4 | CharacterGLM | CodeGeeX | 多模态大模型 | Text-Embedding

东方医院-数字中医大模型应用成果

客户问题：

1.中医存在专业知识传承难和技能培养难等挑战，中医面临失传困境；

2.随着人们健康意识的提高，中医兼具较低副作用和成本的优势，可以提供个性化健康服务，中医领域健康需求迅速增加，由于中医对医生综合性要求高，知识和经验要求高，优质中医服务供给面临较大挑战。

解决方法：

1.基于大模型能力构建了医疗垂直领域的问答功能，支持对医疗、健康问题进行智能化知识问答；

2.数字中医大模型可以根据症状生成中医处方，并提供处方主治症候医学解释等辅助诊疗功能。

知识问答准确率	智能处方合格率	部分疾病预测准确率
93%+	92%+	80%+

教育行业痛点及解决方案

教育行业客服场景痛点及解决方案

售前接待与咨询客服场景普遍存在以下痛点：（1）**线上多渠道咨询接待效率不高：**咨询来源官网、APP、小程序、热线电话等多个渠道，咨询科目众多，正式接待之前无法精准分流，服务效率低；（2）**服务数据分散难以集中管理和分析：**部门间信息流通不畅，数据孤岛严重，数据统计分析困难，无法集中管理和分析；（3）**获客成本高：**线上获客成本越来越高，售前获客线索留资率提升难；（4）**高峰期人工客服接待压力大：**在特定的时间段内客户咨询和投诉的数量会显著增加,人工客服负荷加大等。

美洽全渠道AI客服解决方案

全渠道接入

企业私域：
- 官网
- App
- 小程序
- H5

主流广告平台：
- 百度
- 360
- 搜狗
- 腾讯广告
- ……

社交新媒体：
- 微信客服
- 小红书
- 抖音
- 快手
- ……

海外社媒：
- Facebook
- Instagram
- WhatsApp
- Telegram
- Line
- Email
- ……

语音/400电话

业务场景：售前接待、咨询服务

AI独立接待

文本AI 机器人
- 文本生成
- 信息抽取
- 意图识别
- 内容加工
- 角色扮演
- 知识检索

有效对话率
获线留资率

语音AI 机器人
- 声音复刻
- 通话打断
- 方言理解
- 低延迟通话
- 多语言

响应速度
服务效率
客户满意度

人机协同

AI 辅助
- 话术改写
- 自动标签
- 对话总结
- 信息抓取

质量提升
效率提升

集成知识库生态 一键上传 上传即用

内置行业标准知识库 简易操作 轻量维护

丰富的文档格式支持 一处更新 全局同步

AI 分析
- AI 辅助决策
- AI 数据分析

自主学习
- 历史知识学习
- 标注学习

Agent Studio

Agents | Workflow | Proxy

LLM Gateway

基础模型层

美世教育大模型获客机器人应用成果

客户问题1：线上多个渠道咨询接待，效率不高，服务数据分散无法集中管理与分析

客户问题2：线上流量越来越昂贵，如何提升售前留资率，成为首要需求

客户问题3：高峰期人工客服接待压力大，需要分配人力处理夜间咨询量

解决方法：

针对以上问题 美世教育先后采用美洽全渠道AI客服解决方案以及2024年最新发布的大模型获客机器人，解决了多渠道客户咨询接待统一处理、集中分配管理与数据分析问题，同时接入AI大模型获客机器人，夜间独立接待，安全释放人工坐席。

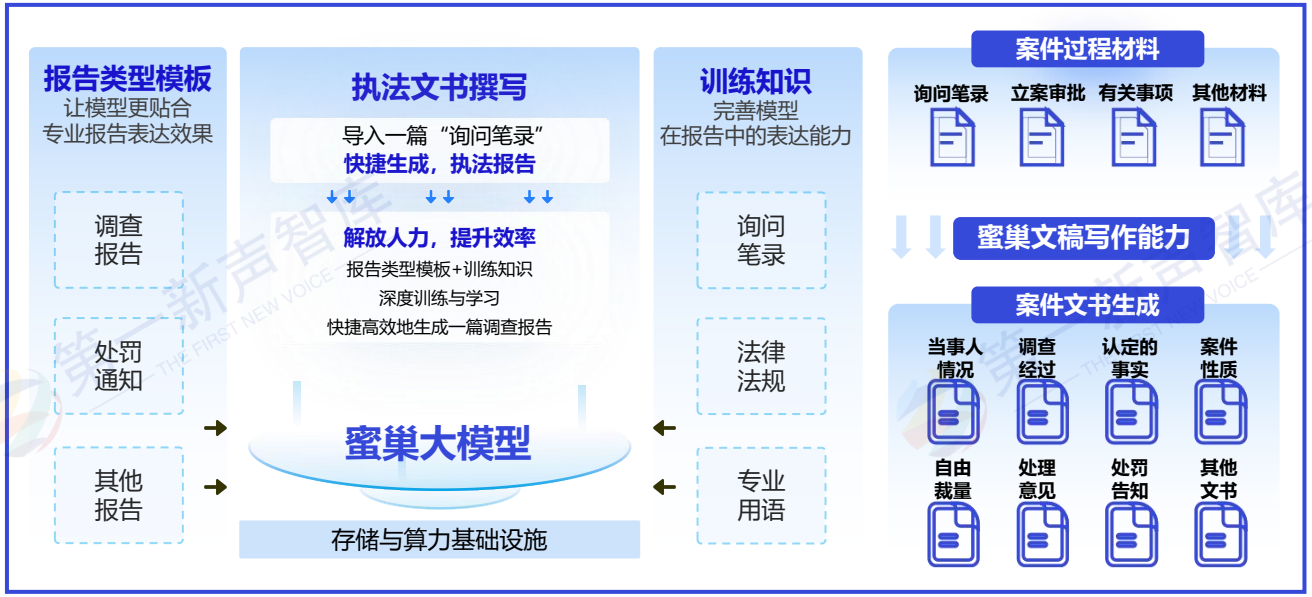
AI带来的效率提升	售前获线留资率提升	全渠道客户消息一个工作台管理智能分流，数据分析，提升运营效率
400%	58%	

政务领域市场监督场景痛点及解决方案

市场监督管理场景痛点及解决方案

市场监督管理应用场景当前存在的普遍痛点包括：（1）**数据分散、统计困难**：如大量案件材料难以统一管理和分析，传统的数据统计方法效率低，且难以进行直观、全面的分析等；（2）**知识库更新不及时**：市场监管领域知识库需根据法律法规、政策变化和业务指南不断更新，当前缺乏统一的更新标准和规则，容易导致信息冗余、混乱或不一致；（3）**执法文书写作耗时长、易出错**：首先人工编写文书耗时费力，且案件业务复杂、法规政策条款众多，信息易缺漏、错误，其次文书格式要求严格，业务人员撰写时易出现不规范的情况；（4）**咨询回复慢、投诉处理效率低**：一方面法规政策多，人工回复慢，市民理解难；另一方面信息检索困难、查询结果不准确、回复效率低等问题，容易引发市民投诉。

蜜巢大模型-市场监督管理解决方案



某区市场监督管理局应用成果

客户问题：人工撰写文书耗时费力，且涉及业务复杂、法律法规多，信息易缺漏、错误……

解决方法：基于国产化环境私有化部署，依托大模型能力，构建市场监管领域的专属知识库，可以辅助生成符合要求的调查报告、行政处罚决定书等文书，同时保障文书的规范性与准确性，有效减轻文书撰写压力。

提升80%	私有化部署	提升至75%
报告效率	信创适配	内容可用率

零售消费行业痛点及解决方案

零售消费行业痛点及解决方案

消费与零售行业传统业务模式当前面临主要问题包括：（1）**客户洞察不足**：数据收集受到限制，缺乏有效的工具来收集和分析客户数据；（2）**市场预测困难**：传统方法难以准确预测市场趋势和消费者行为；（3）**客户体验不佳**：服务个性化缺失，无法满足客户的多样化需求，购物流程繁琐，环节多，体验差；（4）**营销效率低下**：营销策略缺乏创新，难以吸引和留住客户，且营销活动缺乏针对性；（5）**竞争压力大**：市场变化快速，市场竞争日益激烈，许多零售商没有意识到数字化转型价值。

Marketingforce迈富时消费与零售解决方案

客户服务与体验优化	营销与销售策略	库存与供应链管理	数据分析与决策支持
智能客服系统 解决客户呼入流程繁琐的问题。提供7x24H智能客服服务	**精准营销与个性化推荐** 分析顾客画像、交易数据等，实现精准营销和个性化推荐	**AI仓储助手** 计算标准库存计划，监测商品库存数量，并自动发起采购请求或发送提醒	**门店经营助手** 让员工通过语音对话的方式，实时查询所需要的数据
智能外呼场景 支持智能外呼场景，真人交互体验，助力提高转化率	**新客户拓展与老客户维系** 挖掘会员购物数据，精准识别复购机会，实现个性化营销	**顾客声音（VOC）分析** 通过深度分析销售数据，智能推荐热销商品并优化商品陈列布局，提升销售效率	**智能选品与陈列** 分析顾客反馈，优化产品和服务，提升顾客满意度

消费与零售解决方案应用成果

客户问题：
1.员工技能提升缓慢，培训成本高昂；
2.培训效果难以量化，“老带新”效率有限；
3.零售场景中导购员对客户的需求洞察不足，缺乏有效沟通方式；
3.智能化辅助不足，无法快速满足消费者需求。

解决方法：
1.智能对练：模拟销售场景，提升导购能力；
2.知识管理：标准化内容库，支持实时学习；
3.AI助手：语音识别、意图解析，快速响应客户。

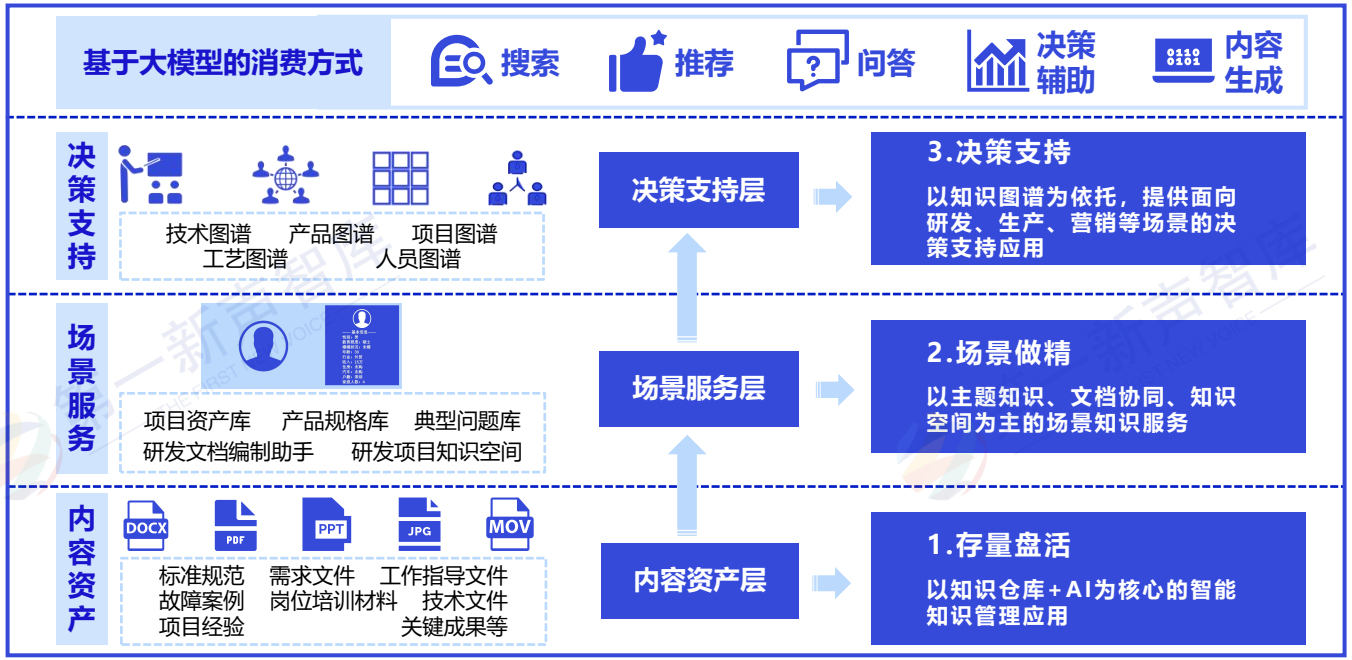
到店率提升	业绩增长	企微沉淀
50%	175%	280%

制造行业痛点及解决方案

制造行业知识管理场景痛点及解决方案

制造型企业，产品开发生命周期长，跨部门协调沟通困难，项目成果多，在知识管理方面通常面临以下痛点：**（1）研发知识难以查找：**研发资料存放零散，知识共享存在壁垒，难以查找，没有体系；**（2）业务骨干依赖度高：**产品研发过度依赖几位核心技术人员，缺乏隐性经验挖掘，忽视了团队能力的提升，知识流失风险大；**（3）生产良品率难以提高：**过往经验和案例总结不到位，质量问题得不到及时的解决，导致问题重复发生；**（4）营销与研发生产脱节：**产品相关资料不全，营销人员赋能不够体系，导致营销业务开展受阻；**（5）知识缺少场景化应用：**不能根据应用场景进行精准知识推送和嵌入、知识和业务两张皮；**（6）知识运营难以持续：**缺乏科学的知识运营方法、缺乏相应数据支撑及智能工具，知识管理推进三分钟热度，难以持续。

蓝凌基于aiKM的制造企业知识管理解决方案



知识管理建设应用成果

需求：赛力斯始创于1986年，是以新能源汽车为核心业务的技术科技型汽车企业。公司业务涉及新能源汽车及核心三电等产品的研发、制造、销售及服务。数十年业务发展中积累了大量知识资产，在知识管理上有几大核心需求：① 整合多业务系统知识，打破知识孤岛，实现知识资产的科学管理、高效获取；② 引入智能化应用，实现研发领域的高效知识协作，提升研发效率；③ 强化知识运营，助力知识的持续提炼萃取。

措施：基于蓝凌aiKM构建全新知识管理平台，包括：基于社群空间构建研发专业社群，强化同行协助、知识共创及经验沉淀；积极探索AI智能助理创新应用，实现智能搜索；基于赛力斯特有研发知识，对接大模型，满足特定业务领域的智能问答等。通过各项知识应用的搭建及运营，实现了“四个一”的效果↓

- 构建1站式知识分享平台
- 打造1个团队共创社群空间
- 实现1分钟找到知识/专家
- 1天问题得到初步解答

PART FOUR

发展趋势及挑战

04

技术趋势

AI大模型四个主要技术方向

当前AI大模型技术发展呈现四大趋势：（1） **Scaling Law面临挑战，大模型研究重点从预训练转向后训练；**（2）**算力平台和模型创新紧密耦合，提升大模型创新效率；**（3） **MoE架构开始广泛应用于推动模型性能和效率提升；**（4）**大模型工具链不断完善加速大模型应用研发与落地。**

01. Scaling Law面临挑战，大模型研究重点从预训练转向后训练

随着互联网文本数据的耗尽，预训练阶段的Scaling law面临挑战，大模型研究关注焦点从预训练阶段转移至后训练阶段，提升数据质量、提升模型复杂逻辑推理能力、降低成本并减少幻觉成为大模型研究工作的重点。

02. 算力平台和模型创新紧密耦合，提升大模型创新效率

“大模型+大算力+大数据”是推动大模型创新和能力提升的重要路线，大模型创新效率提升需要算力芯片、大模型训练和推理加速框架、高质量多模态数据集等紧密耦合共同推动。

03. MoE架构开始广泛应用于推动模型性能和效率提升

MoE架构鼓励在模型设计和训练过程中采用创新方法，有助于促进AI领域内的多样性和创新，MoE架构具有平衡大模型训推成本和计算效率等优势，适合处理大规模数据和复杂任务，已成为谷歌、OpenAI、阿里、腾讯等企业控制成本、提升模型性能、应对大模型“价格战”的新方向。

04. 大模型工具链不断完善加速大模型应用研发与落地

大模型工具链包括训练工具、推理工具和应用开发工具，工具链的不断完善对构建大模型服务体系，应对大模型训推复杂性的挑战和降低大模型开发和部署的门槛起到较强的推动作用，不断完善工具链是推动大模型研发和落地的重要方向。

市场应用趋势

融合应用软件、智能助手和AI Agent是AI大模型市场应用的三个主要方向

大模型商业化落地需要找到合适的产品方向

融合应用软件

- 应用软件厂商将更专注于AI大模型应用场景的探索以及与现有应用的融合，未来大模型厂商将会承担绝大部分的底层算法开发优化工作，应用软件厂商则会更专注在应用场景，以及与现有AI大模型更深度的融合应用；例如美洽将大模型融入全渠道客服解决方案，并发布大模型客服机器人，大幅提升了产品性能，提高了终端用户的接受程度；
- 工具软件与AI的融合能够优化用户体验与生产效率，提升产品竞争力，带给用户“新奇感”，提升用户生产效率，由于短期接入大模型的试错成本较低，市场接受程度高，工具软件厂商对于AI大模型接入抱开放态度，AI大模型融合应用的功能或将成为工具软件的增量付费点。

智能助手

- 大语言模型提升了智能助手的自然语言理解、生成能力和多模态能力，通过执行各种任务，提升用户的生活和工作效率；秘塔、蓝凌软件、动悦信息等开发了基于大模型的知识助手，泛微、致远互联、蜜度等开发了基于大模型的办公助手等产品；
- 当前基于大模型的智能助手已经具备较丰富应用场景，例如天工 AI，可用于包括写作、娱乐、PPT、图片生成、英文学习等应用场景；通义的应用场景包括写作、语音识别、升学、金融等；豆包在情感陪伴、职场办公、教育学习等场景中表现突出。

Agent

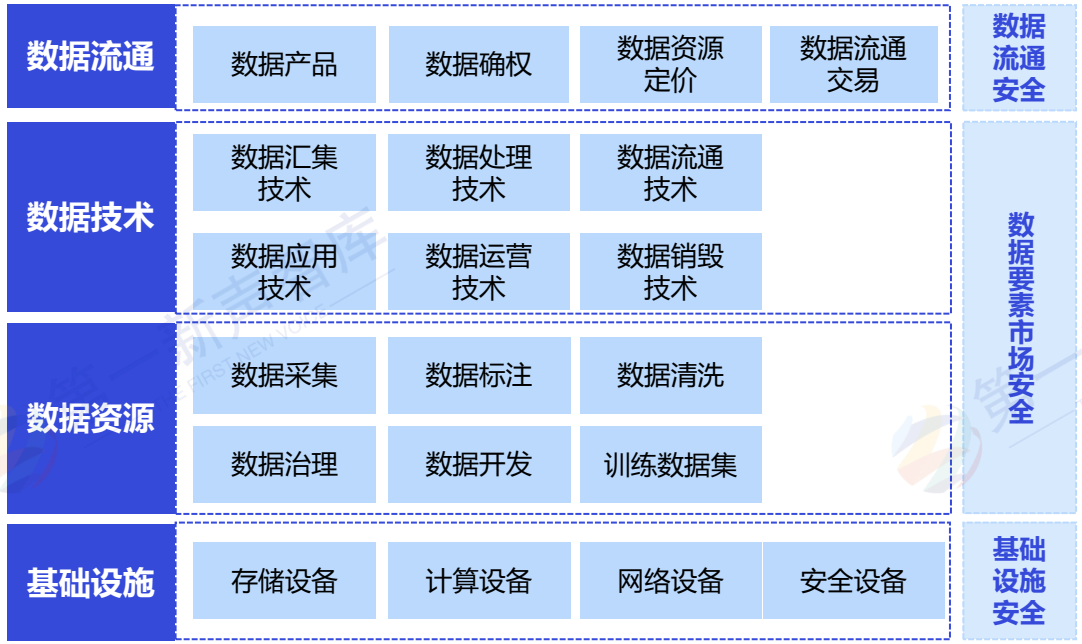
- AI Agent作为一种满足企业智能化需求、打通业务场景的AI大模型产品化落地形态，可承接日益复杂的提质增效需求，帮助强化内外部协同效能；
- 多模态大模型能利用大量异构的数据资源提升应用的效率和能力上限，利好 AI Agent 发展。多模态大模型能提高智能体的工作效率，赋能单个智能体能完成更多复杂的任务，有效减少智能体数量和任务中的交互次数。智能体有望解决跨行业、跨领域的复杂问题和各类长尾场景；
- 目前，Agent研究开发已经包括多模态信息识别与理解技术以及群体智能技术，这将有望加速人工智能从感知向认知的转变。

面临的挑战

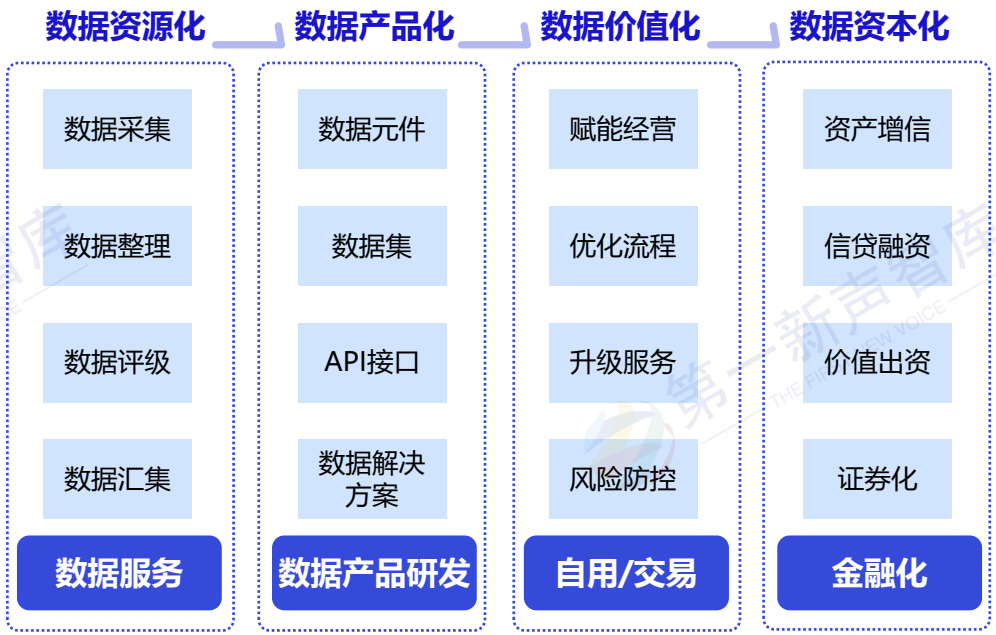
缺乏高质量数据集是大模型商业落地面临的关键挑战

AI大模型应用面临公开可用的高质量中文数据资源相对较少的挑战。当前，高质量数据集缺乏的主要原因包括：（1）**国内专业数据服务产业处于起步阶段，整体资源投入较少：**当前数据存在规范性不足、质量不高、安全性存在隐患等问题；高质量数据集的产生要求投入大量资源对数据进行收集、清洗、标注、验证等工作；（2）**数据交易体系还未形成规范和标准：**数据交易体系流通存在数据权属配置不清晰、交易体系薄弱、缺乏通用的数据定价规则、未形成统一的流通规则、数据安全治理和监管体系薄弱等问题；（3）**私域数据流通难：**训练垂类大模型所需行业和场景数据需要更为专业的行业知识，涉及企业私域数据，出于隐私保护等原因，此部分数据获取较为困难。

数据流通体系架构



数据生产和流通路径



研究团队

研究团队

总顾问

Michael Fang
第一新声
研究院合伙人

Michael Fang，第一新声研究院合伙人

曾就职于Gartner/百度/Oracle，英国帝国理工学院硕士。Michael先生从事 IT B2B 行业8年多，拥有厂商策略、咨询机构多方工作经验，熟悉Gartner及多家咨询公司各类资源构成逻辑、方法论和思路。曾在Gartner高科技行业客户团队服务过包括云计算、安全、数据管理与治理、AI&大模型、企业软件/协同办公、SaaS（BI、营销科技）、科技媒体等领域客户。

总策划

姚毅
第一新声
创始人兼CEO

姚毅，第一新声创始人兼CEO，毕业于中国人民大学。

第一新声研究院《2024年中国信创产业研究报告》、《2024年中国数据库市场研究报告》、《2024年中国交通运输行业数字孪生市场研究报告》、《2024年度中国CIO数字化产品选型白皮书》、《2023年中国快消企业数字化产品应用与实践报告》、《2023年中国信创产业研究报告》、《2023年中国服装供应链数字化应用与实践报告》等报告总顾问。

研究团队

- **报告执笔：**第一新声高级分析师 东君
- **报告审核：**第一新声创始人 姚毅、第一新声合伙人 Michael Fang
- **报告校对：**第一新声 金磊、皓月
- **合作咨询：**请联系第一新声BD Sherry（微信号Sherry_199909）
- **特别感谢：**蜜度、美洽、Marketingforce迈富时、蓝凌、东信营销科技、容联云、泛微、万兴科技、网易云商、北电数智、宁畅、火山引擎、MiniMax、云从科技、新华三、智谱

第一新声
— THE FIRST NEW VOICE —

关注第一新声公众号

合作联系人