IDC | 百度智能云 智能时代基础设施

AI原生开启金融智能新未来

——金融行业大模型应用落地白皮书

目录

前言

从“感知推理”到“自主进化”，算法技术突破进入深水区。2025年，大模型算法的核心跃迁是从“被动处理任务”转向“主动进化策略”，金融行业作为数据密集型和计算密集型的典型应用场景，迎来了深度变革的历史机遇。全球头部玩家通过算法创新直接解决金融场景的“长文本、高实时、强专业”痛点。OpenAI GPT-5：强化“长文本因果推理”能力（支持10万token以上上下文），突破金融机构对“超长篇幅风控/投研文档”的处理瓶颈，Google Gemini 2.0：升级“多模态动态交互”算法，实现“文本-图表-数据”的实时联动——高盛用其构建“动态利率走势模型”。AlphaEvolve自主进化算法：通过“生成式策略优化（GSO）”实现模型自动迭代。国内，百度文心4.5和X1系列模型、DeepSeekV3\R1等大模型，正以多模态+长思维链推理+智能工具调用执行架构融合，实现“能思考、会落地”的大模型。金融行业拥有独特、高质量、大规模的行业数据，核心护城河已不再是“应用好某个开源模型”或“落地单一应用”，而是要构建“场景-算法-数据”的深度协同体系，构建知识壁垒+行业场景深度融合，训练出真正好用的Agent，实现核心业务场景AI原生化改造。

从技术尝鲜到价值优先，“核心业务与AI的融合深度”已成为金融机构的核心竞争力。大模型凭借对非结构化数据向量化处理以及强大的意图理解和推理能力，在面向员工的场景中优势明显，如知识问答、内容生成（金融报告辅助撰写）、智能办公（投研资料汇总）等；在面向客户的业务场景中，尤其是对深入业务应用场景（信贷、风控、营销）以及对实时性要求较高的场景（实时反欺诈、秒级授信）中，目前面临准确率较低、延迟反馈等问题。专精模型结合金融合规规则库、动态风险因子库，并通过领域数据定制与任务特定优化（如反欺诈模型的算法重构），即可实现深度场景适配。需明确的是，通用大模型在金融专业领域存在天然短板：意图理解不准确、专业知识覆盖不足、问答准确率、幻觉率、可解释性均未达到金融场景的“生产级要求”，IDC认为，未来，为满足金融业务多样化要求，在复杂语义理解领域，大模型将持续发挥优势，专业业务领域将由专精模型提供服务，通用模型与专精模型协同管理与适配的AI解决方案将成为主流。

从“技术组件”到“业务赋能平台”，大模型开发工具链降低金融机构的AI使用门槛。2025年，大模型工具的核心升级是“从‘技术导向’转向‘业务导向’”，通过低代码/无代码平台让金融机构快速构建“贴合自身业务的智能体”。金融机构对智能体进入核心业务等需求越来越旺盛，其在智能投研/投顾、信贷决策、风险管理等核心场景中将持续创造更多价值。持续迭代支持MCP/A2A的智能体开发平台，以及SFT工具链、场景优化工具链，以满足金融机构的场景创新应用需求。同时在模型管理方面，IDC指出，2025年“通用模型+专精模型”的协同管理将成为主流，工具平台的核心价值是“降低金融机构的AI使用门槛”——不再要求金融机构具备“顶级算法团队”，而是通过低代码、模块化工具，让业务人员也能“用AI解决业务问题”，工具从“技术组件”升级为“业务赋能的桥梁”。

从“数据驱动”向“知识驱动”跃迁，数据飞轮已成为金融应用AI原生化关键要素——唯有将零散数据转化为可复用的结构化知识，并形成“业务-数据-模型”的闭环，才能让大模型真正适配金融领域“高合规、高精准、高动态”的核心要求。金融机构的数据飞轮建设目标是通过打通全链路数据流，实现数据与业务的双向驱动。金融机构正在对金融数据按照敏感度分级，构建可信数据环境，满足合规要求；通过跨模态数据整合与关联分析，实现内外部数据协同，打破金融数据壁垒；构建高质量向量知识库、打造高价值知识工程与场景化数据沉淀来缓解高价值数据稀疏的现状。数据飞轮的构建将促进模型在知识广度、推理深度、领域专业性和鲁棒性等多维度实现系统性升级，使得金融智能系统能够快速响应业务需求变化。

从通算向智算演进，规模化异构算力管理已成为大幅提升算力效率的核心路径。随着大模型向GPT-5等万亿级参数演进，训练所需算力呈指数级增长，算力架构的“成本-能效平衡”能力愈发关键——以异构计算集群、多芯混合训练为代表的方案，因能兼顾高性能与低成本，已成为企业应对超大规模模型算力需求的核心竞争力。针对不同参数量级的模型场景，需精准适配算力方案，实现“算力资源与业务需求”的最优匹配：百亿参数模型场景：单机单卡即可完成推理与微调任务，是性价比最优的选择，而更高算力密度、更大显存的算力机器，则在模型训练微调场景与复杂推理中更具效率优势。而在千亿/万亿参数模型场景，DP（数据并行）+EP（专家并行）分离的大集群部署方案——通过将数据拆分与专家层分工解耦，可成倍数提升算力利用效率，是突破超大规模模型“算力瓶颈”的必选路径。

第一章

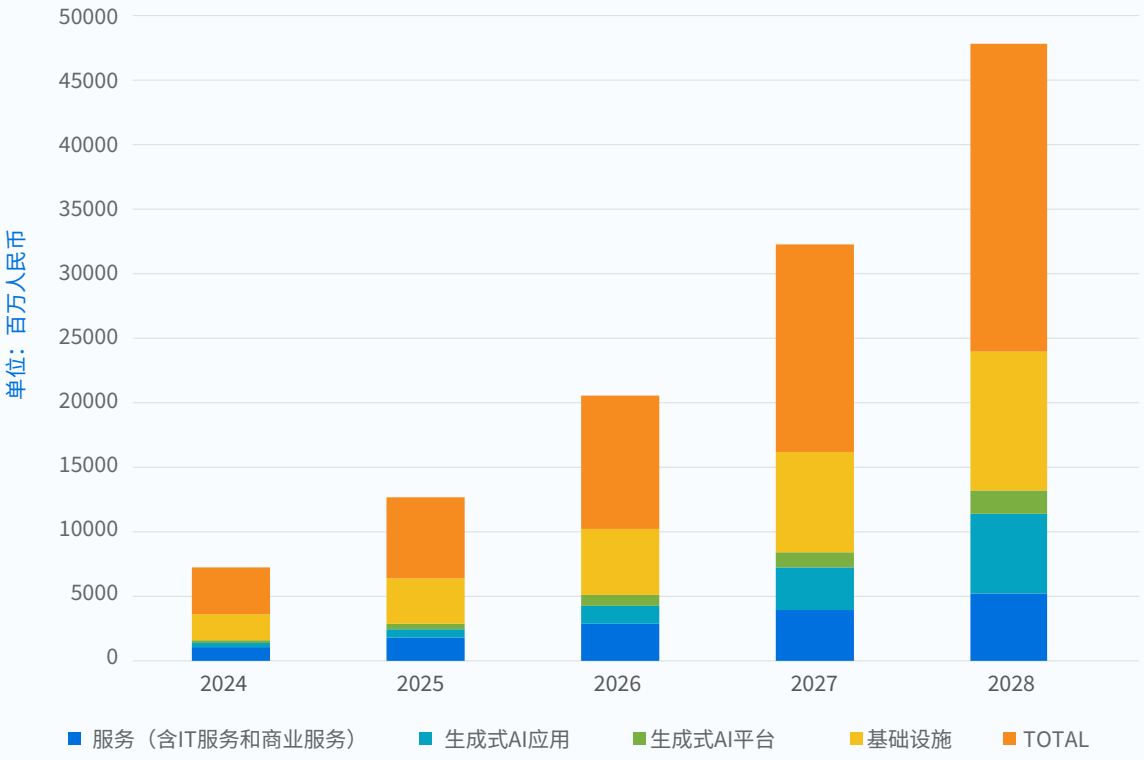
大模型开启金融行业全新智能时代

1.1 大模型驱动金融机构全面加速智能化转型

政策层面，我国已给出了明确的指导意见。2024年1月，金融监管总局等七部门联合印发《推动数字金融高质量发展行动方案》，提出布局先进高效算力体系，强化模型和算法风险管理；2024年12月，我国金融监督管理总局印发了《银行保险机构数据安全管理办法》，为银行保险机构规范数据处理、保障数据安全、促进数据开发、完善监管效能等方面提供了全面、细致的规范标准；2025年7月31日召开的国务院常务会议，审议通过了《关于深入实施“人工智能+”行动的意见》（以下简称《意见》）。“深入实施”标志着“人工智能+”行动正式从政策倡导迈入规模化、商业化落地阶段。

我国金融IT投入持续增加，根据IDC数据，2024年中国银行业IT投资规模达到1,693.15亿元，同比增长3.6%，预计在2028年将达到2,662.27亿元。2024年中国金融行业生成式AI投资规模为36.26亿元，预计到2028年投资规模为238.04亿元，增幅达到556.5%。

图1 2024-2028中国金融行业生成式AI投资规模预测



来源：IDC，2025

在智能时代背景下，我国各类金融机构均加码大模型投入，且各有侧重。国有大行以自主可控为核心，优先保障算力底座自主建设，兼顾千亿级模型再训练与多智能体协同；股份制银行平衡算力成本，侧重模型场景化微调；区域性银行关注低成本算力租用、复用，追求“开箱即用”，保险机构重点关注决策模型与大模型模型配合使用，提升核保核赔效率；证券与基金公司低时延交易与智能投研，侧重高性能算力与金融蒸馏模型，通过A2A与MCP协议构建生态。

表1 不同规模金融机构对大模型的投入偏好与应用策略

比较维度	国有大行	股份制银行	区域性银行	保险	证券/基金
典型代表	国有六大行	招商、中信、浦发、兴业等	城商行、农商行	人保财险、太保财险	券商（中信、华泰、国泰君安等）
核心诉求	自主可控底座	工具链优先	API调用开源模型	核保核赔准确度提升，大小模型配合	低时延交易
算力投入策略	自建智算中心	私有化部署算力，混合云	API直接租用银联或云厂商算力，或与总部共享算力	私有化部署算力，提升高可靠性	自建GPU小集群+混合云
参数规模偏好	千亿规模以上模型+再训练	30-70B中等规模	7-30B小模型	10-30B视觉、LLM模型	10B以上金融蒸馏模型
数据体系建设	全栈数据治理，场景导向的数据标准	数据加密与高质量数据标注	外部数据集，数据不出域	保单、票据多模态数据处理，核保知识图谱	投研数据融合，低时延交易数据
智能体关注	多智能体协同，与核心业务捆绑	复杂场景智能编排	预置智能体模板，快速上线，降低技术门槛	核保智能体定制开发	投研智能体

金融大模型开启了金融智能时代的新篇章。随着政策加码，金融大模型技术升级，应用场景的不断丰富，新旧智能时代转换的拐点将至，金融行业的全新智能时代将完成从“工具导向”到“超级生产力”的跨越。

1.2 强推理和多模态、多个模型深度配合与内外部协同的智能体推动金融走向智能化

IDC认为，“强推理+多模态”是当前人工智能技术发展的关键方向。仅仅“看懂”多模态数据并不足够，医疗和保险场景等复杂场景更需要较强的因果推理能力。例如：AI不仅要识别票据金额和项目，还要推断这些项目是否与患者诊断、治疗方案一致；在保险定损中，AI需要结合事故图像、维修价格体系、历史理赔数据，推理出最合理的赔付金额。“强推理+多模态”技术通过整合视觉、文本、空间等多维信息与高级逻辑推理能力，正深刻重构AI对物理世界的理解范式，从而满足其在复杂场景中的应用需求。

某保险公司——多模态+强推理辅助智能理赔

某保险公司推出基于多模态技术及强推理能力的“车险人伤智能定损机器人”，实现了伤情诊断与赔付标准的自动生成处理，仅需上传伤情照片与索赔材料，即可精准分析伤情、精准计算理赔金额，实现快速赔付，同时还可为伤者提供康复建议等人性化服务，极大提升了该类案件的理赔效率。自2024年3月正式上线启用以来，人伤智能定损机器人的单证分类及伤情识别准确率分别达到95.6%和88.3%。

“多个模型”深度配合是增强决策精准度、推动业务创新的关键。大模型适用于对语义理解和自然语言处理要求较高的场景，如智能客服、智能创作、智能营销等，提升深度推理与非结构化数据的处理效率；决策类的小模型专注于对结构化数据精准判别，在快速响应与细分专业场景中有天然优势。IDC认为，大小模型的深度配合，是满足金融机构对多样复杂场景中的模型应用需求、提升金融业务价值的重要方式。同时，通用模型与专精模型相互结合与灵活适配，也是降低模型运行成本，提升模型应用效果的重要策略。

某股份制银行——业务场景的AI化升级

某股份制银行在财富等业务场景中率先部署AI智能助手，通过大语言模型的知识理解能力与小模型的数据处理优势深度结合，实现了服务模式的智能化升级，能够深度理解客户口头表述中的潜在需求，例如当客户提到“希望稳健增值”时，AI助手不仅能识别风险偏好，还能结合市场行情自动生成包含国债、同业存单等低波动产品的配置方案。该应用显著提升了客户经理的服务效率，使专业财富规划服务得以覆盖更广泛的客群。

“内外部协同的智能体”将在复杂的金融业务场景中创造显著价值。内部智能体主要服务于金融机构内部运营，满足内部数据安全与合规要求；外部智能体聚焦零售与对公用户，为用户提供个性化服务，增强用户体验。内外部协同的智能体可以减少金融机构“内部业务闭环”与“外部生态联动”的割裂现象，通过内、外智能体的能力互补与流程协同，可以解决单一智能体难以覆盖复杂业务场景的痛点。

某国有银行——打造多智能体协同的智能研发体系

某国有银行通过强化大模型软件工程长思维链、动态决策和意图理解能力，建成具备需求理解与拆分、方案设计、代码生成、问题修复以及IDE工具调用、命令执行功能的研发垂直领域智能体群，各智能体通过分布式决策、调用路由、知识共享等机制相互协作，形成一支高效AI研发团队，实现AI程序员根据需求自主生成原型工程代码的能力，为金融业务的创新带来突破。该项目的落地使得团队单位时间编码效率提升约23%，月人均完成需求项（feature）增长30%，仅编码环节24年增效价值4069.9万元。

1.3 金融领域正加速迈向基于AI原生的智能重构阶段

金融行业正在经历从"工具赋能"向"智能重构"的战略转型，AI不再仅仅是提升效率的辅助工具，而是成为重构业务模式和生产关系的核心驱动力。

从用户需求来看，随着数字原生代成为主流客群和数字化渗透率的持续提升，用户对金融服务的期望发生了根本性转变。他们更加看重超个性化服务，期望获得一对一的个性化服务，而非标准化的产品推荐。而AI大模型在客服、产品推荐等场景的应用，显著提升了客户满意度与忠诚度。AI大模型能够理解用户特定场景下的金融需求，并提供恰如其分的支持。例如，当用户表达"想给自己买养老金"的需求时，AI能在几秒内生成相关方案。

从行业发展来看，IDC认为，随着AI大模型所带来的技术底座重构、交互方式变革等在行业中的深化，AI原生应用已成为金融科技演进的核心方向。与传统金融应用中简单嵌入AI功能不同，AI原生应用是从设计之初就以AI为核心驱动而构建的系统，其每个组件和交互流程都深度整合了人工智能能力，形成了自我演进、持续优化的生态体系。

广发证券通过易淘金APP的AI原生化升级，率先实现了从"综合交易服务工具"向"全天候智能投资伙伴"的跨越，开启了"千人千面、所思即所得"的智能服务新范式。IDC认为，服务模式重构是AI原生应用对金融行业最深刻的改变。传统金融服务依赖于标准化产品和人工服务，而AI原生应用使得超大规模个性化服务成为可能。

从技术发展来看，一方面，多模态+强推理技术已成为提升AI原生能力的关键。智能体能够整合文本、图像、音频等多种数据类型，显著提升了客户服务精度和风险管理能力。另一方面，智能体架构的成熟是金融AI原生应用发展的关键突破。其作为一种能够自主感知环境、分析信息、做出决策并采取行动以实现特定目标的系统，使得金融服务从被动转向主动，其能够主动规划、分解任务并协调执行复杂金融操作，正重塑金融机构的运营模式和客户体验。

第二章

从“单点探索”迈向“战略深化”金融行业大模型落地面临多重挑战

金融客户对金融行业大模型的关注在不同时期聚焦在不同领域，关注重心经历了算力基础设施、模型训练平台、模型参数规模、提示词工程、知识工程以及智能体运营等阶段，目前金融机构开始关注大模型数据标准、安全体系以及投入产出策略，在落地过程中面临着诸如算力难调用、数据飞轮难打造、模型与场景难适配、智能体与业务难以深度关联、安全合规体系不完善、ROI难衡量、人才难匹配等挑战。

图2 大模型在金融领域落地挑战

2.1 异构算力管理复杂，算力调度缺乏灵活性

算力、模型作为数字时代新的操作系统、基础设施普惠化和平权化，面向AI原生应用的算力应用要求算力管理动态化适应不断变化业务场景需求、智能体和模型技术持续演化。因此对于银行典型AI应用开发、大模型训练开发、AI模型统一管理部门，需要构建兼容能力强、具备技术领先的大模型训推加速云原生机制的异构算力管理平台。随着大模型应用的展开，异构算力环境下管理复杂度剧增。一是异构AI框架之间存在技术壁垒，模型在不同框架间迁移转换时，需攻克兼容性问题并重新调试参数，转换成本居高不下；二是早期银行采用算力卡单卡独占的使用模式，当训练或运行的模型规模较小时，单卡算力无法被充分利用，算力资源浪费明显；三是千亿模型集中式部署的使用会带来高昂算力的使用，PD分离分布式成为千亿模型运用的最优方案。

表2 不同使用阶段金融机构在算力管理应用的挑战

使用阶段	使用挑战
异构算力管理	• 算力部分需要适配多种芯片，确保OS、内核、驱动等端到端兼容性 • 存储部分要打通多类型存储链路，保障存储层的高性能和安全策略 • 大规模节点间网络架构不合理，导致不能满足低延迟高可靠的通信需求
大模型训练和推理加速	• 千卡GPU长时间并发训练，频繁的硬件故障没有合理的容错机制保障，导致训练有效时长不高 • 复杂的异构芯片规格、多样的任务类型以及昂贵的基础设施，需要丰富的资源分配和调度策略 • 训练/推理存储加速技术储备不足，难以快速闭环整体生命周期
训推一体	• 流量如何进行弹性容缩，监控推理场景CPU使用率或以定时的方式，按照流量监控的方式讲行伸缩，并规划训练任务的抢占
千亿模型PD分离	• 千亿模型PD分离如何快速部署，运维，容灾 • 多机缓存如何进行KV cache监控

2.2 高质量数据价值难以挖掘，飞轮效应尚未形成

高质量数据价值难以挖掘

金融行业积累了大量具备高准确性、完整性与时效性的优质数据集，但在面向大模型落地应用时，其价值释放仍面临显著挑战。一方面，金融领域的高质量数据包含大量非结构化数据（如信贷申请材料、理赔影像、票据图片、客服通话录音等），这些数据因场景高度碎片化，需经复杂预处理（如OCR/ASR转写、实体对齐）才能构建统一语义表示，导致大模型训练与调优效率显著低于通用数据；另一方面，数据安全与隐私约束下的流通壁垒也限制了数据的共享和流通，这使得部分高质量数据无法在大模型的生态系统中得到充分的利用。因此，高质量数据的挖掘受限于数据处理难度与数据安全约束，导致其难以转化为支撑大模型应用的关键资源。

数据飞轮尚未形成

数据的采集、清洗、标注、回流及模型再训练需依赖强健的数据管道与算力支撑，然而当前多数机构仍存在高人工参与度问题，导致反馈迟滞，难以实现敏捷迭代。尽管金融机构已建立数据安全策略并开展验证性实践，但因数据合规要求（如未授权数据、金融安全数据与隐私数据需“数据不出域”管控），大模型在调用金融数据时面临多级隔离限制，致使“数据→模型→业务→数据”的飞轮效应难以运转。

表3 不同类型金融机构在数据领域面临的挑战

金融机构	数据挑战
国有大行	• 自建智算中心面临数据主权与全栈治理难题 • 千亿级模型再训练缺乏场景导向的数据标准 • 业界缺乏统一的大模型导向的数据治理标准
股份制银行	• Prompt工程缺乏高质量标注数据 • 多模态数据需要统一处理框架
区域性银行	• 算力租用模式下数据出域风险 • 字段加密、多模态数据治理存在技术短板
保险	• 非结构化数据（理赔影像）向量化能力
证券/基金	• 投研数据融合难度较高 • 量化交易低延时与数据一致性冲突 • 知识工程能力（知识图谱与交易策略）
消金/互金	• 用户数据安全、数据质量与风控管理难题

2.3 通用模型难以满足复杂金融业务的应用需求

通用模型缺乏对于金融业务的深度沉淀。金融业务对精确性的要求远大于通用模型的“概率性输出”。金融业务流程复杂，专业性较强，业务逻辑差异较大（信贷审批需要多系统跳转、银行风控与保险核保），金融业务的强专业属性超出了大模型的逻辑推理边界。因此模型并非缺乏金融知识，而是天然缺乏深度的金融业务沉淀能力，即需要充分掌握金融业务之间的关联，也需要明确金融细分领域的特有规则，这直接提升了通用模型应对复杂金融业务场景的难度。

表4 通用模型与专精模型对比

比较维度	通用模型	专精模型
技术特征	• 参数规模大、开源模型支持API调用 • 通用性强、NLP与语义意义理解能力强	• 量化、剪枝后部署在本地，低延时 • 参数小、再训练可针对场景优化 • 满足可解释性与安全要求
核心价值	• 场景覆盖范围广 • 多模态数据融合度高	• 具备专业金融业务逻辑抽取能力 • 结构化数据精准分析与判别 • 可与金融业务系统深度集成融合
应用短板	• 金融专业适配性不足，对信贷风控规则、投研算法、合规条款理解程度有限 • 模型幻觉与可解释性缺陷 • 实时反欺诈、低时延场景响应速度慢	• 场景覆盖局限，难以迁移 • 复用率低，迭代成本高 • 数据依赖度高，需要高质量数据 • 工程化复杂度高，需要与金融业务深度融合并形成自动化流程
场景适配	• 非决策类、辅助类工作场景 • 金融专业度依赖较低的场景	• 决策型、高价值业务场景 • 与业务深度融合的场景

2.4 智能体难以穿透金融系统的业务流程、运营复杂度高

智能体工作流与金融业务流程难以对接。智能体的核心优势在于打破系统壁垒，实现跨系统、跨数据、跨部门的业务流程整合与优化。然而，真正要发挥这一优势，就必须深度嵌入金融业务链条，对业务环节的先后逻辑、数据触发条件、风险监控点有高度的掌握。现有智能体对金融业务（信贷、风控、支付结算、理财、核保核赔等）的细粒度环节、行业特有规则、监管要求等缺乏组件调用能力，在嵌入业务工作流的设计能力上仍显不足。

智能体运营复杂度较高。智能体的有效运行不仅依赖于稳定的模型性能，还依赖于运营人员具备多维度能力——包括AI算法基础、工具调用与协同编排以及对金融业务的理解等。这要求运营团队不仅能掌握智能体调用与编排技术，还能将算法结果与业务目标对齐，包括对数据断点、模型偏差、任务中断等问题及时反馈与修正，复杂度较高。

2.5 安全能力尚待体系化突破，金融机构多持审慎落地策略

模型安全能力仍需加强。大模型在幻觉输出、黑盒不可解释性、版本漏洞与对抗攻击等方面仍存在显著风险，直接影响金融业务的稳定性、客户信息安全和系统性风险防控。当前亟待构建覆盖模型全生命周期的安全标准体系，并依托监管推动其强制落地，以实现更系统、可监督的安全治理。

数据安全能力亟需强化。数据是金融机构构建业务差异化优势的核心资产，也是大模型应用与训练的基础。然而，业务数据需严格存储在行内系统、不可出境，这一要求与数据驱动创新的诉求形成内在张力。数据安全与业务竞争力实则为相互制约、又需协同推进的双重目标。在这一背景下，金融机构必须构建覆盖数据全生命周期的安全体系，建立完善可溯源、可审计的数据治理机制，涵盖采集、加密传输、存储管理、敏感信息分级、权限控制及操作日志审计等环节。尤其在保障交易数据的强一致性、实现信贷数据的穿透式验证、以及维持舆情数据的高时效性方面，需构建闭环式治理框架，在安全可控的前提下最大化数据价值，支撑业务差异化竞争。

内容安全能力持续加固。金融领域中，模型生成的文本、代码、决策逻辑链等输出直接关联信贷审批、风险定价等核心业务，当前金融机构需围绕：内生安全设计、动态对抗演练、长推理链的可信性验证等方面进行加固，以降低业务决策偏差风险。

应用安全能力需要深化。机构需要建立与金融业务强绑定的模型安全管理细则，包括：风控领域（利率/汇率/业务合规）、展业场景（保险/理财/交易反欺诈等）、会计审计等关键环节，实施根据业务场景流程设计的模型调用鉴权机制。

行业标准深度适配仍需加强。除了满足大模型通用安全规范之外，还要深度适配金融行业数据不出域、高实时性以及金融合规的行业三大刚性要求，目前大模型与金融核心业务融合深度不足，尤其是对于风控类场景，需要提升模型与金融行业的深度适配能力。

2.6 模型应用效果难以评估，金融机构对大模型的长期价值尚存顾虑

大模型应用效果难以量化。大模型落地具有长期性与滞后性，金融机构对大模型的投入多为战略性布局，前期需要承担高成本，但直接回报（如新业务带来的营收增长）与间接回报（如运营效率提升、风险降低、用户体验优化）多体现在业务侧，而业务的多样性导致大模型的应用边界较为模糊，因此难以形成统一的测算模型。

长期价值的不确定性加剧了金融机构的顾虑。大模型技术更新迭代快、监管环境变化频繁，使得大模型的长期投入面临被替代或被约束的风险。对于资本敏感、风险偏好较低的金融机构而言，如果短期回报不显著，长期收益又缺乏确定性，便容易产生观望甚至保守的投资态度。

2.7 业-技融合的敏捷组织尚未成熟，复合型人才稀缺

金融机构在推进大模型技术落地的过程中，除了面对技术攻坚与业务效果显性化等挑战之外，同时面临着深刻的组织与人才瓶颈。一方面，业务团队与技术团队之间仍存在着难以弥合的“理解壁垒”，业务部门作为最终使用方，更关注技术能否直接解决具体业务痛点，比如提升信贷审批效率或优化客户服务体验，他们通常以投资回报率和合规要求作为核心评估标准，期望获得立竿见影的科技赋能效果。而科技部门则更注重技术实现的可行性和系统兼容性，需要权衡算力成本、数据质量以及与传统系统的整合难度等。这种思维方式的差异导致双方在需求优先级上难以达成一致，业务部门可能低估数据治理、知识建设等基础工作的复杂性，科技部门则容易陷入技术完美主义的陷阱。更复杂的是，虽然业务部门掌握着验收决策权与评估权，但技术投入往往由科技预算承担，这种状况下，很容易出现“各说各话、各讲各事”的现象，而导致大模型场景用例的实际落地效果不如预期。

此外，人才短缺的问题更为突出，既深谙金融业务逻辑又精通人工智能技术的复合型人才在行业内凤毛麟角，且现有的培养体系难以在短时间内填补这一人才缺口，导致金融机构在大模型应用上陷入“有技术无场景”或“有场景无技术”的两难境地。组织架构的固化与人才储备的不足，共同构成了制约金融机构智能化转型的隐性壁垒。

第三章

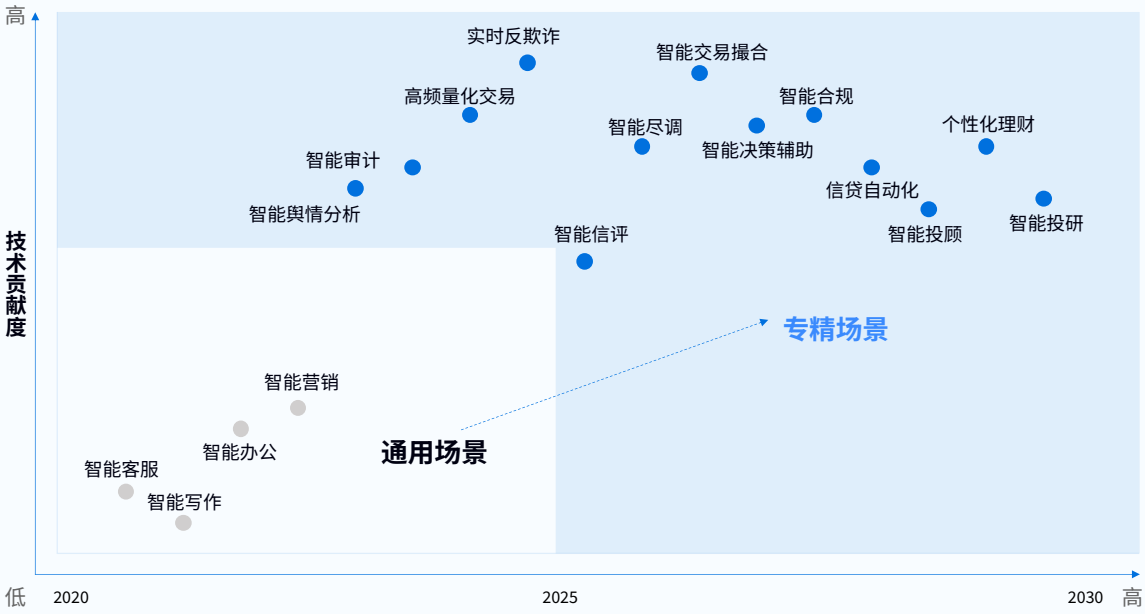
从技术到场景
金融行业AI原生应用的
重构与破局之路

3.1 金融领域呈现出通用场景向专精场景的演进趋势

金融行业大模型的业务场景落地已成为行业智能化转型的关键里程碑。在2023-2024年试点期，大模型多聚焦于单点场景技术验证，尚未形成规模化业务价值突破；伴随技术迭代与场景深化，当前大模型通过整合全域金融数据、深度挖掘细分场景需求，已完成从技术验证向深度金融属性业务渗透的质变，正式开启深度赋能金融核心领域的新阶段。

根据金融行业大模型的技术复杂度（纵轴）与时间线（横轴），IDC将金融行业大模型的应用场景分为通用场景与专精场景。通用场景指的是技术门槛相对较低且具备跨行业复制性强的场景，涵盖智能写作、智能客服、智能办公和智能营销；专精场景是指需深度适配金融业务逻辑的场景，包括智能信评、智能审计、智能舆情分析、智能交易撮合、信贷自动化、个性化理财、智能投研、智能投顾等。随着技术成熟度的不断提升，大模型应用发展的重心正加速由通用基础领域向高价值业务领域的迁移。

图3 通用场景向专精场景演进



通用场景与专精场景对模型的能力要求、落地重点等方面的差异较大。通用场景的定位更加偏重非决策与辅助类业务，对金融知识的专业要求与模型精调需求相对较低，落地重点是工具链轻量化部署，适用于金融通用大模型；专精场景的定位更加偏重决策与高价值类业务，需要深度理解金融业务逻辑，落地重点是可解释性提升与智能体调用，适用于金融专精模型。

表5 通用类场景与专精类场景对比

比较维度	通用类场景	专精类场景
业务定位	非决策、辅助型：信息整理、客服、营销文案、舆情汇总等	决策型、高价值：授信定价、量化交易、合规风控等
模型要求	依赖推理，参数规模较大，优先降低幻觉而非完全消除	深度精调与再训练，可解释性要求高，量化、剪枝后部署在本地实现低延时推理
知识调用	对金融专业知识要求相对较低	需要对金融知识进行工程化管理并通过RAG持续优化
落地重点	轻量化工具链（对接咨询、舆情、客服日志等）	业务场景数据应用，模型可解释性与智能体架构与业务深度匹配
模型推荐	通用大模型	金融专精模型

根据IDC观察，金融行业大模型（token）调用量较大的场景主要集中对话交互类和内部运营类场景。例如，智能客服（如信用卡业务咨询、理财产品咨询）、智能投研、以及内部运营助手（如内部知识库问答、政策制度查询、智能陪练助手）等场景，这类场景对数据隐私及安全合规要求较低，且具有高频交互、数据密集型等特点。

3.2 AI原生能力重构——体系化适配金融行业智能需求

随着模型落地应用场景从通用型向专精型演进，此过程中不同类型、体量的金融机构对模型的差异化需求显著，因此需要针对性匹配建设思路。差异性主要聚焦于算力、数据、模型开发、智能体开发与应用以及场景适配五个层面，如国有大行在算力层面更加偏重自主可控，需要自建多芯算力集群；区域性银行更加关注算力成本，因此多采用算力租赁的方式；证券/基金公司更加关注投研模型与产业图谱动态更新；保险公司需要应对理赔高峰并实现保障条款生成零幻觉，提升核保效率等。

满足各类金融机构的大模型落地需求的关键是对AI能力进行原生重构。非AI原生仅能实现业务局部优化，但具备AI原生应用的金融机构完成的是从算力、数据、模型到业务层面系统进化。基于各类金融机构的挑战，以及各类金融机构对大模型的差异化需求分析，我们提出了构建AI原生应用策略（AI native strategy）。

AI原生应用是指围绕基础设施、数据体系、技术架构、模型应用与业务场景等环节，都以AI为核心，让每个环节架构在AI的价值发挥之上。只有当金融行业的算力、数据、模型与业务目标均围绕AI做原生级重构时，才能系统地解决诸如异构算力调度难度大、模型可解释性不佳、工程化难题难以解决、安全合规要求难以满足、投入产出策略不清晰、人才短缺等挑战。AI原生应用包括AI原生的基础设施、AI原生的数据平台、AI原生的模型平台、AI原生的智能体平台以及AI原生的金融场景适配五个层级。

图4 AI原生应用架构图

AI原生的计算基础设施

AI原生算力基础设施的核心逻辑是“以AI工作负载为中心”：通过动态适配实现“算力与业务同频”，需搭建兼容多芯片（英伟达、国产芯片）、多AI框架（如TensorFlow、PyTorch、飞桨（PaddlePaddle）等），多参数场景（十亿到万亿）等异构算力管理平台，提升算力效率。针对百亿及以下模型，需单卡、单机实现训推一体，针对千亿参数模型的高算力需求，需采用PD/EP（数据并行+专家并行）分离的分布式部署方案—通过将“数据拆分”与“专家层分工”解耦，成倍数提升算力利用效率，彻底解决“集中式部署成本高昂”的痛点。构建算力共享机制，

让金融机构内部不同部门（如AI应用开发部、大模型训练团队、模型统一管理部门）及分支机构集约复用算力资源，降低小规模业务场景的算力投入成本，实现“基础设施普惠化”——例如分支行的“本地化客户服务模型推理”无需单独采购算力，通过总行算力池共享即可满足需求；白天推理、晚上训练的潮汐算力弹性混部架构实现算力的分时复用。

AI原生的数据平台

AI原生数据平台是金融机构基于AI知识需求重构的数据底座，针对非结构化数据沉睡、数据链路断点、高质量数据供给不足等痛点，通过多模态识别与跨模态关联激活零散非结构化数据价值，依托RAG技术+高质量向量知识库打通“行业外部数据（宏观政策/行业案例）+机构内部客户数据（信贷/行为/交易）”链路，实现“数据-知识”转化以提升模型专业度，同时统一传统数据体系与大模型数据体系，沉淀可复用资产，最终推动机构从“数据驱动”升级为“知识驱动”，助力打造“数据越用越准、价值指数增长”的数据飞轮，成为释放大模型价值的核心数据燃料库与知识发动机。

AI原生的模型开发平台

AI原生的模型开发平台能够贯穿模型训练-微调-部署-安全-运维等环节，实现多个模型协同。AI原生的模型开发平台为金融机构提供完整的工具链与各类微调版本以及原生的上下文支持能力，金融机构可以根据业务需要构建专精+通用模型矩阵，提升模型复用能力，降低模型部署门槛。

AI原生的智能体开发平台

AI原生的智能体应用开发平台能够让金融机构快速构建出会思考、会执行、会进化的智能体，大幅降低对技术人员的依赖，这种技术普惠大幅降低了使用门槛。让不会写代码的业务人员能快速创建智能体应用。能够基于金融机构复杂的业务链条提供针对性的编排与搭建方案，也能为金融机构提供丰富的智能体模板，如为保险公司提供核保智能体提升智能核保与理赔效率，为证券/基金公司提供投研助手智能体提升分析师工作效率。

AI原生的金融场景适配

AI原生的金融场景包括通用与专精两类场景，不同场景适配不同模型，从而精准满足金融业务要求。如通用场景可以满足面向员工从提效到决策的全场景适配需求；专精场景适用于对准确率、时效性、专业度要求高的核心业务以及面向客户的专业领域。

总之，金融机构的AI原生应用需要围绕异构算力管理调度、数据飞轮、通用与专精模型协同、智能体普惠以及金融场景深度适配五个层面展开，才能让金融机构能够真正享受大模型带来的指数级价值。

3.3 七大核心要素助力金融机构打造AI原生应用

对于金融客户而言，除了要关注AI原生应用的策略之外，在落地时还要掌握算力、数据、模型、智能体、安全合规、场景适配与组织人才七个核心要素，每个核心要素包括若干与之相关的二级能力指标。

图5 AI原生应用构建的落地七要素

3.3.1 构建AI原生的算力选型指标与算力共享机制

构建AI原生的算力选型指标，包括：算力密度、合理存算比、混合精度、国产化与稳定性。围绕金融业务场景对AI原生的算力需求（如智能投研、实时反欺诈要求毫秒级响应）提升算力密度，降低模型推理延迟；根据金融多模态数据（如交易流水、保单影像、投研报告文本）的处理需求，确定合理的存算比；用混合精度兼顾金融业务对模型精度（如信贷审批需要较高精度识别风险点）与算力效率（如对非关键图像可以采用较低精度，提升算力效率）的要求；通过统一的异构算力调度平台，实现对多类型算力资源的集中调度与智能分配，并通过分时复用、负载均衡和任务拆分等机制，保证业务高连续性与高稳定性。

金融机构业务（如实时风控、智能投顾、交易撮合等）对延迟极为敏感，需依托AI原生的异构智算平台实现毫秒级响应。在能耗管理方面，算力集群在非业务高峰期（如夜间、休市）往往出现GPU闲置率过高的问题。可通过GPU与XPU的混合调度、分时复用机制，以及云-边-端的弹性协同计算，实现算力资源在不同业务场景下的灵活调配，最大化资源利用率并优化功耗比，降低百万token计算成本，达到绿色算力的目标。机构需要以高效稳定、多芯适配、轻量灵活为企业管理者、运维人员、开发人员等多角色提供丰富的资源调度策略、全方位的故障感知与容错机制、极致的存训推一体化加速、便捷的多芯适配及业务迁移等硬核产品能力，完整覆盖算力应用的全生命周期。针对大规模智算场景，可同时提供容器、裸金属等多种基础设施资源类型，满足企业自建、服务托管等多类建设场景，帮助企业快速、平稳的向新一代智能化、集约化基础设施转型。

3.3.2 基于AI原生的数据平台打造从数据驱动到知识驱动的数据飞轮

构建高质量数据标准，挖掘高质量数据

针对多源异构、高敏复杂的数据特性，金融行业需要构建一套统一、高质量的数据标准体系，建立覆盖全生命周期的治理框架。要实现数据价值最大化，需要建立科学的数据标注体系，并在行业层面制定面向大模型的高质量数据治理标准。针对结构化交易数据，可定义字段级别的标签与数据质量标准；针对非结构化文本与多媒体内容，则需引入语义标注、情感分析等维度。参考某国有银行实践，其已在总分行、数据运营商、支付清算机构之间建立跨机构、多维度的数据评价体系，实现外部与内部数据的分工利用：外部数据在合规前提下用于训练模型，扩展业务洞察的广度；内部数据则更聚焦于日常运营，直接支撑精准营销、信贷风控等核心业务。

面向金融场景构建数据分级分类管理体系

金融机构需要对数据进行敏感度分级管理，例如将身份证号、账户交易流水等定义为高敏感数据，实施更严格的加密与访问控制；将地域、年龄等定义为低敏感数据，允许在更宽松的安全策略下共享与分析。着力整合碎片化的多模态数据，如将分散的贷款记录PDF、投资分析文档、票据影像等归类整理，并结合具体业务场景进行匹配应用。

强化知识工程能力

通过知识工程，金融机构可将数据转化为可计算、可推理、可共享的知识资产。金融机构应利用知识图谱技术构建业务映射网络，挖掘实体之间的复杂关系，例如客户之间的担保关系、账户之间的资金流动路径、交易行为与地理位置的关联等。在此基础上，构建AI知识库，沉淀包括风险识别模型、合规规则、营销策略等在内的高价值知识模块，并将这些模块封装为可复用组件，形成知识工程的最佳实践。

打造数据飞轮，沉淀数据资产。金融机构需要通过与模型交互沉淀结构化与非结构化反馈数据，结合外部合规数据源与扩充数据量；通过特征完成数据脱敏并将数据转换为可用的训练样本或知识，针对高频业务迭代模型，再将优化后的大模型反哺业务，提升模型的行业适配性与可解释性，形成“数据—知识—模型—业务”循环增强的飞轮效应。

3.3.3 基于AI原生的模型管理平台实现模型与场景的深度适配

基于业务场景适配模型。金融机构需要根据业务精准选择专精模型（如风控模型、量化投研模型、智能理赔模型）与通用模型（智能客服、智能营销等），并实现从模型需求定义、开发构建、定向精调，到合规测试（含风险评估、数据合规校验）、安全部署，再到上线后性能监控与迭代优化的模型全生命周期管理。

根据自身资源与技术能力选择模型路线。金融机构在模型选择与路线规划时，应对自身的技术能力、数据资源、预算投入和合规要求进行系统评估，重点关注技术透明度、可定制化程度、运维难度、安全合规水平及模型迭代能力等指标。如国有大行需要在复杂决策场景中适配千亿级闭源与专精模型，走闭源与专精路线，降低模型幻觉；股份制银行需要对复杂场景匹配专精模型，对

通用场景匹配开源通用模型，走混合路线平衡资源利用率；区域性银行直接使用7B-30B开源与通用模型或租用MaaS服务；保险公司为核保核赔专业场景匹配专精模型，围绕核保规则与用户归档等通用场景匹配通用模型，走混合路线；证券/基金公司需要关注RAG，通过专精模型先检索向量库再生成研究内容，满足自动拆解财报、生成投资摘要业务需求。

通过多个模型协同满足业务稳定性与连续性要求。适配金融业务高峰场景（如理财发售、信贷申请峰值），模型需要具备高并发处理能力，保障核心业务（如实时风控、交易决策）的低延迟响应（通常要求毫秒级），避免影响业务运转。同时支持模型故障自动切换（如主模型异常时快速启用备用模型），保障业务连续性。

3.3.4 基于AI原生的智能体平台打造智能体开发、应用与运营的赋能闭环

基于业务流程与勾稽关系编排开发智能体。金融机构需先联合业务部门（如信贷部、投研部、客服中心）开展需求拆解，对内面向员工，对外面向客户，业务专家围绕内外部业务逻辑提取关键流程，在确保数据合规的前提下，算法工程师匹配模型并编排智能体工作流，完成业务流到智能体的精准落地。

智能体应用要嵌入业务系统，提升易用性。金融机构需要将开发好的智能体嵌入业务系统，面向客户的智能客服智能体，需嵌入手机银行APP，当客户咨询业务问题时，智能体能够主动调取客户征信数据，再给出反馈；面向员工的智能体需要嵌入关键工作系统，如信贷审批人员使用的“信贷管理系统”中即可调用信贷审批智能体，无需额外打开智能体平台，避免多系统切换。

提升业务侧的智能体运营能力。金融机构需要实时跟踪智能体流量波动（如高峰时段服务并发量）、关键节点报错（如信贷审核流程中断点）与客户反馈（含满意度评分、需求未满足场景记录）等数据，通过结构化表单与定期复盘，将零散信息转化为可落地的技术优化需求，为智能体迭代提供数据与业务依据。

3.3.5 打造从硬件基础到场景应用的全栈安全能力，筑牢金融安全防线

基于AI原生的异构计算基础设施巩固安全防线。异构算力基础设施，可以将芯片池化，让金融机构实现万卡级弹性调度与训推分时复用，有效突破算力瓶颈；通过RDMA网络与联邦学习机制构建可信网络环境，满足监管与审计要求；通过框架准确表示算子执行所在芯片位置，并对不同型

号芯片的算力进行细化分配，主动感知超时、优先级与数据长度等条件，实现异构芯片调度，降低因芯片调度不当引发的安全风险；通过缓存系统与对象存储（冷数据自动下沉）分离实现智能分层，将冷数据自动下沉到对象存储，减少数据被攻击的风险，保障数据在存储和调用过程中的安全性。

基于AI原生的数据平台保障全生命周期的数据安全。在数据加密环节，尤其是数据存储与传输过程中，借助SSL/TLS协议对数据加密，可以防止数据在传输时被窃取或篡改；在存储时运用对称或非对称加密算法，让金融机构的客户信息、交易数据等得到安全存储；在数据访问控制过程中，可以采用多因素认证方式，如结合密码、短信验证码、指纹识别等，强化身份验证的安全性，避免未经授权的访问，降低数据泄露风险。

通过AI原生的模型开发平台可实现模型安全调用。AI原生的模型开发平台具备多模型协同能力，不仅可以满足算法代码加密传输与容器化隔离要求，还能够将基座模型、垂类模型与专精模型匹配至各类金融业务场景中。此外，AI原生的模型开发平台可以在敏感词过滤、伦理护栏与实时干预方面构建安全评测工具链，建造金融应用与模型之间的安全隔离带，保障安全可信。

基于AI原生的智能体开发平台实现应用安全。智能体应用在与外部协议、工具、环境交互时存在各类隐患，如通过提示词注入攻击，攻击者可修改输入提示词或注入隐藏指令，诱导大模型偏离用户请求，输出恶意结果，进而引发数据泄露、错误操作等问题。Al原生的智能体开发平台内置了金融业务规则库，实现开发工具代码合规性自动校验，能够基于金融业务流程预设操作白名单，绑定工具调用权限，从而为金融机构提供可审计、可追溯的应用安全环境。

AI原生的金融场景与模型深度适配保障业务安全。金融场景复杂多变，安全威胁可能随时出现，模型对金融业务的原生适配可以持续分析业务交易数据、用户行为模式等信息。一旦发现异常，模型能迅速启动相应的防护措施，如阻断交易、发出警报等，可以让金融机构更好地适应复杂多变的金融环境，保障业务场景安全。

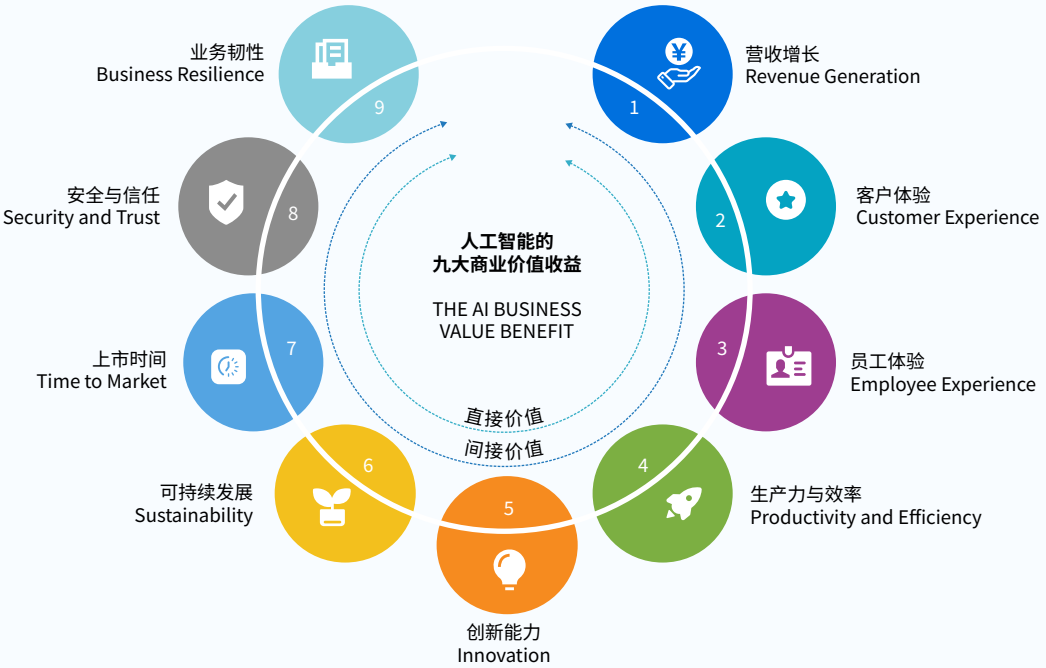
3.3.6 以ROI为核心构建模型价值的评估体系

以ROI为核心，构建模型落地效果评估体系。金融机构在落地大模型过程中，需要明确业务需求、识别关键流程、评估现有技术、分析模型适配、评估潜在收益，并以此构建模型效果评估体系，包括提效、增益、使用频率（MAU、DAU）等维度。其中，ROI作为衡量大模型应用投入与价值创造效果的关键指标，应成为各类金融机构评估模型效果的关键抓手。

成本评估：硬件成本（服务器、存储设备等硬件的采购和维护成本）、软件成本（大模型软件许可与云服务的费用）、人员成本（开发人员与培训费用）。

收益评估：营收增长（通过大模型技术创造的新业务收入）、客户体验（满意度与留存率提升）、员工体验（参与度提升）、生产效率（流程优化与运营效率提升）、创新能力（金融产品开发能力提升）、可持续发展（ESG指标与绿色金融）、上市时间（产品推向市场进程加速）、安全信任（数据安全与客户信任加强）、业务韧性（应对市场波动与不确定性能力提升）。

图6 IDC人工智能九大商业价值收益



来源：IDC，2025

通过场景筛选与指标跟踪进行模型评估。金融机构要以用户为中心，围绕ROI的成本与收益类指标来划分场景落地优先级，并进行动态调整；通过基线对比（大模型场景落地前后对指标进行对比），持续检测并长期跟踪大模型ROI相关指标，确保资源投入能够产生最大的经济效益和业务价值。

3.3.7 建立跨部门协同组织，引入技术合作伙伴，打造复合型人才队伍

对内应当打破传统部门墙，建立以业务价值为导向的跨职能协作单元或虚拟团队，由业务骨干牵头整合技术、数据、风控等资源，形成需求洞察、模型迭代与风险管控的闭环。这种组织创新不仅要求技术人员深入业务前线理解监管逻辑与客户痛点，更需要业务人员具备基础的技术思维，共同将抽象的金融场景转化为可落地的技术方案，例如，可先以设置业务产品经理/技术业务经理等虚拟岗位角色的形式，推动内部开展业技融合。

对外合作则需要建立严格的技术伙伴筛选机制，在伙伴具备领先的全栈大模型技术能力的基础上，重点考察伙伴对金融业务复杂性的专业理解程度，能够综合考虑技术成熟度与业务紧急度的匹配，为金融机构设计既契合整体数字化经营战略、又兼具领先创新方向的大模型应用场景落地方案。同时应当构建动态评估体系，确保技术方案始终与业务战略保持同步，避免陷入“为技术而技术”的误区。

如上文所说，人才队伍建设是破局的关键，金融机构可构建“引进+培养”的双轮驱动模式。在人才引进环节，明确复合型人才的画像标准，优先选拔既熟悉金融业务全生命周期管理又具备算法工程化能力的跨界人才。在人才培养方面，应当设计场景化的成长路径，通过沙盘演练、轮岗实践等方式，帮助员工在真实业务环境中掌握大模型应用价值与实现可能，逐步缩小技术与业务的能力鸿沟。这种人才战略的落地，需要管理层给予足够的资源倾斜和考核激励，才能打破现有组织惯性，真正释放大模型的赋能价值。

第四章

领先实践
金融机构大模型开发与应用案例

4.1 某国有银行——AI PaaS平台让零售业务迈入“秒级”时代

项目背景：零售银行全域升级

作为拥有庞大营运分支机构数量的零售大行，该国有银行近4万家网点像毛细血管般深入城乡，为6.5亿个人客户、18亿账户提供服务。依托“自营+代理”的独特模式，该行把“三农”、城镇居民与中小企业视为核心客群，全力助推中国经济转型。如今，该国有银行正加速实现从“最大”走向“最强”的战略跃迁：通过构建全行级智能AI PaaS平台——“人工智能大脑”，实现所有模型集中调度与统一纳管，驱动智算一体架构快速落地，打通数据孤岛、整合渠道资源、协同批零业务、优化全域运营，最终建成开放互联的数字生态银行。

落地实施：AI原生应用遍地开花

通过携手百度智能云，该国有银行以“AI PaaS”为技术底座，迭代建设，在国内大型商业银行中率先完成首个“全行级统一机器学习平台”全面落地。该平台聚集“地基”夯实：引入百度百舸算力集群与千帆大模型引擎，打通多个总行与分行的数据壁垒，形成从数据采集、特征工程、模型训练、版本管理到上线运维的“端到端”闭环。

该行持续进行平台升级、拓展场景建设，借助生成式大模型能力，孵化出测试用例分类、货币交易机器人、金融领域对话生成、金融领域辅助文档分析、金融领域投诉分析等多款AI原生应用。在项目实施上，统一平台多期迭代，逐渐向功能更完善、场景更独立、流程更智能、生态更开放的方向发展，构建了不同业务场景独立应用、打通流程智能化、赋能业务数字化、延展智能业务生态的“金融全脑”平台。

应用效果：实现了从模型到业务的全面提升

智能风控：大幅降低人工依赖

零售信贷从“5分钟”迈入“10秒”时代：自动化审批秒级完成，模型可按天迭代，信用卡、个贷等14亿账户的风险分池建模，由43天缩短至10小时即可上线。平台还为成本报账、人力、法务等系统提供预测服务，全年400万笔报账影像智能识别，大幅减轻财务审核压力。

数据智能：从建模到合规的数据体系优化

平台对接全行六大主题数据集市，一键完成数据拉取、清洗、特征衍生和统一建模。30余家省级分行已基于该底座上线营销获客与产品推荐；金融市场部也借此把市场数据与交易流水融合建模，实现实时评估交易成本、识别潜在风险。

模型开发：打造智能化战略中枢

内置的高性能数据引擎与建模引擎，把亿级信用卡样本的清洗和分析从“按月/周”缩短到“按小时”；AI集群现已承载18个核心业务系统、3大主管部门、14个支撑部门和30余家分行的模型训练与推理，成为该行智能化战略落地的“中央处理器”。

IDC案例点评

该国有银行将AI PaaS与业务战略同频规划，而非作为单点项目进行推进，通过统一数据治理与模型治理框架，实现“边缘场景—中心大脑”的双向赋能。同时，该行借助百度百舸+千帆底座，将GPU、NPU异构算力资源池化，大幅降低了训练任务周期，将模型上线周期从月缩短至天，并快速打通了数据集市与多个总分行核心业务系统，解决了数据庞杂、业务割裂的问题。

IDC认为，模型即服务是未来银行快速落地AI大模型的关键。该股份制银行不仅凭借AI PaaS完成了数据拉齐，而且持续深化大模型在复杂产品（财富、资管、托管）中的垂直微调，构建了行业级模型即服务的基准，并率先在乡村振兴、绿色信贷等监管重点场景中落地服务内容，持续巩固了该行“普惠+科技”双标杆地位。该行不仅代表了国有银行AI规模化落地的先进水平，也为全球零售金融的模型即服务提供了可复用、可扩展、可度量的全新参考。

4.2 重庆农商行——依托百度智能云企业级金融AI中台，打造代码规范的最佳实践

项目背景：加速智能化转型，破解AI建设难题

作为全国农商行体系的领军者，重庆农商行（以下简称“该行”）积极响应金融行业智能化转型趋势、持续深化人工智能技术应用。早在2018年，该行便已构建了涵盖人脸识别、语音合成、AI数字人等技术能力的智能服务体系，广泛应用于智能外呼、手机银行等业务领域。然而，面对AI技术迭代与业务需求多元化挑战，原有分散式AI建设模式的弊端日益凸显，主要表现为：算力孤岛、模型复用率低、开发效率不足。为解决这些痛点，该行亟需构建一套统一的智能化基础设施，以支撑全行的数字化转型。

落地实施：搭建金融级AI中台，赋能智能化开发

金融级AI中台：全生命周期管理平台

该行依托百度智能云企业级金融AI中台解决方案，搭建了大规模智能服务基础设施，形成了一套完整的智能模型全生命周期管理平台和服务配置体系。该平台通过私有化部署，有效整合了异构算力资源池，支持主流AI框架和各类模型（包括LLM），面向行内提供从数据处理、模型开发、模型训练、模型评估到模型推理部署等AI开发全流程支持，为前台构建了敏捷的、业务导向的智能服务体系。

核心功能包括：

- **算力资源统一管理：**构建异构算力资源池，实现统一管理与弹性分配。
- **模型训练一站式服务：**提供模型/算法库的统一管理与复用，支持一站式开发、训练、评估和微调。
- **高效推理与服务：**通过标准化API/微服务接口，实现秒级弹性扩缩、多模型灰度发布等，显著提升在线推理能力。
- **灵活服务编排：**提供可视化应用编排能力，支持AI服务的高效组合与快速迭代。

图7 某农商行AI中台

大模型知识库：新一代智能问答助手

基于AI中台和端到端应用开发工具链，该行搭建了统一的知识体系与智能问答助手，为总分行提供创新的标准化、高性能、高精度的大模型知识问答应用级服务。在知识文档解析方面，集成了通用文字识别技术，对各类word、pdf文档进行解析与切片，生成知识片段。同时，结合RAG技术，优化了传统问答流程，实现了知识的自动化扩充与精准检索，大幅提升了问答准确率。

代码助手：开创智能化开发新范式

依托百度文心快码（Comate）产品，该行实现了全栈智能化开发。智能编码技术能够自动补全、实时检查、生成单元测试，甚至完成复杂的业务逻辑生成与验证。通过引入代码助手，该行建立了代码规范的最佳实践，显著提高了软件开发的效率、质量和可靠性，并降低了人工调试成本。

应用效果：拓展业务边界，重塑金融科技竞争力

通过AI中台与代码助手项目的实施，该行成功构建了“基础设施+场景应用”的双轮驱动模式，实现了模型复用率和开发效率的显著提升。这不仅为全行的数字化转型提供了可复制的方法论，更重塑了其在金融科技领域的竞争力。未来，该行将持续拓展AI应用生态，从代码助手、员工知识问答等场景，逐步扩展更多业务领域，持续迭代并升级AI产品，进一步释放技术潜力。

IDC案例点评

该农商行智能化转型成功的关键在于百度智能云的金融级AI中台赋能。AI中台整合了分散的AI算力与模型资源，实现了异构算力池化管理和模型全生命周期管理，在显著提升资源复用率与开发效率的同时引入代码助手，实现了全栈智能编程辅助，大幅降低了人工成本。此外，在模型落地应用过程中，通过私有化部署与模块化设计（算力层/训练层/推理层/MaaS层），支持从基础模型训练到复杂业务编排的灵活扩展，也为多场景AI应用深化与场景扩展预留了发展空间。

IDC认为，技术资源整合能力、场景适配与安全合规体系建设，是该农商行顺利转型的成功因素。通过AI中台一体化解决方案解决了过往分散建设与AI碎片化的问题，通过RAG实现了金融知识的自动化萃取，通过AI开发工具链的统一管理，最终实现了端到端的工具链整合与场景落地。凭借百度全栈技术能力、金融场景深度适配及安全合规体系，该农商行实现从分散式AI到统一智能平台的升级，既验证了中台架构在金融复杂业务环境中的适配性，也为行业提供了可复用的“技术底座+场景应用”双轮驱动新范式。

4.3 泰康保险集团股份有限公司——AI综合解决方案大幅提升核保核赔自动化率

项目背景：以ROI为出发点探究AI综合解决方案与业务场景深度适配

泰康保险集团股份有限公司（以下简称泰康集团）以解决实际业务痛点为导向，弱化单一技术标签，构建“大小模型协同+AI工程支撑+场景深度绑定”的综合解决方案，在保险核保核赔、康养服务、中后台运营等场景实现降本增效，其“问题牵引型”落地路径与“ROI优先”的实施策略，为保险行业AI技术规模化应用提供了可借鉴的实践案例。

落地实施：多模型+AI工程+深度共创实现了场景化穿透

多模型矩阵打造AI工程

泰康集团采用“通用大模型+专精小模型”组合策略。基础能力依托百度文心大模型进行保险领域适配（优化保险术语理解准确率大幅提升），同时针对细分场景开发专精模型，如核保场景的病历结构化模型（融合OCR与文本抽取技术）、理赔场景的反欺诈规则引擎，大幅提升了结果可靠性。

大模型早期应用存在“重技术轻场景”的现状，通用模型在保险严肃场景中表现出准确率天花板低（如核保规则匹配准确率不足）、结果一致性不高等问题，泰康集团投入了80%的技术力量优化AI工程，建立了“模型一致性校验机制”，通过规则引擎与模型输出并行比对，降低大模型幻觉。

与技术服务商深度共创，打造“AI产品经理牵引”的运营模式

泰康集团与百度深度合作，共建保险行业首个全链路知识平台，集成搜索引擎、向量化检索、切片编辑等技术，支撑知识助手的精准响应；联合开发医疗影像解析、财务票据识别等专精模型，弥补内部技术短板。

在组织层面，推行“AI产品经理牵引”模式，组建业务与科技交叉团队，通过弱矩阵管理推动跨部门协作，重点解决中间层阻力问题，确保AI工具在业务流程中落地。

应用效果：打通了从效率提升到价值重塑的量化闭环

在核保核赔业务场景下，数据处理周期明显缩短

围绕保险核保核赔场景，引入大模型对病历进行结构化抽取，处理周期从4周缩短至1周，周期缩短三倍，效率提升3倍。同时，核保流程实现了大模型初步结论+小模型规则校验+人工复核三阶流程，大幅降低了人工成本。

在康养服务场景下，档案生成效率大幅提升

围绕康养服务场景，尤其在健康档案生成、照顾计划制定等业务流程中，实现了客户健康数据自动汇总（涵盖体检、诊疗、生活习惯等维度），档案生成效率大幅提升，并通过智能体推送个性化建议（包括慢病干预方案），档案无需人工复核，深度适配了康养场景对容错率的弹性需求。

在中台运营场景下，知识助手覆盖多个业务领域

在中台运营场景中，开发了智能交互工具，支持语音指令完成差旅报销、会议预订等操作，中台的27个知识助手覆盖了3000+内勤和数万保险代理人，实现快速条款查询与规则匹配，大幅减少了系统切入切出的成本。

IDC案例点评

泰康集团的AI实践展现了保险行业以业务为根本的思考逻辑，回归业务价值本身，通过AI综合解决方案替代大模型单一路径，避免技术投入与业务价值脱节，这种业务价值创造导向的思路与ROI策略，为业界提供了技术落地的可行性框架。

IDC认为，泰康集团在AI工程领域的投入深刻影响了其业务的创新能力。泰康集团将80%的精力投入在技术难度最高的AI工程优化方面，通过模型交叉验证、知识工程支撑、流程自动化等手段，弥补了大模型在准确率和一致性上的短板，证明了“大模型工程化能力”是当前大模型落地的关键突破点。

此外，金融机构与技术厂商深度共创将成为主流合作趋势。泰康集团与百度的共创模式解决了技术资源不足问题，而“AI产品经理牵引”的组织调整则突破了内部阻力，说明技术落地不仅是技术问题，更是生态与组织的系统性变革。未来，随着智能体技术的成熟与成本进一步降低，需求牵引与ROI优先的策略将成为金融行业大模型应用的主流。

4.4 银河证券——大模型拓宽证券业务边界

项目背景

银河证券是中国最大的国有证券公司之一。公司根植中国资本市场20余年，服务中国及“一带一路”沿线超1700万客户，客户托管资产超5万亿元，已发展成为国内分支机构最多、亚洲网络布局最广的投资银行之一。近年来，公司深耕机构业务，倾力打造“天弓”品牌，致力于为广大实体企业和金融机构提供专业化的服务。

场外衍生品是服务机构的重要业务，可以为机构提供定制化的风险管理产品。目前各家券商都非常重视该业务的发展。

对于场外交易场景来说，头部券商致力于帮助客户快速处置交易询报价指令，提高运营服务效率，使得在固定的交易时间内转化更多交易。

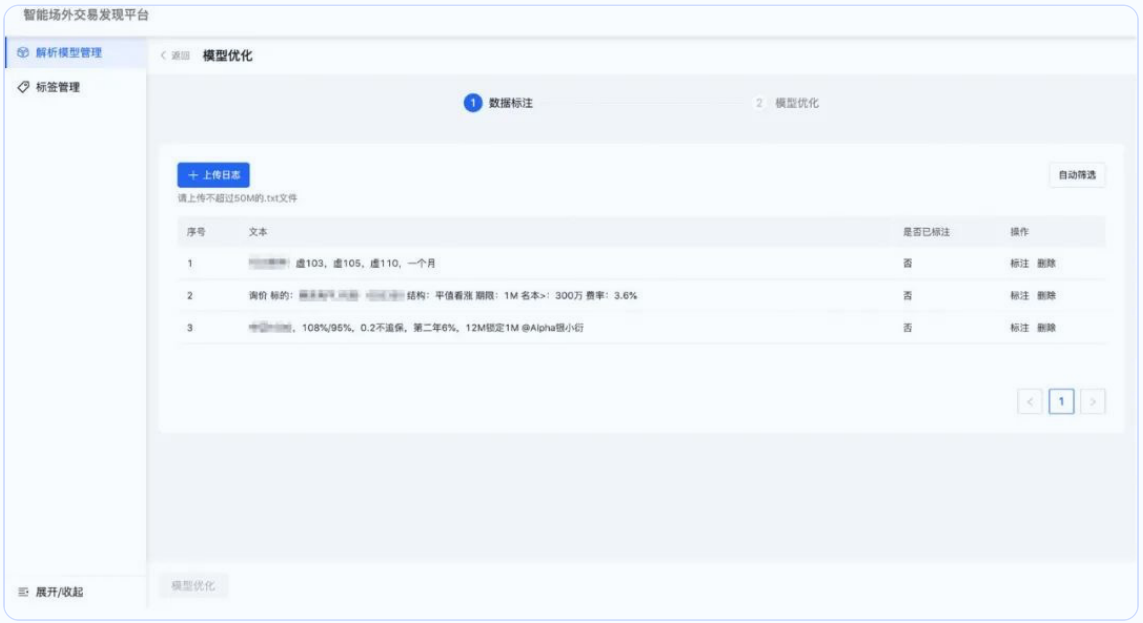
落地实施

场外衍生品业务是银河证券机构业务中非常重要的一环，在当前业务需求与日俱增的市场环境下，其逐渐成为众多头部券商竞争的主阵地。

针对机构业务服务响应滞后、业务运营成本高等问题，银河证券和百度智能云通力合作，基于场外交易解决方案先进的金融行业应用大模型底座，构建了“百度智能云金融智能场外交易平台”。该平台能够通过将交易询报价业务全流程自动化，取代查询、手工回复、信息确认等人工操作，形成从意图识别、询报价回复和多轮会话到交易转化的闭环，帮助银河证券实现了场外衍生品业务运营智能化，有效提升对客服务效率的同时，显著优化机构客户的满意度。

- **内置非标准化数据解析模型助力快速展业：**智能场外交易发现平台大模型泛化能力优异，通过少量的样本训练就可以达到不错的模型效果，目前已支持香草、雪球等股票期权及债券交易的自动询报价服务。

- **模型统一管控快速响应新业务：**智能场外交易发现平台支持解析模型自助优化，可进行自主标注、训练、调优及模型效果监控，使模型可以快速响应新业务、新资产标的。



实现智能体智能会话提升信息获取效率：利用大模型多轮会话能力，根据多轮交互的内容进行问答。通过对短时态记忆的建模，能够跨多轮对话上下文进行语义理解，精准捕捉用户的真实意图和需求状态，提升交互的自然流畅度。支持智能体智能调度，对系统内插件进行工作流调度执行。提供高度模块化的插件调度引擎，可根据如交易、托管外包等不同业务场景调用定制化的功能插件，确保系统的灵活性和稳定性，快速响应业务需求。

应用效果

该项目上线后，降本增效成果斐然，报价能力大大提高，客户体验大幅提升，交易量随之增长。2024年9.26行情爆发，机器人创造了单日下单新纪录。整个系统有力的支持了客户数量和合约数量的爆发式增长，同时保障了业务的风控合规满足监管各项要求。

通过大模型的多轮对话能力，支持历史文本记忆功能，显著提升机器人在订单查询、交易等多轮对话场景中的记忆与理解能力，用户体验满意度提高20%＋，服务效率提升30%＋，知识库问答准确率从69%提升至98%。

该项目也为整个行业积极贡献了成功经验。以该项目为主要研究内容的课题获得2022年中国证券业协会优秀重点课题，相关成果已经整理成论文发表在《金融纵横》、《中国证券》等专业期刊，并获得多项行业奖项。

IDC案例点评

作为服务机构客户、高净值客户的重要工具，证券公司的场外衍生品业务具有产品定制化强（合约期限、交割方式、结算价格）、专业性要求高（定价模型复杂、风险计量技术难度大）的特性。百度凭借领先的AI技术能力，为该证券公司搭建了智能场外交易发现平台，推动业务流程实现自动化与智能化。同时，深度适配优化多种模型到自研投顾平台，客户从询价到下单的转化率大幅提升，满足了投资者千人千面的财富管理需求。

IDC认为，以大模型为核心的AI技术可以大幅拓宽证券业务边界，助力证券行业个性化展业。百度与该证券公司的合作，是通过大模型技术优化了场外业务流程，用智能体实现了各类业务模块调度。未来，大模型会快速适配诸如主题基金、资产证券化、行业舆情实时分析等业务需求，围绕OTC交易策略为用户提供更加个性化的投顾服务与投资组合方案。

第五章

金融行业大模型落地建议：多方协同构建“战略-支撑-生态-监管”四位一体保障体系

金融行业大模型落地需要金融机构、模型厂商、产业生态方以及监管部门共同努力，明晰大模型对自身乃至全行业的战略意义，设计从模型选型到落地应用的战略顶层架构，勇于把握AI技术浪潮奔涌而来的机会，积极推动金融创新。

5.1 金融机构：构建“战略精准-执行适配-风控闭环”的系统能力

- **设计3-5年战略规划：**制定金融行业大模型从选型到落地的顶层规划，涵盖算力建设、数据体系、模型适配、智能体应用、安全合规、ROI评估与组织人才支撑，量化技术与业务考核指标并建立全员共识。
- **制定实施计划与路径：**在战略规划阶段制定三到五年路线图，明确各个阶段的资源分配。国有大行以自主可控为核心，分阶段构建全栈能力；股份制银行平衡成本与效率，聚焦场景化落地；区域性银行复用开源模型，有效提升业务效率；保险机构聚焦核保核赔，强化多模态能力；证券/基金公司聚焦低时延与投研智能化场景的模型策略。
- **战略与合规风险管理：**通过季度复盘降低运营风险，在数据合规、审计合规与个人隐私保护等方面保障大模型应用安全。

5.2 技术服务商：提供“算力效能-平台易用-模型工程化-场景赋能”的全栈支撑

- **打造金融级算力效能平台，**通过异构芯片调度管理、大小模型分布式调度，潮汐算力混合部署，跨机构资源共享，实现算力使用效率的总成本领先。
- **建设应用（智能体）开发平台：**一方面，构建依托低/零代码开发环境与组件化扩展能力的应用开发平台，降低技术门槛，加速场景化应用开发和创新。另一方面，通过建设模型管理平台，提供基于基础模型的模型精调、推理服务、模型优化、模型压缩、prompt工程等全流程的工具链，以有效满足复杂业务场景对模型能力的个性化需求提升大模型工程化能力：围绕算法研发、模型训练、行业垂直化等构建技术护城河，确保大模型具备稳定性、安全性与可控性，在知识工程、工具链管理、智能体运营、安全运维等关键环节提供技术支持。

- **深度赋能垂类业务应用：**联合金融机构拆解垂类业务痛点（如信贷审批效率低、反欺诈误判高、投研信息碎片化），构建符合监管合规要求的数据处理机制（如联邦学习、数据脱敏）、开发场景化工具（如智能风控决策系统、投研问答智能体）、强化模型可解释性与结果溯源能力，将模型能力转化为解决具体业务问题的方案。

5.3 产业生态：共建“标准统一-产学研协同-产业链联动”的协同体系

- **共建大模型标准：**标准组织、国家智库、评测机构、行业自律组织等机构需要推出模型评测、金融数据、安全合规等标准，构建标准开发工具生态，深度建立行业共识并实现资源共享。
- **深化产学研合作：**高校、科研机构与金融机构共同推动技术创新与应用，打造先导性、开放性的交流平台，金融机构与高校建立金融行业大模型联合创新实验室，围绕模型幻觉抑制、小样本风控等难题进行前沿学术与技术攻关，缩短模型应用从实验室到商业化的进程。
- **产业链深度协同：**算力基础设施与云服务提供商需要提供充沛算力；数据服务商与模型厂商需要满足数据监管要求并构建“数据飞轮”；金融机构需要围绕大模型战略聚焦价值创造并设定ROI指标。通过模型平台化、服务产品化、安全合规化、应用嵌入化以及生态协同化的方式构建“标准先行-算力适配-数据打通-场景共创-监管护航”的产业链协同机制。

5.4 规范引领：强化“政策引导-工具迭代-标准牵头”

坚持“规范与创新并重”：出台鼓励政策引导产业基金投向算力适配、算法攻关等领域，同时防范潜在风险；深化监管工具应用：利用大模型提升违规识别、风险预警的穿透性与有效性，秉持包容审慎理念引导健康发展；牵头标准建设：组织制定大模型能力、数据、风控等标准体系，在数据采集规范、算法审计、模型风险评估等方面给出明确指导，协同推动合规落地。

结语：金融行业大模型落地需“机构定战略、服务商给工具、生态聚合力、行业立规则”，多方协同将“技术能力”转化为“业务价值”，最终实现从“模型可用”到“产业好用”的规模化突破。

第六章

关于百度智能云
金融行业“双智能 双引擎”方案

百度智能云金融行业“双智能”“双引擎”方案

随着全球数字化浪潮的加速演进，金融行业正站在一个由人工智能（AI）技术，特别是大语言模型（LLM）驱动的深刻变革的十字路口。传统的业务模式、服务渠道和运营效率面临前所未有的挑战与机遇。百度智能云在服务客户过程中沉淀的，从算力芯片应用架构，提出一种前瞻性的“双智能双引擎”架构体系。该体系以“智能数字员工”与“智能对客服务”为两大核心应用（双智能），并由“百度智能云千帆AI开发平台”与“百度百舸AI计算平台”两大核心技术基座(双引擎)提供动力，系统性地重塑银行、保险、证券、基金等金融机构的业务流程与价值创造方式，旨在为金融行业的智能化转型提供一套全面、可行、高效的战略蓝图。

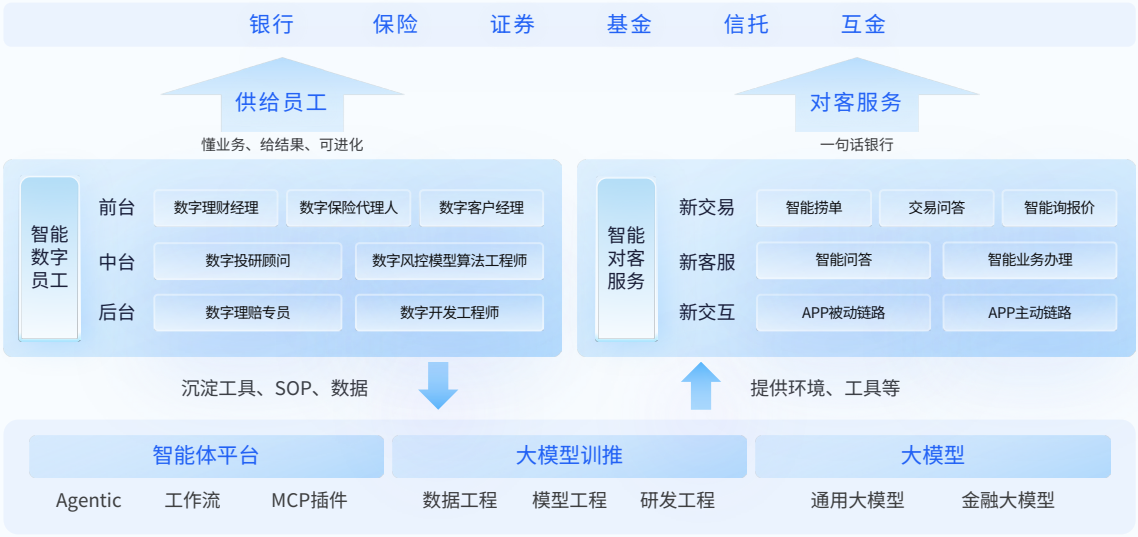
6.1 “双智能”应用层——重构金融服务新体验

6.1.1 智能数字员工：打造金融机构的超级生产力

智能数字员工是AI驱动的虚拟劳动力，它们深度融入金融机构的各个业务环节，承担起高复杂度、高知识密度的工作，成为人类专家的得力助手。

双智能：从“降本增效”到“创新增长

打通技术到业务价值的最后一公里



前台数字员工

- **数字理财经理：**能够7x24小时分析海量市场数据、宏观政策和研究报告，为客户生成个性化的资产配置建议，并辅助投资经理进行深度研究，极大提升投研决策的效率与覆盖面。
- **数字保险代理人：**能够自动挖掘和推送潜在客户，将场景挖掘效率提升500%+。同时，作为全能业务助理，它能自动回复高频问题，并协助完成复杂任务，业务助理助推效率提升90%+。内置10万+专业知识库，使代理人的知识储备提升95%+，并能生成个性化的营销内容，使营销手段更丰富。通过模拟实战环境，智能培训系统可将新代理人的培训周期缩短50%+。可根据客户画像提供精准建议，让代理人面客准备更充分。简化线上投保流程，使保单成交更简单。
- **数字客户经理：**主动洞察客户需求，进行全生命周期的客户关系管理，提供千人千面的产品营销和服务支持，深化银行与客户的连接。

中台数字员工

- **数字投研顾问：**高华证券与百度智能云从去年年初开始在证券投资的核心场景进行深度合作，投入力量共同研发了基于大模型的指数化股票投资系统，依托百度千帆大模型平台，借助提示工程、思维链设计，去模仿专业投资者的思考逻辑，根据公开信息形成指数组合决策，属于我国业内首创。双方合作研发的最新研究成果——华证高度大模型新质生产力指数。这是一个科技成长类指数，与红利类的稳健50恰好形成互补。新指数通过大语言模型技术，将上市公司在生产、销售、研发、投资等维度上的公开信息与权威政策文件语料进行匹配与分析，筛选出深入践行新质生产力发展理念并且成长性强的上市公司用来构建投资组合。截至7月底，大模型新质生产力指数近五年全收益指数年化收益达13.7%，不仅大幅超越同期中证科技100指数0.4%的年化收益，在回撤控制方面也展现优势——最大回撤较中证科技100指数降低近18个百分点，再次展现出大语言模型在选股领域的巨大潜力

- **数字风控模型算法工程师：**在金融风控建模任务中，特征工程始终是影响模型性能的核心环节。传统做法多依赖人工经验与规则构建，虽能产生一定区分度的特征，但在大规模序列化、多维度的交易数据下，人工方法的效率与覆盖度明显不足。应用智能体方案，建模效率可由数月完成特征工程小时级别，极大提升提取的特征的IV效果，保证模型抓违约人群的能力。

后台数字员工

- **数字理赔专员：**传统理赔核算方案耗费的人工成本和时间成本巨大，且核算过程难以按照指定形式向客户呈现，赔付结论可读性较差。通过数字理赔专员，大大节省了人力成本和时间成本，且案例与赔付规则的公式匹配、公式计算、案例赔付总结等过程可以按照指定形式清晰呈现给客户。

- **数字开发工程师：**可以理解业务需求，自动生成和优化代码，构建和迭代风险控制模型，将金融机构的模型开发与软件工程效率提升至新的量级。

通过部署智能数字员工，金融机构不仅能实现显著的降本增效，更能将宝贵的人力资源从重复性工作中解放出来，专注于更具创造性和战略性的高端价值活动。

6.1.2 智能对客服务：开创全场景智慧交互新时代

智能对客服务旨在利用AI大模型，打造一个无缝、统一、高度智能化的客户交互中枢，重塑服务体验。

- **新客服务：**在获客环节，通过智能对话机器人提供全天候在线咨询，精准解答客户疑问，引导客户完成开户、申请等流程，提升转化率。
- **新APP体验：**将金融APP从一个功能菜单的集合，升级为一个“有思想”的智能金融助手。用户可以通过自然语言对话，直接办理业务、查询信息、获取投资建议，实现“所说即所得”的极致便捷体验。
- **新交易场景：**在交易过程中，嵌入智能风控提醒、市场机会解读和交易策略辅助，让每一次交易都伴随着专业的智能决策支持，提升客户的投资成功率和满意度。

6.2 “双引擎”技术基座——驱动智能应用的强大动力

如果说“双智能”是金融智能化的上层建筑，那么“双引擎”就是其坚实可靠的底层基础，确保AI应用能够被高效开发、稳定运行和持续迭代。

6.2.1 百度智能云千帆AI开发平台——一站式企业级大模型开发与服务中心

千帆大模型平台为金融机构提供了从模型到应用的全链路工具与服务，是连接底层技术与上层业务的桥梁。

百度智能云千帆ModelBuilder

百度智能云千帆ModelBuilder能够基于国产化算力资源，实现从数据管理、模型开发、部署上线到在线测试的AI能力研发与应用全生命周期建设和管理。在数据管理方面，可以有效地处理大规模的数据，支持不同类型数据处理等功能；在模型开发方面，提供丰富的预置算法，包括市场领先的开源大模型、百度文心一言大模型、百度千帆中文增强大模型等，同时提供高效、稳定的开发环境，支持多模态、多类型任务、大模型等多种模型开发需求；在部署上线方面，支持多种部署方式，如在线部署、离线部署等，可以灵活地部署在不同的环境中；在线测试方面，可以支持实时在线的测试场景，可以快速地定位大模型调优方向，提高模型的质量和可靠性。

千帆大模型平台的普及使得这些自动化训练技术更加普惠。通过提供易于使用的工具和接口，使得广大开发者和研究人员能够轻松地使用自动化训练技术，而不需要深入了解其背后的复杂原理。平台具备以下关键能力：

- **多种微调方法：**通过全量更新和LoRA自动调整参数，减少人工干预，提高训练效率。
- **可视化界面和工具：**提供易于使用的可视化界面和工具，方便用户管理和监控模型训练过程。
- **丰富的预置算法：**集成市场上领先的开源算法，预置丰富的小模型案例，快速部署服务体验效果，降低用户使用门槛。具备业内优势的开源大模型，Llama、Qwen、Deepseek等。

大模型+小模型部署愈加复杂，体系化工具是应用实现利器

随着大模型和小模型的广泛应用，大模型的复杂性要求更高的计算资源和更精细的调优，而小模型的多样性则带来了更灵活的应用场景和更高的部署需求。它们的部署过程变得日益复杂，需要高度的技术专业知识和有效的工具支持。在这种情况下，体系化工具成为实现顺利部署和应用的利器，为开发人员提供了关键的支持。

首先，大模型和小模型的部署涉及到多个环节，包括模型转换、优化、推理引擎的选择等。针对这些复杂的任务，千帆大模型平台-燧原定制版模型部署工具提供了一套完整的解决方案，通过集

成各种功能模块，简化了部署流程。这使得开发人员能够更加专注于模型设计和业务逻辑，而无需过多关注底层实施的技术细节。

其次，部署工具在跨平台部署方面发挥了重要作用。由于不同硬件平台和操作系统的差异，将模型顺利部署到各种环境中变得复杂而具有挑战性。模型部署工具通过提供通用的部署接口和适配层，使得模型能够在多种环境中运行，从而提高了模型的可移植性和通用性。

百度智能云千帆AppBuilder

作为企业级AI原生应用开发平台，百度智能云千帆AppBuilder是实现业务价值闭环的关键。它是连接底层技术与上层业务的桥梁。其核心在于开创了大模型驱动应用开发的新范式，极大地降低了AI应用的开发门槛。

- **零代码与代码态并行：**为业务人员提供零代码的GUI交互界面，通过简单的“拖拉拽”和对话式配置，三步即可完成应用的创建与分享；同时，为专业开发者提供可编程的完整开发套件、工具链组件和工作流，支持更复杂、更定制化的应用开发。
- **组件化与生态化：**平台提供丰富的预置组件，并设有“组件广场”，鼓励开发者共享和调用，形成繁荣的应用生态。通过连接知识库、数据库、大模型和各类API，可以快速构建出功能强大的金融领域智能体，加速创新落地。

核心优势：

- **应用效果领先：**内置企业级全链路检索增强与应用框架，能够实现效果分析、效果反馈和效果调优的实时闭环，确保问答准确率高达90%以上。
- **组件工具丰富：**预置超过60个AI能力组件，深度覆盖政务服务、营销办公、研发生产等主流业务场景，工具自动编排准确率超过90%。
- **产品开放易用：**通过零代码/低代码开发模式，并配套全栈课程，极大降低使用门槛，将开发效率提升30倍。支持多渠道分发与集成，快速满足各类业务线集成需求。
- **全面支持国产化适配：**支持私有化部署，彻底解决数据安全问题。全面支持信创，满足国产化诉求，适配主流国产芯片、操作系统及数据库。

百度智能云千帆金融行业大模型：千帆慧金

金融场景效果更好，模型应用灵活度更高

QianfanHuijin-70B-32K

QianfanHuijin-Reason-70B-32K

百度千帆慧金大模型

针对行业应用中的专业需求，百度以金融行业为试点，正式推出**千帆慧金金融大模型**。该模型基于海量金融语料深度训练，构建金融专用合成数据管线，优化算法策略，并提供知识增强大模型和推理增强大模型两类模型，每类模型分别提供8B和70B两个版本，支持最长32K上下文输入，覆盖金融行业多数场景。

在金融领域Benchmark评测中，千帆慧金金融大模型综合表现领先，百亿参数模型得分超过千亿参数的通用模型。在金融销售赋能场景中，相比通用模型，千帆慧金金融大模型能更完整地列出贷款材料清单、拆解工作流程，并明确风险管控要点，展现出深厚的行业知识与推理能力。

AI搜索

当前，企业在信息检索与决策支持中普遍面临着四大问题：

- 自建大模型知识库存在信息更新滞后、时效性差的问题，难以覆盖热点事件和突发舆情，在面对复杂、专业的问题时，大模型可能因知识库限制而输出过时或错误的答案；

- 企业内部知识边界局限，缺乏对产业上下游和全球趋势的外部信息补充；
- 员工在处理外部信息过程中需通过外网终端进行人工搜索与筛选，加工成本高、效率低；
- 搜索结果来源杂乱、质量不一，难以保障内容准确性和安全合规性。

针对以上问题，百度搜索可以即时捕捉并提供最新数据，弥补时效性的不足；智能搜索生成结合大模型和基础搜索的能力，通过搜索拓展知识范围和大模型自身的总结推理能力，提升输出的准确率。

百度搜索适用于企业内部有大模型，希望将搜索作为实时数据源，需要“原汁原味”素材自己来加工的场景；智能搜索生成适用于企业希望接口不是给链接、而是直接给答案的情形，并且答案要保证权威性和结构化。常见的应用场景如下：

知识问答助手

知识问答助手可作为企业内部知识库的有力补充，有效解决因知识治理混乱导致的检索不准确、知识更新不及时等问题。知识问答助手通过接入大模型与百度搜索能力，支持用户获取与问题相关的网页列表和原始内容，弥补自建大模型缺乏外部实时数据导致回答不准确的不足。依托百度搜索的分钟级更新能力，知识问答助手能够提供具备高度时效性的内容，有效提升问答系统的智能化和实用性。

客户经理助手/客户助手

企业可以在内部或者面客的系统或者APP中引入智能搜索功能，打造客户经理助手或客户助手。通过提供生活类信息查询、新闻浏览、热点事件追踪等功能，可以有效降低了员工与用户获取信息的成本，还能在持续使用中培养用户习惯，从而提升系统和APP的整体使用率和用户粘性。

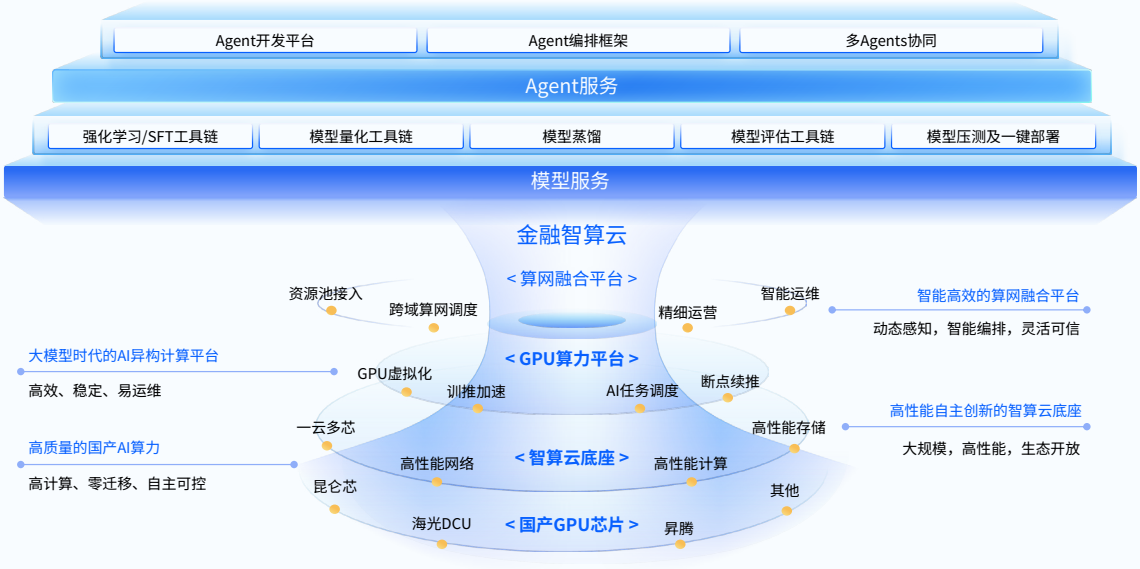
写作助手

写作助手通过集成搜索功能，帮助用户在撰写内容时快速获取所需素材、案例、数据或背景信息，可以极大降低创作过程中的信息搜集成本与时间消耗。无论是用于新闻撰写、市场报告、营销文案还是社交媒体内容生成，写作助手都能够实时提供权威、丰富且多样的信息来源，激发创意灵感，支持结构搭建，满足多样化的创作需求。结合大模型能力，还可对搜索结果进行初步摘要与结构化提炼，为写作者提供更具参考价值的内容支持。

6.2.2 百度百舸AI计算平台——坚如磐石的AI算力底座

算力与模型作为数字时代新的操作系统与基础设施，正朝着普惠化与平权化方向发展。面向AI原生应用的算力应用，要求算力管理具备动态化能力，以适应不断变化的业务场景需求，同时应对智能体与模型技术的持续演化。因此对于银行典型AI应用开发、大模型训练开发、AI模型统一管理部门，需要构建兼容能力强、具备技术领先的大模型训推加速云原生机制的异构算力管理平台，帮助金融机构在数智化转型中抢占先机，在确保安全合规的前提下，高效地开展业务创新和智能升级，**解决算力高效分配问题，完成千亿模型PD分离动态部署，潮汐算力训推一体，在降低算力使用成本的同时，提供算力高效分配机制。**

大模型时代，AI原生的金融基础设施的建议



算力管理平台

随着通用大语言模型（LLM）和金融垂直大模型的迅猛发展，人工智能正在深刻重塑金融行业的业务模式、风控体系和客户服务体验。金融机构在拥抱大模型机遇的同时，逐步走向算力精细化管理的方向。

异构资源管理

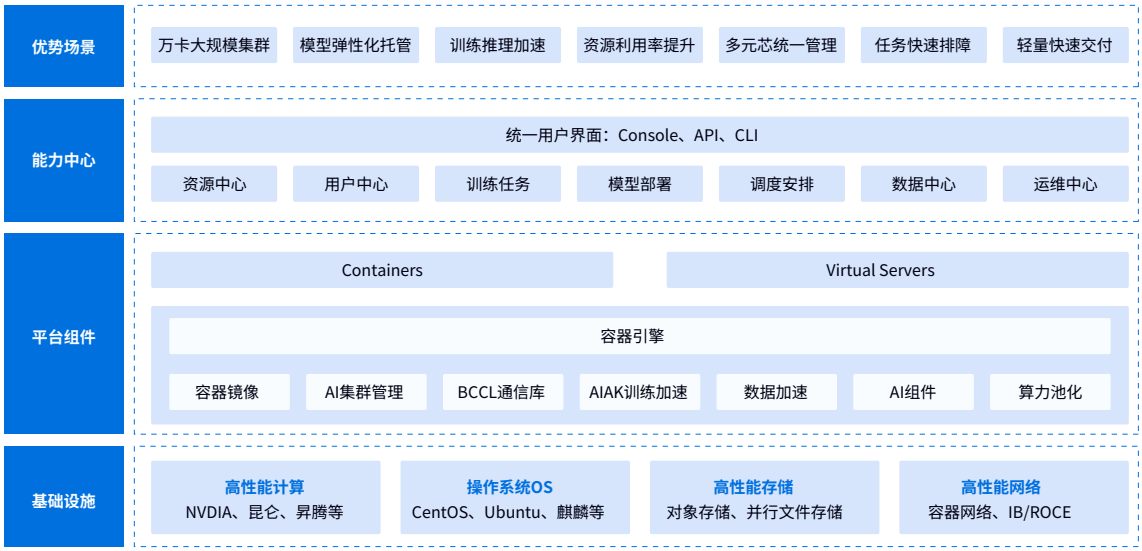
百度百舸AI计算平台通过高性能计算、存储、网络、集群管理、训推框架，为大模型场景下的各种任务提供高效的计算能力和数据处理能力，确保任务的高效执行。支持各类AI加速卡，如NVIDIA、昆仑、昇腾等高性能计算卡，提供强大的计算能力以满足大模型场景下各种训练、推理任务需求。并且支持零成本适配原生PyTorch/TF环境、Hugging Face架构大模型开箱即用、ONNX自动优化。支持CentOS、Ubuntu、麒麟等多种操作系统，通过多操作系统的兼容性，确保应用能够在不同环境中顺利运行。提供并行文件存储、对象存储等高性能存储解决方案，无论是大规模训练数据还是模型参数都能高效读取，确保数据处理的流畅性。同时支持容器网络、IB/ROCE等高性能网络技术，确保数据传输的低延迟和高带宽。

训推一体

AIAK支持多种主流大模型的训练加速，例如Llama、Qwen、Baichuan、Mixtral等系列模型的Postpretrain和SFT微调场景。通过优化算法和提升计算效率，AIAK能够显著提升训练吞吐量和多卡训练加速比，减少训练时间。可提供推理加速镜像，支持Llama、Qwen、Baichuan等系列模型的推理加速。通过并行优化、显存优化和算子优化，AIAK能够显著提升推理吞吐量，降低推理延迟，提高模型的实时性。拥有模型权重格式转换和并行策略切分工具：AIAK支持模型权重从Hugging Face到Megatron框架的相互转换。此外，还支持Megatron框架下模型权重按照不同的DP（数据并行）、TP（张量并行）、PP（流水线并行）并行策略进行切分，方便用户根据硬件配置和任务需求进行灵活调整。同时支持并行策略自动搜索工具，能够根据用户的硬件环境和模型特性，自动搜索最优的并行策略，帮助用户快速进行性能调优，以达到该配置下的最优性能。

监控运维

在AIHC PRIVATE中，支持用户一键开启容错，覆盖了训练进程Hang、训练心跳失联、训练进程报错异常退出、Pod被误驱逐等场景的故障感知定位&自动恢复能力，可完成训练异常感知，提供了强大的训练异常感知能力，能够检测到任务退出、任务假死、运行缓慢等常见故障场景。特别是对于难以识别的任务hang场景，百度百舸AI计算平台基于百度内部大量的最佳实践制定了指标体系，可以及时发现问题。进行容错判断，基于其资源池的自动故障隔离能力，能够检测任务所在节点是否发生故障。一旦检测到故障，平台会自动隔离该节点，并触发任务容错流程。同时可完成任务异常自动恢复，针对节点故障导致的任务异常场景，千帆异构算力管理平台会尝试通过重调度训练任务的能力，快速恢复任务。具体来说，当检测到节点故障时，平台会自动隔离故障节点，并将任务重新调度到健康的节点上继续运行。



关于 IDC

国际数据公司（IDC）是在信息技术、电信行业和消费科技领域，全球领先的专业的市场调查、咨询服务及会展活动提供商。IDC帮助IT专业人士、业务主管和投资机构制定以事实为基础的技术采购决策和业务发展战略。IDC在全球拥有超过1100名分析师，他们针对110多个国家的技术和行业发展机遇和趋势，提供全球化、区域性和本地化的专业意见。在IDC超过50年的发展历史中，众多企业客户借助IDC的战略分析实现了其关键业务目标。IDC是IDG旗下子公司，IDG是全球领先的媒体出版、会展服务及研究咨询公司。

IDC China

IDC中国（北京）：中国北京市东城区北三环东路36号环球贸易中心E座901室

邮编：100013

+86.10.5889.1666

Twitter: @IDC

blogs.idc.com

www.idc.com