

AI应用创新与商业化： 新生态与新基建

行业研究部

©2026 iResearch Inc.

ABSTRACT

摘要



AI应用创新与商业化：本报告中AI应用创新包括已有应用AI化升级、AI原生应用创新，并不局限于AI原生App。当前，各行各业AI化势不可挡，AI应用持续向消费级和企业级市场拓展。对话式AI加速了AI应用落地，而Vibe Coding、CUI等技术创新，加速AI应用开发模式革新。与此同时，AI应用商业化落地日益面临全球化、安全合规新规则。



AI应用创新的关键基础设施：AI应用创新的关键基础设施可概括为算力、模型、数据、平台、安全合规五层。其重要性体现在减少模型幻觉、知识资产沉淀与数据治理、上下文管理、Agent协同，以及数据隐私、内容风控、权限控制和审计追踪等方面。



融云：AI时代，融云作为全球AI通信解决方案提供商，定位为打通大模型商业化“最后一公里”的关键角色。依托自研的AICP（AI Communication Platform）基座，产品服务能力覆盖数据、平台、应用、安全合规层，率先布局A2H（Agent-to-Human）与A2A（Agent-to-Agent）通信新品类。



发展趋势：预计未来AI应用创新将深度融入场景，在A2A协同新生态、端侧智能、多模态等方面趋势特征将进一步凸显，具体体现在产业协同从单点人机交互迈向多智能体互联网络，端侧算力与轻量模型推动本地部署，语音、视频、文本等多模态与实时交互增强等方面。同时，AI结果稳定性、安全合规与全球化适配等方面仍面临挑战。

CONTENTS

目录

01 AI应用创新概况

Overview

02 AI应用创新的关键基础设施

Key Infrastructure

03 融云：打通大模型商业化“最后一公里” 的关键角色

Infra for AI-Native Era

04 AI应用创新未来展望

Trends

01/ AI应用创新概况

Overview

AI应用门槛加速下降，赋能存量和增量市场，全面渗透各行各业

当前，受AI供给增强、推理成本下降与接入门槛降低的影响，AI加速深度应用于各行各业。一方面，AI推动CRM、客服、搜索、办公、供应链等存量升级；另一方面，面向不同应用场景，行业涌现出一批AI原生应用，带动增量创新。从市场需求来看，消费级与企业级的AI需求数字化系统同步扩展，AI从单点功能试用逐步走向常态化使用，AI正由“问答工具”进一步演进为任务处理入口，并加快向各行业渗透。



AI成本与门槛快速下降

过去两年，AI的能力供给、推理成本和接入便利性都发生了结构性变化，AI化不再只是头部公司的实验，而正成为各类应用和系统的标准升级方向

- 以GPT为例，2025年8月发布的GPT-5，API首发定价为每百万token输入1.25美元、输出10美元，较2023年3月发布的初代GPT-4（输入30美元、输出60美元）分别下降约96%和83%



两个价值维度下的存量升级与AI原生创新

“提效”价值维度下，AI赋能CRM、客服、搜索、办公、供应链等数字化系统。“创新”价值维度下，涌现一批AI原生应用

- Microsoft 2025 Work Trend Index提出Dow的供应链Agent预计首年可节省数百万美元
- AI原生应用包括ChatGPT、KIMI、Cursor、即梦AI等



消费级与企业级需求同步扩散

AI化全面渗透C端和B端市场，不再局限在单一市场、单点创新，C端用户习惯形成和B端ROI驱动采购正同时发生

- Google在2025年3月表示，已有超过10亿人使用AI Overviews
- McKinsey《The state of AI in 2025》显示，88%的组织已在至少一个业务职能中常态化使用AI

对话式AI成为重要入口

对话式AI将自然语言转化为用户意图，加速AI应用落地

对话式AI是指以自然语言及语音、多模态信息为主要媒介，支持用户表达意图、持续澄清需求、获取内容或触发任务执行的AI交互方式。对话式AI正在从AI应用的“问答界面”，演进为人机协作和Agent任务执行的“意图入口”，在推进AI应用开发和商业化落地过程中发挥了重要作用。其核心价值在于以自然语言承接用户意图，并与GUI、多模态输入、Agent、业务系统、工具调用和人工服务等共同形成可控的交付闭环。

通用型对话AI助手

ChatGPT、豆包、Kimi、Claude等产品以对话为主要交互框架，用户可通过自然语言或多模态输入完成问答、创作、分析与协作。此类产品推动用户形成以自然语言表达需求、与AI迭代协作的使用习惯



应用内嵌式对话入口

电商AI导购、银行智能客服、办公助手、游戏NPC及医疗问诊等，将对话作为用户调用AI能力的意图入口。其核心业务功能并非对话本身，但对话式AI可降低功能发现和复杂操作的门槛



对话体验驱动的垂直产品

在AI陪伴、角色互动、语言学习、部分AI家教产品等垂直产品中，对话式AI本身构成核心体验和价值交付，产品竞争力主要体现为角色人格、情绪反馈、内容连续性和互动质量



对话驱动的任务型Agent

用户以自然语言发起目标或任务，Agent结合模型推理、工具调用、外部系统连接和必要的人工确认，完成多步骤流程并返回结果。对话式AI承担任务发起、过程澄清、授权确认和异常处理的功能，在任务执行与结果交付中实现价值



除通用型对话AI助手更多依赖大模型的推理能力之外，其他三种产品都需要在大模型的基础上，进一步进行产品设计、开发、部署、运营等，需要更多的工程化落地

AI应用创新市场环境向好

AI工具迭代降低创新门槛，市场应用进入场景验证阶段

当前AI应用创新正由“技术可得”加速迈向“场景验证”。从配套供给来看，基础模型、编程生成、Agent编排、内容生成及RAG评测等工具链持续完善，显著降低了产品原型开发、团队组织与方案交付门槛。从市场需求来看，政府对于AI接受度提高、政策支持持续加码，企业与个人用户对AI的认知、接受度和使用频率同步提升，带动相关产品进入更深入的应用验证阶段。

AI技术工具不断迭代，持续降低AI应用创新门槛

分类	代表性产品
基础模型/API 与托管	OpenAI Gemini ANTHROPIC deepseek Qwen 百度千帆 Tencent Hunyuan
AI编程与应用 生成	CURSOR Claude Code CODE BUDDY Qoder CN TRAE
Agent/工作 流编排	n8n LangGraph Dify Coze AI 百炼智能
设计/内容生 成	Canva Midjourney Eleven Labs WPS AI 剪映
后端/RAG/评 测运维	supabase LangSmith zilliz milvus Weaviate
对于AI应用 创新的影响	降低产品原型试错成本 简化开发流程/团队 缩短方案交付周期

终端用户对AI的认知和接受度不断提升

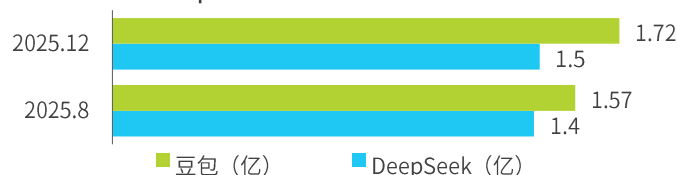
政府

- 多地政府部门已开始把大模型接入政务系统，例如深圳、北京等地将DeepSeek等大模型接入政务系统，用于公文写作、政策解读、政务服务等场景
- 2025年1月，国家出资成立国家人工智能产业投资基金，基金规模达到600.6亿元人民币

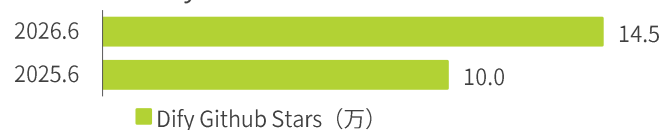
企业&个人

- 根据QuestMobile、GitHub Stars数据，消费者、开发者对于AI的应用和关注不断提升

DeepSeek和豆包MAU变化



Dify GitHub Stars变化



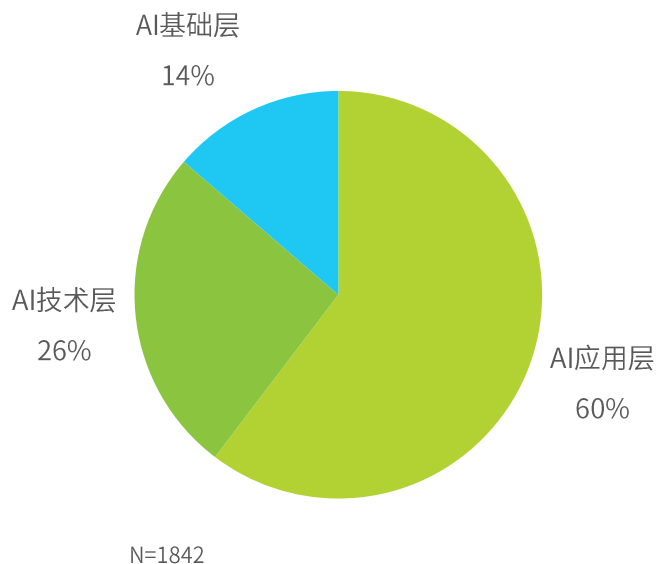
AI创业企业加速应用创新

技术迭代驱动应用创新，应用层成为AI创业最活跃的战场景之一

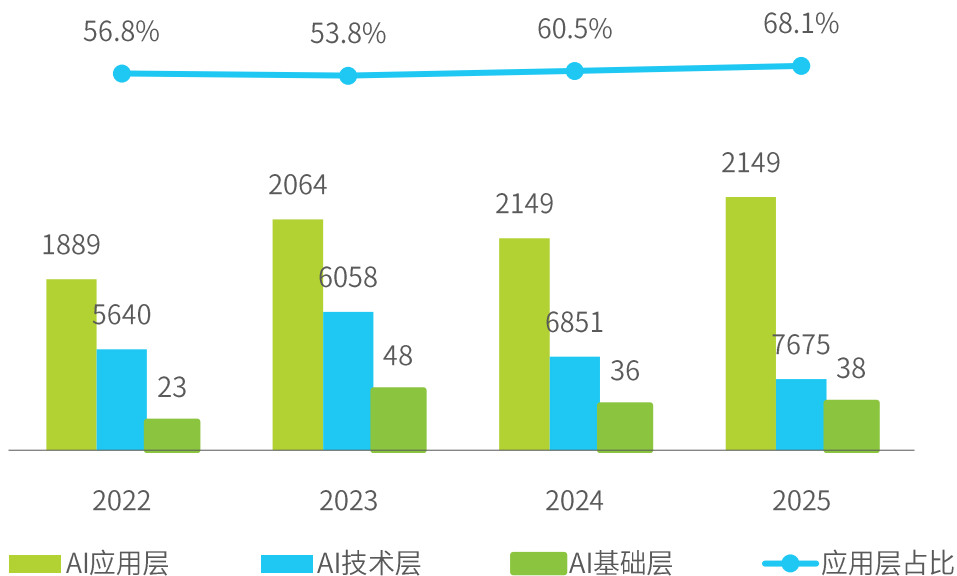
近些年，随着模型能力、开发工具和行业方案持续成熟，技术迭代正更快向产品化与场景化创新传导，推动创业重心由底层能力建设加速转向面向具体需求的应用落地。根据对于2020至2025年间AI创业企业在基础层、技术层、应用层*分布情况来看，近些年AI创业企业主要集中于应用层，整体占比约为60%，明显高于技术层和基础层，可见应用层已成为当前AI创业最活跃的方向之一。从2022至2025年期间来看，应用层企业数量虽有波动，但占比始终保持在53.8%至68.1%的较高区间，2025年进一步升至68.1%。

应用层是AI创新创业最活跃的战场景之一

2020-2025年AI创业企业分布



2022-2025年AI创业企业分布与AI应用层占比



注：此处AI基础层、技术层、应用层是行业通用划分口径：基础层指的是为AI提供算力、数据和基础工具支撑的底层软硬件设施；技术层指的是研发通用AI技术能力和算法模型的环节；应用层指的是将技术层的通用能力与具体行业场景结合，形成面向最终用户的产品、解决方案或服务。

来源：IT桔子，艾瑞咨询整理分析绘制。

AI应用产品：消费级应用

AI演变为基础能力，内容创作、工作提效、情感陪伴是核心需求场景

当前，C端用户对AI应用的态度已经从早期猎奇使用回归理性的实用性选择，其中内容创作、工作提效和情感陪伴是消费端的三大核心场景。全球TOP100 AI应用月活数据显示，一方面，ChatGPT、豆包、DeepSeek等通用对话AI助手的活跃程度持续保持在高位，另一方面，图像/视频生成、工作提效类工具、AI搜索引擎及情感陪伴等细分领域的AI应用也呈现出百花齐放的态势。随着AI能力逐步演变为一项基础能力，图片/视频编辑、教学辅导、饮食健康等纷纷在原有功能上叠加AI模块，以提升用户体验、增强用户粘性并构建差异化竞争力。

2026年6月全球TOP100 AI应用分类型MAU合计

单位：百万	MAU	代表应用
通用AI对话助手	2492.2	ChatGPT、豆包、DeepSeek
AI图像生成/编辑	342.1	即梦AI、Remini、FaceApp
AI搜索/浏览器	222.0	夸克、Perplexity、纳米AI搜索
AI辅助工作提效	156.3	百度文库、DeepL、Notion
AI视频生成/编辑	97.5	PixVerse、Sora、可灵AI
AI角色/情感陪伴	64.3	Talkie AI、星野、猫箱
AI音乐/音频	48.3	Suno、Donna
AI学习/教育	24.3	Gauth、Question.AI
其他	23.8	PictureThis

注释：1) 含全球iOS、海外Google Play及国内Android数据，2) 各赛道MAU为该赛道内APP加总不去重的结果。
来源：AI产品榜·全球总榜（6月版）。

2025年上半年全球新增AI功能的各类APP数量



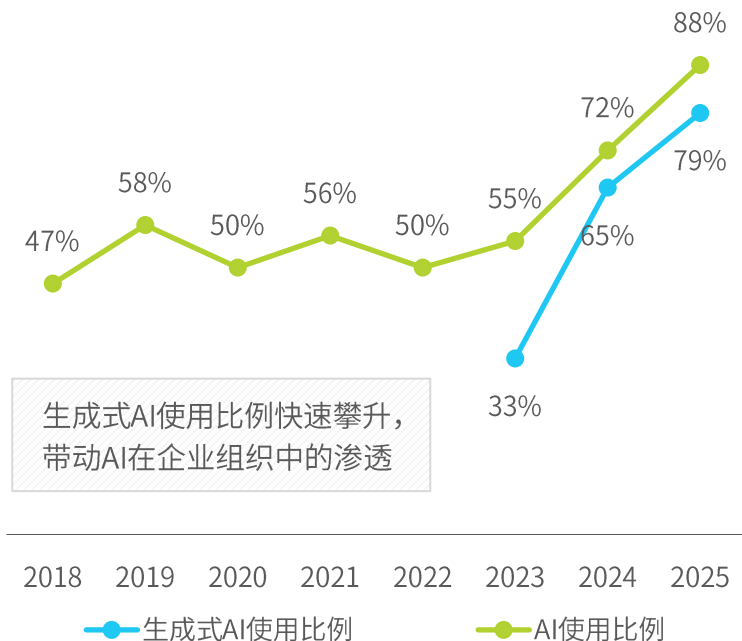
注释：1) iOS和Google Play应用程序应用程序分别计算，2) 剔除AI助手和AI内容生成类，仅考虑非AI原生应用叠加AI功能的情况。
来源：Sensor Tower《State of AI Apps Report 2025》

AI应用产品：企业级应用

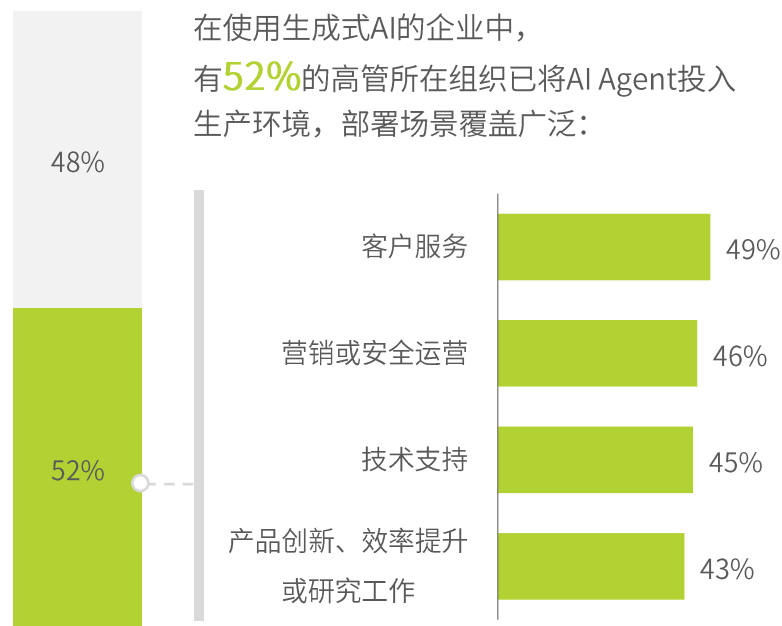
Agent加速向企业端渗透，嵌入企业业务流程，提供自主行动能力

调研数据显示，2025年生成式AI在企业组织中的渗透率由2023年的33%快速攀升至79%，而在使用生成式AI的企业中，过半数已经将AI Agent投入生产环境。相比基础的对话式AI，Agent具备自主行动的能力，能够实现从任务拆解到流程执行的端到端处理，成为企业端AI的核心落地载体。目前，生成式AI率先在客服、营销、研究等强调知识密集交互的场景中实现规模化落地。与此同时，随着AI应用进一步接触企业知识、客户数据和核心业务流程，企业组织对数据安全、权限控制、审计追踪和部署方式的要求也在提升，私有化部署、混合云部署、本地化数据管控等能力正逐渐成为AI应用落地的重要前提。

2018-2025年全球企业AI使用比例



2025年全球企业AI Agent应用情况



注释：企业AI使用比例为至少在一项业务功能中使用AI的组织比例。
来源：麦肯锡《2025年人工智能现状报告》。

注释：Q1：企业内部是如何使用AI Agent的？N=3466；Q2：贵公司已部署的AI代理用于哪些用例？N=1814。
来源：Google Cloud《2026年AI Agent趋势报告》。

AI应用开发：Vibe Coding

Vibe Coding降低开发门槛，个人和小团队成为重要创新主体

在Vibe Coding技术创新驱动下，代码生成、调试与原型搭建能力持续增强，产品开发由传统“写代码”逐步转向以需求表达、快速生成和迭代验证为核心的流程，显著降低AI应用创新团队配置与技术门槛；产品经理、设计师及行业专家也能更深度参与原型构建。相关调研数据显示，实际使用AI编程的人数占比持续提升，个人开发者和小团队正成为AI应用创新的重要力量。



核心带来3方面的变化

交互方式变化	从“写代码”变成“说需求、改提示、看结果”
开发流程变化	从“先设计清楚再编码”，转向“先生成原型再迭代”
角色边界变化	产品经理、设计师、行业专家、非科班创业者，也能直接参与原型构建

代表性产品

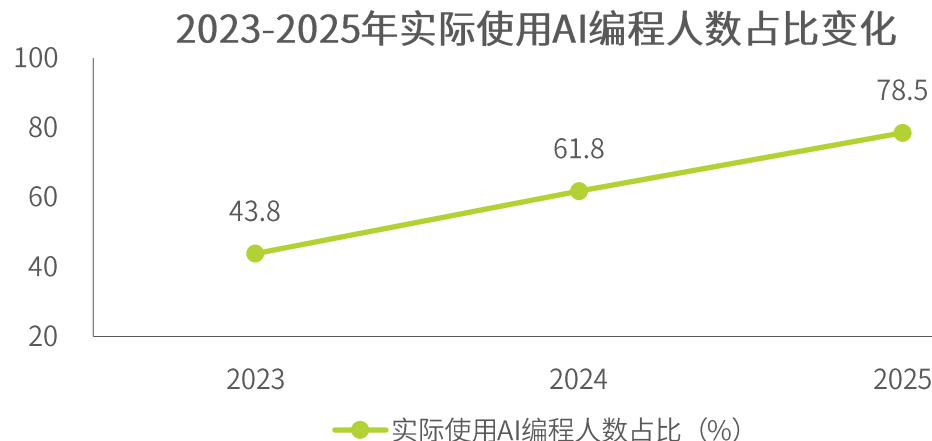
国外	Cursor、Claude Code、Replit Agent、Lovable、Bolt、v0
国内	TRAE、Qoder CN、MeDo、CodeBuddy

使得团队最小有效规模显著下降

原来一个产品至少要“前端+后端+测试+运维”的基础配置，现在其中不少工作被AI压缩为“1个会提需求的人+1个能校验的人”

- 据美国Carta*《Founder Ownership Report 2026》，2025年Carta平台上的初创公司中，约有36%由单创始人领导，相比2024年的31%有明显上升
- 据中国《2026中国OPC行业白皮书》显示，截至2026年4月，全国OPC（One Person Company）数量已突破1600万家，2025年上半年新注册增速高达47%

Stack Overflow开发者调查显示用AI开发的人群基数在急速扩大，如下图所示：



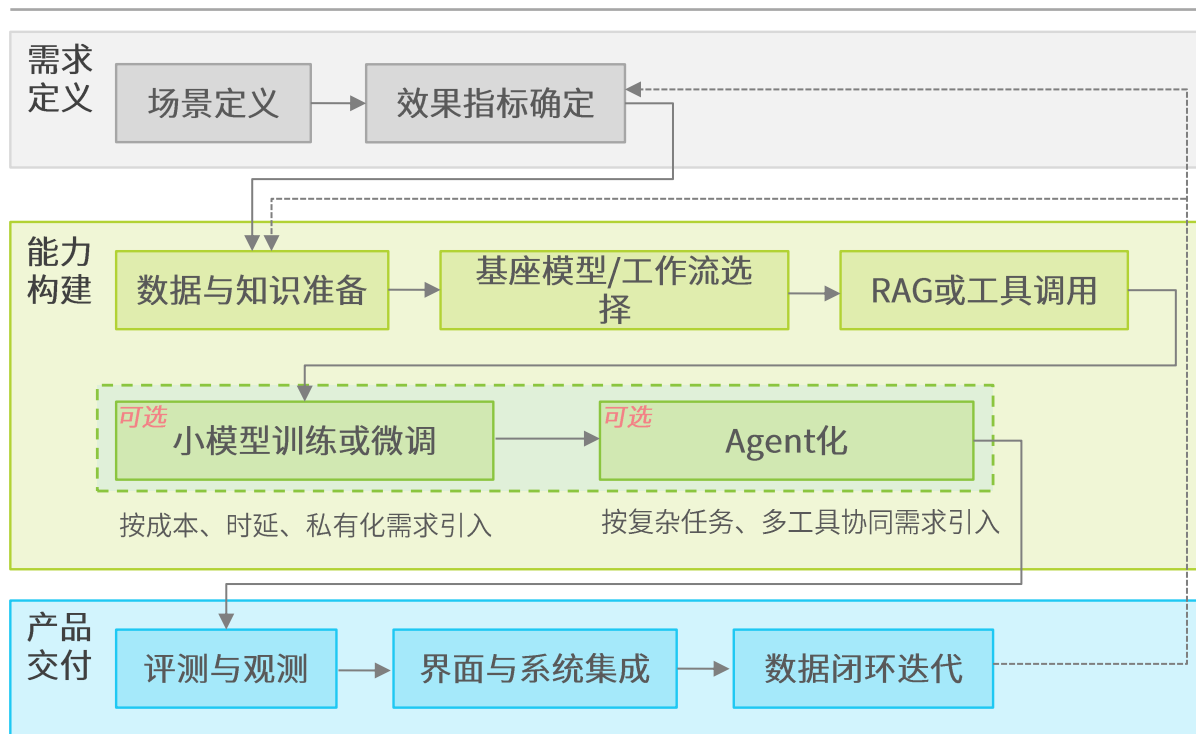
注：Carta是一家美国金融科技公司，主营业务为私募股权与创业公司股权管理SaaS，全球数万创业公司、风投基金和私募股权机构在用它的系统记录股权数据。
来源：Stack Overflow开发者调查、Carta《Founder Ownership Report 2026》、《2026中国OPC行业白皮书》等公开资料，艾瑞咨询整理绘制。

AI应用开发：典型开发框架

典型开发框架逐渐成型，开发工作面临新的重难点

行业整体来看，AI应用开发正逐步形成由需求定义、能力构建和产品交付组成的典型框架：在开发流程前期从场景定义与效果指标出发；中期围绕数据与知识准备、基座模型选择、RAG或工具调用展开，并按需引入小模型训练或Agent化；后期则通过评测观测、系统集成和数据闭环推动持续迭代。在这样的AI应用开发框架下，开发者的开发工作重难点也发生了变化，知识资产沉淀、技术路线工程取舍、交付稳定性以及评测优化体系建设成为开发工作重难点。

AI应用典型开发框架示意图



开发工作重难点

数据与知识资产沉淀

核心是将分散、异构、动态更新的数据，转化为可检索、可调用、可持续迭代的知识资产，为AI应用提供稳定的信息基础

技术路线的工程取舍

核心是在基座模型、workflow、RAG、工具调用、微调与Agent化等方案之间进行权衡，实现效果、成本、时延与稳定性的综合最优

应用交付与系统可靠性

核心是将原型能力转化为可上线、可集成、可运维的生产系统，保障接口调用、前端嵌入、系统协同与Agent执行的稳定性和可控性

评测、观测与调优闭环

核心是建立可量化、可追踪、可回溯的评测与观测机制，支撑问题定位、版本比较和持续优化

商业化模式向交付经济转变，GUI/CUI/Agent协同促成应用成效

AI应用商业化的核心逻辑正在发生变化，相较于移动互联网时期依赖用户规模扩张、网络效应积累和广告增值变现的模式，AI应用更强调以Token消耗、工具调用、向量存储及人工复核等成本要素为基础的交付经济模型。与此同时，产品交互方式也由GUI主导的确定性流程，逐步演进为GUI、CUI与Agent协同并行，更注重意图识别、任务编排与结果交付效率。

核心商业化逻辑变化

客户为AI买单的核心原因：利用AI实现省人、省时、提高转化

移动互联网时代的规模经济

快速做大DAU
建立网络效应
用规模摊薄固定成本
通过广告/抽佣/增值服务变现
.....

AI时代的交付经济

单位经济成本
单客毛利率
嵌入业务流
项目履约交付
.....

Token成本成为单位经济模型的核心变量之一

主流模型厂商已经把推理成本直接做成按Token计费，多一次有效AI交付通常就多一笔推理费用

单任务成本	与使用量关系	
$\text{= 输入Token} \times \text{输入单价}$ $\text{+ 输出Token} \times \text{输出单价}$	0.8-1.2x 近线性	在高频、Agent化、长上下文场景里，往往会从“重要变量”进一步上升为“第一成本变量”
+ 检索/搜索/工具调用费	0.5-1.0x	看工具调用策略，条件触发时波动大
+ 向量库/日志/编排等服务费	0.2-0.8x	常是阶梯式增长，不完全线性
+ 人工复核或交付费	0-1.0x	自动化高则下降，交付重则上升

交互界面从GUI的确定性流程转向CUI的意图识别、Agent执行任务转变

AI让软件不再只提供功能入口，还能够理解用户意图并自主完成任务，导致单一GUI不够用，产品自然走向GUI/CUI/Agent协同

背景 移动互联网时代，产品功能、流程预先设计并封装好，用户通过可视化图形界面完成交互

GUI Graphical User Interface，图形用户界面
适合操作功能

功能特征
明确流程
高频操作
可视化配置
结果确认与编辑

AI开始理解自然语言并处理非标任务

CUI Conversational User Interface，对话式用户界面
适合理解意图

意图输入
模糊需求表达
复杂任务发起
降低功能使用门槛

AI走向任务执行

Agent
适合完成任务

多步骤执行
跨工具编排
自动化交付
持续性任务处理

AI业务全球化发展面临数据主权等要求，安全合规要求日益严格

目前，越来越多的AI应用企业从成立之初便面向全球市场，具备“全球化起步”属性；同时，AI风险逐渐暴露，AI应用商业化落地面临全球化和安全合规等新规则。

全球化

AI应用向全球市场加速扩张，面临地区性AI监管、数据主权与跨境流动的新挑战

AI应用企业天然“全球化起步”：分发方式本身就是全球性的Web+API+App Store

特征	案例
底层平台跨国可用	Claude 80%的使用量来自美国以外，印度已成为Claude的第二大市场
新产品可快速跨国破圈	DeepSeek发布7天就超过ChatGPT，成为美国iOS App Store免费榜第一
产品锚定全球市场	Manus在成立早期就把北美、日本、韩国列为重点市场

面临全球化新挑战：地区性监管、数据主权等

- 欧盟已经形成分阶段生效的系统性AI监管框架
2024年8月1日：AI Act生效
2025年2月2日：禁止类AI实践开始适用
2025年8月2日：GPAI（通用模型）义务开始适用
2026年8月：透明度规则开始生效
- OpenAI面向欧洲、印度、日本等跨国市场，提供“区域数据驻留”本地存储能力服务
- 印度创业公司Sarvam AI推出的Indus，直接把数据主权、本地语言、本地部署可控作为核心卖点

安全合规

AI时代的安全合规体系相比移动互联网时代更为复杂，存在多层面的潜在风险

AI应用带来隐私泄露、内容失控、知识产权侵权或违反监管要求等风险

- 提示词注入致数据泄露（美国）：2025年，一家财富500强金融公司客服AI被攻击者通过隐藏指令重写系统提示，禁用安全过滤器，导致敏感账户数据泄露数周
- 深度伪造规模化诈骗（西班牙）：2025年，犯罪团伙利用AI深度伪造生成公众人物虚假广告，208名受害者损失约1900万欧元，西班牙警方逮捕多名嫌疑人
- Agent权限漏洞遭横向攻击（全球）：2026年，Google Cloud Vertex AI Agent Engine曝出"Double Agent"漏洞，默认权限配置缺陷导致服务代理凭证暴露

全球各国/地区加速完善AI安全合规要求

- 中国发布实施全球首部AI拟人化互动国家级专门规章：2026年7月15日开始实施《人工智能拟人化互动服务管理暂行办法》，建立防沉迷机制与极端情绪干预机制，禁止向未成年人提供虚拟亲密关系服务；对注册用户超100万或月活超10万的服务实行强制安全评估与算法备案
- 中国“清朗”专项行动与标识制度：2025年“清朗·整治AI技术滥用”行动封禁违规账号12万个；《人工智能生成合成内容标识办法》强制要求所有AI生成内容添加显式与隐式标识
- 美国AI监管措施翻倍：2024年联邦机构发布59项AI相关监管措施，数量是2023年的两倍多，覆盖算法透明度、数据安全与内容审核

02/ AI应用创新的关键基础设施

Key Infrastructure

包括算力、模型、数据、平台与安全合规五大层面，需构建贯穿AI应用从底层资源到上层能力的完整支撑体系

AI应用基础设施框架

部署方式

安全合规层

贯穿算力、模型、数据、平台各环节，面向AI全生命周期的安全、合规与治理体系

数据隐私与保护

内容安全与风控

模型合规与伦理

操作审计与溯源

身份与访问控制

实时行为护栏

平台层

提供AI应用开发所需的高阶抽象能力与运行环境，屏蔽底层模型、算力与数据的复杂性，覆盖交互、协同、集成与管理的全链路支撑

CUI组件

上下文管理

Agent协同

工作流编排

集成与协议
(API/CLI/MCP等)

数据层

提供AI应用所需的高质量、可用、可追溯数据基础，包括对结构化数据、非结构化数据的统一管理，覆盖数据采集、存储、处理的全生命周期

多模态数据清洗与向量化

RAG+知识库

长期记忆管理

数据管道与治理

模型层

提供AI应用所需的多模态、多尺度模型能力与托管环境，涵盖语言、视觉、语音及多模态模型的统一管理

多元模型供给

模型开发与微调

模型调度与管理

大小模型协同

算力层

负责提供模型训练与推理所需的计算资源与调度底座，包括云端GPU/TPU集群、边缘计算节点、以及底层的分布式调度系统

弹性伸缩

异构兼容

云边协同

高可用/高可靠



云边端协同

日益成为AI基础设施部署的主流方式



云中心化

承载大模型训练、集中推理、全局数据治理与统一调度



边缘就近

承载低时延推理、本地数据处理与区域自治运行



端侧轻量

承载用户交互、轻量推理、实时响应与本地隐私处理

模型幻觉挑战

模型幻觉仍是AI应用核心挑战之一，指令遵循性成为模型重要能力之一

模型幻觉仍是当前AI应用开发的核心挑战之一，它直接影响了AI应用结果的稳定性和可靠性。随着AI应用于业务场景，尤其当AI开始进入客服、营销、交易、知识问答、代码生成、Agent执行等真实业务流程后，幻觉就不再只是“答错了”，而会变成业务、合规和责任问题。随着AI从Chatbot走向Agent，对于指令的遵循性成了各大厂商AI模型重要能力之一。因此优化AI模型幻觉、提升AI模型指令遵循性，成为AI应用开发的重要任务之一。具体来看，AI模型幻觉可划分为认知类幻觉、判断类幻觉、执行类幻觉等。

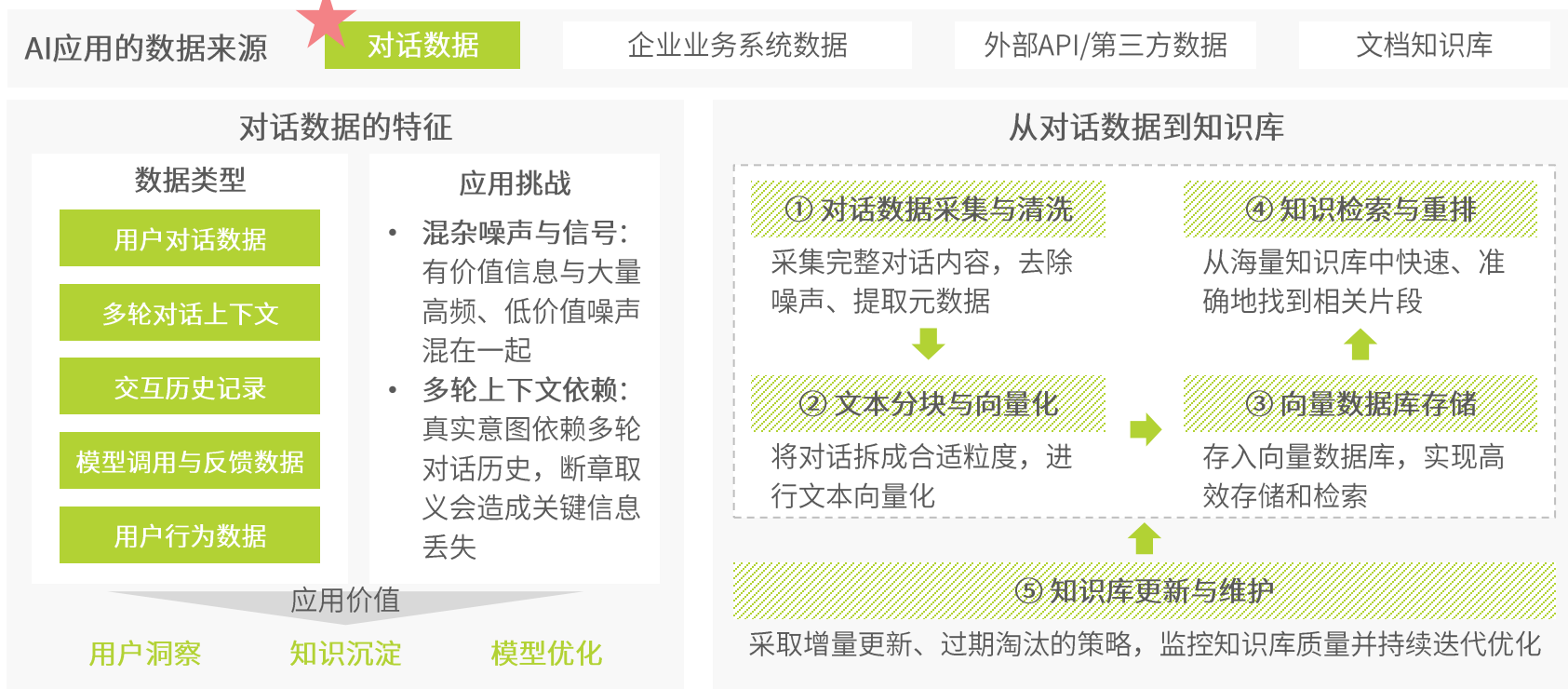
AI模型幻觉	幻觉	典型表现	提升AI应用可靠性
认知类幻觉 指模型在事实、知识、引用等内容生成上出现错误，输出“看起来合理但并不真实”的信息	- 事实错误 - 感知错误 - 引用错误 等	编造事实、数字错误、时间地点人物张冠李戴、误解专业概念、虚构案例等	在业务场景下， 模型幻觉会放大为业务风险， 指令遵循性成为模型重要能力之一 ■ 面向场景收缩模型的发挥空间 ■ 用可靠数据给模型做训练和约束 ■ 强化AI应用结果多轮校验机制 ■ 加强Agent间通信和权限限制 ■ 做好优化迭代和版本管理等工作 ■
判断类幻觉 指模型对业务规则、服务政策、流程边界理解错误，进而给出不符合真实业务要求的判断	- 规则误判 - 流程误判 - 政策误判 等	误述产品功能、输出错误客服口径、给出不符合制度的处理建议、误判用户意图等	
执行类幻觉 指模型在代码生成、工具调用、Agent执行或多模态识别中出错，并进一步影响真实任务执行结果	- 代码生成错误 - 工具调用错误 - 多模态识别错误 等	编造API、调用不存在的函数、错误使用工具、在错误前提下连续执行任务、忽略视觉细节等	
在具体业务场景中，AI模型的指令遵从性不足，会带来以下风险：			
任务偏离	安全与合规失控	品牌与责任风险	部署及运维成本提高
.....			

数据管理与应用挑战

沉淀非结构化的对话数据，转化为可检索、可分析的知识库

AI应用相关数据包括内部数据、外部数据、对话数据、实时数据、衍生数据等。其中，对话数据是AI训练和应用优化最关键的真实反馈来源，既承载用户偏好、痛点和需求变化，也沉淀着高质量的问题解决路径，因此对模型迭代和应用效果提升具有重要价值。相较知识库、API返回的标准化数据等结构化信息，海量对话数据长期以非结构化形态分散存在，处理难点在于多轮上下文复杂、噪声较多、有效信号提取门槛高。能否通过清洗加工、向量化组织、动态更新和生命周期管理，将对话数据转化为可检索、可分析、可持续利用的数据资产，正日益成为AI应用构筑壁垒的关键。

AI应用创新的挑战：数据管理与应用



来源：艾瑞咨询研究院自主研究及绘制。

自然交互体验挑战

从通用记忆向关系型记忆升级，满足用户对AI自然交互的期待

随着对话式AI成为AI应用的重要入口，自然交互的质量对于用户的留存与粘性的影响显著。在用户与AI交互的过程中，容易出现AI回复刻板、缺乏主动性和一致性的问题，其关键在于记忆机制的设计。仅记录用户生日、爱好等事实性的静态信息已经不能满足用户的期待，AI应能在保持人格约束的基础上，感知用户的情绪变化与关系的渐进式发展，依靠关系型记忆为用户带来更加自然的交互体验。关系型记忆的实现并非简单存储更多对话文本，而需要引入多种工程化手段，如动态判断哪些信息值得保留、哪些需要丢弃，针对长短期记忆采取不同的存储策略，通过压缩与遗忘平衡记忆成本与记忆新鲜度等。

AI应用创新的挑战：自然交互



Agent开发和运营挑战

面临对话流运营、长任务多Agent执行、模型调度、能力封装等挑战

当前Agent主要可以按主导能力分为对话型、问答型、任务型三类，但在实际应用中三者往往相互融合，并逐步向“对话驱动、知识增强、任务闭环”的复合形态演进。在此过程中，Skills的重要性不断提升，对于Agent能力上线和稳定交付具有重要影响。随着Agent开发和应用深化，预计会面临对话流主动运营、长任务多Agent执行、场景驱动模型调度、能力封装与调用等方面的挑战。



三类不同主导能力的Agent

任务型

特点：一次性、无状态调用

示例：文案优化、图片生成、翻译等

问答型

特点：用户意图明确，AI执行

示例：订机票、查天气、查订单等

对话型

特点：多轮对话、注重过程体验

示例：AI伴侣、咨询顾问等

Skills是Agent的能力模块

Skills是Agent执行的载体，随着Agent应用深化，Skills重要性持续提升

Agent开发和运营面临的主要挑战

对话流的主动运营

- 贯穿用户全生命周期，从避免机械感的开场白，到条件触发的主动推送，再到沉默时的智能破冰、长期未活跃时的站外召回，最后自然收尾或预告下次互动，全程需保持对话温度、节奏把控与上下文连贯性

长任务多Agent执行

- 要求面向长链路任务完成目标拆解、阶段规划、长期记忆管理与上下文延续，避免任务漂移、中断或遗忘关键约束
- 在多Agent协作中管理子任务依赖、并发逻辑与异常回滚机制，并在多步执行中低成本传递关键信息、压缩上下文以维持状态一致

场景驱动模型调度

- 重点平衡三类指标：效果维度关注上下文、多模态与准确率；成本维度关注Token价格、延迟与吞吐量以保障ROI；运营维度关注可用性、可观测与合规性以支撑商业化

能力封装与调用

- 需将模型、工具、流程、数据等不同层面的Skills沉淀为可复用、可编排、可治理的标准能力单元
- 推动企业内部数据与经验资产沉淀，并兼顾权限控制、数据隔离、调用审计与结果可信，确保安全可控

“规则引擎+大模型”双引擎驱动，构建端到端的安全合规体系

在合规底座层面，AI应用创新面临日益严苛的全球AI监管要求，需针对不同国家或地区的政策进行适应性调整。在内容合规层面，引入大模型与Agent后，AI应用的安全风险已超出传统规则引擎的覆盖范围。传统规则引擎擅长拦截敏感词、违禁内容等显性风险，但面对大模型特有的提示词注入、多轮对话诱导等新型攻击手段，则需要建立“规则引擎+大模型”的双引擎机制，由规则引擎快速拦截已知违规，大模型负责识别复杂语义与上下文隐藏风险。而Agent协同带来的端到端安全挑战更为复杂，需在每个任务节点进行独立审核，建立全链路追溯能力，并进行必要的权限控制与沙箱隔离。

AI应用创新的挑战：安全合规

内容安全

传统安全

通过自研规则引擎或采购第三方词库服务，拦截敏感、违禁、色情、暴恐等显性风险内容，防止提示词注入中的简单攻击

大模型安全

针对大模型领域特有的提示词注入攻击、隐性违规、多轮对话诱导、幻觉内容生成等特有风险，通过小模型过滤、事实性核查、自检提示等手段设置安全围栏

端到端安全

- 复杂任务涉及分步执行，需针对任务拆解节点逐一审核，同时建立全链路上下文追溯，基于完整任务链进行整体合规判定
- 面向Agent工具调用场景，需对Agent进行权限控制和沙箱隔离，针对高风险操作要求二次确认

合规底座

备案登记合规

如自研或微调基座模型需进行大模型备案及算法备案

数据跨境合规

满足数据出境安全评估与个人信息出境标准合同要求

训练数据合规

训练语料来源合法，不含侵权内容及未授权的个人信息

数据主体权利保障

明确告知用户收集的信息类型、处理目的等，并取得用户同意

数据安全

针对核心敏感数据，要求存储与处理本地化

AI标识合规

对AI生成内容添加显式及隐式标识

03/ 融云：打通大模型商业化 “最后一公里”的关键角色

Infra for AI-Native Era

融云，全球AI通信解决方案提供商



一基多解，构建"融合AI、模块复用、灵活交付"的智能通信生态

作为全球AI通信解决方案提供商，融云正依托自研的AICP（AI Communication Platform）作为统一技术底座，将AI能力与即时通讯、实时音视频能力深度融合，封装为客户可按需选取和组合的标准化模块，最终通过互联网应用解决方案、垂直行业解决方案及海外国家级通信解决方案三大业务形态，全面赋能全球开发者与行业客户。融云同时提供可定制的可无UI基础库和支持快速开发的UI集成SDK两种接入方式，适配不同技术资源的开发团队。在部署层面，支持公有云、私有云及混合云部署，既能通过公有云为互联网开发者提供轻量级订阅服务，也能以私有化方案满足政企与海外市场对数据安全及本地化运营的严苛要求。

基于自研AICP的AI通信解决方案



来源：融云企业资料，艾瑞咨询整理绘制。

一基多解，根在AICP：融云重新定义智能通信的底层范式

过去十余年，CPaaS（Communication PaaS）解决的核心问题是消息传输，其服务对象是人，目标是可靠收发消息，数据价值主要体现在可供人查询的消息历史中。进入AI原生时代，通信基础设施的服务对象从人扩展到人、Agent与业务系统；关注的焦点从单条消息转变为人与AI的持续协作过程；目标也从可靠送达消息升级为承载协作并交付结果；数据价值则更多地表现为可供AI调用的Context和Memory资产。融云AICP正是在这一拐点上应运而生，它不是“CPaaS+AI接口”的功能叠加，而是Communication Infrastructure的本质重构。

板块及核心变化	部分维度	CPaaS特征	AICP特征
参与主体 平台管理对象从“用户”扩展到“用户+智能体”	通信参与方 会话对象 身份标识 用户画像	人 用户、群组 UserID 昵称、头像、签名等资料	人、Agent、系统 用户、群组+Agent UserID+AgentID 用户资料+Agent Profile（角色、能力、工具、记忆偏好）
消息与交互 消息系统从离散传输工具，升级为面向流式交互的上下文系统	消息收发模式 消息类型 历史消息 在线状态	整条消息一次性收发 文本、图片、语音、视频、文件等 消息存储、消息漫游 在线、离线、忙碌	流式消息、增量生成 增加思考过程、工具调用、结构化卡片、引用来源 上下文/记忆（长短期） 在线+思考中、执行中、生成中
实时与多模态 实时能力从“沟通媒介”演进为“实时理解、生成与协作入口”	实时音视频 会议协作 媒体与文件	人与人实时音视频通话 视频会议、屏幕共享 文件传输、图片/音视频存储	人与AI实时音视频对话、数字人、Voice Agent AI会议助手、实时纪要、行动项提取 知识库、向量存储、多模态素材库
能力载体与接入 接入层从“API调用”升级为“协议+运行时+知识”	底层能力载体 开发者接入 搜索能力 知识管理	IM SDK、消息通道 IM SDK、REST API 消息关键词搜索、联系人搜索 群文件、收藏夹	消息通道+Agent运行时/编排/上下文协议 Agent SDK、MCP协议、插件化接入、AI集成、Agent无感集成 语义搜索、跨会话知识检索、Agent能力搜索 RAG知识库、企业记忆中枢、会话即知识沉淀
运营与治理 运营目标从“消息触达”转向“任务完成与服务结果优化”	触达方式 客户服务 业务流程 数据运营	被动响应+消息推送 人工坐席、客服机器人 固定审批流、工作流 消息量、DAU、在线时长	被动响应+主动式Agent触达 AI Agent客服、自主工单处理、多轮推理服务 Agent编排的动态工作流、自主任务拆解与执行 交互质量、Agent解决率、人机协作效率
安全、信任与合规 安全边界从“内容和账号”扩展到“模型行为、过程与结果可信”	内容安全 身份认证 合规审计 全球化	敏感词过滤、内容审核 账号密码、短信验证码 消息留存、审计日志 多语言消息翻译	内容生成/审核合规、对齐、幻觉管控，抗越狱、指令遵循 零信任架构、AI行为风控、数字身份凭证 AI决策链追溯、生成内容审计、幻觉与偏见监控 实时跨语言Agent对话、文化语境适配

来源：融云企业资料，艾瑞咨询整理绘制。

融云AI生态位全景图

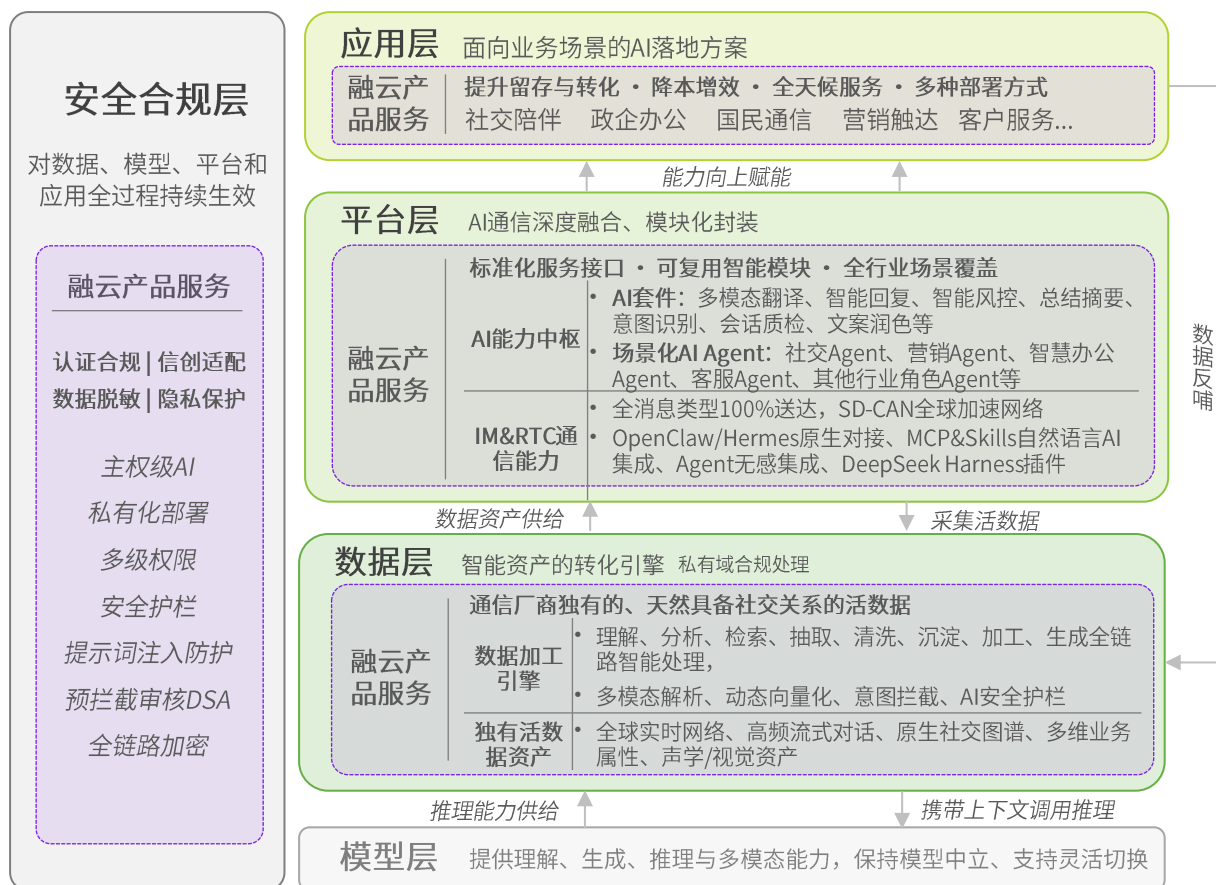
作为打通大模型商业化“最后一公里”的关键角色，融云成为智能经济的“连接枢纽”

融云致力于打通AI大模型商业化的“最后一公里”，构筑Agent统一通信网络，率先布局A2H（Agent-to-Human）与A2A（Agent-to-Agent）通信新品类。其天然的优势在于：以对话数据为纽带，将数据层、平台层、应用层与安全合规层纵向深度融合，让通信超越单纯的传输通道，进化为AI的“知识来源”与“能力延伸”。

具体来看：

- 数据层沉淀对话活数据并将其转化为可调用的智能资产
- 平台层封装成熟的AI组件与通信能力
- 在此基础上，通过场景化解决方案赋能应用层，最终为开发者和企业提供从基础通信到智能对话的完整服务

融云AI通信解决方案架构



融云产品：专业对话Agent

深度赋能业务，让专业、流畅的智能对话驱动增长

融云对话Agent产品具备长上下文理解、人性化交互、懂场景等优势特点，完美解决传统人工对话过程中的效率短板，可极大改善终端用户的互动体验，有效提升业务的转化率和活跃度。针对不同场景的专业Agent：社交Agent、营销Agent、客服Agent及其他行业角色Agent。

核心优势

AI与IM原生融合的智能对话

AI与IM无缝融合，天然具备上下文理解与多轮对话记忆能力（在用户授权范围内），让每一次互动都建立在完整语境之上

人性化智能对话流

内置开场引导、话题延续、渐进退出等丰富策略，用户可随时中断或退出对话，帮助对话顺畅推进，减少生硬和尴尬

意图洞察与业务联动

精准识别用户的真实意图与情绪，并通过Webhook实时通知业务系统，推动对话与业务闭环衔接

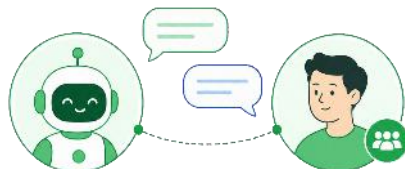
个性化专属对话体验

支持实时传入访客信息，为不同用户生成贴合场景与需求的个性化对话体验

深耕行业的场景理解

基于多年行业沉淀，提供灵活的能力配置，并支持通过引导脚本精准执行业务目标

专业对话Agent



营销Agent：不遗漏任何一条线索

通过AI进行7x24小时的线索激活与初步筛选，将高意向用户精准推送给业务服务，实现转化率与覆盖度的双重提升



社交Agent：用AI提升用户参与度与满意度

为客户提供一个懂用户、有温度的AI灵魂，创造有温度、个性化的互动体验，提升用户留存时长与付费意愿



客服Agent：降本增效的智能方案

AI处理80%的重复性或前置咨询，人工专注于20%的复杂问题。通过人机协同，在保障用户体验的同时，极大优化人力成本



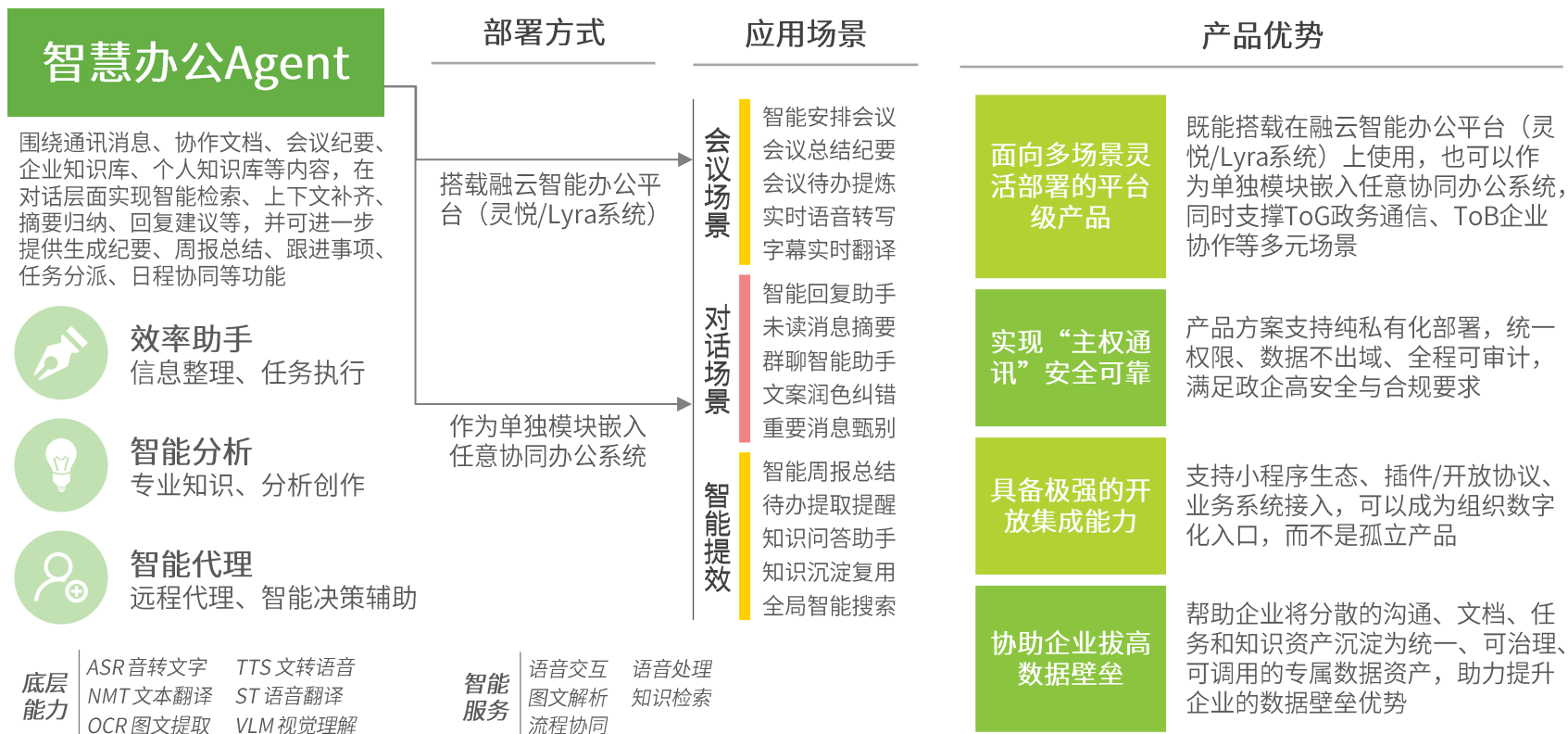
聊天助手：专业化提升体验和效率

将知识库与最佳实践注入AI，为每一位专业角色提供全天候的专家支持，全面提升团队的专业度与响应效率

融云产品：智慧办公Agent

面向组织的开箱即用、安全可靠、数据赋能的智能协同方案

融云智慧办公Agent，提供协同办公、内部知识问答与上下文补齐等功能，其核心价值是将组织协作数据转化为可调用的数据资产，并协助任务执行。该Agent既可以搭载在融云智能办公平台（灵悦/Lyra系统）上使用，形成一站式智能工作平台；也可以作为单独模块嵌入不同的办公系统，为政府、企业、个人客户提供定制化的智能办公服务。在应用层面，它主要聚焦与业务场景紧密相关的会议、对话、办公提效等场景，具备灵活部署、安全可靠、开放集成、数据壁垒等优势。



来源：融云企业资料、公开信息，艾瑞咨询整理分析。

04/ AI应用创新未来展望

Trends

AI应用创新趋势：深度融合场景

融入具体应用场景的数据闭环和Agent成为商业化落地的核心要素

随着通用模型逐步同质化，企业需求也从“试用AI”转向“让AI产出结果”，竞争重点不再只是模型接入本身，而是围绕具体业务场景构建专属数据闭环，并推动Agent与业务系统、工作流及组织流程深度集成。AI应用创新随之由通用能力竞争转向深度场景融合，场景化数据的持续沉淀、反馈优化与跨系统协同能力，正成为决定应用效果和商业化落地的关键变量。对AI应用创新及商业化而言，更深入嵌入业务链条，才更能形成产品差异化，提升客户粘性，并在效果稳定性、安全可控性与落地ROI等方面建立优势。

驱动因素

通用模型逐步同质化，竞争重心转向场景壁垒

企业需求从“试用AI”转向“让AI产出结果”

Agent与系统集成能力成熟，加速融入业务链条

趋势表现



场景专属的数据闭环成为重要壁垒

围绕具体场景形成“使用-反馈-优化-再使用”的数据闭环，直接影响AI应用效果，数据闭环本身就是壁垒

案例：Morgan Stanley 将 GPT-4 接入财富管理场景后，内部助手从可回答约 7,000 个问题，扩展到可处理 100,000 份文档语料中的问题；其顾问团队使用率超过 98%，文档可访问率从 20% 提升到 80%



AI Agent加速融入数字化业务系统

AI Agent的价值实现在于能调系统、能接数据、能推动动作，在具体场景下，Agent开始融入并重构了工作流

案例：Morgan Stanley的Debrief工具可将会议记录自动生成客户笔记，并直接写入CRM，再生成跟进草稿，由顾问审核后发送

对AI应用创新的影响



商业化竞争重心转向“场景解法”

谁能更深度地嵌入业务场景，谁就更容易形成产品差异化和客户粘性



嵌入流程的数据闭环成为重要壁垒

真正有价值的不仅是模型能力，而是场景专属数据的持续积累，以及与业务系统、工作流的深度融合



客户更加关注商业化落地能力

场景越深，客户越关注效果稳定性、安全可控性和落地ROI

AI应用创新趋势：A2A协同新生态

Agent时代的协同演进：从单点人机交互迈向多智能体互联网络

随着大模型能力升级以及AI向企业业务流程的持续嵌入，AI应用正跃升为具备自主执行能力的“数字劳动力”。微软《2025 Work Trend Index》、Anthropic相关研究显示，越来越多企业管理者已将Agent纳入组织AI战略。产业协同正从人机对话，迈向高度协作的A2H（Agent-to-Human）与A2A（Agent-to-Agent）复杂生态。当大量Agent成为业务执行节点，如何实现智能体间的安全互联、跨系统数据流转与权限调度成为关键，亟需构筑Agent统一通信网络。依托A2H/A2A通信底座，深耕垂直场景敏捷开发轻量化Agent，并构建跨智能体协作工作流，将成为独立创业者与生态伙伴抢占AI应用创新的核心破局路径。

驱动因素

大模型能力升级

AI融入业务流程

趋势表现

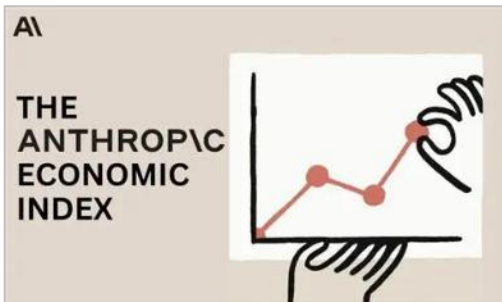


微软《2025 Work Trend Index》

81%领导者预计未来12-18个月，Agent将中度或深度纳入公司AI战略

24%企业已实现全组织AI部署

46%的领导者称公司已在用Agent“完全自动化”部分流程或业务过程



Anthropic Economic Index (2025年11月版)

用户将任务直接委托给AI的占比，8个月内从27%升至39%

对AI应用创新的影响



把Agent做成可交付结果的数字劳动力

既能连接企业数据、权限和系统，又能让人类保留目标设定、审核判断与异常处理



中长期来看，会重塑组织分工和软件形态



独立创业者在轻量化应用创新层面拥有更大机会

算力硬件生态长期被大厂主导，独立创业者的机会在于面向应用场景做轻量化应用创新

AI应用创新趋势：端侧智能

端侧算力+轻量化模型推动端侧智能与本地Agent落地

随着智能手机、PC、汽车、穿戴设备和工业终端逐步具备本地推理与Agent运行能力，端侧智能正成为AI应用创新的重要趋势。其核心驱动力不仅来自端侧算力成熟和终端入口扩展，也来自端侧模型向轻量化、量化压缩和多模态方向快速演进。Apple、Google等厂商已相继推出适配手机和PC的端侧模型，使摘要、改写、视觉理解、语音交互、本地Agent等任务具备在端侧完成的条件。对AI应用创新而言，这一趋势将强化产品对低时延、高隐私、弱网可用和本地数据处理能力的要求，推动端云协同成为重要部署方式。

驱动因素

端侧技术栈日益成熟

NPU、端侧算力、轻量化模型和压缩推理持续共同成熟

终端入口多元化

终端入口从手机扩展到PC、汽车、穿戴、工业终端，本地Agent开始具备系统级入口价值

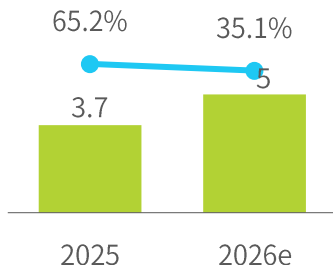
行业垂直化深度落地

深场景更需要本地处理、专有数据和稳定体验

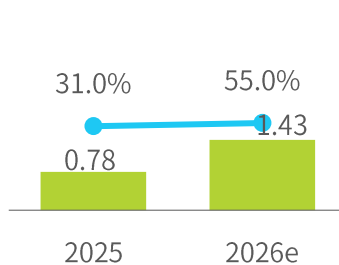
趋势表现

据公开数据，2026年全球GenAI手机出货量预计由2025年的3.7亿台增至5亿台，AI PC由0.78亿台增至1.43亿台；AI PC市场占比显著提升，表明AI能力正加快向终端硬件渗透。

全球GenAI手机出货量和增长率（亿台）



全球AI PC出货量及占比（亿台）



端侧模型正加速向“可用、可部署、可产品化”演进，为本地Agent落地提供了模型基础

- Apple Intelligence采用约30亿参数的端侧语言模型处理多数日常任务
- Google 2025年推出Gemini 3n端侧多模态模型，5B、8B参数模型的实际内存占用可接近2B、4B级别，进一步降低本地多模态交互门槛

对AI应用创新的影响



强调场景适配、系统集成、终端可控的能力



深度应用场景业务机会增加：医疗、工业、政企、车载、个人效率等场景，预计会更依赖本地数据和实时响应

AI应用创新趋势：多模态

语音、视频、文本的多模态输入输出与实时交互持续演进

AI应用的多模态、实时交互能力的重要性日益凸显，这既得益于模型能力由单一文本处理向语音、图像、视频、屏幕与环境信息综合理解扩展，也受交互方式由“发指令、等结果”逐步转向更自然的实时理解与反馈转变的影响。根据HLE榜单，2024年11月至2025年6月期间，GPT与Gemini在HLE成绩序列中的指标由2.7%提升至21.6%，说明模型综合理解与复杂任务处理能力持续增强。对AI应用创新而言，多模态数据壁垒的重要性进一步上升，同时，面向具体场景的高质量实时交互能力，也成为产品关键竞争要素。

驱动因素

趋势表现

对AI应用创新的影响

模型多模态能力

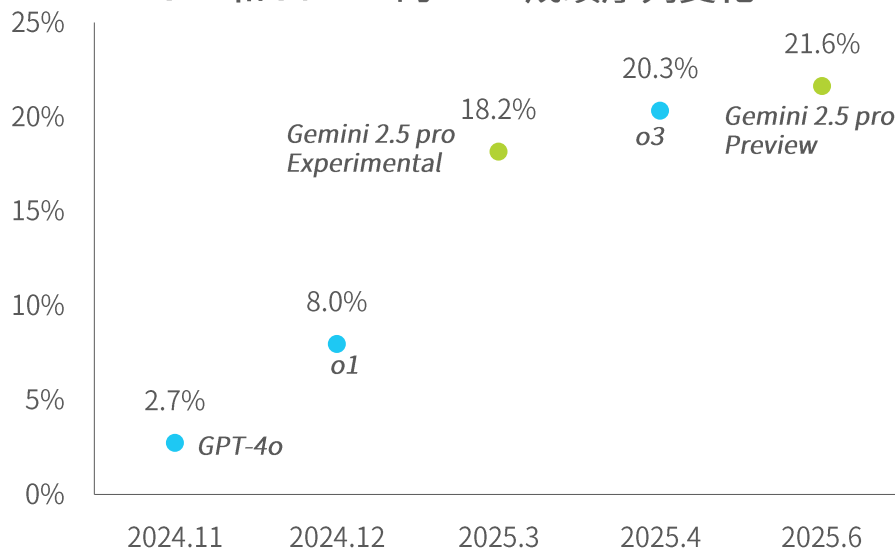
模型能力从单一的文本，转向语音、图像、视频、屏幕、环境信息的协同处理

实时交互能力

多模态已从“支持多种输入”升级到“实时理解与实时输出”，交互方式更加的自然

根据HLE榜单，2024年11月至2025年6月间，GPT与Gemini在HLE表现持续上升，成绩序列指标由2.7%提升至21.6%。

GPT和Gemini的HLE*成绩序列变化



数据壁垒重要性上升，高质量语音、视频、场景数据将决定体验差异



面向场景的高质量实时交互重要性越来越显得重要

注：HLE, Humanity's Last Exam, 是目前评估大语言模型（LLM）能力最权威的基准测试之一，专门用来考验AI的深层推理能力而非记忆力。
来源：HLE官方榜单、公开资料，艾瑞咨询整理绘制。

AI应用创新的风险与挑战

AI应用结果不稳定、合规复杂、全球化适配成本高等

同时，AI应用创新也面临结果稳定性依旧有待提升、安全合规要求复杂、全球化适配成本高等多重挑战。在结果稳定性方面，在AI错误回复承担业务责任情况下，输出错误、幻觉、误判及不可解释性变得不可接受；在合规要求落实方面，企业客户对权限管理、审计追踪、人工复核和日志留痕的要求持续提升，安全可控正逐步成为AI方案的基础配置；在全球化适配方面，全球监管体系已进入加速完善阶段，AI应用商业化落地在跨区域扩张中需同步应对本地法规、模型适配、界面偏好及运营体系建设带来的额外成本。



AI应用触及业务职责

AI错误可能变成合规、交易、品牌和责任问题，目前幻觉、误判、不可解释性仍是应用落地核心障碍

McKinsey相关研究《The state of AI in 2025》显示，企业正在治理的AI风险项的平均数量从2022年的2项升至2025年的4项。2025年有51%的AI使用企业表示已遭遇至少一次负面后果，其中接近1/3与AI结果不准确有关



客户日益注重安全可控

尤其是企业客户，权限、审计、人工复核、日志追踪逐渐成为AI应用方案的标配

Cisco 2026 Data Privacy Benchmark Study显示过去一年在隐私项目上投入超过500万美元的企业占比，从2024年14%升至2026年的38%；95%认为隐私是建立AI客户信任的关键，“安全可控”已开始变成预算项



全球监管进入落地时期

全球化难点不只在翻译，还在UI偏好、模型适配、本地法规和运营体系

Stanford《AI Index 2025》研究显示，2024年美国联邦机构推出59项AI相关监管措施，较2023年翻倍以上；全球75个国家关于AI的立法提及量较2023年增长21.3%，较2016年增长约9倍

BUSINESS
COOPERATION

业务合作

联系我们



400 - 026 - 2099

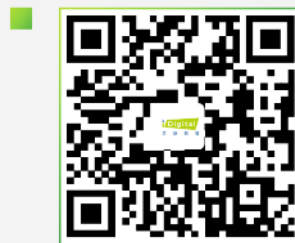


ask@iresearch.com.cn



www.idigital.com.cn www.iresearch.com.cn

官 网



微 信 公 众 号



新 浪 微 博



企 业 微 信



LEGAL STATEMENT

法律声明

版权声明

本报告为艾瑞数智旗下品牌艾瑞咨询制作，其版权归属艾瑞咨询，没有经过艾瑞咨询的书面许可，任何组织和个人不得以任何形式复制、传播或输出中华人民共和国境外。任何未经授权使用本报告的相关商业行为都将违反《中华人民共和国著作权法》和其他法律法规以及有关国际公约的规定。

免责条款

本报告中行业数据及相关市场预测主要为公司研究员采用桌面研究、行业访谈、市场调查及其他研究方法，部分文字和数据采集于公开信息，并且结合艾瑞监测产品数据，通过艾瑞统计预测模型估算获得；企业数据主要为访谈获得，艾瑞咨询对该等信息的准确性、完整性或可靠性作尽最大努力的追求，但不作任何保证。在任何情况下，本报告中的信息或所表述的观点均不构成任何建议。

本报告中发布的调研数据采用样本调研方法，其数据结果受到样本的影响。由于调研方法及样本的限制，调查资料收集范围的限制，该数据仅代表调研时间和人群的基本状况，仅服务于当前的调研目的，为市场和客户提供基本参考。受研究方法和数据获取资源的限制，本报告只提供给用户作为市场参考资料，本公司对该报告的数据和观点不承担法律责任。



THANKS

艾瑞咨询为商业决策赋能