FROST & SULLIVAN
沙利文

头豹
LeadLeo

中国金融大模型市场追踪报告，2024年

人工智能、大模型、金融大模型

2025年6月

头豹研究院
弗若斯特沙利文咨询（中国）

观点摘要

01

■ **2024年中国金融大模型市场规模达28.66亿元，但源于传统金融机构对SLA与交付稳定性要求提升，2024年H2市场增速不及预期**

2024年金融机构在客服、营销、合规等传统场景中的AI需求持续上升，市场整体呈现同比增长趋势，但下半年市场规模并未显著高于上半年。其核心原因在于，金融客户对模型响应速度、SLA达成率与交付质量提出更高要求，部分原通过公有云调用的项目因无法满足金融级部署标准而被迫转向硬件部署，从而拉长交付周期，反映出金融行业AI落地的高门槛与技术严苛性。

02

■ **2024年金融大模型MaaS市场竞争焦点转向平台能力，平台型云厂商正构建“金融AI操作系统”占据战略高地**

2024年，MaaS市场已从模型比拼迈向平台构建力竞争。云厂商通过集成多个开源/闭源与自研模型、统一SDK封装、模型商店分发等手段，打造模型能力调度中台，帮助金融机构以低门槛调用不同模型应对客服、营销、风控等场景，显著降低迁移与整合成本。相比之下，第三方推理云及单一模型厂商受限于生态构建与标准化能力，难以承载金融机构对“大模型中台”的长期期待。

03

■ **2024年金融大模型标准化产品（71%）占据主导地位，2025年后软硬一体机有望在私有化部署中打开新一轮增长空间**

2024年，标准化产品因交付效率高、部署门槛低、覆盖场景广，成为金融机构建设大模型能力的首选路径。进入2025年，随着DeepSeek、通义等优质开源模型推动大模型轻量化、推理成本下降，软硬一体机将在本地化与合规需求突出的场景中进一步释放潜力。其可复制性强、支持灵活交付与快速迭代，有望在私有化部署中加速扩展，成为金融机构“降本、安全、闭环”目标下的重要选型方向。

观点摘要

04

■ **2024年金融大模型在客服场景落地最广，2025年办公效率类场景（尤其代码）的部署意愿上升最显著**

2024年，客户服务类场景因落地成本低、响应效益明显等成为大模型在金融机构中覆盖最广的场景。进入2025年，办公效率类场景（如代码助手）被多数金融机构列为重点推进方向。同时，目前金融机构在不同场景（客服、风控、运营、合规、投研）中部署多个模型，面临数据孤岛、算力浪费、推理能力碎片化等问题。行业普遍对“统一智能中台”的需求逐步增强。未来具备通用推理能力与行业专精能力的大模型，将成为支撑多业务横向融合与纵向任务协同的底座。

05

■ **2025年预计成为AI Agent 在金融行业大规模落地的元年，MOA 架构将成为技术标准化与协同能力突破的关键支撑形态**

金融行业正率先在客服、代码等任务明确、规则清晰的场景中落地AI Agent，当前的技术路径也正由单点工具调用，逐步迈向多智能体协同。随着ReAct提示词、Planner调度器与外部API调用能力成熟，Agent已具备完成分析、生成与合规检测等闭环任务的潜力。在此过程中，以MOA（Mixture of Agents）混合智能体架构为代表的多Agent协同执行模式正逐步取代单体模型，成为金融AI系统主流工程形态。MOA架构强调模块化、可控性与任务解耦，适配金融行业高流程复杂度与强合规要求。进入2025年，随着Anthropic等厂商推出MCP等通用协议，AI Agent正加速向跨系统协同演进，为“多Agent分布式协作 + 金融AI操作系统”打下基础。

章节一 金融大模型产业洞察

- **伴随金融行业智能化需求集中释放、技术突破加速大模型实用化，2024年成为金融大模型应用爆发的关键拐点，市场规模实现超100%的高速增长**

2024年，金融机构对智能投研、智能风控、智能客户服务等场景的需求集中爆发，推动大模型快速规模化落地。得益于多模态融合与知识蒸馏技术的突破，金融大模型在参数量级、推理成本与隐私保护能力上实现同步提升，推理成本持续下降，更好满足高频交易、财务分析等场景的实时性与合规性要求。同时，大模型在金融文本生成、复杂语义理解、多轮对话智能客服等领域表现出显著优越性，进一步加快了金融机构内部应用部署与价值释放的节奏，支撑2024年市场规模的翻倍增长。

- **2024年中国金融大模型市场中，阿里云，百度智能云以及商汤科技位列市场的前三位，营收占比分别为33.2%，19.3%和10.9%**

阿里云在金融大模型领域凭借深厚的金融行业积累、丰富的标准化产品体系以及领先的合规保障能力，稳居市场第一。依托于通义系列模型及千问原生开源体系，阿里云构建了覆盖智能客服、智能质检、智能投研等多个金融核心场景的标准化解决方案，能够快速响应金融机构对于实时性、准确性、可解释性与合规性的多重要求。同时，阿里云通过深度参与金融行业标准制定与联合创新项目，持续强化自身在金融数据安全、隐私保护、审计追溯等方面的能力。

百度智能云围绕千帆平台与文心系列大模型，确立了金融行业一站式智能体平台战略，展现出强劲的市场增长势能。通过智能体化架构设计，百度智能云打通了智能投顾、合规质检、智能投研等多个复杂业务场景，提供从模型调优、应用开发到一体化交付的全流程能力，显著提升了金融机构对智能化转型的适配度与使用体验。产品层面，百度千帆平台集成了丰富的大模型资源与训练微调工具，支持金融行业多样化、差异化的业务定制需求，尤其在场景适配性与二次开发灵活性方面表现突出。

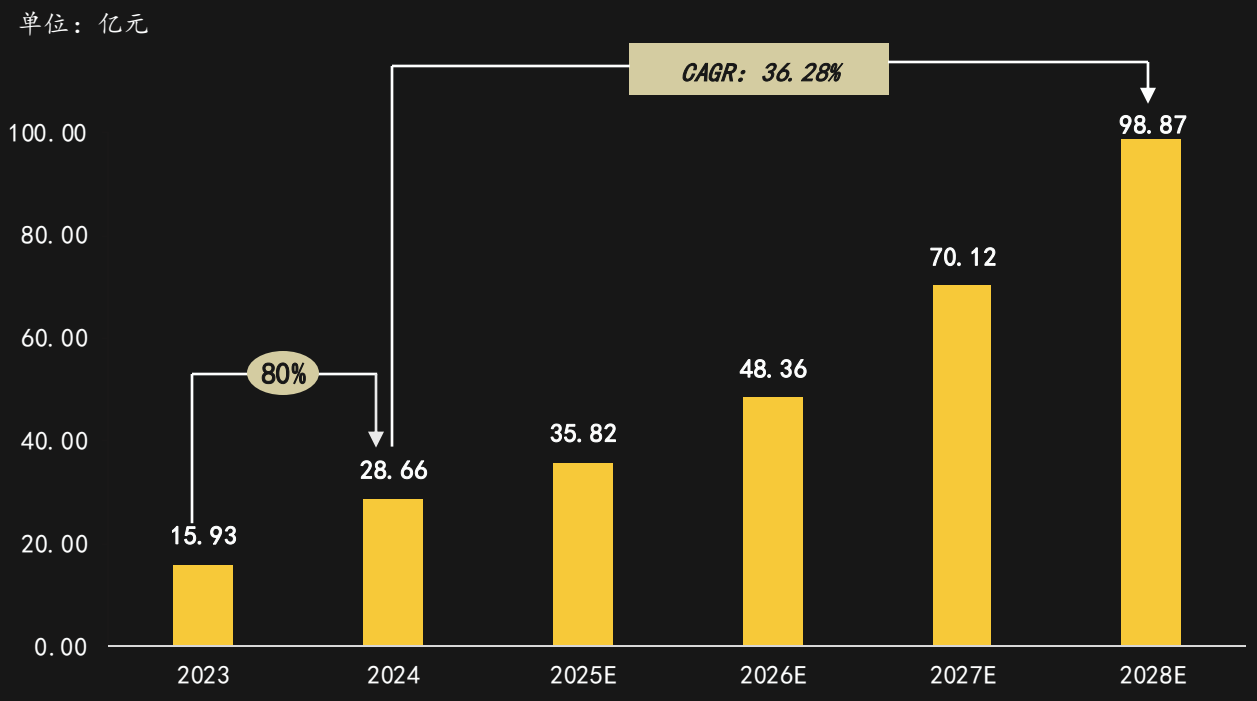
商汤科技凭借领先的多模态大模型能力、自研智算中心基础设施以及软硬一体化交付模式，持续扩大在金融行业大模型市场的差异化竞争优势。基于自建的SenseCore智算中心，商汤科技能够在算力调度、数据治理与模型训练优化方面实现端到端的资源掌控，大幅提升了金融机构在私有化部署、数据安全隔离与模型定制训练中的灵活性与合规性。

中国金融大模型产业洞察——市场规模

关键发现

2024年中国金融大模型市场规模同比增长80%，达28.66亿元，标志着金融大模型应用进入爆发期。未来五年，市场有望保持36.28%的年均复合增长率，推动智能投研、代码助手等场景的快速普及

中国金融大模型市场规模，2023年-2028年



- **伴随金融行业智能化需求集中释放、技术突破加速大模型实用化，2024年成为金融大模型应用爆发的关键拐点，市场规模实现80%的高速增长**

2024年，金融机构对智能投研、智能风控、智能客户服务等场景的需求集中爆发，推动大模型快速规模化落地。得益于多模态融合与知识蒸馏技术的突破，金融大模型在参数量级、推理成本与隐私保护能力上实现同步提升，推理成本持续下降，更好满足高频交易、财务分析等场景的实时性与合规性要求。同时，大模型在金融文本生成、复杂语义理解、多轮对话智能客服等领域表现出显著优越性，进一步加快了金融机构内部应用部署与价值释放的节奏，支撑2024年市场规模的翻倍增长。

- **随着DeepSeek、通义等优质开源大模型出现，以及以一体机为代表的软硬一体交付模式兴起，2025年后金融大模型市场将进入工程化、规模化落地的新阶段**

2025年起，金融行业对大模型的需求将从早期试点转向批量部署，DeepSeek等新一代模型在推理效率、知识融合、长文本理解与推理能力方面显著突破，极大提升了模型在智能投研、合规质检、智能客户经营等核心场景的落地效果。同时，以DeepSeek一体机为代表的软硬一体化交付形态快速普及，不仅降低了模型部署的技术门槛与运维成本，还助力金融机构实现更高标准的数据本地化治理与隐私保护要求，加速推动大模型成为金融数字化转型的基础设施，带动市场进入新一轮高速扩张期。

来源：沙利文、头豹研究院

中国金融大模型行业研究——发展趋势

关键发现

未来金融行业对大模型的需求不再是“能不能回答问题”，而是“是否具备稳定、合规、结构化地支撑复杂任务”的能力。从任务型向平台型、从输出型向解释型、从工具型向底座型演进，将是金融机构在未来3年采购与部署大模型时的主旋律。

金融机构对大模型的未来需求趋势分析

向推理增强能力跃迁

过去的大模型多聚焦在通用问答与知识生成，难以支撑金融领域对可解释、可追溯、符合规则的推理链条的需求未来模型需具备：结构化的推理输出格式；可审计的中间判断过程；对复杂金融逻辑的链式推理能力。

构建统一智能中台能力

当前金融机构多模型割裂部署，造成数据孤岛、算力浪费、能力冗余。未来将转向统一底座模型：横向支撑客服、投研、合规等多业务场景，纵向贯通问答、分析、判断、建议等任务，形成高效智能协同系统。

拥抱高质量数据闭环机制

大模型能力的上限由数据决定。面对金融数据敏感、标注困难的问题，机构将关注具备数据闭环能力的模型：支持自动合成高质量推理样本，内建数据验证机制，并通过飞轮模式实现持续增强与低成本迭代。

升级为多模态融合推理

金融应用正从文本理解扩展到报表图表、语音记录、视频风控等多模态内容。未来大模型需具备跨模态理解、图文协同推理、语音问答等能力，以支撑投研可视化、语音合规审计、智能报告生成等新兴场景。

追求稳态响应与合规可控

金融场景对模型输出的稳定性与合规性要求极高。未来大模型需支持格式可控、响应稳定、日志可溯源、责任可归属，能实现“拒答机制”“幻觉抑制”等能力，满足核心系统对稳态输出的严苛要求。

强化混合云部署支持

随着数据安全与监管要求加剧，金融机构将倾向选择适配国产算力和具备混合云能力的大模型解决方案，确保模型在合规、高性能与成本可控之间取得平衡，实现安全可信落地。

来源：沙利文、头豹研究院

章节二 金融大模型应用洞察

■ 客户体验和营销场景已成为大模型在金融机构中渗透最深的两个高频应用领域

2024年金融大模型在“客户体验与互动”场景的渗透率高达68%，在“市场营销与营销资产生成”场景的渗透率达到59%，均明显高于其他场景，且同比增长显著。这表明金融大模型在提升客户响应效率、优化客户沟通内容、支持多轮交互等方面已取得实质性成果。尤其在客服机器人、虚拟助手、广告文案生成等场景中，金融大模型极大地降低了内容产出与客户支持的人力成本，成为前台业务中最具性价比的金融大模型应用方向。

■ 收益合规、文档处理与代码生成类任务正快速崛起，成为大模型渗透的新增长极

2024年金融大模型在“文档处理”和“代码辅助生成”两个场景的渗透率分别达到了19%和22%。相比之下，2023年由于技术瓶颈，这两个场景完全未被金融大模型渗透。与此同时，“合成数据生成”的渗透率从2023年的5%显著提升至2024年的16%。这一增长表明，金融大模型在处理和生成结构化与非结构化混合文档方面的技术成熟度有了明显提高。这些中台任务通常具有文本重复率高、审核压力大、合规标准明确等特点，因而非常适合利用金融大模型进行自动化处理和初步生成。这种技术应用能够显著提升合规部门、法务团队以及IT开发人员的工作效率。

■ 大模型正由外围辅助向核心业务环节深入渗透，赋能金融决策与风控管理

2024年，金融大模型在“定价、风险管理与承保”场景的渗透率从3%大幅上升至8%，而“交易与投资组合优化”场景的渗透率也从2%增长到9%。这一变化表明，金融机构不再满足于仅将金融大模型应用于辅助性岗位，而是开始探索将其引入核心决策环节，赋予其AI判断力。这些场景对数据质量、模型解释能力以及风险控制能力的要求更高。金融大模型在这些领域的快速渗透，反映出其在处理复杂金融逻辑和高价值任务方面的可信度和效果已逐步获得认可。

■ 企业搜索与知识管理类应用快速普及，金融机构正在构建大模型驱动的内部智能信息系统

2024年，金融大模型在“企业搜索”场景（如内部文档和政策信息检索）的渗透率从13%提升至29%，增长幅度颇为显著。这一趋势显示，金融机构越来越重视对知识库、流程文档以及监管政策等非结构化内容的结构化挖掘与应用。通过将大模型与文档系统和知识图谱相结合，金融机构能够开发出面向员工的智能问答助手或监管合规搜索引擎。这种方式不仅显著提高了信息调用的效率，还有效打破了“人找知识”和“知识找人”之间的内部信息流动壁垒。

中国金融大模型行业应用——主流落地工程

关键发现

大在金融行业大模型应用通常分为三种技术路径：Prompt Engineering适合快速上线、覆盖率高，Workflow架构用于拆解复杂任务、实现多Agent协同，模型微调则适用于高价值场景下的深度定制，但面临通用性损耗等方面的技术挑战

中国金融大模型主流落地工程路径分析

方案1：Prompt Engineering（PE 工程）

- 写提示词＋设置参数来控制模型表现；
- 不动模型本身，约90%客户以这种方式落地应用。

方案2：Workflow架构

- 将一个复杂任务拆解为多个子任务；
- 每个子任务由一个 Agent 调用基础模型完成；
- 最终由 Workflow 将多个 Agent编排为一个完整应用流程。

方案3（少数情况）：模型微调(FT)

- 只有在打样项目或高价值客户需求下才会考虑；
- 极高门槛，需要解决：
  - 高质量数据生成；
  - 幻觉问题控制；
  - 通识能力损耗后的补偿；

✓ **Prompt Engineering（提示词工程）：**Prompt Engineering 是当前生成式AI在金融行业最主流的应用方式，其核心是在不改动模型本体的前提下，通过设计提示词和参数配置来引导模型输出目标结果。该方法部署门槛低、适应场景广，适合快速验证效果并实现业务对接。据行业实践观察，约90%以上的客户选择以此方式落地初期应用，覆盖场景包括文本生成、问答摘要、客户服务等。

✓ **Workflow架构：**Workflow 架构适用于任务链条较长、逻辑路径复杂的场景，其方法是将一个金融业务任务拆解为多个子任务，并由多个独立 Agent 分工调用基础模型，最终通过工作流引擎将各子任务编排为端到端流程。例如，在投资研究或理赔审核等多步骤任务中，Workflow 架构可实现信息提取、规则判断、摘要生成等多模块协同，从而提升准确性与业务可控性。此类架构更强调系统工程能力，适用于金融客户中具备一定开发能力或平台集成能力的团队。

✓ **模型微调（FT，Fine-tuning）：**模型微调属于高投入、精度导向型的应用路径，通常仅在高价值客户项目或具备长期训练计划的机构中使用。该方案需构建高质量训练数据集，并解决模型在幻觉、理解能力、通用性损耗等方面的技术挑战。

来源：沙利文、头豹研究院

中国金融大模型行业应用——落地痛点

关键发现

金融大模型的落地进程正面临数据壁垒、部署成本、监管不确定性及人才协同等多重挑战，亟需在数据治理、系统对接、合规审查与组织机制上实现系统性突破

金融机构部署金融大模型痛点分析

数据壁垒与合规压力

数据分散在多个系统中，整合周期长、质量难控，同时金融数据涉及高度敏感信息，需满足严格合规要求，数据调用受限

技术成本与工程瓶颈

金融大模型部署需高性能算力支撑，训练和推理成本高；同时需与多个系统深度集成，接口对接复杂，运维压力大

法规模糊与风险不明

当前监管规则滞后，大模型在核心场景中的合法性尚存灰区；黑箱机制难以满足审计与可解释性要求，合规风险显著

人才稀缺与组织不协同

AI人才招录难、留存难，业务与技术存在沟通障碍，组织内部缺乏推动AI落地的联动机制与决策魄力

01 *数据整合难与合规要求高，构成金融大模型部署的前置“拦路虎”*

金融机构内部数据长期呈现出强“烟囱化”结构，分散于多个系统且存在治理割裂，数据清洗、权限调取、字段映射等环节成为部署初期最大瓶颈。即便项目立项，往往需数月协调跨部门数据融合。同时，金融数据的敏感性决定其必须符合《个保法》等法规要求，模型调用需满足脱敏、安全审计等标准流程，进一步放缓部署节奏。

02 *部署难度高，成为金融机构将大模型真正用于业务场景的最大瓶颈*

虽然许多金融机构已完成大模型的试点验证，但要将其正式部署到业务系统中却面临重重障碍。一方面，大模型需要强大算力与稳定系统支持，部署过程涉及数据接口打通、权限管理、实时推理等多个环节，运维复杂度远高于传统模型；另一方面，金融机构现有的IT架构多为封闭、分散的“烟囱式”系统，不同系统间缺乏统一标准，导致模型难以快速对接CRM、风控、交易等核心平台。

03 *监管滞后与可解释性缺失，使核心场景应用长期处于“观望区”*

当前对于金融大模型的监管制度尚未系统化覆盖大模型在投顾、风控、授信等关键场景中的应用边界，合法性与责任归属问题存在灰区。由于大模型天然具有“黑箱”特征，难以解释每一次输出逻辑，若无法满足“可追溯、可审计”的监管要求，将被视为算法歧视或操作不当。

04 *AI人才与组织机制“上下断层”，造成技术认知与业务推动“双低效”*

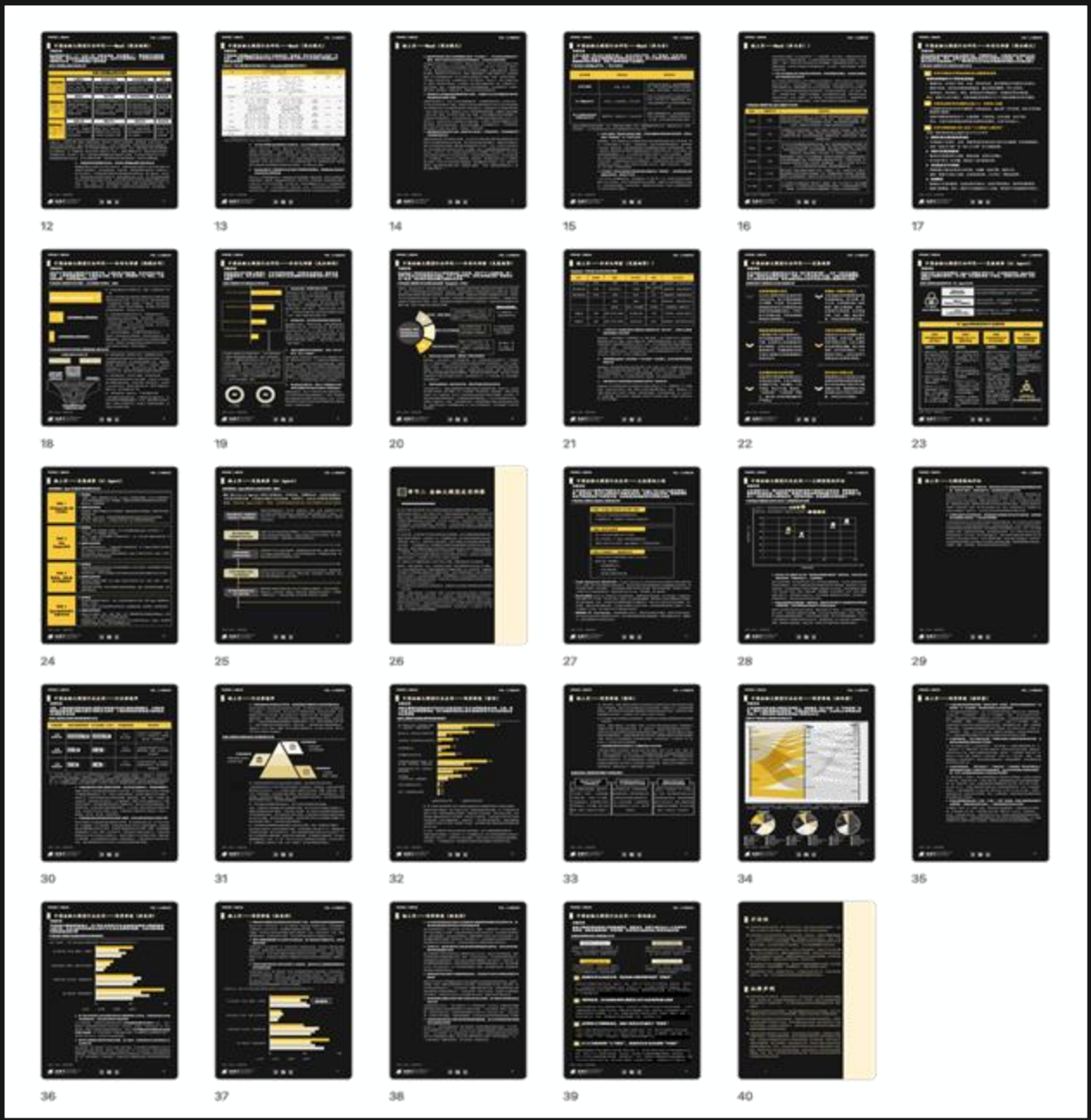
金融机构普遍缺乏具备AI建模与工程化能力的复合型人才，同时难以提供与科技公司匹配的薪酬与成长空间，导致人才引入与保留双重受限。与此同时，业务团队对大模型的理解仍停留在“热词阶段”，而技术人员又常无法精准把握业务场景，协同效率低下。此外，管理层对大模型潜在风险持谨慎态度，POC项目往往难以转化为业务主线，组织内缺乏明确的“AI战略牵引机制”，导致探索止步于初级试点。

来源：沙利文、头豹研究院

中国金融大模型市场追踪报告，2024年

沙利文联合头豹聚焦于“金融大模型”前沿领域，从“高价值场景、强泛化能力”的模型底座出发，依托“多智能体架构与数据工程体系”提供技术支撑，并以“可监管、可控可用、安全可信”为关键要素，深入剖析金融大模型在银行、保险、证券、互联网金融等核心行业中的典型应用路径与最佳实践

报告完整版登录www.leadleo.com 搜索
《中国金融大模型市场追踪报告，2024年》



方法论

- 头豹研究院布局中国市场，深入研究19大行业，532个垂直行业的市场变化，已经积累了近100万行业研究样本，完成近10,000多个独立的研究咨询项目。
- 研究院依托中国活跃的经济环境，从纵深防御、快速响应、轻量化部署等领域着手，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，不构成任何证券或基金投资建议。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，概不构成任何广告或证券研究报告。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。本报告所载的资料、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，头豹可发出与本报告所载资料、意见及推测不一致的报告或文章。头豹均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。

报告完整版登录www.leadleo.com

搜索《中国金融大模型市场追踪报告，2024年》

首席分析师

袁栩聪

15999806788

oliver.yuan@frostchina.com

研究总监

李庆

13149946576

livia.li@frostchina.com

www.frostchina.com ; www.leadleo.com

https://space.bilibili.com/647223552

https://weibo.com/u/7303360042

©弗若斯特沙利文咨询（中国）

©头豹研究院