极客邦科技
双数研究院
GEEKBANG D & D RESEARCH INSTITUTE

极客时间 | 企业版

2026 年中国企业 AI 应用场景报告

为企业实现 AI 价值化落地
提供全景式指引

「水木人工智能学堂」

水木AI知识荟 & 交流社群📢

📚每日分享行业报告、行业资讯等！

🔗链接海量AI行业精英！

🎉不定时进行名校名企行活动！

🚀足不出门，尽在水木AI知识荟！

🔥扫码添加小编微信，免费进水木AI交流群

交流社群

去噪星球

去噪星球 每日仅需0.5元

公众号：水木人工智能学堂

目录

CONTENT

01 企业端 AI 应用场景分析

01

多模态模型技术演变：开启原生模态整合

2025 年多模态模型技术迎来突破性发展。长久以来，多模态模型的理解和生成技术发展相对独立，并形成了两种不同的架构探索路径。但2025年，我们看到了以GPT-5、Gemini3、 Bagel、VEO等开始探索统一理解和生成底座的多模态模型的迅速发展。同时我们也观察到了图像、语音、文字模态的技术路线的相对成熟，视频模态模型发展仍以视频理解和视频生成相对独立的技术发展路线为主。

以扩散架构为主的多模态模型路线

- **技术关键**：以扩散机制为基础，通过迭代去噪过程实现高质量图像生成，同时融入多模态上下文（如文本、图像嵌入）实现理解能力。
- **优点**：擅长生成，生成图像质量高、细节丰富；可根据提示词进行多样化创作（风格、编辑等）；训练过程相对稳定。
- **缺点**：推理速度慢、部分模型训练时学习信号稀疏，对不同长度的输出适配不好；部分模型依赖外部框架，开源支持有限。
- **2025年发布的代表模型**：MMaDA（PU&PKU、2025-05）、FUDOKI（HKU&华为，2025-05）、Muddit（PKU&中国典型&NUS&PU、2025-05）

以自回归架构为主的多模态模型路线

- **技术关键**：基于 LLM 的自回归架构，将图像转化为序列Token，通过预测下一个Token的目标统一建模文本与视觉模态。
- **优点**：和大语言模型结构相通，能灵活进行跨模态推理；支持图文交错生成（比如边写文字边插图片）；部分模型用连续令牌，不会丢失图像原始信息。
- **缺点**：模态对齐依赖令牌器质量，设计难度高。
- **2025年发布的代表模型**：TokLIP（腾讯 ARC Lab&中科院自动化所等、2025-05）、Selftok（华为、2025-05）、 UniTok（字节&CUHK、2025-02）、 UniFork（上海AI实验室、2025-06）、 OmniGen2（北京人工智能研究院、2025-06）、 Qwen-Image（阿里、2025-08）、Ming-Omni（蚂蚁、2025-06）SkyworkUniPic（昆仑万维、2025-08）

扩散与自回归混合统一的多模态模型路线

- **技术关键**：融合自回归的序列推理优势与扩散的视觉生成优势，文本令牌自回归生成，图像令牌多步去噪生成，通过双向注意力或共享骨干网络实现跨模态融合，平衡文本语义控制与图像视觉保真度。
- **优点**：兼顾文本语义可控性与图像生成质量；支持复杂任务
- **缺点**：架构复杂，训练与推理成本高；模态融合难度大，易出现文本 - 图像对齐偏差；
- **2025年发布的代表模型**：Mogao（字节、2025-05）、Bagel（字节、2025-05）

趋势预测：原生全模态加速成型，世界模型迎来首轮技术收敛周期

- 我们预测，2026 年，原生多模态能力成为 AI 的标配，原生全模态模型加速落地，多模态理解与生成逐步融合。
- 世界模型技术路线迎来首轮收敛，跨模态统一底座开始形成，为具身智能、自动驾驶等应用的认知、推理与预测提供系统化基础。

超级智能体将成为产业落地与业务重构的真正执行者

当下，多模态大模型作为企业 AI 应用的核心技术底座，在编程、医学诊断、心理咨询等多个领域，已经稳定超过大部分专业人士，智能不再是瓶颈；AI 不再只是被动回答的 Chatbot，而是具备能动性的超级智能体，会自己设定子目标、调用工具、协作完成任务。

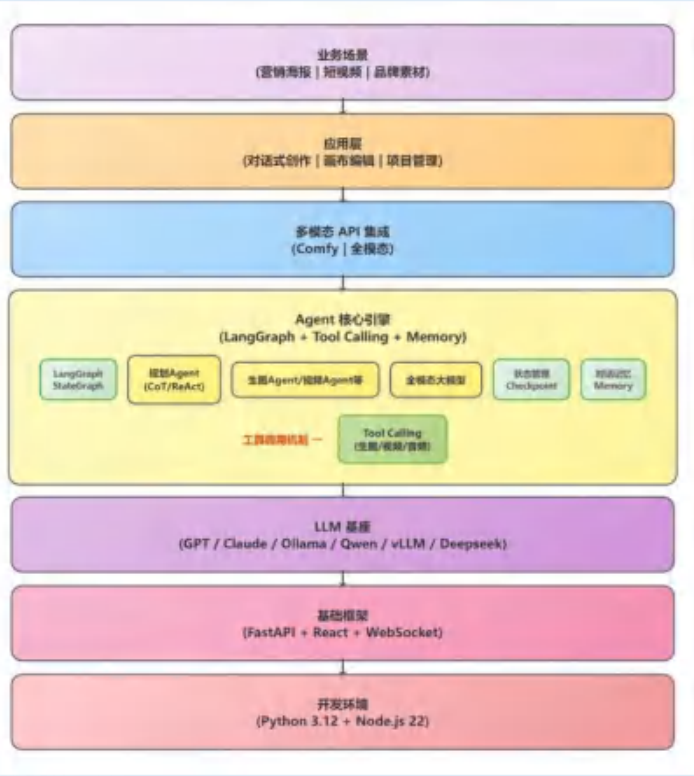
效果涌现 VS 价值涌现

单模型在推理、写作、对话上的能力惊艳：会写代码、会写论文、会诊断、会聊天。
但很多还停留在 Demo、效率工具层面：好看，好玩，但不一定好赚。

当 AI Agent 渗透业务全流程、形成“数据 → 模型 → 决策 → 反馈”的闭环时，开始出现可量化的业务价值：
知识资产变成企业新的“资产负债表科目”
流程智能化成为新的核心竞争力，而不是简单“多一个工具”

多 Agent 协同，让复杂业务可以被系统性拆解，比如金融领域的投研 Agent、风控 Agent、合规 Agent 共同工作，尽调周期缩短 风险评估更可追溯。

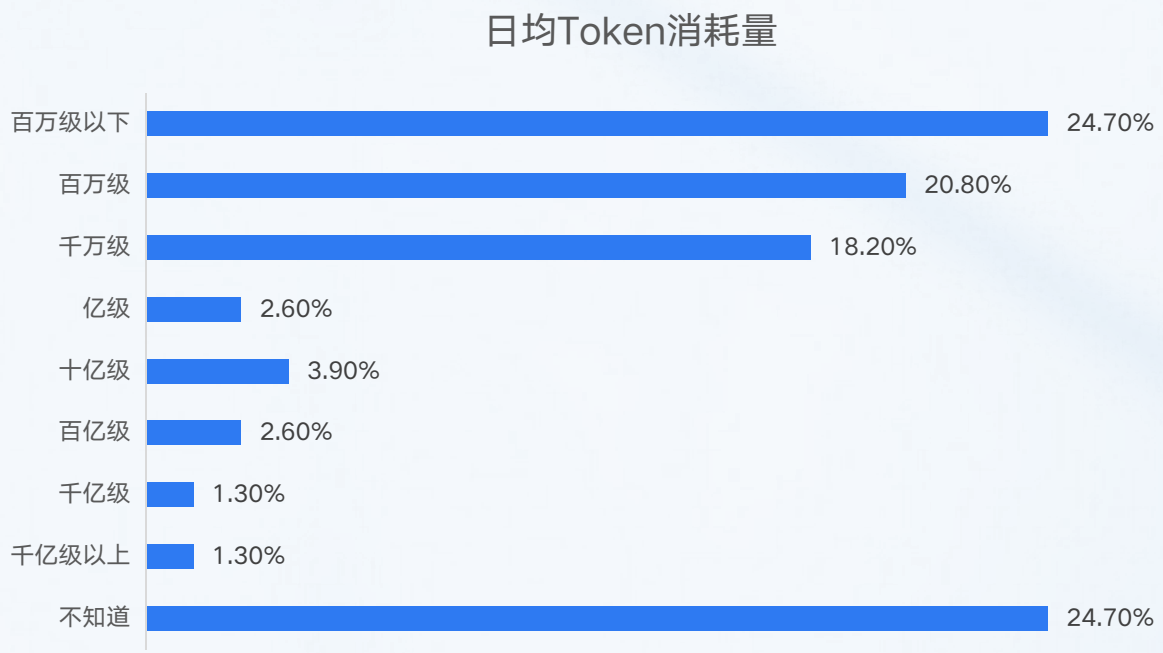
背后依托的是 Agentic Infra / 智能体基础设施：
沙箱 + 资源调度：
上下文与记忆系统全链路可观测

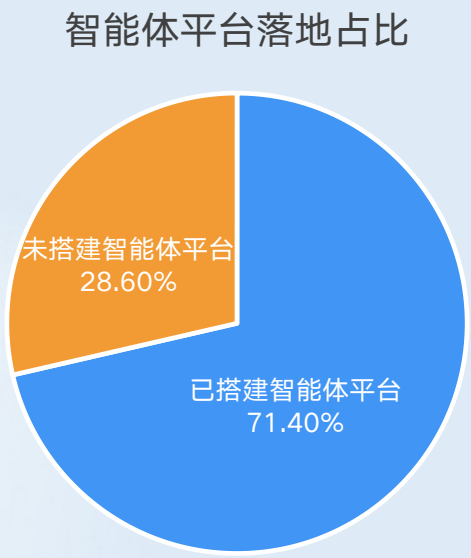


从单一模型能力的“效果涌现”，走向多智能体系统的“价值涌现”，超级智能体将成为产业落地与业务重构的真正“执行层”。

当前通用聊天、通用创作类 AI 仍是企业端 AI 应用场景的主流

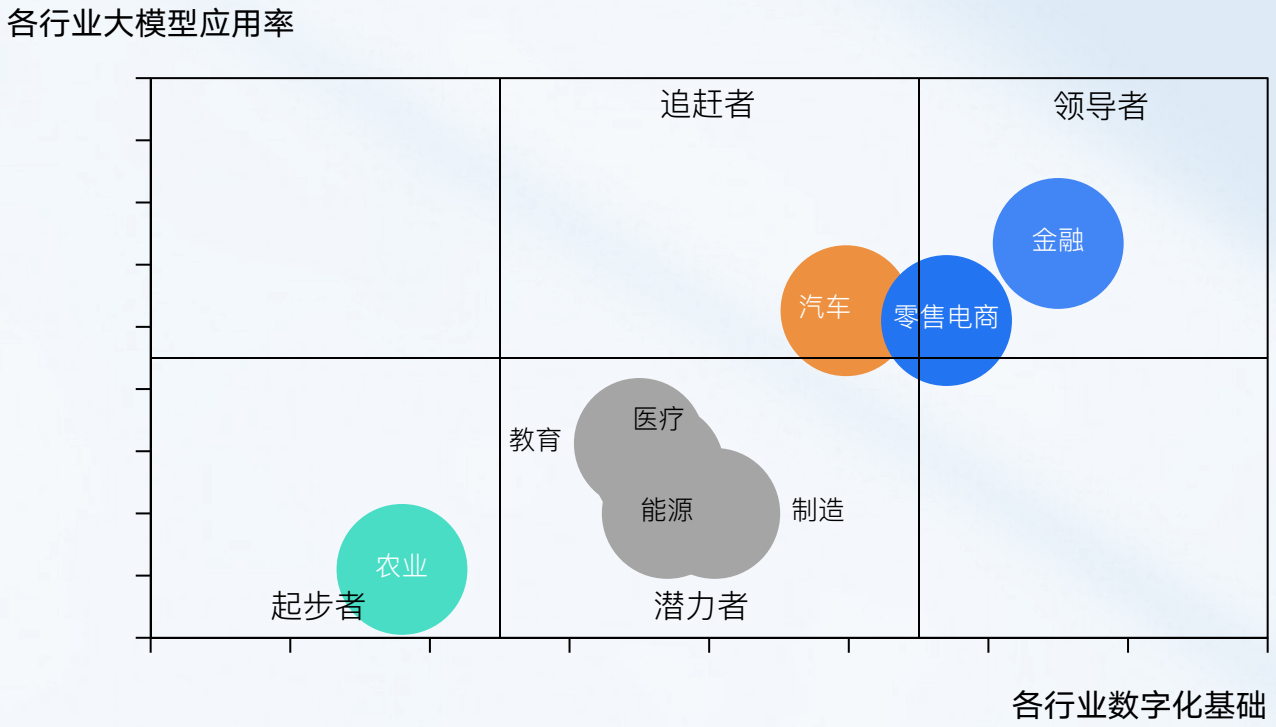
- 根据【2026年中国企业 AI 人才与组织发展报告】的数据显示，75.3%的企业有明确的 Token 消耗量感知，同时有71.4%的企业表示已搭建智能体平台，大模型在产业端的广泛应用由此得以验证。
- 根据我们测算，日均百万级 Token 消耗量，处于大模型应用的“规模化验证期”水平，既非个人试错级（日均万级），也未达企业生产级（日均亿级），最典型、最匹配的应用场景是企业内部协作与办公助手。例如，部门级 AI 助手、内部知识库问答、会议纪要总结、周报 / 汇报生成、合同初审、内部流程查询，小微企业微信客服、品牌私域社群答疑、商品咨询回复。
- 按照日均 Token 消耗量在百万级以下的情况测算，可以推论*通用聊天、通用创作类 AI 仍是企业端 AI 应用场景的主流*。





行业数字化基础同大模型应用率基本呈现正相关关系

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。



数字化基础同大模型应用率基本呈现正相关关系

- 数字化基础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

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

成功案例覆盖的主要行业场景

通过 AI 应用优秀案例评选等活动渠道，我们征集到近千份企业AI应用案例样本，主要覆盖以下行业场景：

金融

- 债券交易智能体
- 金融商保新业智能拓展系统
- 信托智能风控专家系统
- 金融客户协同管理与债务处置系统
- 智能审计模型助手

能源

- 新能源风电智能管控系统
- 燃气智能客服系统
- 燃气管网数据治理与 AI 智算大脑
- 燃气管道缺陷检测系统

智慧医疗

- 语音电子病历、病历自动生成
- 医院智能化服务
- 数字医生体系
- 病历质控与医保控费
- 三医协同智能化
- AI 医疗质控体系
- 智能理赔与医疗审核

制造与出行

- 制造领域 AI 智能体
- 出行领域 GUI 智能体

零售

- 熬胶业务多模态 ChatBI 系统
- 智慧营销矩阵
- 智能价格巡航引擎
- AI 问数智合平台（经营分析）
- 门店商品规划管理系统
- AI 零售现金舞弊审计系统

制药

- 医药行业质量 AI 助手
- 阿胶数据深度挖掘与 AI 大模型
- AI 赋能中成药医学策略制定系统
- AI 问诊与营销驾驶舱
- 医药商业合同质检 AI+
- 合同管理智能 AI

建筑地产

- 智能语音工牌（销售）
- AI 客流与商业空间安全管理系统
- 建筑设备设施运行管理智能体

办公协同

- 智慧员工（办公 AI 助手）
- 智能问数系统
- 数字归因分析引擎（数据分析）
- 多维 AI 财务智能体（财务报销）
- 采购文件合规审查智能体

AI 落地成功的五大主要场景

经过对所有案例样本的归类分析，我们发现 AI 落地成功率较高的五大核心业务场景类型，均具备“痛点刚需、数据可及、价值可量化、落地门槛低”的共性特征。

五大场景

效率提升型

- 核心特征：业务流程中存在大量标准化重复工作，跨系统数据不通导致效率低下，且数据格式相对规整。
- 成功关键：无需复杂技术创新，通过数据打通+轻量化工具即可落地，业务团队接受度高，落地周期短。

风险管控型

- 核心特征：业务中存在明确的风险损失，风险行为有清晰的数据特征可捕捉，需实时或准实时响应。
- 成功关键：风险损失与AI防控效果直接挂钩，投入产出比清晰；风险特征相对固定，模型训练难度低、迭代成本小。

精准决策型

- 核心特征：长期依赖人工经验决策，已积累大量结构化数据，需通过算法优化决策精度。
- 成功关键：数据积累充足，决策效果可通过业务指标验证；算法可基于现有数据快速训练，落地周期短且易迭代。

全链路协同型

- 核心特征：业务流程成熟且标准化，AI可嵌入全链路实现端到端协同，而非单点赋能。
- 成功关键：流程标准化降低AI适配难度，模块化设计便于分阶段落地；全链路协同能放大AI价值，避免单点高效、整体低效。

合规保障型

- 核心特征：处于强监管领域，需满足数据隐私保护或业务合规要求，且合规成本高、风险大。
- 成功关键：采用成熟的隐私计算技术，无需重构现有数据体系；合规效果可量化，符合监管要求且不影响业务效率。

02 重点行业分析

02

金融行业 AI 应用场景分析

- 凭借高数字化基础，金融行业的AI应用场景更为丰富，与核心业务的融合深度也更为紧密。同时，受强监管特性约束，风险管控成为行业推进 AI 落地的核心驱动力。
- 从落地进程来看，智能客服、资讯整理、营销内容生成等轻量化场景，因开发成本低、见效周期短，成为金融机构的首批发力方向。随着外挂知识库的接入与优化，大模型逐步与核心工作流深度融合，推动金融AI从试点验证迈入生产级应用阶段，进而在风控、欺诈检测、审计、债券交易等关键业务环节实现规模化落地，为业务提质增效提供核心支撑。

效率提升型

智能问数系统：某金融企业构建的业务分析智能体，通过打通金融、业务、财务多维度指标与报表，支持自然语言问数、归因分析、报告生成等功能，可用于金融条线经营监控、客户分析、资产质量查询等自助式数据分析场景。

风险管控型

信托智能风控专家系统：某信托企业构建的全域 AI 风险超级智能体，通过融合大模型与传统 ML 模型，在反洗钱场景调用孤立森林和图神经网络，使风险响应速度预估提升 25% 以上。

精准决策型

金融商保新业智能拓展系统：某金融企业融合内外部数据构建的客户拓展智能体，可通过 AI 多模态**解析**与关联规则算法挖掘销售线索，实现精准营销与线索自动化。已关联 3 万现有客户，发现 4 万潜在客户，预计新增 8,000 名客户及千万元保费规模。

全链路协同型

智能债券交易系统：某金融企业搭建的债券交易智能体，已覆盖 500 + 交易群，聊天解析准确率可达99%以上，使交易链路从 4－6 小时压缩至分钟级，人效提升 10 倍以上，客户年化超额收益预计提升 1.5%－2.5%（预计 4.5 亿）。

合规保障型

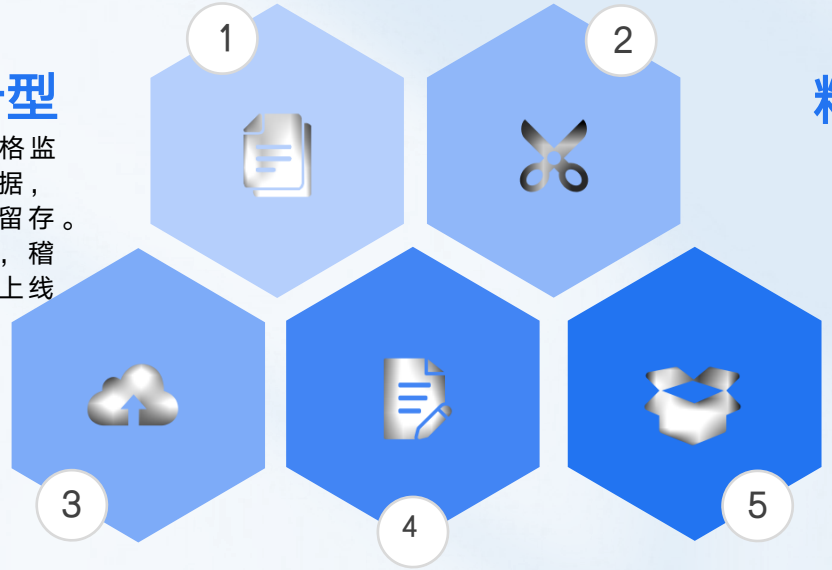
金融客户协同管理与债务创新处置系统：某金融企业搭建的风控超级智能体，采用联邦学习 + 图神经网络（GNN），在数据不出域前提下实现跨机构联合建模，解决金融数据孤岛问题；构建联合清收标签系统，输出还款能力、共债风险标签，并通过 A/B 测试验证有效联系率与催收效率提升。

金融行业AI应用标杆案例

零售行业 AI 应用场景分析

- 由于零售领域兼具消费场景密集、SKU 周转频繁的行业特性，降本增效成为企业核心诉求；同时，不同业务场景的数据可及性差异也决定了终端服务与基础运营场景优先突破，商品与渠道管控则呈现快速渗透的态势。
- 从落地进程看，客流分析、导购赋能、价格监控等轻量化场景因见效快、落地门槛低率先普及；随着多模态数据与大模型能力深度融合，AI 全面渗透选品、运营、风控、营销全链路，在门店精细化管理、全渠道合规、消费者精准运营上形成规模化价值，推动零售从经验驱动走向数据与智能双驱动。

零售行业AI应用标杆案例



效率提升型

智能价格巡航引擎： 某零售企业搭建的价格监控智能体，通过 RPA 自动化抓取价格数据，实现零售渠道价格监控、窜货预警、证据留存。已覆盖四大电商平台与上百家线下经销商，稽查效率大幅提升，违规行为下降，维护线上线下一体化零售价格体系与渠道秩序。

精准决策型

商品规划管理系统： 某零售连锁企业搭建的门店运营智能体，通过算法工具实现品类角色划分、店群归类、商品汰换建议与 DPP 成本拆解，优化选品科学性、利润水平与库存结构，推动零售从经验决策转向数据驱动决策，解决高库存、高损耗、毛利下滑等行业痛点。

风险管控型

AI 驱动的零售审计反舞弊系统： 某批发零售企业搭建的审计超级智能体，面向连锁零售门店，利用 YOLO 与 OpenCV 实现纸币识别，比对 POS 交易与视频时间戳，识别现金舞弊行为。单店 200 小时视频可在 8 小时内完成分析，识别置信度 98%，挽回经济损失、降低人工审核成本，实现零售门店风险前置管理。

合规保障型

AI 客流声购物服务器： 某商业地产开发企业搭建的商场运营超级智能体，通过在商场、店铺部署 AI 摄像头，追踪顾客动线、停留时长、进店行为，构建零售业务漏斗模型，数据采用匿名 ID，符合隐私合规要求。该智能体可用于支持招商定价、品牌调整、营销活动评估、零售额预估、店铺经营优化等场景，同时实现跌倒监测、垃圾滞留等安全管理需求。

全链路协同型

基于 5Ga 的商业空间客流分析： 某商业地产开发企业搭建的商场运营超级智能体，融合运营商定位与 AI 图像识别，实现商场三维客流分析，可用于高峰应急调度、精准营销、新品牌选址优化等场景，从而提升进店率、销售额与商场运营效率，适用于大型商业零售体精细化运营。

能源行业 AI 应用场景分析

- 受限于行业特质及发展现状，能源领域 AI 应用场景主要指向降本、增效、及控险三大核心目标。
- 从案例样本来看，设备巡检、客服响应、能耗监测等场景，因贴合一线刚需、开发周期短、成效可快速量化，成为能源企业的主要发力方向。随着多模态数据融合、数字孪生与隐私计算技术的优化，大模型逐步与能源生产、管网调度、应急处置等核心工作流深度融合，进而为能源行业降本增效、安全运营、绿色低碳提供核心支撑。

效率提升型

燃气领域智能客服：某燃气企业搭建的客服智能体，包含智能导航、伴理同行、远程助手、智能交互、AR 复盘五大功能，可使接通率从 8.7% 提升至 93%，年节约人工成本近 2000 万。

精准决策型

新能源风电智能管控系统：某新能源企业搭建的设备运维智能体，通过构建设备、运维、环境三个知识子图谱，使运维效率提升31.2%，预计年节省 8205 万元，增发电收益 8.35 亿元，减碳 202.31 万吨。

风险管控型

燃气管道缺陷检测系统：某燃气企业搭建的管道维护超级智能体，利用无人机与巡检机器人搭载摄像头、雷达、UWB 等设备采集图像 + 定位数据，通过自主研发的轻量化深度学习模型进行语义分割与缺陷识别，实现腐蚀、断裂等管道损伤高精度检测，识别准确率达 90.8%；月均节省成本约 26,000 元，巡线距离增加 433 公里。

全链路协同型

城市燃气管网数据治理：某燃气企业搭建的运维超级智能体，基于门站、调压站及供气末端设备互联与业务系统数据互通，融合可视化治理、数字孪生与人工智能算法，通过汇聚监控、GIS、客户及远传系统数据，完成 4,000 余公里静态管线与 25 万用户动态数据治理。

能源行业AI应用标杆案例

制造行业 AI 应用场景分析

- 受限于制造行业的数字化程度，制造领域 AI 应用主要集中于生产辅助与工艺监控场景，质量管控与供应链协同类应用正加速渗透，核心制造与全流程智造场景在部分领域（例如汽车行业）逐步规模化推广。
- 从落地进程来看，设备状态监测、生产过程可视化、工艺参数记录、合规巡检等轻量化场景，因改造难度低、业务价值直观、数据易获取，成为制造企业的优先落地方向。随着多模态感知、数字孪生、知识图谱与行业大模型的深度融合，AI 逐步与研发设计、生产制造、质量检测、供应链调度等核心生产流深度绑定，进而在配方优化、智能质检、柔性排产、缺陷检测、全链路溯源等关键制造环节形成规模化效益，为企业实现降本提质、精益生产与智能制造转型提供核心支撑。

制造行业AI应用标杆案例

AI 应用场景的未来趋势预测

- *从通用聊天、通用创作类 AI，到深度绑定核心业务*

企业端 AI 应用正从通用聊天、通用创作类的工具化探索，加速迈向深度绑定各业务板块核心场景的价值化落地。不再局限于会议纪要总结、文案生成等轻量化辅助，而是精准切入生产调度、运营提效、研发创新、合规审查、风险防控等关键环节，直面企业经营中的真实难题。通过多模态数据融合、多智能体协同、知识图谱赋能等技术，AI 深度嵌入业务全流程，成为破解效率瓶颈、降低运营成本、防控业务风险、驱动创新增长的核心支撑。

- *新的 ROI 评价维度浮现*

企业端 AI 应用的价值衡量正浮现新的 ROI 评价维度，不再局限于传统的人力成本节约、工时缩短等直接财务指标。新维度更聚焦 AI 对业务全链路的深度赋能价值，比如知识资产沉淀、技术架构复用带来的研发效率增益，也涵盖员工幸福度指数提升、组织能力升级等产生的间接价值。这种多维度的 ROI 评价体系，更全面地捕捉了 AI 从工具化应用到业务化嵌入的价值增量，让企业能更精准地评估 AI 投入的长期回报，推动 AI 应用从短期试点向规模化落地持续进阶。

- *服务对象以普通业务人员为主，核心目标是“降门槛”而非“替代人”*

通过技术赋能降低专业工具与数据资源的使用门槛，AI 正在成为业务人员的协作伙伴而非竞争对手。它将复杂的算法逻辑、跨系统的数据整合、专业的规则校验等技术环节封装成简单易用的交互界面，让非技术背景的员工无需掌握复杂技能，就能高效完成数据分析、合规审查、客户服务等工作。人机协同的模式既帮助了业务人员从重复性劳动中抽身，使其聚焦于决策、创新等更高价值的工作，也通过降低技术使用门槛，真正实现全员数字化赋能。

- *成果具备知识沉淀与资产化价值，构筑企业“护城河”*

企业端 AI 应用的核心成果，正从单一的效率提升工具，升级为具备知识沉淀与资产化价值的核心数字资产。通过 AI 技术的深度应用，企业在长期经营中积累的专家经验、业务规则、行业洞察、合规标准等隐性知识，被转化为可复用的知识库、可迭代的垂类模型、可推广的SOP模板。这些沉淀的知识资产不仅能跨部门、跨场景复用，降低重复开发成本，更能通过“数据→模型→反馈→优化”的闭环持续进化，成为企业独有的竞争壁垒。这种知识资产化的价值，让 AI 应用突破了单次项目的短期效益，形成可传承、可增值、可规模化的长期核心竞争力。

03 AI 应用的成功范式

03

企业成功落地 AI 的范式

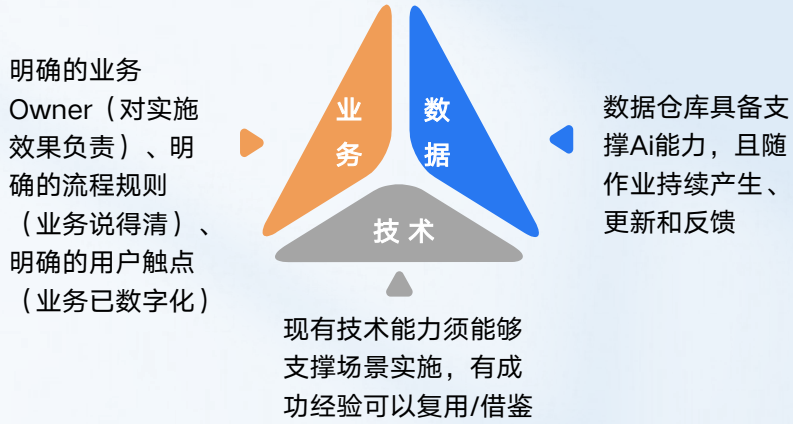
企业落地 AI 不是简单的技术堆叠或者场景复刻，它需要从组织能力、技术底座到行业 Knowhow 的全方位资源支撑。

行业knowhow	*AI场景化落地*	金融风控、债券交易、客户拓展、离散制造/财务领域、人力领域等
AI应用框架	*复合式AI技术组合*	AI Agent、AI应用开发平台、企业知识库RAG、数字员工运营、模型装配
平台底座	*AI应用运行的PaaS*	统一技术栈、流程组织权限、工程化体系、MLOps、大模型
数据工程	*AI的燃料系统*	数据治理体系、数据标注能力、知识图谱、向量化技术等
安全合规	*AI落地的护城河*	访问控制、数据隐私、模型可解释、伦理合规等
组织能力	*技术落地软基建*	AI架构、战略对齐、文化建设、品牌打造、人才梯队

备注：参考资料【InfoQ多智能协同嵌入核心业务流程工程化设计与实践】

成功案例的共性--适配 AI 落地的场景

- *可量化的商业价值*：AI赋能必须直接指向降本、增效、增收、控险四大核心业务目标，项目均有可核算、可对比的量化指标，无明确价值的场景坚决不落地。
- *场景成熟度匹配*：在业务、数据、技术三个维度上是否具备足够的成熟度，从而保证AI技术在该场景的顺利落地。
- *可持续性运营*：须有明确的运营目标、可衡量的运营数据指标，以及支撑持续运营的组织、机制和资源。



成功案例的共性--以智能体为核心的技术框架

- *以智能体为核心组织形式，支持协同与调度*。几乎所有中大型项目都采用多智能体架构，而非单一功能模块，实现复杂任务拆解、分工与协同。
- *所有案例均采用四层智能体认知闭环*：1）感知：理解文本、代码、业务数据、用户意图；2）规划：基于规则、知识图谱或模型进行任务分解与推理；3）执行：调用工具、API、工作流或模型完成操作；4）反思：沉淀经验、优化模型、持续学习。
- *所有案例均深度融合领域知识以抑制大模型幻觉*，手段包括：构建统一语义图谱；基于本体论建模业务对象-关系-行为；训练垂类大模型。

感知

理解文本、代码、业务数据、用户意图

规划

基于规则、知识图谱或模型进行任务分解与推理

执行

调用工具、API、工作流或模型完成操作

反思

沉淀经验、优化模型、持续学习

成功案例的共性--工程化体系持续降本

- *需优先解决 AI 幻觉与可靠性问题。*例如在金融风控、医疗诊断等对准确性要求极高的领域，通过 “大模型 + 传统算法 + 规则引擎” 的混合架构，结合 RAG 检索增强、知识图谱校验等技术，从数据输入、模型推理到结果输出全链路建立校验机制；同时，通过模块化设计与持续的反馈闭环，不断修正模型输出偏差，避免因幻觉导致的决策失误，最终实现 AI 应用在生产环境中的稳定运行，真正为业务降本增效、防控风险提供可靠支撑。
- *搭建可复用中间层与平台化架构，降低重复开发成本。*

中间层与平台化架构

构建领域适配中间层、统一指标平台、标准 SOP 流程库、知识模板，提升跨产品、跨场景模型复用率

模块化设计

采用模块化设计，支持快速迭代与品类、区域拓展

技术底座

形成统一技术底座，大模型、隐私计算、数据治理工具、智能体框架可横向复制

Skills：应对 MCP 带来的上下文爆炸，降低 Token 消耗

Skills 是 Anthropic 在 2025 年 10 月 16日 正式推出的模块化能力系统，通过将特定领域的专业知识、工作流程和最佳实践封装成可复用的指令集，解决大模型应用的效率与稳定性问题。Skills 被认为是对传统工作流与提示工程的迭代优化。

◆ 对传统工作流的核心优化

降低 Token 消耗，缓解上下文瓶颈：有效解决大模型上下文窗口有限的问题，减少冗余 Token 占用，提升模型运行效率，降低使用成本。

提升任务执行稳定性：封装人类专业经验形成标准化流程，强化大模型的指令跟随能力，让模型输出更可控、结果更确定，避免随机化、错误化输出。

兼顾灵活性与泛化能力：区别于硬编码的固定逻辑工作流，Skills 采用文字描述实现流程定义，适配更多场景，泛化能力显著优于传统固定代码流程。

◆ 构建知识工程新范式

降低应用搭建门槛

为知识结构化提供直观载体，非算法、非技术团队也能快速构建 AI 应用，打破技术壁垒。

减少智能体无效开销

将高频、确定性的基础原子能力封装后直接调用，省去 Agent 对简单任务的拆解、推理成本，提升整体响应速度。

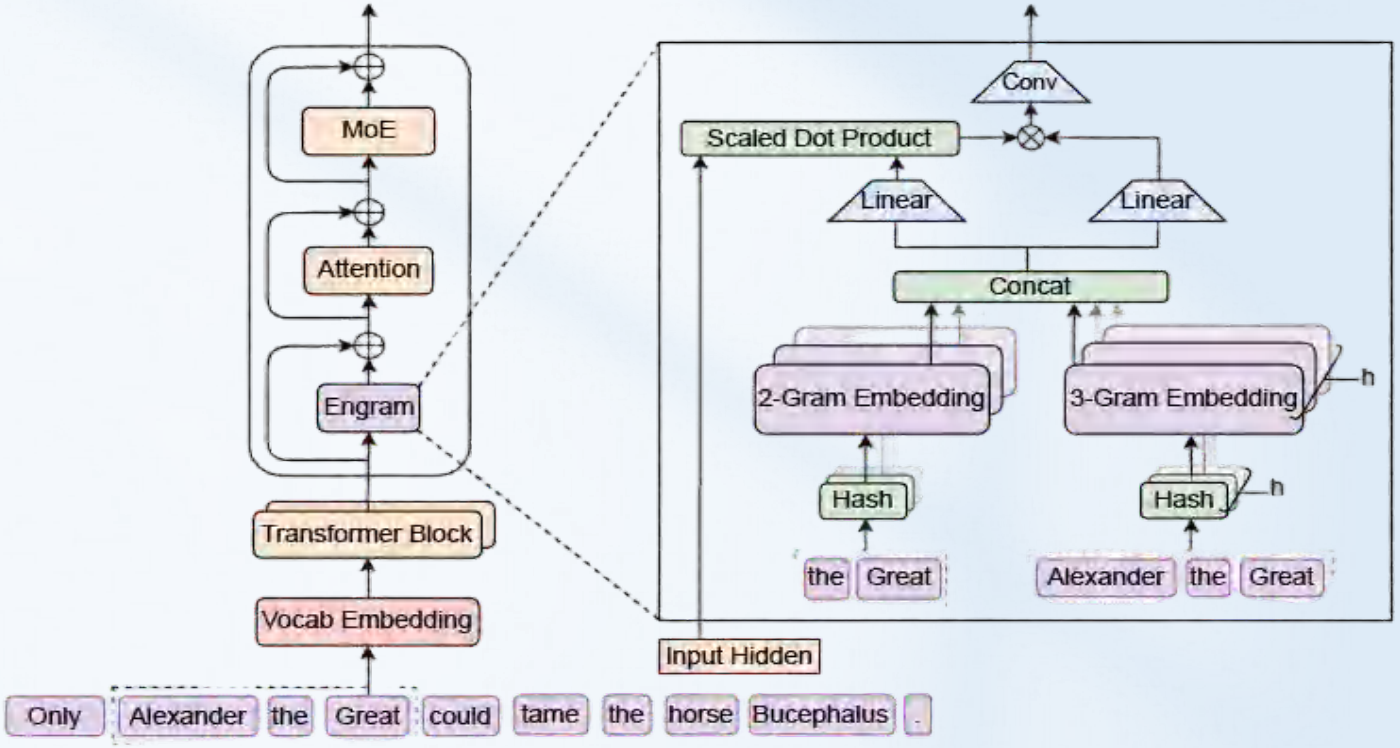
与 MCP 形成互补生态

Skills 负责成熟、标准化的流程任务，MCP处理长尾、复杂、需要深度推理的任务，二者协同覆盖全场景需求。

DeepSeek：实现知识与推理解耦，大幅降低推理成本

DeepSeek从模型架构层面探索重构思考和记忆的方式，打破传统大模型知识存储和推理计算绑定的模式，实现知识与推理的解耦。

- 大幅降低推理成本与延迟：通过将部分知识迁移至 Engram 记忆模块，主干模型可以缩小体积，同时维持原有性能水平。轻量化的主干模型配合Engram 知识表，推理速度更快、算力消耗更低，为企业低成本、高精度部署大模型提供可行方案。
- 实现企业知识内嵌融合：区别于当前外挂式 RAG 知识库，Engram 可实现企业专属知识的内挂式整合，提升模型响应的一致性与专业性。



OpenClaw：本地优先、低门槛部署、可自主执行的数字员工

- OpenClaw（Clawdbot、Moltbot）是 2026 年初由 Peter Steinberger 主导推出的开源、本地优先 AI 智能体网关框架，核心定位为具备自主执行能力的个人数字助理，通过“本地运行 + 多渠道接入 + 工具执行 + 持久记忆”的一体化架构，解决了当前绝大多数 Agent 框架普遍存在的竞态bug、上下文溢出、执行混乱等痛点。
- 其核心执行范式为事件触发--任务规划--工具执行--状态持久化--循环迭代。

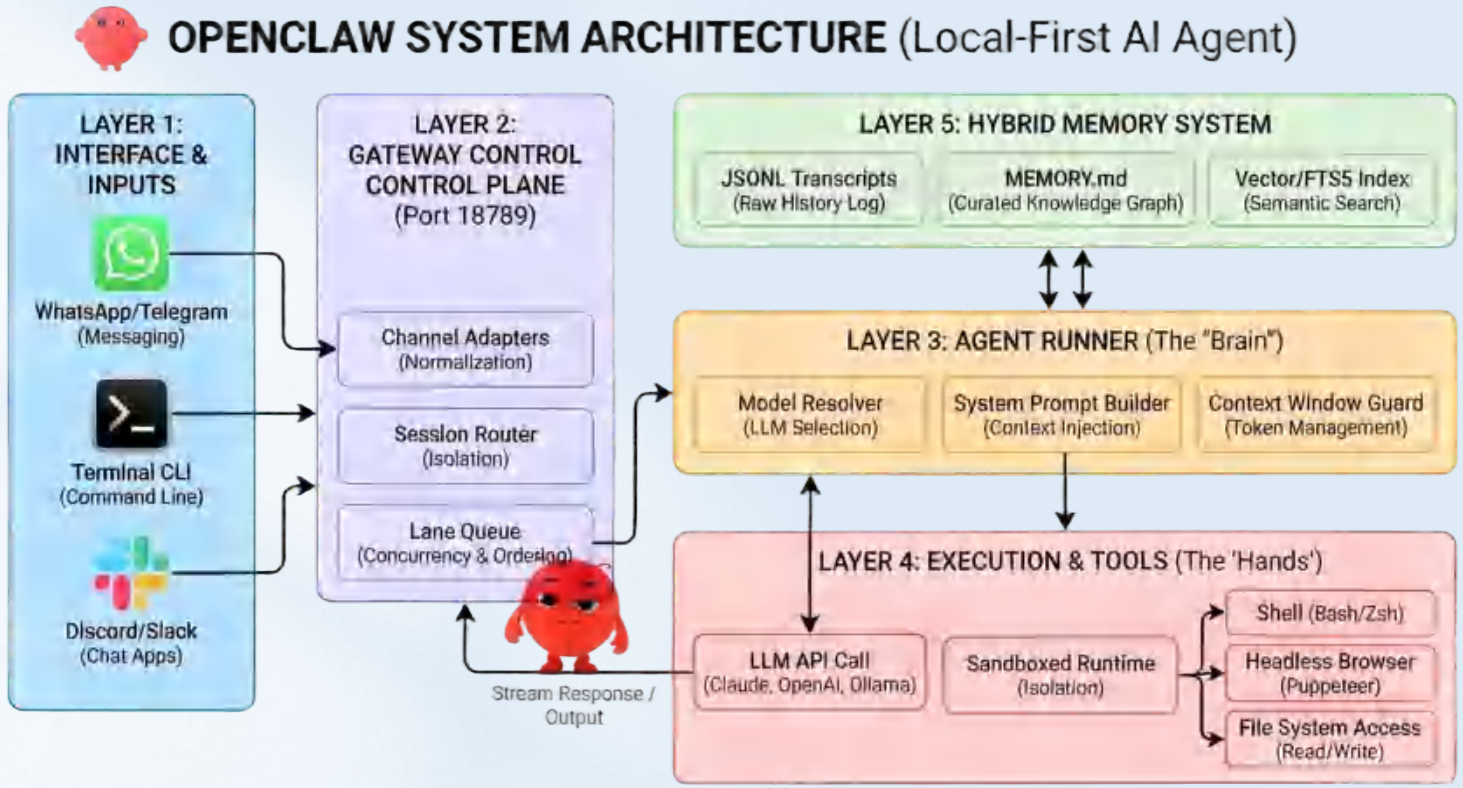
本地优先：突破了传统大模型 “云端托管、数据不可控” 的固有范式，构建了全链路本地数据处理架构，从根源上消除了数据泄露与合规风险。

低门槛部署：采用 Node.js 轻量架构，支持在 PC、服务器、边缘设备上快速部署，大幅降低使用门槛。

生态开放：深度集成 MCP 协议，无需定制开发，大幅降低了生态扩展成本。

多渠道接入：将交互入口从专用 App 迁移至用户日常使用的聊天软件，降低了交互门槛，重构了人机交互的入口范式。

持久化记忆：基于本地向量检索与全文索引构建持久化语义记忆库，可存储用户偏好、历史任务与业务规则，支持上下文关联与主动提醒。



工程化未来关键趋势

知识工程推进：采用 Spec 驱动研发模式，以详细规格说明书驱动 AI 生成代码，提升研发精准度与效率。

运维智能体演进：从静态工作流逐步过渡至单 Agent，最终实现 Multi-Agents 协同运维。

微服务 Agent 框架：采用 “自主多 Agent 协作 + 图编排工作流” 混合架构，支持 LLMAgent、Chain、Graph 等执行引擎，内置 RAG、长短记忆、可观测模块。

Agent Sandbox 构建：基于容器技术搭建 Agent Sandbox，保障测试与运行环境的稳定性。

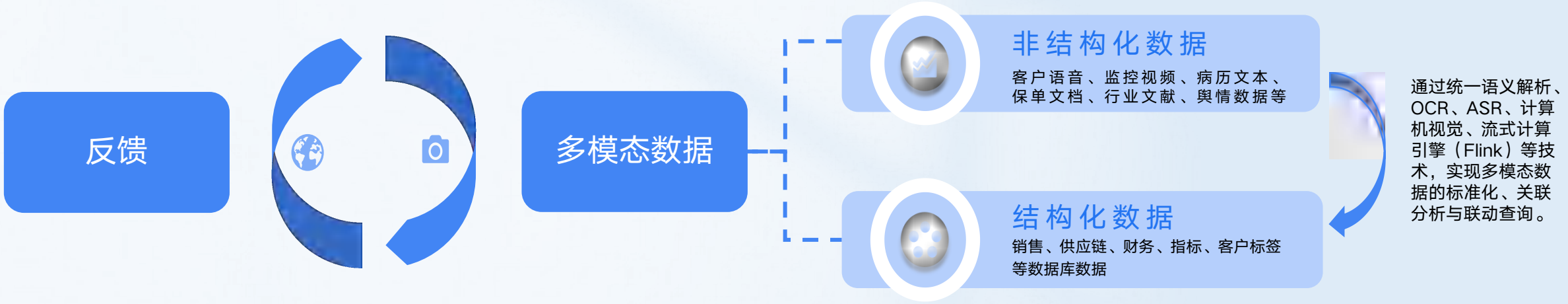
基础模型优化：重点突破多模态大模型打造与优化，同时开展 AI + 全链路压测实践，保障系统稳定性。

成功案例的共性--多模态数据融合

- 多模态数据融合成为标配，打通结构化与非结构化数据，促使 AI 全面感知业务场景，并深度嵌入核心业务。
- 建立数据与反馈闭环，实现持续迭代进化。业务场景中产生的多模态数据持续输入模型，经算法处理后输出决策建议；用户使用反馈、业务结果偏差等信息反向回流，用于优化模型参数、完善知识库与规则库。

趋势预测

- 数据架构升级：构建向量数据湖，通过元数据 + MCP 打造智能数据架构，支持多模态结构化、高质量 Context、多租户隔离、智能冷热分层，满足 AI 对高质量数据的持续需求。
- 工具链层面：整合 OCR、ASR、流式计算引擎等，形成标准化数据处理流水线，实现从原始数据到 AI 可用数据的高效转化。



成功案例的共性--合规安全优先

合规安全优先，主动适配监管。所有涉及核心业务、财务、客户、研发数据的项目，均采用可控大模型 + 本地部署，杜绝敏感数据外流。

基础设施层

强调国产芯片兼容（昇腾、寒武纪等）、混合部署（公有云/私有云/边缘）、智能调度、全链路闭环（从资源接入到计费运营），核心指标为算力利用率、PUE、训练稳定性。

智能体平台层

所有平台均支持国产AI芯片（昇腾、寒武纪、天数智芯、燧原等）及国产操作系统（麒麟、统信、OpenEuler）

通用大模型底座

通义千问、DeepSeek 本地化部署

隐私合规

数据采集采用告知同意、音频加密变音、权限管控等措施；医药、金融类项目严格保证结论可溯源

知识沉淀

构建企业私有知识库、知识图谱、案例库、规则库，数据不出域

InfoQ 技术大会 2026 年规划

极客时间 | 企业版

极客邦技术会议是旗下 InfoQ 中国主办的全球性专业技术盛会矩阵，核心定位为 “扎根社区、服务中高端技术人群与科技驱动型企业，促进创新技术的传播、落地与价值转化”。专为解决企业技术落地的核心痛点而生，通过汇聚字节跳动、阿里、Open AI 等海内外头部企业与科研机构的专家，分享前沿技术趋势帮助企业规避重复试错，全方位助力企业提升技术竞争力、实现前沿技术可靠高效落地。

参会咨询　　查看会议

18年
国内最老牌的技术品牌之一

100,000+人
参会者均为中、高级开发者，以及技术管理人员

2,000+
行业类型覆盖互联网、金融&银行、电商、零售、快消等

智能体时代的 AI 人才粮仓模型

极客时间 | 企业版

《智能体时代的 AI 人才粮仓模型 —— 2026 中国企业 AI 人才与组织发展报告》正式发布

如果你正在思考

- 如何培养既懂业务又能与智能体协同的超级员工?
- 如何利用 AI 完成“十五五”规划下企业的全方位升级?
- 如何应对 AI 时代的组织变革?

来自先行者的洞见

或将带来平稳迈入智能体时代的启发

扫码下载电子书

关于技术应用、人才发展、组织升级的行动指南

如果你正在思考

- 如何让 AI 从“部门项目”升级为“企业战略”？
- 如何培养既懂业务又善用 AI 的“新质人才”？
- 如何选择高价值场景，让 AI 产出可衡量、可持续的业务价值？

这份来自先行者的经验，
或许能为你提供 AI 驱动增长的借鉴

扫码下载

《企业 AI 落地实践案例集》