腾讯云 | CAICT 中国信通院

# 智能体驱动的大模型系统工程与产业实践

## 电信运营商与云服务商的合作探索


腾讯云计算（北京）有限责任公司
中国信息通信研究院计算与大数据研究所

2025年9月

# 编委会

**主编**

张晋、栗蔚

**编委（排名不分先后）**

秦若毅、周锐、王何兵、崔永璇、郭枭

马飞、苏越、赵伟博、桑柳

**参编单位**

腾讯云计算（北京）有限责任公司

中国信息通信研究院云计算与大数据研究所

# 版权声明

# 目录 / CONTENTS

# 前言 / FOREWORD

过去的一年，AI大模型持续作为产业各界的研究热点，“百模大战”退潮，资源愈发向头部企业集中。技术上，多模态逐渐成为模型“标配”，图片、语音、视频理解能力快速提升，并不断在性能增强与成本优化上取得突破；生态上，开源与闭源竞争激烈，开源大模型能力日益媲美商业模型，并在开源社区推广了丰富的增值服务，闭源大模型通过控制核心资源巩固地位，主打订阅服务商业模式；应用上，一方面AI大模型与行业深度融合，AI+X赋能类产品大量涌现，另一方面AI大模型创新应用层出不穷，并从单一大模型应用向多模型协同应用发展，AI智能体成为重要研究方向。

AI智能体的出现，改变了原有的以模型训练与推理为核心的AI大模型系统工程框架，自主智能正逐步成为框架中枢，电信运营商相继将AI智能体的研发纳入AI发展布局。本报告从电信运营商构建以智能体为核心的AI大模型系统工程角度出发，剖析电信运营商面临的建设痛点，结合云服务商相关能力与落地案例给出解决方案，并提出了一套智能体驱动的大模型系统工程成熟度评价体系，希望能为电信运营商的建设规划、成果检验提供参考。

未来，腾讯云与中国信通院将持续关注AI大模型、AI智能体的发展与应用，不断探索电信运营商与云服务商的合作机遇、合作模式，为产业协同、应用创新提供新视角。本报告内容仍有诸多不足，恳请各界批评指正。

# 01

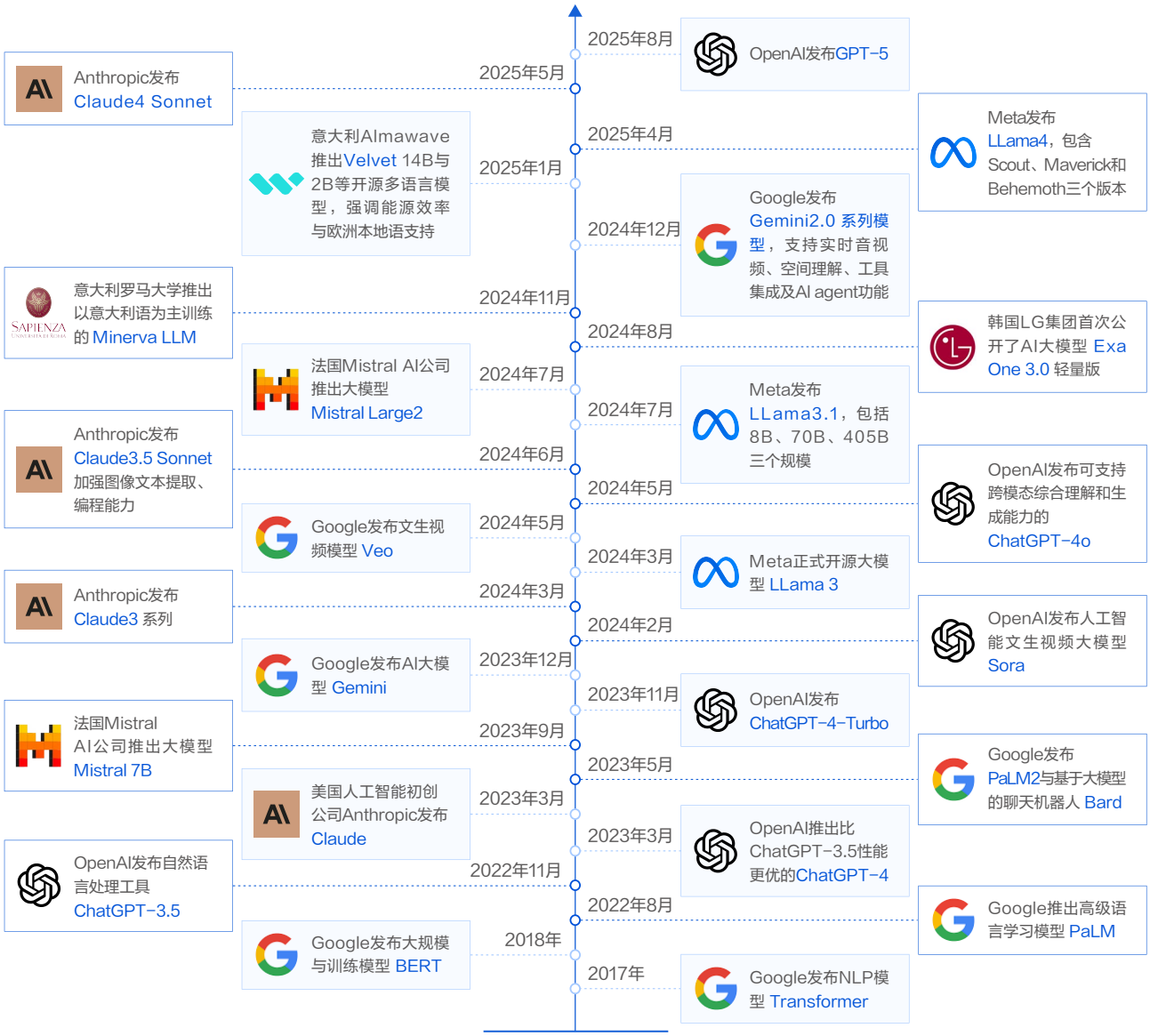
# 智能体驱动的大模型系统工程框架

2024年9月，腾讯云计算（北京）有限责任公司（下称“腾讯云”）与中国信息通信研究院（下称“中国信通院”）云计算与大数据研究所联合发布了《AI大模型应用发展研究报告——电信运营商与云服务商的合作探索》，系统梳理了国际、国内AI大模型的发布时间线及发展特征，对云服务商与电信运营商在此背景下的竞合展开研究，提出了“1+3+N”合作体系，期望为电信运营商与云服务商的合作模式与发展方向提供参考。在过去的一年中，腾讯云与中国信通院持续跟进AI大模型发展动向，梳理其发展趋势与特征，提出以智能体为核心的大模型系统工程框架。

# 1.1 AI 大模型向多模态融合发展，形成全球技术领先者与多元应用生态共生格局

近一年来，海内外不断涌现出新的AI大模型及其版本，但呈现不同的发展特征。

海外，模型体量与效率并重，开放与闭源并行。从发布数量来看，2024年9月至今，AI大模型发布以版本更新为主，新发布大模型数量显著降低，但欧洲个别国家，如意大利等，开始研发自己的大语言模型，强调欧洲本地语言的支持；从模型体量来看，AI大模型在不断突破参数规模上限的同时，也关注到了边缘与终端设备的需求，如Meta的LLaMa 3.1 8B等，已经在探索AI大模型小型化的可能性；从模型性能来看，长上下文能力持续增强，上下文窗口从几十万token扩展到百万级别，更适合长文档分析、代码库理解等场景；从产业生态来看，Meta、Mistral等持续推进开源，OpenAI、Anthropic依旧坚持闭源路线，形成双轨生态。




▲ 海外AI大模型发布情况

国内，技术路线多元，开源热情高涨。从发布数量来看，2024年9月至今，国内各家AI大模型研发方基本保持至少更新一次的频率，持续推动我国AI大模型产业向前发展；从技术演进来看，多模态融合加速，模型普遍支持文本、图像、音频、视频等多模态输入与输出，并在多模态理解和生成上显著进步，推理与工具调用优化能力显著增强，越来越多的AI大模型集成代码执行、信息检索、规划能力，支持复杂任务链路处理，Mixture of Experts（MoE）稀疏架构被广泛采用以提升效率；从产业生态来看，2024年12月DeepSeek V3系列模型的发布与开源引起产业的高度关注，后续腾讯混元、阿里Qwen、百度文心、智谱等都发布了Apache-2.0或MIT的开源版本，共同营造了国内AI大模型良好的开源交流氛围。


▲ 中国AI大模型发布情况

## 1.2 AI 大模型发展从重训练转为重推理、重应用，AI 智能体迎来黄金发展时期

**AI大模型早期研究聚焦模型训练。**

过去几年，AI大模型的发展重心主要集中在参数规模的不断扩展和训练数据的海量积累上，以追求更强的语言理解和生成能力。典型如GPT-3、BERT等模型，通过数十亿到百亿参数及海量预训练数据实现了自然语言处理能力的质的飞跃。然而，通过模型体量增长带来的性能提升局限性逐渐显现，同时带来的训练成本与时间的攀升也决定了能进行“规模竞争”的企业越来越少，单纯注重模型训练已难以满足市场对实时性、精准性和参与度的需求。

**重推理成为AI大模型建设方发展新趋势。**

重推理强调模型在推理链条上的优化，包括增强模型的逻辑推理、多步骤思考、因果分析和工具调用能力。2024年以来，OpenAI发布的GPT-4.1和GPT-4.5，显著提升了推理能力和上下文理解深度，支持高达百万token的长上下文，使模型能处理更复杂的推理任务。例如，GPT-4.5能够在法律咨询、医学诊断等专业领域进行多步骤推断，展现出超越传统问答的智能水平。谷歌的Gemini2.0系列引入了“Thinking”模式，通过链式推理和工具调用进一步提高了解决复杂问题的能力，推动了智能体架构的实现。

**应用方不断创新AI大模型应用新范式。**

伴随AI大模型推理能力的提升，其带来的潜在商业价值日益显著，越来越多的应用方入局探索AI大模型的创新使用场景，促进AI大模型从语言生成工具，转变为多场景、多模态的智能助理。

例如，OpenAI通过插件（Plugins）和API将大模型深度集成到办公套件、搜索引擎、编程辅助工具中，提升了企业和个人的生产力。微软将GPT模型嵌入到Office 365和Azure云服务，实现了智能文档编辑、代码自动生成和数据分析的无缝对接。Meta推出的LLaMa 3.2和轻量级版本，适配移动端和边缘设备，推动了模型在实时交互和低功耗设备上的应用。

AI大模型的发展重心向推理与应用的转变，不仅反映了技术进步的内在需求，也契合了实际应用场景对智能化水平的更高期待，为AI智能体（AI Agent）的出现奠定了基础。

**从技术上看，AI大模型能力与计算基础日益成熟。**

大模型技术突破，2018年以后，Transformer架构、预训练—微调范式（如BERT、GPT系列）不断推动自然语言处理、多模态感知和推理能力提升，AI逐渐具备了更通用的、更强的上下文感知能力，这为智能体提供了“通用大脑”；多模态与工具调用，GPT-4V、Gemini、Claude 3.5等多模态大模型出现，让AI能理解和处理文字、图片、音频、视频等多种信息，结合API调用和插件机制，使得AI不仅具备“说”的能力，还具备了“做”的能力；算力服务与云服务发展，GPU、TPU、智算中心和分布式训练框架（Megatron-LM、DeepSpeed等）的普及，降低了构建智能体的门槛，云平台（AWS、Azure、腾讯云、阿里云、华为云等）提供按需算力和AI服务，使得智能体可以随时部署和扩展。

**从需求上看，用户需要AI从回答问题到完成任务。**

用户需求升级，早期的AI大模型以实现“聊天”或“回答问题”的能力为主，伴随AI大模型能力的升级，用户和企业更希望AI能主动完成任务、管理流程、调用外部系统，这就需要从“对话模型”进化为“任务型智能体”；复杂任务驱动，无论是跨平台的数据抓取、金融风控的实时响应，还是工业生产的设备监测，都需要AI具备长时记忆、环境感知、动态决策与自动执行能力；自动化与效率红利，企业希望通过借助AI智能体减少人力投入，提升业务自动化率，例如客服智能体可以7×24小时处理多语言客户请求，研发智能体可以自动分析代码、生成测试用例并提交等。

**从产业来看，AI应用向智能体发展已具备基本市场条件。**

产业竞争方面，OpenAI发布ChatGPT引发全球AI产业加速竞赛，各大科技公司（Google、Meta、腾讯、微软、阿里、百度、字节）纷纷推出大模型与智能体框架，形成技术迭代的高压态势；开放生态方面，Deep Research、LangChain、AutoGPT、Meta AgentBench等开源项目降低了智能体研发门槛，开发者可以快速搭建多工具、多角色、多步骤协作的智能体系统；政策伦理方面，各国政府逐步制定与完善AI监管、隐私保护和安全标准，鼓励在可控范围内推动AI应用落地，为智能体的普及提供政策边界与安全框架。

相比于单一大模型对话，AI智能体被视为能够自主感知、推理、规划和执行的“全链条AI实体”，是推动AI从“工具”走向“助手”乃至“协作伙伴”的关键一步，正引领AI进入更高阶段的智能化时代。

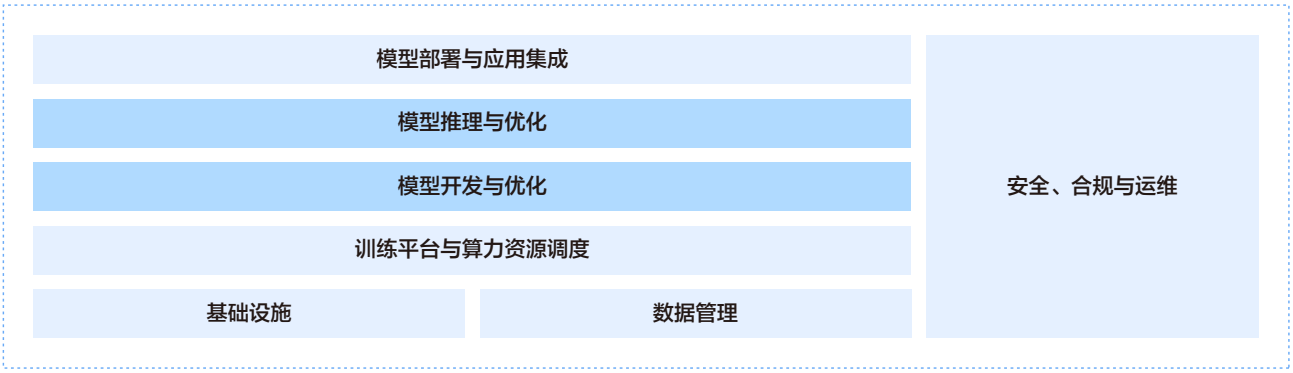
# 1.3 智能体驱动大模型系统工程框架演进，自主智能成为框架中枢

系统工程作为一门跨学科的综合性工程管理与技术方法，致力于通过系统化、结构化和规范化的手段，规划、设计、开发、集成及运维复杂系统，确保各子系统和组件有机协同，达到预期的性能指标和目标。它强调全生命周期管理，涵盖需求分析、系统架构设计、实施执行、测试验证、部署维护等各个环节。系统工程的核心理念是整体最优化，注重跨领域协作与风险控制，尤其适用于解决大型、复杂、动态变化的工程难题。

在人工智能领域，系统工程的概念同样重要。随着人工智能技术的快速发展，特别是大规模预训练模型的兴起，单一模型的训练和推理已无法满足复杂应用场景的需求。AI大模型系统工程应运而生，它不仅涵盖了模型的设计与训练，更涵盖了算力调度、数据管理、模型部署、应用集成、安全保障等多维度的协同管理，形成了一个覆盖端到端流程的完整生态体系。

**AI大模型系统工程以模型训练和推理为核心，注重计算资源的高效利用和模型性能提升。**

其框架主要由以下几个关键部分构成：




▲ AI大模型系统工程框架

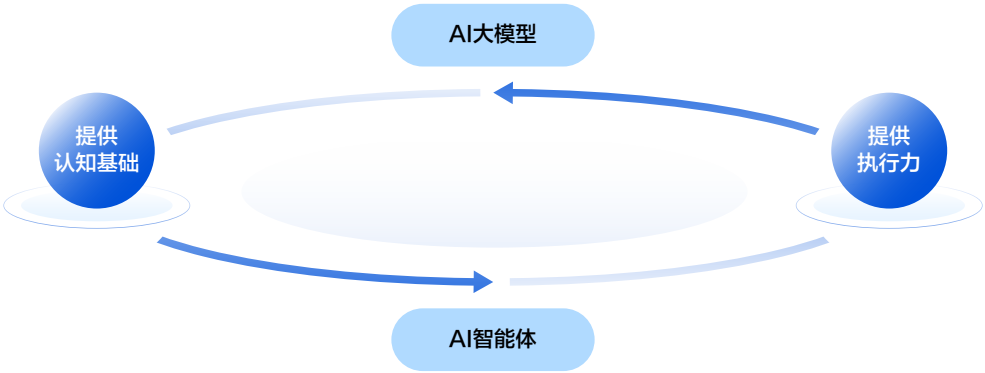
- 基础设施：包括高性能计算集群、分布式存储、高速网络等支撑模型训练与推理的必备基础设施。
- 数据管理：负责采集、清洗、标注和存储大规模、多样化的数据资源，保证数据质量和安全，为模型训练提供可靠基础。
- 训练平台与算力资源调度：包括以算力为主的基础设施资源调度和分布式训练框架，支持模型的大规模并行训练及快速迭代。
- 模型开发与优化：涵盖模型设计、训练调优、量化剪枝、知识蒸馏等技术手段，提升模型性能和推理效率。
- 模型推理与优化：构建稳定、低延迟的推理系统，支持多场景、多模态的在线或离线服务。
- 模型部署与应用集成：将大模型能力嵌入具体业务场景，如智能客服、内容生成、辅助决策等，实现智能化赋能。
- 安全、合规与运维：保障数据隐私、模型安全、系统稳定运行，并根据法规和行业标准进行合规管理。

AI大模型系统工程较少关注模型的主动性和持续智能化能力。然而，随着AI智能体技术的兴起，这一传统框架正在发生深刻变革。AI智能体作为一种具备感知、决策、执行和学习能力的自主系统，不再满足于被动响应用户输入，而是能够主动感知环境、规划多步骤任务、调用外部工具与服务、并根据反馈动态调整行为。这一特性对大模型系统工程框架提出了新的要求：

- 系统架构更加模块化与动态化：智能体驱动的系统不再是简单的模型输入输出流程，而是包含多个协作智能体、多样化工具接口和实时环境感知模块的复杂网络。
- 任务驱动的闭环管理：系统能够持续监控任务进展、自动纠错和优化执行路径，形成端到端的闭环控制，提升执行效率和结果质量。

- 多智能体协同与分工：通过多智能体的协作机制，系统能够处理更为复杂和大规模的任务，支持跨领域、多模态信息融合和决策。
- 实时学习与自适应能力：智能体能够基于新数据和环境反馈不断优化自身模型和策略，实现持续进化。

智能体驱动大模型系统工程的变革不仅体现在技术架构上，更深层次地影响了系统的设计理念和应用模式。具体来说，智能体与大模型技术形成了相辅相成的关系：



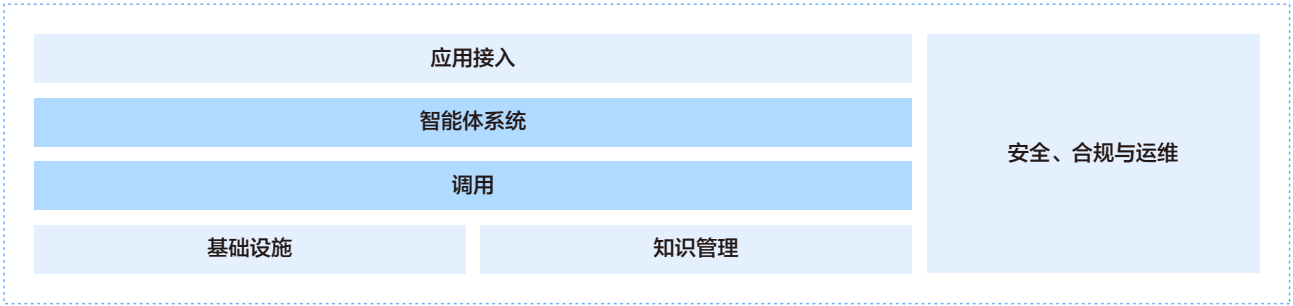

▲ AI大模型与AI智能体关系

- 大模型为智能体提供认知基础：大模型凭借其强大的语言理解和生成能力，成为智能体的“脑力核心”，支撑其复杂推理、多模态交互和知识检索。
- 智能体将大模型能力落地为执行力：智能体将大模型的认知能力转化为实际行动，能够调用工具、访问数据库、执行代码，完成多步骤任务。

这种相互促进的关系催生了以智能体构建为核心的新型AI大模型系统工程框架。

**智能体驱动的大模型系统工程框架以自主智能为中枢，整合大模型、工具链、数据流和应用环境，形成一个高度自治、可扩展、面向多任务的智能系统生态。**

它不仅实现了从“模型驱动”向“智能体驱动”的范式跃迁，更为AI系统的智能化、自主化和协同化发展奠定了坚实基础。其框架主要由以下几个关键部分构成：




▲ 智能体驱动的大模型系统工程框架

- 基础设施：包括高性能计算集群、分布式存储、高速网络以及对应的集群调度能力，以支持AI智能体的基本稳定运行。在此基础上，传统AI大模型系统工程框架中的训练、推理能力整合为推理服务/引擎，与其他算、存、网能力共同作为AI智能体的基础设施。
- 知识管理：涵盖数据系统、向量数据库、记忆存储、知识增强等能力，与基础设施层并列，支撑智能体持续学习与自我进化，促进智能体群体协作与群体智能，支撑智能体演进出区别于AI大模型的自主智能能力。
- 调用：调用层是连接智能体任务规划、调度和底层能力的中间桥梁，支撑智能体实现多模型、多工具、多服务的操作。
- 智能体系统：智能体系统是框架的核心，涵盖上下文管理、多模态处理、自适应决策、多智能体协作等能力，负责任务理解、规划、执行、记忆管理，以及与模型和外部工具的交互。
- 应用接入：应用接入层最靠近用户，将智能体能力封装成可直接使用的应用接口，并提供多种接入方式、交互形式和场景化适配。它既是用户入口，也是业务系统与智能体交互的桥梁。
- 安全层、合规与运维：保障系统合规、可靠、可控运行的关键部分，覆盖从用户访问到模型调用、再到数据与内容输出的全链路安全，防止数据泄露、滥用与攻击，同时满足隐私与法规要求。

AI大模型系统工程框架从以训练、推理为核心向以智能体为核心转变，代表着人工智能发展从“算力+模型”驱动向“任务+智能体”驱动的范式转变，AI智能体正在成为连接模型能力与真实业务价值的核心枢纽。

## 1.4 从单一问答演进为多步骤任务闭环，电信运营商智能体正向行业定制与多智能体协同发展

近1-2年，国内外智能体技术快速发展，已经涌现出了一批典型的产品。按照智能体能力范围，可分为通用智能体、垂直领域智能体、智能体框架以及平台级智能体系统四类。除了云服务商与人工智能企业之外，电信运营商也在AI智能体领域走在前列。

- 通用智能体：通常依托于企业自身的大语言模型，具备较为广泛的任务适应能力，适合个人和企业的多领域应用。例如OpenAI GPTs（Custom GPT），基于GPT-4/4o，可自定义知识库、指令和工具调用；中国移动灵犀智能体2.0，依托MoMA智能决策引擎和主从智能体协同架构，拓展全场景服务能力，尤其注重出行、生活、家庭、办公、通信五大高频场景支持。
- 垂直领域智能体：专注某一行业或场景，深度优化工具链与领域知识。例如腾讯在ToC（面向消费者）方面发布了超过10个智能体，涵盖生活、学习、工作等多个场景；中国联通聚焦家庭场景，基于其AI服务平台，孵化“通通”智能体，通过账号融合、数据融通，使“通通”成为用户的智能数字分身，打通家庭和个人、大屏和小屏的畅享体验。
- 智能体框架：通常不直接提供“成品”智能体，而是支持开发者通过框架构建定制化AI智能体。例如腾讯Cognitive Kernel-Pro智能体框架，采用多模块层次化架构设计，支持使用Python作为行动语言，打破了智能体构建依赖平台预定义能力的限制。
- 平台级智能体系统：具备多个智能体协作能力，形成一个“AI团队”，完成大型复杂任务。例如Hugging Face Agents，可基于Transformers按需组合NLP、CV、语音模型完成任务；中国电信星辰智能体平台以星辰大模型为技术底座，覆盖语义、语音、视觉、多模态四大领域，开放超过300种场景化AI算法能力，集成智能体开发、低门槛模型训推、数据飞轮等功能模块，支持用户进行智能体构建。

总的来说，智能体的历史发展呈现从单一推理向任务闭环与多步骤决策演进、从通用能力向垂直行业深度定制拓展、从独立模型向多智能体协同与生态化发展的特征。

# 02

# 智能体驱动的大模型系统工程成熟度评价体系

根据市场调研与实际落地经验，基于以智能体为核心的AI大模型系统框架，本报告提出了一套智能体驱动的大模型系统工程评价体系，从六个维度评价系统的构建与运维水平，并划分了五个成熟度等级来对评估结果进行描述。

## 2.1 评价维度

智能体驱动的大模型系统工程成熟度评价的六个维度分别为：

- 基础设施与模型能力：反映智能体系统的“核心引擎”能力，包括算力利用能力、模型规模、算法创新、推理效率和自适应学习能力等，是系统价值和竞争力的直接来源。
- 数据与知识管理能力：关注系统背后的“信息底座”，衡量数据管理和知识构建能力，确保智能体有充足、可靠、可更新的数据支持。
- 多样调用能力：关注连接性、调度性、鲁棒性、效率性与自适应性。
- 智能体系统架构与工程化能力：衡量系统设计是否科学合理、模块化、可扩展、可靠及智能体系统的工程化水平和可操作性，确保智能体能够在复杂环境中稳定运行并支持后续升级优化。
- 应用价值与商业化能力：关注系统的实际应用能力、场景适配性、用户体验和商业化价值，体现技术成熟度与业务价值的结合能力。
- 安全合规与运维能力：衡量智能体系统在稳定性、安全性、合规性、伦理性方面的成熟程度，确保系统可持续、可信赖、对社会和用户负责。

## 2.2 评价指标

每个维度选取8个指标评估其成熟度，指标及描述如下表所述。

| 基础设施与模型能力 | 数据与知识管理 | 多样调用能力 | 智能体系统架构与工程化能力 | 应用价值与商业化能力 | 安全合规与运维能力 |
|---|---|---|---|---|---|
| 基础设施建设水平 | 数据来源覆盖与多样性 | 多模型接入能力 | 系统模块化设计 | 业务场景覆盖率 | 模型输出安全性与法律合规性 |
| 模型规模与参数量 | 数据质量管理与清洗流程 | 多工具调用能力 | 可扩展性与可复用性 | 系统对业务流程优化能力 | 数据隐私保护措施 |
| 模型算法创新性 | 知识库/知识图谱构建能力 | 多服务集成能力 | 服务化/微服务化支持 | 用户体验与可用性 | 系统访问控制与身份管理 |
| 多模态能力 | 数据标注与反馈机制 | 任务分解与调度能力 | 数据流与模型调用效率 | 成本效益分析与ROI | 安全漏洞检测与响应机制 |
| 自适应学习与在线学习能力 | 数据安全与隐私保护 | 调用链路管理能力 | 系统容错与可靠性设计 | 多场景集成能力 | 伦理风险评估与管理 |
| 推理速度与资源效率 | 数据生命周期管理 | 错误处理与容错能力 | 多智能体协同能力 | 市场适应性与快速迭代能力 | 模型版本管理与回滚机制 |
| 对复杂任务的解决能力 | 数据可追溯性 | 调用效率与资源优化能力 | DevOps/MLOpS流程成熟度 | 客户反馈与迭代机制 | 异常监控与预警能力 |
| 可解释性与可控性 | 数据智能推荐与增强能力 | 自适应调用优化能力 | 自动化训练与部署能力 | 产品化及可商业化程度 | 灾备与高可用方案 |

▲ 智能体驱动的大模型系统工程成熟度评价指标

| 评价维度 | 指标 | 指标描述 |
| --- | --- | --- |
| 基础设施与模型能力 | 基础设施建设水平 | 算力、存储、网络基础设施水平，可支持的训练、推理规模等。 |
| | 模型规模与参数量 | 模型的大小、参数数量以及计算能力，对处理复杂任务的潜力影响。 |
| | 模型算法创新性 | 模型使用的算法是否具有创新性，是否突破现有技术瓶颈。 |
| | 多模态能力 | 模型同时处理文本、图像、语音等多种数据类型的能力。 |
| | 自适应学习与在线学习能力 | 模型根据新数据实时更新和优化的能力。 |
| | 推理速度与资源效率 | 模型在给定算力下的响应速度和计算资源利用率。 |
| | 对复杂任务的解决能力 | 模型在复杂、跨领域任务中的表现和成功率。 |
| | 可解释性与可控性 | 模型决策过程是否透明，可被人理解和调控。 |
| 数据与知识管理能力 | 数据来源覆盖与多样性 | 数据是否来自多渠道、多类型，满足模型训练需求。 |
| | 数据质量管理与清洗流程 | 数据是否经过规范清洗和质量控制。 |
| | 知识库/知识图谱构建能力 | 系统是否能构建结构化知识库或知识图谱并支持调用。 |
| | 数据标注与反馈机制 | 数据标注质量及用户/系统反馈用于持续改进的能力。 |
| | 数据安全与隐私保护 | 数据存储、处理和使用过程中的安全和隐私保障。 |
| | 数据生命周期管理 | 数据从采集、存储到使用和淘汰的全流程管理能力。 |
| | 数据可追溯性 | 数据来源和变更是否可追踪，支持审计。 |
| | 数据智能推荐与增强能力 | 系统是否能智能补充、增强数据以提升模型表现。 |
| 多样调用能力 | 多模型接入能力 | 支持接入不同类型与规模的AI大模型（通用大模型、垂直领域模型、轻量化模型等），并能在任务中灵活切换和组合调用。 |
| | 多工具调用能力 | 能否顺畅调用外部工具（搜索引擎、计算引擎、数据库、仿真工具等），并进行任务内的工具链组合。 |
| | 多服务集成能力 | 支持调用云服务、企业服务API、第三方平台（如支付、地图、业务系统）等，实现跨服务的统一编排。 |
| | 任务分解与调度能力 | 具备将复杂目标任务分解为子任务，并根据依赖关系和资源情况进行调度的能力。 |
| | 调用链路管理能力 | 具备对多模型、多工具、多服务调用过程的追踪、监控与管理，确保调用链的完整性、可控性和可回溯性。 |
| | 错误处理与容错能力 | 在调用失败、接口异常、延迟超时等情况下，具备自动重试、降级替代或多路径冗余机制。 |
| | 调用效率与资源优化能力 | 能否在调用中实现计算、存储和网络资源的高效利用，优化响应速度和调用成本。 |
| | 自适应调用优化能力 | 能够根据任务上下文、历史调用效果和实时反馈，动态调整调用策略（如选择最优模型、切换工具链路径、平衡调用成本与性能），实现调用智能化、自适应优化。 |

| | | |
|---|---|---|
| **智能体系统架构与工程化能力** | 系统模块化设计 | 系统功能是否划分为独立模块，便于维护和升级。 |
| | 可扩展性与可复用性 | 系统能否方便地扩展功能或复用已有模块。 |
| | 服务化/微服务化支持 | 系统是否采用服务化架构以支持灵活调用和部署。 |
| | 数据流与模型调用效率 | 系统中数据传输和模型调用是否高效。 |
| | 系统容错与可靠性设计 | 系统面对故障或异常的容忍能力及恢复能力。 |
| | 多智能体协同能力 | 系统中多个智能体能否高效协作完成任务。 |
| | DevOps/MLOps流程成熟度 | 系统开发、训练、部署、运维流程是否规范化和自动化。 |
| | 自动化训练与部署能力 | 模型训练和部署是否可以自动化执行。 |
| **应用价值与商业化能力** | 业务场景覆盖率 | 系统能支持多少实际业务场景。 |
| | 系统对业务流程优化能力 | 系统能提升业务效率和流程优化程度。 |
| | 用户体验与可用性 | 系统的易用性、响应速度和用户满意度。 |
| | 成本效益分析与ROI | 系统投入与产出比，商业价值评估。 |
| | 多场景集成能力 | 系统能否在不同场景或系统间灵活集成。 |
| | 市场适应性与快速迭代能力 | 系统能快速适应市场需求并持续迭代优化。 |
| | 客户反馈与迭代机制 | 系统是否有闭环机制收集反馈并改进。 |
| | 产品化及可商业化程度 | 系统技术能否转化为成熟产品并实现商业化。 |
| **安全合规与运维能力** | 模型输出安全性与法律合规性 | 系统输出是否可避免有害、违法或误导信息，系统和模型是否满足相关法律法规要求。 |
| | 数据隐私保护措施 | 是否遵循数据隐私保护法规和安全策略。 |
| | 系统访问控制与身份管理 | 系统访问是否受控，支持身份认证和权限管理。 |
| | 安全漏洞检测与响应机制 | 是否能及时发现并响应安全漏洞。 |
| | 伦理风险评估与管理 | 系统是否有伦理风险识别和缓解措施。 |
| | 模型版本管理与回滚机制 | 是否支持模型版本管理和快速回滚到稳定版本。 |
| | 异常监控与预警能力 | 系统是否具备监控异常和及时报警的机制。 |
| | 灾备与高可用方案 | 系统在灾难或故障时能快速恢复并保持可用。 |

▲ 智能体驱动的大模型系统工程成熟度评价指标及描述

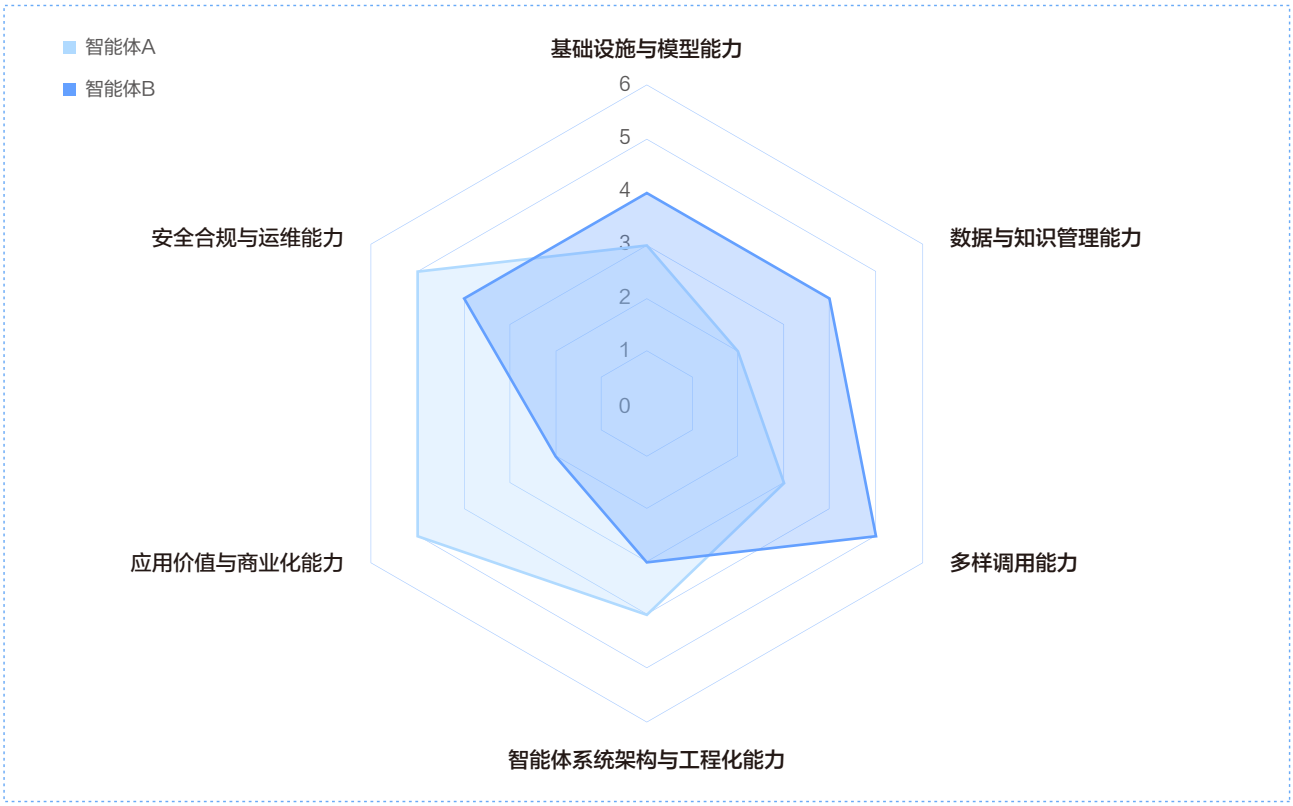
## 2.3 评价结果

在实际使用中，可根据需求或目标为2.2中每个指标设置各级别对应的量化值及分数（例如1–5分），根据其重要性设置权重，通过加权平均计算最终得分，获取评价结果。

成熟度各级别描述如下表所示。

| | |
|---|---|
| Level 1 初始级（Foundational） | 零散搭建，缺乏标准流程，功能靠人工触发 |
| Level 2 可重复级（Repeatable） | 模型与Agent组合形成模块，具备基本流程，人工控制为主 |
| Level 3 已定义级（Defined） | 建立起统一架构体系，有清晰组件定义，任务可分解与复用 |
| Level 4 可量化级（Managed） | 有度量体系，Agent运行有监控、评估与调优能力 |
| Level 5 最优化级（Optimizing） | 系统具备自学习、自动演化、反馈闭环与智能协作能力 |

▲ 成熟度级别描述




▲ 评价结果示意图

# 03

# 智能体驱动的大模型系统工程关键技术和能力

智能体作为具备环境感知、决策制定及动作执行能力的自主算法系统，其发展经历了从基于规则的方法到基于模型的智能体，再到当前基于大语言模型的智能体的演进。传统智能体受限于启发式规则和特定环境约束，难以适应开放和动态场景，而大语言模型凭借其强大的学习和规划能力，显著提升了智能体在复杂任务中的性能。当前，智能体的构建围绕记忆组件、规划组件和执行组件展开，支持多模态信息处理和复杂决策。随着大模型技术的持续优化，智能体将在更多场景中实现落地与创新。

智能体驱动的大模型系统工程关键技术包括智能体开发工具（例如，腾讯云智能开发平台、火山引擎HiAgent、阿里云百炼等）、大模型技术（例如，腾讯混元大模型、阿里通义千问大模型等）、异构算力调度等。核心能力涵盖智能体构建能力、多工具调用能力、生态支持能力以及智能体的灵活部署能力等。这些技术和能力共同推动了智能体在企业和生活场景中的广泛应用与创新。


▲ 智能体大爆发，成为企业创新提效的利器

## 3.1 智能体开发平台能力

智能体开发平台是面向企业客户及合作伙伴的，基于大模型的应用构建平台，结合专属数据，更快更高效地搭建智能体应用，并能通过工作流串联多个智能体，满足更为复杂的应用场景。平台旨在通过知识问答、工作流、Agent等多种形式以及提供基础能力服务，促进高效的大模型应用构建。

### 智能体开发平台具有高时效性

智能体开发平台通过外部知识库连接、知识库增强方案、大模型知识引擎优化、Agent模式灵活性以及知识生产自动化等技术手段，实现了高时效性的智能体开发与应用落地。

外部知识库的快速更新。通过连接外部知识库，智能体能够实时获取最新信息，确保知识库内容的时效性。这种方式避免了传统模型精调中因训练周期长而导致的知识滞后问题。

知识库增强方案。智能体开发平台采用知识库增强方案，结合大模型理解和外部知识库，能够在短时间内更新信息并生成高时效的精准回复。相比模型精调，显著降低了计算成本，同时提升了信息更新的效率。

大模型知识引擎的优化。大模型知识引擎通过多模态解析、检索和问答提取等功能，能够快速处理复杂文档和多轮交互场景，提升知识问答的覆盖率和精准率。这种能力确保了智能体在动态环境中的高效响应。

Agent模式的灵活性。在Agent模式下，智能体能够根据用户指令自主规划任务，灵活调用外部工具和知识库，快速生成符合用户需求的回复。这种模式特别适用于需要高时效响应的生活咨询和百科知识场景。

知识生产与更新的自动化。智能体开发平台支持快速导入企业级知识文档，并通过自动化扩充知识库，大幅缩短了知识生成与更新的周期。这种方式替代了传统的人工问答梳理和对话树设计，显著提升了效率。

### 智能体开发平台能实现可追溯

智能体开发平台通过知识库增强方案、答案来源追溯、调试信息展示以及反馈与排查机制，全面实现了智能体开发与应用过程中的可追溯性，确保结果的可靠性和透明度，确保了结果的来源可靠，便于追溯。

知识库增强方案。通过结合大模型理解和外部知识库，智能体开发平台确保了结果的来源可靠，外部知识库的连接使得信息更新简单快捷，同时保证了知识来源的可追溯性。

知识库答案来源追溯。在用户端体验中，智能体开发平台支持对答案的来源进行引用展示。对于来源于文档的答案，支持点击查看对应来源切片，确保用户能够追溯答案的具体出处。

搜索来源答案追溯。对于来源于网页地址的答案，智能体开发平台同样支持引用展示，并允许用户点击跳转至原始网页，进一步验证答案的准确性。

调试信息展示。在调试过程中，智能体开发平台会展示当前会话的运行情况，包括运行链路、耗时和过程数据。调试信息中包含每个运行步骤的请求信息和输出结果，如content数据、设定的角色指令提示词、检索到的切片信息等，确保每一步操作都可追溯。

反馈与排查机制。用户可以通过平台反馈问题，平台运营方参考提供的信息进行排查并反馈排查结果，确保问题处理过程透明且可追溯。

### 智能体开发平台需要满足数据安全要求

智能体开发平台通过敏感数据保护、知识库增强方案、调试信息与反馈机制、系统管理与调用统计、原子能力与API安全以及多租户功能与数据隔离等多重措施，全面保障了数据的安全性，敏感数据无需参与模型训练，降低了泄露风险。

敏感数据保护。智能体开发平台采用知识库增强方案，确保敏感数据无需参与模型训练，降低了数据泄露的风险。这种方式避免了传统模型精调中因数据参与训练而可能引发的安全问题。

调试信息与反馈机制。在调试过程中，智能体开发平台会展示当前会话的运行情况，包括运行链路、耗时和过程数据，确保每一步操作都可追溯。用户可以通过反馈机制提交问题，后台运营人员将参考提供的信息进行排查，确保问题处理过程透明且安全。

系统管理与调用统计。智能体开发平台提供详细的调用统计和用量统计，支持查看资源消耗的总数和调用明细，包括用量统计、并发统计和知识库容量统计。这些功能帮助企业监控和管理数据使用情况，确保数据在可控范围内安全运行。

原子能力与API安全。智能体开发平台以API形式提供原子能力接口，支持具有开发能力的用户自行搭建大模型应用，拓展大模型能力边界。通过严格的API调用管理和权限控制，确保数据在传输和使用过程中的安全性。

多租户功能与数据隔离。智能体开发平台支持多租户功能，不同主账号之间数据隔离，确保数据不互通。通过角色管理和权限设置，进一步细化了数据访问控制，防止未经授权的数据访问和操作。

### 智能体开发平台能带来使用成本降低

智能体开发平台通过知识库增强方案、知识生产与更新的自动化、大模型知识引擎的优化、原子能力与API服务以及多租户功能与资源隔离等多重措施，全面提升了成本效益，帮助企业在大模型应用中实现降本增效，相较于模型精调，显著降低了使用成本。

知识库增强方案降低计算成本。相较于传统的模型精调，智能体开发平台采用知识库增强方案，显著降低了大模型计算成本。通过连接外部知识库，平台能够快速更新信息并生成高时效的精准回复，避免了模型精调中因训练周期长而导致的高成本问题。

知识生产与更新的自动化降低运营成本。平台支持快速导入企业级知识文档，并通过自动化扩充知识库，大幅缩短了知识生成与更新的周期。这种方式替代了传统的人工问答梳理和对话树设计，显著降低了人力成本。

## 3.2 基础大模型能力

基础大模型是智能体的核心驱动力，首先，基础大模型赋予智能体强大的自然语言理解、推理和决策能力，使其能够自主感知环境、制定决策并执行任务。其次，大模型的能力覆盖了复杂任务处理、多模态信息理解以及长期记忆管理等方面，显著提升了智能体的性能。

然而，随着大语言模型的规模不断扩展，其在训练和部署过程中对于计算资源的消耗急剧增加，如何优化大语言模型智能体系统的资源效率是当前面临的一个巨大挑战。此外，基础大模型在支撑智能体系统时，需要进行针对性的优化与适配，例如在理解复杂指令、处理长期记忆信息等方面，大模型的表现需要进一步优化与改进。尽管大语言模型智能体在虚拟仿真任务中已经取得了重要进展，但是在真实世界仍存在较多的应用问题，例如将智能体应用在机器人上时，机器人的硬件在很多时候并不能准确地工作。

基础大模型同智能体之间有多方面的关系，包括：组件的关系、赋能的关系、构建的关系、效率的关系。处理好这些关系，才能更好的做智能体开发和运营。


▲ 模型能力快速提升，AI原生应用兴起

### 基础大模型是智能体的核心组件

基础大模型为智能体提供强大的自然语言理解、推理和决策能力。智能体通过调用大模型，自主感知环境、制定决策并执行任务。具体而言，基础大模型赋予智能体处理复杂任务、理解多模态信息以及管理长期记忆的能力，使其在开放和动态场景中表现出更高的适应性和扩展性。基础大模型通过记忆组件、规划组件和执行组件的协同工作，使智能体能够有效感知环境、制定决策并执行规划的动作，进而完成相应任务。此外，大语言模型智能体还能够根据环境的实时反馈动态调整自身的行为策略，进一步提升其在复杂场景中的应用效果。然而，随着大语言模型规模的扩展，其在训练和部署过程中对计算资源的消耗急剧增加，优化大模型智能体系统的资源效率成为当前面临的一个关键挑战。

### 基础大模型赋能智能体高适应性

基础大模型为智能体赋能，使其能够在开放和动态场景中表现出更高的适应性和扩展性。电信运营商业务的大模型能力覆盖了复杂任务处理、多模态信息理解以及长期记忆管理等方面，显著提升了智能体的性能。在多模态信息理解方面，大模型能够同时处理文本、语音、图像等多种形式的数据，例如在智能客服场景中，结合语音识别和自然语言处理技术，实现更精准的客户需求理解和响应。此外，基础大模型通过记忆组件管理智能体的长期记忆，使其能够存储和检索历史交互信息，从而为客户提供个性化的服务体验，例如根据客户历史记录推荐合适的套餐或解决重复性问题。在开放和动态场景中，智能体可以实时感知网络状态变化，动态调整资源配置，确保网络稳定性和服务质量。通过知识库增强方案，大模型显著提升了智能体的扩展性和效率，使其能够快速接入最新的行业知识库，实现知识的实时更新和高效利用，从而降低运营成本并提升服务能力。

### 基础大模型面临资源消耗增加和成本控制的问题

基础大模型的资源消耗是智能体系统面临的重要挑战。随着大模型规模的扩展，其在训练和部署过程中的计算资源消耗急剧增加。对于单个智能体来说，通常每次动作行为都需要对大语言模型进行调用，导致整个过程中产生了较高的调用成本。进一步，在多智能体系统中，当需要多个大语言模型智能体协同工作时，资源消耗问题更为严重，导致当前的多智能体系统往往不能扩展较大规模的智能体数量。效率问题已经成为制约其在智能体系统广泛部署的一个重要因素。

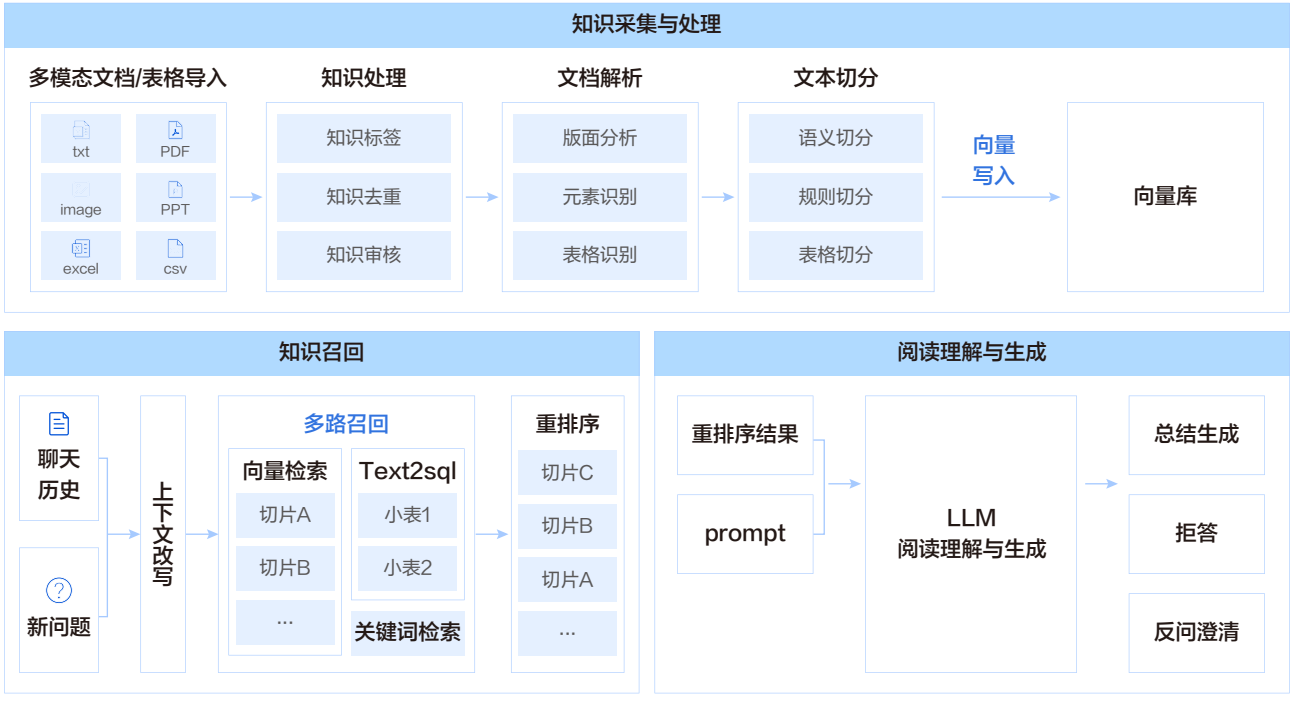
电信运营商的大模型应用场景包括：智能客服、网络优化、故障诊断等，这些场景对计算资源的需求尤为突出。例如，在智能客服场景中，大模型需要处理大量的自然语言交互，实时生成响应，这对计算资源的要求极高。此外，电信运营商的业务通常涉及大规模的数据处理和多智能体协同工作，进一步加剧了资源消耗问题。

为了应对这一挑战，需要采取多种优化策略。首先，可以通过模型压缩和量化技术减少大模型的参数量和计算复杂度，从而降低资源消耗。其次，采用分布式计算和边缘计算技术，将计算任务分散到多个节点，提高资源利用效率。此外，还可以通过知识库增强方案，利用已有的行业知识库减少大模型的训练和推理成本。

## 3.3 检索增强生成能力

搜索增强生成，简称RAG（Retrieval-Augmented Generation），是一种结合信息检索与生成式人工智能（如大语言模型）的技术框架，旨在通过动态检索外部知识库中的相关内容，增强生成模型（LLM）的输入信息，从而提升回答的准确性、时效性和可解释性。其核心思想是让模型在生成答案前“查资料”，解决传统LLM的知识固化、幻觉问题及专业领域覆盖不足等缺陷。

RAG通过整合外部知识库和优化检索策略，显著提升了智能体在业务中的决策能力、信息整合与生成能力，以及动态调整与优化能力。




▲ 采集、处理、检索信息再交由大模型处理

### RAG为大模型供给信息

信息检索与生成结合：RAG通过检索外部知识库中的相关信息，将其作为上下文输入给大模型，从而提升生成内容的质量。其工作流程包括三个关键步骤：将语料库划分为离散块，构建向量索引，并根据与查询和索引块的向量相似性来识别和检索块。

外部知识库整合：RAG通过引入外部权威信息，显著提升大模型内容生成的准确性和丰富度，减少错误和臆测。生成的内容可以追溯到具体的信息源，提高透明度和可解释性。

### RAG提供检索增强能力

检索增强策略：RAG通过增强有针对性的检索策略和改进索引方法来弥补朴素RAG的缺点。它实施了预检索和后检索策略，并采用了滑动窗口、细粒度分割和元数据等技术来改进索引方法。

模块化RAG框架：模块RAG结构具有高度的灵活性和适应性，可整合各种方法增强功能模块，解决特定问题，支持多模块间串行流水线或端到端训练方法。

语义表示与对齐：RAG中的语义空间对于查询和文档的多维映射至关重要，建立准确语义空间的方法包括块优化和管理外部文档的微调嵌入模型。

### RAG强化智能体思考能力

增强决策能力：RAG为基于大型语言模型的自主智能体提供更广泛的信息访问能力，增强其决策和问题解决能力。智能体可以使用RAG从自己的外部记忆中检索相关信息，以增强其理解和决策能力。

提升信息整合与生成：智能体将RAG获取的信息与内部知识库结合，生成更全面、准确的答案。例如，在账单查询场景中，工作流可以调用计费系统API，生成精准的账单信息并返回给用户。

动态调整与优化：智能体根据RAG获取的最新信息动态调整其行为策略。例如，在网络优化任务中，智能体可以实时获取网络状态数据，并动态调整资源配置。

## 3.4 工作流生成能力

工作流是一种通过可视化画布和多种节点（如大模型节点、参数提取节点、条件判断节点等）编排复杂业务流程的工具，旨在实现稳定且可控的业务效果，确保每个流程节点的准确性和可解释性。工作流的实现原理和关键技术主要体现在通过自动化编排和分层模型设计，实现复杂业务流程的高效管理和执行。

### 工作流的工作原理

自动化编排：工作流通过自动化编排体系，将复杂的业务流程分解为多个可执行的原子任务，并通过流程驱动的方式调用各任务的能力，实现端到端的业务开通和运维。

分层模型设计：工作流采用分层模型设计，拉通业务、服务、网络、资源的端到端设计过程，确保每个环节的任务能够高效执行。

意图编排：工作流通过意图编排，将业务对网络的诉求转化为网络可理解、可配置、可度量、可优化的对象和属性，完成从模型转换、仿真验证、业务发放到策略闭环的自动化工作流。

### 工作流的关键技术

业务编排器：业务编排器抽象底层能力，提供开放API给运营门户，实现业务端到端检测及订单流程管理。南向对接云侧及网侧编排层，支撑实现业务开通流程的全程贯通。

网络编排器：网络编排器通过分层模型设计、原子能力设计、网络服务编排设计和编排任务设计，解决入云网络配置和大网资源拉起难的问题和挑战。

融合编排器：融合编排器通过业务驱动和技术驱动，实现产品组合营销驱动资源编排能力的形成，支撑市场营销的快速响应与方案构建。

工作流引擎：工作流引擎内置多层次的网络功能模块，通过Restful接口的方式对外提供能力，支持原子API的调用顺序编排，实现基于租户网络的意图服务化接口。

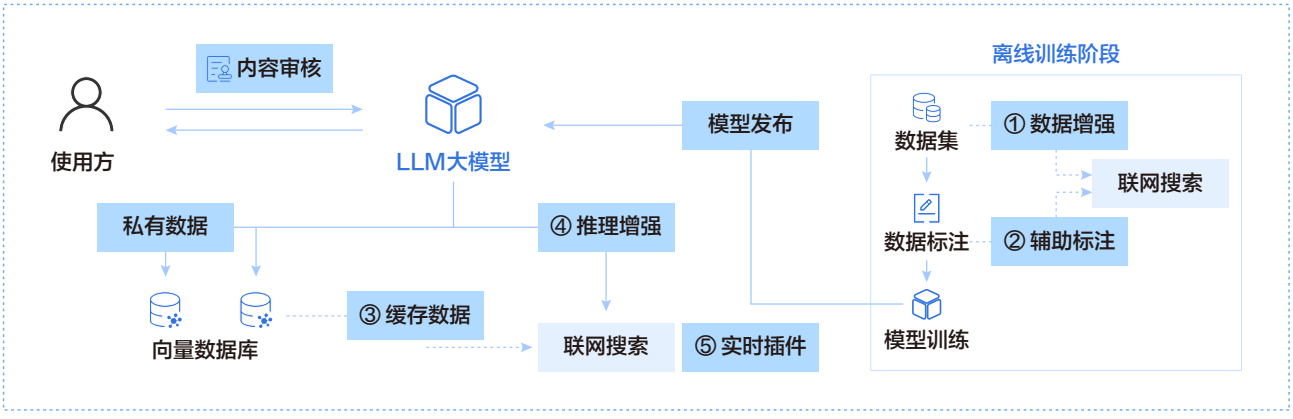
策略闭环引擎：策略闭环引擎支持通过策略引擎实现灵活的闭环处理，实现网络的高度自治，例如在检测到租户网络时延超限时，触发路径重优化。

仿真验证引擎：仿真验证引擎通过控制面/数据面的仿真验证能力，为网络变更评估提供客观、可靠的定量依据，保障连接服务意图编排设计的正确性。

智能体通过工作流编排任务流程，同时将工作流获取的信息与内部知识库结合，生成更全面、准确的答案，如在账单查询场景中，工作流调用计费系统API，生成精准的账单信息并返回给用户。此外，智能体根据工作流获取的最新信息动态调整其行为策略，例如在网络优化任务中，工作流调用网络数据分析节点，实现故障自动诊断和资源优化，智能体实时获取网络状态数据，并动态调整资源配置。

## 3.5 联网搜索能力

智能体开发与联网搜索在数据方面是相互补充关系，主要体现在通过联网搜索功能可以增强智能体的实时信息获取能力和答案丰富性。智能体在配置联网搜索功能后，能够对接搜索引擎，提升对时事问题的回答能力。例如，当知识库中没有相关答案时，智能体可以通过联网搜索获取最新的信息，从而提供更全面和准确的响应。这种能力对于电信运营商尤其重要，因为其业务涉及广泛的用户需求和动态变化的行业信息，联网搜索功能可以帮助智能体更好地满足用户对实时信息的需求，提升服务质量和用户体验。此外，联网搜索功能还可以与智能体的知识库问答能力结合，形成检索增强生成（RAG）框架，进一步提升智能体的回答准确性和效率。




▲ 联网搜索架构图

### 联网搜索获取多类型信息

实时信息：联网搜索能够获取最新的实时信息，例如天气预报、新闻动态、市场数据等，这对于需要最新知识的任务至关重要。

外部知识库：通过联网搜索，智能体可以访问外部知识库，获取特定领域的专业知识或历史数据，从而丰富其知识储备。

多模态信息：联网搜索不仅限于文本信息，还可以获取图像、音频、视频等多模态数据，帮助智能体更全面地理解任务需求。

### 智能体应用联网搜索信息

任务规划与决策：智能体利用搜索获取的信息进行任务规划和决策。例如，在规划旅行时，智能体可以通过搜索获取目的地的天气、酒店信息等，并基于这些信息制定最优方案。

信息整合与生成：智能体将搜索获取的信息与内部知识库结合，生成更全面、准确的答案。例如，在回答用户问题时，智能体可以结合搜索到的实时数据和历史知识，提供更可信的响应。

动态调整与优化：智能体可以根据搜索获取的最新信息动态调整其行为策略。例如，在网络优化任务中，智能体可以实时获取网络状态数据，并动态调整资源配置。

### 联网搜索提升智能体使用体验

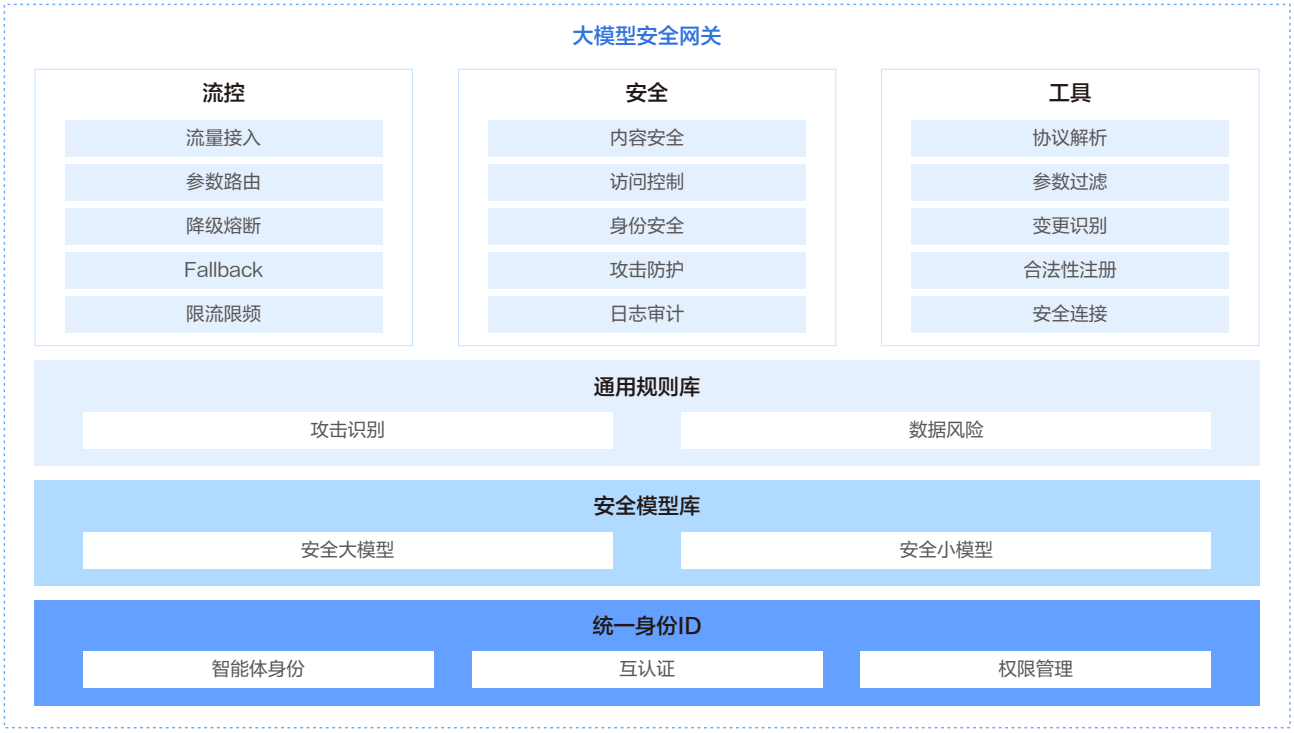
降低错误率：通过获取实时和准确的外部信息，联网搜索能够显著降低智能体的幻觉和回答错误率。

扩展应用场景：联网搜索使智能体能够处理更多依赖外部信息的任务，例如智能客服、市场分析、医疗诊断等。

提升用户体验：通过提供更准确、实时的信息，联网搜索能够提升智能体的服务质量和用户体验。

## 3.6 智能体安全管理能力

智能体开发的安全问题呈现多维渗透特征，需从数据、模型、行为三方面构建纵深防御体系。大模型安全网关通过多层次防护机制，连接智能体、模型与服务，实现统一治理与高效协同，确保AI规模化应用中的关键风险得到有效控制，有效平衡安全性与效率，成为智能体规模化落地的关键基础设施。未来，随着多模态技术与自适应算法的融合，安全网关将进一步提升智能体应用的可靠性与可信度。




▲ 大模型安全网关框架

智能体开发过程中，数据泄露与滥用风险突出，敏感信息（如用户隐私、企业机密）因权限管理粗放易被窃取，训练数据污染还可能导致模型输出偏离预期；恶意攻击手段多样，包括提示注入、工具调用劫持等，易诱导智能体执行危险操作；模型行为失控问题显著，幻觉输出或越权决策可能引发医疗误诊、金融欺诈等高风险事件，责任界定困难；此外，第三方工具链投毒、沙箱逃逸等生态协同风险，进一步加剧了安全防护的复杂性。

### 大模型安全网关构建全链路防护体系

大模型安全网关通过“全链路拦截+动态防御”机制，系统性解决上述挑战：在输入输出环节，采用多模态审核与生成内容过滤，阻断恶意指令与有害输出；数据安全层面，通过动态脱敏与零信任访问控制，限制敏感数据访问范围；针对模型与工具风险，实施MCP协议防护与模型对齐加固，防御投毒攻击并约束行为合规性；同时，结合流量管控与溯源、自适应策略更新等动态防御手段，实现从数据、模型到行为的全层级安全覆盖。

### 典型场景验证安全网关实效

在金融行业，安全网关通过内容审核与数据加密保障交易合规，拦截欺诈转账；医疗领域，沙箱隔离医疗数据防止隐私泄露，模型对齐避免误诊建议；政务场景中，网关审核政策文件生成内容，确保符合法规要求。实践表明，安全网关能有效平衡智能体应用的效率与安全性，是规模化落地的关键基础设施。

# 04

# 智能体的产业实践

智能体的产业实践已渗透至教育行业、工业制造、医疗健康、物流交通、电商零售、教育政务等核心领域，形成规模化落地态势。在工业领域，智能体能将全球供应链决策效率大幅提升；医疗领域，睡眠健康智能体提供全病程管理服务，医院在探索多智能体协同的AI医院模式；物流与交通领域，整合导航、外卖等多个智能体实现“一句话订餐”，自动驾驶汽车通过具身智能完成复杂路况决策。头部互联网企业如腾讯、阿里、火山等，通过开放智能体开发平台，加速垂直场景渗透。当前，智能体正从单点工具向系统化解决方案演进，成为驱动产业智能化转型的核心引擎。

# 4.1 应用场景

### 智能体在电信运营商行业的应用

智能体在电信运营商产业实践中已广泛应用于：网络运维优化、客户服务和垂直行业赋能。在移动网应用中，AI技术用于物联网端到端质差识别与定位、智能基站节能、无线网络异常小区发现，以及VoLTE语音质量评估。在业务服务中，包括智能语音交互、家宽视频内容智能推荐和内容分发网络智能调度。在垂直行业如电力、医疗和工业领域，通过5G网络切片实现智能运维，并支撑运营商自智网络目标。这些实践显著提升了运维效率、用户满意度和业务收入，同时降低成本和资源浪费。

### 智能体在教育行业的应用

课程大纲自动生成，通过文生文能力，输入课程主题或关键知识点，智能体基于预训练教育数据生成结构化的课程大纲，包括章节划分、学习目标、课时分配等。结合结构化内容助手功能（如生成Markdown/SVG格式），输出可直接用于教学的标准大纲。例如，教师输入“高中数学函数专题”，智能体生成包含函数定义、性质、图像及应用等模块的完整大纲，并标注重点难点。

教学案例自动创建，案例生成助手：输入知识点（如“化学反应速率”），智能体结合学科知识库生成贴近实际的应用案例（如实验设计、生活场景类比）。交互式优化：支持通过分步骤Prompt（提示词）调整案例复杂度，例如要求“增加一个跨学科融合案例”，对现有案例自动生成变体，为同一知识点生成不同难度的教学案例等。

课件配图智能生成，文生图能力：输入描述性文本（如“细胞分裂示意图”），毫秒级生成符合教学场景的配图。实时画板编辑：支持绘制轮廓草图并添加文本描述，智能体结合轮廓与提示词生成精准配图，例如标注生物器官位置、课件演示中生成“动漫风格物理力学图示”、“化学分子结构3D渲染图”等，显著提升课件吸引力。

产业实践效果方面来看，引入智能体后，教学案例生成效率提升80%，课件配图制作时间减少90%。通过知识库匹配（如教育文献库）确保案例准确性，例如，输入“量子力学基础”时优先调用权威物理教材内容，结合教学大纲自动匹配配图，实现课件内容与视觉素材的同步生成。

### 智能体在医疗行业的应用

诊前环节，精准引导与高效预诊。交互式预问诊，基于大模型的智能体通过自然语言对话收集患者主诉、病史、用药禁忌等信息，生成结构化预诊报告供医生参考。智能导诊与预约，根据患者症状推荐科室及医生，支持分时段预约。例如腾讯健康AI预问诊系统助力深圳市人民医院提升服务效率，系统准确率87%，月均使用超2万人次；京东健康“康康”助手可精准匹配医疗资源，缩短患者决策路径。

诊中环节，临床决策支持与流程优化。通过多模态大模型辅助诊断，大模型整合文本、影像、基因等多源数据，辅助医生诊断。例如，金域医学基于腾讯云的算力平台和大模型接入能力，持续构建和迭代“域见医言”大模型，开发智能体应用“小域医”，广泛服务于基层医疗机构。同时，“小域医”在智能问答、智慧报告、安全用药等场景中与医疗产品多维融合，为基层医疗机构提供更完善的解决方案，助力分级诊疗落地。

诊后环节，延续性服务与健康管理。智能随访管理，全周期患者管理系统支持医生制定随访计划，患者上传居家监测数据（如血压、血糖）。结合物联网硬件实现动态健康监测，实现个性化风险评估。

### 智能体在传媒行业的应用

智能写作智能体，能实现语义级校对与创作赋能，智能体通过规划、记忆、工具调用能力，实现写作流程优化。深度审校、编辑体例勘误智能体能结合企业历史体例规范文档，对文章进行语义级纠错，输出勘误结果及修改建议，解决传统工具无法处理的语义错误。创作辅助，提供全流程写作辅助（构思建议、内容润色、修缮检查），提升企业员工创作效率。

多场景“特专精”智能体，支持产业大会报道等专题快速构建。

内容审核智能体，能通过规则与语义的协同校验，整合规则引擎与大模型理解能力。动态规则解析，待审核内容通过加密请求匹配规则库，由LLM分析违规可能性并返回理由（如敏感内容标注依据）。

长效学习机制，向量数据库存储规则库，支持持续更新审核策略；短期记忆缓存任务上下文，提升复杂场景适应性。

视频创作智能体，实现全流程智能化生产，大模型视频创作引擎是驱动智能体自动化处理的关键。一键成片工具，整合音视频AI技术，实现文生视频、视频转译、图片动态化。

## 4.2 面临挑战

**智能体的技术发展，面临大模型稳定性不足、多智能体协同不顺等问题**

模型准确性与稳定性。当前主流大语言模型在自然语言生成方面表现出色，但在具体场景中容易产生事实错误的“AI幻觉”。对于涉及业务流程、数据分析、合规判断等高风险任务，准确性成为智能体系统的首要保障目标。

上下文理解能力。复杂任务往往依赖于对业务流程、用户角色、历史操作等上下文的深度理解。若上下文注入机制不完备（如函数调用时序错误），将导致模型输出偏离意图，影响任务执行的正确性。

可解释性与可控性。智能体行为的不确定性对IT管理带来挑战。运维人员和审计人员需要理解模型为什么输出某个答案、调用了哪个插件、用了哪些数据，但目前智能体的可解释性仍然较差。

多智能体协同。随着智能体数量增加，通信和计算复杂度呈指数级增长。如何有效地进行多智能体协同，避免信息孤岛和通信兼容性问题，是一个重要的技术挑战。


▲ Multi-Agent：通过Handoff的自由范式协同多个Agent

### 智能体的应用过程，面临三大痛点：环境适应、数据质量、行业规范

- 复杂环境适应性。智能体在复杂现实环境中的适应能力不足，尤其在理解复杂业务逻辑和执行多任务时表现不佳。企业在将智能体与现有业务系统对接时面临技术和成本挑战，且在特定行业的应用能力有待提高。
- 数据质量与数量。数据质量和数量对智能体的性能至关重要，低质量或不完整的数据会影响其决策效果。企业并没有为智能体做好数据准备，数据收集和整理工作困难。
- 行业标准化与规范。智能体在不同行业的应用标准和规范尚未完善，增加了应用难度。例如，金融、能源、制造等对可靠性、安全性、专业性要求极高的行业，智能体的落地并非坦途。

### 智能体生态发展，面临初期教育成本高、生态建设慢等问题

商业化落地困难。市场对智能体的期望与实际应用效果存在差距，企业对其收益回报不满意。智能体的商业化落地困难，企业需要明确其在业务流程中的价值。

生态建设与接口标准。目前缺乏统一的API标准和开发框架，限制了智能体的进一步发展。企业在引入智能体架构时，面临技术和成本挑战，生态建设也相对薄弱。

员工抵触与文化因素。员工的抵触往往成为技术实施的重大障碍，特别是那些使用此类技术的员工，如果他们不相信AI有能力承担任务，或者认为它对自身的工作构成了威胁，抵触情绪将尤为强烈。

### 智能体的演进，面临标准建设滞后，试错成本较高等问题

与大模型深度融合。将大型语言模型（LLM）作为智能体的认知核心，实现复杂任务规划、知识整合和自然交互。

具身多智能体系统。结合机器人技术，使智能体能够在物理世界中协作，提升其在动态环境中的适应能力。

统一标准与生态建设。需统一智能体定义与行业标准，促进产业规范化；优化生态体系，搭建多方协同创新平台，推动产业融合。

## 4.3 解决方案

通过不断的技术创新、生态建设和政策支持，智能体在克服这些挑战方面取得了显著进展。未来，随着技术的进一步发展，智能体将在各个行业中发挥越来越重要的作用。

### 加大多模态技术的投入，应对技术挑战

模型准确性与稳定性。当前主流大语言模型在自然语言生成方面表现出色，但在具体场景中容易产生事实错误的“AI幻觉”。通过多模态技术（图像、视频、文本等）结合，提高模型的准确性和稳定性。此外，提供从数据处理、模型训练到知识融合的全流程支持，帮助企业提升智能体的性能。

上下文理解能力。复杂任务依赖于对业务流程、用户角色和历史操作的深度理解，若上下文注入机制不完备，将导致模型输出偏离意图。引入工作流模式和全局视野的Agent，支持智能体的逻辑编排和动态调度，增强智能体对复杂上下文的理解能力。

多智能体协同。随着智能体数量增加，通信和计算复杂度呈指数级增长，如何有效进行多智能体协同是一个重要挑战。通过模块化的方式提供大模型、开发工具和数据服务，支持多智能体的协同工作。该平台帮助机器人具备感知世界、规划任务和自主决策的能力。

大模型结合多模态技术提高了模型的准确性和稳定性，基于混元、星辰等大模型研发的智能体开发平台可进一步提供从数据处理、模型训练到知识融合的全流程支持，帮助企业提升智能体的性能。平台通过引入工作流模式和全局视野的Agent，支持智能体的逻辑编排和动态调度，增强了智能体对复杂上下文的理解能力。具身智能开放平台（Tairos）通过模块化的方式提供大模型、开发工具和数据服务，支持多智能体的协同工作，帮助机器人具备感知世界、规划任务和自主决策的能力，从被动执行指令的机械体进化为主动适应现实世界的智能生命体。

### 提供智能体开发平台，提升“应用”供给效率

复杂环境适应性。智能体在复杂现实环境中的适应能力不足，尤其在理解复杂业务逻辑和执行多任务时表现不佳。构建标准化的强化学习框架和提供云端算力支持，降低学术和产业界在研究与应用中的门槛。此外，通过零代码支持多Agent的转交协同方式，简化智能体的应用流程。

数据质量与数量。数据质量和数量对智能体的性能至关重要，低质量或不完整的数据会影响其决策效果。对接主流数据库和更全面的数据源，确保智能体能够获取高质量的数据，进行有效的训练和决策。

开悟平台通过构建标准化的强化学习框架和提供云端算力支持，降低了学术和产业界在研究与应用中的门槛，进一步推动了技术创新。同时通过零代码支持多Agent的转交协同方式，简化了智能体的应用流程；对接主流数据库和腾讯文档等更全面的数据源，确保智能体获取高质量的数据，进行有效的训练和决策。

### 根据企业特点，积极参与商业、标准、文化三个生态建设

- 商业化落地困难。市场对智能体的期望与实际应用效果存在差距，企业对其收益回报不满意。提供从数据处理到部署上线的全流程支持，帮助企业快速实现智能体的商业化落地。此外，通过黑客松Agent应用创新挑战赛等赛事，鼓励开发者探索智能体的应用场景。
- 生态建设与接口标准。目前缺乏统一的API标准和开发框架，限制了智能体的进一步发展。提供统一的开发环境和接口标准，支持多模型和工具的协同工作。这些平台帮助开发者高效构建AI原生应用，并实现智能体在不同平台和渠道的分发。
- 员工抵触与文化因素。员工的抵触往往成为技术实施的重大障碍。通过简化智能体的使用流程和提供丰富的培训资源，降低员工的学习成本，提升其对智能体的接受度和使用意愿。

### 效果验证与可持续性挑战

缺乏评测框架。智能体效果依赖人工测试，试错成本高。内置应用调试、评测、发布工具链，支持效果量化。

长期知识沉淀困难。业务数据分散，难以转化为模型可用知识。通过实际应用积累数据集，支持智能体工作流沉淀数据。

# 05

# 未来发展趋势

技术创新方面，智能体围绕自主智能与自适应系统持续发展。即指系统不仅具备高性能计算能力，还能自主感知环境变化、调整行为策略，并在复杂场景中实现持续优化的能力。

自主感知与决策能力增强。未来系统将结合多模态感知（文本、语音、图像、传感器数据）与高级认知算法，使智能体能够主动理解环境变化并做出决策。例如，在智能制造中，生产线智能体可实时监测设备状态、产能波动和供应链变化，并自动调整生产计划，减少人工干预，提高生产效率。

自适应学习与持续优化。大模型与智能体系统将具备在线学习和自适应能力，通过动态数据反馈不断优化模型参数和决策策略。强化学习、元学习及迁移学习将成为关键技术支撑，使系统能够应对未知任务和快速变化的环境。例如，金融风控智能体可根据市场波动和新兴风险实时调整策略，保持高准确性和鲁棒性。

端-云协同与资源动态调度。自主智能和自适应系统需要高效利用算力资源。未来技术创新将侧重于端-云协同计算、异构算力调度和模型压缩优化，实现系统在高性能中心与边缘设备间的智能负载分配。这不仅提升运行效率，还降低成本，使智能体系统能够在多场景、多环境中高效部署。

自主智能驱动工程创新。技术创新不仅局限于算法，也推动系统工程方法论升级。全生命周期闭环管理、自动化训练与部署、智能监控与优化，将与自主智能能力紧密结合，使系统工程从静态管理向动态、自主、可持续发展演进。

发展机遇方面，AI大模型向普惠应用发展，全球创新环境趋于平权。随着智能体驱动的大模型系统工程在性能和效率上的持续突破，尤其是在算力成本和训练成本显著降低的背景下，全球范围内创新环境正朝着更加平权的方向发展。这一趋势使得不同国家、不同规模的企业均有机会参与到智能体应用创新中，打破了传统上技术垄断与资源集中带来的壁垒。

过去，顶级大模型的训练和部署需要庞大的计算资源、海量高质量数据以及丰富的资金投入，这使得少数技术巨头掌控了大部分AI技术发展的话语权和市场优势。而如今，随着云计算、分布式训练框架、开源模型和优化算法的普及，智能体系统工程的门槛大幅降低。像Meta的LLaMa系列模型开源、Hugging Face平台提供的预训练模型与工具链，都为中小企业和科研机构提供了极大便利。

这种成本下降带来的“平权”效应，不仅体现在技术资源的普及上，还体现在创新空间的开放上。企业不再仅仅依赖大规模训练来提升性能，而更多聚焦于基于智能体架构的应用创新，如行业定制智能体、跨领域多智能体协同、智能辅助决策等。新兴市场国家和地区可以利用本地特色数据和需求，设计出贴近本地实际的智能体应用方案，从而实现差异化竞争和快速迭代。

生态协同方面，构建智能体驱动的开放协同生态系统。智能体系统工程不仅是一套技术集成和应用解决方案，更是一个多方协作、多维度融合的开放生态体系。

未来，智能体系统工程将打破传统“闭环”模式，转向开放、模块化、标准化的生态架构。这包括电信运营商、云服务商、开源社区、产业联盟、终端用户等多方参与，形成技术创新、资源共享和应用落地的协同网络。以LangChain、AutoGPT为代表的开源智能体框架，正在成为推动这一生态活跃度的关键引擎，使得更多开发者和企业能够快速构建和迭代智能体应用。

生态的开放性使得智能体能够灵活调用各类异构资源，包括不同厂商的大模型、工具API、数据服务及边缘计算节点，实现跨域、跨平台的能力协同。这种跨界融合不仅提升了智能体系统的适应性和扩展性，也催生了更多创新业务形态，如智能体辅助的智慧城市、数字孪生、跨行业智能决策支持等。

与此同时，生态中的协同创新机制也将成为技术突破的重要驱动力。通过多主体共享技术进展、联合研发新算法和共建行业标准，智能体技术的演进速度和质量将大幅提升。政府、企业和学术界的紧密合作，将推动智能体在安全、伦理、隐私保护等方面的规范建设，保障生态健康有序发展。

腾讯云小程序