

行业简报

2025年

中国银行大模型部署实践：

DeepSeek如何优化银行业的算力资源与运营效率

企业标签：DeepSeek

China Large Model Industry

中国大規模モデル産業

报告提供的任何内容（包括但不限于数据、文字、图表、图像等）均系头豹研究院独有的高度机密性文件（在报告中另行标明出处者除外）。未经头豹研究院事先书面许可，任何人不得以任何方式擅自复制、再造、传播、出版、引用、改编、汇编本报告内容，若有违反上述约定的行为发生，头豹研究院保留采取法律措施、追究相关人员责任的权利。头豹研究院开展的所有商业活动均使用“头豹研究院”或“头豹”的商号、商标，头豹研究院无任何前述名称之外的其他分支机构，也未授权或聘用其他任何第三方代表头豹研究院开展商业活动。

研究框架

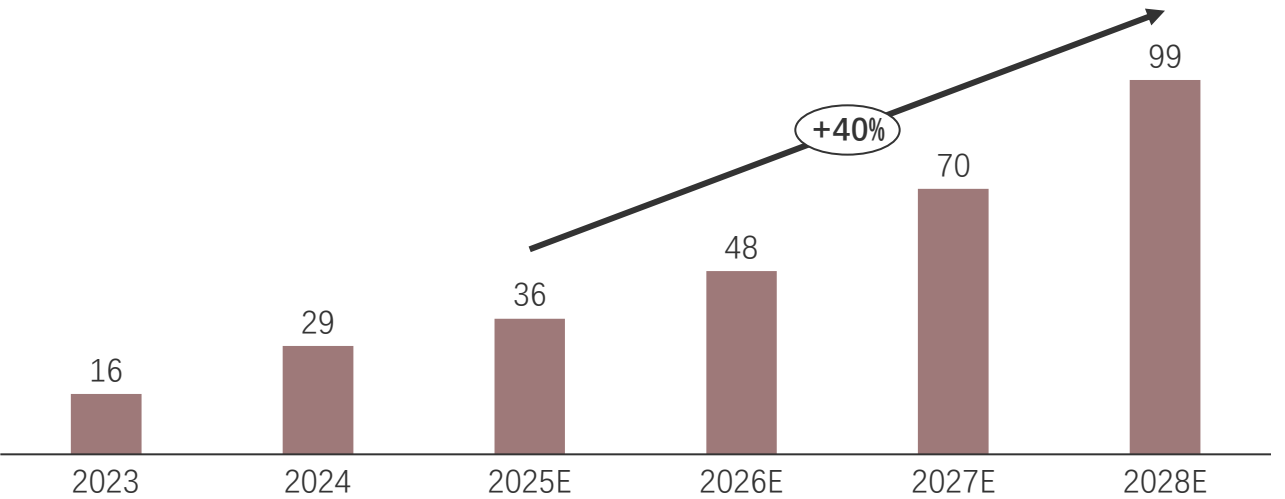
◆ 中国银行大模型部署实践	-----	3
• 金融大模型发展现状	-----	3
• 银行大模型招投标情况	-----	4
• 应用落地主要场景	-----	5
• 应用落地实践效果	-----	6
• 优化路径分析	-----	7
• 发展机遇	-----	8
◆ 业务合作	-----	9
◆ 方法论及法律声明	-----	10

■ 中国银行大模型部署实践——金融大模型发展现状

- 当前金融大模型发展虽需求旺盛，但受SLA、合规与本地部署要求约束，整体仍处于技术与交付能力磨合的关键阶段。中国银行等头部机构正通过构建平台化、可控的大模型体系引领行业落地标准

金融大模型发展现状

■ 中国金融大模型市场规模



- 2024年中国金融大模型市场规模达28.66亿元，同比增幅显著，但全年增长结构出现“前高后低”的阶段分化，市场在经历上半年快速放量后，于下半年出现增速趋缓，未能延续上行势头。
- 这一变化表面上与AI在金融场景中的需求增长趋势相悖，实则反映出大模型在金融行业落地过程中面临的“结构性瓶颈”与“系统性挑战”双重制约。一方面，随着金融机构对AI依赖度提升，客户对模型响应速度、生成准确性、SLA达成率与交付稳定性提出了显著更高的门槛，传统“模型即服务”的交付范式难以满足银行、保险、证券等机构对稳定性、可控性与审计可追溯性的刚性要求。在此背景下，项目部署逐步从公有云迁移至本地私有化、软硬一体化的架构方案，带来交付流程复杂化、集成周期拉长与成本结构上移，尤其影响到中小型AI厂商的实施节奏与交付能力。
- 另一方面，监管对于数据合规、安全边界、模型可信的要求日益强化，催生出一系列技术和工程门槛，如训练数据来源可溯性、推理路径可解释性、模型行为审计等要求，在短期内加大了大模型在金融行业的系统集成复杂度。这导致即便整体需求旺盛，项目落地从决策到交付的周期显著拉长，部分机构出现“预算在手但项目搁置”的停滞状态。
- 尽管如此，市场长期发展逻辑并未逆转。2025—2028年市场仍将保持年均复合增长率40%的中高速增长趋势，到2028年总规模将达99亿元。这表明当前放缓更多是一种“技术适配与治理能力磨合期”，而非行业拐点。尤其是大型金融机构正通过构建本地智能中台、统一模型底座与数据资源调度平台，逐步搭建可持续、可管控的大模型运行体系，为行业下一轮规模化增长夯实基础。

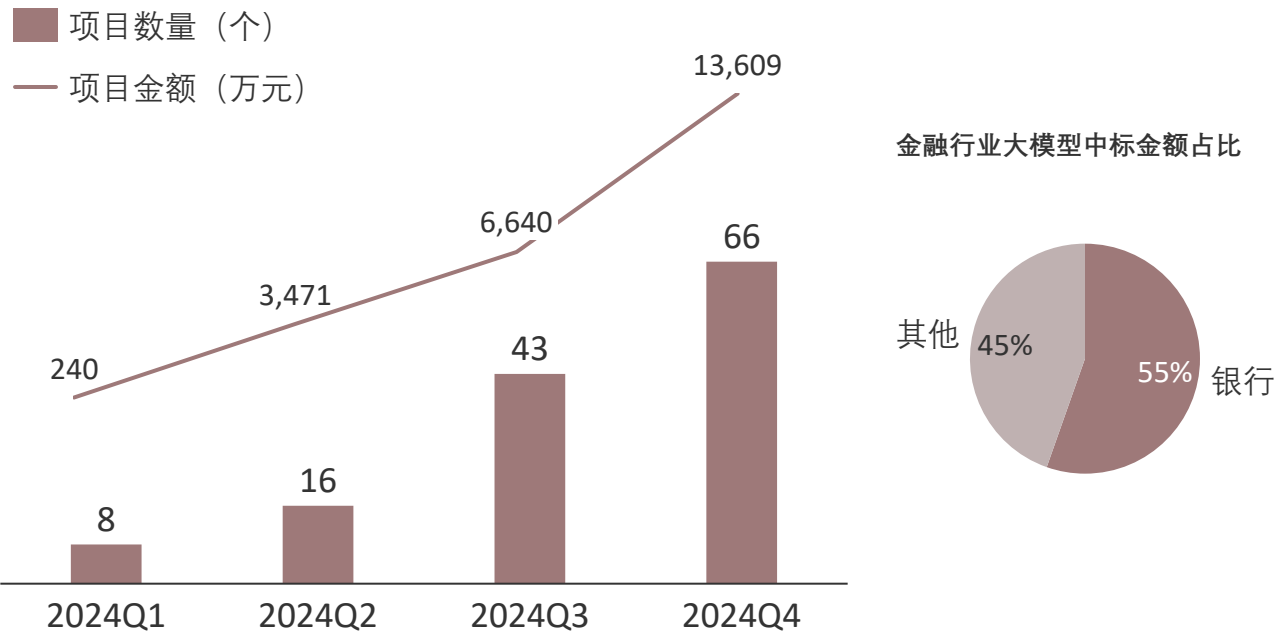
来源：专家访谈，企业公告，头豹研究院

中国 银行大模型部署实践——银行大模型招投标情况

- 银行已成为推动金融行业大模型商业化落地的主阵地，不仅在数量和金额上形成示范效应，更在架构设计、场景沉淀与人才组织机制上率先完成转型，为行业提供可复制的实践样本

银行大模型招投标情况

金融行业大模型中标金额及数量，2024年



- 2024年银行业大模型招投标市场呈现显著升温趋势，全年共完成133个中标项目，总金额超过2亿元，尤其在下半年呈现爆发式增长。这一趋势背后，反映出银行业在完成早期探索性部署后，正加速进入以业务条线为主导的系统化扩展阶段。
- 以中国银行为代表的头部银行已在客服、风控、审贷、合规、财务、投研等多条业务线上推动大模型深度嵌入。一方面，中国银行依托自身数据体量与系统稳定性优势，构建具备私有化部署能力的专属语言大模型，探索包括智能客服、智能质检、远程开户审查、自动问答、信贷反欺诈、财务公式生成在内的高复杂度场景，显著提升运营效率与客户体验；另一方面，其在治理架构上同步建设“算法可信、数据安全、流程可控”的应用框架，推动从IT项目式部署向组织级能力平台转型。值得注意的是，在金融行业整体大模型中标结构中，银行占比高达55%，不仅体现出银行在资金投入上的领先地位，也显示其对大模型技术的实际吸收转化能力最强。
- 与保险、券商等子行业仍处在产品验证和局部试点阶段相比，银行尤其是国有大行，正在率先构建以大模型为核心驱动的智能运营体系，重塑从数据获取、业务处理到客户交互的全链条逻辑。这一趋势标志着银行正在从“规则化自动化”向“泛认知智能化”演进，将大模型视为继信息系统、云计算之后的第三代关键基础设施。

来源：专家访谈，企业公告，头豹研究院

中国 银行大模型部署实践——应用落地主要场景

- 大模型正在从底层重塑银行运营逻辑，以智能交互、流程自动化、精准风控与数据驱动决策为核心，全面推动银行由“数字化叠加”迈向“智能化重构”

银行大模型应用落地场景

应用落地主要场景	客户服务类	业务流程优化类	风险管理类	营销与客户管理类	数据管理类	决策支持类
细分场景	智能客服 辅助客服	智能合同质检 自动化估值对账 OA 文档处理 财务指标及绩效 公式辅助生成	风险管理与 欺诈检测 风险评估	精准营销 辅助营销 信贷审批与 客户管理 产品推荐	数据分类 分级	数据分析决策 智能运维辅助 理质效提升

- 客户服务类：银行通过大模型构建多模态智能客服体系，显著提升复杂问题理解与个性化响应能力，降低人工成本；同时在辅助客服中引入语境分析与情绪识别，为人工审核生成回复建议，增强客户沟通的效率与体验，大模型正从工具转向认知型客户管理中枢。
- 业务流程优化类：在合同审核、估值对账、文档生成等流程中，大模型推动复杂文本识别与公式生成自动化，提升效率与准确率，优化人力结构。特别是在高频场景下，具备高识别率与低部署门槛，助力中小银行实现智能化普惠转型。
- 风险管理类：大模型提升银行对欺诈行为与信用风险的识别预警能力，支持反洗钱等复杂场景；同时引入社交媒体等非结构化数据，拓展风险识别维度与实时性，强化模型在复杂高频场景中的泛化与实战表现，构建智能风控新基座。
- 营销与客户管理类：通过客户行为建模与语义理解，大模型驱动精准营销与个性化推荐，提升转化率和客户参与度；在信贷审批与客户分层中，强化数据处理与决策能力，推动中小银行由经验驱动向算法驱动转型。
- 数据管理类：大模型替代人工进行字段分类与敏感度识别，构建标准化、自动化的数据分级体系，提升治理效率与安全性，奠定模型构建与合规监管的数据基础。
- 决策支持类：借助对多源数据的实时分析与建模能力，大模型优化报表生成、运维响应与运营决策流程，助力管理层实现从数据洞察到智能决策的全面升级，推动银行运营逻辑智能化重构。

来源：专家访谈，企业公告，头豹研究院

■ 中国银行大模型部署实践——应用落地实践效果

- 银行大模型落地的实质是一次以智能决策与流程重构为核心的范式迁移，推动银行从局部数字化向组织智能化跃迁，重塑其竞争结构与能力边界

银行大模型应用落地实践效果

应用落地主要场景	实践效果
客户服务类	智能客服：响应时间缩短30%，常见问题自动处理率90%，满意度提升15%。 辅助客服：基于客户特征优化响应策略，提升非标准场景处理效率。
业务流程优化类	合同质检：识别准确率达96%，效率提升12倍。 合规检查：审批效率提升200%。 估值对账：识别准确率超90%，显著减少人工。 OA文档处理：效率提升50%，误差降低40%。 财务指标归因：误差下降60%，人力成本大幅节省。
风险管理类	欺诈识别：风险预警准确率提升40%，风险处置覆盖面扩大30%。 风控评估：识别准确率提升15%，风险识别更及时。
营销与客户管理类	精准营销：转化率提升25%，参与度提升20%。 辅助营销：一线营销识别率提升15%，历史数据利用率提高50%。 信审风控：审批时间缩短50%，远程信贷满意度提升20%。 产品推荐：多轮对话识别准确率达90%，服务链数字化提升显著。
数据管理类	数据分级：分类准确率提升80%，结果一致性95%以上，周期开销显著降低。
决策支持类	分析效率：报告生成效率提升60%。 运营辅助：处理效率提升30%，成本降低25%。 管理决策：多源数据支撑，报告效率提升40%。

- 银行大模型应用正从“场景嵌入”迈向“能力重构”，其本质不是简单提升某一环节效率，而是通过通用智能能力对全流程进行重塑重排。在客服、审批、风控、营销、投研、数据治理等多个核心业务中，大模型打通感知—理解—判断—执行的智能链条，使银行摆脱原有“人工作业+系统堆叠”的割裂模式，转向高度集成、灵活响应、持续优化的智能流程体系。
- 同时，模型对非结构化数据的理解处理能力大幅提升银行对客户、风险和市场的认知能力，成为推动组织能力跃升、产品体系重塑和运营逻辑转型的关键底座。这标志着银行业从“数字化运营”进一步迈入“智能化驱动”阶段，竞争的核心逐步转向模型能力与数据闭环能力的强弱。

来源：专家访谈，企业公告，头豹研究院

■ 中国银行大模型部署实践——优化路径分析

- DeepSeek可为银行业在AI时代构建低成本、高效率、强合规的算力调度与运营体系。对于算力压力日益突出的银行客户而言，深度接入DeepSeek不仅是技术降本手段，更是组织AI化转型的关键抓手

银行大模型部署实践的优化路径分析

- **背景：**银行业面临的算力与运营双重挑战在金融科技加速发展的背景下，银行业广泛应用大模型、智能客服、风控建模、反洗钱监测等AI场景，带来以下挑战：
算力资源紧张：大模型推理需占用GPU等高性能算力，成本高、波动大。
运营效率分化：模型调用频率不稳定，资源利用率低，容易出现“峰值浪费+谷值空转”现象。
合规可控压力：需满足数据安全、模型可解释性、调度透明等监管要求。

- **优化路径一：基于MoE架构实现推理成本下降与吞吐提升技术机制**
技术机制：
DeepSeek-R1与DeepSeek-VL系列均采用稀疏激活的Mixture-of-Experts（MoE）架构，仅激活部分子模型参与推理。
对比传统全量激活模型，可将单次推理所需GPU计算量降低60%-80%，同时支持并行多任务调度。

- **优化路径二：部署轻量模型与端云协同，提升资源弹性与场景覆盖**
策略要点：
针对高频低复杂度任务（如常见咨询、模板回复），DeepSeek提供参数裁剪版LoRA模型或Tiny模型部署至本地或边缘节点。
通过与云端大模型联动，构建“轻量本地 + 精准云端”分级调度体系。

- **优化路径三：构建模型调用中台，实现调用管理、精度控制与成本透明化**
中台能力建设：
接入DeepSeek原生工具链（如Prompt Cache、调用日志分析模块、动态调度器），建立“模型调用中台”。
支持调用行为记录、算力使用审计、API负载分层调度，形成可控、可优化、可评估的运营机制。

来源：专家访谈，企业公告，头豹研究院

■ 中国银行大模型部署实践——发展机遇

- 银行发展大模型不仅是技术升级，更是组织能力跃迁、客户关系重塑与经营逻辑转型的关键路径。具备数据基础、场景优势与合规刚需的银行应尽快布局大模型核心能力，抢占未来“模型原生银行”的制高点

银行大模型发展机遇

趋势驱动

■ 趋势驱动：银行迈入“模型原生”时代

当前，大模型技术正由通用生成迈向行业专用，在数据密集、流程复杂、合规刚性的金融领域，其渗透速度正在加快。驱动因素主要包括：

- 算力与模型双下沉：开源大模型能力增强、推理成本下降，使银行具备自主训练或定制微调能力。
- 政策与监管释放空间：监管科技（RegTech）鼓励AI辅助合规；人民银行、银保监等机构提出“智能风控、安全可控”的明确导向。
- 数据资产价值重估：银行拥有大量结构化和非结构化数据，成为驱动金融大模型训练的独特优势。

典型场景

■ 大模型嵌入银行核心业务全流程

- 智能客服升级：支持多轮对话、情绪识别与全渠道统一接入，提升客户响应效率与服务体验。
- 营销智能化重构：借助个性化推荐与生成式脚本能力，实现精准营销与内容自动化。
- 风控能力增强：多模态数据理解能力驱动智能反欺诈、授信审批与行为监测，提升风险识别准确性。
- 合规审查提效：自动解析监管文件与审计数据，辅助生成合规报告，降低合规人工成本与风险。
- 投研决策加速：高效处理非结构化信息，支持舆情分析、策略生成与研报摘要，提升投研响应速度。

差异化价值

■ 银行具备构建专属大模型的先天优势

- 数据资源领先：银行拥有高频、结构完备的交易、信贷与客户数据，可构建专属金融语料与知识库。
- 系统流程规范：银行IT体系成熟，业务逻辑清晰，便于模型嵌入原有系统、实现模块化集成部署。
- 合规能力扎实：具备完备的风险审计、数据脱敏与访问控制机制，为模型私有化部署提供安全保障。
- 落地路径清晰：具备AI治理基础与IT协同机制，有能力实现从模型微调到组织融合的闭环演进。

战略价值

■ 发展大模型对于银行的战略意义已远超工具价值，可概括为四大核心价值：

- 构建智能中枢：大模型将成为连接客服、风控、合规、营销等多部门的统一智能平台，实现“多点任务+一体调用”。
- 驱动业务重构：引导银行从传统流程驱动型组织向“Prompt驱动型”组织演进，业务设计与模型反馈形成闭环。
- 强化组织韧性：通过智能决策与自动响应机制，提升银行对黑天鹅事件、市场波动的反应力。
- 构筑差异化品牌：智能化服务体验、定制化资产建议将成为未来银行客户粘性的关键来源。

来源：专家访谈，企业公告，头豹研究院

业务合作

会员账号

可阅读全部原创报告和百万数据，提供PC及移动端，方便触达平台内容

定制报告/词条

行企研究多模态搜索引擎及数据库，募投可研、尽调、IRPR等研究咨询

定制白皮书

对产业及细分行业进行现状梳理和趋势洞察，输出全局观深度研究报告

招股书引用

研究覆盖国民经济19+核心产业，内容可授权引用至上市文件、年报

市场地位确认

对客户竞争优势进行评估和证明，助力企业价值提升及品牌影响力传播

行研训练营

依托完善行业研究体系，帮助学生掌握行业研究能力，丰富简历履历

报告作者



袁栩聪
首席分析师



王利华
行业分析师

• service@leadleo.com

业务咨询

- 客服电话：400-072-5588
- 官方网站：www.leadleo.com



商务咨询与深度合作

深圳办公室

广东省深圳市南山区粤海街道华润置地大厦E座4105室

邮编：518057

上海办公室

上海市静安区南京西1717号会德丰国际广场 2701室

邮编：200040

南京办公室

江苏省南京市栖霞区经济开发区兴智科技园B栋401

邮编：210046



方法论

- ◆ 头豹研究院布局中国市场，深入研究19大行业，持续跟踪532个垂直行业的市场变化，已沉淀超过100万行业研究价值数据元素，完成超过1万个独立的研究咨询项目。
- ◆ 研究院依托中国活跃的经济环境，研究内容覆盖整个行业的发展周期，伴随着行业中企业的创立，发展，扩张，到企业走向上市及上市后的成熟期，研究院的各行业研究员探索和评估行业中多变的产业模式，企业的商业模式和运营模式，以专业的视野解读行业的沿革。
- ◆ 研究院融合传统与新型的研究方法，采用自主研发的算法，结合行业交叉的大数据，以多元化的调研方法，挖掘定量数据背后的逻辑，分析定性内容背后的观点，客观和真实地阐述行业的现状，前瞻性地预测行业未来的发展趋势，在研究院的每一份研究报告中，完整地呈现行业的过去，现在和未来。
- ◆ 研究院密切关注行业发展最新动向，报告内容及数据会随着行业发展、技术革新、竞争格局变化、政策法规颁布、市场调研深入，保持不断更新与优化。
- ◆ 研究院秉承匠心研究，砥砺前行的宗旨，从战略的角度分析行业，从执行的层面阅读行业，为每一个行业的报告阅读者提供值得品鉴的研究报告。

法律声明

- ◆ 本报告著作权归头豹所有，未经书面许可，任何机构或个人不得以任何形式翻版、复刻、发表或引用。若征得头豹同意进行引用、刊发的，需在允许的范围内使用，并注明出处为“头豹研究院”，且不得对本报告进行任何有悖原意的引用、删节或修改。
- ◆ 本报告分析师具有专业研究能力，保证报告数据均来自合法合规渠道，观点产出及数据分析基于分析师对行业的客观理解，本报告不受任何第三方授意或影响。
- ◆ 本报告所涉及的观点或信息仅供参考，不构成任何投资建议。本报告仅在相关法律许可的情况下发放，并仅为提供信息而发放，概不构成任何广告。在法律许可的情况下，头豹可能会为报告中提及的企业提供或争取提供投融资或咨询等相关服务。本报告所指的公司或投资标的的价值、价格及投资收入可升可跌。
- ◆ 本报告的部分信息来源于公开资料，头豹对该等信息的准确性、完整性或可靠性不做任何保证。本文所载的资料、意见及推测仅反映头豹于发布本报告当日的判断，过往报告中的描述不应作为日后的表现依据。在不同时期，头豹可发出与本文所载资料、意见及推测不一致的报告和文章。头豹不保证本报告所含信息保持在最新状态。同时，头豹对本报告所含信息可在不发出通知的情形下做出修改，读者应当自行关注相应的更新或修改。任何机构或个人应对其利用本报告的数据、分析、研究、部分或者全部内容所进行的一切活动负责并承担该等活动所导致的任何损失或伤害。