腾讯云 | 腾讯研究院 Tencent Research Institute | Gartner

企业级智能体
产业落地研究报告

从场景试点到规模化应用实践

引言

人工智能的发展正迎来一个决定性的转折点，过去，AI在大多数场景下扮演着“辅助工具”的角色，辅助人类优化信息检索、内容生成与数据分析；而如今，一个全新的范式正在崛起——智能体（AI Agent），正推动AI从“辅助工具”向“自主生产力”发生深刻的身份跃迁。这场变革的核心在于，AI不再仅仅是响应指令的被动执行者，而是进化为能够自主理解目标、规划路径、调用工具并与物理或数字世界交互的“数字员工”。想象一下，未来的企业营销与运营人员不再需要“手把手”地执行繁琐的跨系统操作，只需用自然语言表达一个战略目标，由智能体组成的“虚拟团队”便能自主协作：市场分析智能体负责抓取并分析竞品动态与用户画像，创意智能体生成多版本的广告文案与视觉素材，投放智能体则自动在各大平台创建并优化广告活动，最终由数据分析智能体生成一份完整的复盘报告。这标志着人机协作的边界被彻底重塑。

这场变革的背后，是大型语言模型在推理、规划与工具调用能力上的飞跃。一方面，先进模型的“思维链”（Chain-of-Thought）与“反思”（Self-Reflection）机制，赋予了智能体类人的规划与纠错能力，使其在面对复杂任务时，能够自主拆解步骤、评估中间结果并动态调整策略，自主完成复杂任务。另一方面，模型原生工具调用能力的成熟，让智能体获得了连接外部世界的“双手”。通过无缝调用API、数据库与各类应用程序，智能体得以将模型的“思考”转化为对外部世界的真实“行动”，无论是查询实时航班信息、执行一笔线上交易，还是控制一台工业机械臂，都成为可能。这种“大脑（自主规划）+双手（工具调用）”的协同结构，构成了智能体的核心，使其具备了真正意义上的任务闭环能力。

随之而来的是智能体应用形态的百花齐放。在个人生活领域，它正从简单的聊天机器人演变为无所不包的“数字伙伴”，能够管理你的日程、筛选信息、处理邮件，甚至在你授权下完成订餐、购物等生活琐事，逐步成为个性化的“生活操作系统”。在企业运营中，智能体以“嵌入式”或“产品化”的形态，深度融入营销、客服、研发、财务等核心业务流。从处理海量高频咨询的“高效助手”，到串联多个系统完成复杂流程的“执行专家”，再到辅助进行市场分析的“决策专家”，智能体的角色愈发多元且关键。更有甚者，多个智能体构成的协同网络，正以“虚拟项目组”的形式，自主完成软件开发、市场研究等复杂项目，预示着一种全新的组织形态与生产关系正在形成。

与此同时，智能体的能力边界正从数字世界向物理世界延伸。当智能体的“大脑”与机器人、自动驾驶汽车、智能家居等硬件深度融合，具身智能便应运而生。它不仅能“想得明白”，更能“动得精准”，在复杂的物理环境中完成导航、操控与交互任务，推动AI从“数字大脑”走向“现实代理人”。这不仅将深刻改变制造业、物流、养老等行业的面貌，也为通用人工智能（AGI）补上了与物理世界互动的关键一环。

智能体不仅是一项技术的演进，更是一场生产力的革命。它将人类从重复性、流程化的工作中解放出来，让我们得以专注于更具创造性与战略性的思考。本报告将深入剖析智能体的核心能力、应用场景、技术挑战与未来趋势，为企业提供一份清晰的路线图，共同迎接由智能体驱动的、人机深度共生的新纪元。

CONTENTS

目 录

CONTENTS

目 录

PART 04

智能体先锋实践

CONTENTS

目 录

PART 04

智能体先锋实践

CONTENTS

目 录

PART 05

智能体发展展望

01

智能体概念

智能体的定义与形态

智能体(AI Agent)带给人们最大的想象空间，在于其“自主完成工作”的能力。在过去，AI更多地被视为一种“生产工具”，辅助人们完成各种任务；而如今，随着AI Agent的发展，AI正逐渐从生产工具演变成“生产力”本身。从本质上来看，AI Agent是由自主性（Autonomy）与行动力（Action）共同构成的智能系统，可形象概括为“大脑+手”的协同结构。“大脑”不仅要能自主思考，还应能与环境交互，并根据环境变化动态调整自身行为策略；“手”则需要根据“大脑”的指令直接完成工作(例如Deep Research)，还能使用外部工具（例如Tool calling）。其行为不再是静态响应，而是包含规划、执行、调整的完整循环，从而实现真正意义上的任务闭环。根据其架构和组成方式，AI Agent可分为狭义和广义两类：

狭义智能体（AI Agent）强调在无需持续人工干预的情况下，实现自我学习与优化，具备高度的环境适应与泛化能力。其核心是模型本身具备原生工具调用与任务闭环执行能力。

广义智能体系统（Agentic AI System）则更具包容性，泛指一切能够感知环境、决策并执行任务以达成目标的系统。它通常依托“模型推理能力（Reasoning）+任务指令（Instruction）”构成“引导式自主（Guided Autonomy）”，并通过“工作流（Workflow）+工具调用（Tool Use）”实现“预定义行动（Pre-defined Action）”。

但我们不应该过分扩大化Agentic AI system的概念。我们认为，“行动”（Action）应该成为现阶段AI Agent的最低定义。AI Agent不应仅以“能输出内容”作为标准，而需满足“能自主调用工具并对外部世界产生结构性影响”的基本条件。最简单的例子就是“行动”不等于“回答”。“模型生成一句文本”是语言反应，而非行动本身；只有当系统将该输出转化为操作——例如发出请求、调用搜索、写入数据库、控制物理设备——才能构成真正的“Do”。

因此，一个AI Agent应该具备至少两个核心特征：第一，能调用模型以外的外部工具：这表明它不仅限于语言处理，还能通过搜索、数据库、API等接口扩展自身能力边界；第二，能自主执行完整任务链：即具备从目标识别、任务拆解、步骤规划到动作执行的闭环能力，且可在无持续人工指令干预下推进任务。

需要说明的是，“狭义”与“广义”并不是互斥关系。在可预见的时期内，单一的狭义AI Agent难以独立解决所有问题，实际商业落地更多体现为Agentic AI System的混合形态：既包含具有AI Agent能力的模型，也依赖外挂的工作流和工具协同。

	狭义AI Agent	广义Agentic AI System
	大脑（Autonomy）+ 手（Action）	大脑（Guided Autonomy）+手（Pre-defined Action）
主要实现方式	包含工具与环境的 端到端模型训练	模型思考能力+任务指令+工作流+工具调用
核心特征	自主规划与反馈调节	执行预设任务
自主性	高，无需持续人工干预，自我学习优化	低，遵循预设规则，具备一定决策空间，依赖Prompt
系统能力	感知环境、动态适应	响应输入、完成指令
工具调用	原生工具调用能力	以工作流规定外部工具调用
泛化能力	可将知识经验应用到全新情境	可在类似任务中复用，但适应新情境能力有限
示例	AutoGPT、多模态任务AI Agent	自动推荐系统、AI客服

表：智能体的狭义和广义定义

如果说狭义AI Agent是模型能力，那么Agentic AI System更是一种产品能力，是一种新的服务形态。当前AI Agent系统也自然地呈现出多元化的形态：

类型	形态	示例	特征说明
产品型AI Agent	AI Agent本身就是一种产品 以APP、网页、小程序等形态呈现	ChatGPT Agent、Replika Cursor、腾讯元宝	AI Agent即交互界面与使用入口
嵌入型AI Agent	AI Agent嵌入App、平台或企业系统 是对现有系统/产品的一种能力提升	Office Copilot、钉钉智能助理 企业微信智能机器人、腾讯会议AI小助手	AI Agent作为某一功能模块存在
多AI Agent协作系统	多个AI Agent协同完成任务	Devin、AutoGPT、Manus Genspark、Claude Research	内部模块分工明确，具备协作能力
隐形AI Agent	以插件、推荐系统等形式无感嵌入到 用户的使用链路中	智能搜索、日程推荐	用户无感知、系统替代行为

表：智能体的主要产品形态

尽管AI Agent形态存在差异，但其根本标志是“行动”能力——它必须能调用外部工具，并自主执行完整任务链，而不仅仅是生成文本或回答疑问。也正因此，AI Agent 才得以超越传统AI工具，成为新一代生产力变革的核心驱动力。

智能体的能力界定与分类

智能体的能力界定

我们对AI Agent的核心要求是“能干活、能落地、能实战”，这意味着其能力界定和分类不能停留在抽象层面，而必须依托可验证的评测体系，并以“世界真实性”和“行业适配性”为核心标准。然而，现有的技术测评标准仍难以全面满足这一需求。

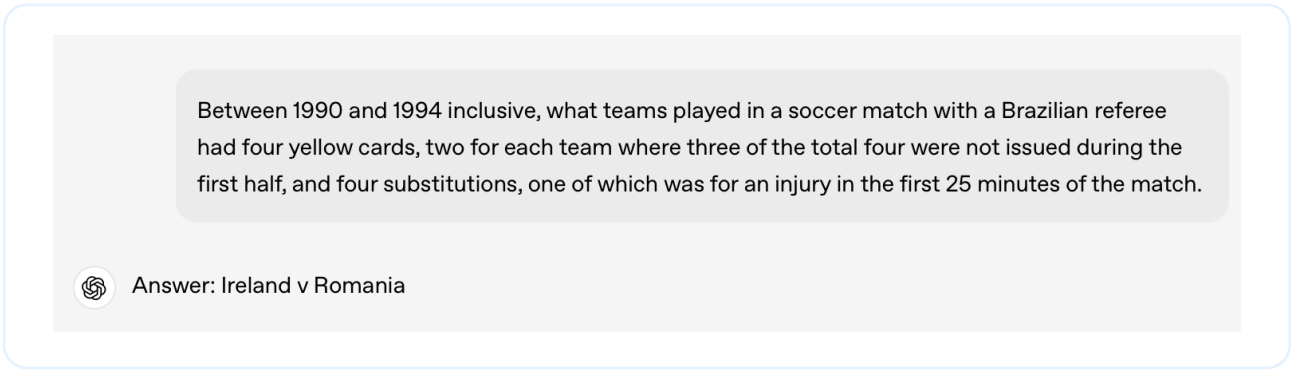
按照评测内容，目前做法大致分为三类：

第一类是模型基础能力测试，主要评估基础知识水平（如MMLU）、多模态理解能力（如MMMU）、长上下文能力（如

MRCR）、工具调用能力（如ToolBench、APIBench）以及规划和多步推理能力（如GSM8K、MATH、HotpotQA）。

第二类是通用AI Agent任务测试，侧重考察AI Agent在配备环境和工具的情况下解决多样化问题的能力，但测试范围相对有限。例如，GAIA侧重多模态理解、网页浏览和工具调用；AgentBench在统一环境中提供多种任务，测试跨领域适应性；OSWorld、OmniACT、AppWorld则在真实或准真实操作系统环境中评测AI Agent多步操作能力，体现更完整的系统级表现。

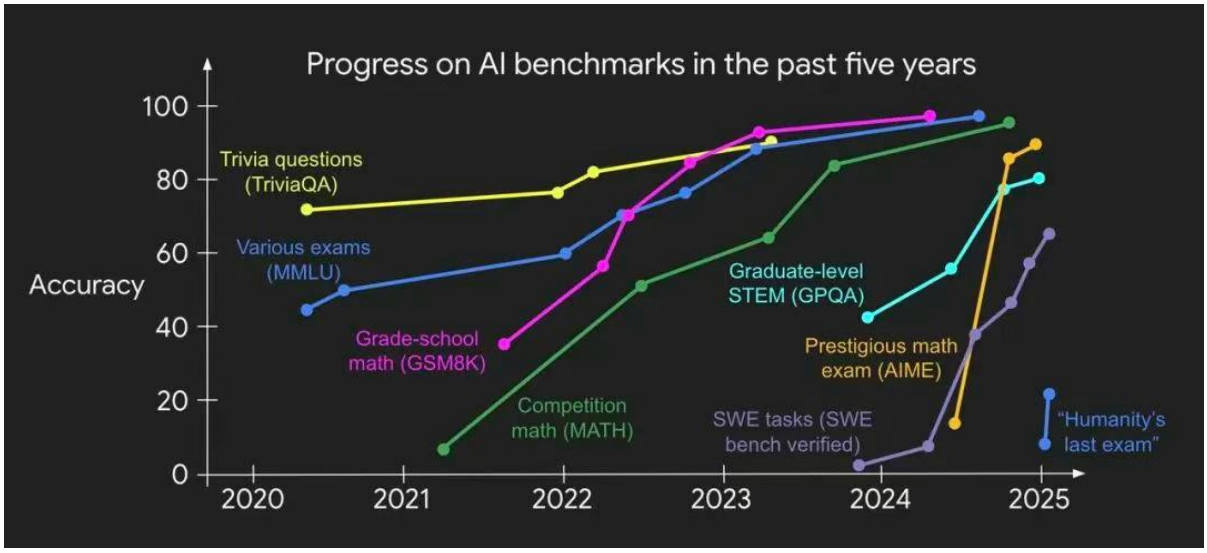
第三类是特定领域的端到端任务测试，针对行业场景构建更接近真实应用的评测。例如，在搜索能力上，OpenAI推出的BrowseComp聚焦于复杂的信息搜索任务，这些任务经过难度筛选，要求AI Agent进行多步搜索且答案不出现在首页；在软件工程上，SWE Bench及其变体基于GitHub代码仓库的真实问题单（Issue），验证AI Agent生成的代码能否解决问题；OpenAI提出的SWELancer则更进一步，通过为AI Agent分配“真实定价的外包任务”，模拟更贴近商业环境的复杂要求，并量化AI Agent的经济价值。不过，这些测试仍不能完全覆盖真实场景的复杂性。



图：BrowseComp题目举例

按评测方式来分，同样可分为三类：其一，只测最终输出 (Final Response)，只验证最终答案是否满足需求；其二，测中间过程，适用于合规、严谨性要求高的场景。包括逐步动作（Stepwise）评测，验证AI Agent每一步的对话、调用和工具执行；以及更高级的完整执行轨迹（Trajectory）评测，分析AI Agent的动作序列是否合理，并与“最优路径”对比；其三，相对评测，即通过大规模投票或对比来判断结果优劣，如Agent Arena。

然而，当前多数评测都停留在“考试型”范式，在简化的抽象场景下设置有明确环境和答案的考题，比如做题、翻译、下棋等，客观上把AI训练成了“做题家”。再难的题，AI刷榜的速度也越来越快。现有评测方式最大的问题在于：更难的题目往往不解决真实问题，而真实问题又难以抽象成可验证的题目。因此，如何把真实世界的场景任务工程化为可复现、可量化、可对比的评测，从而指导AI Agent的进步，这是AI Agent下一阶段最重要的问题之一。



图：在“考试”的测试形式下，AI的刷榜速度越来越快

一个例子是AndonLab的Vending-Bench，让AI Agent来经营自动收货机，目标是赚更多的钱。在简化的测试中，AI扮演供应商和顾客来测试AI Agent的经营能力，Claude 4平均可以赚到4倍的启动资金，而Anthropic让员工扮演真实的顾客来测试时，发现有各种各样的问题，包括给错收款账户、滥发折扣码、亏本卖货等等，险些破产。这个测试为我们评价AI Agent能力提供了思路，但它的评测效率和能给出的反馈数据非常有限。

另一个例子是红杉的X-Bench，这个测评的目标是好的：招聘和达人营销领域的实际任务。但从实际落地的角度，其测评内容还是在行业背景下的搜索子任务，而不是端到端的解决问题，相当于行业中的Junior/实习生面对的任务。更致命的是，其中的众多子任务都只列了概念，而实际上“无法测评”。

Evaluation Category	Task Type	Realizable	Testable	Included
Client Needs Communication	Client Needs Analysis	✓	✓	Y
	Business Communication	✗	✗	N
KOL Search	KOL Search	✓	✓	Y
	KOL Screening	✓	✓	Y
	KOL Traffic Prediction	✓	✓	Y
KOL Match	Ad Requirements & Solution Negotiation	✗	✗	N
	Placement Price Negotiation	✗	✗	N
Ad Customization	Ad Copy Design	✓	✗	N
	Video Production	✓	✗	N
Video Review	Video Audit	✓	✓	N
Content Distribution	Distribution Effect Monitoring	✓	✗	N
	Strategy Adjustment	✓	✓	N

数据来源：美国汽车工程师协会（SAE）；美国国家公路交通安全管理局（NHTSA）。

退一步看，我们对AI Agent的要求不是刷题，而是能够真正落地应用。在具体场景中，完成任务的效果不仅依赖于AI本身的能力，更取决于其与环境、与人的配合。因此，我们需要跳出对“绝对智慧水平“的追逐，从与人类配合的能力这个角度，重新思考AI Agent的分级标准。

在这一点上，一个可直接参考的对象就是自动驾驶的分级体系。在自动驾驶分级中，主要依据“人类责任逐步减轻”原则，按照驾驶员与车辆在不同阶段所承担的责任范围来界定能力边界。这种分级方式既考虑了技术能力的迭代升级，也兼顾了人与机器的协作关系。

智能体能力的五个层级

在AI Agent分级时，同样可以“人与智能体之间的协作边界”为核心，明确各等级下“AI Agent应擅长什么”与“人类不可替代什么”。由于AI Agent的本质是“数字劳动力”，其价值在于替代或扩展人类能力，因此还可以借鉴人类职业成长的路径（被动执行→项目助理→初级项目负责人→专业骨干→领导者），来构建分级框架。

基于以上思路，我们构建了AI Agent能力分级的五个层级：基础响应与流程执行（L1）→流程范围内自主（L2）→全自主决策（L3）→环境驱动与创造（L4）→组织与领导（L5）。

与此对应，智能体实现任务的方式也呈现出演进趋势：知识库问答、工作流、大模型自主规划和多智能体协同。不同的AI Agent类型对应着不同的技术要点。在AI Agent能力发生变化的同时，AI Agent的类型也会不断变化：在L1阶段，智能体以知识库问答和工作流为主；从L2起，智能体能力进入狭义AI Agent的范畴，规划能力、协同能力以及自主使用工具的能力成为关键；当能力达到L5水平后，多智能体协同类AI Agent成为常态，展现出类似“组织与领导”的能力。

现在的绝大多数AI Agent，都还在比较初级的水平，仅能在具体任务中发挥作用，离“独立干活”还有一定距离。

处于L1阶段的AI Agent仅仅是被动的执行者。它依赖人类的指引（各种形式的prompt，或固定好的工作流）来行动，能够在理解意图后给出回应、完成任务，但完全无法判断答案的正确性，也不会思考下一步要做什么。这类Agent虽然能够调用知识和工具完成任务，但是这些能力都是工作流中人预先设定好的流程与标准，本质上还是靠人的决策和执行。这类AI Agent通常负责大工作流中的某个环节，主要价值是把人从重复性劳动中解放出来，例如基础版Chatbot（Deepseek、日常对话场景的元宝、豆包、基础版本的ChatGPT/Gemini等）、图片/视频生成的基础工具、智能客服系统、法大大等法务领域的合同生成与修改建议工具等。

当AI Agent进入L2时，才真正符合狭义上的AI Agent定义。它不再完全依赖工作流完成任务，而是能在既定工具和流程范围内，进行一定的规划，按部就班独立完成任务。L2的Agent在关键的决策与动作执行时，必须由人介入。这就像职场新人：你丢给他一个目标，他能自己列计划、找数据、生成报告，但最终方案是否合理，仍然需要你来拍板。典型例子是OpenAI、Gemini的DeepResearch这类“通用AI Agent”，它们能自主完成全流程，但遇到重要抉择时，还是会拉人一起商量；需要说明的是，L1分类下的产品也可能具备L2级别能力的功能模块，比如在高考填志愿场景下，元宝能够根据高考考生的需求，自主调用高考信息查询、高考院校推荐等工具，为考生筛选出匹配的院校和专业，也是Agent等级达到L2的体现；L2的进步在于不再依赖预设规则，而是像真人一样“见招拆招”。

L3的智能体已经具备“初级项目负责人”的特征。L3和L2最大的区别在于方案规划的步骤不再依靠人类，自主规划、自主收集信息和寻找工具的能力进一步提升；此外，L3级别的Agent还会边干边优化，甚至主动检查工作成果，仅在最关键的环节需要人的决策，以及最终环节靠人类验收。在整个工作过程中，L3更少地依赖人的介入。当前的AI Agent类产品（如Flowith 2.0，MiniMax M1，ChatGPT Agent等）正展现出从L2向L3演进的明显倾向，在执行任务时减少人工介入的频次、增加自我反思与迭代的动作。然而，从整体表现来看，尚无任何产品能在所有任务场景中稳定实现L3级别的核心能力。在需求调整、任务范围拓展、结果优化等环节，这些产品仍需依赖人类明确的指令输入，完全脱离人工指令的自主任务闭环尚未形成。

L4阶段的智能体，则更像一个“能独立发现问题的同事”。它能主动观察环境、发现问题，甚至不用等你派活，自己就能规划要做什么，人类只需在最终环节验收成果。与L3的最大区别在于，L4具备了环境理解能力和自主决策能力，不再依赖人工派活，而是能根据环境变化主动识别工作需求。在这一层级，多智能体协同的特征开始显现。例如，一个虚拟的电商管理AI Agent：它可以自主访问公司文档、数据及会议记录，全面掌握关键信息；基于这些信息自主规划分析任务，定位业务痛点并制定解决方案；随后调用数据分析工具，整合多平台数据，精准识别出具体问题（如“华东区库存告急”），并设计出补货计划或投放策略调整方案。整个过程中，AI Agent能独立完成数据收集、问题诊断和方案设计等核心环节，仅将需要人类决策的关键节点（如大额采购审批）交由人工处理。

到L5，智能体则演化为“团队领导”。它不仅能单干，还能组织其他AI甚至真人一起完成复杂项目。想象一个AI项目经理，它能根据公司目标拆解任务、分配资源、协调不同部门的AI Agent和人类员工，最终带着大家完成一个商业计划。此时的AI已经从“工具”升级成“伙伴”了。此时的AI Agent必须具备与其它AI Agent合作的能力，形态变为了“多智能体协同”。

整体来看，当前AI Agent市场呈现明显的阶梯式发展特征：绝大多数产品仍停留在L1-L2级别，依赖人工指令或预设流程完成辅助性工作，是工作上的“好工具”、“好帮手”；少数被归为L3的产品，实则多为L2到L3的中间态，在自我评估、持续优化的主动性上，尚未严格达到L3的标准。而随着技术在自主决策、环境感知等能力的突破，AI Agent将向更高级别跃迁，未来有望真正实现从“辅助工具”到“数字伙伴”的跨越，在各行业释放更大价值。

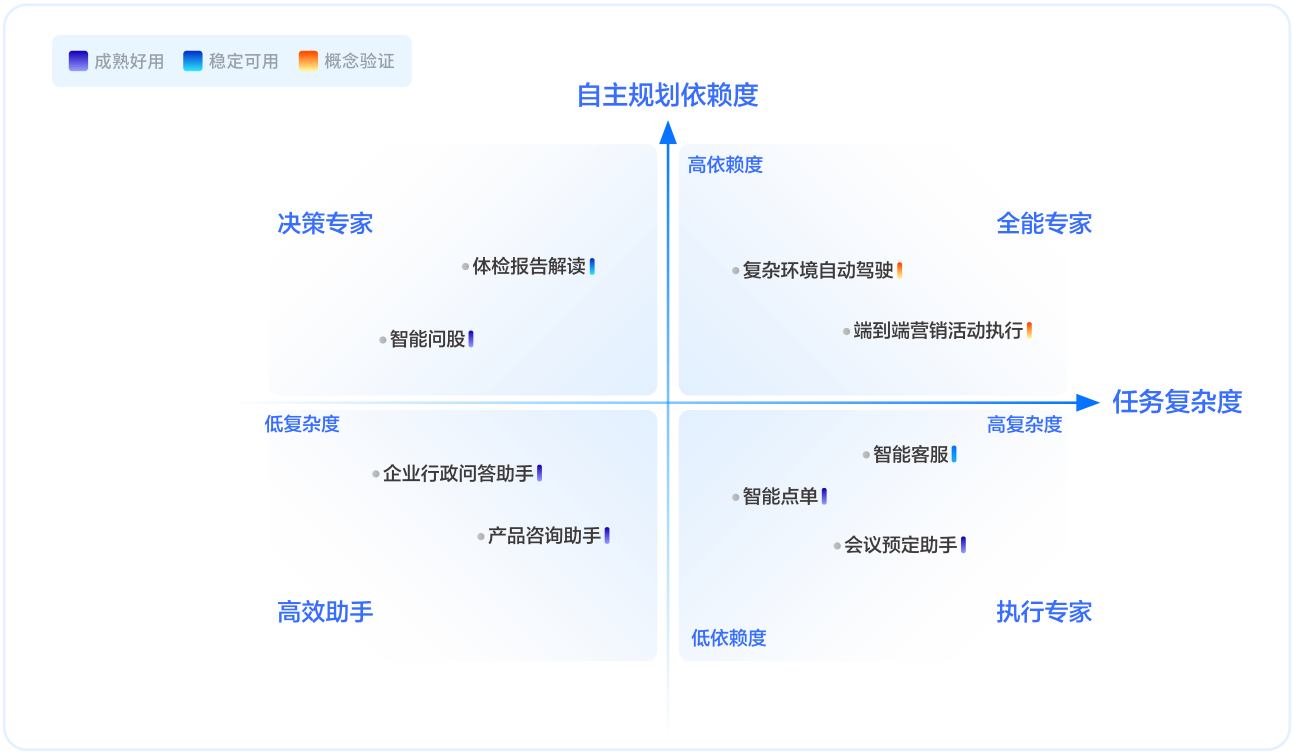
02

智能体场景盘点

在当今智能化转型的浪潮中，从辅助客服、自动化内容创作，再到复杂的决策支持，智能体的应用场景日益丰富，智能体正从一个前沿技术概念，迅速演变为企业提升效率、开辟新业务模式的强大工具。企业迫切希望将智能体应用在业务流程中，期待能够通过应用智能体提升运营效率、降低运营成本并实现业务创新。然而，智能体在企业场景的落地还处于早期，企业没有成熟可借鉴的场景建设成功经验，如何精准识别可以成熟落地、产生实质性商业价值的智能体是企业管理层面临的极大挑战。

智能体场景罗盘

本报告提出了“智能体场景罗盘”，紧密围绕“企业场景”和“智能体特性”这两个核心要素，为企业提供一个可以清晰识别出智能体的落地成熟度的分析框架，帮助企业制定自己的智能体建设规划。本章将详细阐述这一分析框架的构成，并以此为基础，深入剖析智能体在罗盘不同象限中的价值，为企业的智能体建设提供一份切实可行的规划指南。



图：智能场景罗盘
智能体场景罗盘由横纵 2 个坐标轴、4 个场景象限以及分布在象限中的场景构成

罗盘的横轴为场景的“任务复杂度”

暨智能体完成一个场景任务所需执行的步骤、系统和协同的复杂程度。越靠近横轴负方向复杂度越低，这类场景中的任务越简单且独立。它们通常表现为：步骤少，调用少，处理量小，依赖度低。智能体在此主要扮演高效的单点工具，其价值在于提供即时、精准的服务。越靠近横轴正方形复杂度越高，这类场景中的任务通常复杂且需跨系统协同。它们通常表现为：步骤多，调用多，处理量大，依赖度高。智能体在此主要扮演复杂流程的执行专家，其价值在于将端到端的复杂业务流程封装为一键式服务。

罗盘的纵轴为场景对智能体的“自主规划依赖度”

越靠近纵轴负方向，智能体的自主规划依赖度越低，可按照既定规则/流程运行。它们通常表现为：指令清晰，流程稳定，简单对话，通用知识。智能体在此主要扮演忠实的执行者或知识查询者，其价值在于高效、精准地完成既定任务。越靠近纵轴正方向，智能体的自主规划依赖度越高，其决策因素复杂多变、可能性无法穷尽。它们通常表现为：指令模糊，流程多变，复杂对话，专业知识。智能体在此必须具备强大的自主规划、决策与学习能力，其价值在于独立应对和解决开放性问题。

横纵轴构成了“高效助理”、“执行专家”、“决策专家”和“全能专家”4个场景象限，象限中场景的不同颜色代表了不同的场景成熟度，颜色越深代表场景成熟度越高。以下，我们也将选取四大象限中的典型场景，对场景的执行路径、场景价值和落地策略进行解析。

◆高效助手

该象限的场景具任务流程和规则明确，自主决策依赖度低，且执行路径简单。智能体在此主要扮演“高效助手”的角色，其核心价值在于快速响应、知识检索和重复性任务的自动化。

场景示例：企业行政问答助手

场景说明：“企业行政问答助手”是部署于企业内部协同平台（如企业微信）的智能机器人，它扮演着一个24小时在线的、面向全体员工的共享服务中心（SSC）虚拟客服。该智能体的核心任务，是自动应答来自员工关于行政、IT、财务、人事等方面的海量、高频、重复的咨询，例如“如何报销差旅费？”、“IT权限申请的流程是怎样的？”等。它通过即时提供标准答案，将人工客服从繁琐的重复性问询中解放出来。

执行路径：此场景的对智能体的自主决策依赖度低，员工的提问虽然形式多样，但核心意图高度可预测，且所有答案都来自于一个确定的、内部定义的知识库（如公司的差旅报销政策文档）。智能体无需进行复杂的判断和自主决策。此场景的任务复杂度低， 智能体在此执行的任务是一个简单的“查询-响应”操作。它仅需接收用户输入，在内部知识库中进行搜索匹配，然后直接返回标准答案。整个过程操作步骤单一，无需跨系统协同，即可完成任务。

场景价值：通过部署此类问答助手，企业能够极大地降低SSC部门的人工客服成本，同时显著提升员工获取内部信息的效率和体验，实现降本增效的双重目标。

落地策略：现阶段，该象限的应用场景大多数采用工作流和知识问答型智能体。它们无需复杂的自主规划能力，只需遵循预设的规则或在特定知识库中进行搜索。

从企业落地视角来看，该象限的场景是企业智能体应用的首选切入点。其技术门槛相对较低，能够实现快速部署和快速验证，使得企业能够在最小的风险下，迅速实现可见的价值，并为后续更复杂的智能体应用积累宝贵的成功经验和内部信心。

◆执行专家

该象限的场景自主决策依赖度低，但其任务流转复杂且执行路径冗长，通常涉及多个系统和部门的协同。智能体在此扮演“执行专家”的角色，其核心价值在于将复杂的业务流转进行智能串联和整合，把一个长链条任务封装为一键式操作。

场景示例：智能会议预定助手

场景说明：“智能会议预定助手”是一个能够将复杂的会议预定流程自动化的智能体。它能让用户通过一个简单的自然语言指令，如“帮我预定明天下午3点和张三、李四的会议，讨论项目A”，自动完成一系列跨系统、跨部门的操作。它能够代替员工，高效地查询会议室空闲情况、确认参会人日程、发起预定并发送会议邀请及提醒，将繁琐的流程简化为一个无缝衔接的自动化服务。

执行路径：此场景对智能体的自主规划依赖度低：员工的请求（如“帮我预定明天下午3点和张三、李四的会议，讨论项目A”）意图清晰、规则确定。智能体无需进行创造性的判断，只需严格遵循预设的执行逻辑。同时任务复杂度高， 智能体在此扮演一个复杂的流程协同者。它需要进行多步骤的智能编排，依次调用不同的API接口。

场景价值：通过智能体的跨系统协同能力，将一个繁琐、低效的复杂流程，转化为一个简单的自然语言指令，实现了跨越式的效率提升。

落地策略：现阶段，该象限的应用主要聚焦于任务编排与跨系统协同。这里的智能体需要更强大的流程编排能力和多系统API调用能力。

从企业落地视角来看，该象限的价值在于实现跨越式的效率提升。企业需要重点关注智能体的集成和协同能力，通过将过去分散的业务流程整合成无缝的服务，实现效率的质变和业务的优化。

◆决策专家

该象限的场景自主决策依赖度高，需要智能体基于复杂且动态的环境进行深度分析和决策，但决策后的执行路径相对简短。智能体在此主要扮演“智能参谋”或“决策辅助者”的角色，其核心价值在于提供专业级的分析洞察与决策建议。

场景示例：智能问股

场景说明：“智能问股”是一个能够处理海量、动态的金融市场信息，并为用户提供专业级分析洞察和决策建议的智能体。它能够自主筛选并整合公司财报、实时新闻、行业动态、市场情绪等各类非结构化和结构化数据，进行复杂的逻辑推理和趋势预测。通过将繁重的认知工作自动化，它能帮助用户在极短时间内获取高质量、有针对性的信息，从而提高投资决策的质量和效率。

执行路径：此场景对智能体的自主决策依赖度高：用户的需求（如“分析一下某只股票近期上涨的主要驱动力”）是开放式且动态的。智能体需要从海量的、非结构化的数据源中自主筛选信息，并进行复杂的逻辑推理、情感分析和预测。其信

息模糊度高，且决策依赖度高。任务复杂度低，尽管分析过程复杂，但智能体在此阶段的执行路径却非常简短。它只需接收用户指令，在后台完成分析后，生成一份结构化的报告或一份投资建议。整个过程是一个“分析——输出”的单向流程，无需与多个后端系统进行复杂的交互或执行很多步骤操作。

场景价值：通过智能体对高不确定性信息的深度处理，极大地释放了股民和分析师的认知瓶颈，使他们能够专注于更深层次的策略制定和最终决策，从而提升了决策的质量和效率。

落地策略：该象限的智能体通常需要强大的数据分析、逻辑推理和自主决策能力，以应对高不确定性的挑战。

从企业落地视角来看，该象限的价值在于为企业的核心业务提供专业级的智力支持。虽然这类智能体不直接执行业务操作，但其提供的洞察与建议是企业获得竞争优势的关键，是企业从“业务效率提升”迈向“战略决策赋能”的重要标志。

◆ 全能专家

该象限代表了智能体应用的高级形态。场景具有极高的不确定性，且任务执行路径复杂且漫长，涉及多系统、多步骤的协同。智能体在此扮演“全能专家”的角色，其核心价值在于自主规划、自主执行，并对复杂任务进行全生命周期管理。

场景示例：端到端营销活动执行

场景说明：“端到端营销活动执行”是一个能够自主规划并执行复杂营销全流程的智能体。它能够独立处理高度不确定的市场数据、用户行为和社交媒体情绪，自动生成个性化的营销创意和投放策略。该智能体能通过智能编排，自动在不同平台上创建、发布和监控广告，并根据实时效果数据自主调整投放策略，实现从策略制定到效果优化的全流程自动化和自我运行。

执行路径：此场景对智能体的自主决策依赖度高，智能体需要自主分析动态变化的宏观市场趋势、复杂的用户行为数据，甚至捕捉非结构化的社交媒体情绪。这些都属于高度不确定的信息源，需要智能体进行复杂的自主分析、判断与创造性规划。任务复杂度高，智能体在此扮演一个多步协同的执行者角色。它需要进行复杂的智能编排，并依次执行多项任务。

落地策略：该象限的智能体应用通常需要大模型强大的自主规划能力与多智能体协同能力。一个任务可能由一个主智能体进行宏观规划，再由多个子智能体分工协作完成具体执行。

从企业落地视角来看，该象限代表着企业的长期战略目标。它需要企业在技术、数据和组织架构上进行全面的升级与投入。虽然门槛极高，但其价值在于为企业打造全新的竞争壁垒，实现业务模式的质变。

当前处于“高效助理”、“执行专家”、“决策专家”象限的智能体场景成熟度相对较高，处于“全能专家”象限的场景大多还在技术验证期，落地应用较难。企业可以使用智能体场景罗盘，判断智能体场景所处的象限和落地成熟度，从而构建自己的智能体建设规划

智能体百大场景

我们也梳理了智能体在金融、零售、文旅、教育、医学、出行、互联网等14个行业上百个具有代表性的智能体应用场景，致力于帮助各行业企业快速、直观地了解智能体在不同业务领域的落地形态，企业可以从中遴选合适的场景推进智能体应用落地。

03

智能体技术/产品方案解析

1.智能体产业应用技术挑战

智能体正从概念走向实践，然而，任何一项革命性的技术落地都非坦途。与传统IT项目不同，智能体的落地挑战不仅源于技术本身，更在于其与企业既有业务、数据、系统和安全体系的深度耦合。本报告将系统性地剖析智能体在企业应用落地的六大核心挑战，希望帮助企业以清晰的风险意识，开启智能体建设的征程。

训推成本:以架构升级破解经济性难题

智能体应用正将大模型从“对话生成”推向“自主执行”的复杂业务场景，然而，这一跃迁也带来了严峻的成本挑战。一方面，大模型因其庞大的参数量与数据处理需求，对训练和迭代的资源要求极高；另一方面，智能体的工作流包含感知、规划、工具调用、反思等多个复杂步骤，单次任务的Token消耗量远超简单问答，导致推理成本急剧增长。同时业务负载的“潮汐效应”——高峰期需求激增，低谷期资源闲置，与传统的静态算力部署模式形成尖锐矛盾，使得企业陷入“用不起、跑不动”的困境，算力投资仿佛掉入“黑洞”，难以转化为可控的业务效能。

◆挑战

智能体落地的成本困境，本质上是其复杂工作模式与传统算力架构之间的不匹配，我们可以将其拆解为四个层面。在基础设施层，智能体执行复杂任务时，往往需要在多卡、多节点间进行并行计算与频繁通信。传统的网络架构可能成为瓶颈，一旦通信效率跟不上，整体性能便会严重受限，导致计算资源利用率大幅下降，无形中增加了时间与机会成本。在算力调度层，企业内部的算力资源往往呈碎片化分布，缺乏统一、智能的调度机制。这导致GPU等宝贵资源利用率低下，大量算力处于闲置或低效运行状态，成本无法有效摊薄。在服务部署层，传统的模型服务部署框架在应对智能体带来的高并发、长序列请求时，容易出现推理队列积压，导致服务延迟飙升，严重影响用户体验与业务连续性。在模型框架层，大模型本身计算密集、内存消耗巨大，若缺乏针对性的底层框架优化，硬件的潜力就无法被充分释放。这不仅仅是算法问题，更是复杂的系统工程挑战。这四个层面的挑战相互交织，共同构成了智能体应用落地的成本壁垒。单纯依靠“堆硬件”的粗放式投入，不仅成本高昂，更无法从根本上解决效能问题。真正的破局之道在于对整个技术栈进行系统性的重构与优化，将算力从“成本中心”转变为驱动业务增长的“效能引擎”。

◆解决思路

为应对上述挑战，行业正积极探索从IaaS到PaaS的全栈协同优化方案，通过系统级的技术整合，实现端到端的降本增效。

构建AI原生的弹性基础设施：

通信优化：采用专为AI负载设计的网络通信技术（如RDMA的升级与优化），解决跨节点通信瓶颈，确保大规模集群的近无损扩展，让数据在计算节点间高效流转。

智能调度：引入全局智能调度平台，实现对多云、多地域、多异构算力的统一编排。通过“潮汐调度”等模式，在推理任务低峰期将闲置算力自动分配给训练或精调任务，实现算力资源的动态“削峰填谷”，将整体利用率最大化。

打造高效能的模型服务与推理框架：

先进服务架构：采用“Prefill/Decode分离”（PD分离）等先进部署架构。针对任务的不同阶段（如长文本理解的Prefill阶段和逐字生成的Decode阶段）采用不同的并行策略（如张量并行、专家并行、数据并行等），最大化利用计算与显存资源，在不影响精度的情况下大幅提升吞吐率，降低单位请求成本。

模型深度优化：在模型框架层进行深度优化。这包括但不限于：使用模型量化（如int4/int8）技术压缩模型体积；重写关算子以适配最新硬件特性；应用多令牌预测（Multi-Token Prediction, MTP）等技术，通过一次前向计算预测多个Token，显著提升生成速度。这些优化技术的组合，能将硬件潜力压榨到极致。

通过上述体系化的架构升级，旨在将“堆砌算力”的粗放模式，转变为“精耕细作”的效能模式，让每一份算力投入都产生最大价值，最终使智能体应用真正“跑得动、用得起”，从昂贵的“奢侈品”变为企业数字化转型中不可或缺的“生产力工具”。

模型性能：幻觉与泛化性的双重困境

智能体在应用中面临的性能挑战，核心在于通用大模型的知识局限性和生成机制的固有缺陷。一方面，通用大模型虽然在海量公开数据上表现出色，但其在垂直领域的专业知识存在缺失，导致泛化能力不足。另一方面，大模型基于概率生成内容，存在固有的“幻觉”问题，即生成看似合理但与事实不符或缺乏依据的信息。在需要自主执行任务的智能体场景中，幻觉可能导致错误的决策和危险的行为，影响系统可靠性。

◆挑战

通用大模型的知识是“静态”且“宏观”的。当面对特定企业的内部流程、专业术语、最新业务数据或实时信息时，其知识存在“鸿沟”。例如，一个智能体需要理解最新的内部风控政策，或根据最新的临床指南辅助决策，但通用模型原生并不具备这些知识。当其面对不熟悉的问题时，会倾向于“一本正经地胡说八道”，产生幻觉。更深层次的问题在于，幻觉在智能体场景中会产生“放大效应”。在简单的问答场景中，幻觉可能只是生成错误信息，用户尚能自行判断。但在自主执行的智能体中，幻觉可能导致智能体调用了错误的 API、执行了不当的操作，甚至引发安全事故。

◆解决思路

为应对模型幻觉与泛化性的双重困境，业界普遍采用一套多层次、相辅相成的技术组合，而非单一的解决方案，以系统性地将通用大模型锻造为精准、可靠且安全的专业智能体。

在应用层，通过检索增强生成（RAG）为模型外挂一个动态更新的“事实大脑”。该技术通过构建“检索-生成”的两步模式，在处理任务时先从企业内部数据库、产品文档等可信知识源中精准定位相关信息，再将其作为权威上下文注入提示词，引导模型进行有据可依的回答。这不仅是解决知识性幻觉、确保信息时效性的最直接手段，也相当于为模型提供了“开卷考试”的条件，从根本上降低了其“凭空捏造”的风险。

在模型层，通过微调（尤其是参数高效微调PEFT方法，如LoRA）技术，使用企业高质量的私有数据对模型进行二次训练，我们能够将其内部参数“校准”到特定行业的语境和业务逻辑上，提升模型在专业领域的泛化能力，使其不仅能理解行业术语，更能掌握独特的任务流程、沟通风格与决策模式，使其行为更贴合企业需求的核心环节。RAG与微调并非互斥，而是常常协同使用，前者提供事实，后者优化处理事实的方式。

在对齐层，通过人类反馈强化学习（RLHF）为模型植入一套安全、可靠的行为准则。RLHF通过引入人类评估者对模型的输出进行偏好排序，训练一个“奖励模型”来模拟人类的价值观和安全标准，再以此为指引，持续优化模型，使其生成的内容不仅正确、专业，更要确保其有益、无害、负责任。这是解决更深层次的、涉及偏见与安全风险的“幻觉”问题的关键手段，是确保智能体成为一个值得信赖的数字化合作伙伴的重要防线。

安全防护：从模型到基础设施的立体防御

智能体应用面临的安全挑战是系统性的，涵盖了从模型、应用到基础设施的各个层面。首先，智能体应用本身引入了传统网络安全无法有效应对的新型威胁，如提示词注入、敏感信息泄露和不当的 API 调用。其次，智能体在处理和存储海量敏感数据时，面临数据滥用、隐私泄露和法规不合规的风险。最后，模型和数据部署在云环境或边缘设备上，面临供应链漏洞、运行时攻击和物理入侵等基础设施安全威胁。

◆挑战

智能体的安全风险贯穿其整个生命周期，而非局限于应用层面。在数据准备阶段，攻击者可通过投毒样本污染训练数据，改变模型行为。在部署推理阶段，恶意用户可通过构造恶意提示词来绕过安全防护，获取模型内部信息或导致敏感信息泄露，即所谓的“越狱”攻击。此外，攻击者还可能通过探测模型参数信息，实施模型窃取，侵犯知识产权。除了模型本身的风险，智能体作为应用系统，其 API 接口和业务逻辑也成为新的攻击面。例如，攻击者可以通过构造恶意 API 请求，绕过安全策略，进行非授权访问或数据滥用。尽管每个应用都会各自建立安全能力，但这种零散的防护体系可能存在“千里之堤，溃于蚁穴”的风险。传统的网络安全工具（如 Web 应用防火墙 WAF）难以检测和防范这些针对 AI 模型的复杂攻击。这些攻击利用的是模型的语言逻辑漏洞，而非传统的网络协议或代码漏洞。例如，WAF 虽然可以限制 API 请求速率以防范模型抓取，但无法识别恶意提示词注入。因此，智能体安全防护的本质挑战是“非线性”的，解决方案必须从单一的技术防护转向覆盖全生命周期、多维度的治理与技术协同。

◆解决思路

为构建智能体安全防线，必须建立一个集基础设施安全、模型安全、数据安全和应用安全于一体的多层级的纵深防御体系。

在基础设施安全方面，应采用零信任架构和微隔离技术。通过严格的身份验证和访问控制，确保只有可信的人员和应用能够访问核心算力与数据，同时通过微隔离抑制“东西向”横向渗透，将潜在攻击的危害范围限制在最小。

在模型安全方面，建立常态化的对抗性测试和红队演练机制，模拟越狱、投毒、模型窃取等攻击，提前发现和修复漏洞。同时，部署运行时入侵防范系统，实时监控智能体调用行为，对异常 API 请求进行识别与阻断。

在数据安全方面，从源头进行数据净化和脱敏处理。实施全面的数据丢失防护（DLP）策略，对智能体工作流中的敏感数据进行实时扫描、分类和过滤，防止敏感信息在模型输出中意外泄露。此外，还需通过加密和严格的访问策略，保护模型和数据集等核心资产，防止其被盗窃或篡改。

在智能体应用安全方面，需要针对应用层面的特有风险进行防护。这包括通过 API 安全审计和运行时入侵防范系统，实时监控智能体调用行为。例如，可以设置安全策略，对高频访问或跨系统接口调用等异常行为进行识别与阻断，杜绝第三方滥用 API 导致的数据泄露。同时，通过部署虚拟补丁方案，可以在不中断业务的情况下快速修复针对大模型的复杂攻击。

数据治理：打破数据孤岛 构建可信知识底座

智能体要发挥价值，必须依赖高质量、可信赖的数据。然而，企业内部数据普遍存在两大核心挑战：首先，数据质量参差不齐，来源庞杂，普遍存在重复数据、格式不统一、信息缺失和逻辑冲突等问题。其次，不同业务部门的数据通常存储在各自独立的系统中，形成难以打破的“数据孤岛”。因此，由于业务指标和数据定义缺乏统一标准，导致智能体在面对同一问题时，因数据口径不一而给出相互矛盾的答案，陷入“有数无洞察”的困境。

挑战

“有数无洞察”是智能体应用中普遍存在的业务痛点。当一个智能体需要整合不同部门的数据来生成一份综合报告时，如果财务数据和销售数据对“新客户”的定义不一致，它将无法给出可信的、一致的分析结果。这暴露了底层数据治理的根本性缺陷：技术上的“数据孤岛”与业务上的“语义鸿沟”是核心矛盾。单纯的数据清洗只是解决了物理层面的问题，而未能解决对数据“认知不统一”的深层挑战。数据清洗本身也是一个复杂且高成本的过程，不仅仅是技术问题，更需要业务知识和经验的介入。例如，对于重复数据，需要根据主键和业务含义来判断是否真的重复；对于缺失值，需要根据业务场景和保真度要求，判断是采用统计填充、机器学习预测还是人工补全。这使得数据清洗难以自动化，成为智能体获取高质量知识的巨大障碍。

解决思路

解决数据治理问题，需要自下而上地构建一个完整的、可信的知识底座。

首先，建设企业级数据治理中心。该平台应集数据接入、清洗、转换、质量监控与元数据管理于一体。它应具备自动化数据去重、格式统一和信息补全能力，为智能体提供高质量的、可信赖的原始数据。

其次，在数据治理中心之上，构建统一语义层与指标平台。该语义层将复杂的底层数据源抽象为业务人员易于理解的业务概念和指标，如“客户订单”、“用户活跃度”等。所有业务指标都在指标平台上进行统一管理，确保所有智能体和应用在调用数据时都遵循同一口径，彻底解决“有数无洞察”的问题。这种从“数据”到“信息”再到“知识”的治理进阶路径，是智能体实现商业价值的基石。

知识解析：高效检索 告别“一本正经地胡说八道”

为真正解决大模型“一本正经地胡说八道”的问题，必须构建一个覆盖知识解析、检索与理解的全链路智能体系，而非仅仅停留在文本检索层面。这一体系面临三大核心挑战：前端的知识源解析不准，即如何处理图文混排、版式复杂的多模态文档；中端的知识检索不全，即如何应对用户口语化的模糊输入，并在海量、异构的知识库中实现精准召回；以及后端的知识理解不深，即如何满足严谨场景下对复杂推理、多步问答及无关信息判断的高要求。这三大挑战环环相扣，共同决定了最终问答体验的可靠性与精准度。

◆挑战

这些挑战在实际应用中构成了严峻的技术壁垒，其深度远超表面。在解析层面，企业的核心知识往往沉淀在扫描版的PDF、图文并茂的报告或截图等非结构化载体中。传统解析工具在面对这些复杂版式时，常将表格拆解成无序文本，或将图表下的关键注释与正文混淆，导致知识的结构化信息丢失，从源头就注入了“残缺”的知识。在检索层面，矛盾尤为突出。用户“上季度华东区A产品的销售额怎么样？”的口语化提问，可能需要系统同时理解“上季度”的时间范围，定位到“华东区”的地域数据，并从一个上万行、数百列的 SQL 数据库中精准查询。传统 RAG 采用的固定长度文本切片（Chunking）机制，不仅会割裂上下文，更无法与结构化数据库进行有效交互，从而导致检索结果要么遗漏关键信息，要么返回大量不相关的文本片段。在理解层面，挑战在于深度推理。例如，回答“负责‘凤凰项目’且后续调往欧洲部门的项目经理，他在新岗位的第一个任务是什么？”这类“多跳问题”（Multi-hop QA），需要系统先找到项目经理（第一跳），再查询其调动记录（第二跳），最后找到其新任务（第三跳）。这种逻辑链条的追踪能力，是简单的文本相似度匹配完全无法企及的，模型若缺乏对知识间关联关系的深度理解，便只能给出臆测的、不可信的答案。

◆解决思路

为系统性地攻克上述难题，行业前沿的解决思路正推动 RAG 架构从固定的流程向具备自主规划与调用能力的 Agentic RAG 框架演进，形成了一套更智能、更精细的解决方案。首先，为攻克解析难题，业界普遍采用先进的 OCR 大模型。这类模型不仅能提取文字，更能理解文档的版面布局，精准还原表格、标题和段落的结构化关系，确保知识在数字化之初就保持高保真度和可用性，为后续所有环节奠定坚实基础。其次，为化解检索困境，Agentic RAG 会智能地分析用户意图：当识别到需要查询结构化数据时，它会自动调用 Text2SQL 模块，将自然语言转化为精确的 SQL 查询语句；面对非结构化文档，则会启动由向量检索、关键词检索、摘要检索等构成的混合检索引擎，实现广度与深度的平衡。更关键的是，所有初步召回的结果都会经过一个 Reranker（重排序）模型的二次精筛，确保最终喂给大模型的是最核心、最相关的上下文。最后，为实现深度理解与推理，GraphRAG 技术正成为关键突破口。通过将分散的知识点构建成知识图谱，模型得以洞察实体间的深层关联，从而从容应对“多跳问题”等复杂推理场景。同时，为模型注入无关知识拒答和模糊问题主动澄清的高级交互能力，正使其从一个被动的“答案生成器”升级为一个能思考、会提问、负责任的“智能知识伙伴”，彻底告别“一本正经地胡说八道”。

业务流程耦合：从“助手”到“执行者”的集成路径

智能体的最终价值在于深度融入企业核心业务流程，从“聪明的助手”升级为能推动业务运转的“可靠执行者”。实现这一跨越的关键瓶颈在于业务流程的深度耦合，这不仅涉及技术层面如何将智能体无缝嵌入企业现有复杂且异构的 IT 系统，更涉及流程层面如何科学地界定人机权责边界，设计出高效协同的新型工作流。

◆挑战

新旧系统间的技术耦合是基础性障碍 。现代企业的 IT 环境是一个由新旧系统交织而成的复杂网络，包含了 OA、ERP、CRM 等传统系统及大量的自研系统，这些系统形成了难以逾越的“数据孤岛”和“流程断点”。智能体要实现端到端的任务执行，就必须在这一复杂环境中穿梭，但若无法与现有系统高效、安全地交互，它就会沦为一个“外挂”工具，而非“原生”成员，当智能体无法深度嵌入员工日常依赖的企业协作平台时，其价值将大打折扣 。与此同时，人机协作的流程耦合是决定应用成败的另一关键 。挑战在于设计一个既能发挥智能体效率，又能保障人类关键决策与监督的混合工作流 。例如，智能体可以起草合同，但最终决策必须由人完成 。如果人机权责边界、审批节点与交接流程设计不当，不仅无法提升效率，反而可能因新的混乱与风险导致项目失败。

◆解决思路

为实现智能体与业务流程的深度耦合，需要从技术集成和流程设计两个层面协同推进。

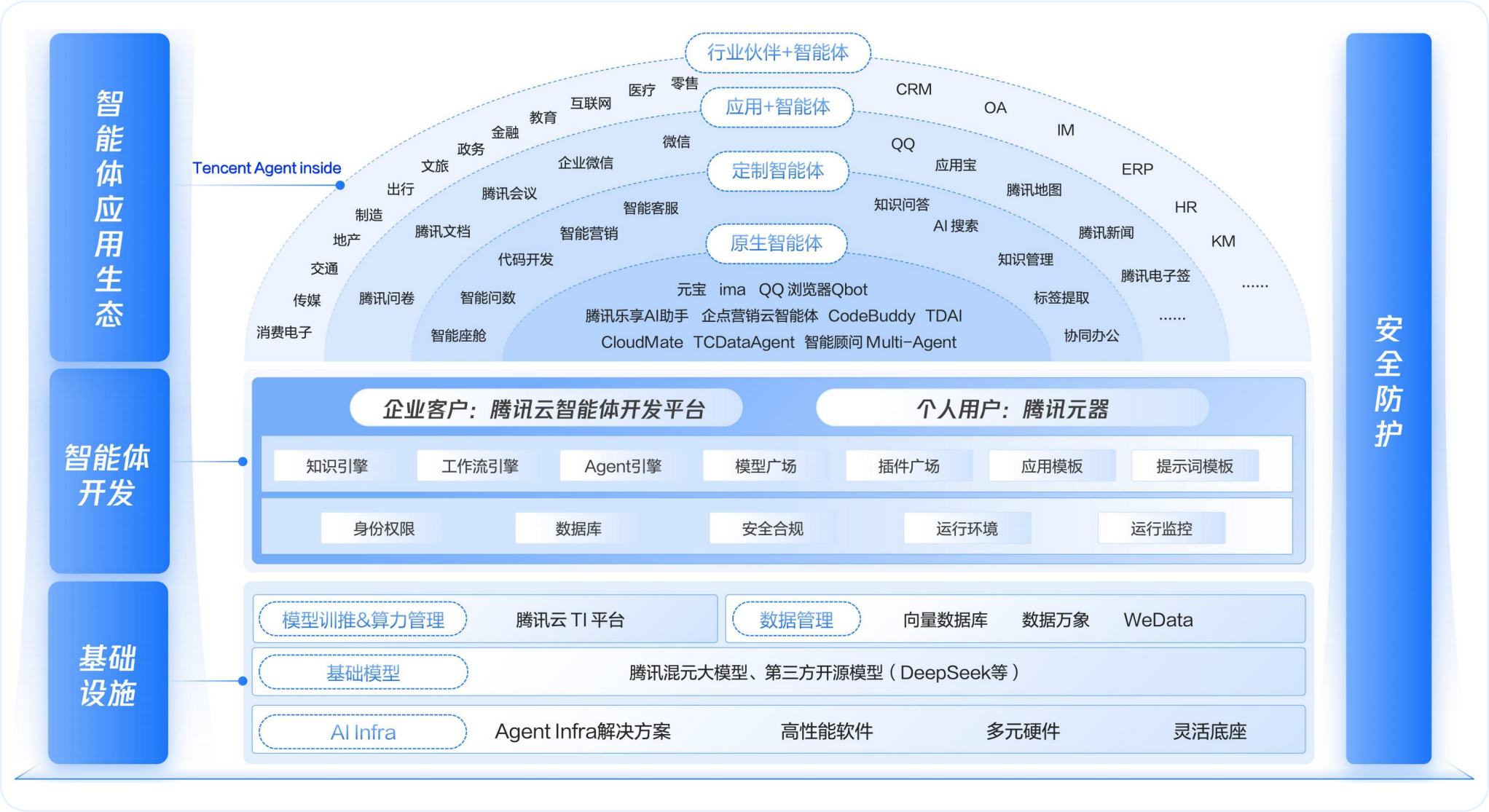
在技术集成方面，以开放接口集成为基础，实现技术层面的无缝嵌入 。解决新旧系统耦合问题的核心，在于构建标准化的开放接口集成模式。智能体不应作为一个孤立的应用存在，而是应该通过标准化接口，与企业协作平台和核心业务系统深度耦合。这使得智能体能够实时感知工作上下文(如群聊中的任务指令),并直接在当前平台调用其他业务系统的能力(如创建日历、发起审批)，最终将执行结果反馈至当前对话窗口。这种“感知 – 决策 – 执行 – 反馈”的闭环能力，能够彻底打通系统间的壁垒，实现从“对话”到“执行”的无缝衔接。

在流程设计层面，以工作流编排为核心，实现人机协作的清晰界定 。针对人机协作流程耦合的挑战，关键在于引入可视化的工作流编排平台 。通过该平台，业务专家和 IT 人员可以将一个复杂的端到端任务，拆解为一系列标准化的子任务节点 。其中，部分节点可配置为由智能体自动执行的系统调用(如“查询订单状态”),而关键节点则可设定为“人工审批”环节。这种方式将抽象的“人机协同”理念，转化为清晰、可控、可追溯的业务流程图。它不仅明确了智能体与人的权责边界，还使得整个协作过程透明化、规范化，从而在确保业务安全合规的前提下，最大限度地释放智能体的自动化价值。

技术与流程的耦合并非孤立的两条线，而是相辅相成的双螺旋。强大的技术集成能力为灵活的流程设计提供了基础，而科学的流程设计又为技术能力的释放规划了安全的航道。二者共同构成了智能体融入业务的核心路径。

2.腾讯云智能体战略全景图

我们系统梳理并全面开放从模型、平台到产品的智能体构建路径，帮助企业构建「懂客户、会决策、能执行、高可靠」的智能体



3.腾讯云产品方案

智能体应用

腾讯企点营销云

◆智能营销的特征与本质

从“千人一面”到“一人一面”的进化革命

在AI与大数据深度融合的2025年，营销运营已不再仅仅是“建个社群、发发优惠券、做做客服”的初级动作。真正的智能营销，正在经历一场由技术驱动、以用户为中心的深刻变革。这场变革不仅体现在工具和流程的升级上，更体现在营销思维与价值逻辑的根本跃迁。

其核心体现为四大特征——敏捷、自动、精准、个性，这四大特征并非孤立存在，而是相互支撑、层层递进，共同构成了AI时代智能营销的“能力四维+体验顶点”，推动企业从“被动响应”走向“主动创造”，从“广撒网”走向“精耕细作”。

◆腾讯企点的MAGIC数智运营增长方法论

在AI技术深度融入营销场景的2025年，企业营销正从“数字化执行”迈向“数智化增长”的关键跃迁。传统的营销方法论已难以应对用户需求碎片化、消费路径非线性、内容偏好动态变化等新挑战。

基于此，腾讯企点提出MAGIC智能营销增长方法论——以AI为驱动、数据为燃料、用户为中心的全链路智能营销新范式。

MAGIC方法论贯穿用户生命周期的每一个触点，通过五个环环相扣的智能化环节，实现从“千人一面”到“一人一面”的精准营销升级：

M－发掘需求（Mine）

结合全域数据，“发掘”用户即时的真实需求。

依托CDP与AI数据分析引擎，整合公私域行为数据、社交互动、内容偏好及企业知识库，构建动态更新的360°用户画像。AI模型实时“挖掘”用户显性与隐性需求，识别购买意图、兴趣迁移与情绪波动，实现对用户状态的敏捷洞察与预测。

A – 编排旅程（Architect）

“编排”用户旅程，定制商品/服务、权益、内容、渠道的匹配策略。

基于需求洞察，AI自动“编排”个性化的用户旅程。系统智能匹配商品组合、权益激励、内容素材与触达渠道，并动态规划最佳执行节点。无论是新客转化、复购唤醒还是品牌种草，均可实现策略自动化、路径最优化。

G – 生成内容（Generate）

“生成”个性化、多场景、多触点、多模态的沟通内容。

借助AIGC技术，根据用户画像、场景节点和沟通偏好，“生成”多模态、高适配的个性化内容。无论是社群话术、朋友圈文案、短视频脚本，还是专属海报与推荐语，AI均可实现秒级批量生成，兼顾创意质量与规模化效率。

I – 互动触达（Interact）

“互动”触达、实时对话、陪伴式运营。

通过智能客服、社群机器人、企微助手等多触点、多智能体协同，在营销关键触点实现“互动式触达”。AI支持多轮对话、情感识别与上下文理解，提供7x24小时陪伴式服务，在提升用户体验的同时，高效推动转化与关系深化。

C – 核查复盘（Check）

“核查”数据与用户评论，并输出复盘报告。

每一次营销活动结束后，AI自动“核查”核心指标、用户反馈与内容表现，结合归因分析与NLP情感判断，生成可视化复盘报告。不仅评估效果，更提炼策略优化建议，形成“执行-反馈-进化”的闭环增长飞轮。

MAGIC方法论的本质，是将AI深度嵌入营销全链路，让运营更敏捷、触达更精准、服务更个性、决策更智能。它不仅是工具升级，更是企业私域运营思维的重构——从“流量运营”转向“用户价值运营”，从“经验驱动”迈向“AI驱动”。在数智化增长的新纪元，MAGIC正成为企业实现魔力增长的核心引擎。

先“发掘”需求，再“制定”策略，接着“生内容”，实时“互动”，最后“核查”复盘——MAGIC让营销魔力增长。

腾讯企点的“MAGIC”智能营销增长解决方案，不仅仅是数据治理和产品的升级，还是智力、知识的升级，将腾讯企点多年服务各行业企业的行业方法论、全域增长方法论，通过AI技术，提炼为共享的行业大模型、知识库、智能体、专家模型等，实现全链路智能营销，帮助更多企业实现数智化驱动的高效增长。

◆腾讯企点营销云产品方案

腾讯企点数智化升级：面对营销链条长、角色多、策略复杂的问题，腾讯企点推出新一代产品形态——“营销云智能体”。

- **基于混元+DeepSeek模型双引擎：**提供强大推理与生成能力。
- **营销知识RAG技术+营销服务MCP技术+CustomerAI：**深度融合业务数据、营销技术产品、营销知识库及场景化预测及决策能力，构建企业专属智能体体系。
- **全链路子Agent协作：**基于multiagent架构，可调用包括人群圈选Agent、旅程编排Agent、内容生产Agent、企微互动Agent、效果分析Agent等多个智能角色。

帮助企业实现AI时代的数智化增长。

智能运维专家

◆挑战&痛点

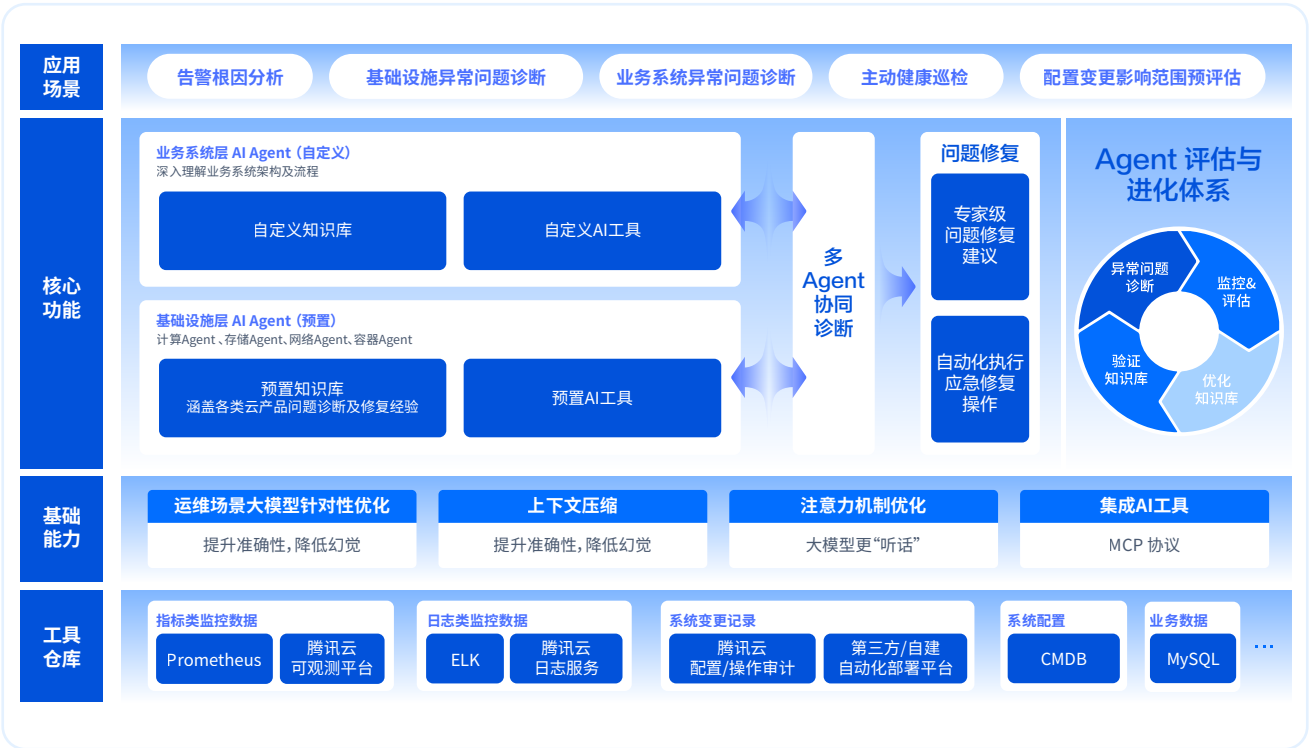
云计算时代下，企业业务系统运行在多云服务中、并多使用微服务架构，虽提升弹性与扩展性，却显著增加运维复杂性。核心挑战如跨层诊断困难（异常跨越计算/存储/网络/应用层）、海量数据处理难（指标/日志/Trace/变更记录繁杂）、多维度异常源（基础设施至代码/数据库/配置错误均可引发连锁故障）及效率与稳定性压力（需快速修复保障业务连续性）。这就要求运维工具需具备跨领域的全链路诊断能力，智能化推理和分析，且能无缝融入现有运维体系。

传统的 AIOps 工具可解决上述部分问题，但往往存在诸多不足，如强规则依赖（预设规则难解跨多服务及层级故障）、智能缺陷（缺乏动态推理能力需人工干预）、 适应性差（难匹配个性化业务架构）。

因此，用户亟需具备自主推理、跨域联动、多工具协同且无缝集成的新型智能运维方案应对挑战。

◆产品方案

腾讯云推出的智能运维专家，以大模型为核心，构建多Agent协同的智能诊断体系，实现从基础云服务到业务系统层的全链路异常分析。



核心能力：

- **智能（Smart）：** 基于大模型，动态规划排障路径，智能推测根因。
- **专家（Expert）：** 沉淀腾讯云多年运维经验，提供专家级分析与修复建议。
- **高效（Efficient）：** 多 Agent 协作，快速定位并提出解决方案，大幅提升效率。
- **系统守护者（System-Keeper）：** 7×24 小时监控与诊断，保障业务稳定。

技术架构：

垂直领域Agent 分别作为各自单一领域专家，解决特定运维问题，包含三大组件：

- **知识库：** 整合云服务与行业经验，支持业务定制。
- **诊断工具 ：** 查询监控指标/运行状态/变更记录，支持集成第三方工具（如MCP）。
- **大模型诊断框架 ：** 基于ReAct机制，结合推理加速、注意力机制等减少模型幻觉，提升诊断准确率。

多Agent协作机制 实现跨领域全链路诊断：

- **协同诊断 ：** 自动调用关联领域 Agent 进行跨层级、跨服务联合排查。
- **上下文共享 ：** 多Agent共享监控数据、中间结论与知识库，避免重复查询。
- **动态规划：** 基于ReAct机制实时调整诊断流程，智能分配分析任务。
- **最终汇聚：** 聚合各Agent结论，输出完整故障根因链路与修复建议，形成专家级的最终诊断报告。

◆ 场景&价值

典型应用场景：

- **基础设施层排障：** 使用计算/存储/网络 Agent诊断基础设施层异常问题，例如安全组配置异常导致网络故障等。
- **业务系统诊断：** 支持用户定制业务知识库与 Agent，如电商支付、库存、订单等模块的异常诊断。
- **复杂链路异常分析：** 结合监控、日志、变更数据，自动构建排障流程并定位根因。

价值体现：

- **快速故障定位：** 缩短诊断时间，避免长时间业务中断。
- **提升运维效率：** 减少人工排障压力，释放人力成本。
- **降低风险与损失：** 预防连锁反应，保障核心业务稳定运行。
- **无缝接入现有体系：** 兼容企业既有工具与平台，降低改造成本。

TDAI Agent Service

◆挑战&痛点

当前，企业在数字化转型过程中普遍面临技术运维、数据分析和系统治理三大瓶颈。

- **技术运维：**风险SQL隐蔽性强，传统手段难以提前识别；开发与DBA协作不畅，优化依赖上线后反馈，成本高响应慢。
- **数据分析：**传统方式依赖个人经验，难以实时捕捉数据变化，静态查询模板导致业务洞察滞后。
- **系统治理：**流程被动低效，缺乏全链路预防机制，且缺少智能工具支持，难以实现自动风险识别与优化。

◆产品方案

核心架构与定位

TDAI Agent Service覆盖数据库DevOps与数据洞察两大体系的智能体，依托自研数据库垂类大模型（DB LLM）、全域上下文（Context）及工具集（Tools）三大基础设施，系统推进AI技术在数据库治理与数据价值挖掘中的深度集成与应用。

智能体分类与功能体系

数据库 DevOps

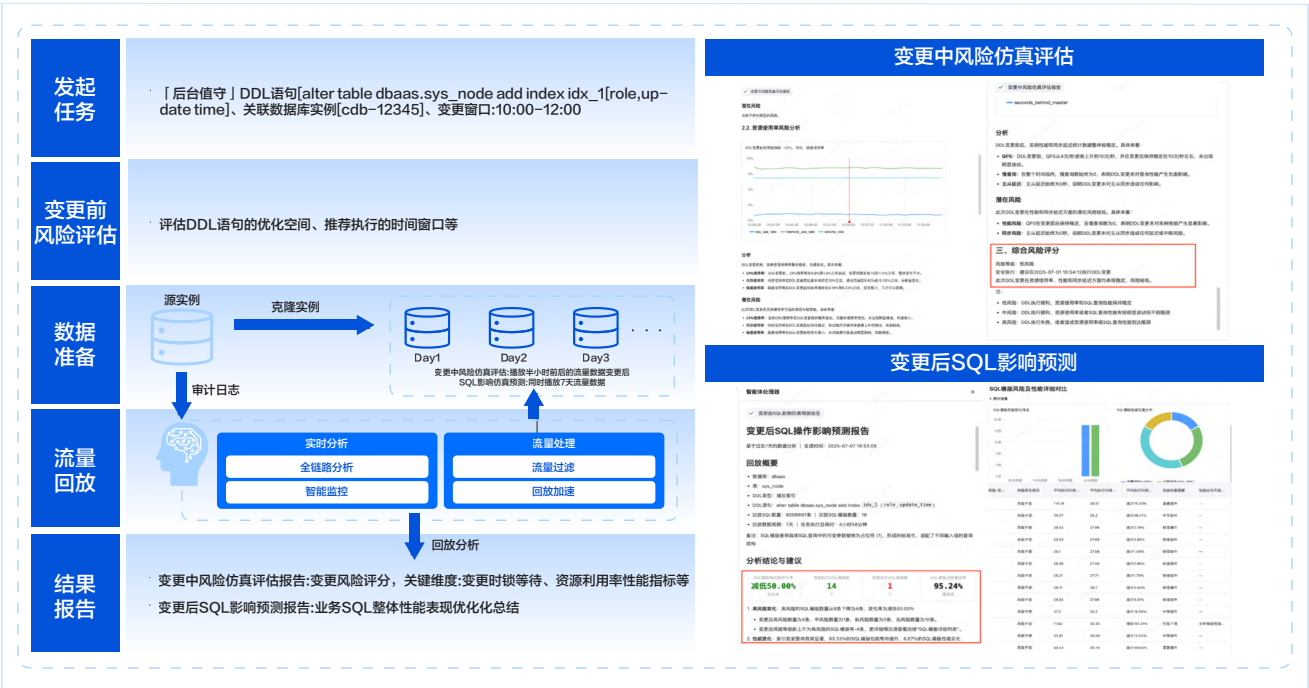
在数据库DevOps方面，致力于构建“AI驱动的智能治理”新范式，通过三大智能体实现闭环治理。

SQL事前风险预测智能体

在部署前通过静态扫描和ORM行为建模，结合C1模型进行SQL语义还原，实时识别全表扫描、索引失效等风险，提供行级优化建议，从源头拦截问题SQL。

DDL变更风险评估智能体

借助流量回放与沙箱仿真，克隆生产环境并注入历史日志，评估锁竞争、主从延迟等隐患，支持以7天流量基准仿真验证性能变化，24小时内完成风险评估。



高负载止损值守智能体

提供24×7实时监控与自动告警，在CPU或业务异常时触发动态干预，基于TOP SQL实现自动限流/Kill、根因分析及优化建议推送，达成RTO＜120秒的主动防控。

数据洞察

在数据洞察方向，系统构建了从“人找数据”到“数据找人”的主动赋能体系。

- **资源规划智能体：**集成数据库状态与业务特征，实现智能资源调度与闲置回收，支持13周多地域预测和自动水位管理，并通过多维度看板辅助故障定位与效率优化。
- **业务洞察与报表分析智能体（规划中）：**拟进一步拓展数据自动挖掘与洞察生成，完善“治理-运营-洞察”一体化技术闭环。

◆场景&价值

TDAI Agent Service聚焦两大核心场景，为企业提供智能化数据解决方案：

场景一：风险SQL治理

针对SQL缺陷引发的性能故障、变更失误及运维效率低下等问题，提供全生命周期智能治理方案。通过SQL事前风险预测、DDL变更智能评估及自动高负载止损等功能，有效提升SQL质量管控水平，大幅降低生产故障率，保障数据库稳定运行和业务连续性。

场景二：企业数据洞察

针对企业数据分析中感知滞后与表面化的痛点，通过自主业务理解、自主动态感知与主动决策推送等核心能力，实现从海量数据到业务决策的端到端感知-分析-推送，从“人找数据”升级到“数据找人”。帮助企业把握商机、预防风险，提升决策与业务转化效率。

数据分析智能体

◆挑战&痛点

随着GenAI技术的不断发展，我们相信AI Agents很快将成为企业劳动力的重要组成部分，释放员工的宝贵时间，使其专注于业务面临的更高价值的挑战。Data Agent是一种专门的AI Agent类别，通过结合企业数据和AI Agent工具来主动提供更准确、更可靠的数据洞察，辅助企业决策。

尽管模型的质量在不断提高，推理成本不断降低，我们发现企业在部署可信赖的AI Agent系统方面面临着以下共同挑战：

- **准确性：**在企业应用中，对AI Agent输出的质量要求很高；在财务等关键业务功能中，错误的容错率很低。

- **受控的数据访问：**AI Agent需要能够访问各种各样的数据源，以便其能够在业务背景下可靠地运行，这些数据源包括非结构化（例如文本、音频）和结构化（例如表、视图）数据源，它们通常分布在多个系统中。

◆产品方案

为了解决这一问题，我们在今年推出腾讯云数据分析智能体服务（TCDataAgent），旨在为企业提供一个全托管的智能体服务，用于整合、检索和分析结构化&非结构化数据，帮助用户更直观的理解数据，并提取有价值的洞察，同时企业用户也可以方便的基于腾讯云数据分析智能体（TCDataAgent）构建高质量的AI Agent应用。

TCDataAgent可以正确理解用户意图，主动规划任务、使用工具来执行任务，并通过反思结果来改进响应。在执行任务时，TCDataAgent 会使用NL2SQL、NL2Py、AI Search、XPark等原子能力，同时结合大语言模型（LLMs），进行分析并生成答案。同时，TCDataAgent兼容标准MCP、A2A等协议，可以方便的被集成到第三方AI应用。

腾讯云大数据在数据分析领域经过多年沉淀，积累了一系列可供TCDataAgent调用的高质量数据分析工具，例如：结构化数据处理（TCAnalyst，支持自然语言交互），非结构化数据处理（AI Search、Document AI），高性能计算引擎（Meson），统一分布式计算框架（XPark）等。在这样的基础上，搭配统一的元数据服务TCCatalog和统一湖存储系统TCLake，腾讯云TCDataAgent可以为用户提供高质量、高性能的数据分析智能体服务。其核心优势如下：

1.结构化数据处理：

与仅依赖模式匹配的典型text-to-SQL系统不同，TCAnalyst使用语义模型将业务术语映射到底层数据。这种方式在涉及多表关联的复杂业务场景中，有效提高了NL2SQL的准确率。

2.非结构化数据处理：

ES是原生的混合检索服务，通过关键词搜索和向量搜索，能够为非结构化数据（例如：文本、音频、图片等）提供大规模、高质量，低延迟的数据检索服务。结合“智能搜索开发”提供的embedding，rerank等原子服务，能够轻松搭建RAG框架，支持创建智能问答应用，同时我们也提供智能文档处理（Document AI）的功能，帮助用户快速提取文档中结构化数据及文本内容。

3.高性能计算：

Meson是腾讯云自研的融合、开放、智能的新一代高性能计算引擎，覆盖批处理、交互式分析、机器学习等多种场景，为TCDataAgent提供底层高性能计算保障。

4.统一分布式计算框架：

XPark是腾讯云自研的高性能分布式计算框架，兼容Python生态中常用DataFrame+ML接口，提供一体化的数据分析、数据预处理、模型训练和模型推理能力。与TCDataAgent集成，通过自然语言进行数据分析、预测和辅助决策。

◆场景&价值

典型用例一：供应链预测

某东南亚公司使用 AI 模型进行需求预测与销量预测，优化库存与供应链管理。但AI算法编写、调优复杂，模型周期长，预测效果不佳，且销售预测与需求预测数据协同性差，导致库存积压或短缺频发。

TCDataAgent 协助用户实现AutoML 流程，缩短预测周期，并根据业务数据和场景持续迭代优化预测效果。针对预测结果，有效协同销售数据与需求数据，输出建议支持决策，促进企业降本增效。

典型用例二：视频智能搜索

在电视台等传媒行业中，积累了大量的节目视频，在移动互联网下，很多人希望可以智能化地检索往期内容，查看相关节目视频。TCDataAgent支持用户构建视频RAG，将视频转换的文本存入TCDataAgent知识库，使用TCDataAgent智能搜索能力，检索出相关节目片段，关联视频智能反馈给客户，支持交互式对话和相关推荐。

典型用例三：微信读书“AI 问书”

微信读书 是一个流行的在线阅读平台，为数亿用户提供海量的书籍、漫画、公众号内容，及在线听书等服务。ES支持书籍内容的智能检索，平台可以形成对搜索词的完整理解和认知，来支持开放式问题回答、并支持书籍引源、猜你想问等丰富的互动能力。

智能体开发

腾讯云智能体开发平台

◆挑战&痛点

当前，企业在智能体应用构建和落地中面临多重挑战：

1、知识碎片化与低效管理：

企业文档形式复杂（如跨页表格、图文混排、多模态内容），传统解析技术难以准确提取非结构化数据，导致知识库更新滞后且检索效率低，最终导致解析及问答效果差。

2、业务场景复杂化：

客服、金融、教育等领域需处理多样化咨询问题，但传统AI系统依赖人工配置流程，灵活性和扩展性不足，多轮对话不准确，配置成本高。

3、数据孤岛与安全风险：

多源数据分散在独立系统中，缺乏统一治理，且敏感信息（如金融数据）易因技术漏洞泄露。

4、智能化应用门槛高：

大模型落地需专业团队处理结构化、非结构化数据清洗、模型调优和流程编排，中小企业难以承担技术及成本压力。

5、效果难以评估：

大模型应用落地过程中，没有好的工具和评测框架，试错成本高。

6、智能化应用复杂度不断提升：

在AI Agent模式下，面对日益复杂的任务场景，单一智能体的能力日益受限，难以有效解决诸如场景耦合度高、并行处理困难等问题。

◆产品方案

提供LLM+RAG 、Multi-Agent、Workflow等多种应用开发框架，预置精选官方插件及MCP插件，支持应用配置-应用调试-应用评测-应用发布一站式工具链，助力企业降低大模型应用构建门槛。

1、面向企业服务严谨场景，提供自研RAG算法和最佳实践，同时提供问答对生成/校验、保守回复等多种运营工具，适用于严肃严谨场景，如政务服务、政策咨询、智能客服等。平台内置最佳实践流程，只需导入文档/问答对，即可即可让大模型对接企业知识，达到更稳定和精确的知识问答效果。适用于企业知识服务、产品咨询等严肃问答场景。

2、面向的企业多模态知识，提供领先的文档解析效果和图文混排处理能力：强化了对图文表混排文档的处理能力，通过OCR解析大模型提升识别精度，有效处理标题、公式、页眉页脚等文档元素。广泛适用于各类行业文档知识处理，如图文混排零售/出行等说明书、研究报告等。

3、支持Multi-Agent多智能体模式。由大模型根据提示词自主拆解复杂任务和规划路径，模型主动选择和调用工具，并能够主动纠错和反思，回复效果更灵活。且可在单AI Agent的基础上添加更多AI Agent，从而让应用实现多个AI Agent协同调用来响应复杂任务。

腾讯云智能体开发平台提供了【自由配置转交】【工作流编排转交】【P&E协同模版】多种Multi-Agent构建方式：

- 【自由配置转交】支持零代码多Agent协作转交，用户可简单设定转交关系，实现“专家协同体系”。

- 【工作流编排转交】基于工作流中的Agent节点，利用工作流画布，实现通过固定流程编排Agent，确保任务执行流程稳定可控。创建Multi-Agent模式应用后，只需将Agent协同方式调整为工作流编排，即可在对话界面对整个工作流进行调试。

- 【Plan-and-Execute（P&E）协同】是一种预设的协作模式，它将任务明确分解给规划（Plan）与执行（Execute）两类独立Agent。与其他协同方式相比，P&E协同预置了特定的Agent角色、协作逻辑及记忆管理机制，目前仅开放少量配置项供用户调整。该模式内置了多Agent协作的最佳实践。通过周密的规划与高效的多Agent协作，能够生成结构严谨、细节丰富的成果。它特别适用于对实时性要求不高，但内容深度与完整性要求较高的场景，例如：深度分析报告生成、复杂网页生成等。

4、提供工作流全新解决方案。即可通过可视化拖拉拽的方式编排不同的原子能力（如大模型、知识库、插件等），节点全面清晰，支持零代码构建复杂业务流，且针对业务中灵活的多轮对话难题，提供多参数提取、参数回退等优势能力。适用于有复杂业务流程的企业服务场景，如寄快递、挂号等。工作流编排协同基于工作流中的AI Agent节点，利用工作流画布，实现通过固定流程编排AI Agent，确保任务执行流程稳定可控。创建Multi-Agent模式应用后，只需将AI Agent协同方式调整为工作流编排，即可在对话界面对整个工作流进行调试。

5、支持一键调用微信支付、位置服务等官方及MCP插件与自定义扩展插件，极大提升系统扩展性，灵活满足各类组合场景需求；提供多层权限体系及全流程管理能力，为企业用户提供高可靠、易运维的智能体应用体验。

6、提供了全方位的智能自动化评测能力，助力企业高效、客观、地衡量大模型应用效果。支持裁判模型打分、代码打分、规则打分等多种评估方式，按需选择适合的打分方式，支持将繁琐的评测工作流自动化，实现对智能体性能的快速、批量、标准化评估，显著提升迭代效率与评测客观性。

◆场景&价值

场景案例1：金融业务提效（重庆农商行）

金融产品知识繁杂、传统金融服务效率不足、风控精准度有限以及数据价值挖掘不充分。借助腾讯云智能体开发平台的联网搜索、RAG与知识库能力，重庆农商行基于腾讯云智能体开发平台内置的DeepSeek模型推出智能助手“AI小渝”，形成三大解决方案：1）智能风控——动态识别欺诈行为，提升风险预警能力；2）场景金融——搭建分钟级响应智能客服，结合知识库数据提供个性化财富管理建议；3）数据决策——通过大模型挖掘行内金融数据价值，优化信贷评估与市场策略。效果上，“AI小渝”显著提升了员工协同办公效率，推动金融服务智能化升级，并强化了其“数字农商行”的品牌价值。

场景案例2：医药零售服务（大参林）

大参林作为医药零售行业龙头企业，面临内部办公协同效率不足、垂域业务场景知识响应慢、经销商培训赋能低效等痛点。依托腾讯云智能体开发平台能力构建了行业首个医药专属AI知识库系统"AI小参"。有效解决了传统大模型与企业知识脱节、垂域效果差等难题，实现毫秒级精准响应：在销售场景，药品查询效率提升80%；在办公场景优化跨部门协同；通过分析60万条用户反馈数据辅助决策。目前，AI小参已在全国范围内直营店与加盟店上线使用，并持续推动AI小参从知识问答到销售助手再到决策引擎的全方位进阶，塑造医药零售行业智能问答新标杆。

场景案例3：政务知识问答服务

政务服务机构面临高频咨询压力。

（1）缺乏维护人员：服务渠道线上服务平台建设滞后，传统AI效果差，无人维护，无法满足特定需求；
（2）知识管理断层：知识文档健全，但文档分类、维护缺乏专人管理，知识没有形成体系；
（3）同质化问题多：高频次、同质化政策、查询等问题多，服务人力紧张，传统AI解决率低；
（4）数据驱动不足：问题处理过程无量化追踪，难以做分类分析，并进行服务优化改善，亟需高效构建大模型智能应答体系，提升民生服务质量。

腾讯云提供大模型智能体开发平台，助力西南某省会公积金构建大模型智能应答体系，提升服务效率，并实现从知识服务到服务反馈、服务优化的智能化发展闭环。经过实践认证，西南某省会公积金通过大模型智能化服务了90%以上高频标准化问题，准确率预期达95%以上；构建了端到端智能服务流程闭环，整体提效40%。

基础设施

模型训推-腾讯云 TI 平台

◆挑战&痛点

大模型训练和推理是支持智能体落地的关键环节，企业在大模型训练到应用落地的过程，面临“算力管理难、数据处理难、训练效率低、推理成本高”的四重挑战：

- **异构算力纳管复杂，算力利用率低:**企业需管理多厂商不同架构、型号的算力，并将这些算力用于多个训练推理业务，算力调度挑战大，且离线业务潮汐特征明显导致资源利用率低。

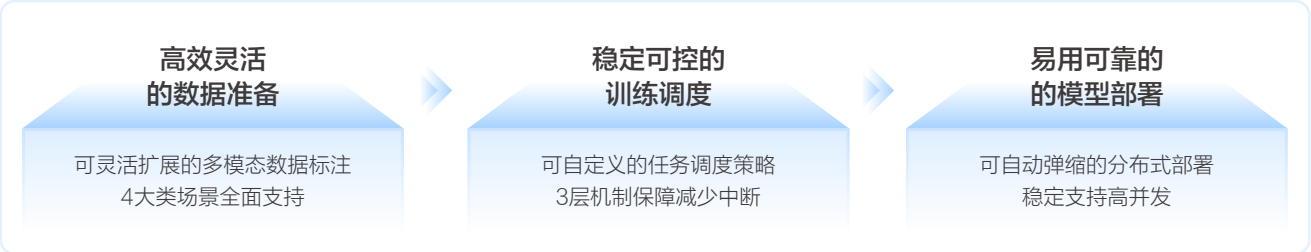
- **数据准备流程繁琐，处理灵活度不足：**大模型领域高速发展，精调需处理文本、图像等多模态数据，现有工具多依赖固定处理流程，无法满足灵活数据处理需求，导致数据准备周期长。

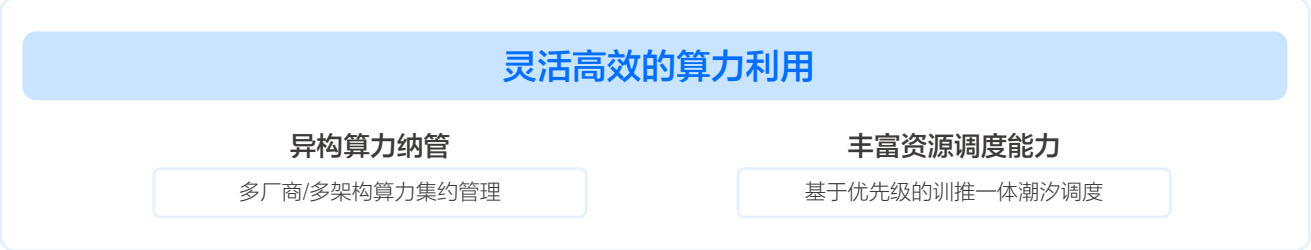
- **大规模训练稳定性差，故障恢复困难：**多机多卡训练中，机器故障、网络中断、Pod异常等问题时有发生，单节点故障可能导致数天训练成果丢失；同时，企业内多用户间可能产生资源不合理抢占。

- **推理算力需求大，应用成本高昂：**大模型推理需使用大量算力资源，成本相比于传统AI模型显著增加，需在原生模型基础上进行多种性能优化，从而合理控制成本。

◆产品方案

针对上述挑战，腾讯云TI平台以“面向实战的大模型训推”为核心目标，构建“算力+数据+训练+推理”四位一体的解决方案，通过技术创新破解痛点：



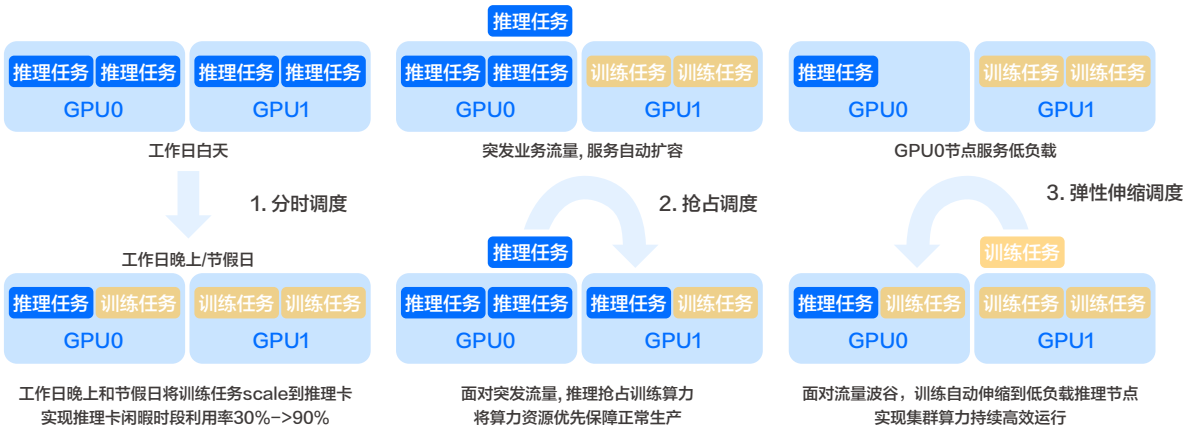


- **以“灵活数据构建+自定义数据标注”降低数据准备门槛。**在数据构建方面，平台提供可扩展的数据处理流程（Pipeline），内置原始数据分析、数据清洗、数据去重、Prompt优化、训练格式转换等五个步骤的数据处理能力，以开源代码形式内置于开发机，支持灵活修改。在数据标注方面，平台突破传统“固定标注操作台”的限制，支持用户自定义数据集Schema，自动生成自定义操作台，覆盖多场景的复杂标注需求，大幅提升数据准备效率。

- **以“三层容错机制+自研加速框架”确保训练稳定高效。**在训练稳定性方面，平台构建三层保障机制：机器故障自动迁移、异常Pod驱逐与重调度、断点自动续训，有效应对节点故障、磁盘异常及系统中断等问题，保障训练任务持续运行。在训练性能方面，内置自研Angel训练加速框架，相比开源框架性能提升50%以上，大幅缩短模型训练周期。

- **以“自研推理加速+多维服务管控”提升推理性能与效率。**在推理加速方面，平台自研Angel推理加速框架，针对DeepSeek模型的推理性能较开源版本提升超20%。在服务管控方面，提供健康检测（存活、就绪、启动）、Token级精细化限流与多维度LLM监控能力，全面覆盖首Token延时、Token吞吐量等核心指标，帮助企业实现高稳定、高可控的推理服务部署。

- **以“智能调度 + 故障诊断与恢复” 提升资源利用率。**在智能调度方面，平台提供灵活可配的排队与优先级抢占策略，适配不同业务场景；同时引入潮汐调度方案，日间优先保障推理服务，夜间自动将训练任务调度至空闲推理节点，实现分时复用与资源高效流转。在故障诊断与恢复方面，支持快速定位问题根因，并恢复异常节点，确保任务/服务的稳定运行。



◆场景&价值

场景一：头部无人机企业大规模多机多卡训练

某头部无人机企业原自行管理算力与存储，训练效率低，面临200T数据下的Transformer训练瓶颈。TI平台集成TurboFS存储与GPU算力，提供Notebook调试、多机多卡训练及RDMA与TI-ACC加速功能，支持VScode快速转换任务，训练性能提升40%，大幅提升迭代效率。

场景二：金融大模型精调

某互联网金融客户需训练投顾与投研大模型，但缺乏底层工程能力。TI平台提供从数据管理到服务部署的全流程支持，包括30+数据处理方法、多模态标注与稳定任务调度，支持单任务长时运行。客户可专注算法效果，通过高效纳管与实时监控提升资源利用率，加速模型落地。

场景三：泛互行业大模型服务部署

某互联网公司需在自有集群部署高并发大模型服务，原面临资源争抢与性能不足。TI平台统一调度多台GPU服务器，实现分布式部署与弹性伸缩，显著提升DeepSeek等模型的推理吞吐与稳定性，支持知识库、代码助手等多场景低延迟高并发需求，提升资源利用率与服务可用性。

◆挑战&痛点

在席卷全球的数字化浪潮中，数据已不再仅仅是信息技术的副产品，而是驱动业务创新、优化决策流程、构建企业核心竞争力的关键战略资产。然而，绝大多数企业在从海量数据中挖掘洞察、实现价值转化的进程中，仍普遍面临以下核心挑战：

1、数据理解成本高

数据的业务含义模糊不清，大量复杂数据难以被业务人员直观理解，更难以被AI系统准确识别与处理，导致数据“看得见却用不好”，严重制约了智能化分析与自动化应用的落地。

2、计算口径不统一

关键业务指标分散于多个系统，同一指标在不同场景下计算逻辑不一致，导致“数出多门”、结果矛盾。若依赖大模型进行自动计算，其对统计逻辑的理解也可能偏离实际业务口径，进一步影响决策的准确性与可信度。

3、分析灵活度不足

传统指标开发依赖固化报表或ADS表，分析维度预先设定、难以动态调整，无法满足业务人员或AI Agent在不同场景下进行多维度灵活组合分析的需求。例如，当需要按客户地域或等级对DAU进行细分分析时，往往因缺乏预置维度或需二次聚合而受阻，难以支撑敏捷的数据探索与快速响应。

这些问题正成为企业释放数据潜能、迈向智能化运营的关键瓶颈。唯有构建统一、语义清晰、灵活可扩展的统一企业语义体系，才能真正让数据“活起来”，驱动业务持续创新与增长。

◆产品方案

WeData以指标平合和治理中心为落脚点打造统一语义层（Unity Semantics），实现指标口径定义和元数据管理标准化，解决企业因数据标准不一、口径混乱导致的数据解读偏差和沟通障碍，显著提升数据资产的可管理性和业务含义的清晰度，极大降低数据理解成本。

1、语义层组成元素

语义层是一个面向数据消费者设计的跨组织、统一整合的企业业务数据视图，它通过搭建一座桥梁，将技术复杂、结构繁琐的原始数据转化为易于理解的业务语言，让数据变成“人人可懂”、“机器亦可读”的企业数据资产。

语义层不仅仅是简单的数据映射，而是由概念、关系、指标、维度四大关键元素构成：

概念：通过将技术字段翻译成业务语言，使数据“说人话”。例如，“customer_id”变为“下单客户ID”，“order_date”变为“下单时间”，降低业务人员的理解门槛。

关系：建立数据之间的关联，如同订单与客户、产品的联系，确保数据分析有逻辑性和依据性。

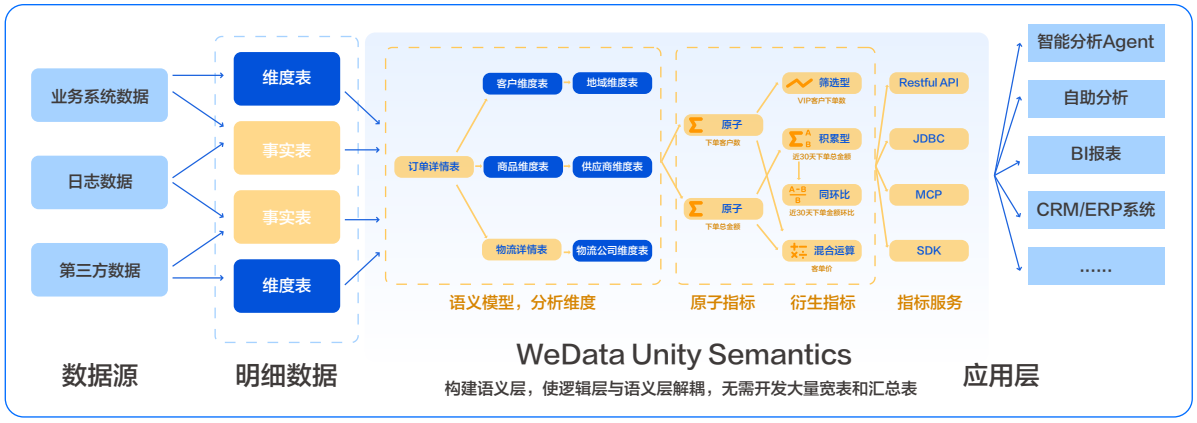
指标：统一计算口径，保证同一指标在不同系统中的一致性，如明确DAU的定义和去重规则，实现“Single source of truth”。

维度：提供清晰统一的分析维度，如客户等级、地域等，支持业务人员进行灵活多维交叉分析，真正实现敏捷用数。

这四大元素共同构建了一个可理解、可信任、可复用的语义层，不仅让数据从“能看”转变为“好用”，而且极大地提升了数据的价值转化效率，助力企业在数字化转型道路上更进一步。

2、语义层产品架构

强大的语义表达能力：通过业务概念与术语定义、数据关系建模、复杂指标的统一定义与开发、以及标准化维度建模，构建面向业务的统一语义模型。将技术语言转化为可理解、可复用的业务语言，实现数据资产的语义化、标准化表达。



自适应查询加速能力：支持智能路由、多级缓存、物化视图等技术，对用户取数请求进行自动路径优化与执行策略调度。通过动态查询加速机制，显著提升响应速度，保障高并发场景下的高效分析体验。

生态应用连接能力：对语义层数据构建统一知识索引，支持多模态混合检索与语义理解。开放MCP、RESTful API、JDBC、SDK等多种接口，无缝对接BI工具、 AI Agent、报表系统等生态应用，实现灵活接入与深度集成。

多源异构数据融合能力：基于统一数据目录，整合跨系统、跨源、异构的数据资源，进行逻辑化组织与编织，构建企业级统一数据逻辑图谱，打破数据孤岛。全面支持StarRocks、Doris等主流OLAP引擎，实现高效、灵活的数据存储与计算。

◆ 场景&价值

场景一：作为AI Agent准确取数的“法宝”

语义层支持开放MCP（Metrics & Context Protocol）服务，可供企业内部的AI Agent调用，以获取准确、一致的业务数据。例如，营销策划人员可通过AI Agent的自然语言交互界面直接提问：“本周各类会员的DAU表现如何？”。

此时，AI Agent将依托统一语义层中预定义的营销相关指标语义，通过调用MCP服务，自动识别出该问题对应的指标为“DAU”，分析维度为“会员类型”，并将其转化为标准的语义查询语言（SemQL），向语义层发起精准请求。系统返回准确的DAU计算结果后，AI Agent可进一步完成趋势分析、归因解读与可视化呈现。

这一机制将传统的Text-to-SQL模式升级为“Text-to-Metric”的语义化取数方式，大幅降低因自然语言理解偏差或SQL生成错误导致的数据幻觉风险。通过统一口径、统一语义的指标服务体系，真正实现“一次定义，处处可信”，显著提升数据获取的准确性与效率，让业务人员能够更专注于洞察发现与决策创新。

场景二：构建企业指标服务中台

WeData Unity Semantics作为企业业务指标服务的统一中台，通过统一语义建模、标准化数据服务等，实现了数据口径的一致性、数据使用的便捷性和数据分析的准确性，为企业数字化转型提供了坚实的数据基础和强大的支撑能力。各个业务系统再也不需要各定义开发指标，从根本上解决数据口径不一致的问题：

统一指标配置与管理

通过语义建模，各个业务系统无需再各自定义和开发指标，而是可以直接在统一语义层中配置指标计算口径和分析维度；统一管理指标和维度的生命周期，确保所有系统的数据口径完全一致，从根本上解决了“数出多门”的问题。

高效的数据服务接口

提供Restful API、JDBC等多种方式，将统一配置的指标和维度便捷地提供给BI（商业智能）、CRM（客户关系管理）等各个业务系统使用。这种标准化的数据服务接口不仅简化了数据接入流程，还提升了数据使用的灵活性和复用效率。

向量数据库

◆挑战&痛点

在人工智能大模型与智能体应用迅猛发展的时代浪潮下，众多企业都在积极拥抱这一变革机遇，企业内部的数据形态和搜索需求正在经历革命性变化。图像、文本、音视频等多模态数据量正以前所未有的速度持续激增。面对海量数据，如何高效处理并充分赋能AI Agent应用，进而构建契合 AI Agent 时代需求的新一代数据基础设施，已然成为众多企业高度关注的问题。

◆产品方案

要实现AI Agent真正在企业落地，关键在于实现企业数据与大模型的深度结合——向量数据库正是大模型访问企业数据的「必备方案」，大模型+向量数据库也成为了企业落地AI应用的「最佳搭档」。

通过将企业内部的结构化和非结构化数据转为向量并存储于向量数据库中，即可构建企业内部的数据枢纽，不仅突破了传统检索的鸿沟，更构建起连接多源异构数据与AI Agent应用的智能桥梁，成为推动下一代搜索技术演进的最佳实践，实现到对企业搜索、智能推荐等应用系统进行全面升级，并结合大模型LLM的能力，实现效率、用户体验等方面质的飞跃。

腾讯云向量数据库是国内“首家”获得权威机构（中国信通院）认证的企业级自研分布式数据库，源自腾讯集团多年技术沉淀，稳定运行于腾讯内部100+核心业务线（例如腾讯视频、腾讯会议、QQ音乐等国民应用），每日支撑超过8500亿次向量检索请求，可支持千亿级向量规模存储、五百万QPS及毫秒级查询延迟；向量数据库作为AGI时代的“数据枢纽”，专门用于在搜索/推荐和AIGC场景中提供文档、图片、音视频等非结构化数据的存储检索服务，是大模型落地AI应用的“最佳拍档”。

◆场景&价值

背景：某教育头部客户大力发展AI Agent老师业务，期望将内部沉淀的教材数据和教辅数据与AI Agent结合，打造AI智能教师，发挥海量课程、试题资源的数据价值，快速、准确地回答学生提问，提高在线教学效果，减少人力成本。

难点：在实际落地过程中，因为内部数据大多为非机构化数据如图片、文档、视频，存量数据不能做准确、快速地召回；初期使用使用传统关键字检索方案，但学生提问较泛且不标准，导致效果测试效果只有60%准确率。客户在初步调研测试后使用腾讯云向量数据库作为AI Agent老师的数据底座，向量数据库支持的语义级检索方式作为检索核心。

效果：

1.上线后，经客户测试认证效果，将所有数据向量化后存入向量数据库可以覆盖95%场景，且效果优于关键字检索方案。

2.从AI Agent老师上线后，大幅提升传统老师工作效率，同期服务学生数量提升1.5倍，且学生好评率提升20%。

3.因向量数据生成需要选择Embedding模型并搭建集群，资源开销较大且需要人力维护，最终选用腾讯云向量数据库内置的Embedding功能，业务接入效率提升80%，客户成本降低10倍，至上线以来未出现过稳定性问题。

腾讯云智算

◆挑战&痛点

随着能力升级和即将预见的爆发，AI Agent也正面临着大规模极速延展、用户体验、隐私安全等技术挑战：

一是AI Agent需要更快的大脑，负责理解任务、规划步骤并生成指令。大模型参数量呈指数级增长，导致模型文件大小和分发加载时间显著延长（以Deepseek为例从下载到加载需要1个小时），传统弹性扩容方案难以适应快速增长的流量需求。同时参数量的增长导致计算耗时激增，这与互联网AI应用的即时响应体验背道而驰。这种延迟在智能客服、代码生成等高频场景尤为突出。如何保证模型推理质量的同时，减少重算代价实现秒级响应，以及提高多机交换效率实现更快的吐字效率，已成为行业期待突破的技术瓶颈。

二是AI Agent需要更可靠的手，负责调用工具链、操作外部资源。当AI Agent拥有高度自主性与执行能力后，高度依赖外部工具（浏览器、PC应用、Code Interpreter、MCP等）时，需保障工具的注册发现、调用的稳定性、执行轨迹可追溯，多工具、多会话、多步骤的执行过程也增加了调试难度。此外，AI Agent过高的自主执行权限可能造成不可逆的破坏（如批量删除业务数据），跨系统调用的认证复杂性等。

◆产品方案

在AI模型推理层面，一个高效的推理服务需要满足这两大要求——扩容快、响应快，才能把资源利用率尽可能提升。腾讯云推出HAI推理集群，基于在云计算、分布式存储、高性能网络等基础设施方面的深厚积累和场景优化，提供即开即用、安全稳定、免运维模型云上专属推理集群，大幅提升云上资源的性能和利用率，最终实现降本增效。其中包括：

模型服务启动快

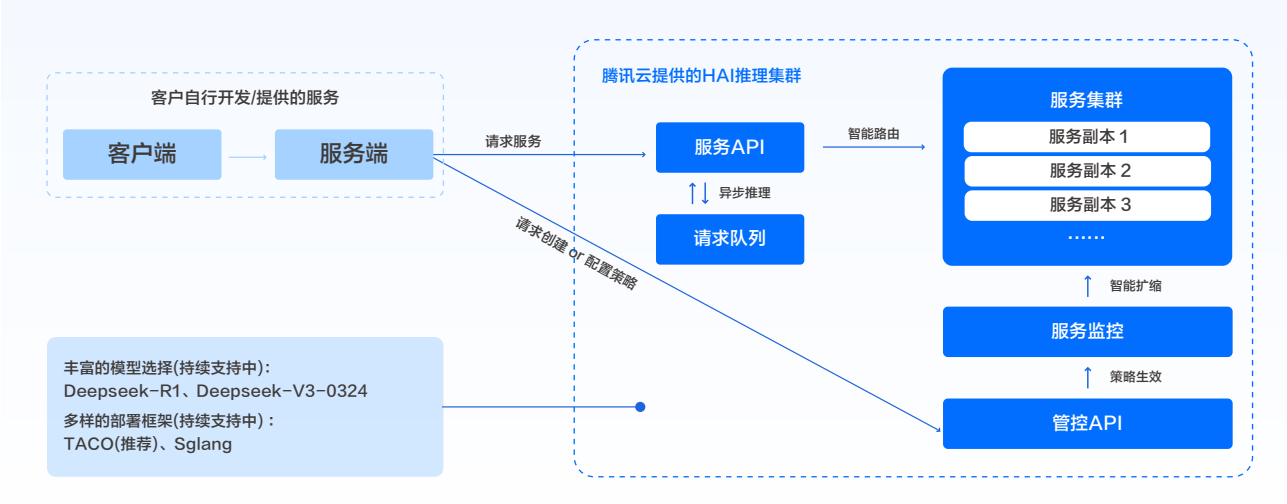
模型启动时间影响业务拉起的时效。模型文件大小随着参数量指数增长，导致推理服务冷启动和扩容时的模型下载和加载时间长，压缩模型下载和加载时间变得尤为关键。

基于RDMA的去中心化模型文件加速分发创新优化，针对星脉网络进行了多端口流量负载均衡和基于交换机亲和性的节点选择优化，实现多节点扩容时在3秒内完成DeepSeek满血版权重分发，是友商传输性能的5倍，大规模扩容整体服务拉起时间从30min缩短至34秒。

KV缓存复用率高

首字延迟短取决于能否减少重复的计算。在多轮对话中，推理过程通常存在大量的上下文计算，导致首字响应慢进而影响在线业务的用户体验，进而需要一个全局缓存池存储上下文或相似的计算结果，以减少重复计算量。

持久化存储和集群级共享的KV Cache，模型推理实例能够优先使用已缓存KV，直接进入Decode阶段，减少Prefill重算代价，缓存命中率提升50%以上，业内最佳。



在模型推理基础之上，针对AI Agent的安全执行环境。腾讯云重磅推出新一代企业级智能体基础设施引擎AI AgentRun，为企业在生产环境中构建、部署、运行大规模AI Agent应用提供极简运维、全链路安全的基础设施。其关键能力包括：

安全沙箱弹得快

沙箱创建效率关乎AI Agent执行任务的效率。需要超高并发的安全沙箱与秒级启动速度，否则无法支撑密集任务和资源弹性调度。

基于会话隔离、秒级启动、超高并发的Serverless运行时Cube，为编码助手、GUI-Agent、自动化测试、Agent RL等场景提供安全、弹性、低延迟的运行环境。支持浏览器、PC、Code等多类型沙箱，实现工具安全接入与环境级防护。沙箱冷启动时间相比业内提升18%，最高支持每分钟1万个沙箱。

◆场景&价值

针对模型推理效率和AI Agent工程化的痛点，腾讯云智算持续打磨调度能力，演进成为 更贴近AI Agent的AI Infra，帮助Agentic AI从“实验室”走向“生产级”。

典型业务场景：

- 搭建AI团队+整体部署推理服务周期通常不短于3个月以上，而业务侧通常需要快速构建模型推理服务。
- 第三方模型API通常共享资源模式，存在稳定性、可靠性、数据安全性等不可控风险，需要专属独立的计算环境。
- 编码助手类AI Agent需在安全的IDE或Code Interpreter沙箱中执行编译、调试与代码运行。
- GUI-Agent需在浏览器或桌面应用沙箱中操作企业内部系统完成流程自动化。
- 自动化测试AI Agent需在隔离的测试机沙箱中批量运行UI/接口测试并生成报告。
- 基于Agent RL的智能体训练需在大规模并发的仿真环境沙箱中进行任务执行与环境交互，以获得高质量的策略学习数据。

TDAI Agent Infra

◆挑战&痛点

企业在构建数据库智能体（Agent）时面临三大核心挑战与痛点：

- **垂类模型能力不足：**对复杂ORM操作理解偏差大，SQL漏报率超30%，优化建议命中率低于40%，且输出不稳定，难以内置企业专属规则，依赖外部规则引擎导致成本高；

- **记忆系统与业务脱节：**现有系统缺乏与企业元数据、私有数据及血缘的深度融合，制约智能体对业务背景的理解与实际价值；

- **工具调用缺乏标准化：**缺少统一接口与调度协议，导致实例克隆、流量回放等工具协同困难，系统复杂且维护成本高。

◆产品方案

核心架构与定位

TDAI Agent Infra包括数据库垂类大模型（DB LLM）、全域上下文（Context）及工具集（Tools）三大基础设施，有效提升TDAI Agent Service在任务中的连贯性、上下文理解力以及个性化服务，是企业智能体规模化落地的必备基础设施。

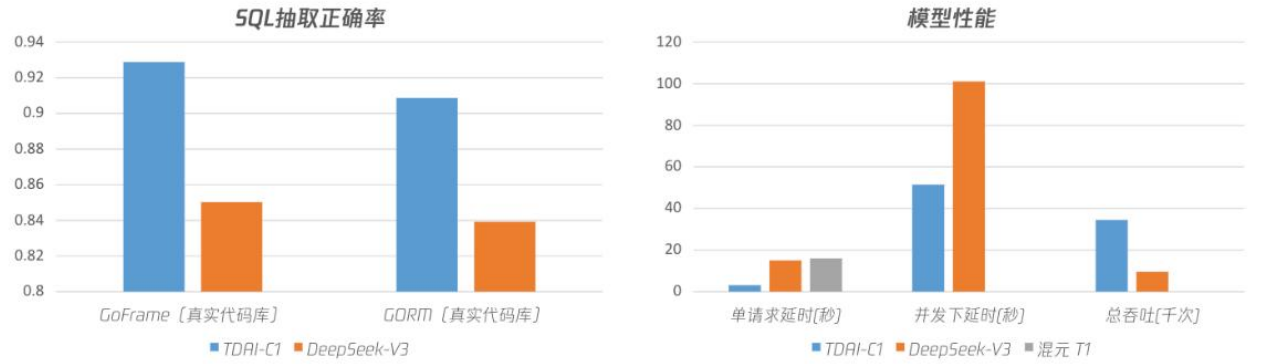
TDAI进一步构建了面向To B场景的智能体基础设施

- ✓ 数据库大模型〔DB LLM〕：通过自研Code2SQL模型〔C1〕、智能诊断模型〔D1〕、智能优化大模型〔O1〕，将**SQL抽取准确率提升至90%+，并对SQL诊断推理持续深度优化，诊断结果高置信度**；
- ✓ 全域上下文〔Context〕：整合Memory〔长短期记忆〕、DeepSearch〔深度检索〕、Catalog〔数据目录〕，构建企业级数据中枢，实现**企业数据与智能体记忆的深度融合**；
- ✓ 工具集〔Tools〕：基于**MCP协议**封装上层智能体所需的原子能力〔如实例克隆、流量回放〕，为智能体提供"从规划到执行"的工具支撑。

基础设施分类与功能体系

数据库大模型（DB LLM）

自研TDAI-C1、TDAI-O1、TDAI-D1模型，通过代码整体建模和ORM框架优化，SQL抽取准确率超过90%，兼容Golang、Python等主流开发语言与框架。引入数据库运行时数据，实现对SQL风险的多维度量化评估，在同等硬件环境下推理吞吐量达到通用模型的4-7倍。

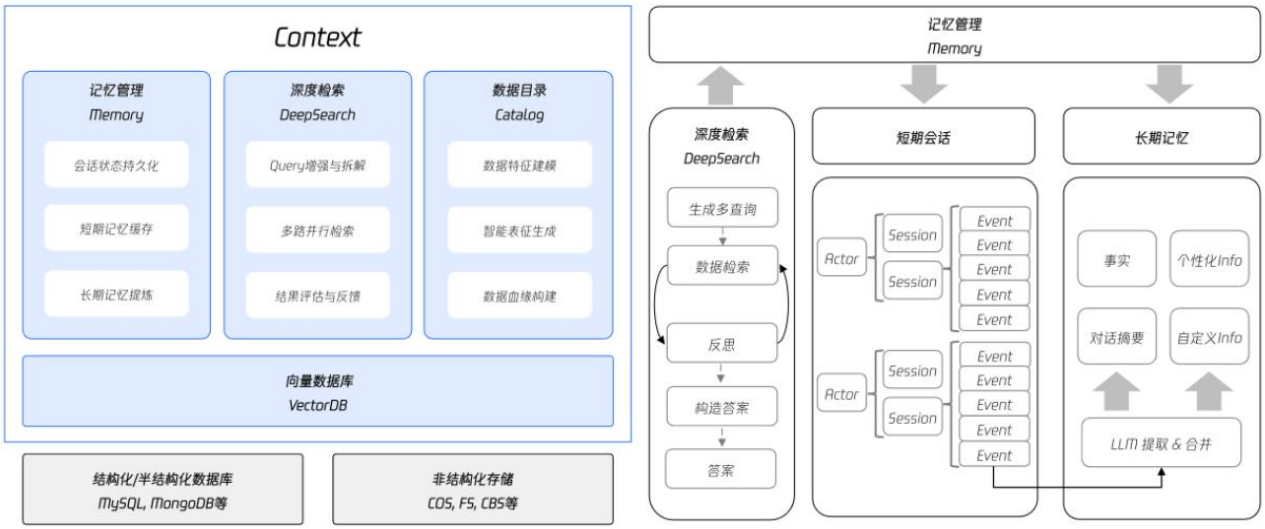


C1基于真实代码库的SQL抽取能力优于通用大模型，模型性能全面占优

- ✓ ***SQL抽取正确率：****通过代码整体建模、注入ORM框架特性知识等技术进步，自研模型在复杂代码解析中实现**端到端准确率超90%，较通用大模型提升5-8%，有效解决多度caller、隐式调用链等漏报难点。***
- ✓ ***性能：****在资源规模相同的条件下，Code2SQL自研小模型可支撑通用大模型**4倍以上的吞吐，以及1/5的延迟。***
- ✓ ***最小化部署：****在资源受限时，Code2SQL自研小模型**最低可通过单张H20部署**，以低成本提供服务，且不受部署环境限制。*

全域上下文系统（Context）

由Memory、DeepSearch和Catalog三大模块组成，构建企业级数据中枢与记忆基座。Memory负责历史交互信息的存储与检索，Deep Search通过自然语言处理实现精准信息定位，Catalog自动生成元数据描述并构建企业级数据地图，全面提升数据定位与SQL执行的准确性。



工具集（Tools）

基于统一的MCP协议，实现工具的标准化接入与协同调度。涵盖数据治理工具（如SQL抽取模型、索引优化建议器）、资源调度工具（实例克隆、弹性扩缩容）和智能增强工具（集成上下文模块），为智能体提供从决策到执行的完整能力支撑，助力企业实现数据库运维的自动化与智能化转型。

◆场景&价值

TDAI Agent Infra聚焦三大核心场景，为企业提供智能化数据解决方案。

场景一：智能数据库治理与运维

通过垂类大模型实现风险SQL精准识别、慢查询优化与合规审计。将SQL抽取准确率从不足70%提升至90%以上，慢查询优化建议采纳率从40%提升至82%，并内置企业专属合规规则，显著提升数据库安全性与运维效率。

场景二：基于企业全域数据的决策支持

打破智能体与企业私域数据间的隔阂，通过Memory、Catalog、DeepSearch构建统一数据中枢，使智能体深度理解多源异构数据，提供更具业务价值的分析洞察，释放私域数据潜力。

场景三：智能体自动化工作流执行

通过MCP协议提供标准化工具集（如实例克隆、流量回放），使智能体从“顾问”升级为“工程师”，实现跨云、跨数据库引擎的自动化操作，形成“感知-分析-决策-执行”完整闭环，提升运维效率与一致性。

安全防护

挑战&痛点

企业在拥抱大模型与AI时，业务和IT面临严峻安全挑战：数据是AI的“燃料”，但敏感数据在训练、推理过程中极易泄露，如通过提示词注入或不安全的API接口。AI模型的“黑箱”特性使得其决策过程难以追溯，提示词注入等新形态的AI攻击模式，可能引发合规风险与业务损失。MCP协议作为智能体生态中的“通用连接器”，也带来了协议碎片化、传统应用接入安全性控制、身份认证和统一权限管理等一系列安全问题。AI智能体（AI Agent）的自主行为扩大了攻击面，凭证管理、越权访问等问题凸显，对现有安全体系构成严峻考验，亟需可信、可控、可靠的AI安全防护。

产品方案

1、大模型数据安全保护

数据分类分级与脱敏：在数据接入和处理阶段，对敏感数据（PII/PHI、商业机密）进行精准识别、分类分级，并采用国密加密、脱敏、数据匿名化等技术，在保证数据可用性的前提下，最大限度降低泄露风险。特别是在构建RAG知识库或微调模型时，确保输入数据的合规性。

安全的数据流转与存储：建立覆盖数据全生命周期的安全防护，包括传输加密、存储加密、访问控制。针对向量数据库等新兴存储，需制定专门的安全策略。

2、AI基础设施和运行环境保护

通过AI大模型防火墙(LLM-WAF)，可以对智能化应用的边界建立全面的应用和API保护能力，对用户输入(提示词)进行恶意指令检测敏感信息过滤，对模型输出输出进行合规性审查，提供多模型、多场景、高并发环境下的全链路防护能力。LLM-WAF还支持实时检测并拦截针对大模型的算力滥用、提示词攻击及数据泄露风险，助力企业构建可信、稳定、可持续的大模型服务生态。

AI态势管理（AI-SPM）是保护大模型基础设施和运行环境，检测大模型攻击面和漏洞的安全管理系统，通过外部攻击面检测、AI基础设施软件组件成分管理、结合威胁情报和漏洞检测，及时发现和处置安全风险。

3、AI模型与应用安全保护

天御大模型安全网关，作为企业AI应用落地的安全中枢，连接智能体、模型与服务，实现统一治理与高效协同，并通过多层次防护机制解决AI规模化应用中的关键风险。

身份与权限管理：为AI智能体分配独立的、最小权限的凭证，避免凭证硬编码；建立严格的访问控制策略，限制其操作范围和数据访问权限。

行为监控与审计：对AI智能体的行为进行实时监控、记录和审计，及时发现异常活动和潜在威胁，如越权访问、数据外泄等。

全链路安全管理：支持精准识别并处置输入指令、生成内容、运行环境等大模型场景下的各类安全风险，包括恶意指令注入、工具投毒攻击、访问敏感信息等，同时具备弹性的管控粒度，支持基于MCP server以及MCP tools维度的策略配置，为用户提供丰富的、全面的的安全防护能力。

MCP协议封装：提供标准化协议转换能力，支持将传统API接口一键封装为标准化MCP服务，同步生成Client SDK实现AI Agent无缝集成，有效降低开发门槛，助力企业存量业务快速融入MCP生态。

网关流量控制：提供基于MCP Server的路由配置、正反向代理及全维度流量管控能力，可基于参数、版本、协议、负载做动态调度，实现多版本、多模型、多工具动态调度。

多维防护体系：基于开放式安全能力中台，提供标准化接口快速集成内容安全、数据安全等第三方服务组件。用户可灵活配置文本/图片多模态内容审核、结构化数据安全检测等场景化防护策略，构建全面的大模型安全治理体系。

4、腾讯云优势：

腾讯云在能够为企业构建安全可信的AI应用提供坚实保障：

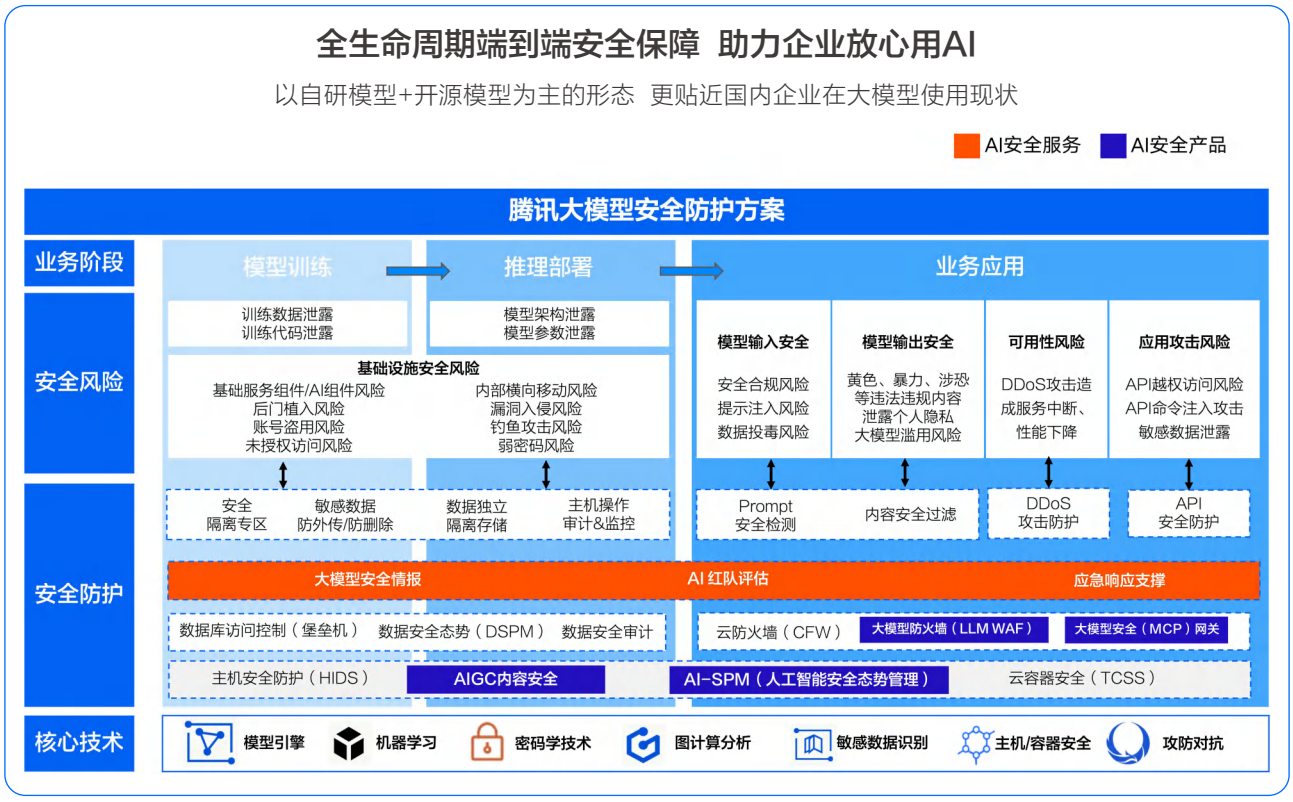
全栈安全能力与经验：腾讯云拥有从基础设施安全（IaaS）、平台安全（PaaS）到应用安全（SaaS）的完整安全产品矩阵和深厚的攻防实战经验，能够为AI系统提供多层次、纵深化的安全防护。

领先的数据安全治理实践：腾讯在海量数据处理和安全治理方面拥有丰富经验，其数据安全产品（如数据安全中心、数据库审计等）能够有效保障AI训练和推理数据的全生命周期安全。

强大的AI技术积累与生态：腾讯混元大模型等自研AI技术与行业解决方案相结合，能够提供更懂业务场景的AI安全能力，例如针对性的内容安全审核、AI智能体行为分析等。

云原生与开放集成：腾讯云平台提供云原生的安全服务，易于集成和扩展，能够与企业现有IT架构和第三方安全工具无缝对接，助力企业快速构建符合自身需求的“数据+AI”安全体系。

通过这些优势，腾讯云致力于帮助企业在充分享受AI智能体技术红利的同时，有效管理和控制潜在安全风险，实现可持续的智能化转型。



◆场景&价值

场景案例一：企业智能体应用的安全防护

某制造企业使用AI智能体监控生产线数据，并自动管控设备参数。大模型在规模化部署中面临注入攻击、敏感数据泄露、MCP Tools投毒等多重安全威胁，企业需要构建从输入到输出全链路的智能体安全防护体系。通过部署天御大模型安全网关，建立针对大模型的多维攻击防御，识别及拦截处置提示词注入、工具越权访问、异常流量等攻击行为，并内置敏感数据识别规则，脱敏身份证号、电话号等高危字段，保护企业系统和信息安全。实时追踪智能体的操作指令和数据访问行为，一旦发现智能体尝试访问非授权数据或执行超出权限的操作，系统会立即告警并阻断，防止因智能体被恶意利用导致生产事故或数据泄露。

场景案例二：零售企业智能客服系统的大模型安全防护

某连锁零售企业上线了基于大模型的智能客服系统，为消费者提供订单咨询、退换货政策解答及商品推荐服务。为防止用户输入、大模型输出中包含个人敏感信息，或恶意用户发起提示注入、诱导攻击，以及模型被误导生成违规、敏感或不当内容，该企业接入了腾讯云LLM-WAF大模型应用防火墙方案。在输入阶段，系统可实时识别手机号、地址等隐私数据

并提示用户，同时检测并拦截潜在的注入攻击与提示词诱导行为；在输出阶段，对大模型生成结果进行内容审查，防止泄露个人敏感信息或输出不合规表述。系统还支持对风险交互内容进行日志留存与策略优化闭环，助力企业在提升客服智能化体验的同时，保障合规运营与品牌声誉安全。

同时通过部署AI-SPM，对于企业大模型基础设施和智能应用生产运行环境进行全面的风险管理和安全监控，结合外部暴露面管理、威胁情报、漏洞检测，综合评估和运营AI应用和基础设施的安全态势，完成管理和响应闭环，保障业务安全稳定运行。

4.企业智能体建设：面向未来的分阶段战略规划

在智能体技术飞速演进的时代，企业智能体已从概念探索迈向规模化应用的关键拐点，其作为能自主理解、决策并执行任务的数字员工，正重新定义企业运营的效率与智能上限。然而，智能体在企业级场景中的稳定性、安全性与成本可控性仍需持续验证。这意味着，企业必须摒弃“毕其功于一役”的激进思维，转而采纳渐进式、可迭代的战略规划，需要清晰的分阶段目标、实施路径和资源投入。

阶段	核心目标	关键行动	预期结果
短期（0~6个月）	试点验证，建立信心	聚焦2~3个核心场景，打造标杆应用，初步验证智能体的业务价值。	成功上线2~3个智能体，初步验证智能体价值，建立团队协作流程。
中期（6~12个月）	平台赋能，场景拓展	建设智能体开发平台，提升智能体业务应用覆盖广度和垂直领域专业深度。	完成企业级智能体开发平台构建，实现多个核心业务场景的智能体落地应用，形成成熟的智能体落地流程。
长期（12个月~24个月）	夯实底座，生态融合	构建智能体底层技术栈，促使智能体与企业业务、流程、组织的耦合共生。	构建完整的智能体技术底座体系，实现与企业业务的深度融合，最终形成一个自我进化、持续创新的内生智能生态。

因此，我们为企业管理者提供一个从短期、中期到长期的阶梯式规划框架，企业可根据自身情况参考调整，制定适合自身情况的智能体建设规划。

短期规划（0~6个月）

核心目标：试点验证，建立信心

本阶段的核心目标是快速落地核心场景，验证商业价值。企业应聚焦于能够快速见效、投入产出比高的场景，通过效率提升百分比、响应时间缩短、人工成本节约等量化的指标，验证智能体技术能带来真实的业务收益。

关键行动：聚焦核心场景，打造标杆应用

首先，在此阶段企业需要识别出技术成熟度高、可快速落地的智能体场景，参考智能体场景罗盘，选择执行复杂度低、自主规划依赖度低的场景，典型的场景包括企业行政问答、员工专家助手、产品营销客服、售后咨询客服等；其次，此阶段建议企业优先采用轻量化的落地方案，如选择成熟的垂类智能体应用软件（SaaS产品）、调用公有云大模型API或利用低代码智能体开发平台快速构建智能体原型；同时/最后，企业应组建包括业务、产品、技术在内的跨部门试点团队，明确智能体试点效果评估指标，上线后持续收集反馈优化智能体。

预期结果：试点应用成功落地，验证智能体价值，建立协作流程

短期成功实现2~3试点场景智能体的上线，明确其可量化的业务价值；同时形成业务部门提需求、IT技术部门实现、最终用户做反馈的敏捷开发与运营闭环，初步建立智能体建设跨部门协作流程；成功的智能体试点也会激发企业更多业务部门的需求，提升企业智能体建设的信心与势能，为中长期智能体建设规划与投入提供有力支撑。

中期规划（6~12个月）

核心目标：平台赋能，场景拓展

本阶段的核心目标是完成企业级智能体开发平台的构建，并以此为核心引擎，实现对业务场景的深度赋能与广度拓展。企业应将战略重心从点状的试点成功，全面转向平台化能力的沉淀与复用，通过统一的技术底座支撑智能体应用在更多业务领域“开花结果”，实现规模化价值。

关键行动：建设开发平台，提升应用广度与深度

首先，企业在此阶段的中心任务是建成统一的智能体开发平台，该平台需要整合模型服务、知识库、工具集、开发组件等核心能力，形成一个可快速响应业务需求的“智能体工厂”，从而为全公司的智能体应用提供统一的创建、管理和运营支持。其次，依托于平台能力，企业应在两个维度上推动智能体应用的拓展：一是追求“覆盖广度”，将短期已验证成功的“高效助手”类智能体（如行政问答）快速复制到人事、财务、IT等更多职能部门；二是追求“垂直深度”，聚焦核心业务线，打造流程更长、专业性更强的“执行专家”与“决策专家”类智能体，解决更复杂的业务挑战。最后，企业必须围绕平台建立一套成熟的智能体落地流程与治理规范，涵盖从需求评估、开发上线到后期运营优化的全生命周期管理，确保规模化应用的高质量与可持续性。

预期结果：建成企业级平台，形成成熟落地流程

中期规划完成后，企业应成功构建一个功能完备的企业级智能体开发平台；基于该平台，在多个核心业务场景成功落地智能体应用，显著提升业务效率与专业化水平；同时，在实践中打磨并形成一套标准、成熟的智能体规模化落地流程，为下一阶段的生态化融合打下坚实的基础。

长期规划（12~24个月）

核心目标：夯实底座，生态融合

本阶段的核心目标是将智能体能力全面内化为企业的核心竞争力，通过夯实底层技术栈，实现智能体与企业业务、流程乃至组织文化的深度融合与“耦合共生”。在这一阶段，智能体不再是简单的赋能工具，而是作为企业数字化机体的“原生器官”，共同参与到价值创造的每一个环节，最终形成一个自我进化、持续创新的内生智能生态。

关键行动：构建技术底座，促进智能体与业务共生

首先，企业需要将中期的“开发平台”全面升级为更稳固、更强大的“智能体底层技术栈”，这意味着要在算力、数据、模型三大支柱上进行体系化投入，构建企业级的算力基础设施、自动化数据治理体系以及支持复杂模型训练、精调与推理加速的训推平台，为实现“全能专家”等高阶智能体应用提供不竭动力。其次，关键行动是打破技术与业务的壁垒，强力推动智能体与企业核心业务流程的深度耦合，通过重塑甚至颠覆现有工作流，将智能体嵌入到研发、生产、营销、服务的全价值链中，使之成为流程的“智能中枢”。最后，企业需要自上而下地推动组织变革，构建适应“人机协同”的新型组织架构与工作模式，培养员工驾驭智能体的能力，让智能体与员工形成优势互补、高效协作的共生关系，真正激活组织的内生智能。

预期结果：构建完整技术底座，形成内生智能生态

长期规划的成功落地，将标志着企业构建起一套完整的智能体技术底座体系；实现智能体与企业业务、流程和组织的深度融合，智能体成为驱动业务创新的核心引擎；最终，在企业内部形成一个能够自我进化、持续创新的内生智能生态系统，构筑起面向未来的、难以逾越的核心竞争壁垒。

通过短、中、长期的阶梯式布局，企业不仅能有效管理风险与投资，更能确保智能体建设与技术进步和业务战略的动态发展保持同频，最终稳健地迈向全面智能化未来。

04

智能体先锋实践

文旅：华住集团打造7x24小时“全能酒店管家” 用AI智能体重塑酒店服务

项目背景及痛点

华住集团作为全球第四大酒店集团，旗下拥有31个酒店及公寓品牌，华住会会员规模达2.88亿。华住集团始终处于酒店行业数字化、智能化创新的前沿，通过技术驱动服务体验升级，目前已构建了覆盖预订、入住、服务与管理全流程的智慧酒店生态，其便捷友好的数字化服务体验已成为行业标杆：

为了进一步提升客户体验，华住集团希望借助大模型与智能体等前沿技术，进一步升级客户服务模式，将数字化优势延伸至更智能化、个性化的服务全链条。其目标有以下三点：

- 深化宾客全周期体验：在现有高效自助服务基础上，利用AI技术构建更集中、更主动的信息交互渠道，确保宾客在任何时段都能获得即时、流畅、一致的高品质服务体验。
- 赋能前台运营新高度：希望进一步依托智能体技术与数据赋能，提升数字化系统对宾客需求的闭环处理能力，深度优化前台服务流程，为宾客提供更高效、更贴心、且高度个性化的卓越服务体验。
- 构建数据驱动的服务洞察新范式：当前华住已经构建了对用户服务过程、常见问题和需求偏好的数据分析与运营优化闭环，华住希望在此基础上，使用大模型技术对数据进行更深层的挖掘与智能分析，为服务的持续优化与个性化推荐提供更强大的决策支持。

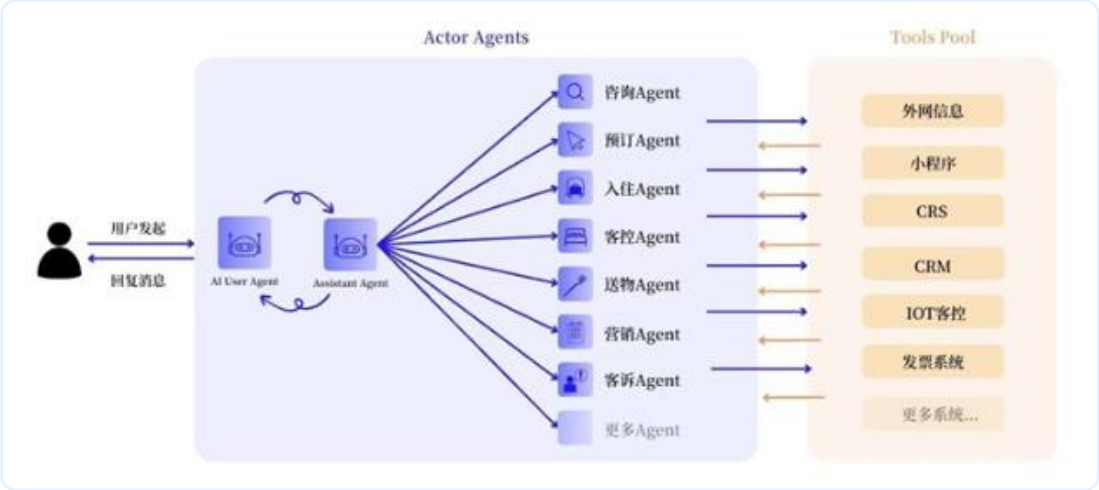
在此背景下，华住集团与腾讯云达成合作，共同深耕智能体领域，以尖端技术赋能服务创新，构建更智能、更主动、更贴合宾客需求的下一代服务新模式。

项目方案

腾讯云与华住集团深入合作，旨在打造酒店赛道首个大语言模型智能体开发平台标杆， 提供LLM+RAG、multi-agent、工作流、MCP组合的全链路行业集成平台，综合大模型和传统小模型能力（如语音识别、语音合成等），支持外呼、400、住中客服等主要服务渠道的高效智能化改造升级。

重点针对最复杂的住中客服场景，腾讯云希望支持华住打造7x24小时永远在线的“智能管家”。腾讯云依托工作流及AI Agent开发范式，将细分的客服场景AI Agent化，提高宾客对话的多意图识别效率与准确性，并结合MCP插件实现酒店业务系统对接，让AI Agent们不仅能“说”、还能“干活”。例如送物AI Agent，宾客跟AI提出需求确认后，AI能自动调用系统生成工单，并调用机器人完成送物上门，提高客房服务效率。

智能体技术尚处早期，真正实用落地需要做大量的工程调优工作。在华住的项目中，腾讯云充分发挥大模型 toB能力与智能体开发应用经验，有效解决落地中的若干关键障碍问题，典型如：1）全局AI Agent支持意图跳出，让用户可以快速切换服务需求，不用等一个服务完成才能发起另一个服务需求；2）自研意图模型youtu-intent-pro，能够更稳定、准确识别用户意图，特别是用户一个问题中包含多个意图的情况。



未来随着大模型能力提升和智能体普及，酒店智能客服将不再只是“客服”。AI智能体能够为所有住客和员工提供“全能”的AI助手，一个AI入口就能调用各类应用，住客还能搜索推荐、导航、预定酒店外的其他服务，酒店的员工和加盟商等则可以获得经营分析、业主服务等智能服务支持。

落地效果

- 降本提效客服：智能体能7x24小时处理常见问询和客需，有效提升客服的响应速度和问题处理效率，同时解放人力、优化人力资源配置，让员工专注于更复杂、高情感价值的服务，提升客户服务的整体水平。

- 增强宾客体验：智能体能提供即时、一站式、个性化服务，提升宾客服务体验和满意度。特别是青年和商旅客户，对数字化、智能化的自助服务接受度较高。

- 提升服务管理运营效率：智能体能有效沉淀和分析宾客数据，为酒店优化服务项目、进行精准营销、开展个性化服务、提升运营效率等提供有力支持。

- 探索增收渠道：智能体能扩展周边推荐、商品服务订购等功能，帮助酒店拓展多元营收渠道，提升非客房收入。

医疗：将时间还给医生 将生机留给患者
迈瑞x腾讯云“启元”大模型重塑重症诊疗范式

项目背景及痛点

加强重症医学医疗服务能力建设，是健康中国建设和卫生健康事业高质量发展的重要内容，是构建优质高效的医疗卫生服务体系、重大突发事件救治体系的重要举措，对于维护人民生命安全和身体健康具有重要意义。

根据卫健委规划，到2025年末，全国重症医学床位达到15张/10万人，可转换重症医学床位达到10张/10万人，相关医疗机构综合ICU床医比达到1:0.8，床护比达到1:3。

重症医学涉及医院多科室，且临床环境复杂，信息不对称普遍存在，不利于医疗救治的协同。且重症患者病情都相对严重，对质量效果和质量的及时性要求更高。然而我国目前能够承担重症治疗的医护人员非常紧缺。且根据丁香园的调研，50%以上的住院医生每天花4个小时以上写病历，甚至还其中还有相当一部分医生写病历的时间超过七小时，挤占了大量的医患沟通和救治的时间。

项目方案

在重症病领域，难点主要源于重症知识的复杂性、多样性和实时性。重症病知识涉及的人体构造复杂，多个系统和器官的相互作用，往往涉及多个部位的病变，需要全面理解这些结构及其功能关系。同时，重症领域涉及的检查检验指标繁多，患者通常伴随大量复杂的生理和生化指标和辅助检查，这些指标的变化趋势和相互关系对治疗决策至关重要。患者治疗过程中，涉及的药品、机器使用频繁且情况复杂，对医生来说记忆负担大。

腾讯对此构建知识图谱，录入全量重症知识，预制检验检查指标、药品等信息的映射关系。在知识录入阶段，就会将知识进行整理，拆分，结构化，再转化到特征空间，存储到知识库中。此外，为了提高大模型在重症医学领域的能力，我们也用重症知识对大模型进行了生成式训练，保证大模型对重症医学的理解能力和表述的专业性。

腾讯联合迈瑞医疗，发布了全球首个重症医疗大模型——启元重症大模型。作为一款具有“重症思维”的“AI队友”，启元能高效处理病情数据，接管文书工作，让医生专注于治病救人。

大模具备四大功能：

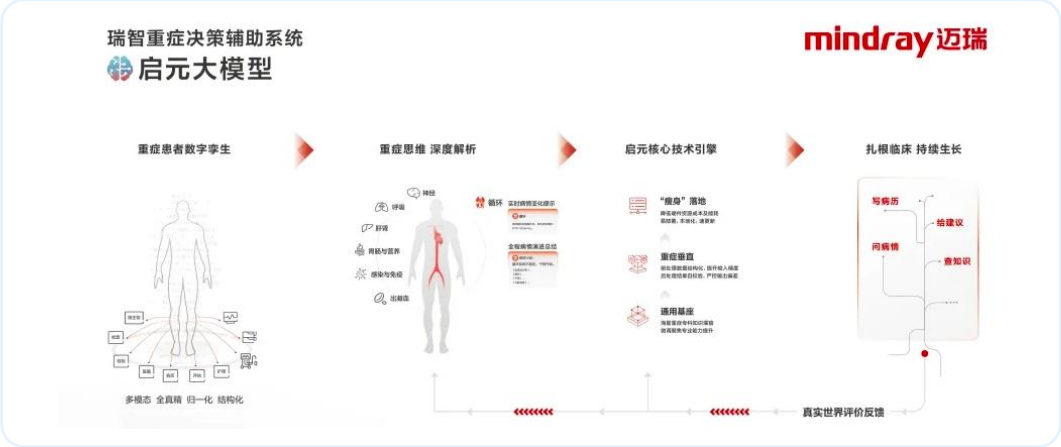
- 病情问答：在5秒内梳理患者病情历程，提取关键指标，生成数字画像，预测病情趋势，并提供治疗建议，帮助医生快速决策。

- 病历撰写：整合诊疗数据后，启元还能高效生成条理清晰、格式规范的病历，1分钟内完成整个撰写过程，效率提高30倍。

• 知识查询：基于九大亚组重症医学知识图谱，启元可精准定位关键知识，分析准确率高达95%，为医生提供高效、权威的决策支持。

• 诊疗建议：结合患者数据和重症思维训练，启元提供个性化治疗方案，助力医生快速制定精准的诊疗计划。

凭借这四大核心功能，启元重症大模型希望能最大限度地减轻ICU医生的负担，为医生赢得更多时间，也为患者争取更多生机。



作为技术基座，腾讯通过混元大模型为启元提供超强“脑力”——万亿级参数和7万亿Tokens，并采用大量医学数据——涵盖285万医学实体、1250万医学关系，以及98%的医学知识和文献，训练出“懂医学”的医疗大模型。

同时，基于大模型知识引擎的RAG能力，腾讯帮助迈瑞搭建面向医生的检索应用，辅助医生快速、精准地检索最新的重症知识和文献。比如，包括呼吸系统、心血管系统、神经系统、消化系统、内分泌与代谢、泌尿系统、感染性疾病等在内的重症基本知识、预防、诊断、用药禁忌、营养支持等信息，为医生节约时间。

腾讯还进一步联合迈瑞医疗，通过“边用边学”的反馈机制，鼓励医护评估和反馈模型输出，不断校准逻辑、补足短板，逐渐形成贴合临床需求的“重症思维”。

此外，通过模型量化、蒸馏和压缩技术，腾讯为启元“瘦身”，让它在医院现有计算资源上就能高效运行，实现临床“用得上、用得起”；支持本地化部署，降低使用门槛，还进一步保障了数据隐私。

落地效果

腾讯与迈瑞医疗合作，共建的医疗大模和重症知识问答系统，实现了良好的效果。

一是实现目标导向的重症治疗。过去设定重症治疗的目标靠经验和感觉，现在通过大模型，可以把过程管理、统计分析、策略调整都可以量化起来，实现精确的目标管理。同时基于软硬件结合，把设定的目标真正传递到硬件设备上，显著提升治疗质量、协同性和精准度。

二是实现患者形象的数字孪生。高度还原患者模型，还原病人病情演变过程，预测病情演进可能。在数字化的世界里，帮助医生做辅助决策支持。

三是全时段的监护预警，帮助医护人员减少非核心工作时间的浪费。

四是针对重症的集束化治疗、标准化治疗。帮助年轻医生快速适应重症临床工作，降低学习成本。

五是对科研工作创建、管理、入组、前瞻性分析、回顾性分析进行系统管理，便利科研成果转化，将科研成果尽快应用到临床。

出行：AI 驱动 安全护航
一汽丰田用大模型打造专家级汽车服务智能客服

项目背景及痛点

一汽丰田成立于2003年，是中国领先的合资汽车制造商，累计销量超1000万辆。持续加码智能化转型，构建覆盖研发、生产到服务的全价值链数字化体系。然而，在面向用户（车主）的咨询服务场景中，如车辆使用疑问、日常故障处理、保修保养政策等C端场景，现有的客服系统仍面临效率不足的问题。目前，大量基础性问题依赖人工客服解答，导致运营成本较高，且响应速度难以满足用户需求。企业直接部署大模型时，仍面临着以下痛点：

企业专属知识脱离：直接部署大模型时，模型缺乏对一汽丰田专属知识（如车型参数、维修手册、保修政策）的深度理解，导致回答不准确或无法提供针对性解决方案。

回答过于宽泛：大模型的回答往往偏向通用性解释，而非针对具体车型或品牌的政策。例如，用户询问“发动机故障灯亮起可能原因”，通用模型可能列举多种可能性，但无法结合一汽丰田车型的常见故障库给出精准诊断建议。

垂域业务场景解决效果不佳：汽车咨询服务涉及大量专业术语和流程（如保养周期、零部件更换标准），通用大模型缺乏行业知识库支撑，难以提供符合品牌标准的答案。

数据安全与合规压力大：企业发展安全是基石，守好网络安全底线，保障数据安全与合规是关键。

项目方案

腾讯云智能体开发平台助力一汽丰田知识解析及知识库运营，打造大模型汽车智能客服，优化用户体验和服务效率，实现企业全域降本增效。

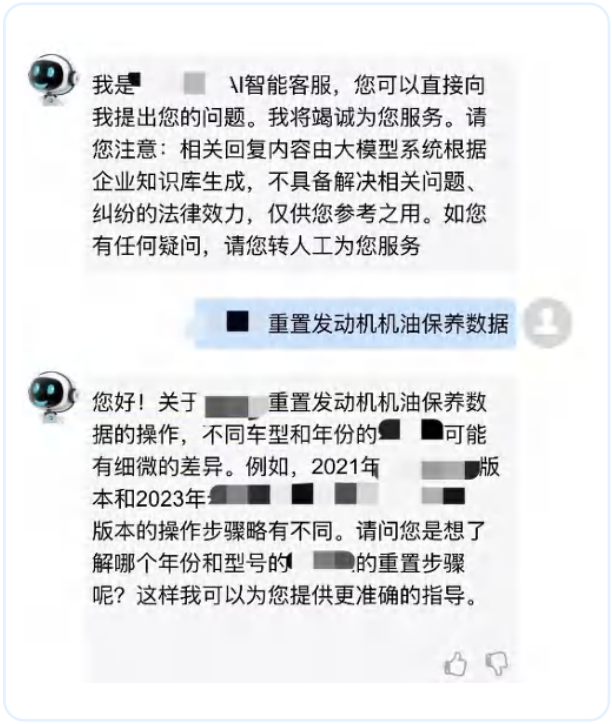
基于腾讯全链路自研LLM+RAG技术，将一汽丰田的车型手册、维修指南、保修政策等非结构化数据转化为可检索的知识库，运用OCR大模型、多模态理解模型和长文本embedding等能力，提升知识检索精准度，支持用户上传车辆故障图片或说明书截图，自动识别关键信息（如故障码、保养记录），并结合多模态理解模型提供可视化解答，解决知识处理与答案生成的技术难点，提升服务的准确度和效率，确保其回答基于企业专属知识，避免大模型的“幻觉”问题。同时基于腾讯云智能体开发平台，对历史客服对话记录进行挖掘，提炼高频问题、优质解决方案等知识库信息，作为企业知识库的有效补充，进一步丰富专业客服知识体系，降低知识库的整理工作量。

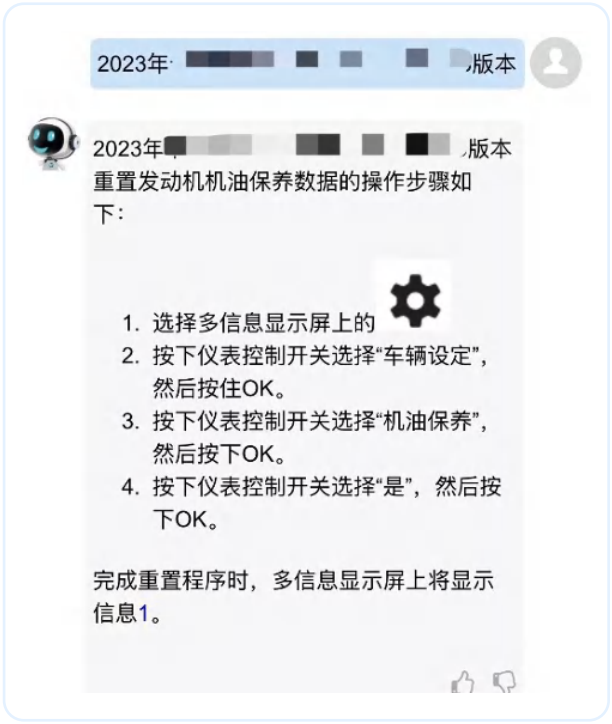
在安全层面，建立数智化安全平台，在整体汽车云安全建设上，一汽丰田充分考虑到混合云、去中心化办公模式带来的治理复杂度，结合腾讯云安全能力，构建了一套云端融合的数智化平台。平台具备多种能力，能够服务于基础安全、数据安全、安全运营建设，为护航重要业务系统起到了关键性作用。

落地效果

借助腾讯云智能体开发平台，在快速接入行业领先的大模型能力后，在一汽丰田APP/小程序/官网/公众号等终端上线大模型服务，率先在智能客服等业务场景深度应用，为用户提供24小时的准确快捷的服务体验，智能在线客服机器人独立解决率已从37%提升为84%，月均自动解决客户咨询问题1.7万次，覆盖车辆使用指导、故障诊断、保修政策查询等高频率场景。并有效减少人工客服介入，使客服团队更专注于复杂问题处理，显著提升了客户满意度的同时，助力企业全域降本增效。

在安全领域，一汽丰田数智化安全平台已经接入多云、多办公环境的安全告警信息，能够做到安全事件RTD（平均检测到响应时间）≤2小时。平台已支持公司内数十个业务系统的数据分类分级，并通过免改造的方式落实了数十个库表的加密、脱敏策略。





零售：伊利集团用智能体打造智能导购新体验 激活全域营销新动能

项目背景及痛点

伊利导购智能体

挑战：导购员缺少编辑社群营销内容的能力，导致大量社群促销信息单一，社群没有提供用户感兴趣的内容，导致社群信息打开率低，有效信息无法触达用户。

项目方案

伊利导购智能体

核心举措：提供高价值内容，建立用户信任，提升群活跃度。

利用腾讯智能体，结合【商品卖点风格+文案方法论选取+多视角评估】，选择最贴合的文案给出建议。

区分三个视角对文案进行评估：

用户视角：用户是文案的主要锚定视角。只有考虑用户本身的需求、痛点和收益，才能吸引其注意，并引发下单动机。

品牌官方视角：品牌视角主要关注商品的核心价值是否传达到位，能否有效吸引用户转化。

专家视角：文案专家的视角主要关注文案的结构、内容和语言表达，以及是否可结合三者凸显商品价值，达成转化目的。

评价维度		用户视角 关注商品价值，是否满足需求解决我的实际问题	文案专家视角 关注文案结构是否清晰、内容是否体现商品价值，语言是否足够吸引人等	品牌官方视角 关注文案的实际转化效果
商品内容 60%	健康与安全： 1. 健康与安全认证 2. 营养价值与成分透明	• 这款商品的健康水准如何，有没有官方背书？ • 这款商品的成分有哪些，营养价值高吗？	• 通过有机、绿色食品认证等，能显著提升信任感。	• 是否出示权威认证，体现我们的专业性。 • 产品优势一定要突出。
	场景代入感	• 我在实际工作和生活中，哪些情况下需要它？ • 饮用它时，会让我想到什么样的画面？	• 场景一定要具象、有画面感。 • 例如：针对具体人群（宝妈/健身党等）、具体时段（早餐/运动后）、情感场景（家庭温馨/职场充电）等，具象画面的体现。	• 该场景一定要让用户代入其中，产生共鸣，才有益于转化。
	活动与福利	• 现在购买有什么优惠吗？ • 我买这么多总要送我点礼品吧 • 什么，只有10个名额，我要赶快下单。	• 限时限量，制造稀缺感。 • 赠品价值可视化（比如，赠品标价）。	• 关注活动效果，体现活动投入的价值，要真正有所转化。
	用户口碑与权威推荐	• 也用这款商品的用户感觉怎么样，他们推荐不？ • 我喜欢的 KOL 也在喝这款？	• 需要有其他消费者、权威专家、或社交媒体、KOL 的背书，增强信任值。 • 使用真实口碑评价。 • 有数据支撑。	• 宣传、投流等成本需要体现价值，起到转化作用。
	品牌价值	• 品牌有什么样的价值理念，和我的价值观一致吗？	• 价值具象化：企业社会责任（碳中和牧场/助学公益等）、专利技术、环保理念等展示	• 体现我们的品牌责任与理念，激发消费者共鸣。
	感官体验	• 这款商品的口感如何？质地是粘稠的还是比较稀？ • 包装适合送礼吗？	• 针对细分人群，描述口感、包装、质地等感官方面的特点； • 增强风味、口感等记忆点	• 这部分有差异化的产品，值得提出来，有卖点。
语言表达 40%	说人话 非专业术语	• 虽然都是专业营养知识，但讲得都是大白话，我能看懂。	• 词藻简洁通俗，不能使用晦涩语句。 • 可以考虑嵌入"控卡""轻负担"等平台热搜词。	• 关注整体语言表达对用户的吸引度和转化效果。

落地效果

伊利导购智能体

经过ABtest，导购的人均单产对比原有方式，销售订单单产提升26.02%，销售额单产提升20.4%，社群商品活跃度明显提升（商品卡点击提升15.7%）。

零售：绝味食品 绝味会员全链路智能化营销

项目背景及痛点

企业简介

绝味食品股份有限公司（以下简称“绝味”“绝味食品”），总部设在湖南长沙，是一家聚焦卤味赛道及连锁加盟体系的运营和管理的公司，为国内现代化卤制食品连锁行业的领先企业之一。2017年3月17日，绝味食品在上海证券交易所成功上市，股票代码：603517。主营鸭脖、鸭翅等卤味产品，以“鲜、香、麻、辣”为特色，卤制技艺融合湘菜工艺，核心产品鸭脖以卤香入味著称。绝味食品在全国拥有超过1万家门店，年营收超过60亿元人民币，每日服务数百万消费者。

案例背景

企业在精细化运营中通过简单的数据分析和A/B测试实验，已经较难提升业绩结果。绝味食品在2024年启动基于AI+数智化营销技术、对会员营销平台进行智能化升级，以实现对近亿消费者的精细化运营，以进一步提升营销转化效果。绝味食品在和腾讯企点合作之初，就希望借助大模型和算法的能力，带来更好的人、商品、权益、内容、触点等元素的组合。

- 实现对数千万用户的数据资产沉淀，清晰洞察用户消费特征，构建符合绝味业务发展的人、货、场标签体系；
- 构建和优化会员营销策略，规划和落地用户生命周期运营、活动运营的自动化SOP策略；
- 基于AI智能体，实现对会员个性化的精细营销策略和触达，以进一步提升营销智能化程度的同时，提升会员营销转化效果。

客户痛点

- **清晰的洞察用户：**历史数据分散，无法高效回收数据，从而精准洞察用户画像；
- **精准营销：**历史营销链路长、效率低、数据回收难；
- **提升营销效果：**因为无法清晰洞察用户，沉淀“有效策略”困难，落地的营销动作效果不明显：
  - 亿万费用投入难见水花
  - 有效策略少活动靠运气
  - 轰炸式触达消费者不买单

项目方案

◆MAGIC智能营销执行亮点

- **全域数据接入打通：**公域（抖音+外卖）+ 私域数据（门店、小程序）接入1.2亿会员，并打通各渠道用户身份，让用户在各渠道的数据可汇总，并用于发掘需求（Mine）阶段的分析画像以及运营策略制定；

- **精准营销：**通过MAGIC各环节的打通，打通会员、优惠券、积分、企微、公众号、小程序等用户触达触点渠道，发掘需求（Mine）、编排旅程（Architect）、生成内容（Generate）、互动触达（Interact）各环节智能执行，让运营在一个系统即可完成，并完成实时数据回收核查复盘（Check）。上线半年至今，触达5700万人、23亿人次、4000+条策略；

- **提升营销效果：**绝味和腾讯企点共创中国零售连锁首个AI会员智体，全自动一人一面精准营销。AI会员智能体系统内含人群圈选、人券匹配、人货匹配、内容生成（叠加腾讯生态数据模型）、旅程编排等子Agent，最终串联成落地运营策略，并自动执行。

落地效果

全国首家通过大模型跑通营销业务全流程闭环，采用全AI组VS人工组进行对比，效果提升显著：

消费者更感兴趣——触达-点击率：1.8倍

消费者更愿买单——触达-点击-支付转化率：2.4倍

企业营收更佳——触达-点击-支付金额：3.1倍

用户资产留存好——企微好友删除率：降低47%

金融：东吴人寿 智能体技术助力保险全周期服务体系智能化升级

项目背景及痛点

东吴人寿保险股份有限公司成立于2012年，是国内第一家在地级城市设立的全国性寿险公司，业务范围覆盖人寿、健康、意外险等保险产品。随着东吴人寿近年来业务覆盖面的扩大与产品的普惠升级，传统保险业务流程已难以满足日益增长的业务。投保、核保、理赔等核心业务流程仍大量依赖人工，导致处理效率低下、响应缓慢，无法满足客户对时效性的要求，同时也带来较高的运营成本。

东吴人寿积极推进保险业务数字化转型，致力于通过智能化技术为客户创造更优质的服务体验，同时提升企业运营效率，以满足客户对高效、精准金融服务的需求。

项目方案

东吴人寿通过建设AI能力平台，推动智能体应用在赋能客户服务、提升员工效能等关键场景的落地，并不断迭代优化以强化价值输出。为支持这一转型，腾讯云助力东吴人寿建设公司级统一的大模型智能体开发平台，并探索在更多业务场景中的应用。

- 智能体开发平台部署：建设公司的统一智能体开发平台“东吴天枢大模型知识引擎”。采用私有化与云上环境混合部署模式，构建RAG、工作流等多种模式的业务智能体应用。私有化环境主要支撑数据安全合规要求高的业务场景，而云上环境支撑数据可公开、数据量大、并发高的业务场景。

- 业务应用场景的搭建：“东吴天枢”平台支持东吴智脑问+系列（问综合、问产品、问数据和问制度）、苏惠保智能理赔、智能知识库系统、东吴数据分析平台等业务场景的上线，实现多业务场景的智能化升级。

落地效果

东吴人寿以“东吴天枢大模型知识引擎”为底座搭建AI能力中心，并将其能力标准化封装为API，深度嵌入关键业务流程，实现业务流程再造与价值链重构，形成一系列东吴特色的智能体矩阵，为业务运营各个环节赋能。

东吴人寿建设完成东吴智脑“问系列”作为企业内部统一知识平台，已支持覆盖16大业务主体域共计3万+内外部文档的知识智能问答，以及300多项关键业务数据项的智能化问数。

- 问综合：智能中枢平台，集成“问系列”各领域能力，实现跨领域信息的智能整合与精准作答；

- 问产品：集成公司内外部近万款产品的产品说明书、条款、费率表、现金价值表等海量信息，支持秒级精准查询与提取；

- 问制度：汇聚公司内控制度、外部监管制度及合规条款，提供全量制度库的多维检索与合规要点的精准解读；

• 问数据：支持全业务口径数据查询，助力用户一手掌握多维数据。

东吴人寿“苏惠保”智能快赔助手通过智能对话方式为客户提供理赔咨询、理赔报案、进度查询与结论解读服务。

理赔咨询服务涵盖保障范围、理赔规则、报案流程等内容，结合DeepSeek V3，支持用户在对话中提出问题并即时获取精准回应。通过大模型语义理解能力，系统可自动解析问题意图并返回清晰解答，显著提升自助服务水平，降低人工客服负担。

理赔报告服务由东吴天枢AI智能体开发平台的工作流引擎驱动，支持客户通过自然语言的方式提交理赔申请。系统自动识别报案意图并提取职业、地址、联系方式等关键信息，生成标准化数据并进入审核流程，实现与保单信息系统、用户信息库等后端系统的自动集成，显著减少人工操作与信息遗漏。

进度查询与结果解读服务支持理赔进度的实时查询，与此同时系统基于医保结算明细与理赔结论，结合DeepSeek R1推理能力自动生成结构化、可解释的结果说明。客户可多轮对话追问，系统结合历史对话与知识库内容，生成个性化的回应，进一步提升透明度与满意度。

借助“苏惠保”智能快赔助手，东吴人寿将该业务理赔处理时效从传统人工审核的3-5天缩短至3分钟内，最快42秒即完成自动理赔，大幅提升用户服务效率。“苏惠保”智能快赔助手已覆盖230万的参保人，预计支撑每年上万案件的自动化理赔，可节省大量的人工成本。

互联网：同程DeepTrip智能助手 用AI重新定义旅行体验

项目背景及痛点

同程旅行作为中国领先的在线旅行服务平台，业务覆盖机票、酒店、火车票、度假预订等众多场景。为帮助用户更好地规划旅行行程，腾讯云助力同程打造的产业级智能体DeepTrip能够通过自然语言描述需求，自动生成行程建议。但面对海量且复杂的用户咨询，需确保每次用户点击时，模型都能快速响应，以及确保模型能支持节假日、活动期等高峰流量的冲击。另外，内部员工（如客服、运营人员）需要快速查询产品信息、政策法规等，但知识分散在不同系统和文档中，检索耗时费力，影响决策效率和客户响应速度。

项目方案

原型验证期：快速迭代，集成DeepSeek最新能力

项目初期，腾讯云紧跟DeepSeek模型迭代节奏，为同程提供全套API接入与功能测试支持，包括算法升级带来的响应提速与新增模块的集成落地，协助完成验证、适配和上线，确保其产业级智能体DeepTrip始终保持行业前沿的能力状态。

产品开发期：算力支持+API调用，实现降本增效

在智能体开发阶段，腾讯云为同程提供了算力支持+模型API服务的双保障。比如针对训练环节，腾讯云提供了高性能GPU资源，支撑DeepTrip对定制模型的训练需求；调用多年沉淀的网络通信优化技术（TRMT），帮助DeepTrip提升多机多卡训练效率，让模型能更快上线新能力；同时，对于模型API的接入服务方面，腾讯云针对API服务构建了整套性能优化体系。包括：自研的API网关，能同时处理海量用户请求；动态限流机制，能智能应对流量激增等情况；Angel加速套件，对模型推理做了深度优化，在保证吐字速度的同时，大幅提升响应速度和调用次数。

上线部署期：弹性扩缩容+专属资源池，保障业务稳定

腾讯云为同程配备弹性算力池以及专属资源池，根据访问量动态扩缩容，自动调度资源，保证用户量剧增的时间段，也能让DeepTrip系统不卡顿以解决高峰流量的冲击，确保用户点击时模型的快速响应。

基于腾讯云智能体开发平台，同程不仅有企业级智能体DeepTrip的落地，同时其企业问答助手已服务同程财务、工程效能等部门。

落地效果

腾讯云助力同程打造的智能旅行助手DeepTrip上线，率先在旅游行业跑通了垂类Agent的开发路径。用户仅需用自然语言描述需求，AI就能自动生成行程建议，整合酒店、交通、美食等内容，从推荐到预订实现闭环。这不仅是一次面向C端用户的体验升级，更是一次面向B端的技术验证：DeepTrip不仅能应对高频访问、复杂调用和实时交互等挑战，还能深入嵌入业务流程，成为真正可用、可持续的生产力工具。

互联网：从“机械应答”到“金牌销售”
驾校通用智能体重构营销客服转化链路

项目背景及痛点

驾校通是58驾考旗下“驾校一点通”专为驾校打造的一站式管理平台APP。运用前沿信息技术与互联网理念，对驾校运营中的招生、营销、培训、车辆、教练、考试、财务及报表等全业务环节，进行一体化信息管理。平台联动智慧学车硬件设备，打通软硬件数据，实现从招生、服务到管理的全流程数字化解决方案，通过一套系统流转学车全流程，助力驾校构建智慧驾培生态圈，提升运营效率与管理水平。

为提升招生客服的工作效率和营销转化，驾校通打造了基于知识库的营销客服问答系统，为营销客服提供 AI 辅助答案生成、智能回复等 AI 能力。但是由于知识库中的营销客服对话记录标准不一、话术混乱，系统知识库内容碎片化，系统检索召回精准度低，导致生成内容效果不佳；另一方面，基于Q&A问答对召回的模型回复机械，缺乏金牌销售所具备的主动营销意识和销售技巧，在用户报名引导、深层需求挖掘、决策犹豫应对等复杂场景中无法有效提供帮助，致使高价值的营销转化环节依高度依赖人工坐席，营销效率提升有限。

项目方案

五八驾校通基于腾讯云智能体开发平台，从知识库优化与检索机制升级两大核心层面构建智能营销客服系统。

在知识库优化方面，对原始金牌销售对话记录进行系统化数据清洗与结构化处理，过滤无效内容并将冗长文档拆分为独立、完整的用户对话单元，形成标准统一的金牌销售对话知识库；同步通过API动态接入优惠信息等可变参数，保持信息的时效性与灵活性，为精准检索奠定高质量数据基础。

在检索机制升级方面，升级RAG增强检索策略，采用 chathistory+userquery 的组合检索策略，增强了角色整体设定，保持人设和原对话信息一致，紧密贴合真实对话场景，显著提升话术召回的准确率与匹配度；同时结合提示词工程的深度优化，模型在答疑之外更能主动模仿金牌销售的话术风格和销售技巧，实现自然流畅且转化导向的智能服务。

目前，升级后的智能营销客服系统已经在驾校通营销客服团队全量上线使用，为营销客服团队提供包括问答答疑、客户关系维护、报名引导等智能问答服务，助力团队工作效率和客户转化率提升。

落地效果

五八驾校通智能客服系统上线后，取得了显著的应用成效。系统能够从知识库中精准召回与用户咨询场景高度匹配的对话，极大提升了上下文的关联性和应答针对性。同时可深度参考召回内容，主动运用营销与逼单话术，有效引导用户完成报名转化，实现了从“被动答疑”到“主动营销”的服务升级。

目前，该系统已成为内部营销客服团队的核心工具，显著提升了销售效率和用户转化率，为驾校行业的智能化服务转型树立了标杆。

教育：从“能解答”到“优解答”
考试宝以AI大模型解锁精准学习新范式

项目背景及痛点

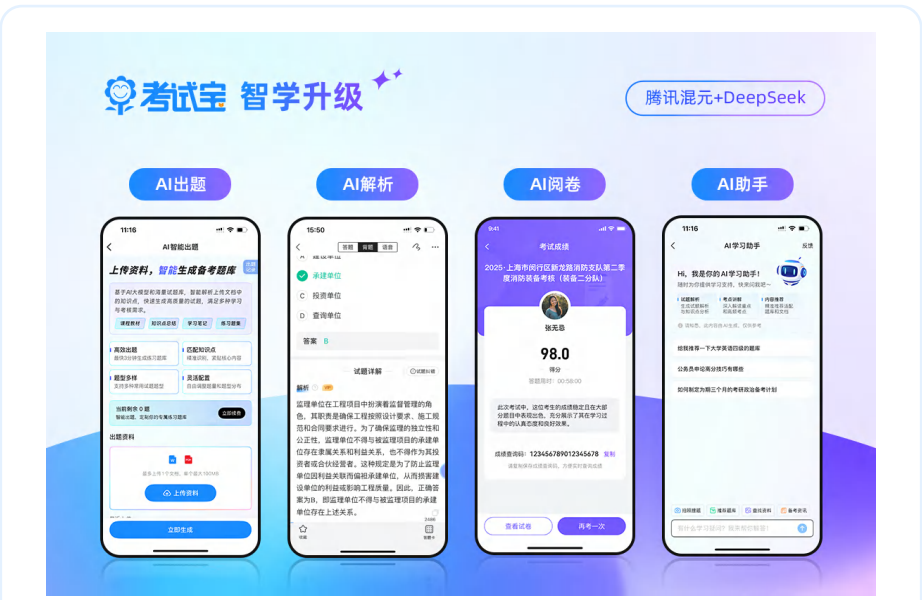
考试宝创立于2019年，是一个服务于职业教育的综合性学习平台。平台题库覆盖10000+考试类目，试题总量超过60亿道，涵盖建筑工程、安全生产、特种作业、考公考编、医药卫生等职业考试全领域，用户总数超过1亿。

项目方案

2024年起，考试宝与腾讯云达成战略合作，借助腾讯混元大模型和混元turbo模型及多项AI原子能力，搭建起AI基础设施，实现为用户提供学-练-考-评全流程智能化服务。

考试宝基于“腾讯混元+DeepSeek”打造了AI助手，突破时间空间限制，随时随地帮助用户解决复杂难题。无论是微积分极限求解、低压电工电力分析，还是机械设计公差计算，输入题干立即获得分步解析、逻辑拆解和知识溯源，通过展示推导过程、解题思路思维导图及关联教材章节与考点频率，提高用户学习效率。AI助手随时随地为用户提供学习助力。考试宝智能阅卷功能，借助AI大模型可自动完成考试阅卷评分，凭借强大的语义理解能力，逐题分析考生答案，精准判断对错，确保评分标准统一、精准公平。

此外，考试宝积极借助腾讯云智能体开发平台底层能力，在学习全流程展开多元探索。在智能出题环节，大模型依据知识图谱与教学目标，自动生成涵盖不同难度层次、题型多样的试题。既可模拟真实考试场景，又能满足学生个性化练习需求，提升其解题能力。在知识图谱构建方面，大模型发挥强大语义理解能力，深度剖析海量学习资料，精准提炼知识点，构建出逻辑清晰、关联紧密的知识图谱。用户借此能直观把握知识脉络，实现高效学习。辅助记忆上，大模型化身“智能导师”，依据学生的学习进度与遗忘规律，提供个性化记忆方案。通过智能提醒、趣味记忆游戏等方式，帮助学生巩固知识，增强记忆效果。



落地效果

在“腾讯混元+DeepSeek”双模助力下，考试宝在题目解析效率、解析质量及成本控制方面实现突破。基于腾讯混元AI能力及DeepSeek长思维链推理能力，考试宝实现单小时可生成20万道解析，专业领域题库知识点精准拆解及步骤可视化呈现，AI解析使单题处理成本直降90%。在内容质量方面，考试宝增加复杂计算题解析模块，支持多步推导验证，升级图文混合题理解功能，工程制图及财务报表等解析不再是难题；语义理解精度也显著提升，长题干逻辑拆解准确率达到90%。从"能解答"到"优解答"， AI不仅给出答案，更提供多种解题思路和知识关联推荐，助力考生举一反三。

未来，考试宝将继续携手腾讯云，不断升级合作场景，深度挖掘双方技术优势与资源潜力。依托腾讯云强大的云计算、大数据及人工智能技术，考试宝将进一步优化平台性能，实现海量试题数据的快速处理与精准分析，为学习者提供全方位、智能化的学习支持，助力他们在知识海洋中乘风破浪，实现高效学习与成长。

政务：邯郸公积金全国首创“边聊边办”数字柜台 重塑公积金服务新体验

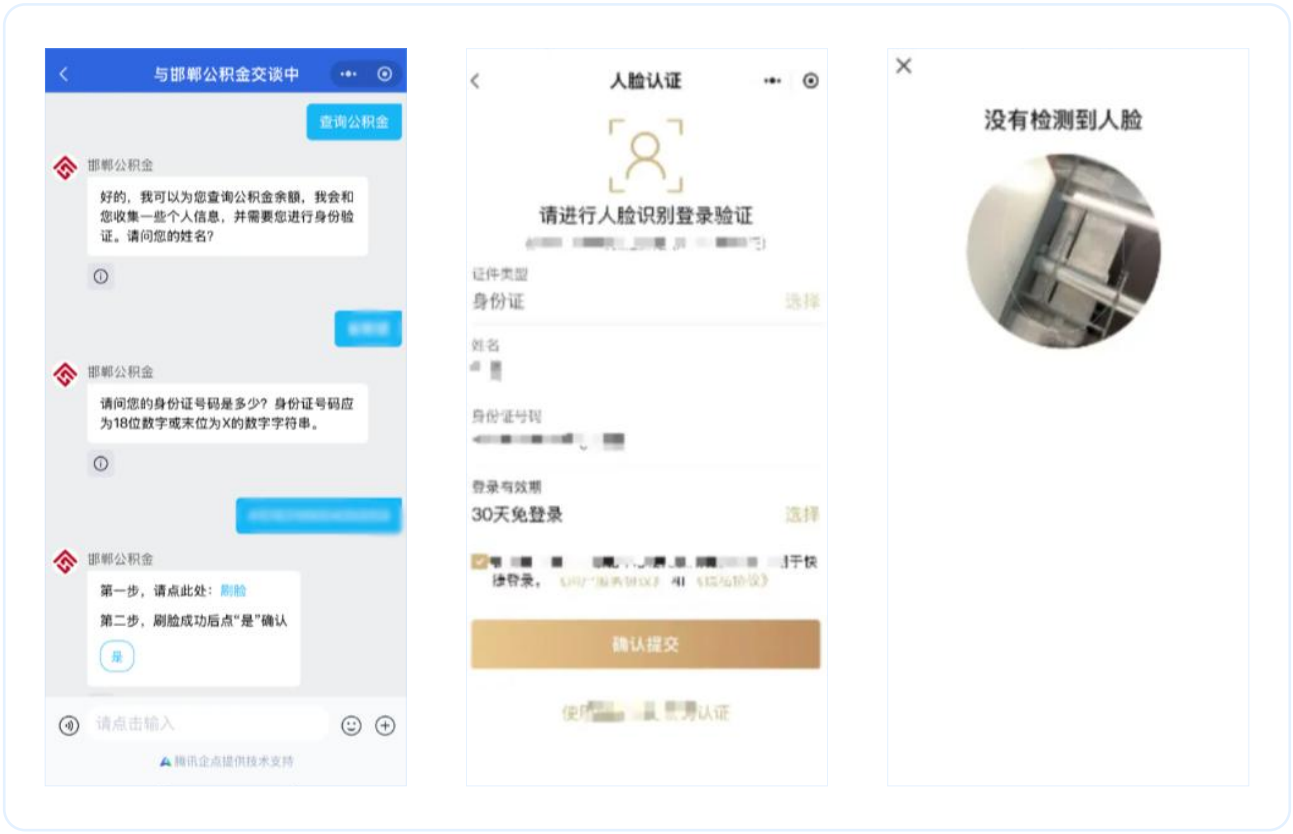
项目背景及痛点

邯郸作为人口大市，每年超60万人次办理公积金提取，业务需求量大且服务场景复杂。随着数字化政务服务的深入推进，传统的线上线下服务模式已难以满足群众对高效便捷服务的期待。线下窗口服务资源有限，市民需要排队、填表、跑腿来完成业务办理，所花时间长、体验差；线上移动端操作流程复杂，从下载App到填表上传资料，对于老年群体来说流程长且门槛高。

项目方案

通过腾讯云智能体开发平台，邯郸公积金在微信公众号里打造全国首创“边聊边办”的“数字柜台”，搭建一整套属于公积金的业务工作流，率先把智能体工作流应用到政务办事场景。AI基于与用户对话，引导操作，自动调用不同身份，验证协议签署等环节，无需来回切换界面，一站式完成全流程服务。这种办理方式充分发挥出了大模型在政务服务领域的价值，让服务更有温度，更有智慧。

“数字柜台”系统省去了App繁重的开发、更新和维护成本，一次开发即可连接到众多业务场景，同步迭代更新，且仅需低成本维护，同步多部门的数据业务。目前已有多项高频公积金业务陆续接入，覆盖退休提取、租房提取、失业提取、贷款申请等多个使用场景，未来预计将在医保、人社、税收等领域不断创新应用，提高民生服务效率。



落地效果

数字柜台上线以后，公积金业务办理效率显著提升，以退休提取业务为例，原本需要15分钟的办理流程，如今缩短至3分钟完成，提速超过80%。正式运行以来“数字柜台”受理职工咨询82079次，完成退休提取边聊边办102笔，偿还公积金贷款提取边聊边办11笔，网厅业务远程帮办487笔，查询公积金账户信息49875人次，深入推动住房公积金业务从“能办”向“好办、智办”跨越式发展。

邯郸公积金在智能体领域的探索应用，不仅为国内公积金服务领域的AI应用创新实践起到重要带动作用，也为教育、医疗等各个行业采用基于大模型的工作流，构建“聊”与“办”融为一体的服务模式提供了可复制的标杆案例。这一探索充分展现了智能工作流在政务服务领域的应用价值，通过“对话即服务”的新模式，实现了服务效率与用户体验的双重提升。

政务：腾讯云助力深圳市政数局人工智能基础平台加速政务大模型应用落地

项目背景及痛点

深圳市政数局作为深圳市政府的重要组成部门，统筹负责全市数字政府建设、政务服务优化、数据资源管理及智慧城市发展的规划与实施。当前，大模型技术为政府治理和公共服务数字化转型升级带来重要机遇，然而，在实际落地过程中，仍面临诸多系统性挑战：

多数业务部门虽然意识到大模型的应用潜力，却普遍缺乏专业开发能力和清晰的建设路径；外部技术供应商亦难以提供集约、可靠的一体化解决方案，致使很多优质场景迟迟难以启动。从模型选型、资源申请、精调优化到最终部署上线，整个流程周期过长，无法敏捷响应快速变化的业务需求，制约了政务智能化的落地速度。此外，由于缺乏统一的建设标准与资源共享机制，各部门在模型、算法、算力及数据层面重复投入、孤立建设，不仅资源利用效率低下，也难以实现集约化管理和规模化复用，极大限制了大模型在政务场景中的推广深度与应用广度。

项目方案

为系统化解以上难题，深圳市政数局与腾讯云等企业携手，共同构建了集约化、敏捷开放的人工智能基础支撑平台。平台通过“1门户+1中心+1中台+1体系”的人工智能应用支撑能力，为政数局统筹、规范全市AI应用建设和归集管理AI数据资产提供抓手，实现对全市AI政务应用建设的“全面赋能、协同集约、标准规范、安全可控”。

其中，中台能力主要依托于腾讯TI平台、智能体开发平台、内容安全等产品进行构建，提供了包括模型精调、知识增强、提示工程、内容安全等在内的全方位工具集，帮助政府部门实现场景需求与模型效能的最佳匹配，同时兼顾建设成果与投入成本的整体优化。腾讯云TI平台，实现异构算力资源的统一纳管、对多种开源、商业及行业模型的统一部署，显著降低模型使用门槛和提升资源利用率。腾讯云智能体开发平台，支撑便捷接入市语料库及各部门自有的知识库内容，可快速构建高质量、低幻觉的政务行业知识体系，为各类上层应用提供可靠的知识支撑，成为基于大语言模型的知识应用构建中枢。 在大模型的使用过程中，腾讯云内容安全产品为业务的合规、安全提供全方位的保障。

落地效果

平台上线后，在多方面取得显著成效。首先，实现了全市AI场景从申报、评估、备案到清理的全流程数字化管理，并引入AI技术自动识别和辅助评估可智能化场景，极大增强了政数局在资源统筹、效能管理和科学决策方面的能力。其次，实现对异构算力的省心管理、模型资源的统一部署、场景开发的多模型支持，实现全市算力及模型资源的集约化管理。最后，通过提供模型服务、知识库管理、提示词工程等一站式基础能力，原本复杂的大模型知识应用构建流程被极大简化，仅需四步操作、五分钟即可快速搭建出可用原型，极大降低了开发门槛和时间成本，全面提升了“AI+政务”应用的开发效率与标准化水平，为深圳市政务服务的智能化升级奠定了坚实基础，也为全国数字政府建设提供了可复制、可推广的先进经验。

政务：AI赋能 智治惠民——宝安区打造政务大模型应用新标杆

项目背景及痛点

深圳市宝安区信息中心作为全区新型智慧城市与数字政府建设的关键单位，承担着制定智慧城市发展规划、推进政务信息系统整合、构建全区大数据体系、保障政务网络安全等重要职责。该中心负责运维区政府各类网络与数据中心，指导智慧平台与电子政务应用体系建设，并持续推动政务大数据资源共享与新技术应用，为全区线上民生诉求平台提供技术支撑。

尽管宝安区在政务数字化方面已有扎实基础，但在人工智能技术的全面融合与应用方面仍面临系统性挑战。一方面，AI能力分散，缺乏统一平台支撑，从开发到部署流程繁琐、周期长，难以敏捷响应业务需求。大量政务数据未被充分转化为可复用的模型能力，制约了“人工智能+政务”的创新与生态构建。另一方面，政务场景高度复杂，政策咨询和民生诉求处理仍主要依赖人工，标准不一等体验差异；内部办公缺乏智能辅助，员工负担重、决策链条在信息整合与协同效率仍有提升空间。面对服务540万人口和94万商事主体的巨大压力，传统模式难以兼顾响应速度、人力成本与服务质量的平衡。

项目方案

为系统破解政务智能化难题，宝安区与腾讯云合作，以构建集约化、赋能型AI中台为核心，推动人工智能在政务领域的深度整合与全面创新。

项目首先基于腾讯云TI平台打造了全区统一的AI模型开发与服务中台，致力于实现AI资源的高效整合与全流程管理。该平台提供从模型训练、评估到部署监控的一体化支持，显著缩短开发周期，提升业务需求的响应速度。通过规范化、流程化的平台设计，不仅实现了资源的集约化使用，也使得不同业务部门能够共享通用AI能力，有效避免重复建设，为政务智能应用奠定坚实技术基础。

在此基础上，项目着力构建可共享、可复用的模型能力库与工具集。通过系统梳理各政务业务中的共性环节与高频需求，将通用能力封装为标准化的AI服务模块，支持跨场景的灵活组合与快速调用。这一机制不仅大幅提升模型利用率和部署效率，也显著降低了二次开发门槛。平台进一步引入持续学习机制，支持模型基于实际应用反馈不断优化迭代，确保能力库的持续进化与实用性的不断提升。

在赋能业务应用层面，项目重点推进“人工智能+政务”的融合创新，围绕政策服务、民生诉求、办公辅助等真实场景，开发了一系列轻量、高效、易用的智能应用。对外服务方面，落地了民生诉求智能问答系统，实现公众咨询的实时响应与精准指引；政策智能办理工具则支持条款解析、合规审查，大幅提升政务服务的透明度和效率。内部办公场景中，推出公文拟办自动生成系统，可快速完成发文前意见起草；政务知识一键检索与员工行政问答功能，为工作人员提供实时、准确的知识支持，有效减轻事务性负担。

通过平台、能力与应用的三层架构，项目不仅打通了从数据到智能的转化路径，也系统实现了政务流程的智能化重塑，为宝安区数字政府建设提供全面技术赋能。

落地效果

宝安区政务大模型项目在实践应用中取得多方面显著成效。政策问答智能应用每天为市民提供超过200次精准解答与办事指引，推动民生诉求平均处理时间从4.21天缩短至2.75天，公众满意度和信任感显著提升。内部办公场景中，系统日均处理超3000次民生诉求摘要生成任务，高效支撑“一件事一次办”服务改革；智能拟办意见功能日均协助生成超300次发文前意见，有效提升公文处理质量与效率。此外，基于政务微信的内部知识问答助手，为工作人员提供实时业务支持，进一步强化了行政协同与决策效能。

该项目荣获多项权威认可，包括入选2024世界人工智能大会，获评中国工业互联网研究院“AI赋能新型工业化创新应用优秀案例”，并夺得“AIC年度商业价值奖”与云鼎奖AI创新奖等荣誉，成为深圳市人工智能示范区建设的标杆案例。通过构建从模型层到应用层的一体化政务智能服务体系，宝安区不仅有效破解了传统政务服务的痛点，更打造了“边问边办、智治惠民”的数字政务新标杆，为全国政务智能化转型提供了可复制、可推广的宝贵经验。

制造：运达能源科技集团 以智能体技术提升风电装备制造和交付效能

项目背景及痛点

运达能源科技集团股份有限公司（股票代码：运达股份SZ.300772）作为中国风力发电领域的拓荒者、创新者、引领者，业务遍布全球五大洲，是全球领先的智慧能源技术解决方案供应商。运达能源科技集团坚持全产业链布局与全球化发展双轮驱动，业务生态涵盖六大板块，包含风电整机装备制造、新能源工程总承包、储能系统产品、新能源电站投资开发与运营、综合能源服务、新能源消纳。然而，随着业务规模持续扩大与项目复杂度不断提升，运达能源科技集团在施工现场装配效率面临一系列挑战。

运达能源科技集团在全国乃至海外设有多处制造与施工基地，装配人员在实际操作中常需查阅大量技术文档、图纸和工艺规范。传统支持方式主要依赖人工查询或通过电话、邮件联系总部专家，响应不够及时，也存在信息传递准确性不足的挑战，直接影响装配效率与质量。此外，资深专家资源覆盖范围有限，难以实时响应多地并发的技术咨询需求。

项目方案

基于腾讯云智能体开发平台，运达能源科技集团与腾讯云共同构建了一套覆盖企业内部运营与外部业务需求的智能应用系统。该系统以智能体开发平台为支撑，深度融合自然语言处理、知识库管理与风险规则引擎，有效服务从生产制造到商业决策的多类场景。

在施工现场装配支持方面，企业部署了装配支持智能体，面向全国制造基地与风电场项目现场人员提供实时、精准的知识问答与操作指导。该智能体整合了产品文档、工艺标准、历史案例等大量专业知识，能够通过自然语言快速理解现场提问，并输出可靠解答与关键步骤指引，显著减少了人员查找资料和等待支持的时间，提升了装配作业的效率与准确性，同时降低了因咨询延迟带来的隐性成本。

落地效果

运达能源科技集团智能体应用平台已于近期上线，现场装配人员通过施工装配智能体即可通过问答的形式快速获取技术指导，预期平均问题解决时间将大幅缩短，操作错误率将明显下降，整体装配效率将得到有效提升。企业也借此减少了对高端专家资源的重复依赖，实现咨询成本优化和人才资源高效配置。运达能源科技集团作为新能源领域的创新实践者，深入推进智能体技术在企业运营与生产制造环节的深度融合。通过构建以AI驱动的智能运营支持体系，公司实现了运营流程的系统化优化与资源协同效率的显著提升，不仅大幅降低了人力与物资成本，更增强了决策响应速度与精准性。依托智能体技术打造了行业领先的应用范式，形成可复制、可推广的解决方案。这一系列实践不仅为本企业高质量发展注入新动能，也为传统制造与能源行业提供了智能化转型路径的重要参考，彰显了科技赋能产业变革的引领作用。

能源：五环集团用 AI 重塑工程管理 赋能新质生产力

项目背景及痛点

中国化学五环公司成立于1958年，前身为化工部第四设计院，现隶属于中国化学工程集团有限公司，是一家拥有工程设计综合甲级资质的国家高新技术企业。公司业务涵盖技术研发、工程科技和实业运营，在氨加工、磷化工、煤化工、石油化工、天然气化工、新材料及工业环保等领域具备显著优势。六十余年来，累计完成大中型设计项目3000余项、EPC总承包项目300多项，其中境外项目200多项，荣获国家及省部级奖项超过320项。

尽管在工程技术与项目经验方面积淀深厚，五环公司在数字化转型过程中仍面临两大核心挑战：一是数据治理方面，大量业务指标依赖手工统计与线下报送，导致数据口径不一、效率低下，难以支撑实时、精准的管理决策；二是知识管理方面，作为典型的知识密集型企业，传统知识传递方式已无法适应大体量、高专业密度、强时效性的知识共享和创新需求，制约了组织能力的持续提升。

项目方案

五环公司与腾讯云携手合作，以应对工程管理中的复杂挑战，共同推动以人工智能为核心的智能化工程管理体系建设。该项目围绕数据治理与知识管理两大核心领域展开系统性升级，致力于打造高效、智能、可持续的企业运营支持平台。

在数据治理方面，项目从梳理核心业务主数据入手，逐步构建统一的企业级指标库，完成对公司全域数据资产的全面盘点与质量治理。通过建立清晰、分层的数据架构和完善的数据管理规范，显著提升了数据的完整性、准确性和可应用性，为后续各类智能应用奠定了可靠的数据基础。这一举措不仅增强了数据的可用性与易用性，也为实现数据驱动的精细化管理提供了坚实支撑。

在知识管理层面，项目依托腾讯乐享平台，深度融合AI助手技术，构建了集知识沉淀、智能检索、内容生成与员工培训于一体的AI知识管理平台。该平台整合了智能问答、自动摘要生成和多模态学习支持等功能，能够帮助员工快速、精准地获取技术知识和操作指引，大幅缩短信息查找和时间成本，提升解决问题的效率。AI智能问答系统不仅可实时响应工程与技术咨询，还能够作为辅助培训工具，帮助员工系统掌握化工等复杂领域的专业知识，从而有效提升组织内部知识流转与传承效果，全面加强企业的核心能力与创新活力。

该项目标志着五环公司在数字化转型道路上迈出关键一步，不仅初步建成AI运营支持体系，实现降本增效，也为流程工业领域提供了可复制、可扩展的智能化管理样板，展现出领先的技术融合与应用能力。

落地效果

五环公司在数据与知识两大维度取得显著成效，通过数据中台整合全域信息，构建覆盖财务、市场、生产等多领域的分析看板，实现包括项目利润率分析、核心风险提示等关键指标的动态管理，为工程项目全生命周期提供坚实的数据决策支持。知识管理方面，依托腾讯云智能体开发平台与腾讯乐享协同赋能，公司建成系统化企业知识库，涵盖近20个部门、2000余个分类和超过4万篇文档，通过精细化标签体系实现知识的结构化管理和高效检索。AI问答功能的深度应用，为员工提供7×24小时技术支持，显著提升知识获取与转化效率。尤为重要的是，系统成功将资深专家的隐性经验转化为可复用、可传承的组织知识资产，进一步强化了企业的核心竞争力与创新能力。

地产：碧桂园服务打造「一问」AI客服机器人 赋能员工效率跃升

项目背景及痛点

碧桂园服务是中国领先的综合物业服务商，业务覆盖住宅物业管理、社区增值、城市服务及商业运营等多板块。公司以物业管理为基石、社区增值为增长引擎、城市服务为战略蓝海，通过科技赋能与生态协同，持续巩固行业龙头地位，管理面积超11亿平方米，稳居行业综合实力TOP1，致力成为“国际领先的科技型综合服务集团”。

碧桂园服务作为服务行业企业，在AI热潮下正从数字化向AI转型，面临一线员工对几十个物业系统的操作问题及业务知识解答需求。传统人工客服存在响应不及时（非7×24小时）、成本高、服务质量不稳定等问题，需通过AI实现高效知识检索与问答。

项目方案

碧桂园服务使用腾讯云智能体开发平台，构建知识库与问答机器人，打造辅助物业员工的智能助手。

- 联动全集团各数字化团队，构建产品知识库，以弹窗挂件形式无缝嵌入各业务系统中（如BOSS物业收费系统、市场拓展系统、大管家系统、碧有单系统等），当前已接入30+个系统，并以此为基础，构建客服机器人，支持各区域数字化群内的答疑，秒级响应，支持7×24小时服务。

- 服务台升级为AI服务台，支持用户下工单时，由AI根据用户诉求进行智能回复或诉求处理，并支持通过多轮对话识别实现人工工单精准转接。最终通过用户评价，客服团队进行定期的知识纠错，持续迭代提升知识质量。

- 建设与发布五大领域专业知识库，典型场景包括公文制度检索、业务知识（如运营、增值、财务、市拓、通用）检索等，累计知识沉淀约8w+。

- 当前试点累计服务问答数3w+，覆盖用户3400+，未来计划扩展至5万员工；规划从内部员工服务延伸至规模效应提升。

架构图

落地效果

碧桂园服务「一问」AI客服机器人，全天候、秒响应秒回

从过去普通客服直接答疑，升级为AI训练师，依托AI知识库，打造懂系统、7*24小时在线的AI客服机器人在线答疑。

一问AI客服机器人试点成果

30000+
服务次数

3400+
服务人数

14.85s
回复处理平均耗时

20.13%
试点区域工单量降幅

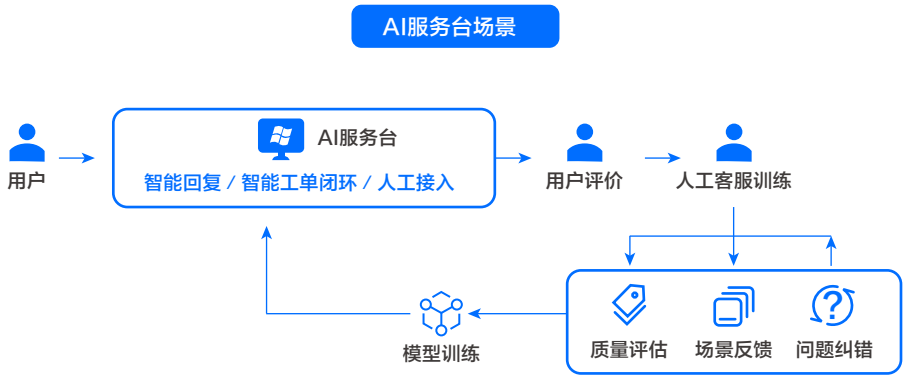
解决方案

	普通客服	AI训练师
工作内容	负责常规问题咨询 指引服务 工单处理 知识文档书写编辑	专注知识库优化与人工复核
服务效能	8小时服务 日均单量30单	7*24小时服务 日均单量300单
响应速度	单个咨询约10分钟响应 单个工单约30分钟响应	单个咨询‹5秒响应 单个工单‹5秒响应
优化提升	受限于抽样比例 存在滞后与覆盖不全的问题 对指导产品优化效果不明显	AI实时分析 100%覆盖全量对话数据

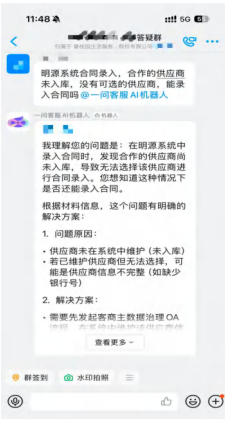
◆产品挂件

◆一问知识库

◆AI服务台



◆客服机器人



物流：DHL用智能体重构跨境智能客服 实现效率与合规双提升

项目背景及痛点

DHL成立于1969年，是全球领先的物流服务商，业务覆盖220多个国家和地区，提供国际快递、货运及供应链解决方案等专业服务，拥有全球最大的航空快递网络之一。国际快递企业客服长期面临复杂业务场景的严峻挑战，随着全球电商和跨境贸易的快速发展，DHL的客户咨询量持续攀升，庞大的信息量和复杂的业务分支不仅造成客服团队沉重的工作压力，也推高了运营成本。主要如下：

业务场景复杂，人工客服压力大：从跨国快递信息追踪到邮寄需求处理，DHL的国际快递业务涉及多环节、长流程的服务链条，包含跨国运输、禁运品查询、时效追踪、计费核算等。客服人员需掌握大量专业知识，其培训成本高，且高频重复问题占用了大量人力资源。

全球化与本土化的双重挑战：由于业务覆盖220多个国家和地区，在语言和文化差异方面，客服系统需支持数十种语言的实时交互，并精准适配不同地区的沟通习惯。同时，由于各地区法规与政策的差异性，客服系统必须确保所有回答都严格符合当地法律要求。

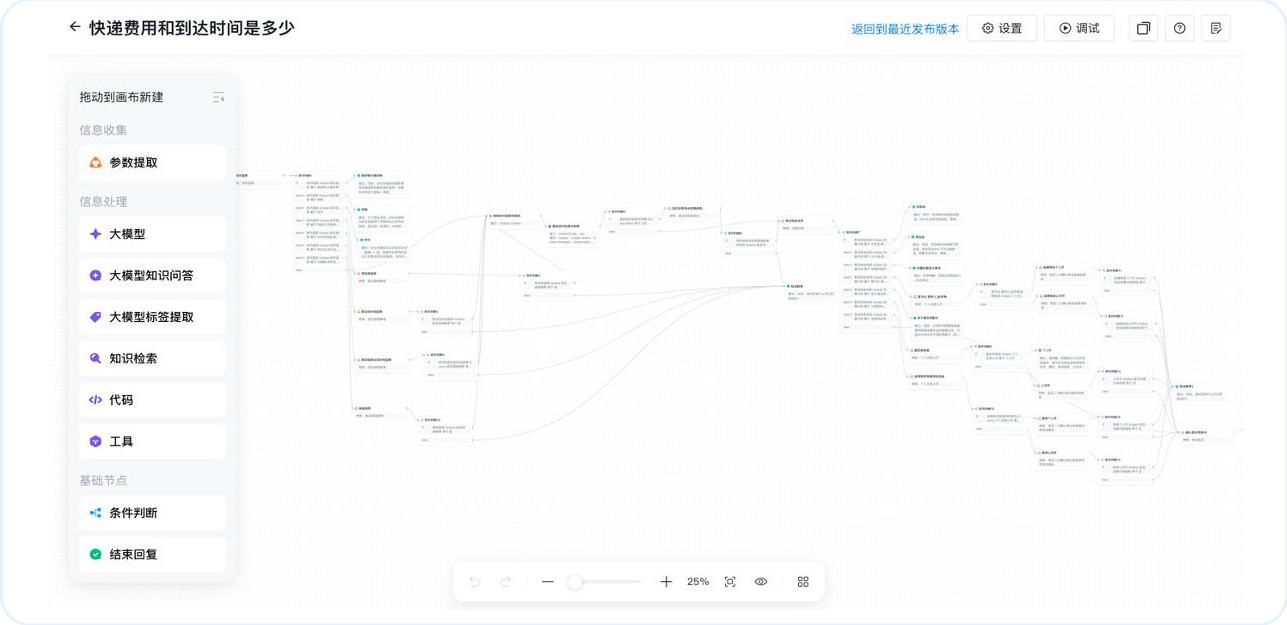
因此，在推进客服体系智能化升级过程中，企业需要构建统一的数字化平台，既要实现客户系统的本土化适配，又要满足全球化管理的协同需求，这对技术架构和运营模式都提出了双重考验。

项目方案

腾讯云智能体开发平台支持可视化工作流快速编排，助力国际快递企业DHL打造智能客服，自动化处理40+类复杂快递任务，节省人工坐席工作流，提升业务服务效率和质量。

通过工作流调用知识库，将知识库查询、多轮对话、业务规则等功能模块灵活组合，构建端到端的快递业务处理方案。在具体应用场景中，当客户咨询寄件服务时，系统能通过多轮对话收集寄件地址、收货地、包裹类型、联系方式等关键参数信息。

通过大模型、代码等高级节点，解决快递费计算等复杂业务场景，结合实时汇率、目的地关税政策、油费附加费等动态因素，生成精准的费用估算。在嵌入合规检查节点后，系统会自动识别禁运物品并提示相关法规要求，确保全球业务合规性。这种高度灵活的自动化工作流架构大幅提升了跨国物流服务的效率和准确性。



落地效果

通过接入腾讯云智能体开发平台，DHL将传统智能客服升级为「大模型客服」，实现了智能客服系统服务能力的全面升级。现已成功部署至企业小程序、官网、公众号等核心渠道，C侧用户通过对话形式进行问答，可帮助自动化处理快递寄件、收件、信息查询、关税计算等超过40类过程复杂且分支较多的任务场景，从而减轻人工客服压力，提升业务服务效率和质量。

升级后显著提升了运营效率，实现了人工维护的知识条数从超900条下降至119条问答，转人工客服绝对数减少200人次/天，机器人解决率从69%提升至74%。

互联网：巨人网络《太空杀》游戏

项目背景及痛点

《太空杀》是由巨人网络研发的一款手游，游戏中玩家从船员、内鬼、中立三个身份进行选择，使用技能赢得比赛。2025年上线AI原生玩法“残局对决”，构建了一个“真人玩家 vsAI智能体vs 真人玩家”的三方智能竞技格局。

游戏在发行上线阶段主要的安全挑战：

1、黑灰产打击频繁：不仅面临监管合规压力，更要直面黑灰产打击，包括核心业务的爆破扫描、漏洞利用、网络爬虫，cc攻击等等。

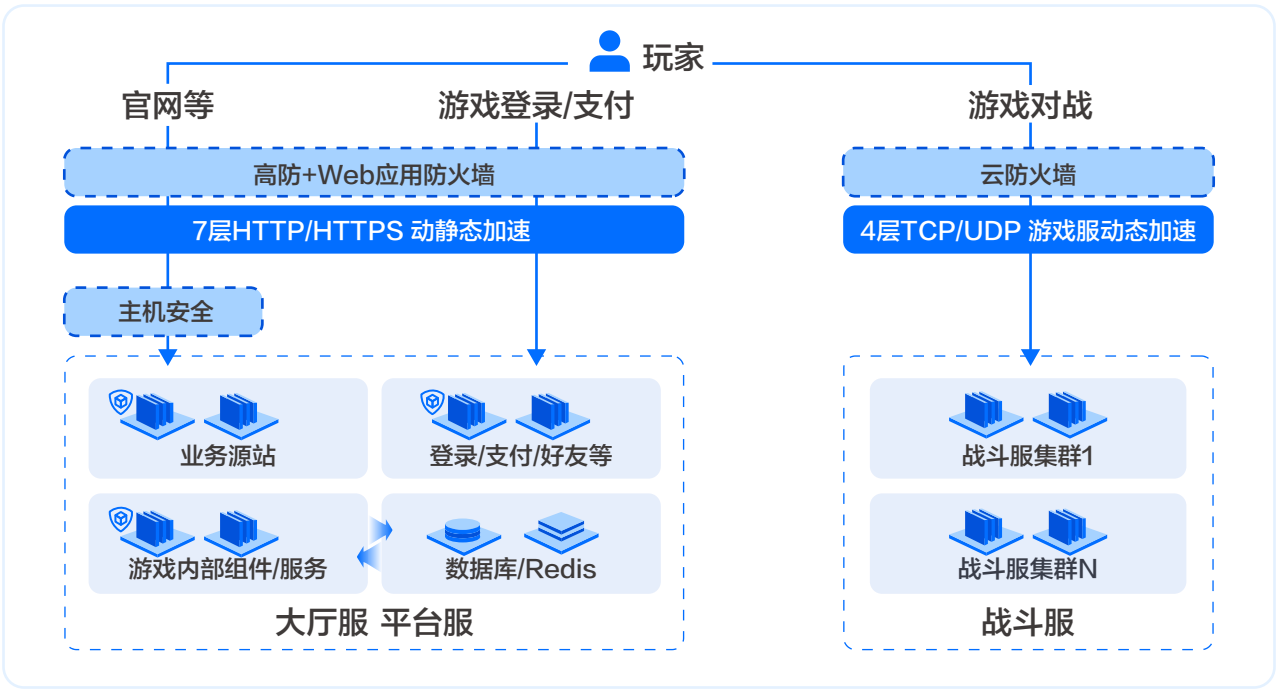
2、网络攻击易导致游戏服务器瘫痪：在发行上线阶段，则可能面临漏洞利用、APT等多种网络攻击，可能导致游戏服务器瘫痪，玩家无法正常登录等问题。

项目方案

1、DDos防护 ：通过高防包、高防IP和定制化策略等多重方式保障游戏稳定运行。

2、WAF防护：使用WAF保护太空杀上线后注册、充值等API接口不被CC攻击实时拦截拉新活动的恶意注册薅羊毛流量。

3、主机安全：客户公有云上全量使用主机安全产品，监控突破边界的入侵行为，有效的控制了病毒木马的植入，对游戏业务稳定起到至关重要的作用。



落地效果

通过体系化安全防护方案，助力太空杀发行、上线阶段稳定运行，成功抵御多次APT攻击。最终以全域联防保障游戏业务稳定运行，护航“AI”新竞技。

互联网：腾讯云助力心言集团打造AI情感陪伴服务重塑心理健康服务生态

项目背景及痛点

心言集团于2011年在北京成立，秉承“以科技服务心灵”的发展理念，经过十余年踔厉奋发，已逐步发展为AI泛心理与情感陪伴领域的创新引领者。2013年，集团正式推出明星产品——测测App（Cece App），为用户提供基于科学量化而又温暖有趣的AI情感陪伴产品及服务，致力于让广大用户“遇见更好的自己”。截至2025年6月，测测注册用户近5200万，签约咨询师超过26000名。

日益增长的用户需求和AIGC在泛心理服务各个场景的深入应用，以及心言集团本身正在推进的全球化进程，让测测App在技术架构和稳定运行上面临着新的挑战。

面对高速增长的用户需求、产品服务的快速迭代以及业务的全球化拓展，测测App亟需一种能够稳定、灵活、按需提供的云计算资源和服务解决方案，以确保高效利用并避免资源浪费和线路中断风险；

大模型训练和推理业务场景，对GPU算力的需求呈指数级增长，因此需要解决自研情感大模型“心元”训练和推理需要的大规模GPU算力问题等。

项目方案

腾讯云凭借全栈的云产品能力和丰富的客户服务经验为心言集团制定了针对性的解决方案，兼顾用云成本和效率：

- 针对用云成本和效率兼顾的问题，腾讯云通过TKE容器服务解决因需求快速变化带来的部署问题，通过TKE智能调度算法和qGPU技术助力客户提升资源利用率和降低成本。
- 为解决心言集团自研情感大模型“心元”对于GPU算力的需求缺口和成本压力，腾讯云提供了GPU资源共享技术和GPU算力在离线混部能力。
- 依托腾讯云星脉网络，提升数据传输效率，助力客户应对高并发的业务场景。

落地效果

- 通过TKE容器服务，测测App可在几分钟内完成部署，彻底解决业务需求快速迭代带来的部署挑战。相较于之前虚拟机部署方案，部署速度提升约50%。通过TKE智能调度算法和qGPU技术，资源利用率提升至90%以上，同时平均成本节约达20%，可轻松应对千万级用户的高并发访问。
- 在算力支撑方面，基于腾讯云异构计算产品有效满足测测持续增长的AI推理算力需求，实现训练成本的降低，提高训练速度。
- 基于腾讯云星脉网络提升数据传输效率40%，助力测测App实现如语音咨询、心理测评、社区互动交流等核心业务的稳定运行，支持千万用户高并发访问。

法律：得理科技打造AI法务助手 重塑企业法务服务新范式

项目背景及痛点

近年来，国家积极推动“人工智能+”行动，深圳“20+8”产业集群政策也将法律科技划入重点发展范畴，为行业智能化注入新动力。然而，众多中小企业在日常经营中仍面临法务资源匮乏、专业法律服务价格高昂、合规流程复杂等现实难题，市场上亟需一款既可靠又易用的智能法务产品。

另一方面，现有法律服务存在显著资源不均衡，很多企业对专业法律科技产品了解有限，常依赖通用型AI工具，难以应对实际业务中的复杂法律需求。法律科技领域竞争加剧，传统律所、科技公司纷纷入局。在这一背景下，得理科技与腾讯混元展开合作，打造具有显著差异化的产品，突破同质化竞争困局。

项目方案

得理科技与腾讯通力合作，充分发挥各自领域优势，联合推出新一代智能法务解决方案，涵盖技术、产品与生态三大层面：

技术深度融合：依托腾讯混元大模型强大的中文理解与多模态能力，结合得理多年积累的5亿+司法案例、法律法规与合同样本，联合训练出法律垂直领域大模型，确保专业知识与AI能力的有效融合。

产品联合创新：基于该大模型，双方共同发布“得理法搜企业版”，提供合同生成与审查、AI咨询、合规检测、文书辅助、律师函自动生成等六大核心功能，全面覆盖企业高频法务需求，实现“一站式”智能法务支持。

生态开放协同：得理法律咨询智能体全面接入腾讯元宝、腾讯输入法及微信生态，用户可通过自然对话方式便捷获取免费法律咨询、类案推送、律师匹配等服务，极大提升法律服务的可及性与使用体验。

落地效果

得理科技与腾讯混元的合作成果获得了法律界的高度评价，其显著降低了企业获取法律帮助的门槛并全面提升服务效率。在技术层面，“得理法搜企业版”实现合同审查95%的准确率，风险识别效率较传统方式提升80%，合同生成与审查达秒级响应，合规咨询响应时间缩短70%。有效帮助中小企业降低50%以上的法务支出，将单次律师函生成成本压缩至传统方式的三分之一，并通过AI实时监测政策动态，帮助企业规避潜在法律风险。此外，借助腾讯元宝、输入法及微信等多端生态接入，得理成功拓展服务覆盖范围，实现法律智能服务的规模化普及与深度赋能。

合同生成

一句话起草，秒出专业合同！

智能高效生成： 依托海量真实合同样本与行业标准模板，高效生成结构严谨、内容精准、格式规范的合同文本，大幅提升合同起草效率与准确性。

条款灵活定制： 支持多种合同类型与业务场景，自动生成关键条款（如付款、期限、违约、保密等），可根据企业实际需求灵活调整，确保合同文本合规合法。

免费体验

05

智能体发展展望

智能体将通过协同、感知和执行能力的不断增强，从单一、静态的应用走向动态、泛在化和具身化的新形态。这将是继互联网和移动互联网之后，又一次重塑社会和商业的范式变革。这一变革的驱动力主要源于三大核心方向的持续突破：一是智能体通过高效的协同网络，从“单兵作战”走向“群体智能”；其次是其通过多模态感知和世界模型，构建对世界的内在理解，实现更高级的推理；最后是其能力的泛在化与具身化，将智能从软件层面延伸至物理世界，并融入各类硬件终端。

智能协同：从单兵作战到群体智能

智能体的价值正从单个任务的执行，向形成高效的协同网络转变。这不仅包括多个智能体之间的分工协作，也包括跨越组织边界、实现全产业链的自动化流程对接。

多 Agent 协同：从工具到团队的价值跃迁

早期智能体多为“单兵作战”模式，依赖预设规则或单一模型，尽管在一些场景表现出色，但面对复杂、多步骤任务时能力受限。当前，行业正迅速向多智能体系统（Multi-Agent Systems）迈进。在这种“智能体团队”模式中，一个复杂任务被分解为多个子任务，并分配给拥有特定能力的智能体来协作完成。

多智能体系统通常采用两种主要架构：中心化网络和去中心化网络。在中心化网络中，一个“主管智能体”负责协调整个流程，分解任务并分配给专业智能体。例如，在软件开发流程中，一个协调智能体负责规划，而编程智能体负责代码生成，代码审查智能体则提供反馈。这种模式类似于人类项目经理，易于控制但存在单点故障风险。相比之下，去中心化网络中的每个智能体都具备一定程度的自主性，直接与其他智能体交流协作。这种模式更具鲁棒性和可扩展性，尤其适用于快速变化的环境。

多智能体协同的价值远超简单的任务分解。其核心在于实现“能力复合”和“情景自适应”。传统自动化依赖预定义的规则，而多智能体系统则通过自主规划和协作来应对问题。例如，一个擅长数据分析的智能体与另一个擅长内容创作的智能体协同，可以自主完成一份完整的报告。更重要的是，当面临未知情况时，多智能体能够通过彼此的沟通动态调整计划。这一能力使智能体从简单的“执行工具”升华为能够应对不确定性的“决策队友”。

跨组织智能体协同：重构产业链生态

智能体的协同能力正突破单一组织边界，迈向重构整个产业链生态的全新阶段。过去，企业间的协作主要依赖API、EDI（电子数据交换）等预设的、刚性的技术接口，本质上是“数据管道”，仅能处理标准化的、可预期的信息交换。一旦遇到供应链中断、需求激增等突发状况，协作模式便迅速退化为邮件、电话等低效的人工沟通。而跨组织智能体协同则是一种范式革命，其目标是打破传统的企业“信息孤岛”，将供应商、制造商、物流商和客户等多个实体连接成一个自组织、自适应的智能网络。在这种网络中，不同企业的智能体不再仅仅是交换数据，而是能够基于共同目标，进行实时的、动态的“协商”、“规划”与“资源再分配”，从而将产业链从脆弱的链条结构，升级为富有韧性的网络生态。

跨组织协同的价值在于重构整个产业的协作模式。它带来效率的显著优化，通过实时数据共享和动态规划，实现精准的库存管理、生产调度和运输物流，从而大幅降低运营成本。也会显著增强产业链的韧性，快速响应市场波动或供应链中断。

感知与推理：走向多维度的世界理解

随着多模态模型和世界模型的快速发展，未来的智能体将拥有更强的感知能力，能像人类一样理解和处理多模态信息，并构建对世界的内在认知。

多模态大模型：智能体的“大脑”升级

原生多模态模型的核心在于统一的架构设计，这推动了跨模态感知与生成能力的飞跃。这项技术不仅能理解复杂的数据，还能基于这些理解快速生成内容，从而极大地拓展了AI的应用边界。原生多模态能够整合来自不同模态的信息，如视觉、听觉和语言，从而对现实世界形成更全面、更细致的认知。这种深度的理解将为AI的规划、推理和决策能力提供更丰富的信息输入和上下文，并有望推动具身智能（Embodied AI）和视觉-语言-动作模型（VLA）的发展。随着这些技术的进步，未来的机器人将具备更强的感知、理解和执行能力，能够更自然地与物理世界互动。

原生多模态模型的高速推理能力将传统“等待-调整-再生成”的串行流程，转变为“所见即所得”的实时闭环体验。这种即时反馈的能力将成为推动下一轮商业创新的核心动力，并在多个行业中展现出巨大的潜力。

在个性化电商场景中，模型可以根据用户的浏览偏好、实时位置和天气等信息，即时生成定制化的穿搭推荐，显著提升用户的购买转化率。例如，腾讯混元图像2.0通过优化模型结构和推理加速技术，将通用图像生成时间压缩至极短，为实时生成个性化推荐提供了技术支持。在XR领域，结合混合现实头显和眼动追踪技术，原生多模态模型能够提供沉浸式的虚拟商品交互体验。用户只需通过简单的眼动或手势，就可以即时改变虚拟试穿商品的颜色或款式，从而大幅提升用户体验和下单率。

在游戏行业，内容创作正在引入即时生成技术，以实现“千人千面”的沉浸式体验。例如，腾讯混元游戏视觉生成平台就涵盖了游戏图像和视频模型能力，能够根据玩家需求快速将创意概念转化为高质量的图片、3D模型，并支持实时画布和AI 2D美术功能，极大地提升了游戏内容的生产效率和个性化水平。

世界模型：赋予智能体预测与规划能力

世界模型是一种能够对物理世界或特定环境的动态变化进行内部建模和预测的技术。其核心思想是，智能体不再仅仅被动响应输入，而是能够在行动前，先在“大脑”中模拟不同行动的后果，从而选择最佳策略。

典型的世界模型由感知模型、记忆模型和控制器组成。这一技术在自动驾驶、工业机器人等高风险领域具有关键应用价值。例如，自动驾驶汽车可以利用世界模型模拟交通动态和行人行为，在虚拟环境中进行大量“排练”以应对突发状况。这使得智能体能够进行高阶规划和策略制定，而不是仅遵循指令，是其从“工具”向“队友”转变的关键一步。

世界模型是智能体自主性、安全性和通用性的核心引擎。它赋予了智能体“知道”如果我这样做会发生什么的能力，这种对因果关系和未来状态的理解，使其能够进行高阶规划和策略制定。此外，通过在虚拟环境中进行大规模模拟和验证，世界模型为智能体的安全部署提供了强大的保障，这在自动驾驶和医疗等对安全性要求极高的领域尤为重要。

执行与应用：智能体的泛在化与具身化

智能体将不再局限于软件层面，而是将能力延伸到物理世界，开始具备了在物理世界中感知、理解、推理和行动的能力，并普及到各种硬件设备中。

具身智能：智能体能力的物理延伸

具身智能（Embodied AI）是指将智能体的感知、推理和规划能力与机器人等物理实体相结合，使其能够感知、理解并与物理环境互动，完成现实世界中的复杂任务。具身智能的兴起，旨在解决劳动力短缺等社会问题，并替代重复、危险或体力密集型任务。例如，腾讯Robotics X实验室推出了专为养老护理设计的“小五”机器人。它集成了混合移动能力（轮子+伸缩腿）和灵巧的触觉手，能够在复杂的家庭环境中安全导航和提供帮助。其AI算法使其能理解人类行为并预测需求，例如辅助老人站立或行走时动态调整支撑力。

具身智能的未来在于构建“智能体+硬件”的平台生态。英伟达等公司强调“仿真优先”的开发模式，通过构建物理精确的数字孪生环境，让智能体在虚拟世界中进行大规模、高效率的训练和测试。这降低了现实世界中训练的风险和成本，为具身智能的快速迭代和安全部署提供了基础设施保障。具身智能的普及也将直接影响劳动力的供需结构，并对养老护理、物流等社会服务领域产生深远影响。

端侧智能体与智能硬件：打造无处不在的智能

随着技术的不断进步，端侧智能体正在成为人工智能发展的新焦点。它们将不再局限于智能手机，而是会覆盖更多样的智能硬件，并与设备深度融合，从根本上改变人机交互的方式和软件生态。

通过模型蒸馏、量化、剪枝等技术，端侧大模型轻量化，原本庞大的语言和多模态模型正被成功“瘦身”，使其能够在资源有限的终端设备上高效运行。这使得智能体具备了更强的自然语言理解、逻辑推理和内容生成能力，交互体验从“指令式”向“对话式”跃升。高能效AI芯片（NPU）逐渐普及，从手机的旗舰芯片到汽车的智能座舱芯片，再到微型的可穿戴设备处理器，专门用于AI运算的NPU已成为标配。这为端侧智能体提供了在本地执行复杂推理任务的算力基础，摆脱了对云端的持续依赖，保证了低延迟和隐私安全。传感器技术的融合与进步帮助现代智能硬件集成了远超以往的传感器矩阵，如视觉、听觉、空间定位、生物特征等，端侧智能体能够低功耗、高效率地调用这些传感器数据，实时、全面地感知用户状态、行为意图和环境变化，这是其提供主动式、情景化服务的前提。

未来，端侧智能体的应用范围将从手机、PC这两个核心，迅速扩展至一个庞大的智能硬件生态，创造一个万物互联且智能的“泛在”环境。智能汽车将从“驾驶辅助”到“主动式智能座舱”演进，端侧智能体将成为汽车的“灵魂”，它不仅是语音助手，更是能融合车辆数据、环境感知和用户习惯的“智能副驾”；例如，当检测到用户略显疲态时，它会主动调整空调、播放提神音乐，并建议在下一个服务区休息；在通勤路上，它能根据实时路况和用户日程，自动规划最优路线，并在车上预先处理工作消息。智能穿戴设备将从“数据记录器”转换为“个人健康管家”。手表、手环、AR眼镜等设备将成为用户的“数字孪生”。端侧智能体能持续分析心率、血氧、睡眠等多维度生理数据，建立个人健康模型。它不再是被动记录，而是能主动发现异常趋势，并结合用户的日程和位置信息，做出智能判断和提醒。在AR眼镜上，智能体甚至

能做到实时翻译、导航指引和信息叠加，成为视觉的延伸。智能家居与机器人将进一步实现“全屋智能协同”。未来的智能家居将由一个中枢智能体统一调度，而非依赖用户逐一控制。这个智能体能识别家庭成员，理解他们的生活习惯。例如，当它感知到主人下班回家时，会自动调节灯光、空调至偏好状态，并让扫地机器人避开行进路线。对于家庭服务机器人，智能体将赋予其理解复杂指令并自主规划任务的能力，如“把客厅的玩具收到儿童房的收纳箱里”。

结语

我们正处在一个由智能体定义的全新技术浪潮的起点。这场变革的深远意义，不仅在于效率的指数级提升，更在于它从根本上重塑了生产力、组织形态乃至人机协作的底层范式。回望过去，信息技术革命将物理世界数字化；而今天，智能体则正在为数字世界赋予自主行动的“躯体”与“灵魂”，使其从被动的“信息容器”进化为主动的“价值创造者”。智能体不再仅仅是回答问题的“智囊”，更是执行任务、驱动流程、连接虚实的“数字员工”。

本报告通过构建能力分级体系与应用场景罗盘，力图为企业在这场变革中提供一张清晰的导航图。我们看到，尽管当前绝大多数智能体尚处于L1的初级阶段，但其向更高阶自主决策与执行能力演进的路径已然清晰。从解放重复性劳动的“高效助手”，到打通复杂业务流的“执行专家”，以及赋能战略决策的“决策专家”，企业可以依据自身业务痛点与数字化成熟度，找到最适合的切入点，循序渐进地释放智能体的价值。

当然，通向未来的道路并非坦途。高昂的训推成本、模型的幻觉问题、系统性的安全风险、深层的数据治理以及与现有业务流程的耦合难题，都是企业在落地智能体时必须正视的挑战。然而，这些挑战绝非不可逾越的天堑，而是技术成熟过程中必然经历的工程难题。通过训推一体的算力调度、端到端的模型精调、多层次的安全防护以及企业级知识底座的构建，我们正在逐步攻克这些难题，为智能体的大规模应用铺平道路。

展望未来，智能体的发展将呈现协同化、具身化与泛在化三大不可逆转的趋势。多智能体系统将以“虚拟团队”的形式，实现超越人类组织效率的复杂任务协作；具身智能将打破数字与物理世界的壁垒，让智能体走进工厂、家庭与城市；而端侧智能的普及，则会让个性化的智能服务无缝融入我们身边的每一个设备，成为无处不在的“环境智能”。

智能体的时代已经到来。它所开启的，不仅仅是一场技术的革命，更是一次关于未来工作与生活的深刻重构。对于每一家企业而言，这既是挑战，更是前所未有的机遇。现在，正是告别观望、主动布局的最佳时机。我们相信，那些勇于拥抱变革，将智能体深度融入自身战略、流程与文化的企业，终将在这场由智能驱动的未来竞争中，构筑起难以逾越的核心优势。


腾讯云
腾讯研究院
Tencent
Research Institute
Gartner