IDC | 中电金信 GienTech

开启智能金融新时代

——中国金融大模型发展白皮书

目录

核心观点

AI大模型成为新质生产力的重要组成部分，国内外科技公司正加速布局

AI大模型已成为新质生产力的重要组成部分，能够大幅提高生产效率，优化资源配置，降低生产成本，为企业高质量发展提供强大的技术支持和动力。当前，美国、中国、日本、欧盟等全球主要地区的科技公司正加大大模型技术的创新及应用。在未来GenAI投资分配上，中国和全球企业几乎都会平均分配在生产力提升应用场景、跨行业水平职能应用场景、以及垂直行业专属应用场景上。

金融行业AI大模型的研发投入和应用较为显著，且具有一定的应用特殊性和应用挑战

近两年，金融行业在AI大模型的研发投入和应用方面亦走在市场前列。根据IDC数据显示，2024年，中国金融行业AI and Generative AI投资规模达到196.94亿元，到2027年将达到415.48亿元，增幅达到111%。同时，金融行业属于信息密集型、风险规避及强监管行业，在推进大模型落地过程中，相比其他领域，金融行业对数据质量、推理准确性及响应速度，以及在管控、合规、安全层面的要求都更高。同时，根据IDC调研数据显示，数据治理、模型治理、以及合规应用是金融机构落地大模型/生成式AI更需求关注的要素。

AI大模型在金融行业的应用场景正从简单到复杂加速分步推进

IDC认为，生成式AI的行业应用往往都是循序渐进的过程，一般是逐渐从内部辅助运营到外部对客提效、从业务边缘到核心，相应地AI对金融机构的价值也逐渐增大。在未来18个月内，支付清算、智能投研、内部研发（代码生成、测试等）、数据分析（报表生成与分析、数据建模、数据决策等）、欺诈/洗钱/威胁监测、资产管理（资产尽调、资产评估及定价等）是金融机构主要的落地场景。在应用流程方面，IDC认为，金融行业生成式AI应用场景的落地可以从场景应用评估与选择（如技术解决方案评估、项目管理及风险评估、投资回报分析）、以及面向场景的工程化落地（如模型选择、技术路线、数据及算力准备、模型训练及调优、以及系统集成与部署、组织协同等）等层面分步推进。

金融机构应根据其资源能力选择不同的大模型应用路径，并需打造多样化的能力要素

当前，不同类型的金融机构在推进大模型的落地中，有着不同的路径选择，其可根据自身战略目标、业务需求、技术能力、资源禀赋、风险偏好来决定是否自主建设、基于已有模型微调，或是采用其他方式来利用GenAI能力。同时，IDC认为，金融机构在落地大模型的过程中，需要综合考虑打造数据价值链管理、模型的选择与部署、AI平台搭建、以及AI治理等要素能力。尤其是在数据价值链管理方面，IDC认为，金融机构的数据价值链管理是生成式AI在金融场景中有效发挥价值的基石，其核心目的是提升数据质量、数据可用性以及确保数据的合规获取，有利于金融机构面向不同的应用场景快速构建高质量的数据集，并为后续金融大模型的规模应用奠定坚实的基础。

◉ 多模态技术、AI智能体、以及通过大小模型协同应用和构建大模型生态资源共享平台是金融机构落地大模型的主要趋势

随着大模型技术的发展，大模型的参数规模也将显著增长，多模态技术及智能体亦将在金融机构中深入应用。一方面，多模态之间的融合将使得AI大模型能更深刻地捕捉复杂场景背景、细节和情感，使其更快的感知和适应场景，并能应用于更加复杂的金融场景。另一方面，AI智能体通过“感知-认知-推理-决策-组织/行动”的闭环，及其在数据处理、智能决策与自然交互等方面的卓越能力，预示着它将在客户服务、业务流程优化及业务效率提升等多个关键领域发挥核心作用，为金融机构带来前所未有的价值创造。此外，IDC认为，通过大小模型协同也能驱动金融机构在更加多样复杂的场景中的应用。同时，通过构建大模型生态资源共享平台，向金融机构提供大模型应用所需的全套资源，是金融机构大规模应用生成式AI的主要路径之一。

第一章

百舸争流
AI大模型发展概述

1.1 AI大模型与新质生产力

当前，人工智能正以前所未有的速度和规模渗透到我们生活工作中。人工智能是数字基础设施建设的重要组成部分，是新一轮科技革命和产业革命的核心驱动力，在人工智能技术的加持下，全球的数字化转型已进入倍增创新阶段，同时以多模态大模型为代表的新型人工智能技术正高速发展。

2024年1月31日，习近平总书记在主持二十届中共中央政治局第十一次集体学习时，进一步强调发展新质生产力是推动高质量发展的内在要求和重要着力点。而以AI大模型为主的新技术，作为各行业的新质生产力的重要组成部分，能够大幅提高生产效率，优化资源配置，降低生产成本，为企业高质量发展提供强大的技术支持和动力。尤其是随着AI Agent的潜力被不断挖掘，以AI Agent为核心的人机协同将为业务洞察与决策提供新的能力支撑，为金融机构构建领先的新质生产力。未来推动金融业逐步走向智能化金融的演化，实现超高数据处理与实时决策的融合，推动着普惠金融、金融供给侧改革、客户体验/个性化服务不断深化。

1.2 国内外AI大模型的发展现状

当前，大模型技术加速发展，美国、中国、日本、欧盟等全球主要地区的科技公司正加大大模型技术的创新及应用。美国在生成式AI方面起步较早，OpenAI、Google DeepMind、Meta等科技公司在生成式AI领域取得了里程碑式的进展。在中国，百度、阿里、华为、腾讯、京东、科大讯飞、字节跳动等科技公司也纷纷发布了基座大模型，且加速推进其在各行各业的落地。而欧盟的科技公司也加速应用生成式AI，且其更倾向于在细分领域（如医疗、金融等）应用生成式AI，而不是开发通用的大规模生成模型。例如，英国的DeepMind是生成式AI领域的重要力量，其生成模型在文本生成、游戏AI等方面表现突出。日本在生成式AI的技术开发上相对滞后，其在机器人和自动化领域具备全球领先的技术实力，但在自然语言生成和通用图像生成等方面，尚未推出具备国际竞争力的大规模模型。

不过，日本的一些企业和科研机构也在逐渐跟进，根据日本政府发表的《信息通信白皮书》表示，未来的增长潜力不容忽视。71.1%的受访者表示，在合适的情况下，愿意尝试使用生成式AI。

表1 主要国家和地区生成式AI发展现状

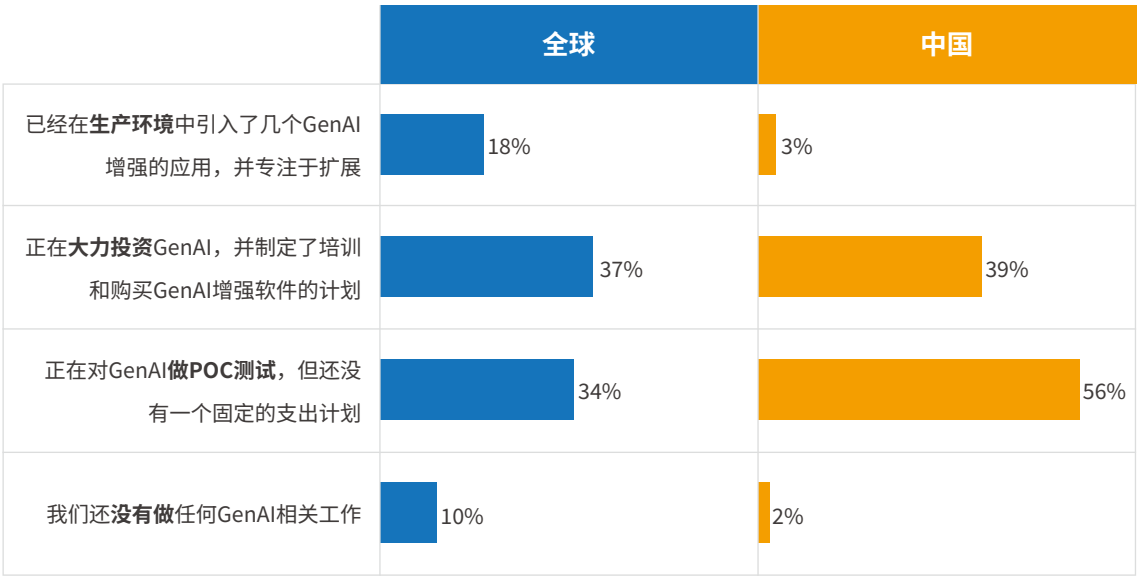
地区	技术实力	应用场景	政策支持/监管
中国	快速发展，百度的“文心一言”、阿里的“通义千问”等语言生成模型快速发展	广泛应用于社交媒体、电商、金融、在线教育、智能客服等领域	政府大力推动AI创新，出台支持政策，关注数据保护与AI伦理
美国	全球领先，拥有OpenAI的GPT-4、 Google Gemini等顶尖模型	广泛应用于文案生成、图像生成、代码生成、音乐和视频创作等	政府出台AI“权利法案蓝图”，探索监管框架，平衡创新与社会责任
欧盟	技术发展相对滞后，DeepMind等公司在细分领域有所进展	主要用于医疗影像生成、数据合成、建筑设计、自动化报告撰写等专业领域	《人工智能法案》强调高风险AI应用的严格监管，注重公平性和透明性
日本	技术进展缓慢，集中于机器人、自动化等领域，缺乏国际竞争力的生成模型	主要应用于动漫创作、虚拟偶像、工业设计、医疗健康等领域	根据日本政府发表的年度《信息通信白皮书》数据显示，日本国内生成式AI的个人使用率和企业使用率都相对较低

来源：IDC根据公开资料整理

1.3 AI大模型应用发展整体现状

大模型作为带动产业/组织服务效率及范式变革的重要技术，已经具备较高的识别准确率和较强的场景泛化性，在多模态的任务下也有明显的突破，全球诸多企业已在金融、电商、能源等行业“试水”。

图1 您的组织目前评估或使用生成式AI（GenAI）的情况如何?

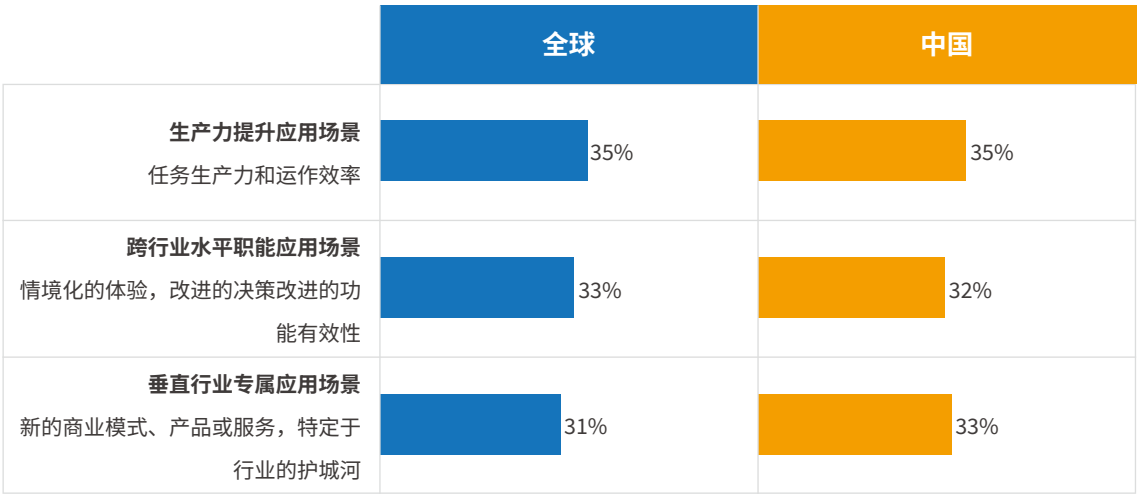


来源：IDC's Future Enterprise Resiliency & Spending Survey, Wave 3, 2024年3月, n = 887
（北美: 363, 欧洲与其它地区: 204, 亚太: 300 [中国: 100] ）

据IDC全球调研显示（如图1），全球已经有18%的企业在生产环境中引入了几个GenAI增强的应用，并专注于扩展；中国的这一比例仅为3%，但中国开始投资或做POC测试的企业达95%。

在未来GenAI投资分配上，中国和全球企业几乎都会平均分配在三类应用场景上（大约各1/3），如图2。

图2 考虑您在未来18个月GenAI的投资，这些投资将如何分配到以下类型的应用场景中?



来源：IDC's Future Enterprise Resiliency & Spending Survey, Wave 2, 2024年2月, n = 896
（北美: 371, 欧洲与其它地区: 225, 亚太: 300 [中国: 100] ）

目前，基础大模型建设已经较为完整，诸多云服务商、AI技术服务商、数据服务商等均推出其基座大模型，且各具特色，未来将会进入大模型应用跑马圈地阶段，行业应用场景数量也将爆炸性地多元化增长，且会逐渐从辅助运营类的业务场景向决策管理场景深入。

第二章

聚焦行业
金融行业大模型概述

金融业在我国经济中举足轻重，金融机构通过提供资金流动和管理服务，为个人、企业和政府的各种经济活动提供必要的资金支持。近两年，金融行业不断地利用新兴技术推进业务高质量发展，尤其是在AI大模型的研发投入和应用方面亦走在市场前列。根据IDC数据显示，2024年，中国金融行业AI and Generative AI投资规模达到196.94亿元，到2027年将达到415.48亿元，增幅达到111%。

2.1 金融行业大模型应用的特殊性

如今，金融科技已经从“立柱架梁”迈入了“积厚成势”新阶段，越来越多的金融机构积极使用AI大模型等新技术助力其实现高质量发展。AI大模型虽在金融行业有较多的应用场景和应用价值，但是，金融行业属于信息密集型、风险规避及强监管行业，在推进大模型落地过程中，相比其他领域，金融行业对数据质量、推理准确性及响应速度，以及在风控、合规、安全层面的要求都更高。同时，根据IDC调研数据显示，数据治理、模型治理以及合规应用是金融机构落地大模型/生成式AI更需求关注的要素。

在数据层面，金融行业处理的数据往往涉及客户的隐私信息、交易记录等敏感数据，数据来源多样且数据质量参差不齐，而数据规模、数据质量和多样性会影响大模型在具体场景应用的效果和性能。若输入的基础数据不准确或时效性较差或存在数据操控问题，那将直接影响模型微调效果，以及模型输出的准确性。同时，训练数据可能存在性别、种族及主观因素等方面的偏见。如果这些偏见被应用到金融决策中，可能导致模型在决策和预测中产生不公平或歧视性的结果，如何解决数据的合规获取及保护信息/内容版权，并合理设置相关的诉讼机制和监管及罚款机制，也是金融机构落地大模型需要解决的问题。因此在数据准备阶段，涉及数据获取、数据脱敏/数据处理、数据清洗和数据标注等复杂工作，在此过程中尤其需要注重数据隐私保护，确保数据安全和符合隐私法规。

在模型层面，金融行业业务复杂度更高，金融领域的决策和分析通常要求精准的回答和实时的响应，对模型推理的推理速度和精度都有较高的要求。如果金融大模型/生成式AI做出虚假的、误导性的陈述，或推理与响应速度较慢，就会造成严重的决策损失和较差的用户体验。在应用人工智能技术时，大模型因其黑盒效应（复杂的模型结构和庞大的参数，难以线性化表达），可解释性、透明性及安全性也有待提高，金融机构需着重解决大模型的安全性和可解释性、透明性，以防范模型和算法风险。

在安全与合规层面，金融领域对于数据安全、监管合规和风险控制具有严格的要求，需要遵守各种法律法规和国家金融监管机构的要求。大模型在应用中必须确保符合风控和合规要求，防止欺诈、洗钱等非法活动，并保护客户利益。同时，随着网络攻击手段的不断升级，大模型在部署和运行过程中需要采取严格的安全措施，防止数据泄露、篡改和非法访问。此外，在当前市场中，围绕提升客户体验、增强数字化经营能力，深度服务客户已成为推动金融机构发展的关键要素，金融行业大模型落地各个环节需要以客户体验为中心，且GenAI应用又促使金融服务模式及客户体验的升级。

2.2 金融行业大模型应用落地面临的挑战

大模型在金融机构中的应用场景广泛，且应用成本较高，所需关注的安全合规问题较多，金融机构需以谨慎的态度去推进大模型的应用落地。如何选择合适的应用场景及如何推进其在金融场景中的有效落地，是当前金融机构在大模型应用落地中重点关注的问题。

- **首先，在应用成本考量方面，**金融机构训练模型需要大量的算力资源，资源调度需要使用更优化的硬件设备来提升训练速度。尤其是在处理千亿级参数的大模型时，对算力的需求更是呈指数级增长，其所投入的成本也较高。根据IDC调研显示，算力限制及技术投入成本高是金融机构在推进大模型/生成式AI过程中的最主要的两大阻碍因素。

- **其次，在应用场景选择方面，**在推进大模型落地时，有哪些大模型应用场景可供选择，如何选择合适的大模型落地场景是诸多金融机构面临的问题，需要重点考虑模型方案（如模型选择、模型适配性、模型能力域及性能、模型更新速度等），并面临很多数据难题（如数据质量、数据可用性、数据安全及合规等）和业务难题（如业务需求、应用价值评估、以及ROI等），在此过程中也需要考虑应用场景的优先级及推进策略问题。

- **同时，在应用路径选择方面，**金融机构在推进大模型落地时，面向不同的应用场景有着不同的应用路径，例如，自主开发和预训练的金融大模型、优化基础模型推进金融业务场景的落地、通过标准化SaaS模式接入GenAI应用，如何选择适合金融机构的应用路径，以及在推进大模型在具体场景应用时所需考虑的要素及所需具有的能力，这些都是金融机构亟需面对和解决的问题。

总体来说，生成式AI虽然可以提供低成本、高价值的解决方案，但在应用成本考量、应用场景选择、应用路径选择等方面仍面临诸多问题。金融机构需综合考虑应用场景选择、成本控制、安全合规等多方面因素，采取科学、谨慎的策略，以实现技术创新与业务发展的双赢。

第三章

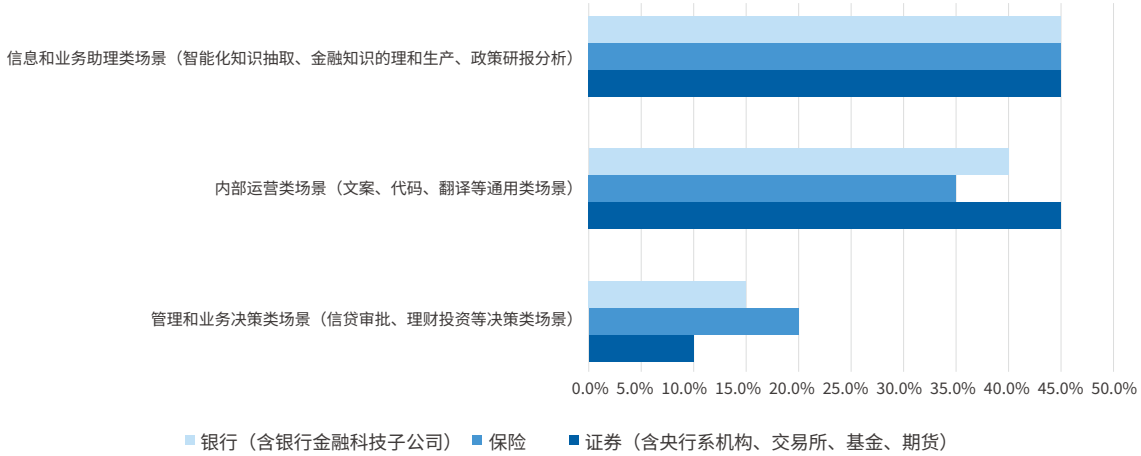
落地进展
大模型催生效率变革
金融行业务实求效

3.1 大模型在金融行业的典型应用场景梳理

金融行业生成式AI通用类应用场景梳理

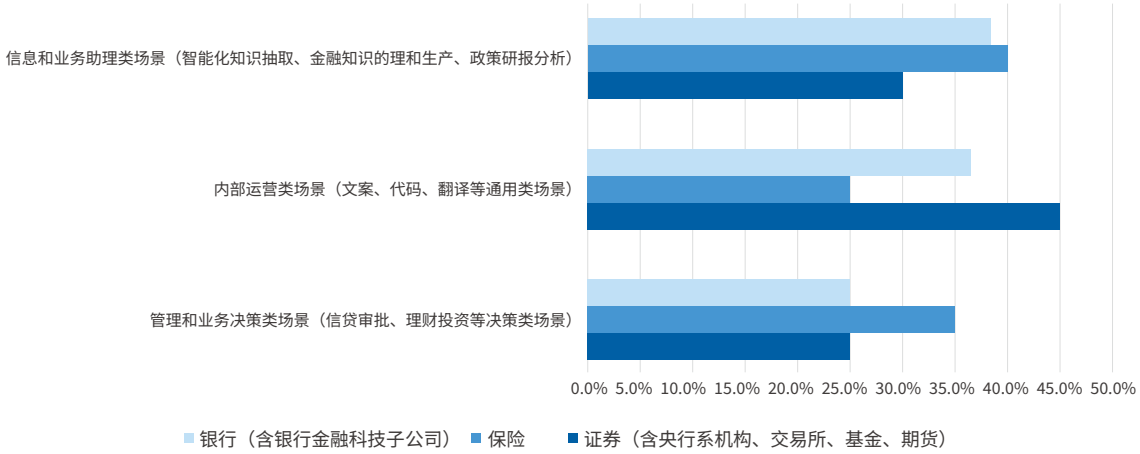
大模型/生成式AI在金融行业应用具有极高的潜力和价值，当前诸多金融机构正以大模型/生成式AI技术的工具辅助、信息处理、业务决策等特性，应用于内部运营类场景（文案、代码、翻译等通用类场景）、信息和业务处理类场景（智能化知识抽取、金融知识的理解和生成、政策研报解读）、管理和业务决策类场景（信贷审批、理财投顾等决策类场景），从而为金融机构带来运营效率提升、产品/服务模式创新、客户体验提升等价值。根据IDC调研数据显示，在当前，信息和业务处理类场景及内部运营类场景是当前金融机构主要的应用方向。而在未来12个月，管理和业务决策类场景的应用比例有所提升，尤其是保险机构在该类场景的应用进程较银行及证券机构相对更快。（如图3、图4）

图3 目前，贵公司应用落地最多的大模型/GenAI用例是什么？



来源：IDC，2024

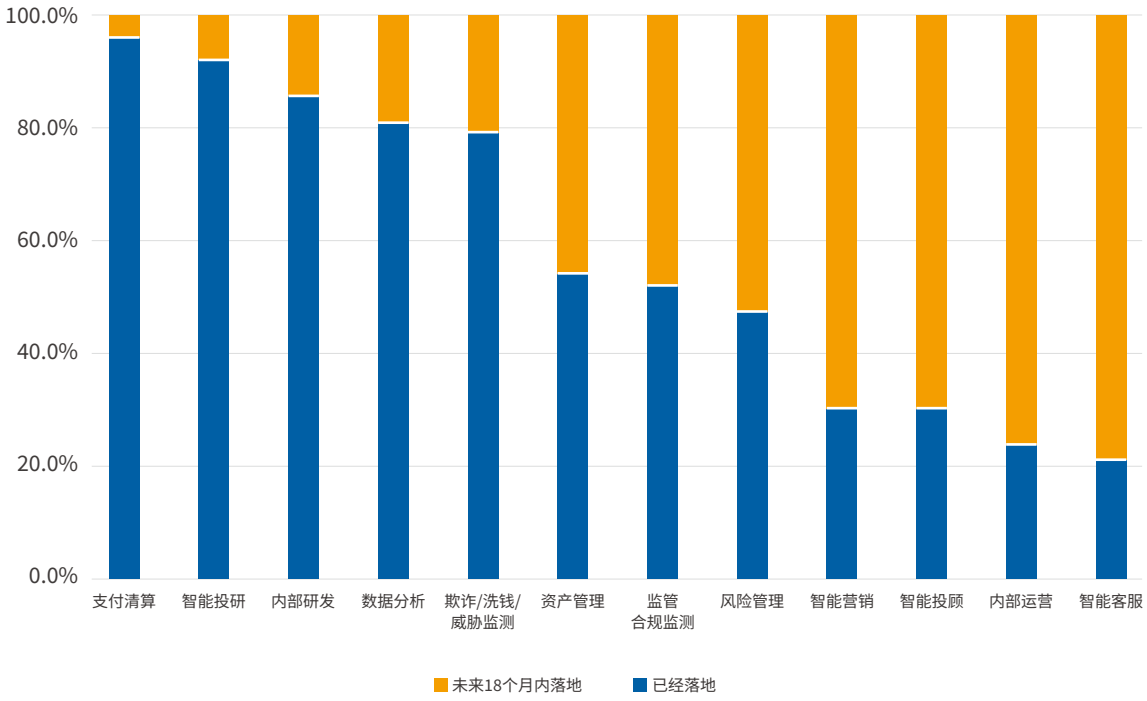
图4 贵公司在未来12个月内应用落地最多的大模型/GenAI用例是什么?



来源：IDC，2024

同时，根据IDC调研数据显示，目前及未来18个月内，金融机构落地大模型/生成式AI的场景按照调研统计比例如下图所示。其中，智能客服、内部运营（搜索与问答、知识管理/内容创作、舆情管理、HR等）、智能投顾/财富管理、智能营销（内容营销、产品营销等）以及风险管理（风险评估、风险识别、风险预警等），是金融机构当前落地较成熟的场景（按照调研比例从高到低排序）。而在未来18个月内，支付清算、智能投研、内部研发（代码生成、测试等）、数据分析（报表生成与分析、数据建模、数据决策等）、欺诈/洗钱/威胁监测、资产管理（资产尽调、资产评估及定价等）是金融机构主要的落地场景（按照调研比例从高到低排序）。

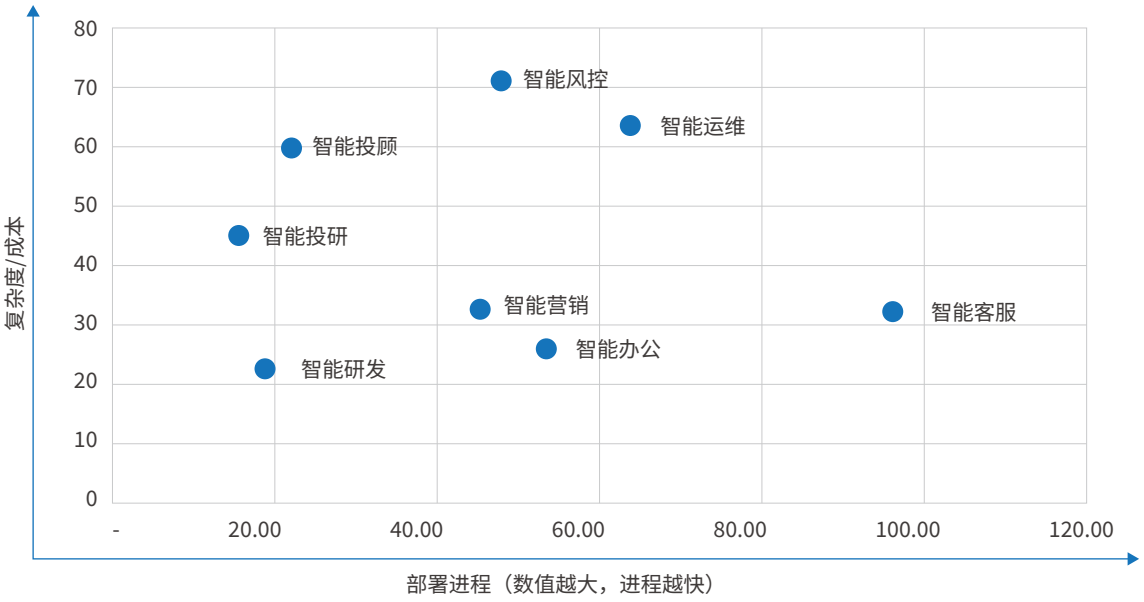
图5 贵机构当前及未来18个月内的大模型/生成式AI应用场景情况



来源：IDC，2024

图6是IDC根据调研结果，并从复杂度/成本、部署进程维度列出了金融行业主要场景的分布图。其中智能客服、智能办公、智能营销等场景落地复杂度较低、应用进程较快；而智能投研、智能投顾、智能风控等场景落地复杂度较高，应用进程也较慢。

图6 生成式AI在金融场景部署进程及应用复杂度概览



来源：IDC，2024

银行业生成式AI应用场景梳理

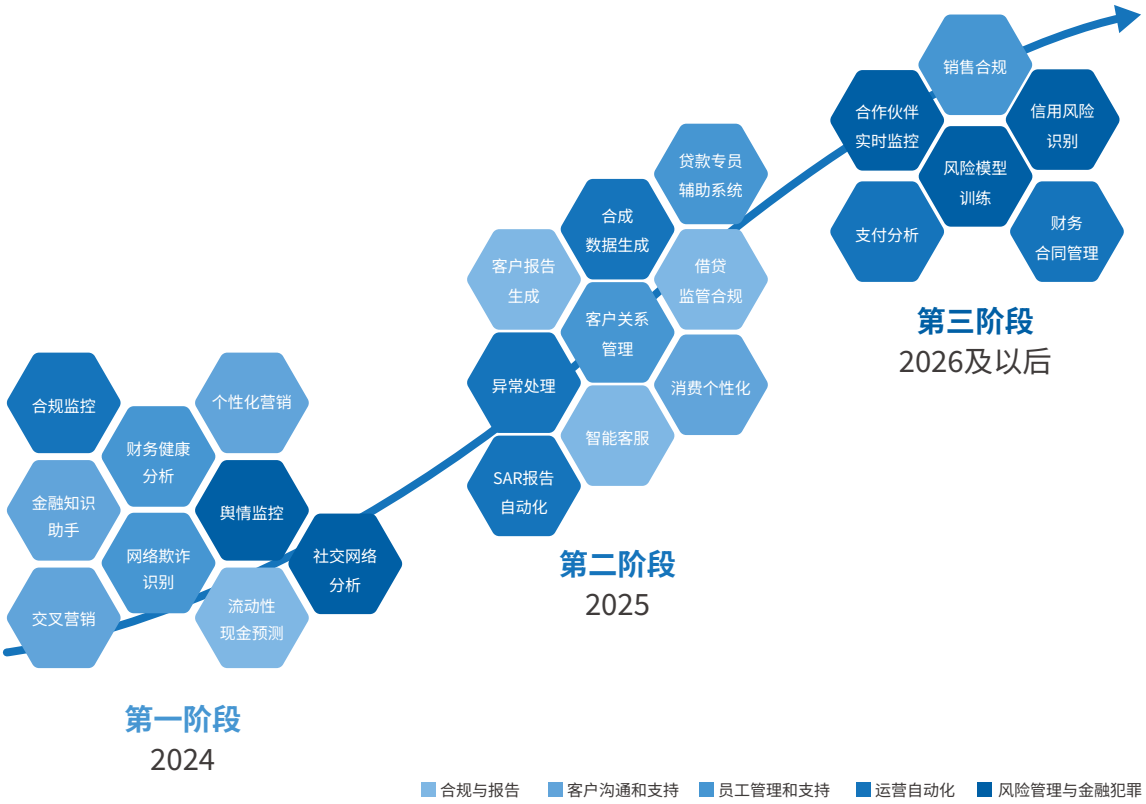
IDC认为，生成式AI的行业应用往往都是循序渐进的过程，一般是逐渐从内部辅助运营到外部对客提效、从业务边缘到核心；相应地，AI对金融机构的价值也逐渐增大。如下图，IDC认为，银行业生成式AI应用可以贯穿到银行业务链条的各个环节，包括从面向员工的管理和支持到面向市场的数字营销和运营自动化、从产品研发到风险合规管理等环节。

- **第一阶段（2024年）的应用场景有：**金融知识助手、财务健康分析、舆情监控、合规监测、网络欺诈识别、现金流动性预测、个性化营销等。以现金流动性预测为例，借助于GenAI，银行可以快速识别、分析和解释市场信号、走势和评论（如监管/决策层的评论），将其转化为准确、可靠和可操作的领先指标，从而为银行/客户提供投资、信贷、流动性和风险方面的建议和决策。

- **第二阶段（2025年）的应用场景有：**合成数据生成、客户报告生成、智能客服、贷款专员助手、客户关系管理、SAR报告自动化等。以贷款专员助手为例，GenAI可以通过访问客户账户历史记录，评估其需求/偏好，并就未来的贷款和其他银行产品提供营销建议。贷款专员在通过多渠道为客户提供信贷审核或信贷产品推荐时，商业银行可以通过相关GenAI应用为其提供培训工具和带有风险提示及营销策略的信息。

- **第三阶段（2026年及以后）的应用场景有：**风险模型训练、支付分析、信用风险识别、财务合同管理/财务预测、销售合规等。以风险模型训练为例，商业银行通过使用包括开放、非结构化数据源、合成数据来预训练或优化风险模型，以便为客户群提供风险决策支持，最大化减少风险。例如，在智能投顾场景，借助于预训练大模型能够对金融文本进行整体认知和理解，消除人为的主观因素，提供客观的投资建议，同时也能不断演进和创新，减少对人工审核的依赖，给出风险警示和解决方案。

图7 银行业生成式AI应用路线图



来源：IDC，2024

保险业生成式AI应用场景梳理

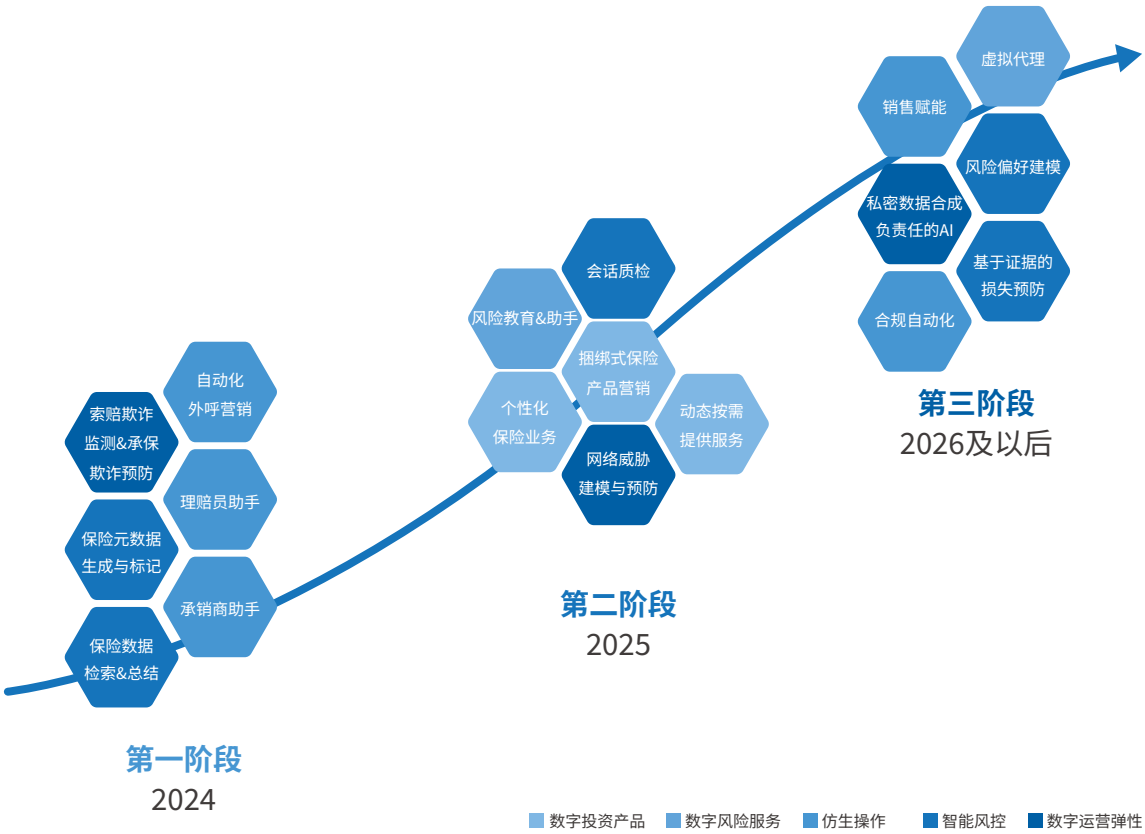
如图8，在保险行业，生成式AI的应用主要分为数字投资产品、数字风险服务、仿生操作、智能风控以及数字运营弹性等类别。其应用主要也分为三个阶段。

- **第一阶段（2024年）的应用场景有：**保险数据检索&总结、保险元数据生成与标记、理赔员助手、承销商助手以及自动化智能外呼和索赔欺诈监测等场景。以保险元数据生成与标记为例，GenAI通过解析和生成元数据层、掌握语义关系和主题标记，学习大量数据集的元数据模式，自动生成符合规范的元数据。这包括但不限于文档标题、描述、关键词、分类标签等，有效减轻了人工标注的负担，提高了整体数据质量。

- **第二阶段（2025年）的应用场景有：**风险教育&助手、捆绑式保险产品营销、网络威胁建模与预防、会话质检、动态按需提供服务等场景。以个性化保险业务为例，保险机构通过开发出基于机器学习和GenAI评估单个商用车风险的工具，工作人员可以分析不同的数据，如车辆类型和事故历史，进行细致的风险评估，从而实现个性化定价和个性化保险产品推荐。

- **第三阶段（2026年及以后）的应用场景有：**私密数据合成&负责任的AI、合规自动化、风险偏好建模、虚拟代理以及基于证据的损失预防等场景。以虚拟代理为例，GenAI通过赋能数字助理提供量身定制的解决方案，彻底改变了个人和企业的风险管理。从评估财产和负债风险到优化投资，它都可以提供个性化的指导，使用户能够做出明智的决策，并有效地保障客户的财产/资产。

图8 保险行业生成式AI应用路线图



来源：IDC，2024

证券与投资业生成式AI应用场景梳理

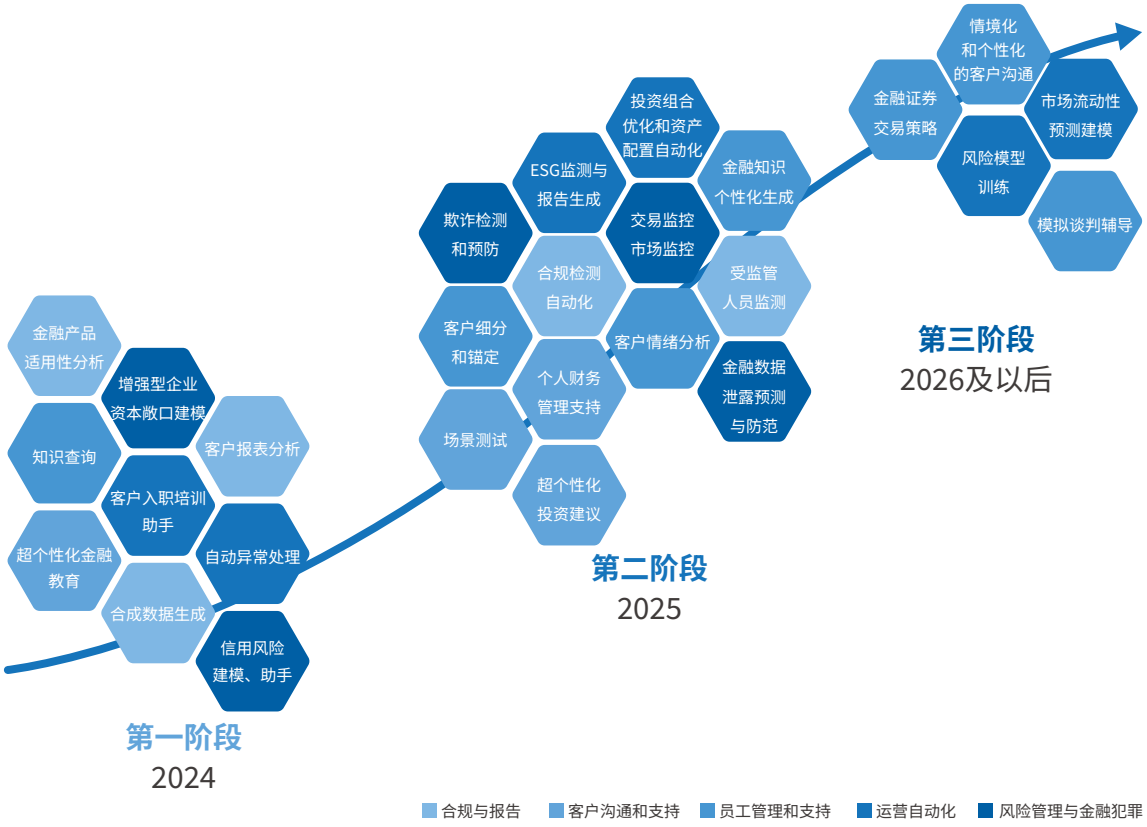
如图9，生成式AI在证券与投资业的应用场景较为丰富，其主要分为合规与报告、客户沟通和支持、员工管理和支持、运营自动化以及风险管理和金融犯罪等类别。在推进生成式AI应用过程中，也分为三个阶段。

- **在第一阶段（2024年）的应用场景主要有：**知识查询、金融产品适用性分析、超个性化金融培训、客户入职培训助手、客户报表生成、合成数据生成、信用风险建模助手等场景。以金融投资产品适用性分析为例，生成式人工智能可以帮助金融专业人员根据模拟场景分析客户数据，例如历史购买金融投资产品的对象特征及风险偏好，分析潜在的投资目标，同时帮助金融机构识别并控制投资风险，根据风险状况生成量身定制的建议，并确保符合适用性法规。

- **第二阶段（2025年）的应用场景有：**超个性化投资建议、客户细分和客户锚定、欺诈检测和预防、合规检测自动化、交易监控&市场监控、投资组合优化和资产配置自动化、客户情绪分析、金融知识个性化生成，以及金融数据泄露预测与防范等场景。以投资组合优化和资产配置自动化为例，生成式AI可以根据不同的市场状况和历史趋势，并根据不同客户的风险偏好、投资期限、流动性需求等，从而生成多样化的投资组合建议。

- **第三阶段（2026年及以后）的应用场景有：**金融证券交易策略、资产智能定价、情景化和个性化的客户沟通、市场流动性预测建模，以及模拟谈判辅导等场景。以资产智能定价为例，生成式AI通过分析市场数据、实时和历史数据以及其他相关输入来生成准确、动态的定价模型和高度精细的估值方法，从而实现智能资产定价。

图9 证券与投资行业生成式AI应用路线图



来源：IDC，2024

3.2 生成式AI在金融行业场景应用流程梳理

在大模型在金融行业落地中，场景选择难是诸多金融机构的痛点，如何选择GenAI应用场景，使AI能力与业务场景无缝融合，让GenAI真正赋能于业务提效、成本节约、业绩提升或体验升级，充分发挥GenAI应用潜力。IDC认为，金融行业生成式AI应用场景的落地可以从场景应用评估与选择（如技术解决方案评估、项目管理及风险评估、投资回报分析）、以及面向场景的工程化落地（如模型选择、技术路线、数据及算力准备、模型训练及调优、以及系统集成与部署、组织协同等）分步推进。

场景应用评估与选择

一般可以从技术解决方案、项目管理及风险评估、投资回报分析等角度评估与选择GenAI的具体金融应用场景。**但是从谨慎的角度，金融机构可遵循由简单到复杂、由内而外、由点及面、逐步推进的原则选择与推进金融大模型的应用场景。**同时通过下述评估方法确定采用GenAI后在相关的金融场景能发挥哪些潜力，并明确自身的业务目标和需求，例如提升客户体验、提高运营效率、降低成本等。

- **技术解决方案评估：**

  主要考虑现有技术可否实现该场景应用的业务目标或愿景，尤其是 GenAI 技术在场景中的应用还面临着一些挑战（如2.2节所述），在技术解决方案的评估方面，亦需重点考虑GenAI是否可以解决金融业务场景中的需求或痛点，现有技术或资源（如基础设施资源、基础模型资源、AI平台资源等）可以解决哪些问题，哪些技术/模块需要自建或外采，哪些需通过与技术合作伙伴共同构建，从而综合考虑业务实施可行性——即概念验证、解决方案的试行版本、前期工作、创新方案。

- **项目管理及风险评估：**

  项目实施决策既要着眼于场景，也要考虑基于GenAI的实施方式，下列因素将有助于确定某个项目或项目集是否行得通：与当前战略的契合度、风险管理、成本及资源、数据及基础设施资源、市场盈利潜力及长期价值，以及监管与合规挑战等。

图10 项目管理及风险评估维度

来源：IDC，2024

- **投资回报分析（ROI）：**

  GenAI 在业务中的实际价值和 ROI 将取决于 GenAI 在实现成果方面的表现。但是，这些成果可能是有形的，即可以用特定KPI衡量的回报，也可以是无形成果，即成果比较抽象，难以通过 KPI 或收益来衡量，但通过网络效应或其他方式带来了可观价值。以知识管理为例，某金融机构将金融和法律相关的数据和深度知识集成到一个基于数据的GenAI应用中，形成特定主体的数据和深度知识，呼叫中心的客服专员使用此工具可以直接回答客户金融和法律领域几乎各方面的问题，因此缩短了平均处理时间，提高了首次呼叫解决率。该应用用到的数据集涵盖数十万份金融和法律文件，使40个国家/地区的多达15,000名员工能够即时访问需要的数据。该解决方案还提高了员工对产品的了解，从而也提高了客服接触点互动中的交叉销售和追加销售机会。

面向场景的GenAI工程化落地

IDC认为，GenAI的场景应用是一项系统工程，涉及模型选择、技术路线选择、数据及算力准备，以及模型训练及调优、系统集成与部署、组织协同等工作。

表2 面向场景的GenAI工程化落地要素及要点

类别	描述	要点
模型选择	金融机构根据自身业务需求和技术实力选择生成式AI模型	• 基础大模型（如GPT系列） • 行业大模型 • 特定场景下的定制化模型
技术路线	深度研发大模型、基于现有大模型进行工程化适配，或直接通过SaaS化模式使用大模型服务	• 自主研发：自主性强，但需高投入 • 工程化适配：快速响应，但定制化受限 • SaaS化：即插即用，但数据安全与定制化需考量
数据及算力准备	数据准备需高质量数据、定制化采标数据、合成数据及企业内部数据；算力资源需充足且稳定	• 数据集多元化减少算法偏见 • 大量定制化数据提升模型精度 • 部署与模型适配的算力基础设施
模型训练及调优	使用训练平台、开发工具、调优工具进行模型训练，并通过微调、提示词工程、RAG等方式优化	• 调整超参数优化模型性能 • 微调、提示词工程等适应特定金融场景 • 提供部署&推理工具保障模型运行
系统集成与部署	将生成式AI/模型集成到业务流程或产品中，选择合适的部署方式	• 自动化、智能化生成式AI功能实现 • 同时根据业务需求和技术环境选择本地化部署或云托管等方式
组织协同	扁平化组织架构、跨部门合作、灵活项目管理方法推动生成式AI应用	• 促进快速创新与决策 • 高效推动AI在各部门中的使用 • 确保技术与业务需求紧密结合

来源：IDC，2024

第四章

金融行业大模型的应用路径与关键能力

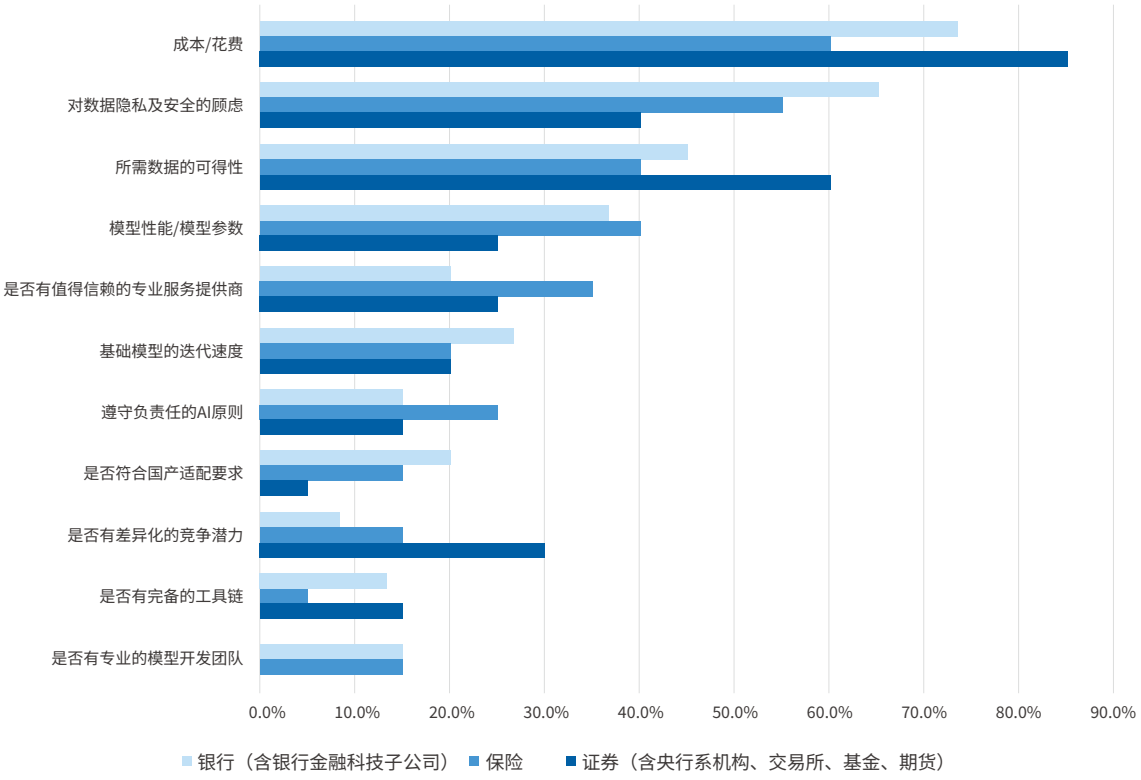
当前，不同类型的金融机构在推进大模型的落地中，有着不同的路径选择。有金融机构自主开发和预训练的金融大模型；基于通用大模型并通过提示词设计、模型微调、检索增强生成等方法，优化基础模型推进金融业务场景的落地；通过API按需接入各类金融大模型或通过标准化SaaS模式接入GenAI应用。金融机构可根据自身战略目标、业务需求、技术能力、资源禀赋、风险偏好来决定是否自主建设、基于已有模型微调，或是采用其他方式来利用GenAI能力。

4.1、金融机构落地大模型的应用路径

路径1：金融机构自主主导开发和训练大模型

该路径往往需要金融机构投入较大的IT资源和人力资源，与之对应的金融应用场景也较为复杂，需要大模型在金融领域专业知识、术语及政策等方面具有专业的理解能力；这些场景对模型精度、模型安全，数据的可用性、丰富性、安全性，以及业务合规等都有较高的要求，故自主开发和专门训练的金融垂类大模型可能会更好地满足这些需求。根据IDC调研数据显示（如图11），金融机构考虑为GenAI构建自有模型的主要考虑因素有成本/花费、对数据隐私及安全顾虑以及所需数据的可得性。

图11 以下哪三项是贵机构考虑为GenAI构建自有模型的主要因素?



来源：IDC，2024

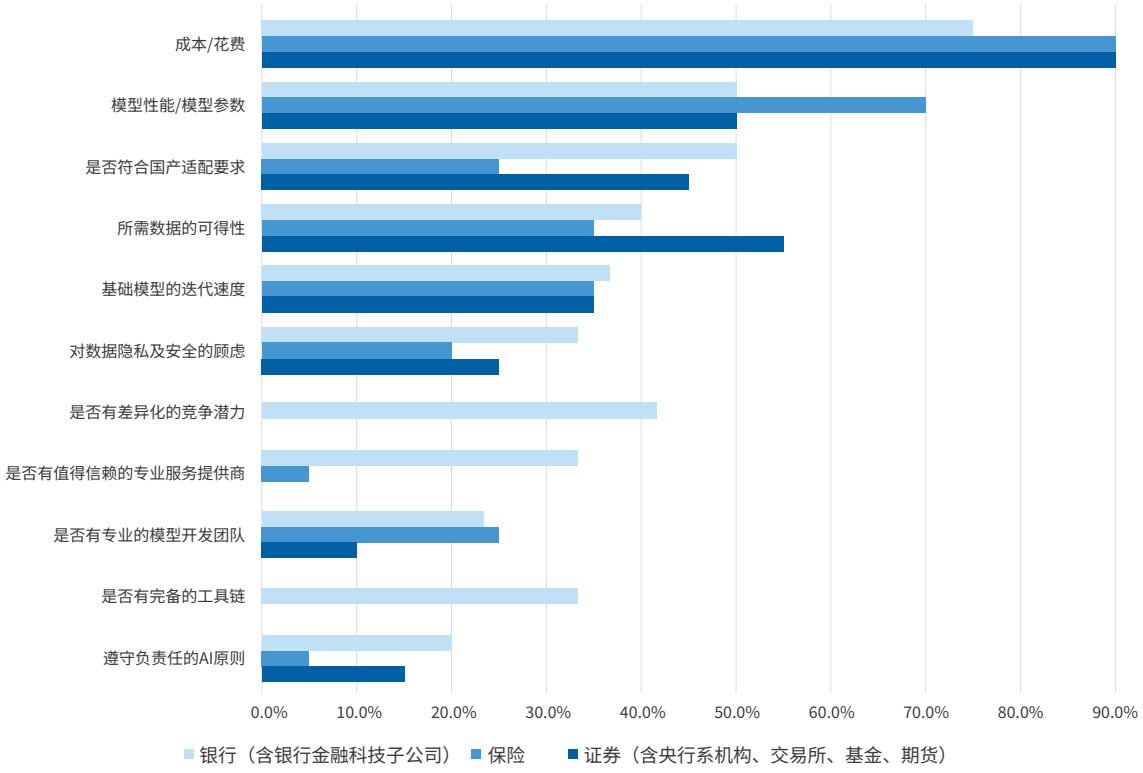
例如，某证券与投资服务公司，基于其海量的数据资源，如宏观经济数据、行业经济数据、企业研究报告、上市公司信息披露等结构化和非结构化数据，以及交易所、政府部门、科研机构、高等院校、专业行业数据公司等机构提供的授权数据，同时为了提高模型的通用能力，该公司自主研发了大模型，从训练语料、训练框架到模型结构的设计，均从零开始、创新性地构建基础模型及金融大模型，并且在预训练阶段就融合了金融领域的语料，而不是在微调阶段。这种做法使得模型在知识学习上更为深入，能够理解金融领域中的复杂关系和概念。此外，为了加速AI能力面向各场景的深入落地，该公司构建了AI开放平台，目前可面向客户提供短视频生成、文章生成、数字虚拟人、智能金融问答、智能语音、智能客服机器人、智能质检机器人、会议转写系统、智能医疗辅助系统等多项AI产品及服务。

路径2：基于通用大模型/开源模型，叠加金融服务领域数据，通过使用参数微调、提示词工程/检索增强生成等方式优化基础模型推进金融业务场景的落地

这种路径是金融机构基于基础模型，通过私域数据集进行模型训练调优，以实现金融机构金融大模型的建设，该方式在特定任务上表现出来的性能和在特定领域知识的深度理解会更强。在IDC的调研中，金融机构为GenAI使用第三方现有模型的主要因素是成本/花费、模型性能/模型参数、以及数据可得性及数据隐私等（如图12）。

尤其是在数据层面，向量数据的管理和合成数据生成是金融机构着需解决的问题。其中，向量数据库等创新技术现在已成为 GenAI 数据价值链的关键部分，包括数据管理流程的整个生命周期,已被用于检索增强生成，因为它们可用于对存储为高维数据的非结构化数据集执行搜索，可以轻松地与现有数据库集成，从而为从LLM实施 GenAI 解决方案提供了一种更灵活、更高效的方式，以高效用于各种应用场景，例如产品推荐、异常检测和业务分析等。而合成数据生成涉及创建模拟真实数据特征的人工数据集。在分析中，它用于解决隐私问题和数据稀缺问题，或模拟各种测试场景。通过生成代表性数据，可以在不泄露敏感信息的情况下开发和完善模型。这种方法可以加速创新，提高模型性能，尤其是在处理有限或敏感的数据集时，该方法特别适用。

图12 以下哪三项是贵机构考虑为GenAI使用第三方现有模型的主要因素?



来源：IDC，2024

在模型训练及调优方面主要通过微调、提示词设计、检索增强生成(RAG)等方法，增强基础模型输出的准确性、知识实时性。

- **微调：使用金融领域数据和人工监督来调整预训练的模型，以提高金融领域的模型性能。**

  选择该方式的考虑因素：具有特定领域性能的高复杂性用例，例如智能投顾、支付分析、风险模型训练等。

- **检索增强生成（RAG）：将金融领域的文档集合与预先训练的模型相结合，使输出情境化，而不涉及LLM。**

选择该方式的考虑因素：金融机构拥有该领域的专有数据和标签数据。适用的应用场景包括内容搜索/金融知识问答、事实调查/欺诈识别/风险监控、内容生成/营销助手等用例。

- **提示工程：使用提示技术来影响预训练模型生成输出的准确性。**

  选择该方式的考虑因素：适用于不需要特定领域的上下文的用例，同时允许用户级控制产生特定任务输出的场景，以便客户可以更加精准地获得其想要的知识。

未来，随着诸多基座模型的开源，以及一系列低成本的微调/检索增强生成等技术的出现，将有越来越多的金融机构，会基于其需求定制专属大模型。

例如，某国有银行从算力、数据、大模型、场景等维度推进大模型的落地：在算力层面，通过构建一体化云原生的异构算力平台来管理和调度多元异构的AI算力资源；在数据层面，围绕“采建管用”闭环，构建大模型训练和持续提升的基础数据闭环；在大模型方面，其基础模型是采用第三方开源模型，包括业界主流的开源模型以及正在做共建和联创的产业大模型，各个基础模型之间可以无缝切换，具有灵活的适配性，而对于金融大模型平台，包括NLP、CV、多模态等大模型，主要是通过组件化（例如微调组件、RAG组件等）方法快速优化各类开源模型和商业模型，以实现各类金融场景的接入，目前已在智能客服、智慧三农、智能营销、智能运营、智能风控等场景进行探索及应用。

路径3：按需接入各类大模型API（按需付费的SaaS订阅模式）

这种路径主要是以SaaS模式面向金融机构提供服务，通过将生成式AI模型能力封装为API服务接口，金融机构以外采订阅的形式，可以通过API将GenAI模型嵌入自研应用软件增强智能化水平，或是基于API创建定制化的全新智能应用，通过嵌入式AIGC应用，进行场景变革或产品升级。

选择该路径推进GenAI应用的场景往往是那些标准化程度较高的通用类场景，金融机构通过API接口的形式将第三方生成式AI内嵌于应用之中，可以开箱即用。比如，将生成式AI内置于HR SaaS软件，实现以智能问答的交互形式服务员工。

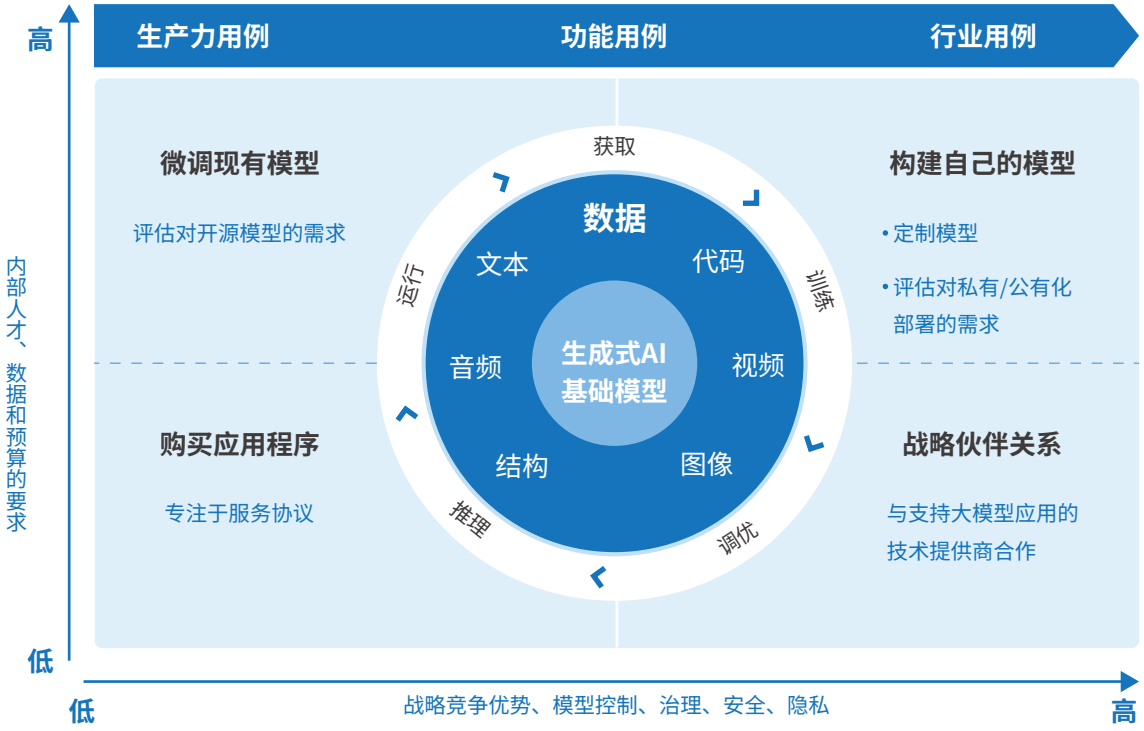
例如，某金融机构通过使用某云厂商合作（即：金融机构外采模式），利用云厂商的全栈解决方案（从计算基础设施到数据库，再到金融场景服务）来帮助金融机构构建其GenAI支持的SaaS，所有这些都在云服务环境中完成。如此客户就无需启动大量工具和服务来开发GenAI功能。该方法在数据保护方面，主要是通过在第三方系统中托管数据库以满足金融机构对数据安全性及合规要求。

路径4：与战略合作伙伴协同推进GenAI的场景落地

金融机构也可以与AI基础大模型厂商、AI大模型平台服务商、以及AI应用开发和集成服务商、咨询/服务商等战略合作伙伴合作推进GenAI的场景落地。例如，在与基础大模型厂商合作方面，通过选择与业务应用场景相匹配的基础大模型，例如如果金融机构需要将大模型应用在面向零售业务的智能客服场景，则其应优先选择在自然语言处理方面有较大优势的基础大模型。在与大模型平台服务商合作方面，通过与AI平台服务商合作构建多种AI工具和能力（如自然语言处理、图像识别等多种能力和工具），用于支持AI应用的开发与部署，并集中管理AI模型、运维和治理确保AI系统的稳定运行和合规性。在AI应用开发和集成服务商方面，通过选择利用厂商在该领域的AI应用开发与集成能力，如计算机视觉（CV）、自然语言处理（NLP）、知识图谱（KG）和文档智能（Doc AI），结合具体场景来为金融机构大模型的落地提供应用开发与集成服务。

综合来看，不同路径的选择对于金融机构内部人才、数据以及预算的要求等各不相同；同时，鉴于金融对专业性要求较高，且需要遵守各类流程和规范，每一种路径对于战略竞争优势、模型控制以及安全和隐私等方面的影响均有不同。**上述各个路径所需考虑/执行的决策要点可以结合图13示例。**

图13 应用路径的决策框架



来源：IDC，2024

4.2 金融机构选择或部署大模型时的关键能力要素

IDC认为，金融机构在落地大模型的过程中，需要综合考虑数据价值链管理、模型的选择与部署、AI平台搭建、以及AI治理等要素。

数据价值链管理：提升数据可用性和数据质量

金融机构的数据价值链管理是生成式AI在金融场景中有效发挥价值的基石，其核心目的是提升数据质量、数据可用性以及确保数据的合规获取，有利于金融机构面向不同的应用场景快速构建高质量的数据集，并为后续金融大模型的规模应用奠定坚实的基础。目前市场中已存在通过构建多模数据管理平台DMS或是通过数据管理全流程解决方案，以更好地满足大模型时代的用数需求。

生成式AI数据价值链负责监督基础模型、微调模型/金融行业模型使用的训练数据的获取、生成、处理和管理，以确保全面、多样化和高质量的训练数据的可用性。同时，生成式AI数据价值链管理能够显著提升模型输出的准确性和安全性。通常，数据价值链往往包括：数据采集、数据存储、数据标注、数据精益、训练数据生成、数据验证、数据保护、数据监控、数据管理、数据分析等。

表3 数据价值链管理要素

序号	数据处理阶段	描述
1	数据采集	从各种内部和外部来源收集原始数据的过程，可能涉及收集大量文本、代码、图像或其他相关数据源。
2	数据存储	支持多种类型的存储需求，包括结构化数据和非结构化数据（如图像、音频和视频等）。需确保可扩展性、性能、数据安全性和隐私保护。
3	数据标注	对选定的数据集进行注释或标记，以加快监督式机器学习过程。虽然不是所有生成式AI用例的必要步骤，但有助于确保数据整合和模型推理的准确性。
4	数据精益	确保嵌入到Prompts中的数据高质量、可靠、及时、准确、完整，并适用于其使用上下文。通过数据清洗和数据质量标准提升数据质量。
5	训练数据生成	准备训练或转换基础模型的精选数据子集，确保数据质量和多样性满足训练需求。
6	数据验证	采用人工审核、多渠道数据验证和用户反馈机制等方法，确保数据准确性，并纠正模型可能出现的准确性问题。
7	数据保护	在推理或模型转换过程中，保护个人或公司敏感信息不被泄露。进行数据安全和隐私分类，确保隐私和安全，促进创新，并保持合规性。
8	数据监控	使用标准、策略和指标实时监控数据使用及流动情况，确保主要活动中的数据隐私和安全性。
9	数据管理	提供处理和访问大量数据的标准工具和技术，如数据管道、数据复制等。确定数据所有权，控制业务域、应用程序和分析中数据的访问、使用和维护。
10	数据分析	运用数据挖掘、统计、商业智能和预测工具来理解数据，为AI开发人员和数据科学家提供重要支持。

来源：IDC，2024

模型层面：模型的选择与部署的考虑因素

在模型选择方面，往往涉及如下问题，例如是否选择开源模型、选择何种开源模型、选择何种模型开发及模型优化方式、如何部署模型等，金融机构可以根据业务需求、任务类型、数据量级、以及业务场景的技术投入和其对安全合规等因素去选择合适的模型。尤其是需要评估大型模型开发与金融业务场景之间的匹配度，了解模型在实际场景中的应用方式及应用价值，确保其符合业务需求（具体可参考3.2节中关于应用场景选择与评估的方法）。

表4 模型选择与部署考虑因素

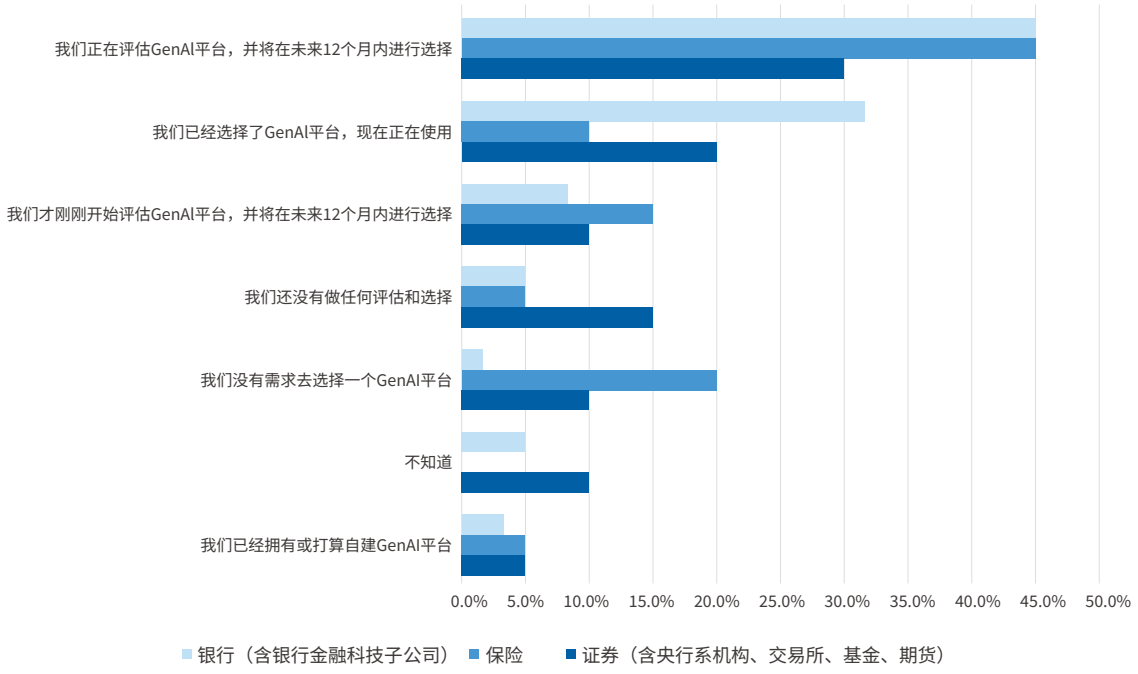
选型方案	考虑因素/优势
开源模型	开源模型通常已经过验证，可以快速部署和测试，能有效降低开发成本，提升开发速度，适用于场景验证及成本投入不高的机构。而在开源模型选择方面，通常需要根据具体需求选择性能最优、功能最符合的模型，同时确保所选模型与现有技术栈和工具链兼容。
闭源模型	开发成本较高，可满足高定制化、安全性需求，例如金融行业部分应用场景对模型的透明度和可审计性有严格要求，这时可能需要闭源模型。
模型部署方式	金融机构可以根据业务类型，如贷款、投资、保险等，以及不同业务场景下的模型应用需求及其对数据安全与合规要求等因素，考虑具体的模型部署方式。通常，私有化部署具有数据安全性高、自主可控性强、可定制化程度高等特点。公有云部署成本最低、灵活性高，但是数据安全性相对较低。
模型优化方式	金融机构在选择提示词设计、微调、检索增强生成（RAG）等方法的考虑因素及场景选择方面可以参考4.1节中相关内容的描述。

来源：IDC，2024

AI平台：从模型管理到应用搭建的一站式开发与服务平台

由于金融行业大模型应用场景丰富，随着大模型与证券、保险、银行业务的融合，将迸发出大量的GenAI应用开发需求，亟需AI平台来提高模型构建及编排效率、应用开发部署效率。根据IDC调研数据显示（图14），大多数金融机构受访者表示，他们已经选择或正在评估使用GenAI平台来帮助其开发、运营和管理GenAI模型及应用。

图14 以下哪一项最能描述贵机构当前或预期使用GenAI平台来帮助您开发、运营和管理GenAI模型及应用?



来源：IDC，2024

领先的云供应商应在其AI平台中添加多种模型，以满足客户对数据集、参数和开放的各种需求。在算法模型库的建设方面，组织需要将开源的算法，自研的算法等都统一管理。在模型训练方面，搭建基于CPU和GPU的分布式训练框架。在模型推理方面，统一实现离线批量预测和在线预测的功能，包含推理加速、资源管

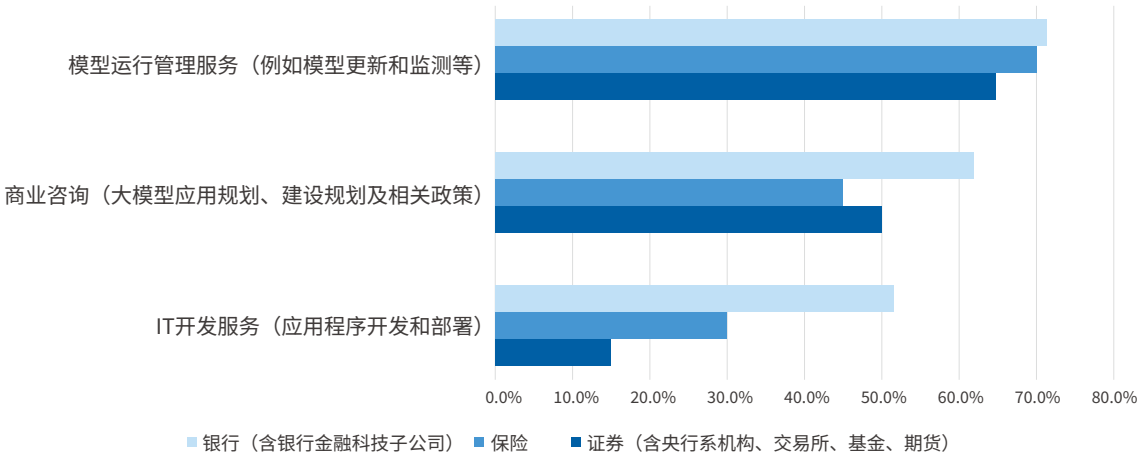
理等。基于这些基础组件的能力，搭建一站式AI开发工作流，从特征筛选到特征处理、样本构建、模型训练调试评估、模型的部署和实验，再到后期模型的运维监控，贯穿算法开发的整个流程。IDC认为，未来生成式AI开发平台将向更普惠的MaaS演进，并加速生成式AI应用的落地。

因此，搭建一套大模型工具链（包括提示词管理、多种 PEFT 微调方法的集成以及一键式RLHF训练工具）以实现面向不同的应用场景实现模型优化，也是十分必要的。这些工具包含大模型优化及应用扩展能力，使大模型能够更有针对性地服务特定应用。例如，某证券与投资公司，在推进大模型落地中所推出的AI开放平台包含模型开发、模型调优、推理加速等能力，目前可面向客户提供短视频生成、文章生成、数字虚拟人、智能金融问答、智能语音、智能客服机器人、智能质检机器人、会议转写系统、智能医疗辅助系统等多项AI产品及服务。

同时，为了提高模型在具体业务场景中的推理准确性及安全性，推理时始终需要高质量、最新的数据，以提高模型返回内容的准确性和相关性。而且，在推理过程中，所使用的信息流也存在暴露敏感信息（个人或公司）的可能，金融机构可通过AI插件来监控模型推理过程中的数据泄露问题。

此外，为了更有效的推进大模型在金融业务场景中的应用，金融机构也可以自建或通过与外部厂商合作，提供面向场景共享复用、开箱即用的组件或能力，以及模型运行管理服务（例如模型更新和监测等），从而加速金融机构的GenAI应用。根据IDC调研数据显示（如图15），大多数金融机构受访者表示，预计在未来12个月内将使用模型运行管理服务（例如模型更新和监测等）来推动大模型的应用落地。

图15 预计在未来12个月内您所在的组织将使用哪种专业服务来推动大模型的应用落地?



来源：IDC，2024

例如工商银行就通过打造适配金融行业的“1+X”工程化解决方案，其中“1”是指智能中枢平台，通过智能中枢的任务感知、决策、执行、反馈等能力实现金融复杂场景的应用；沉淀“X”可共享复用的范式能力，包含多模态知识检索、对话式数据分析、智能化文档编写、交互式智能搜索、陪伴式智能研发等多项金融即插即用的零代码工程化解决方案，以高效赋能于金融业务。

同时，金融机构在选择外部服务商搭建GenAI平台时，可以提供丰富的开箱即用的能力以支撑多样场景的落地，GenAI平台的可扩展性、安全性，以及在金融行业的大规模生成式AI落地经验是其最看重的三项能力。目前，中电金信也帮助多家金融机构开发了GenAI平台，例如其向某机构构建了企业级人工智能研发平台和研发体系，通过沉淀AI原生应用的研发规范，打造AI场景应用的标杆，并建设基于大模型的服务管理平台，在已有数据中台基础之上集成公文文档和外部情报数据，开发了信息情报工作站、智能公文写作两个场景应用；在另一个典型案例中，帮助某头部城商行构建了人工智能融合中台，通过建设大模型平台，实现底层资源统一池化管理、提供一站式的数据工程、模型训练调优、大模型评估、推理加速和提示词工程等工具链能力，有效支撑了全行的大模型应用开发。

AI治理：构建大模型在金融场景稳定且安全的应用保障

组织需要采用 360 度治理视图，包括数据、人员、AI和使用案例。其主要包括模型治理、风险治理、以及满足负责任的AI的要求。

在模型治理方面，其需要解决模型幻觉、模型偏见以及模型的审计跟踪及其可解释性等问题。例如，可以通过增加数据多样性和规模、改进模型结构、建立反馈机制等措施来应对模型幻觉；通过数据清洗和预处理、多样化数据集、公平性评估等措施来应对模型偏见；通过记录模型训练过程、模型版本控制、输出日志记录等措施来加强审计跟踪；通过使用基于prompting 范式的模型解释、基于分类器进行探测等模型解释工具、以及可视化技术等方法来提高模型的可解释性。

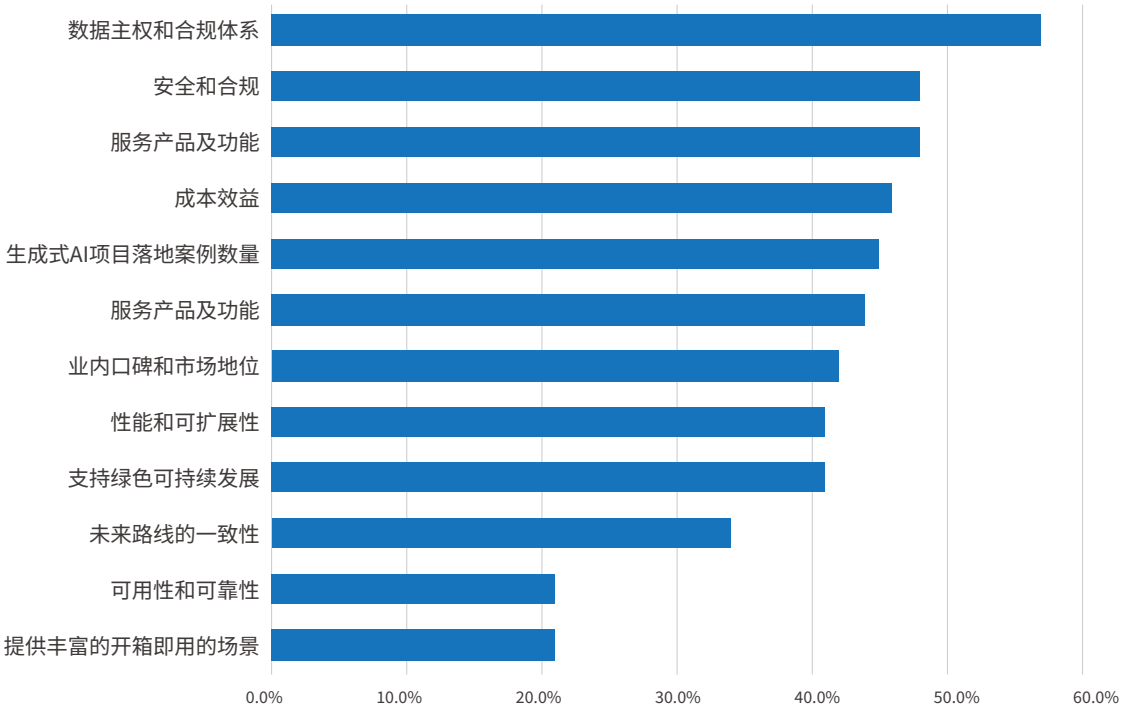
在负责任的AI方面，IDC将负责任的人工智能（RAI）定义为以恪守严谨安全的监管要求、遵循公平透明的行业规则、维护用户为先的价值取向的方式进行大模型和GenAI的设计、开发和部署。金融机构在推进大模型落地中，秉承负责任的AI原则，有助于确保所有操作符合法律法规，减少违规风险和潜在的法律诉讼，并辅助实现维护金融系统稳定的社会价值。金融机构在推进负责任的AI方面，一方面，在开发AI时注重公平性，避免偏见确保AI系统具有透明性，决策可以被解释，另一方面，在AI系统造成伤害发生时，也应有相应的纳入问责和补偿机制。

在大模型应用链条中的风险治理方面，在大模型的应用前先分析预判可能存在的各项风险，金融机构可以将AI模型风险纳入整体风险管理框架，并建立专门的AI监管报送平台、流程和规范，集成多源数据，包括模型训练数据、运行数据、监控数据等，实现数据的统一管理和分析，同时可以根据业务需求，将平台划分为不同的功能模块，如数据采集模块、风险识别模块、报告生成模块等，实现灵活配置和扩展，及时披露模型决策机理、运行逻辑和潜在风险。

此外，金融机构也可以选择合适的技术服务商，通过生态伙伴之间的能力协同，从而加速GenAI加速嵌入金融业务场景。例如，通过与技术服务商（如ISV）合作，加速构建GenAI应用,并通过相关的咨询服务、模型建设及调优、应用二次开发、运

营管理和培训服务等一系列服务，以帮助金融机构更好地利用AI技术，让AI更好地在金融场景中见效。根据IDC调研结果显示（图16），金融机构在面向大模型落地选择技术服务商时，最看重的前三项能力是数据主权和合规体系的构建、金融安全解决方案优势、以及服务产品和功能（如：大数据分析和数据治理等）。

图16 在选择技术服务商方面，目前比较看中下列哪些方面能力?



来源：IDC，2024

尤其是在数据主权和合规体系建设方面，由于在金融领域，数据的敏感性极高，直接关系到客户隐私、商业机密乃至国家安全。因此，金融机构必须确保对其数据拥有绝对的控制权，即数据主权。技术服务商需能够提供完善的数据管理机制，确保数据在采集、存储、处理、传输等各个环节中都能被金融机构有效掌控。此外，随着全球范围内数据保护法规的不断加强（如GDPR、中国《网络安全法》及《个人信息保护法》等），金融机构在选择技术服务商时，会重点考察其是否具备构建和维护符合国内外法律法规要求的合规体系的能力。这包括数据加密、匿名化处理、访问控制、审计追踪等措施，以及及时响应监管要求的能力。

第五章

展望未来
金融行业大模型的
发展趋势

5.1 大模型技术创新与发展趋势

随着大模型技术的发展，大模型的参数规模也将显著增长，如OpenAI的GPT系列模型、Google的PaLM，以及Meta的LLaMA。这些模型将具备更强的推理、生成和上下文理解能力。未来，这种趋势也将继续，其不仅能极大地提高大模型在具体场景中的应用性能，也能提高大模型在具体应用场景中的体验。

多模态模型也将在金融行业普遍应用。多模态模型能够处理文本、图像、音频、视频等多种类型的输入，并且通过多模态融合与跨模态理解，可以更有效地用于各类复杂问题的求解，从而极大地丰富其应用场景。例如，OpenAI的CLIP（将图像和文本映射到相同的向量空间）、DALL-E（通过文本生成图像）、Meta的ImageBind（支持六种模态，包括图像、音频、文本等）等。这些模型能够在图像识别、文本生成、音频理解等多方面表现出色。这种多模态之间的融合也将使得AI大模型能更深刻地捕捉复杂场景背景、细节和情感，使其更快的感知和适应场景，并能应用于更加复杂的金融场景。

AI智能体将迸发更大的应用潜力。AI Agent 是一种软件程序或计算实体，它能够通过传感器或数据接口收集环境信息，运用算法处理数据，制定决策或规划行动方案，并最终执行这些决策以影响环境，实现预定目标。AI智能体将成为下一代平台，从Copilot副驾走向主驾，智能体的加速进化，有望完成 “感知-认知-推理-决策-组织/行动”的闭环，其在数据处理、智能决策与自然交互等方面的卓越能力，预示着它将在客户服务、业务流程优化、市场预测等多个关键领域发挥核心作用，为金融机构带来前所未有的价值创造。为了实现在复杂场景中的业务决策，金融机构也可以引入Agent系统，以处理那些单靠提示词难以实现的复杂业务场景，进而达到高效的自动化决策。尤其是在智能客服、智能质检/陪练等产品创新，以及风险评估、个性化金融咨询以及智能评估场景，为用户提供更加精准、高效的服务。

开源服务与开放生态成为主流趋势。国内外大模型开放平台、开源模型/工具，能有效加速大模型技术演进，金融机构可以在开源基础上进行二次开发，满足其个性化需求，赋予中小金融机构使用前沿AI技术的能力，从而加速大模型在不同金融场景中的广泛应用。IDC认为，未来大模型在金融领域的生态化发展，例如通过构建大模型生态资源共享平台，向金融机构提供大模型应用所需的全套资源，包含算力等基础设施资源、通用大模型及各类专业领域小模型等多样化的模型资源、金融业务的各类场景应用资源，以及连接金融产品和服务终端用户，是金融机构大规模应用生成式AI的主要路径之一。

软硬件、工具之间协同也能优化降低大模型开发和应用成本。金融机构可以充分利用硬件加速技术、优化软件架构和构建灵活的工具链，有效提升计算效率，减少资源消耗。NVIDIA、Google、Apple等公司都在开发专用AI加速芯片，以提高大模型训练和推理的效率。同时，通过分布式训练、跨平台部署以及端到端工具链，使得模型能够更加灵活地适配不同的硬件资源，并可以随业务需求灵活扩展或缩减规模。例如，NVIDIA推出的CUDA和TensorRT配合GPU硬件，不仅加速了训练过程，还在推理阶段进一步提升了运行效率。或通过硬件加速器（如TPU）和软件框架支持的优化算法，模型可以在推理中以较低的计算量实现较高的性能，从而推动大模型高效训练与部署。

5.2 行业应用场景的拓展趋势

当前，阻碍大模型在金融行业应用的主要因素之一是高昂的算力成本。在训练大模型或推理大模型过程中，金融机构需要消耗较大的算力。通过使用分布式计算、云计算资源及高效的硬件设施（如GPU/TPU），将大模型的训练任务分配到多个GPU或TPU上，以并行处理的方式加快训练速度，可以降低模型训练的时间和成本。同时，通过优化模型架构，精简网络结构、减少层数或神经元数量，降低不必要的复杂性，有助于节省资源。通过构建模型管理平台，可以帮助金融

机构以模块化设计方式复用不同任务间的共享组件；通过构建可组合、可重用的模块，不仅可以简化训练过程，还能够方便地迁移模型到其他领域或任务场景中，进一步扩展应用场景、提升服务效能。

未来，随着大模型在预训练及推理过程中算力成本的降低以及模型性能的提升、模型架构的优化，其在管理和业务决策类场景（例如信贷审批、理财投顾等决策类场景）中也将发挥出更大的应用价值，将作为业务决策的辅助决策者，与人类共同完成复杂的分析任务。通过模型提供的数据洞察和决策建议，决策者可以更加客观、数据驱动地做出判断，提高整体决策水平。

此外，通过大小模型协同也能驱动金融机构在更加多样复杂的场景中的应用。金融机构应加强研究和推进大、小模型协同、生成式技术与传统人工智能技术协同，大模型与小模型协同发展；大模型在自然语言处理、计算机视觉等领域展现出强大的能力，但同时面临着高昂的计算资源需求；而小模型则通过精准的数据集和优化算法，能在特定任务上展现卓越性能，甚至超越某些大模型。尤其是在金融行业，大模型与小模型的结合也将为这些行业提供更全面、更高效的解决方案。大模型可以完成高维度、非结构化数据的初步处理，而小模型则在关键节点进行细化分析和快速决策，从而实现协同工作。例如金融机构在反欺诈过程中，既需要识别复杂的欺诈模式，也需要在交易发生时实时做出快速决策。由于大量数据的复杂性，仅依靠大模型会带来过高的计算成本和延迟。此时，如果通过大、小模型之间的协同，大模型对交易数据进行模式分析和风险评估，识别出潜在的欺诈特征；而小模型则在交易时进行快速筛查，验证是否符合风险特征库中的条件，从而进行实时预警，则可进一步实现风险控制和即时响应的平衡。

第六章

关于中电金信

6.1 中电金信公司介绍

金融数智化转型的领导者

成立于1995年，中电金信是中国电子控股的二级企业，作为全球领先的金融数智化咨询及软件提供商，以及重点行业数智化转型服务的专家，我们致力于利用数智科技创造更美好的世界。秉承中国电子“打造国家网信事业战略科技力量”的使命，中电金信通过持续的技术创新和参与国家重大工程，依托丰富的行业场景，构建了新型数字基础设施“源启”。我们为金融及重点行业的数智化转型和安全发展提供全面的技术平台、应用软件和专业技术服务，将中国的数智化转型最佳实践推向全球。

全球布局与专业团队

中电金信汇聚了4万名国内外员工，在全球31个城市设立了交付中心。29年的发展历程中，我们始终专注于行业需求，通过保障安全、加速创新、升级体验和优化运营，为客户持续创造商业价值。我们已经与160余家《财富》500强企业建立了长期合作伙伴关系。

合作共赢，聚焦成果转化

依托中国电子的核心技术优势和组织平台，中电金信联合科技领域的生态伙伴，以市场为导向，以研究院为载体，凭借强大的技术专家服务团队，融合先进技术与创新基因。我们研发形成了从数字基础设施、异构集群管理、AI智算底座到数智化应用的全方位解决方案，推动行业的可持续发展。

国际化视野与服务经验

多年的国际化战略布局和丰富的海外服务经验赋予了中电金信卓越的国际视野。我们拉通海内外及行业间的技术体系和市场，引入国外的先进技术及标准流程，更好地支撑金融及重点行业的数智化转型。同时，我们为中国企业走出去提供通道，以中国数智化转型的最佳实践服务全球。

质量与安全管理

中电金信坚持严格执行质量与安全管理原则，建立了符合国际标准的安全与质量管控体系。我们已通过CMMI ML5、ISO27001、ISO9001、ISO20000等认证，在软件研发的项目管理、质量管理和工程管理等方面均保持行业一流水平。

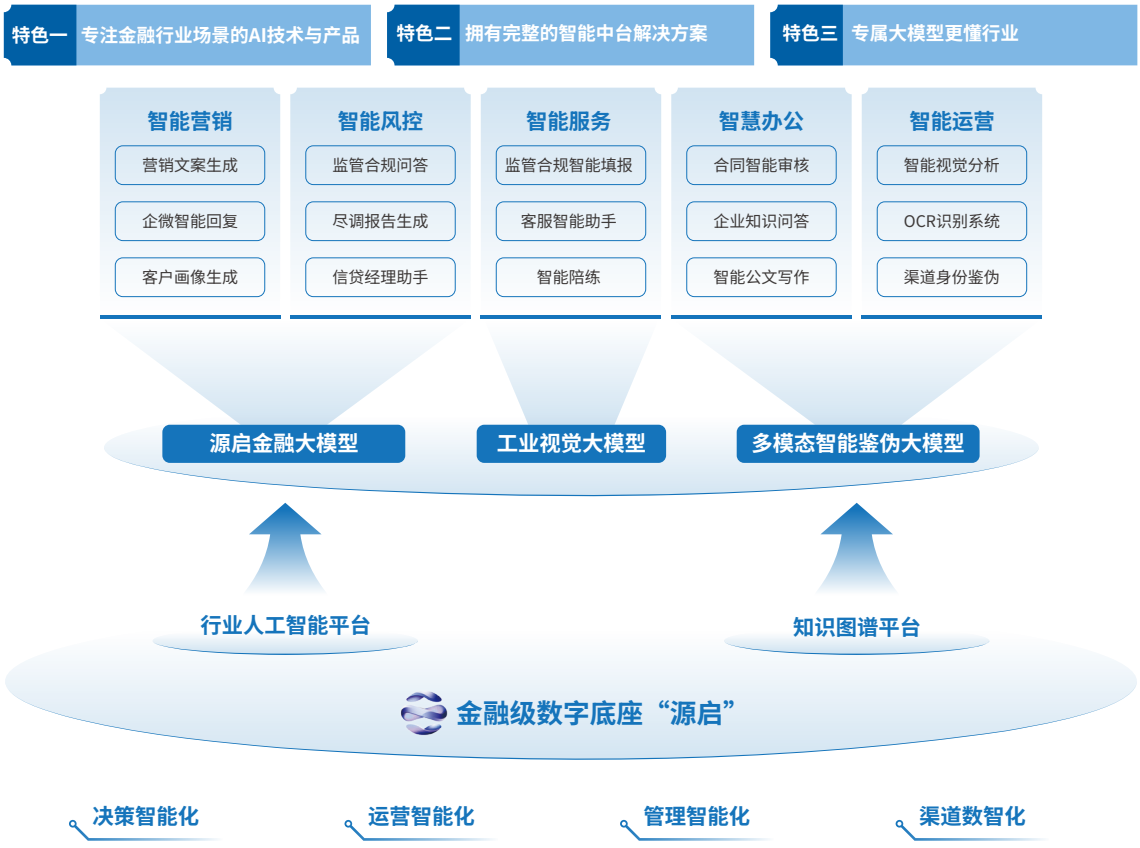
行业领导者地位

凭借出色的行业服务能力，中电金信在行业中确立了领先地位。自2017年起，我们连续7年位列IDC中国银行业解决方案市场第一名，连续10年入选IDC全球金融科技百强（IDC FinTech Rankings Top 100），连续9年入选中国软件和信息技术服务综合竞争力百强企业。

6.2 中电金信人工智能产品及能力介绍

中电金信面向金融、能源等重点行业，可提供智能平台、智能模型和智能应用的整体解决方案，帮助企业智能化转型升级。下图是中电金信AI领域的整体能力全景图，公司在人工智能领域重点围绕人工智能平台、金融大模型、计算机视觉、知识图谱等几大领域进行布局，通过多年的研发和沉淀，相关AI产品和能力已在一大批银行、保险、能源等行业客户中应用落地。

“两大平台+三大模型+N应用领域”，提供智能平台+智能模型+智能应用的整体解决方案能力



来源：中电金信，2024

智能算力底座与基础智算能力供给

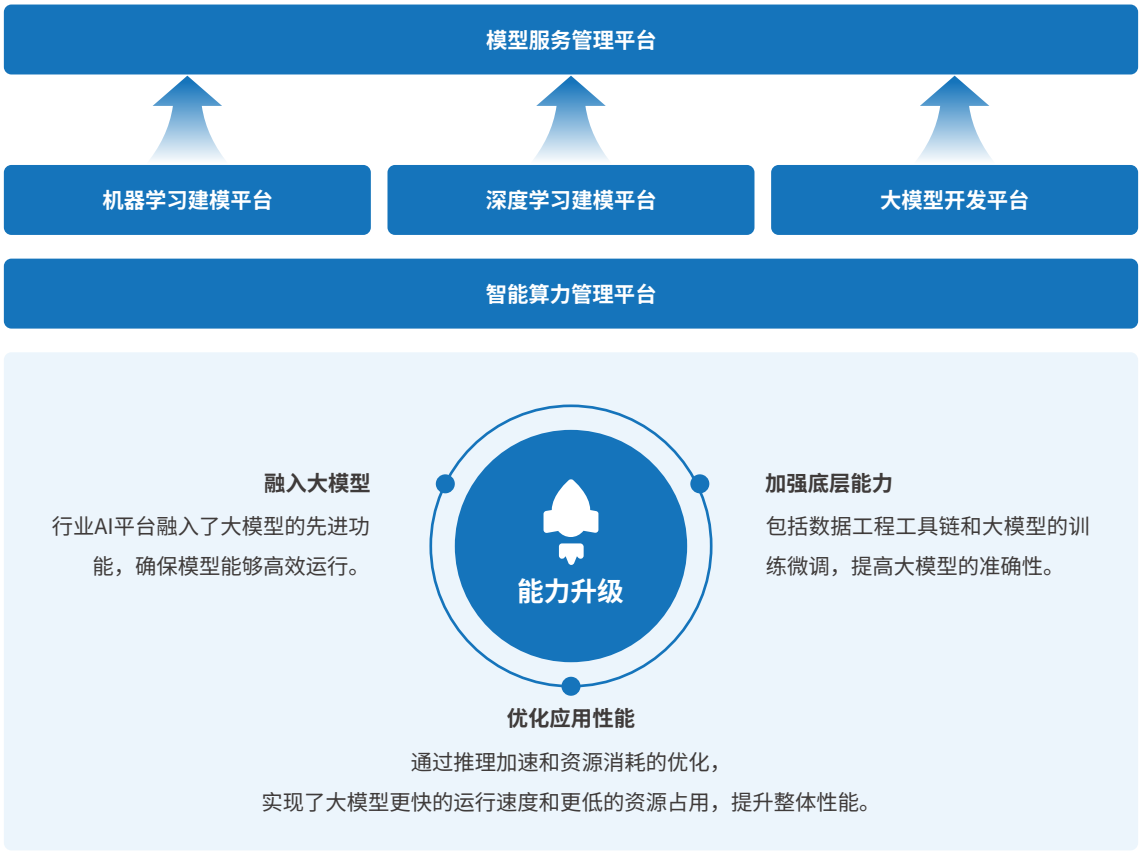
中电金信的智能算力底座是支撑人工智能应用和服务的基础设施，提供高性能、高可靠性和高可扩展性的计算资源。它集成了多种硬件和软件能力，并优化算力资源的接入、管理、调度和编排，形成了一个统一的计算平台。该智算底座能够为各种数据处理任务、训练框架、推理服务、智能应用提供所需的计算资源和软件能力，并显著提升AI工作负载的运行效率。

来源：中电金信，2024

在金融行业，智能算力底座的应用场景包括但不限于大规模的数据处理、智能模型训练、实时推理服务等。它能够为金融客户提供快速响应市场变化的能力，推动智能化转型。例如，在信贷审批、风险管理、智能投顾等场景中，智能算力底座能够提供强大的计算支持，实现秒级的决策响应，提升金融服务的效率和质量。

源启行业AI平台（含大模型平台）

源启行业AI平台是企业级的具有AI模型集中化生产和运营管理的平台型工具，帮助开发人员进行规模化模型开发，实现企业对模型和资源的统一管理和维护。

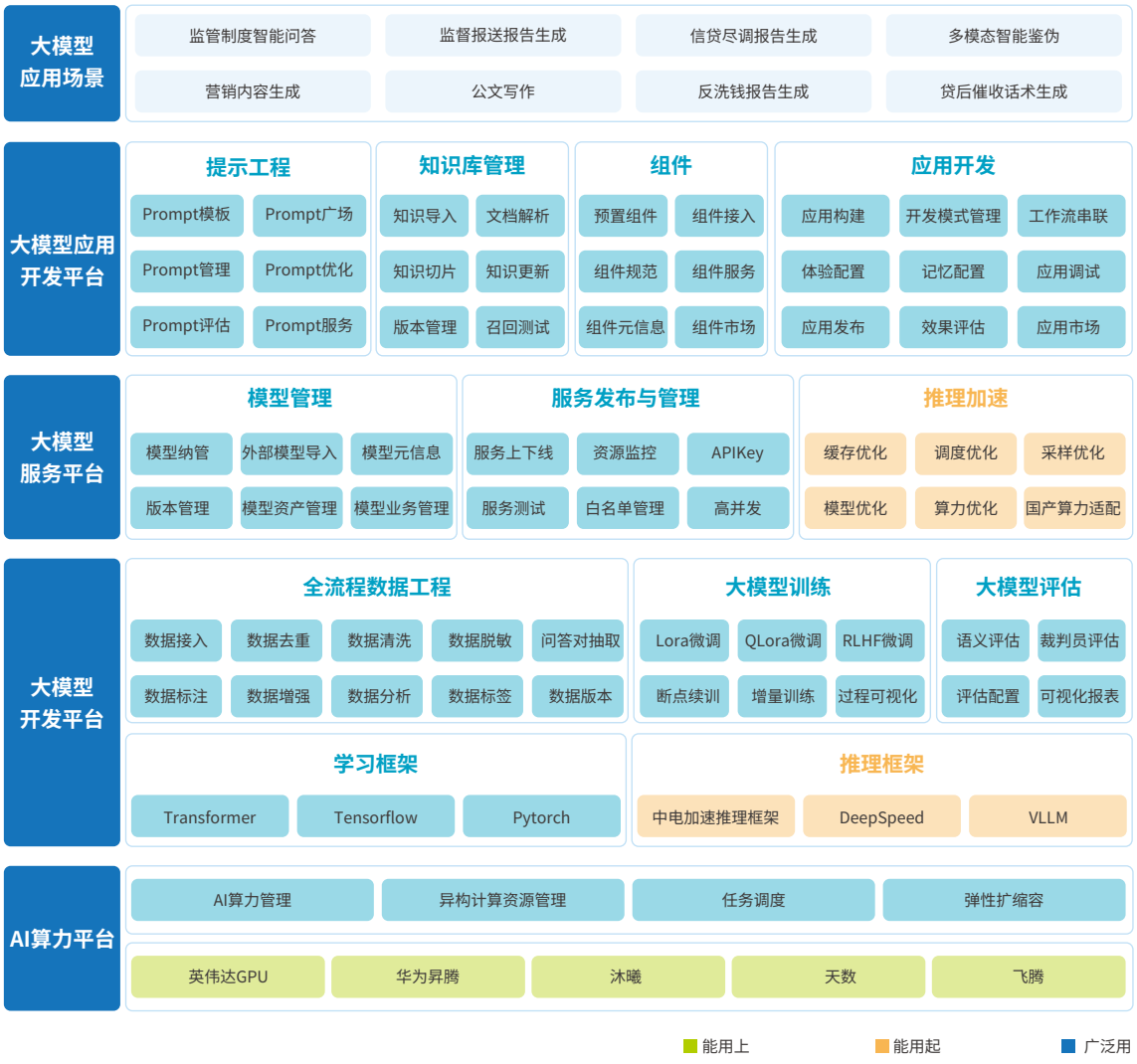


来源：中电金信，2024

其中，中电金信大模型平台涵盖了大模型数据工程、模型开发调优、部署与服务管理的全生命周期能力，大幅度降低了企业引入大模型的门槛。

该产品的核心亮点是：

- 支持完整和丰富的大模型数据工具链，支持大模型指令数据和偏好类标注；
- 大模型推理加速引擎，吞吐量和响应速度大幅提升。



来源：中电金信，2024

平台支持两种大模型训练方式：SFT训练和RLHF训练。除常规训练方式外，还提供断点续训和增量训练功能，支持容灾与模型能力的持续提升。

平台提供大模型语义相似度评估与裁判员评估功能，利用高参数模型作为裁判员评估低参数模型的效果，帮助选择高效、低成本的大模型。

此外，平台支持将大模型发布为在线服务，供业务应用调用，并支持大模型生命周期管理，包括模型文件、元信息和版本的统一管理。用户还可以导入自己的模型进行服务发布。

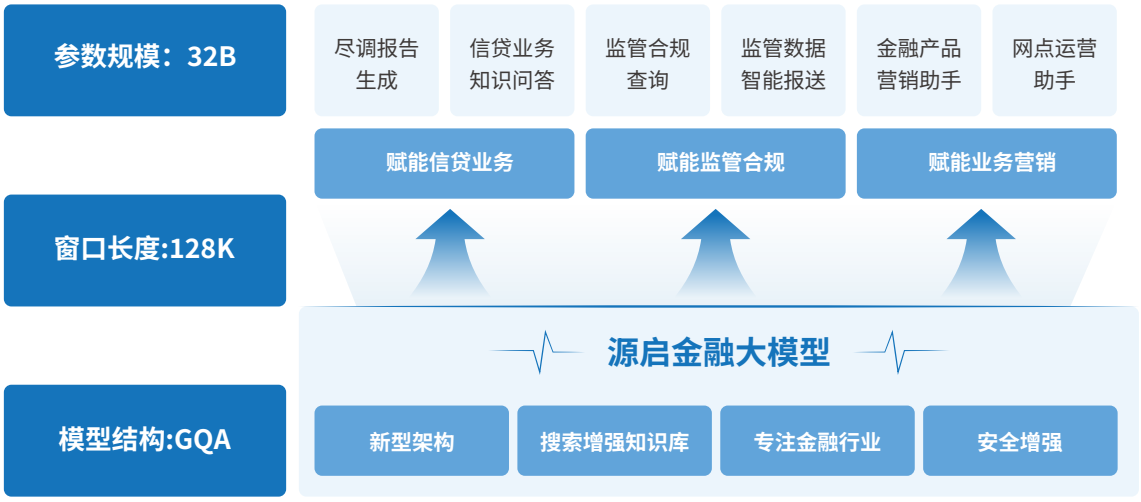
中电金信大模型开发工具链及数据标注服务是一套完善涵盖大模型训练、服务、应用开发的综合大模型平台，通过该平台，可以大幅降低企业应用大模型的成本，加速大模型在企业内的业务应用孵化，实现大模型在企业内的正循环。

源启行业AI平台在金融行业中可以应用于智能营销、智能风控、智能运营、智慧管理和智能服务等应用中，支撑金融和能源等重点行业决策智能化、运营智能化、管理智能化和渠道数智化，为行业带来全场景服务能力和全栈解决方案，助力金融等重点行业数智化转型。

来源：中电金信，2024

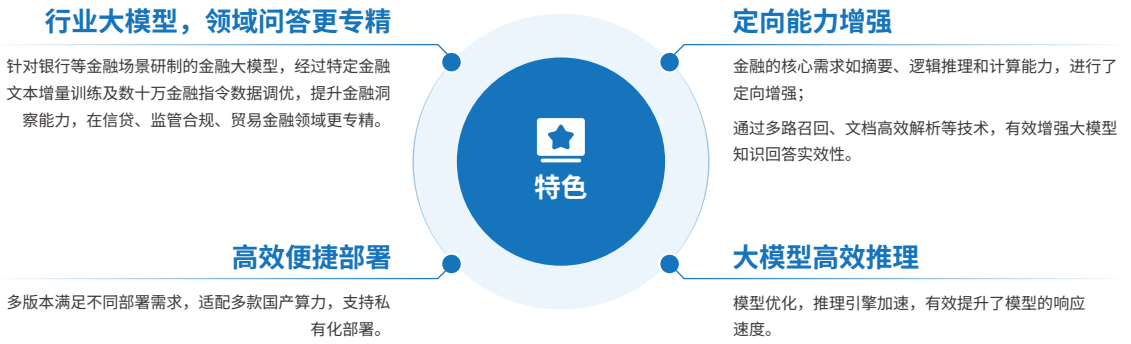
源启金融行业大模型

源启金融行业大模型是中电金信自主研发的L1级金融大语言模型，采用主流的Decoder-only模型架构，并在自有金融数据集上进行多轮训练迭代后得到。该模型能够服务于信贷、证券、银行、监管等多个金融业细分垂域，满足用户的问答、生成、计算等多种需求。



来源：中电金信，2024

源启金融大模型的基本信息如上，320亿参数、窗口长度128K，在金融领域具备以下特色：

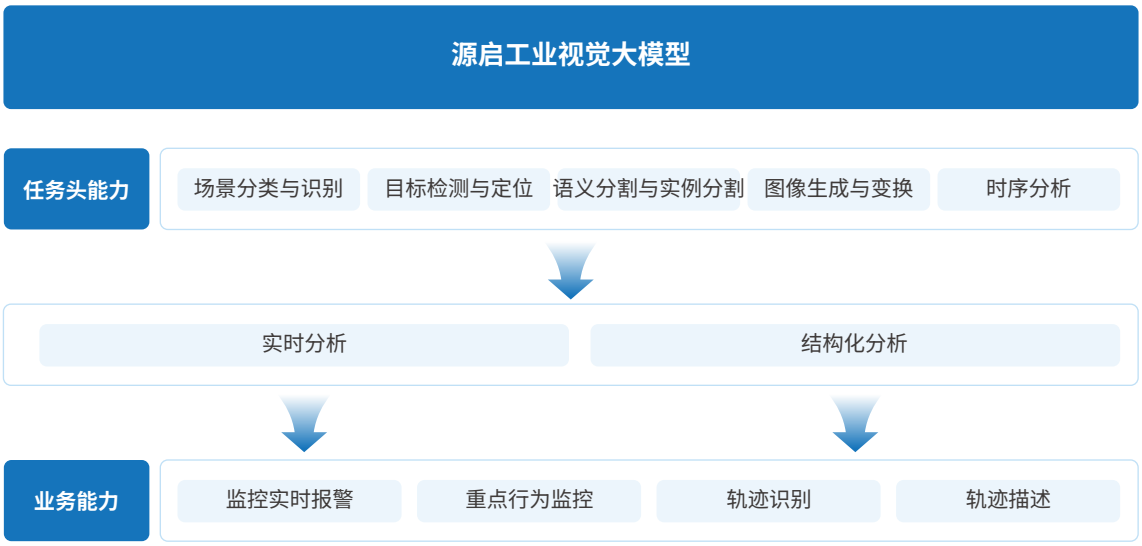


来源：中电金信，2024

源启金融行业大模型在金融行业中可以应用于多元业务场景。通过提供金融单轮、多轮问答、金融文本生成、材料分析总结、推理等能力，源启金融行业大模型能够显著提升金融服务的智能化水平，为客户提供更加精准和个性化的服务。

源启工业视觉场景大模型

源启工业视觉大模型使用了数十万的工业标注数据，采用了ViT/CNN等多种网络架构融合，训练具有更强大的视觉描述能力的模型，通过添加低参数的任务头网络进行微调，从而实现高效的任务性能。这种预训练微调的策略极大提高了视觉模型的性能，并且减少了训练时间和计算资源的需求。



来源：中电金信，2024

源启工业视觉大模型在工业领域中可以应用于质量检测、预测性维护、安全生产监控等场景。通过提供通用的工业行业描述能力，结合微调方法与不同任务头进行迁移学习，成功实现了在业务层面的实时报警分析和基于时序的结构化分析能力，显著提升了工业智能化水平。

源启多模态智能鉴伪大模型

计算机视觉技术的快速发展显著提升了深度伪造技术（如AI换脸、口型编辑、人脸重现等）的视觉效果，虽然为娱乐和媒体行业带来便利，但也带来了媒体文件造假和身份信息滥用的风险，这引发了学术界对媒体文件鉴伪的广泛关注。早期的鉴伪模型主要针对图像数据，采用空间域或频域方法识别伪造痕迹。如今，媒体文件通常包含视觉、音频和文本模态，单纯依赖视觉模态无法准确识别真假。为此，中电金信与复旦大学合作，提出了R-MFDN多模态鉴伪方法，综合分析视觉、音频和文本模态。R-MFDN由多模态特征提取器、特征融合层以及分类任务头组成。多模态特征提取器包括视觉特征提取器、音频特征提取器以及文本特征提取器；特征融合层则结合了自注意力机制、交叉注意力机制和前馈神经网络层；分类任务头包括二分类任务头和多分类任务头。

来源：中电金信，2024

源启多模态鉴伪大模型综合分析视觉、音频和文本模态，通过多模态特征提取器、特征融合层以及分类任务头组成。该模型能够识别伪造内容，包括AI换脸、口型编辑、人脸重现等深度伪造技术，保障渠道金融业务安全。

源启多模态鉴伪大模型在金融行业中可以应用于防范金融欺诈、身份验证等场景。通过识别伪造的音频、视频和文本内容，该模型能够有效降低金融欺诈风险，保护客户资产安全。

知识图谱平台

中电金信知识图谱平台集图谱构建与分析应用为一体的一站式解决方案，实现数据接入、图谱构建、查询探索、挖掘和服务的全流程管理。平台结合自然语言处理技术，支持从结构化和非结构化数据中快速构建业务图谱，具备可视化设计和拖拽式构建能力。

来源：中电金信，2024

知识图谱平台在金融行业中可以应用于智能审计、反洗钱、反欺诈等业务，通过深度挖掘企业和个人间的关联关系，提前识别风险，增强风控能力。在企业知识管理方面，平台支持构建领域知识图谱，提升内部知识搜索效率，优化决策流程。

智能体平台

大模型Agent是基于大型语言模型（LLM）构建的智能体，利用大模型的理解能力调度外部工具，弥补计算能力和知识更新的不足。Agent具有自主理解、决策和执行的能力，能够根据场景和数据灵活调整策略，模拟独立思考过程，逐步实现预设目标。在技术架构上，Agent从面向过程转变为面向目标，通过感知、思考与行动的紧密结合完成复杂任务。

智能体平台中的Agent能赋能多种场景，如个人助理和自动化办公，带来效率提升、成本降低和体验优化。通过集成大型语言模型、检索增强生成技术、自动化数据处理与分析工具，以及定制化的任务规划，Agent可以自动完成特定任务。

智能体平台在金融行业中可以应用于客户服务、风险管理、内部审计等场景。例如，智能体可以作为客户经理的助理，自动完成客户信息的收集、风险评估和投资建议的生成，提升客户服务的效率和质量。

RAG知识问答

中电金信的知识问答系统，依托于强大的自然语言处理技术和大模型技术，能够实现对非结构化信息的深度理解和智能问答。该系统通过自动化的文本数据标注、知识抽取、知识融合、图谱存储和图谱分析，提供全流程的知识图谱构建与服务能力。它能够处理和理解大量的数据，进行复杂的推理和决策，尤其是在金融数据分析领域，能够挖掘数据中的价值，提供革命性的升级。

来源：中电金信，2024

大模型RAG知识库问答系统已广泛应用于金融、能源、制造等行业，支持企业知识问答、智能客服、智能风控、投研支持和设备运维等场景。在金融领域，知识问答系统可以覆盖金融机构营销、渠道服务、风控、办公、研发等前中后台数字化经营关键环节。例如，金融知识助手能够提供即时的监管合规咨询，智能风控系统能够分析客户信用和行为，提供信贷风险评估和反欺诈预警。此外，该系统还能支持数据查询和投资建议，助力证券投研，以及帮助一线工程师精准运维、提高故障定位效率。

智能报告写作平台

中电金信的智能写作平台，基于自然语言处理和AI大模型技术，能够自动实现文档抽取、文档比对、文档摘要和文本审核等多种文本处理功能。该平台能够实时从海量文档和政策法规中检索最新、相关的素材，确保内容准确权威，提升写作质量。

来源：中电金信，2024

智能写作平台在金融业务中具有广泛的应用。它能够为公文写作、营销文案、信贷报告生成等提供智能化支持。例如，在公文写作中，平台能够自动生成符合规范的文档，如红头文件和条例细则，同时提供深度润色和修改建议，提升公文质量。在营销领域，智能写作能够根据市场动态和客户数据，生成个性化的营销内容，提升营销效率和效果。

6.3 中电金信AI大模型在金融行业的服务案例

案例一：某股份制银行模型运行管理平台项目

在数字化转型的浪潮中，某股份制银行面临着模型管理的挑战，需要构建一个全行统一的模型服务平台，以提升数字化经营能力。中电金信凭借其在金融AI领域的深厚积累，中标该银行模型运行管理平台项目，致力于打造一个敏捷部署、统一管理、集中监控的模型服务平台。

实施思路与方案

项目的核心目标是构建一个企业级的模型运行管理平台，实现模型的全生命周期管理。中电金信采用了源启·行业AI平台，整合了AI算力平台、AI计算框架、AI开发平台以及AI服务平台四大产品组件，为银行提供了一个集中化生产和运营管理的平台型工具。通过该平台，银行能够实现模型的快速开发、部署、监控和优化，同时支持模型的统一管理和资源的高效调度。

来源：中电金信，2024

在实施过程中，中电金信首先对银行的现有技术架构和业务流程进行了深入分析，确保新平台能够与现有系统无缝集成。随后，通过源启·行业AI平台的模型开发工具链及数据标注服务，帮助银行构建了专属的金融能力评测集——Gien-FinData评测数据集，进一步提升了模型的准确性和适用性。此外，中电金信还提供了模型性能监控和优化服务，确保模型在实际业务中的稳定运行和持续优化。

落地效应及价值

项目的成功实施，使得该股份制银行在模型管理方面实现了质的飞跃。模型管理平台的建成，不仅提升了银行的数字化经营能力，还为进一步深化合作打下了坚实的基础。通过集中化的模型管理和监控，银行能够更快速地响应市场变化，提高决策的效率和准确性。同时，模型的统一管理和资源的高效调度，也为银行节约了大量的运营成本。此外，通过Gien-FinData评测数据集的应用，银行的模型准确率和响应速度得到了显著提升，客户体验也因此得到了改善。

案例二：某头部城商行AI人工智能融合中台项目

随着AI技术的发展，某头部城商行寻求通过AI中台实现AI能力的全生命周期管理，以提升银行的智能化服务水平。中电金信凭借其在金融大模型领域的技术优势，承接该行AI人工智能融合中台项目，助力银行实现AI能力的全面提升。

来源：中电金信，2024

实施思路与方案

项目的整体定位为全行级AI中台，旨在实现AI能力的全生命周期管理，包括底层资源管理、数据融合管理、模型建模功能、能力融合层、场景运行层、AI门户等功能。在实施过程中，中电金信首先对银行的业务需求进行了深入的调研和分析，确保AI中台能够覆盖银行的所有业务场景。随后，通过源启·行业AI平台的AI算力平台和AI计算框架，为银行提供了强大的计算支持和模型训练能力。同时，通过AI开发平台和AI服务平台，银行能够实现模型的快速开发、部署和监控，以及模型服务的统一管理和调度。

中电金信AI平台提供数据管理、特征工程、模型开发、模型评估、模型部署、服务发布、模型监控、迭代更新等模型生命周期全流程功能，能够有效地支持客户数据创新场景应用开发与上线，应用包括全渠道拓客能力提升、客户满意度调研问卷词云分析、楼盘均价合理性评估等，内容包括数据集成与清洗、数据模型开发与应用集成等。

落地效应及价值

项目的实施，使得该城商行在AI能力方面实现了质的飞跃。AI中台的建成，不仅提升了银行的服务效率和质量，还为银行的业务创新提供了强大的技术支持。通过AI能力的全生命周期管理，银行能够更快速地响应市场变化，提高决策的效率和准确性。同时，AI中台的建成也为银行节约了大量的运营成本，提升了银行的市场竞争力。此外，通过AI技术的深度应用，银行的客户体验也得到了显著提升，增强了客户的忠诚度和满意度。

案例三：某省级城商行金融大模型项目

在数字化转型的大背景下，某城市商业银行（以下简称“城商行”）面临着激烈的市场竞争和日益多样化的客户需求。为了提升银行的数据处理和分析能力、客户服务水平、风险管理精度以及监管合规效率，城商行决定引入中电金信的金融大模型解决方案。该方案旨在构建一个集数据处理、智能分析、客户服务和风险管理于一体的综合性大模型应用平台，以提高银行的整体运营效率和市场竞争力，为客户提供更安全、更便捷、更智能的金融服务。

实施思路与方案

城商行的项目实施分为三个主要阶段：数据构建、模型训练和部署推理。

- **数据构建：**

  城商行与中电金信合作，从多个来源收集包括交易数据、客户互动记录、市场分析报告等在内的专业领域数据集。另外，通过数据质量过滤、去重、隐私脱敏等步骤提升数据质量，确保模型训练的语料库既丰富又准确。

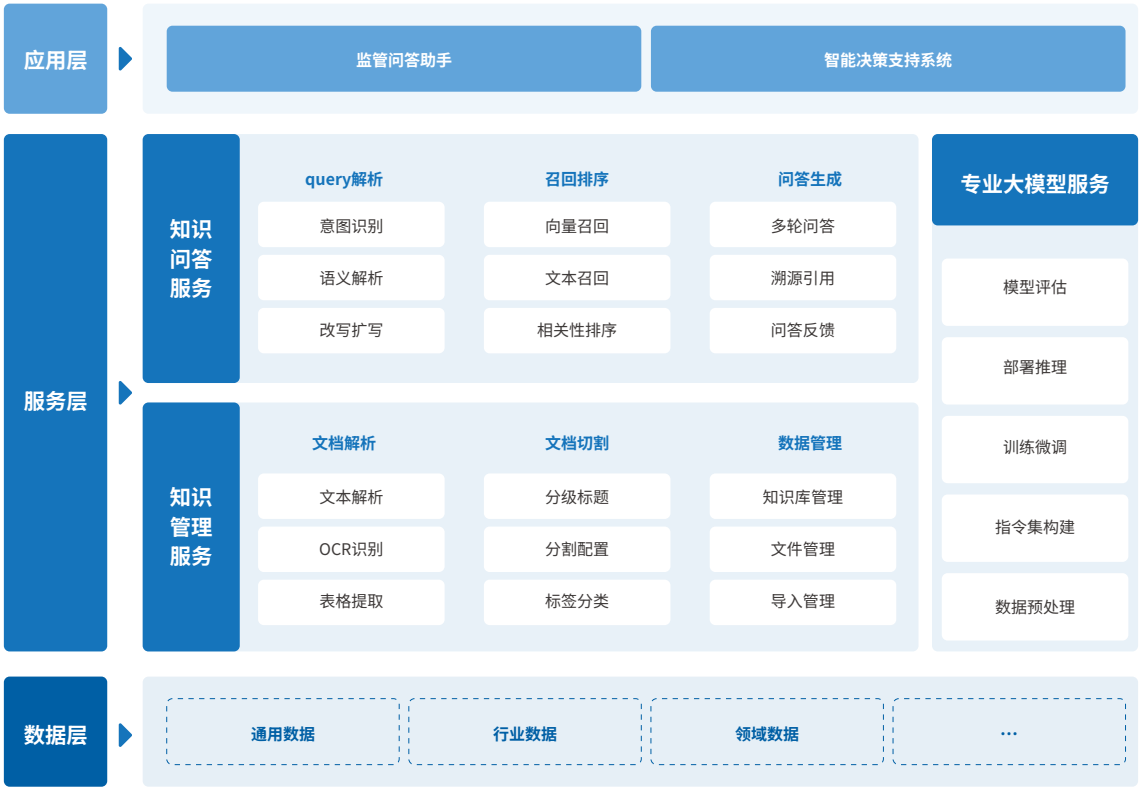
- **模型训练：**

  采用基于Decode-only架构的专业大模型，通过模型微调方法进一步提升模型在金融领域的特定任务表现。同时利用分布式训练技术，提高计算效率，满足大模型训练对算力的需求。

- **部署推理：**

  通过模型压缩和加速方案，优化模型以适应资源受限的设备环境，提高推理速度并降低资源消耗。

在监管问答平台的开发上，城商行利用中电金信的技术，结合检索增强生成技术（RAG），构建了一个集知识管理、知识检索问答和系统管理于一体的智能问答系统。该系统能够提供即时、准确的合规咨询服务，支持上下文理解，并直接索引到对应文档知识段落，确保答案的真实性和透明度。



来源：中电金信，2024

落地效应及价值

通过中电金信金融大模型的成功落地，城商行在多个关键领域实现了显著的改进：运营成本得以降低，业务处理效率和客户响应速度大幅提升，进而推动了收入的增长；同时，风险管理能力的提升有效预防了潜在的金融风险，确保了银行业务的稳健运行。此外，该项目还为银行培养了一批金融科技人才，为行业的数字化转型树立了新的标杆，展现了金融大模型在提升银行竞争力和创新服务中的重要作用。

注：本次调研样本量为100，其中国有商业银行样本量是30，股份制商业银行样本量是20，城商行/区域性商业银行的样本量是10，证券机构样本量是20，保险机构样本量是20。调研对象是金融机构高管、AI部门/数据部门负责人、IT部门相关人员（产品经理、算法工程师、云工程师等）。

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