

极客邦科技双数研究院

InfoQ 研究中心



中国大模型落地应用 研究报告 2025

InfoQ 极客传媒



研究范围说明：本报告聚焦于应用层的发展现状研究（蓝色部分）

应用层

生活服务

- 法律助手
- 医疗助手
- 旅游助手
- 家居助手
- 购物助手
- 市民服务

金融

- 客户交互
- 研究交易
- 内部运营
- 智能风控

社交娱乐

- 角色聊天
- 游戏智能NPC&队友
- AI特效
- 直播互动
- 写真相机
- 同声传译

零售电商

- 实体零售
- 电商

协作办公

- 思维导图
- 会议逐鹿
- 文档
- 编程
- PPT
- 搜索助手

汽车

- 生产
- 研发
- 用车
- 售后&营销

艺术创作

- 网文/小说
- 音乐
- 播客
- 绘画
- 短视频
- 视频

教育

- 学前&其他教育
- 基础教育
- 高等教育
- 职业/素质教育

应用开发层

模型调用

开源模型
闭源模型
微调模型

应用开发工具

提示词工程
RAG
工作流编排

Agent平台

智能体插件商店
智能体模版
智能体运营平台

用户体验设计

Web界面
移动端App
前端交互设计
系统提示词设计



智能体平台
初创企业



垂直领域数
字化厂商



云厂商

模型平台层

模型微调与精调

微调/精调数据库
模型微调

模型部署

模型优化
部署环境准备
微调适配
硬件加速

模型评估

性能评估
安全&伦理评估



开源社区/
工具链企业



垂直领域数
字化厂商



云厂商

模型层

预训练模型

文本模型
视觉语言模型
语音模型
视频模型
图文模型
多模态模型

微调模型

对话模型
推理模型
指令微调模型
Agent模型

开闭源模型

开源模型
闭源模型



通用大模型
创业厂商



垂直领域数
字化厂商



云厂商&传
统AI厂商



新型研
究机构

基础模型层

算力基础设施

GPU等算力芯片
算力集群（智算中心、云算力集群）

数据基础设施

数据采集和处理
数据标注
数据增强
数据治理



AI芯片公司
(GPU&TPU等)



算力集群提供/运营商
(云厂商、智算中心)



数据标注&合
成数据服务商



数据治理
服务商

目录

CONTENT

01 | 中国大模型发展驱动因素

02 | 中国大模型行业应用全景洞察

01

中国大模型发展驱动因素



七年技术进化全景图：大模型进入推理纪元

- 自2017年以来，大模型已历经诸多关键时期，从文本到图文、音频、视频，从单纯的对话到代码和终端。推理成本的大幅下降也助推了大模型技术的广泛应用。在2025年这个阶段，我们看到了文本推理模型的爆发，推理特性也正在向图片、视频等多模态迈进。

多模态融合期

成本普惠期

文本模型进化期

算法奠基期

- 以Transformer为代表的**全新神经网络架构**，奠定了大模型的算法架构基础
- RLHF、SFT、指令微调等**后训练手段**开始出现，大模型推理能力和任务泛化能力得到提升

- 以Midjourney和DALL.E的**图像生成**引发热议
- 以GPT3.5为代表的**超大参数规模模型**陆续上线，国内外大模型厂商陆续完成一轮交卷
- 以Llama模型家族为代表的**开源模型**登上战场
- AutoGPT框架发布，**AI Agent**概念持续发酵，Dify等以Agent搭建为核心业务的商业公司/组织/平台出现
- 《**生成式人工智能管理暂行办法**》发布执行

- 文本模型展开**长上下文窗口**竞争，同时**推理API价格**开始持续下调
- Sora拉开视频生成模型序章，图文模型、视频生成模型、音乐生成模型能力从单文本迈向**原生多模态**探索
- MOE架构模型**频繁发布，架构计算效率优势广受关注
- Devin**代码智能体**概念引发热议，包括Baidu Comate、通义灵码在内的国产**智能编程助手**开始大规模公测
- 手机、PC厂商**开始进行战略调整，更新发布**终端智能体**，出现**大模型小型化**趋势
- Figure 01和Ameca率先尝试**人形机器人融合大模型**

- 截至2024年12月31日，共**302款**生成式人工智能服务在国家网信办完成**备案**
- DeepSeek、豆包、通义、文心等开启大模型**推理成本**降价潮
- O1拉开**强化学习**路线序幕，各家开始尝试更新推理模型
- LLM**交互模式**探索：Claude Artifacts, ChatGPT Canvas, CogAgent
- 视频生成国内成果**迅速追平国际水平：MiniMax海螺AI、快手可灵、字节即梦、生数科技Vidu、智谱清影、爱诗科技PixVerse、腾讯混元

推理模型爆发期

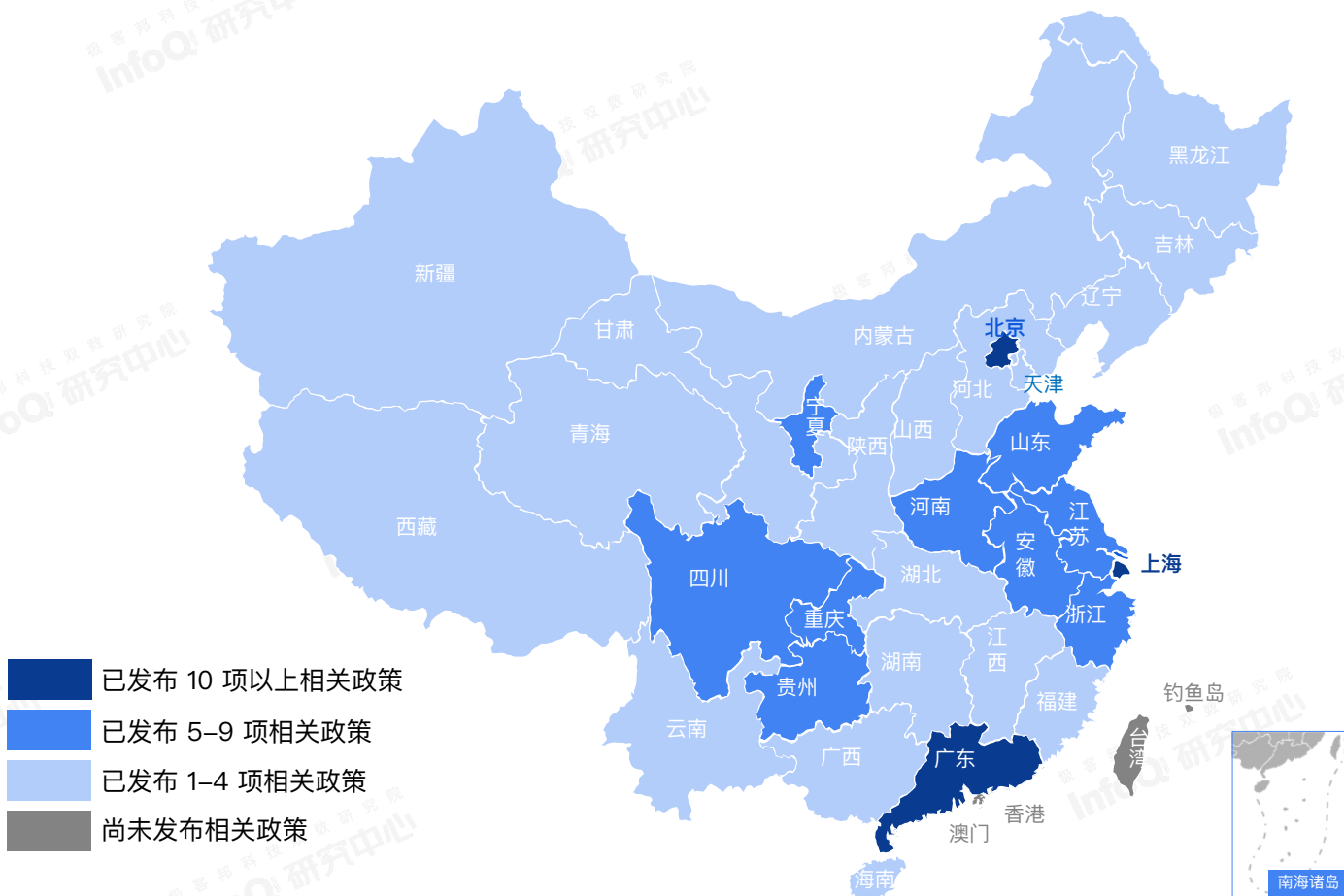
- DeepSeek开源引发热议，k1.5、Grok3先后发布，正式进入**推理模型**时代，视觉推理模型频繁发布
- 未完待续.....

2017-2021年	2022-2023年	2024年上半年	2024年下半年	2025年上半年
------------	------------	----------	----------	----------

数据来源：InfoQ研究中心根据新闻、官方公告和文章等公开渠道整理和分析总结

政策引擎全开：31省大模型产业布局呈现三级阶梯格局，北京上海领跑「全链」发展

2023年-2025年1月大模型相关政策发布情况



数据来源：InfoQ研究中心根据各省市政府官网、政府公文、知乎文章等公开渠道整理和数据分析

- 除港澳台外，所有省份均已发布大模型相关政策
- 北京、上海均已发布接近**20**项大模型相关政策，数量和政策全面性在全国遥遥领先
- 东数西算**范围内的宁夏、内蒙古、贵州、京津冀均在算力建设整体规划中提到支持大模型发展

2025年政府工作报告支持大模型广泛应用

- 11**个省份/直辖市/自治区，在2024年政府工作报告中，提及了大模型/通用人工智能：北京、上海、重庆、安徽、广东、贵州、湖北、江苏、山东、内蒙古、宁夏

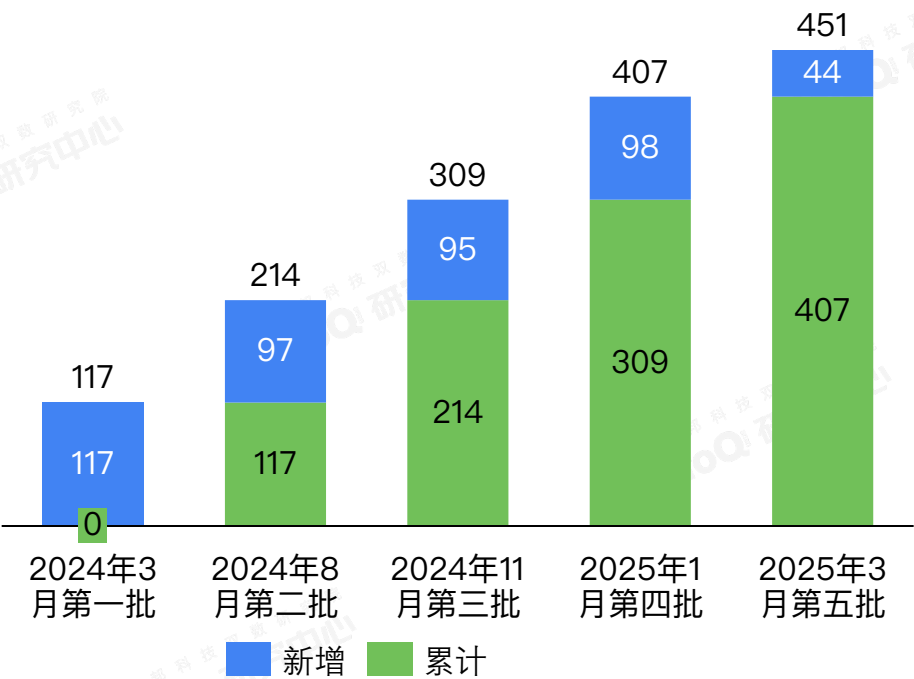


备案情况：451个生成式人工智能服务已完成备案，当前市场处于"技术扩散期"向"场景深耕期"过渡的关键节点

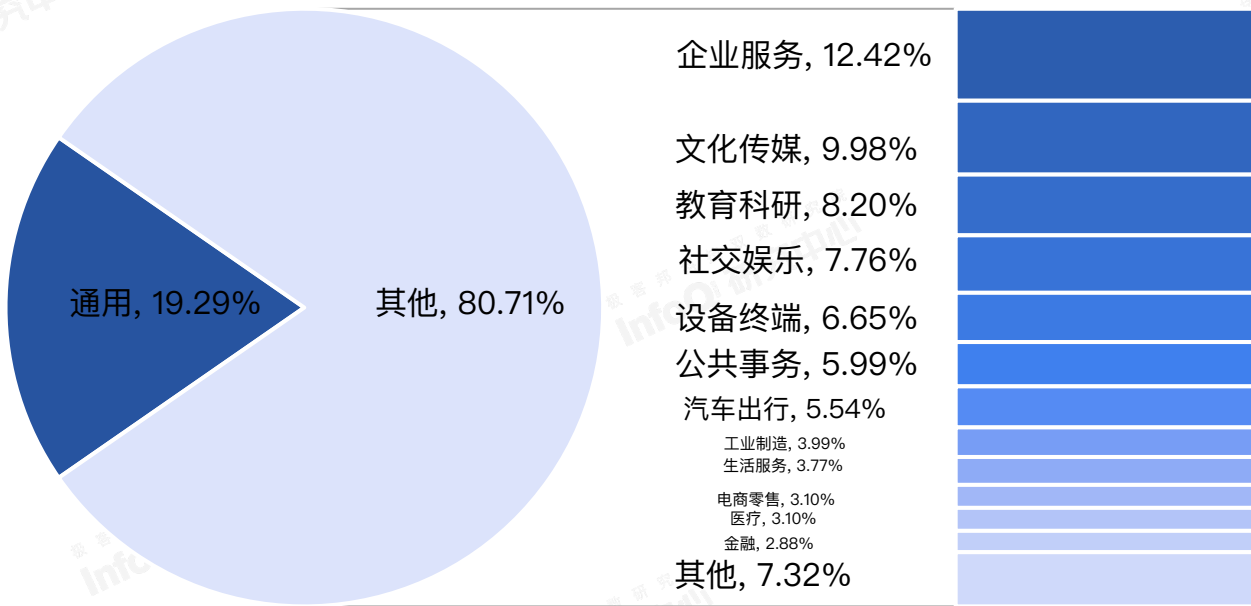
《生成式人工智能管理暂行办法》执行近两年中，五个批次共有451个生成式人工智能服务先后完成备案工作。

从备案生成式人工智能服务的类型出发，通用型仍然持续领跑，但企业服务、文化传媒、教育科研、社交娱乐等垂直领域多点开花，生成式AI正在各类垂直领域印证商业化前景。

各批次生成式人工智能服务已备案数量



已备案生成式人工智能服务类型分布



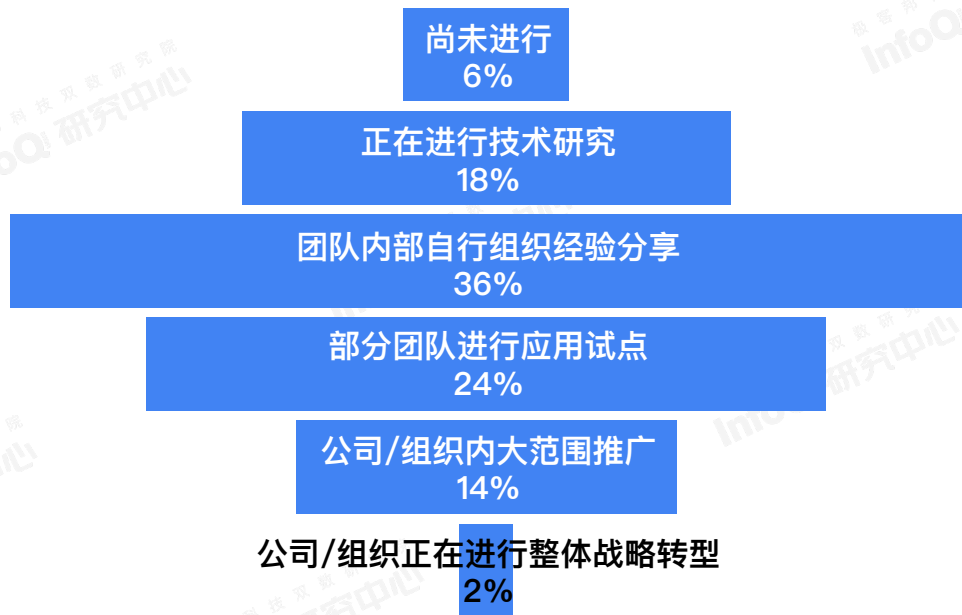
数据来源：InfoQ研究中心根据国家互联网信息办公室相关公告收集整理和分类分析；注：企业服务包括营销、客服、招聘等场景应用



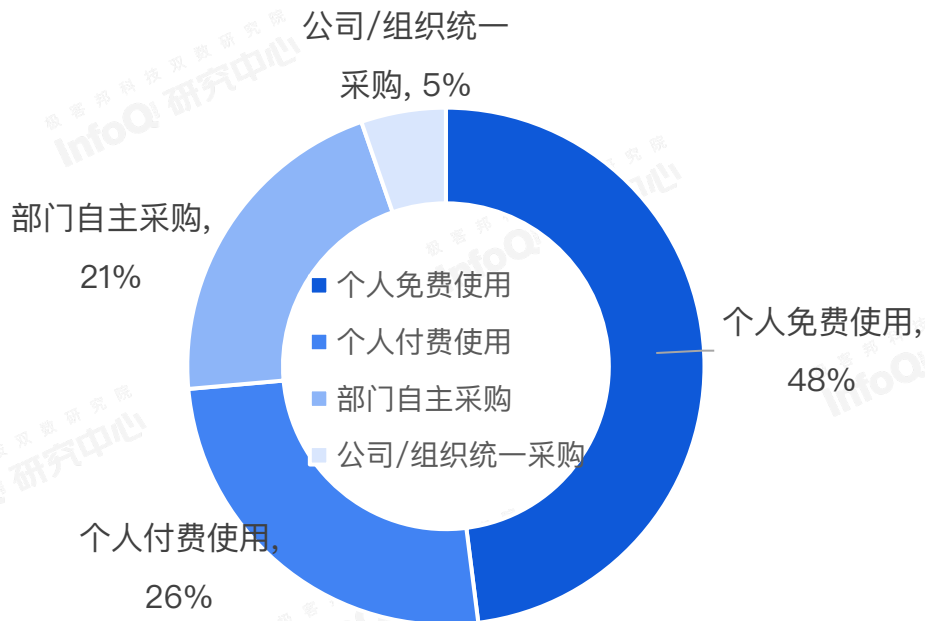
企业应用情况：积极试点，谨慎付费，为成果买单

- 与大模型技术热度相对应的是，企业的谨慎探索，仅14%的公司/组织推行了大范围应用，更多的企业停留在了经验分享和小范围试点。
- 同应用落地现状一致的是，公司在采购方面方面的决策较为谨慎，这也同实际效果的感知较弱相关。这也意味着，大模型的落地仍处于探索阶段，企业对其价值的认知和信任度尚未完全建立，需要通过更多实践和成果来验证其实际效益，从而推动更广泛的规模化应用。

各调研样本所处公司/组织大模型落地阶段分布



各调研样本所处公司/组织落地方式分布



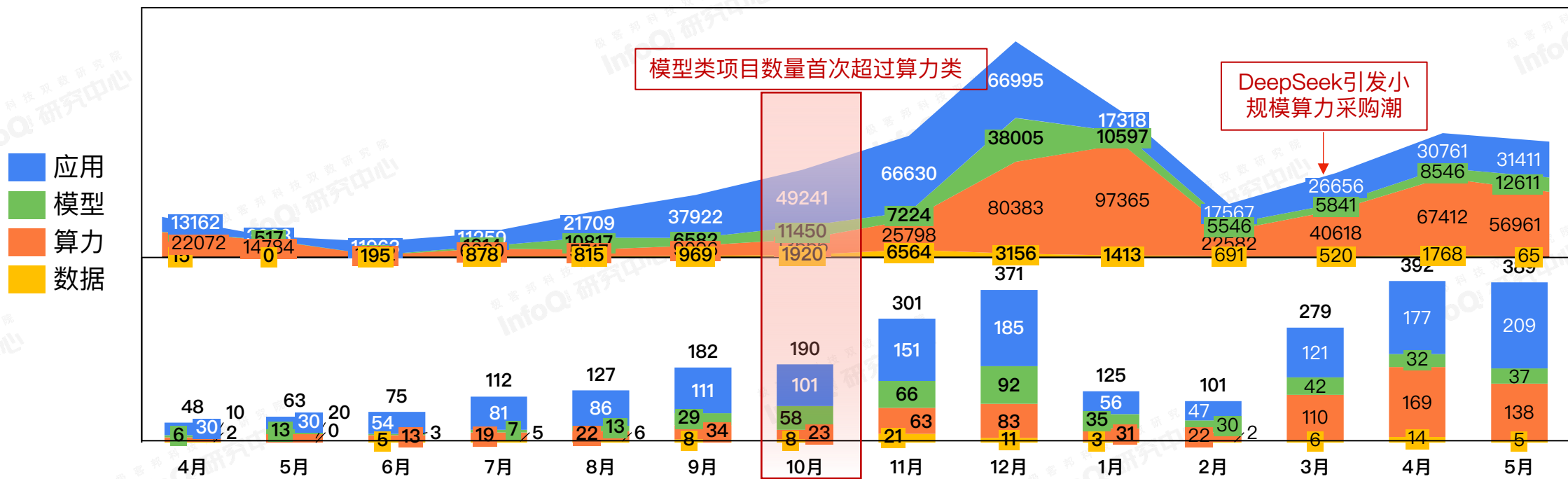
数据来源：InfoQ研究中心2024年7月的用户调研，N=1166



招投标数据：算力与应用领跑，模型类数量超车但价值认可度待破局

- 从整体数据来看，大模型相关招投标在2024年呈现爆发式增长，全年项目总量达1520个，较2023年的190件激增700%，展现出强劲的市场动能。
- 从项目数量和金额来看，2024年，应用类大模型招投标项目占比明显提升，反映出市场对垂直场景解决方案的迫切需求。然而，其平均投资规模仍远低于算力类项目的重金投入。值得注意的是，自2024年10月起，模型类项目数量首次反超算力类，并在随后的4个月中持续保持这一趋势，或将成为市场格局的重要转折点。

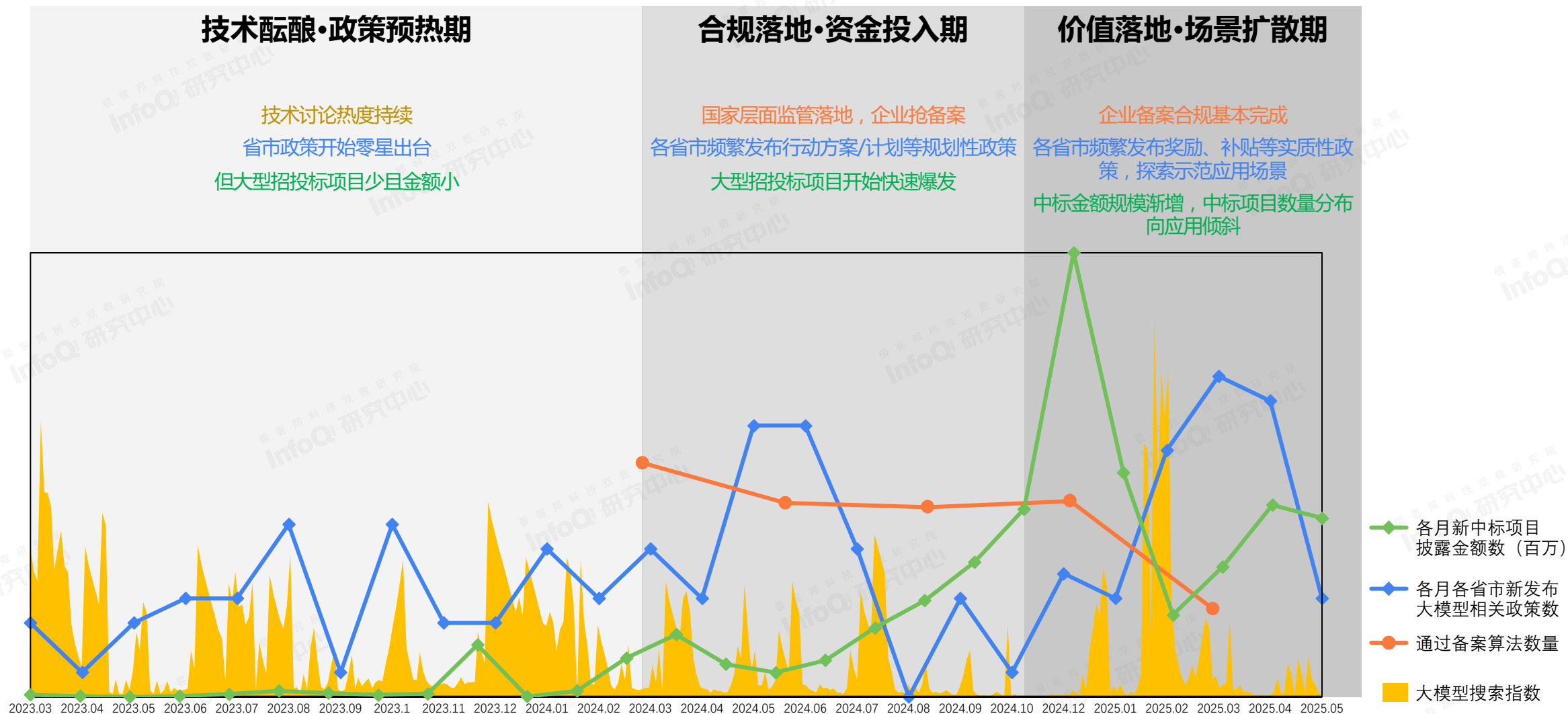
2024年4月-2025年1月大模型相关招投标披露项目金额（上图：万）和中标项目数量（下图：个）示意图



数据来源：InfoQ研究中心根据智能超参数、百炼智能数据整理和分析总结

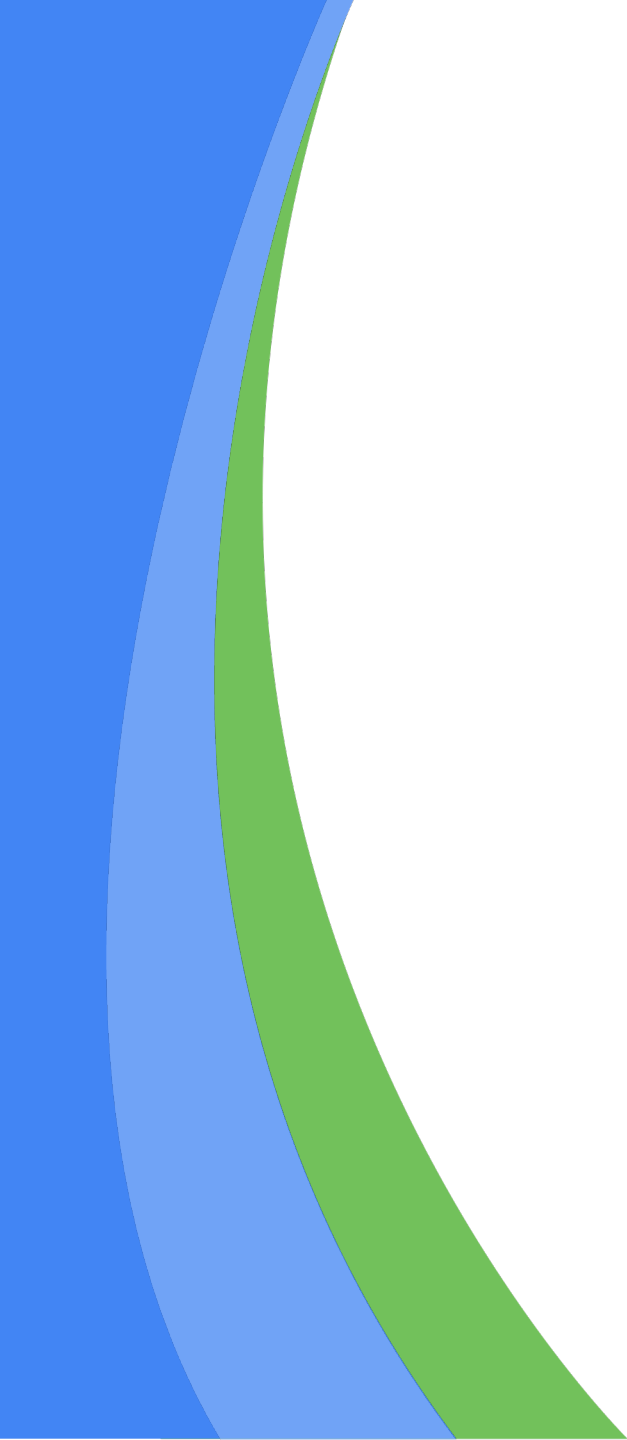


大模型驱动力演变曲线：从技术预热回归商业价值





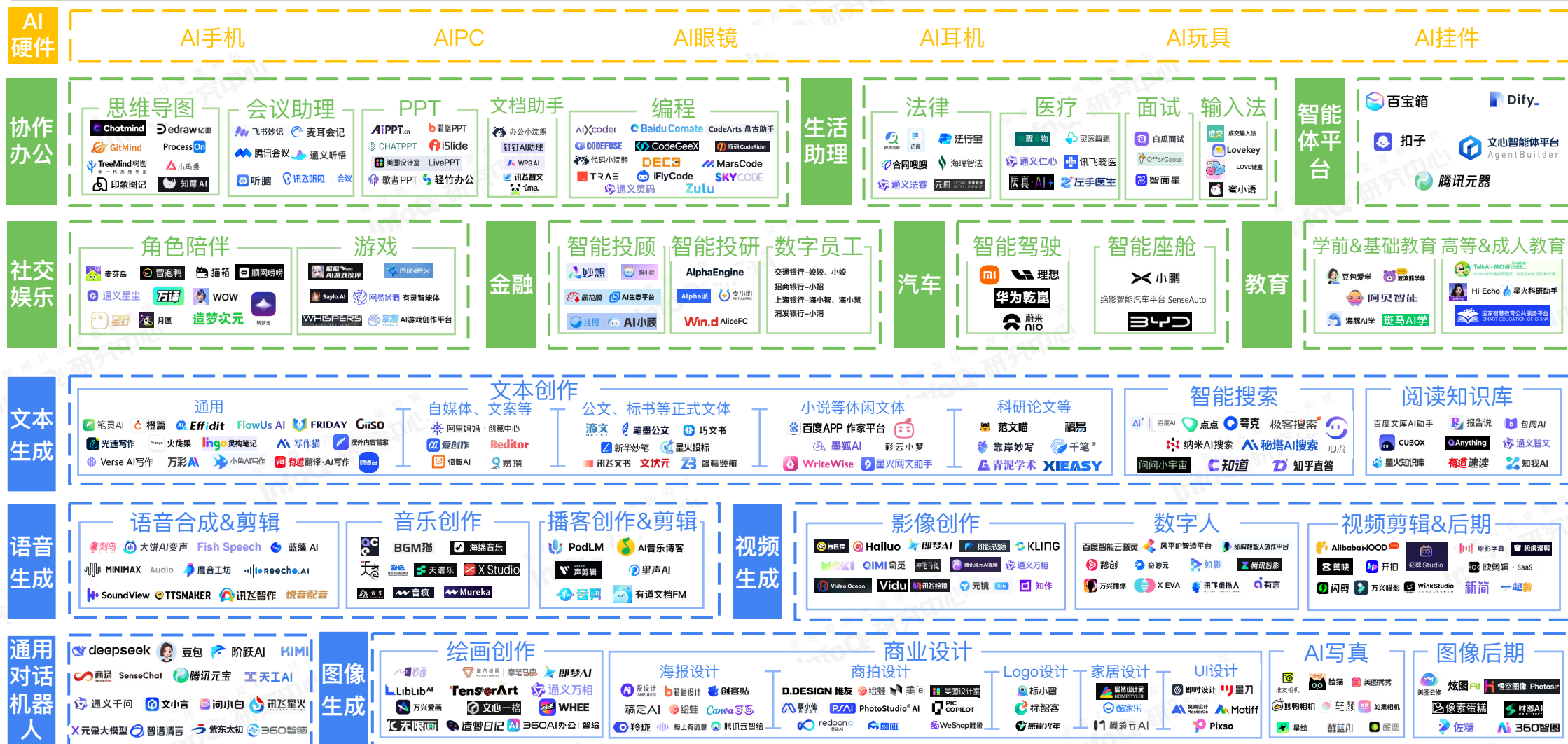
中国大模型行业应用全景洞察



C端大模型产品发展情况解读



大模型应用产品图谱2.0



入库标准：2025年4月仍在运营，存在独立app、网页或者其他消费者能够直接使用的入口

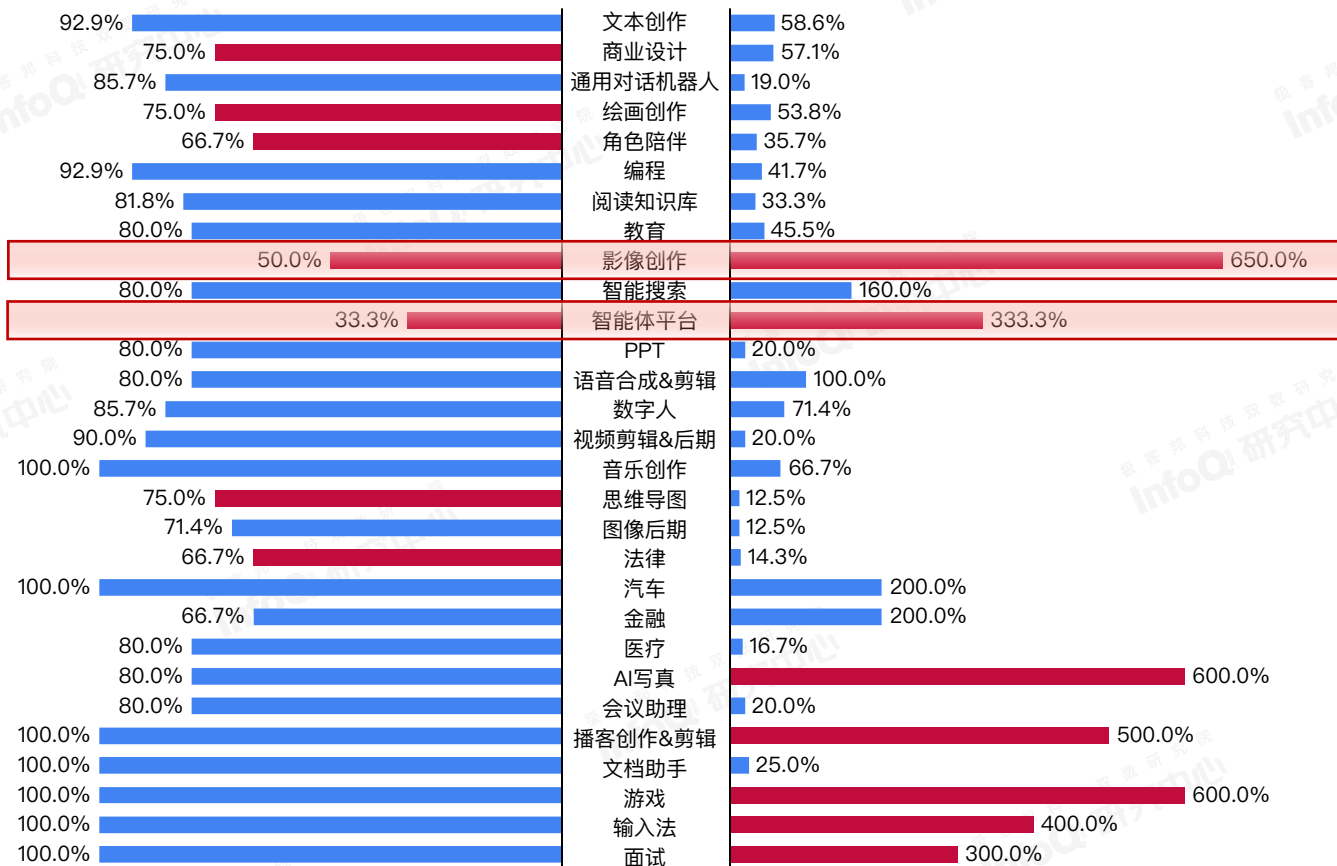


图谱新知：大模型产品进入分化期，超10%在库产品已下架，影像创作和智能体平台新市场正面临激烈竞争，播客、输入法等细分领域产品正在涌现

- 在《2023年第4季度中国大模型季度监测报告》中，InfoQ研究中心梳理总结了截至2023年12月，以独立app或者网页端形式正常运营的229款AI应用。在2025年这个节点，InfoQ研究中心又回顾和更新了InfoQ研究中心AI产品库，并将库内产品扩充至近400款。

存续率

新增率



① 商业设计、绘画创作、角色陪伴、思维导图类的AI应用原有产品竞争激烈，但市场新进入者较少

② 影像创作和智能体平台正在面临激烈竞争，但新市场正在吸纳大量新进入者

③ 编程、PPT、思维导图、会议助理和文档助手进入相对稳定的发展期

④ AI写真、播客创作&剪辑、输入法、面试类AI应用正在涌现大量新产品

12.3%

2023年12月的在库AI产品
已停止运营

65.0%

2025年4月的在库AI产品
增长率

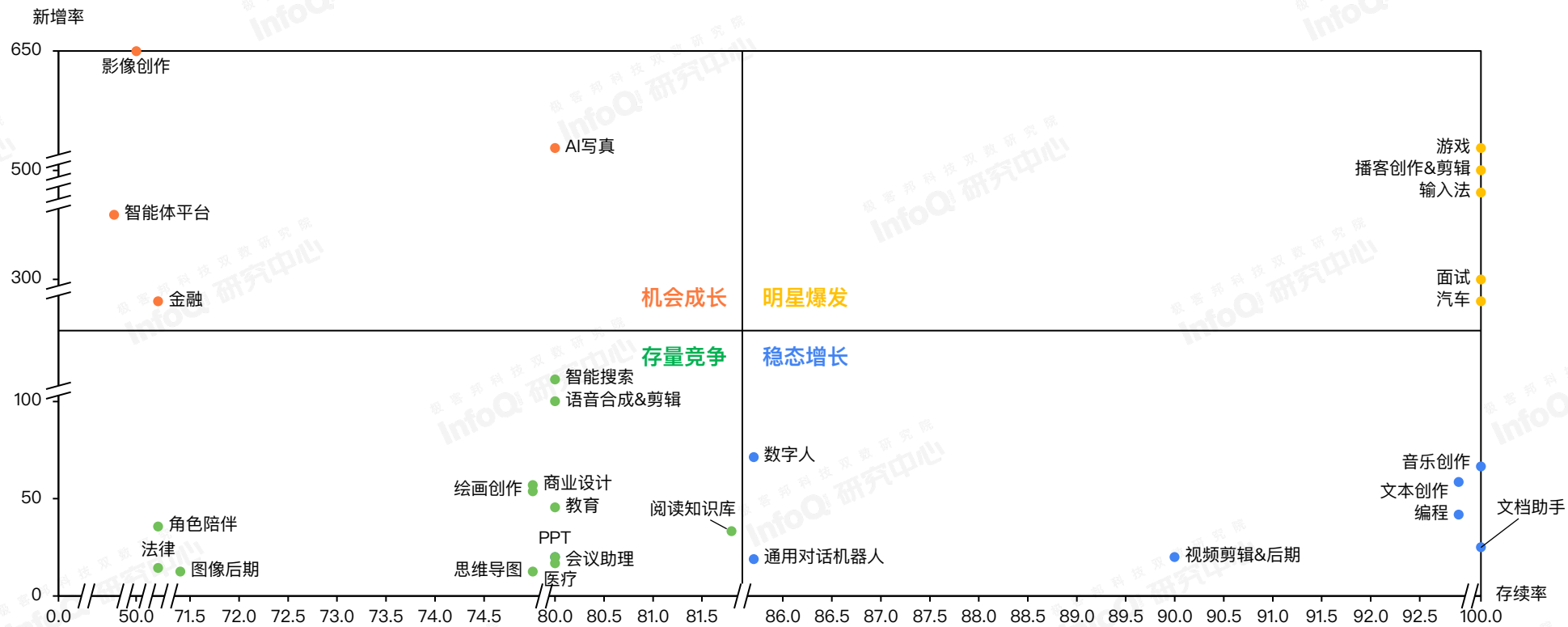
入库标准：2025年4月仍在运营，存在独立app、网页或者其他消费者能够直接使用的入口



C端大模型产品四大赛道：存量竞争、机会成长、稳态增长和明星爆发

- 报告按照产品存续率和新增率将大模型产品分为四个赛道，分别是：

- 低存续 · 低新增的存量竞争赛道
- 低存续 · 高新增的机会成长赛道
- 高存续 · 低新增的稳态增长赛道
- 高存续 · 高新增的明星爆发赛道



存量竞争赛道（低存续·低新增）：竞争焦点向效率和内容生态转变

- 在商业设计、绘画创作、角色陪伴及思维导图AI应用领域，市场已进入存量博弈深化阶段，表现为成熟产品竞争激烈程度持续攀升，但新玩家渗透率显著收窄。同时上线产品功能已经逐渐稳定，这意味着该赛道的产品竞争焦点已经转向效率提升和内容生态建设转变。

产品成型早，现阶段聚焦能力升级

- 以上各类产品，产品成型早，功能萌芽早。在这2年中，产品主要依托底层模型的能力升级，进行功能的补全和提效
- 例如绘画制作中的局部调整，场景固定等优化使用体验的功能上新

竞争激烈且市场新进入者少

- 该赛道产品的存续率均在70%左右，这意味着该领域产品目前正在面临激烈竞争，但同时也意味着首批玩家已站稳脚跟
- 该赛道产品的新增入库率均低于50%，这意味着该领域外部新玩家涌入放缓

竞争焦点向效率提升和内容生态建设转变

- 从流量争夺到体验锁定：在基础功能相对同质的情况下，产品以用户指令遵循能力和任务闭环效率，构建用户替代成本和体验壁垒
- 从技术追赶到生态深耕：通过垂直数据资产沉淀与版权协同网络建立结构性壁垒



机会成长赛道（低存续 · 高新增）：竞争重点在于模型代际适配力 × 场景耦合深度

- 在影像创作、智能体平台等领域，市场伴随着机会也伴随着风险，表现为新玩家持续涌入，但已上线产品近半数无法存活，面临快速洗牌和激烈竞争。
- 同时底层模型能力仍在快速升级，这意味着产品需要快速跟上底层模型的能力升级，以及对用户使用方式和场景有明确拆解和融入。



尽早上线抢占用户心智，但底层模型仍在快速升级

- 底层模型能力仍在快速升级，对现有产品形态仍存在一定颠覆可能性
- 但现有产品的思路核心在于早上线以提前抢占用户认知

竞争蓝海，淘汰率高但吸引力强

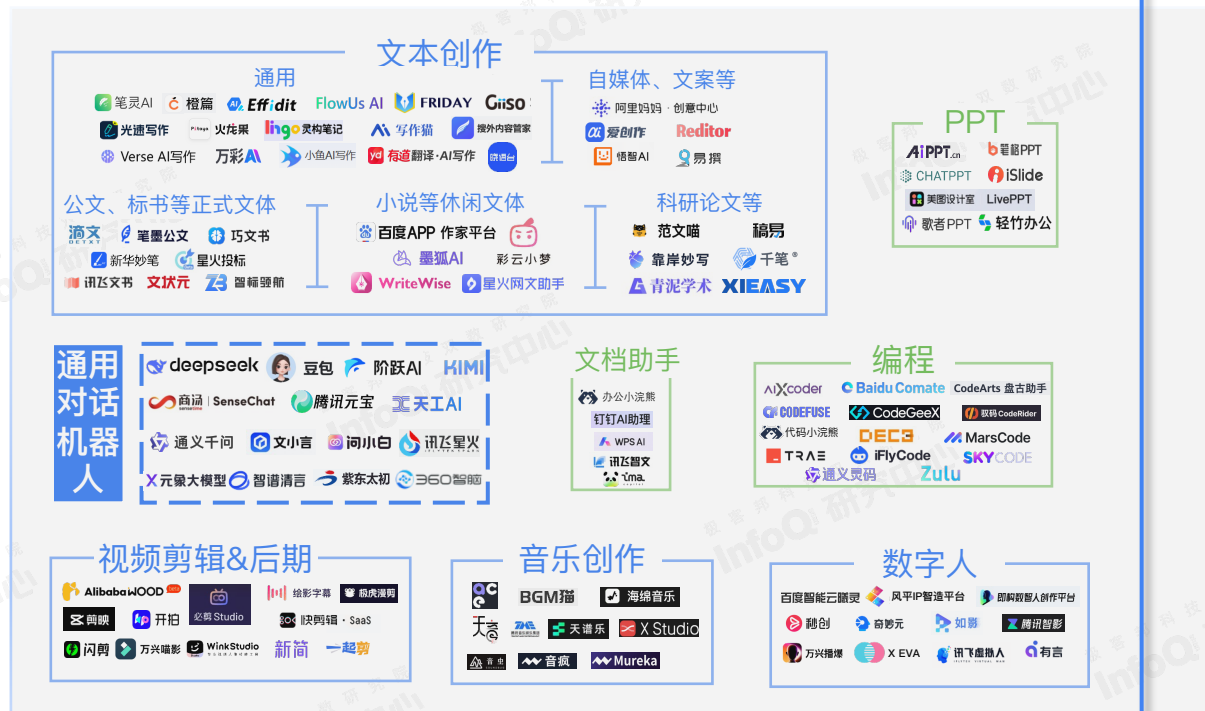
- 存续率仅 50—70%，产品生命周期被极度压缩，这对产品设计提出了高要求，即如何在相似的功能点上留下用户
- 新增率均超过200%，这意味着该领域的广阔增量市场价值被认可

竞争重点在于模型代际适配力 × 场景耦合深度

- 一方面，产品需要能够跟上仍在快速迭代的底层模型升级
- 另一方面，通过对用户使用方式和场景的拆解，将产品融入用户的工作流，深度耦合用户的使用场景

稳态增长赛道（高存续·低新增）：竞争核心在于保障文件安全·流程合规·管理提效

- 在编程助手、PPT生成等效率工具领域，市场已确立稳态增长格局，表现为高存续率与低新增率。竞争核心正从基础功能覆盖转向企业级组织需求满足深度——产品需通过精准建模文件管理、流程协同与安全治理三大刚性诉求，构建工作流路径依赖，驱动工具属性向基础设施演进。



产品形态成熟

- 产品多与协同办公软件集成，产品形态成熟
- 对底层模型的稳定性要求高

竞争趋稳

- 存续率 80—100%，首批玩家基本存活
- 新增入库率 15—30%，仍有新点子诞生，但远低于爆发赛道

竞争重点转向工作流粘性

- 护城河由“模型差异”转为“工作流黏性”，哪个产品最懂企业文件、流程和安全诉求
- 例如，编程助手：IDE 深度插件 + 私有代码语料，形成二次壁垒



明星爆发赛道（高存续·高新增）：产品价值验证和用户留存是核心

- AI写真、播客创作&剪辑、输入法、面试类AI应用正在涌现大量新产品。明星爆发赛道（高存续·高新增）呈现出强劲增长势能。市场中新产品持续涌现，其核心逻辑聚焦于通过产品价值验证与高效获客，构建用户留存基本盘，驱动赛道发展。



产品切入用户细分需求

- 该赛道产品聚焦用户未被满足的垂直场景，例如风格写真、博客自动创作
- 区别于技术驱动型赛道，该赛道产品更强调需求适配度

市场需求正在爆发，新老玩家正在同台竞争

- 存续率 80—100%，首批玩家基本存活
- 新增入库率 > 300%，赛道正在吸纳大量新进入者

竞争重点在于找准用户需求并获得持续的用户留存

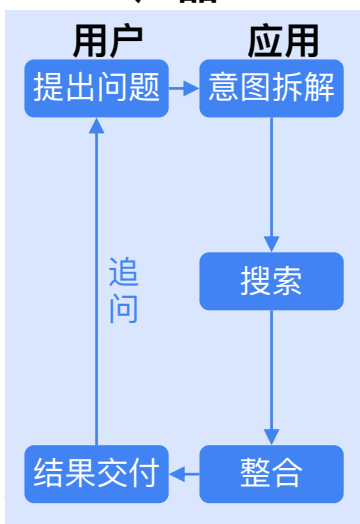
- 一方面，需快速通过市场反馈验证产品核心价值
- 另一方面，依托获客效率与体验设计，将新增用户转化为长期留存，形成“高新增→价值验证→高存续→再获客”的增长飞轮



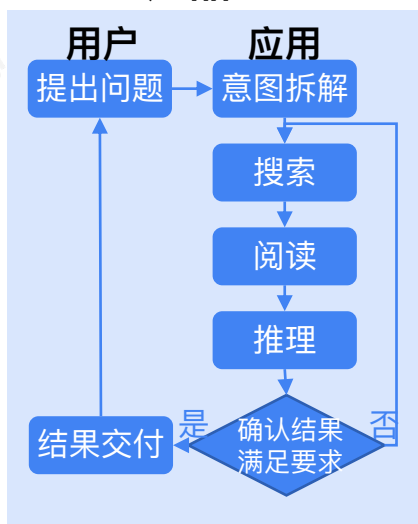
产品新知：DeepResearch和Manus革新搜索交付形式

- 除了以上赛道的共性以外，InfoQ研究中心还发现了一些正在演变的产品类型。
- 25年一季度，伴随着DeepSeek-R1热潮，深度思考模式成功打入用户市场。同时，基于深度思考模型的智能解析能力显著拓展了搜索边界，使AI搜索从传统的信息检索向深度知识挖掘转型。
- 此外，OpenAI-DeepResearch以及Manus带来的静态网页开启了AI搜索交付形态的革新探索，AI搜索从单纯的提供搜索结果交付，到离用户搜索目的更进一步的交付方式进化。

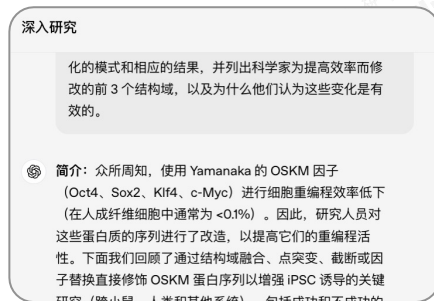
AI产品1.0



AI产品2.0



1



深入研究——研究报告形式交付

代表：OpenAI-DeepResearch

2



搜索总结——UGC内容总结

代表：小红书-搜搜薯

3



静态网页——静态网页形式交付

代表：Manus、秘塔AI

4



其他工具——调用其他插件交付

代表：夸克AI超级框

产品新知：大模型改造经典内容应用，提升运营效率，开创全新用户体验

- 除了AI原生应用的探索，各类经典应用也尝试用大模型改造自身功能逻辑，焕发新的应用活力。



百度文库升级AI功能

- 在百度文库超14亿专业文档资源的基础上，升级定位为“一站式AI内容获取和创作平台”，期望飞越式提升内容查阅与创作效率。
- 学习办公：依托原有文库内容，展开搜索、总结、创作等新功能改造
- 家庭教育：解题、手抄报、绘/画本
- 兼职赚钱：围绕小说、漫画等兼职领域

小红书AI翻译和封面AI生成

- 小红书海外爆火后，依托大模型，一周快速上线翻译功能，并实现对网络热梗、方言等中式表达方式进行准确翻译
- 依托UGC场景，快速上线封面AI生成功能，根据「写想法」、「问大家」、「涂鸦」三个场景，自动生成笔记封面
- 另外，小红书还上线了「搜搜薯」功能，自动对小红书内容进行总结

飞书多维表格接入DeepSeek-R1满血版

- 飞书一直在大模型应用探索的前沿，在宣布接入DeepSeek-R1满血版前，飞书多维表格已经实现了翻译、总结、分类、提取信息、打标等操作
- 在接入DeepSeek-R1满血版后，飞书多维表格可以快速搭建基础管理系统、更精准地生成营销话术等多项功能，为企业内部运营提供了实例模版



终端新知：大模型改造原有终端，新终端时机尚待成熟

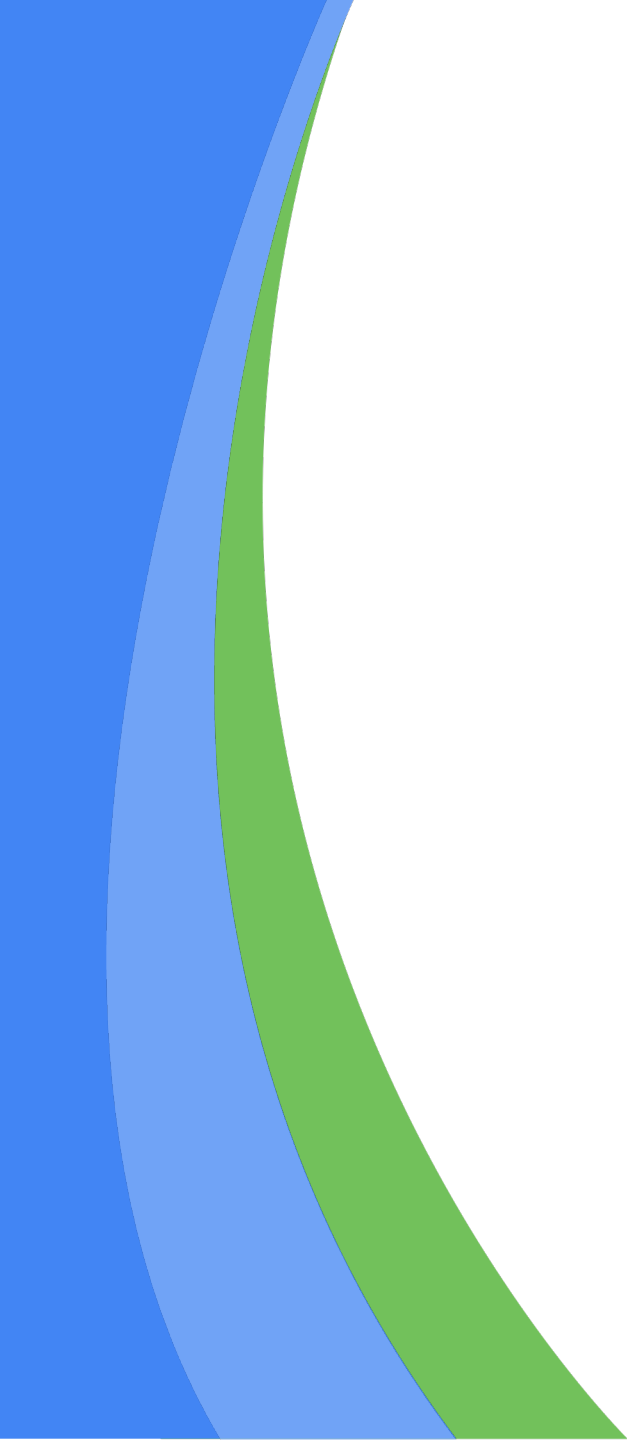
- 在2024年，我们看见了AI眼镜、AI手机、AIPC等原有硬件终端的升级，能够使用电脑和手机的智能体也陆续发布。
- 但AI原生终端，伴随着AI pin的开发商被惠普收购和Rabbit R1受到市场质疑，其进展明显放缓。

眼镜终端大模型接入情况

	雷鸟V3	通义定制模型	场景识别、智能搜索、语音识别
	Rokid Glasses	通义模型	物体识别、智能搜索、实时翻译
	影目科技 INMO GO2	-	实时翻译、演讲题词、离线翻译
	闪极 拍拍镜	数十家模型	实时翻译、语音撰写、AI闪记
	李未可 Meta Lens	自研Wake AI模型	AI伙伴语音交互、快速拍照、智能导航

手机终端大模型接入情况

	 HUAWEI		vivo	oppo	HONOR	
端侧模型	盘古	MiLM2	蓝心	安第斯	魔法	通义
云端模型	盘古 DeepSeek	MiLM2 豆包 DeepSeek	蓝心 DeepSeek	安第斯 豆包 DeepSeek	魔法 百度千帆 DeepSeek	DeepSeek



B端大模型应用发展情况

部署新变：DeepSeek引发一体机采购浪潮，助力大模型快速部署落地

- 从企业级部署方式来看，大型企业更倾向于选择本地部署的方式，端侧部署伴随着设备端硬件条件的成熟，也正在快速拓展。
- 此外，DeepSeek在国内引起舆论关注后，一体机方式因其开箱即用，快速上线的特性，吸引了一批想要快速落地DeepSeek的央国企和地方机构的买单。华为、浪潮、联想等集团也快速推出了DeepSeek一体机产品。

大模型部署方式

企业级部署	本地部署	自有机房
		私有云
		★ 一体机
	云端部署	公有云托管 (API等)
		混合云
	端侧/边缘部署	

DeepSeek一体机采购公告

公告日期	招标单位	项目内容	披露项目预算
2025.3.13	中远海运科技股份有限公司	DeepSeek 一体机及配套服务项目	414万
2025.3.14	数字新疆产业投资集团	DeepSeek 大模型一体机采购项目	
2025.3.14	浙江师范大学	deepseek 一体机集群	350万
2025.3.17	广州医科大学附属脑科医院	Deepseek 一体机采购及软件支持	67万
2025.3.18	日照口岸信息技术有限公司	Deepseek 一体机采购项目	407万
2025.3.19	华润医药商业集团有限公司	DeepSeek一体机/专有云/隔离物理机等部署方案的租赁及运维服务	

DeepSeek一体机供应商

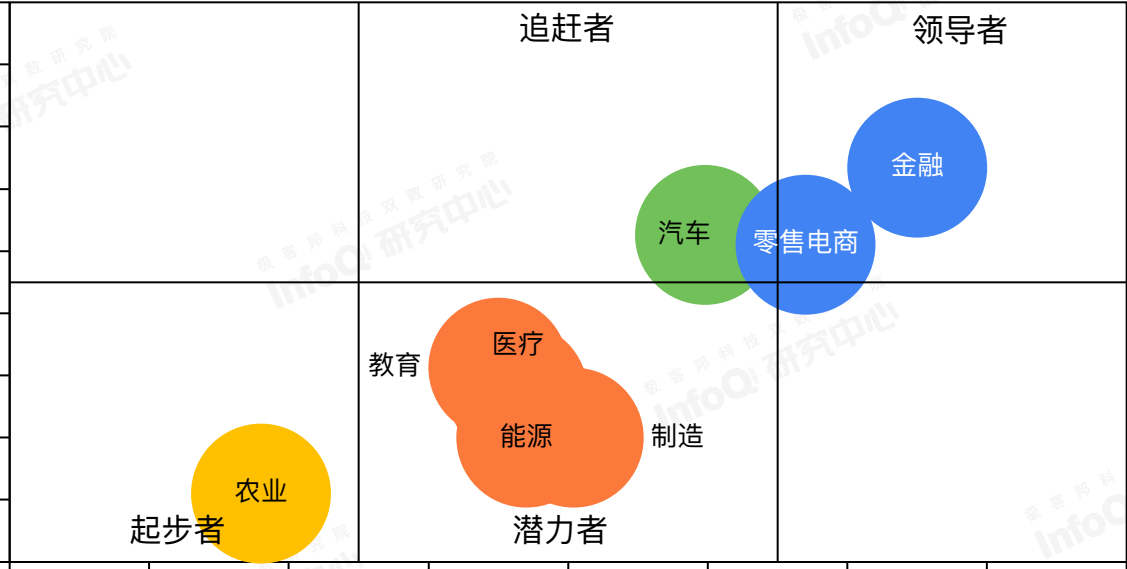
	HUAWEI	华为昇腾
	浪潮信息	浪潮信息
	联想	联想集团
	京东云	京东云
	中国移动	中国移动
	中国电信	中国电信
	新华三	新华三
	中科曙光	中科曙光
	优刻得	优刻得
	科大讯飞	科大讯飞



大模型行业应用现状：行业数字化基础同大模型应用率基本呈现正相关关系，报告选取金融、零售电商、汽车和教育进行进一步解读

- 从整体来看，行业数字化基础同大模型应用率基本呈现正相关关系，其本质是数据治理成熟度与业务系统化能力共同驱动模型落地。
- 在接下来的章节，报告选取金融、零售电商、汽车、教育进行进一步解读。

各行业大模型应用率



各行业数字化基础

数字化基础同大模型应用率基本呈现正相关关系

- 数字化建设中的数据治理和打通，同大模型建设前期所需的数据治理高度相同，这为模型训练提供了可直接复用的高质量数据基础
- 数字化程度高的行业通过长期的业务流程系统化梳理，形成了结构化的业务知识体系和标准化的操作节点，这种业务逻辑的显性化不仅便于大模型快速理解行业特性，更为模型嵌入实际业务流程提供了天然的接口和验证机制

报告从不同象限选取四大行业进行进一步解读

报告接下来从领导者、追赶者和潜力者象限分别选取**金融、零售电商、汽车和教育**进行进一步解读

数据来源：大模型采用率数据来自InfoQ研究中心2024年7月展开的用户调研，N=1166；行业数字化基础来自中关村信息技术和实体经济融合发展联盟、数字化转型指数报告和中国企业智能化成熟度报告等公开数据梳理以及专家评估



中国大模型金融行业应用案例解读

金融行业特征为大模型应用提供丰富可实践场景，同时高安全属性对大模型应用提出更高要求

- 金融行业的复杂度高，但本身多模态数据密集的特点，为大模型的应用提供了大量潜在场景。
- 同时，不同的风控要求和监管环境，也为金融子行业以及细分场景的应用落地产生了不同的影响

业务系统复杂

金融行业整体的业务系统更为复杂，涉及更多业务线、服务种类和客户触点，为AGI技术的应用提供了广阔的舞台

多模态数据密集

金融行业过往良好的数字化基础，让金融行业累积大量的客户数据，这些数据为大模型的技术应用提供了丰富的训练数据源

1

2

金融行业具有大量丰富的可实践场景

3

4

金融行业实践进展速度不一

高监管和合规要求影响场景应用深度

金融业务受到严格的监管，因此在大模型应用落地时需要考虑到合规性和监管融合的问题。不同金融子行业的监管环境对大模型技术的采纳和应用产生影响

优先赋能内部提效

现行应用主要赋能对客服务、数据挖掘、业务助手等环节。在涉及较严监管领域的复杂金融投资建议以及核心分析决策环节主要起到辅助作用。

金融大模型应用场景分析：智能交互与内部运营先行，研究辅助正在快速落地

- 结合金融客户旅程和细分行业场景，在本报告中，我们将大模型在金融行业的应用分类为客户交互、研究交易、内部运营、风控四大类。

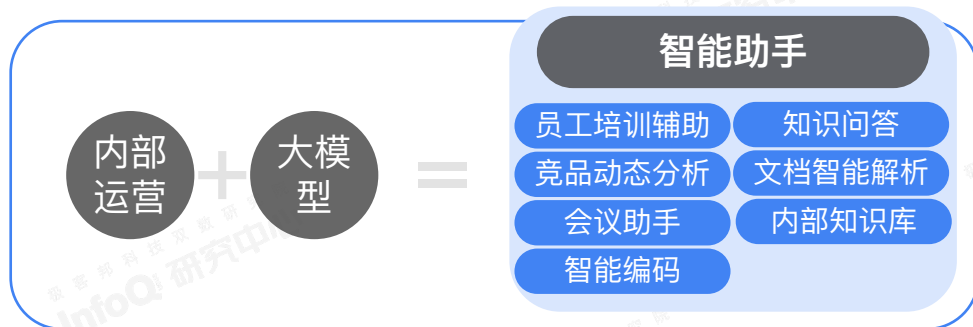
智能交互与服务先行



研究交易正在试点



内部运营先行



浅应用



技术在金融领域应用成熟度模型：致力于运营管理的革新以及业务流程的优化

- 金融行业大模型应用的步伐尚处于探索期，正逐渐向产品测试期发展。金融大模型产品主要被应用于智能客服、智能办公等运营场景，部分头部科技企业在金融核心业务环节有所突破：大模型产品是否能够在金融核心业务中发挥出重要作用，一直是业界热议的话题。

大模型在金融领域应用成熟度模型

应用探索

研究交易-投资交易决策
研究交易-大宗商品交易
研究交易-实时情报分析
研究交易-金融领域专家智能体
风控-员工行为异常检测

产品测试

客户交互-数字员工/智能机器人
研究交易-交易询价助手
研究交易-智能投研
研究交易-量化策略分析及优化
研究交易-市场情绪分析
研究交易-报告生成（较复杂应用）
内部运营-员工培训辅助
内部运营-竞品动态分析
风控-市场风险预警

市场投放

客户交互-智能投顾
客户交互-保险方案规划
客户交互-理财规划
客户交互-智能客服
客户交互-营销内容生成
研究交易-报告生成（简单应用）
内部运营-会议助手
内部运营-智能编码
内部运营-知识问答
内部运营-文档智能解析
风控-信用评估及预测

成熟应用

客户交互-智能外呼
客户交互-资讯整理
客户交互-资讯分析
客户交互-智能产品解读
研究交易-交易执行监控
内部运营-内部知识库
风控-欺诈检测与预防

合作模式：内部提效和研究辅助倾向合作落地，金融行业大模型落地方式同机构规模高度相关

- 从合作模式来看，金融行业在内部提效和研究辅助更倾向于合作共建，在招投标环节也多数注明供应商须有相同领域的落地经验。在智能风控场景下，更倾向于保持技术自主可控。除此以外，金融行业的应用也同机构规模高度相关，六大行以外的商业银行、以及证券业都更倾向于轻量自研+外部合作落地。

交互与服务保障数据自主，但可合作落地

智能交互与服务

数字员工/智能机器人

智能投顾

保险/理财方案规划

智能客服

营销内容生成

智能外呼

资讯整理

资讯分析

智能产品解读

数字员工/智能机器人

广发银行 | CGB

中国人寿集团成员单位

科大讯飞
iFLYTEK

国泰海通
GUOTAI HAITONG

腾讯云

智能客服

浦发银行
SPD BANK

Baidu 百度

中国邮政储蓄银行
POSTAL SAVINGS BANK OF CHINA

新晨科技
XINCHEN TECH

资讯整理&分析

招商银行
CHINA MERCHANTS BANK

阿里云

倾向合作采购

智能助手

员工培训辅助

知识问答

竞品动态分析

文档智能解析

会议助手

内部知识库

智能编码

智能编码

浙商证券

阿里云

华泰证券

科大讯飞

中国邮政储蓄银行
POSTAL SAVINGS BANK OF CHINA

智谱·AI

知识库

CMS 招商证券

智谱·AI

上海银行
Bank of Shanghai

Baidu 百度

研究辅助倾向合作，智能交易应用初期

研究辅助与智能交易

投资交易决策

交易询价助手

大宗商品交易

实时情报分析

金融领域专家智能体

量化策略分析及优化

报告生成

市场情绪分析

交易执行监控

交易询价助手

ICBC 中国工商银行

中国银河证券
CHINA GALAXY SECURITIES

百度智能云

浦发银行
SPD BANK

达观数据
DATA GRAND

报告生成

北京银行
BANK OF BEIJING

中科闻歌

招商银行
CHINA MERCHANTS BANK

阿里云

技术自主可控

智能风控

员工行为异常检测

信用评估及预测

市场风险预警

欺诈检测与预防

欺诈检测与预防

ICBC 中国工商银行

中国邮政储蓄银行
POSTAL SAVINGS BANK OF CHINA

招商银行
CHINA MERCHANTS BANK

平安银行

工商银行：全面推广并应用大模型技术体系，加速金融科技与业务的无缝融合

- 工商银行在2023年已迈出了行业先驱的步伐，率先构建了全栈自主可控的千亿级参数规模的金融大模型技术体系，该体系由“三大技术支撑”、“八大业务领域创新应用”以及“五大应用范式”共同组成，并在客服、营销、运营和风控等多个业务领域得到了广泛应用。工商银行将继续深化在“金融五篇大文章”的实践，致力于打造以大模型为核心的“AI金融”生态系统。

工商银行金融大模型技术体系

八大业务领域

对公信贷

运营管理

远程银行

金融市场

内控合规

人力资源

智能办公

智能研发

五大应用范式

知识检索

智能搜索

文档编写

数据分析

智能中枢

三大技术支撑

高效算法

千亿级金融行业大模型，具有出色的金融知识理解和生成能力

可靠算力

行业首个国产千亿级大模型算力底座，支持千亿级大模型二次训练和大规模并发推理需求

全面数据

将金融数据、企业内部专业数据安全快捷地融入大模型，为业务创新应用提供数据支持



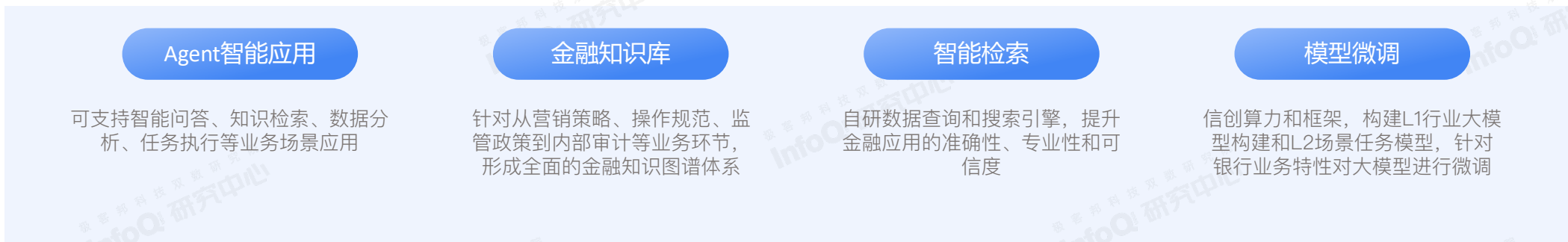
北京银行：深度应用AIGC技术，加快构建大模型应用体系

- 随着银行业务数字化转型的步伐不断加快，人工智能技术正迅速融入到各个核心业务领域。北京银行自2021年起便着手实施“京智大脑”项目，旨在满足不同业务线对智能解决方案的需求，同时不断提高对人工智能核心技能的掌握与应用能力；
- 2023年2月，北京银行开始探索大模型在金融领域的应用，同年4月上线首个大模型问答工具京智助手，8月启动AIB金融智能应用平台建设工作，主要面向营销和销售场景。

北京银行成为最早探索AGI技术应用的银行之一



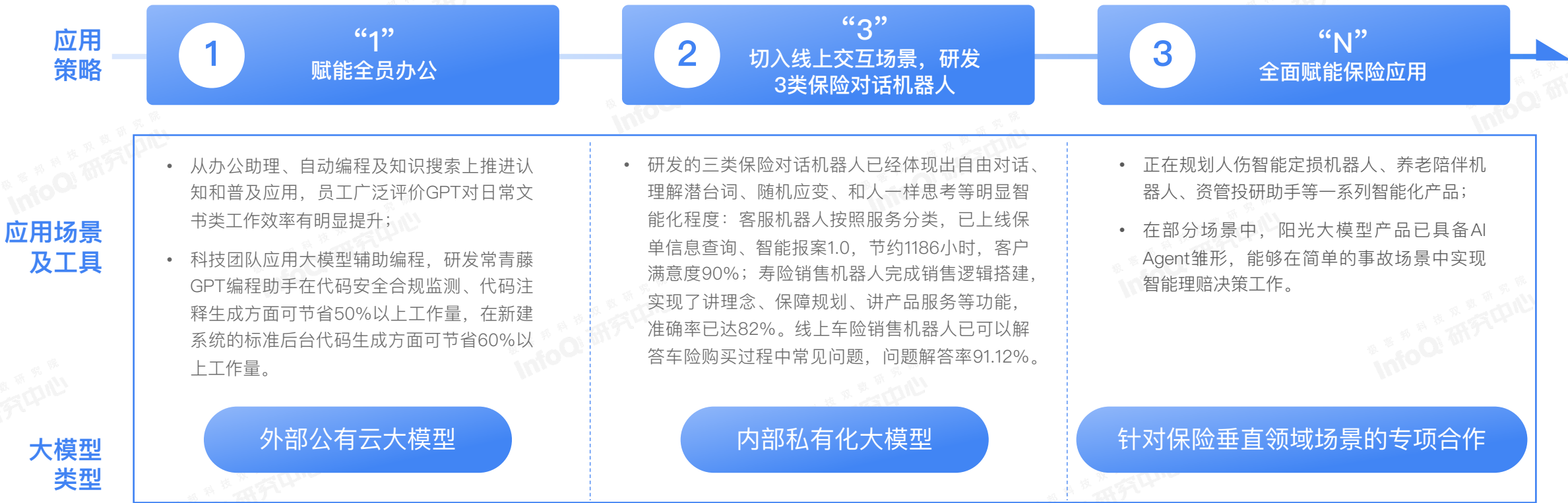
四项技术确保大模型与业务无缝融合，提升灵活性、可解释性、安全性和可适应性



阳光保险：大模型应用的建设遵循从普及应用到垂直深入的策略，全面覆盖业务场景

- 阳光保险已经聚焦大模型进行了诸多尝试，迅速确立了大模型平台与大数据平台的核心地位，深入探索大模型的能力边界：如在双底座建设中积极探索了包括销售、服务和管理等业务领域在内的各种应用场景。此外，阳光保险还基于 Transformer 结构研发了“阳光正言”大模型，未来将继续积极开发高价值的应用场景，最大化大模型技术的潜力和效益。

阳光保险建立“1+3+N”的应用模式





中国大模型零售电商行业应用案例 解读



大模型为零售电商带来个性化消费体验、内部运营效率提升及数据价值激活的潜力，但由于不同模态生成能力及场景价值量化难度的差异，各应用场景的实施动力和深度不尽相同

零售电商行业需要技术创新解决行业痛点

本地市场趋于饱和，零售电商行业的获客成本、供应链运营成本导致竞争进一步加剧，亟需通过技术建立个性化消费体验和内部运营提效

海量数据价值挖掘需求

零售电商行业中，用户行为、商品评论、交易记录、门店运营等非结构化数据占比高，大模型可充分捕捉消费行为中的深层规律

1

2

零售电商行业为大模型的应用提供了大量潜在场景

3

不同模态生成能力存在梯度差

文本问答和营销文案等文本模态生成已基本迈进成熟区，商拍为代表的图像生成稳定性仍有提升空间

4

场景价值可量化驱动

营销、客服、运营等有明确KPI或ROI量化价值的场景优先落地

零售电商行业大模型应用进展速度不一

零售电商大模型应用场景分析：零售电商场景核心升级点围绕效率提升与体验优化

- 总体来看，在零售电商行业的大模型应用正推动着从前端用户体验到后端运营效率的全面革新。围绕效率提升和体验优化，为零售行业带来了深刻的变革。
- 实体零售的重点在于空间限制下的体验升级与库存动态管理，而对于电商而言则是从“流量争夺”到“需求创造”，大模型让平台进化为用户与商品的智能匹配中枢。





技术在零售电商领域应用成熟度模型：内容生成、智能客服、投放推荐成为应用“先锋队”

- 尽管零售电商领域整体上还处于成长阶段，但某些特定应用如智能推荐和个性化营销已经相对成熟，而其他应用如基于Agent的平台商家助手和供应链管理则仍在发展中。
- 生成能力则促进了AI商拍、营销物料生成等全新场景的诞生和发展。

大模型在零售电商领域应用成熟度模型





合作模式：零售场景整体战略自控，但细分场景合作落地；电商平台多依托自研模型快速落地商家服务，同时商拍、跨境服务等创新业态诞生新供应商

整体自控，局部合作

重塑空间价值和即时体验

- 动态库存优化
- 库存营销活动建议
- 动态虚拟货架
- 品牌标识设计
- 用户需求挖掘
- 门店数据分析



实体零售企业在战略层面倾向于自主把控全局，但在具体执行层面，尤其是营销与用户运营等专业领域，实体零售企业普遍引入技术服务商合作

平台自研，创新业态诞生

用户和商家的智能伙伴

- AI买手
- 基于Agent的平台商家助手
- 虚拟模特试穿图生成
- 基于Agent的智能投放
- 产品背景图调整
- 品牌标识设计
- 用户需求挖掘
- 营销物料/话术生成
- 数字人导购/主播
- 商品智能推荐
- 运营话术训练
- 多语言商品文案调整
- 社群内容智能生成

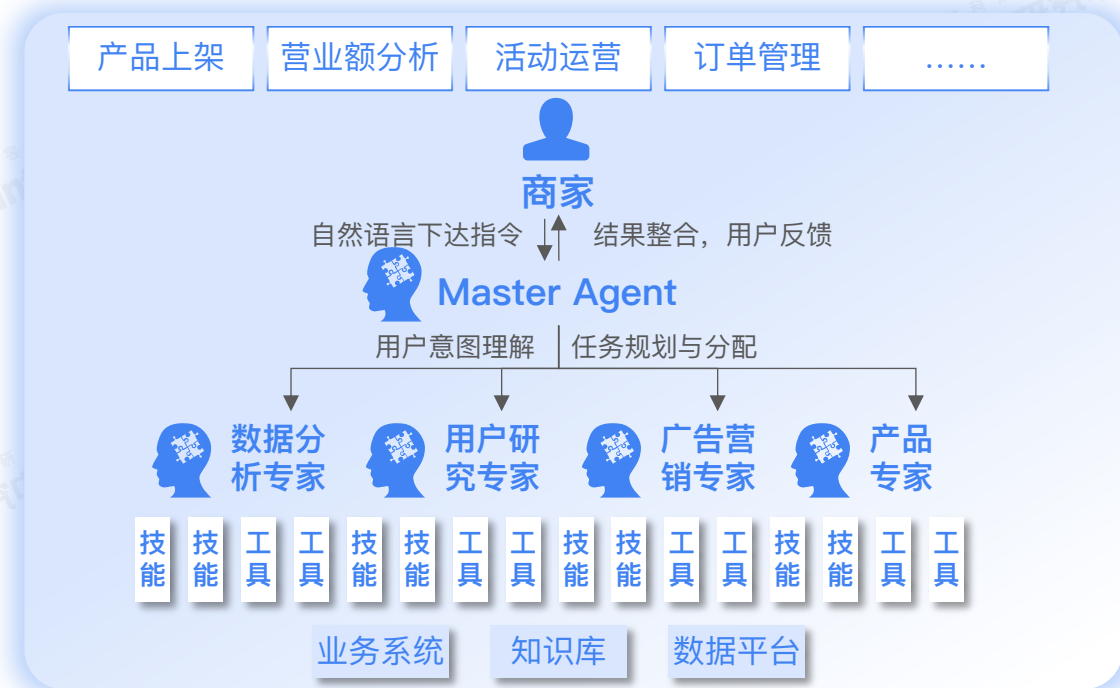
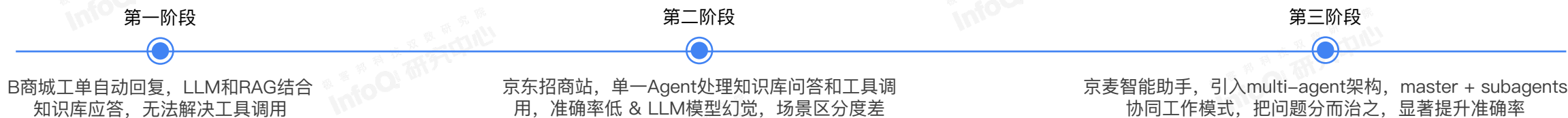
以阿里、抖音、京东为主的平台正在快速自研落地商家服务

此外，以营销、客服为代表的垂类应用供应商也纷纷引入大模型技术，提升产品竞争力

商拍、跨境等解决方案也诞生了一批创新业态供应商

京东商家助手：助力商家打造“更快运营、更好服务、更省成本”的开店体验

- 京东商家助手是京东基于京麦平台，面向商家端推出的智能助手，以帮助商家解决在经营过程中面临的多方面问题。
- 目前，商家智能助手采用的多智能体协同的技术，决策准确率达90%以上，商家单个系统操作时效快至秒级，助力商家打造“更快运营、更好服务、更省成本”的开店体验。



秒级响应，7*24小时在线的定制管家

- 不受限于平台客服的回应时间，自主完成数据查询、运营活动规则查询等操作
- 相当于拥有一个定制的商家知识助理：为商家提供最新的进驻商城的规则条款、营销活动的内容和规则等

多智能体组建专业商家团队，高效经营决策

- 解决商家在电商运营中的复杂决策问题，覆盖从商品发布、订单管理、客服沟通到数据分析的全流程，并为商家提供如营销投放、商机词推荐等经营功能
- 相当于拥有一个定制的商家经营团队：包含数据分析专家、用户调研专家、广告营销专家和产品专家在内的日常经营团队



中国大模型汽车行业应用案例解读



汽车行业为大模型应用提供了大量潜在场景，但受限于安全性要求和用户需求急迫性，应用进展速度不一

- 汽车行业本身产业链复杂度高，但本身多模态数据密集的特点，为大模型的应用提供了大量潜在场景。
- 但对于智驾和生产这类对安全性、可靠性要求高的场景而言，其应用路径和进展都会受到影响。除此之外，智能座舱场景，因为较为明确的用户需求，为大模型在该类场景的应用提供了充足动力。

1 产业链复杂度高

汽车产业覆盖「研发-生产-营销-用车-售后」全生命周期，涉及主机厂、Tier 1/2供应商、经销商、充电运营商、司机、乘客等多元主体

2 多模态数据密集

汽车行业在整个产业链环节积累了海量多模态数据（如传感器数据、图像、语音、文本等），这些数据为大模型的训练提供了丰富的多模态资源

汽车行业为大模型的应用提供了大量潜在场景

汽车行业应用进展速度不一

3 部分场景安全可靠性要求严苛

对于智驾、生产等相关场景而言，大模型需在准确性、时延性、稳定性上有着更高的要求，会影响该类应用的落地进度

4 用户需求驱动

对于智能座舱场景而言，用户逐渐期待更具情感和个性化的服务体验，用户需求结合技术成熟度，驱动主机厂优先探索智能座舱

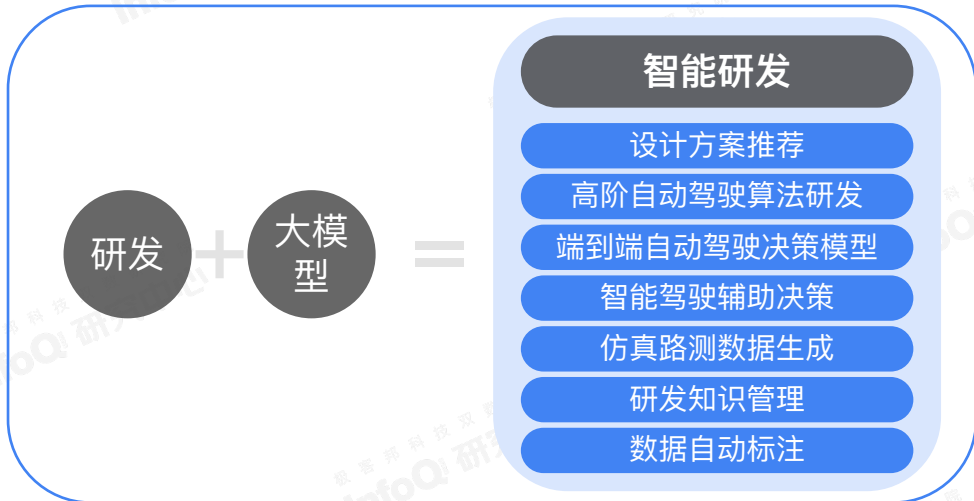
汽车大模型应用场景分析：售后、营销、用车先行，辅助研发，生产场景仍在探索试点

- 结合汽车整体产业链，在本报告中，我们将大模型在汽车行业的应用分类为研发、生产、售后&营销、用车四大类。研发场景虽然对汽车行业至关重要，但受限于技术成熟度，大模型仍然以研发知识管理或辅助决策等辅助作用为主。生产场景而言，大模型在该环节的提效相较于传统AI尚未明确，因此仍在探索试点应用。

售后
& 营销
先行



研发
仍以
辅助
为主



用车
先行



生产
仍在
探索

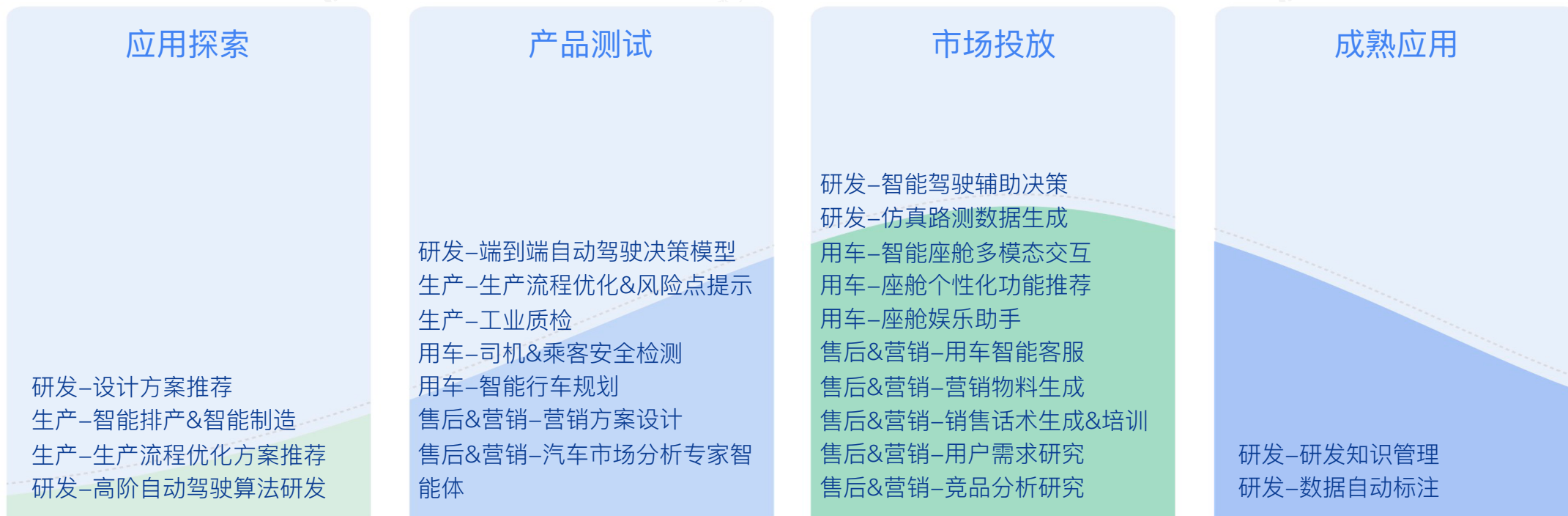




技术在汽车领域应用成熟度模型：辅助研发应用迈向成熟，智能座舱及智能驾驶进展迅速

- 研发知识管理和数据自动标注，是本轮大模型浪潮最开始探索的典型应用，经历近3年的探索，目前已迈向成熟。伴随着视觉、语音模型的技术探索以及终端应用的积累，智能座舱应用已迎来全面升级。
- 此外，以FSD为代表智能驾驶中端到端模型进展迅速，国内部分车型也已逐渐开始以「端到端感知+VLM」的方案陆续展开上车试验。

大模型在汽车领域应用成熟度模型





合作模式：把握智驾核心技术，其余场景可合作可自研，努力找寻技术控制力和商业化效率的平衡点

- 从合作模式来看，除了智驾算法外，售后&营销、智能座舱均有主机厂和科技企业合作以及主机厂自研两条路线。这意味着主机厂在把握核心智驾技术外，在其余的场景中积极寻求合作与自研的平衡，通过差异化路径实现技术控制力与商业化效率的最优解。

售后 & 营销 倾向合作

智能售后&营销

- 营销方案设计
- 汽车市场分析专家智能体
- 智能客服
- 营销物料生成
- 销售话术生成&培训
- 用户需求研究
- 竞品分析研究



深空灵智

用车 自研合作各家路径不同

智能助手

- 司机&乘客安全检测
- 智能行车规划
- 智能座舱多模态交互
- 座舱个性化功能推荐
- 座舱娱乐助手



问界M7
阿维塔12
享界S9



自研



自研

智驾 算法研发是根本，以自研为主

智能研发

- 设计方案推荐
- 高阶自动驾驶算法研发
- 端到端自动驾驶决策模型
- 智能驾驶辅助决策
- 仿真路测数据生成
- 研发知识管理
- 数据自动标注

TESLA



生产 仍在探索

智能制造

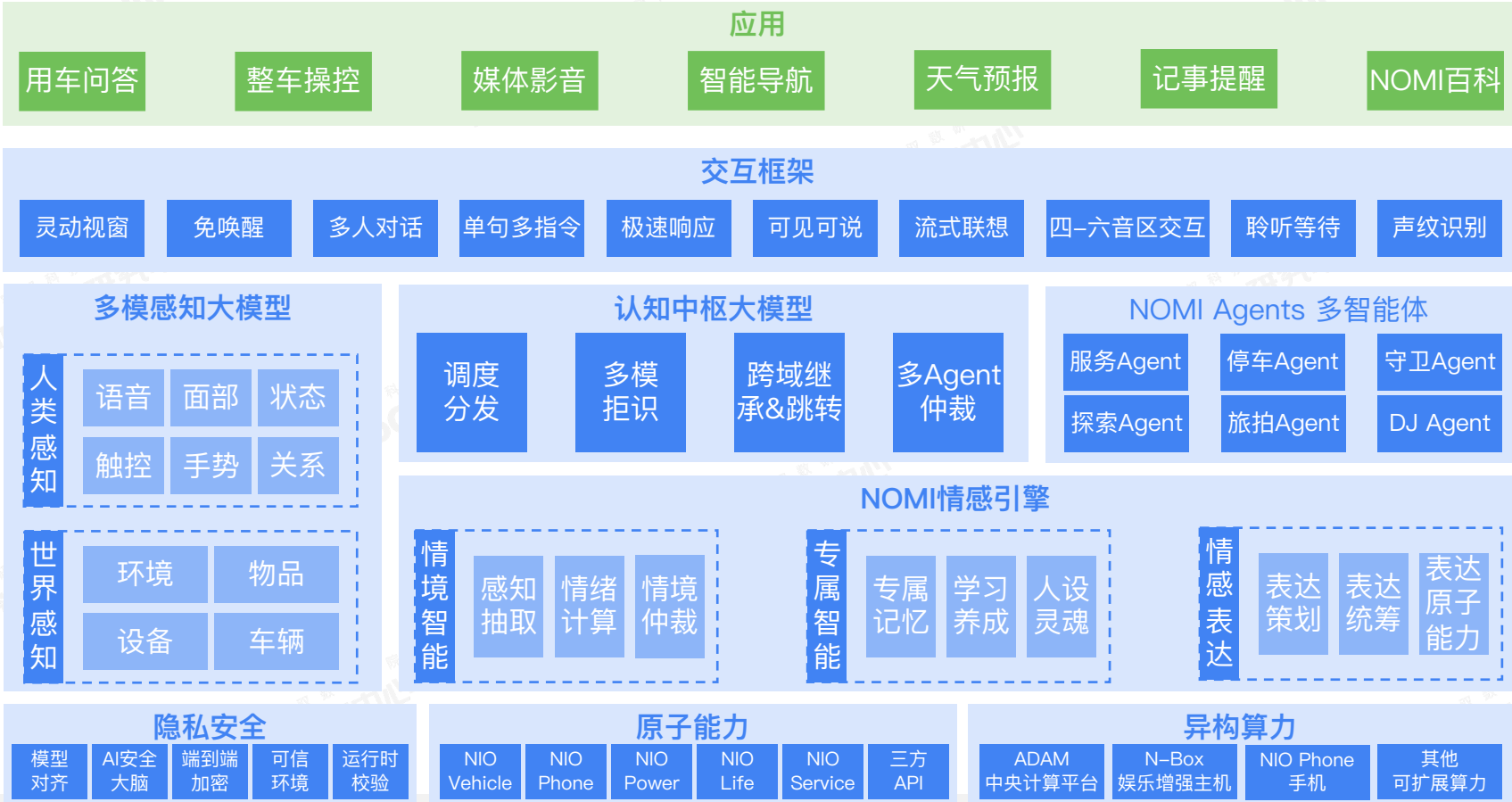
- 智能排产&智能制造
- 生产流程优化方案推荐
- 工业质检





蔚来NOMI GPT：智能座舱从“功能执行”到“情感陪伴”的体验跃迁

- 2024年，蔚来宣布推出NOMI GPT，并于4月正式上车。2024年7月，Banyan 3基于AI大模型能力和原生智能系统，打造了全新的NOMI GPT架构。目前，NOMI GPT通过构建多模感知、认知中枢、情感引擎和多智能体等技术框架，支撑了NOMI GPT在智能座舱中各项应用，



超2000种

技能指令

超96.8%

对话判断准确性

超6290万次

用户互动

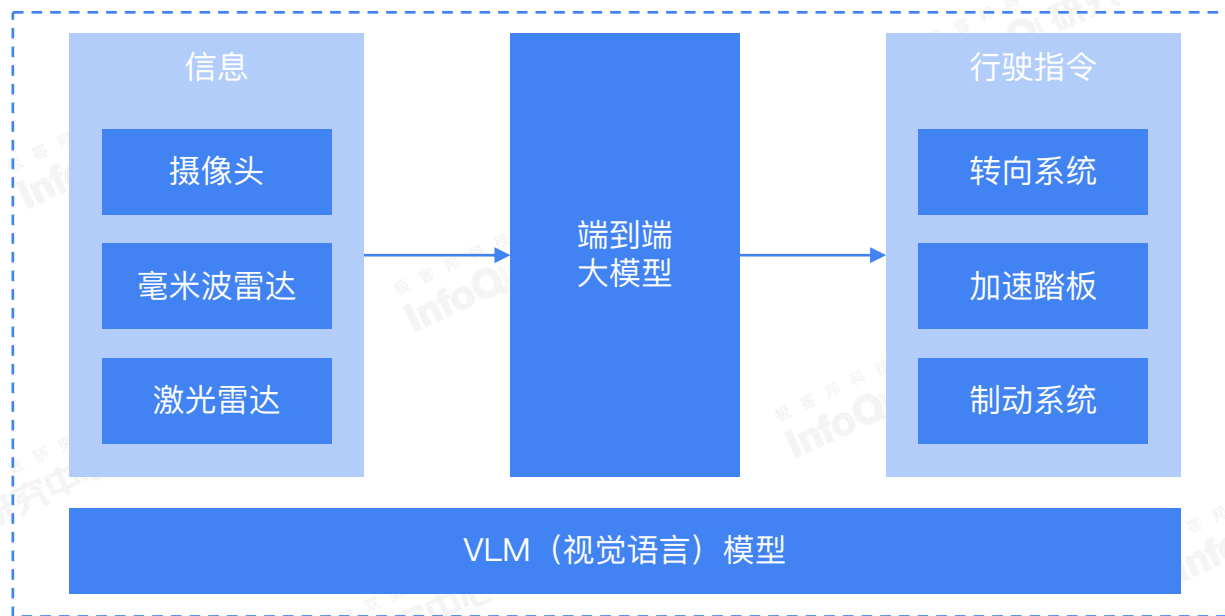


小米汽车：智能驾驶探索端到端大模型+VLM方案，保证下限提升上限

- 小米汽车2024年3月，推出第一款车型小米 SU 7，其智驾方案选择了高精地图+模块化版本；2024年5月迭代至轻图+模块化版本；截至2024年11月，小米官宣其HAD全面接入端到端大模型，12月宣布接入VLM视觉语言大模型。至此，小米汽车智驾方案在半年内经历3次转变，并最终形成端到端大模型+VLM的架构。

- 2024年11月
小米端到端模型已学习**300万**的视频片段
- 2024年12月
开启先锋版推送
- 2024年年底
小米端到端模型已完成**1000万**视频片段的学习
- 2025年2月
开启万人团测试

① 端到端大模型：减少原有感知-预测-规划过程传导中造成的信息丢失和延迟



② VLM模型帮助端到端大模型理解复杂场景和罕见物体，帮助端到端大模型做出更合理的决策

③ VLM模型在识别坑洼路面、施工场景、丁字路口等路面情况后，提醒驾驶员注意



中国大模型教育行业应用案例解读



大模型为教育行业提供了理念实践和数据激活的可能性，但技术应用需要边界约束严谨谨慎对待技术落地

- 首先，大模型为教育行业实践部分教育理念，并且系统激活原有大量沉睡数据提供了条件。但不同类型教育的重点有所区分，同时对于教育涉及的伦理、价值观等内容也需要严格把控。

教育理念的需求驱动

个性化学习与因材施教虽被广泛倡导，但受限于教育资源分配不均与教师能力边界，始终难以规模化落地

1

多模态数据密集

教育过程中产生的文本（教材/作业）、语音（课堂对话）、图像（实验操作）、视频（技能演示）等多元数据，长期分散于孤立系统中，未能形成价值闭环，大模型应用提供将沉睡数据系统激活的动力

2

大模型为教育行业提供了理念实践和数据激活的价值实现

3

教育重点分化

对学前、基础、高等、职业、素质等各类教育类型，教育重点存在分化，对大模型提出了差异化的能力要求

4

教育伦理约束

部分教育场景面向的是尚未形成自我判断和价值观的未成年学生，乃至对成年人，也需避免技术滥用导致认知偏差或心理伤害

教育行业的大模型应用速度不一



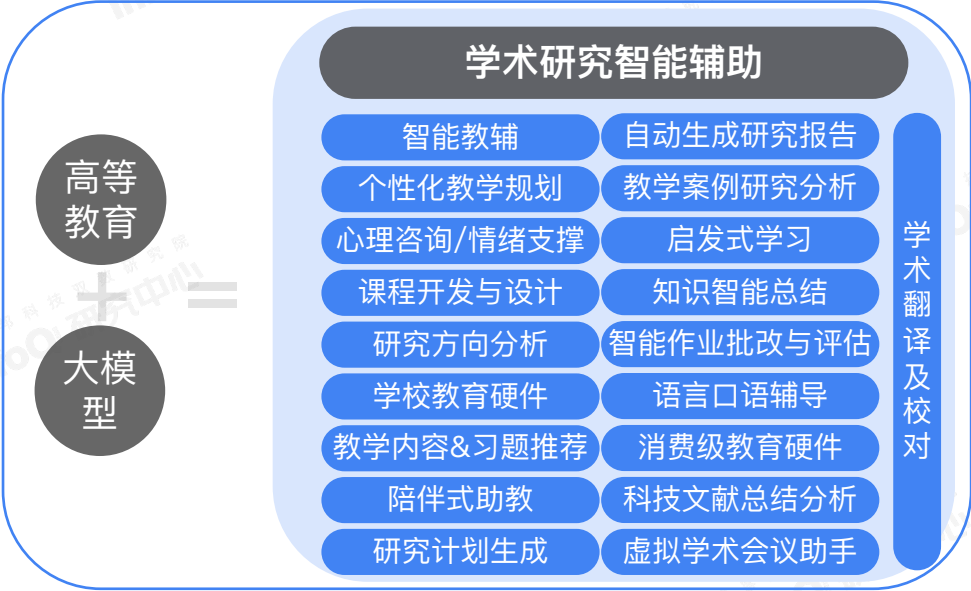
教育大模型应用场景分析：学前教育主打陪伴启蒙，基础教育尝试因材施教，高等教育重点聚焦知识学习和学术研究，职业/素质教育注重知识和实操的结合探索

- 围绕不同的教育类型，大模型所结合的应用场景也有所不同，学前教育主要通过多模态人机互动构建低龄儿童的认知启蒙体系，基础教育定位于个性化、因材施教，应用场景集中于学习内容个性化规划、习题陪伴解答以及优化建议提供；高等教育则更注重启发式学习以及辅助提升科研效率，职业及素质教育更关注将知识与实操相结合的方式。

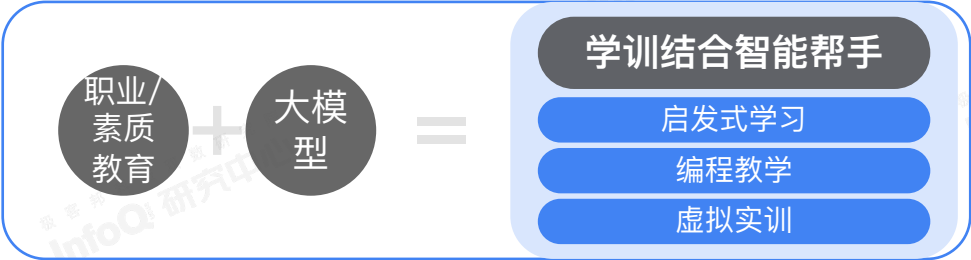
因材施教式学习



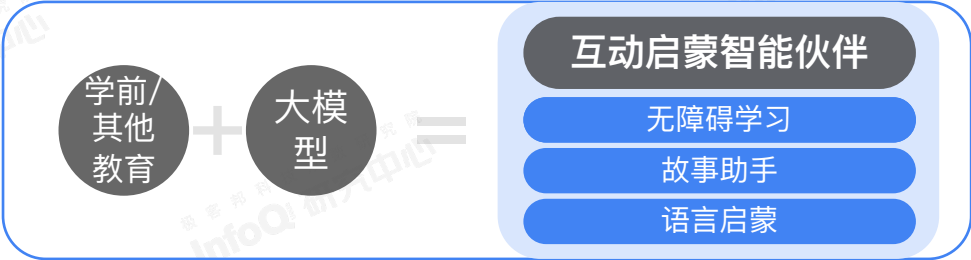
启发思考
科教融合式学习



知识实操



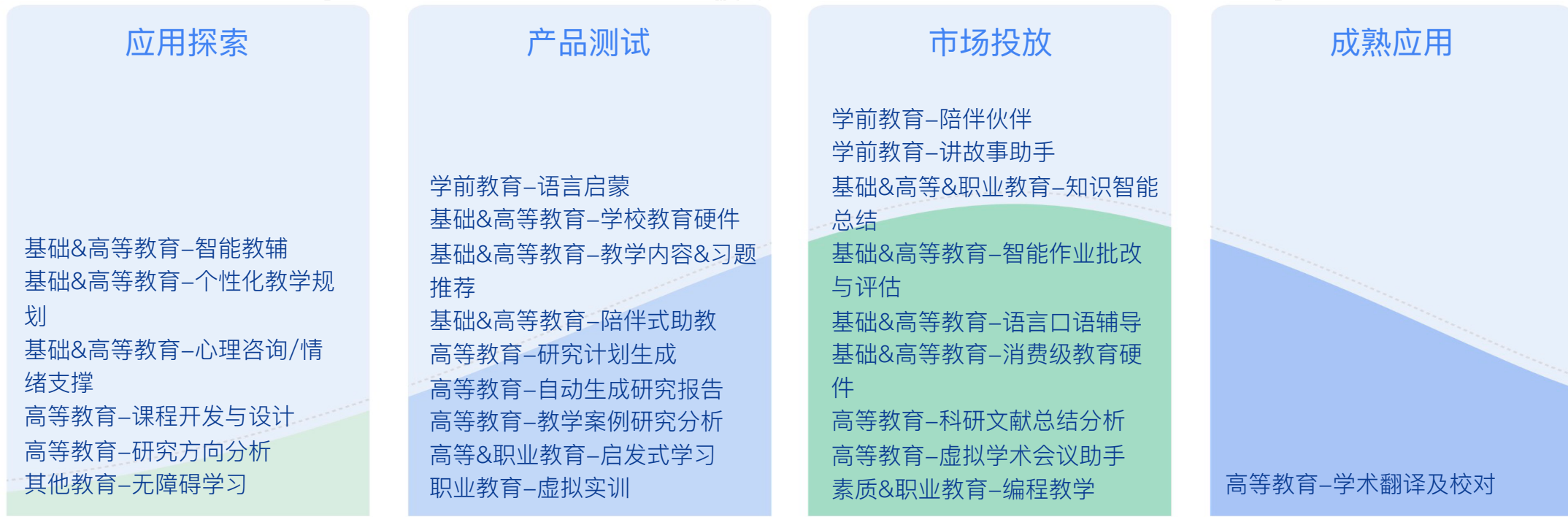
陪伴互动式学习



技术在教育领域应用成熟度模型：大模型在教育行业多担任助手，决策型和全局型

- 整体而言，大模型在教育行业仍主要以辅助、助手的角色尝试应用，例如科研文献的总结和分析，起决策作用和全局型应用仍在探索阶段。例如，课程开发与设计，大模型目前在教学案例研究分析这一单一环节已有XX，但完整的课程开发与设计还需要教师自我根据大模型的产出内容进行调整。

大模型在教育领域应用成熟度模型





合作模式：消费端升级原有终端功能，学校端响应政策号召，多同科技厂商合作探索大模型应用落地

- 消费端，多将大模型作为升级卖点的一部分，例如学习机加入互动陪伴的功能等，但目前存在部分家长对大模型功能的认同度不高。基础教育、高等教育中，学校侧多同积极响应国家政策，同科技厂商合作探索大模型应用落地

教育类型	落地形式	合作形式			
学前教育	玩具、APP等形式	玩具、APP等形式	 淘云科技 阿尔法蛋	 跃然创新 BubblePal	
基础教育	学习机等消费机硬件、 或作为智慧教室解决方案的一部分	学校多同科技厂商合作探索， 或直接进行采购	 科大讯飞 学习机	 作业帮 学习机	 重庆南开高中 自研教育大模型
高等教育	在线平台、独立应用	高校多同科技厂商合作探索	 同济大学 CivilGPT等	 北京大学共同研发 北大智学等	
职业/素质	线上学习平台APP， 实训环境搭建	线上学习平台APP， 实训环境搭建	广东科学技术职业学院牵头研发 知行大模型	商丘师范学院师范生职业技能训练中心 AI实训室	

北大智学：以北大内部作为试点，深入实施教育部人工智能赋能教育行动

- 北京大学软件工程国家工程研究中心联合北京大学教育学院共同研发并推出的基于博雅教育大模型的“北大智学”教学平台和课程教学创新项目。
- 目前，当前项目已进入实际验证阶段，“北大智学”平台和课程已以班级为单位面向北京大学本科生和研究生开展教学试点，未来也将在高等院校、职业院校、高级中学等教学机构展开试点。

北大智学设计体系拆解图



北大智学教师和学生两端界面示意图



- 发挥和运用教育大模型能力，对课程重新进行智能化教学设计
- 通过提示词生成和实时问答进行引导式辅导，让学生思考得更主动
- 通过个性化出题和主动追问实现自适应学习，为学生提供情境化的学习环境
- 通过作业自动评阅和教研助手提升教师教学效率



星火科研助手：立足科研论文场景，提升科研效率

- 科技文献大模型-星火科研助手是中国科学院文献情报中心携手科大讯飞股份有限公司共同研发的智能知识服务平台。它基于科技文献资源，提供“成果调研、论文研读和学术写作”等三大科研助手功能。

2023年10月

2023年12月

2024年8月

2024年10月

2024年12月

平台正式上线，并向公众开放

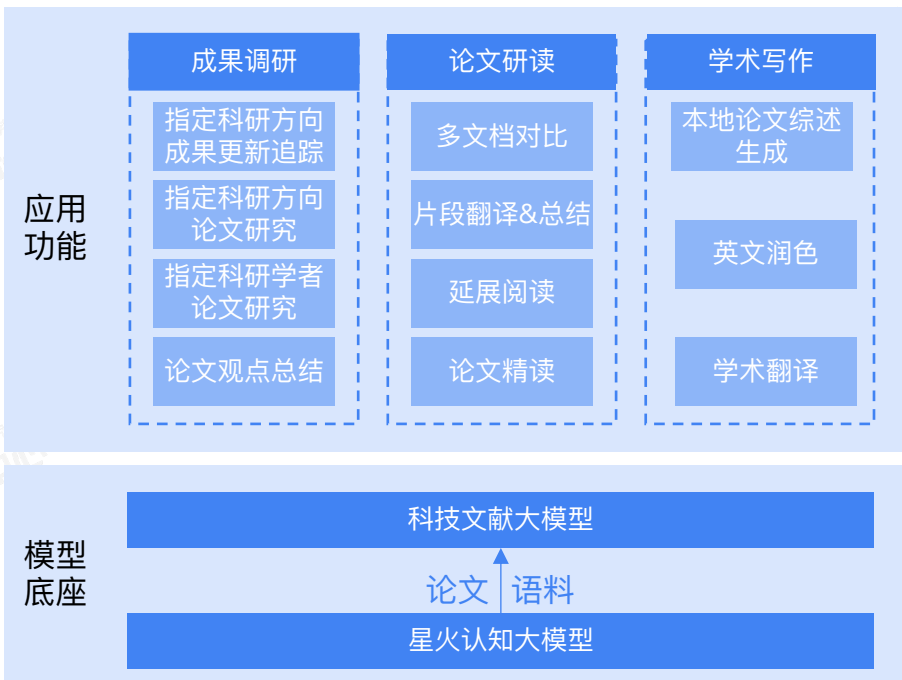
上新多文档对比和综述生成功能

上新领域追踪和多文档问答功能

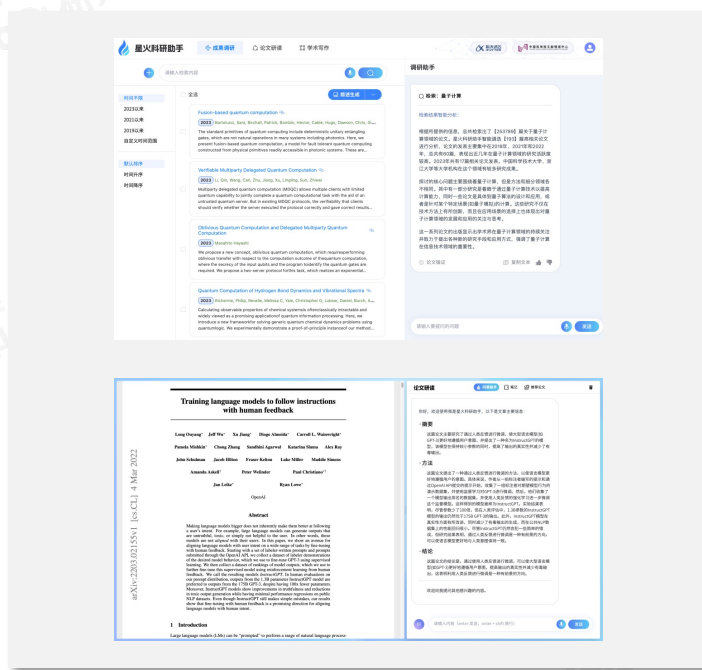
科技文献大模型2.0发布

宣布正在中国科大、中国科学院大连化学物理研究所等科研单位，聚焦化学、化工领域典型场景及下游任务，研制科学知识推演大模型

星火科研助手设计体系拆解图



星火科研助手使用效果示意图



星火科研助手关键成果数据

近100万次的调用量

近100所研究所的落地

90%+的论文研读有效率

90%+的学术写作采纳率



报告预告

InfoQ 研究中心也将继续关注大模型及AIGC领域的应用和产品进展，也欢迎各位行业内的专家就本报告的内容进行交流和讨论，共同助力中国大模型和AIGC领域的发展。

极客邦科技双数研究院
InfoQ 研究中心

中国人工智能AI Agent市场 研究报告 2025

InfoQ 极客传媒

中国人工智能AI Agent市场研究报告2025

历经1年的发展，中国AI Agent市场有哪些新进展？

- 底层模型能力有哪些突破？
- MCP和A2A将带来哪些变化？
- AI Agent 产品形态上又有哪些演进？
- 各行业目前的应用现状又如何？

.....

极客邦科技双数研究院

InfoQ^{ueue} 研究中心

InfoQ 研究中心隶属于极客邦科技双数研究院，秉承客观、深度的内容原则，追求研究扎实、观点鲜明、生态互动的目标，聚焦创新技术与科技行业，围绕数字经济观察、数字人才发展进行研究。

InfoQ 研究中心主要聚焦在前沿科技领域、数字化产业应用和数字人才三方面，旨在加速创新技术的孵化、落地与传播，服务相关产业与更广阔的市场、投资机构，C-level 人士、架构师/高阶工程师等行业观察者，为全行业架设沟通与理解的桥梁，跨越从认知到决策的信息鸿沟。

技术市场趋势洞察



- 市场份额追踪
- 细分市场分析
- 市场规模预测
- 市场分析模型输出

技术市场用户分析



- 用户规模评估
- 用户认知分析
- 用户决策分析
- 用户行为分析

数字化实践趋势分析



- 技术需求洞察
- 技术实践分析
- 应用规划建议
- 发展趋势研判

内容咨询: researchcenter@geekbang.com

商务合作: hezuo@geekbang.com

- 极客邦科技，以“推动数字人才全面发展”为己任，致力于为技术从业者提供全面的、高质量的资讯、课程、会议、培训等服务。极客邦科技的核心是独特的专家网络和优质内容生产体系，为企业、个人提供其成功所必需的技能 and 思想。
- 极客邦科技自 2007 年开展业务至今，已建设线上全球软件开发知识与创新社区 InfoQ，发起并成立技术领导者社区 TGO 鲲鹏会，连续多年举办业界知名技术峰会（如 QCon、ArchSummit 等），自主研发数字人才在线学习产品极客时间 App，以及企业级一站式数字技术学习 SaaS 平台，在技术人群、科技驱动型企业、数字化产业当中具有广泛的影响力。
- 2022年成立双数研究院，专注于数字经济观察与数字人才发展研究，原创发布了数字人才粮仓模型，以此核心整合极客邦科技专业的优质资源，通过 KaaS模式助力数字人才系统化学习进阶，以及企业数字人才体系搭建。
- 公司业务遍布中国大陆主要城市、港澳台地区，以及美国硅谷等。十余年间已经为全球千万技术人，数万家企业提供服务。



促进数字技术领域知识与创新的传播



科技领导者同侪学习社区



数字人才的移动知识库



一站式数字技术学习 SaaS 平台