FROST & SULLIVAN
沙利文

头豹
LeadLeo

中国金融大模型市场追踪报告，2024H1

人工智能、大模型、金融大模型

2025年2月

头豹研究院
弗若斯特沙利文咨询（中国）

观点摘要

01

■ **中国金融大模型市场高速增长，源于大型金融机构的数智化升级**

2023年中国金融大模型市场规模为15.93亿，预计到2024年将增长至38.2亿，增长率预计将达到显著的140%。这一显著增长主要得益于大型金融机构，尤其是银行、保险公司和券商的推动。这些机构逐步将金融大模型视为数智化转型的核心，利用AI技术优化风险管理、客户服务以及智能投顾等业务。与此同时，金融大模型的轻量化部署使得中小型互金机构也能广泛采用，这推动了整个市场的进一步扩展。

02

■ **2024年H1中国金融大模型市场，MaaS占52%，引领中小型机构规模化应用，私有化部署占48%，是大型金融机构首选**

2024年H1，中国金融大模型部署模式中，Maas模式凭借“开箱即用、按需付费”的特点，大幅降低了中小型互联网金融机构的技术门槛和初始投资压力。同时，私有化解决方案因其在数据安全、合规性以及深度定制方面的优势，成为大型金融机构的首选。未来几年，MaaS将继续占据主导地位，持续满足金融企业在降本增效方面的需求。

03

■ **在金融大模型的落地应用中，标准化产品占据了约60%的市场份额，未来几年，标准化产品的市场份额将持续增长，预计2026年将会达到70%以上**

定制化方案的部署周期通常为3至6个月，投入金额从数百万元到千万元不等；而标准化产品通过融合模块化架构与预训练参数解决方案，能够在4至6周内迅速完成部署。这类产品依托成熟的大模型平台，提供强劲的算力、资源调度、模型训练及推理能力，确保了高效的执行效率。未来几年，标准产品将凭借促进云技术与私有化环境高效融合能力，助力客户实现整体架构升级与迭代能力，及应对未来技术变革及不断变化业务需求提供灵活支持能力，推动其市场份额持续增长。

观点摘要

04

■ **金融大模型在客服服务与数据分析等前中台场景赋能显著，但复杂金融决策领域仍需进一步深化**

目前，金融大模型通常融合厂商现有的数据库与大数据平台，优化营销、信贷、客服等核心业务场景的客户体验。同时，在中后台，大模型通过自动化数据处理能提升流程效率和决策速度。然而，对于高复杂度的金融决策场景，如投资组合优化和衍生品定价等，金融大模型的智能化水平仍需进一步深化和提升。未来，更多的金融大模型将依托原生数据中台和智能应用体，在高复杂度业务决策场景提供更精准的支持，从而进一步推动金融行业智能化升级。

05

■ **负毛利时代的百模大战后，未来3年，中国将形成“3-5家闭源巨头+1-2家开源平台”的格局，同时，头部企业将实行开源和闭源的双轨策略**

未来三年，开源模型市场预计仅有1-2家头部企业，能依靠“非大模型业务”的现金流来支撑长期投入。这些企业会战略性地选择开源，以释放技术红利、培育开发者生态；同时，利用闭源模型保护核心技术，实现高价值变现。由于开源模型维护成本高昂，需要万卡算力并持续迭代，因此这些头部企业将会推出“有限开源”模式（如部分模块开源+核心闭源），或者采用“开源即服务”（开源模型必须部署在自家云平台）的方式，以实现商业与生态的双重控制，从而模糊闭源与开源的边界。

章节一 金融大模型产业洞察

- **大模型正从“技术选项”跃升为中国金融数字化发展的“技术基石”**

从2023年到2028年，中国金融大模型市场规模预计将从15.93亿元跃升至131.79亿元，2024年增长率更有望达到140%。这一飞速扩张不仅反映了大模型在金融行业应用的潜力不断被挖掘，也凸显了其从“可选技术”向“核心依赖”的加速演变趋势。技术方面，自然语言处理、多模态数据融合等前沿突破显著提升了金融舆情分析、监管政策解读以及风控建模的实时性与效率；政策方面，数据保护法规与金融科技发展规划的持续发布，为大模型在合规与数据安全层面供了坚实保障，进一步降低了行业采纳的阻力。这些内外部因素的共振，使大模型成为中国金融企业获取竞争优势的关键变量。

- **金融大模型在高频交易、个性化服务等核心场景带来显著价值，MaaS模式的灵活性和低成本则迅速打开中小金融机构的长尾市场**

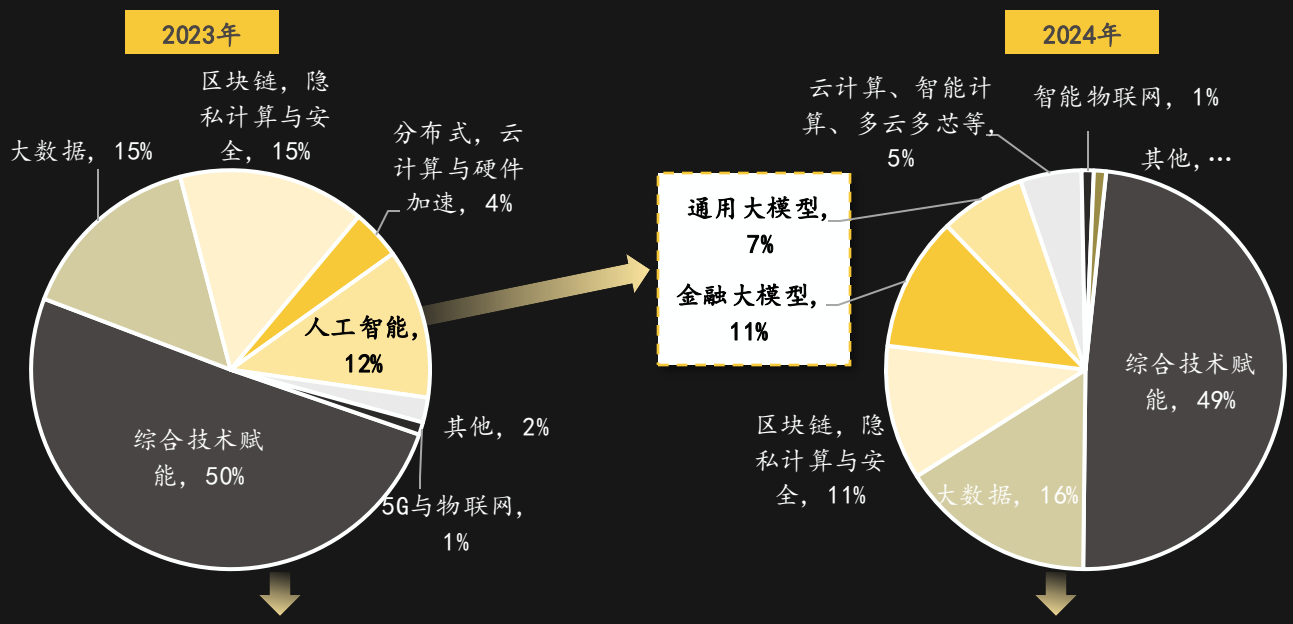
在高频交易场景中，金融大模型能够毫秒级追踪市场波动并生成交易策略，不仅让金融机构的决策速度提升30%以上，也大幅降低潜在风险损失；在个性化服务中，基于多模态数据融合技术的客户画像和风险评估让服务质量与客户满意度同步提升，显著增强了机构的黏性和竞争力。与此相应，部署成本的下行与“即开即用、按需付费”的服务模式共同催生了庞大的“长尾”需求。以往只有大型银行才能负担的金融科技技术，现今部分大模型功能通过MaaS模式落地仅需数十万元，加之MaaS模式能在确保合规和基础功能的同时提供极高灵活度，迅速成为中小型互联网金融企业的首选，占据了52%的金融大模型市场份额。

中国金融大模型产业洞察——发展背景（技术）

关键发现

中国金融科技的数字化进程中，大模型正从“技术选项”跃升为“技术基石”，并逐渐成为金融科技企业获取竞争优势的关键变量。大模型驱动的“智能+”变革正在颠覆传统金融机构的业务模式

中国金融科技企业核心技术要素发展情况，2023-2024年



2024年18%的金融科技企业将AI技术作为核心技术要素，比2023年增长了6个百分点

金融大模型的崛起意味着金融技术的应用将从规则驱动向数据驱动重大转型

传统的金融技术依赖于预先设定的业务逻辑和算法规则来处理交易和管理风险，这种方法在标准化和流程化的环境中表现出色。然而，面对当今金融市场日益复杂的场景以及快速变化的风险环境，这种固定模式逐渐暴露出其局限性。预设规则难以灵活适应新的挑战，导致金融机构在应对未知情况时反应迟缓、效率低下。

相比之下，基于大模型的金融科技解决方案标志着一次革命性的进步。这些模型通过深度学习算法对海量的历史数据进行分析，能够自动识别模式并预测未来趋势。更重要的是，大模型展现出卓越的通用性和迁移能力，它们不仅可以应用于特定的已知情境，而且能够在未定义或新出现的情境中动态调整策略，提供实时的个性化服务和支持。

这种转变不仅仅是技术上的更新换代，更是从根本上改变了金融机构处理信息的方式——从“被动应对”转变为“主动洞察”。金融机构不再局限于遵循既定规则，而是利用先进的数据分析工具提前预见市场动向，并据此制定更加精准有效的商业决策。因此，随着大模型技术的发展，金融行业正迎来前所未有的创新机遇，这将重塑整个行业的业务逻辑和服务模式。

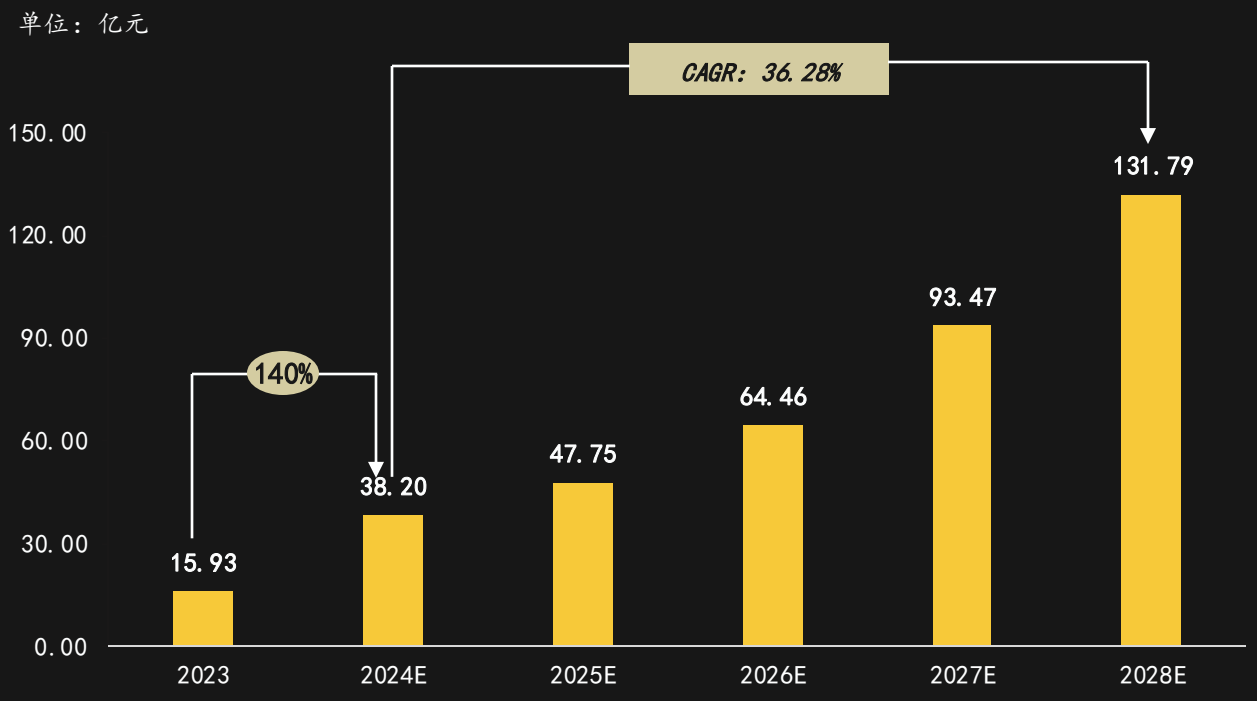
来源：沙利文、头豹研究院

中国金融大模型产业洞察——市场规模

关键发现

随着大模型技术的不断进步和金融行业数字化转型的加速，中国金融大模型市场规模快速增长，预计从2023年的15.93亿元将跃升至2028年的131.79亿元，其中2024年上半年市场规模已达16亿元，预计2024年全年规模将达到140%的增长

中国金融大模型市场规模，2023年-2028年



- **多模态融合与知识蒸馏技术的突破，正引领金融大模型跨越“金融可用性”的临界点，重塑金融服务的未来，推动金融大模型市场规模快速增长**

头部金融大模型不断精进文本、图表及音频的联合分析能力，现已能够深度解析上市公司年报中的非结构化数据与结构化财务指标之间的关联，为财务分析等场景提供了更准确的支持。同时，借助知识蒸馏技术，模型参数量级从千亿精简至百亿，使得推理成本降低了80%以上，充分满足了高频交易场景对实时性的需求。此外，联邦学习技术的进步促进了反欺诈模型的发展，在保障数据隐私的前提下显著提升了模型效果，有效解决了金融机构间的数据孤岛问题。这些技术革新共同推动了金融大模型市场的高速增长。

- **伴随智能客服、风控合规、投研决策等核心场景规模化落地，中国金融大模型市场即将进入价值兑现的爆发期**

金融大模型在智能客服、投资顾问等对话式场景中快速渗透，通过语义识别与知识库匹配，大幅提升客户服务效率并优化客户体验，带来显著的获客与留存价值。另一方面，风控与合规管理是金融机构的“生命线”，金融大模型凭借更精准的风控策略与实时监控能力，有效降低不良贷款与欺诈风险，直接转化为商业价值。此外，在投研与资管环节，金融大模型可借助多模态数据分析与时序预测，深度挖掘市场机会并实现差异化的投资策略，从而提高盈利水平。

来源：沙利文、头豹研究院

章节二 金融大模型部署核心要素

- **稳定性、低延时与高并发构成金融大模型部署的关键技术基础**

金融大模型在部署与实施中需深度融合云原生理念，并非仅仅局限于“上云”或“堆容器”，而应充分利用微服务、容器化和服务网格等技术，并在部署流程中高效整合GPU、NPU等异构算力，以在系统负载动态变化时，实现自动化构建与持续交付，从而确保核心金融业务的稳健运行。同时，针对大模型参数量大、推理计算开销高的特点，需通过剪枝、稀疏激活、混合专家模型及知识蒸馏等技术手段优化模型结构，在保持大模型核心能力的基础上显著缩减模型体积，实现毫秒级响应，保障高并发场景下的服务稳定与快速。此外，面对金融业务中常见的突发流量，还需构建完善的负载均衡策略与资源弹性扩展机制，确保计算资源能够依据实际需求灵活扩容或收缩，有效避免因资源分配不当导致的系统不稳定或延时增加问题。

- **准确性与兼容性是金融大模型的核心价值所在**

金融大模型需先通过单任务指令微调（如情感分析、命名实体识别）确保高精度，再借助多任务与Zero-shot指令微调增强跨任务泛化能力，以兼顾传统任务精细需求与应对突发场景。同时，模型需支持跨平台部署，提供丰富API与SDK，消除数据孤岛，实现系统间协同。此外，高质量数据清洗、标注、对齐及结合金融知识的定制化微调是提升模型准确度和稳健性的关键，过程中需严格遵循合规要求，整合多元数据源。

- **安全性与合规是金融大模型全面落地的前提**

金融大模型在业务应用中必须严格遵守数据安全、隐私保护和内容合规的法规要求，确保客户信息在模型训练和推理过程中不被滥用或泄露，并避免误导性或违规输出。为实现逻辑透明度，采用标签学习对模型推理的关键节点进行标注，使重要决策可被解释和验证，降低模型“幻觉”风险。

- **从基础硬件、云平台、机器学习框架、开发工具链到上层业务应用的全链路整合能力是金融机构考察大模型厂商的核心要义**

金融大模型的高效部署依赖于供应商是否具备从基础硬件、云计算架构、机器学习框架到应用开发工具链的全链路整合能力。只有在这一体系下，才能确保模型在高并发场景中实现低延时、精准决策，并通过智能算力调度优化资源利用率。同时，完整的安全合规体系必须贯穿全链路，从数据隔离、访问控制到日志审计，构建金融风控的技术底座，确保在严格监管环境下稳定运行。供应商唯有在全链条布局中形成自主可控、软硬件一体化的竞争力，才能真正支撑金融大模型在行业中的落地与长远发展。

中国金融大模型部署核心要素——部署选择

关键发现

在金融大模型的部署过程中，供应商必须确保在稳定性、准确性和安全性等方面的坚实保障。实现这些目标的关键在于其具备从基础硬件、云平台、机器学习框架、开发工具链到上层业务应用的全链路整合能力

中国金融机构在评估供应商时核心考量分析

金融机构评估供应商的全链路整合能力		
标准化上层应用与行业解决方案	**金融应用场景标准化产品及案例** 考察厂商是否具有成熟的行业应用案例及针对场景的标准化产品，如风险控制、智能投顾、反欺诈、智能客服等，证明其全链路方案在金融行业的实战经验	**大模型应用深度定制化服务** 考察厂商是否能够根据金融机构的特定需求进行定制化开发，提供专业的咨询、技术支持和后续服务，确保模型解决方案的快速落地和迭代更新
平台层的学习与机器学习框架	**自有平台支持** 考察厂商是否提供一站式的机器学习平台，涵盖数据采集、数据处理、模型训练、验证和上线，形成完整的ML生命周期管理	**算法与模型库** 考察厂商是否拥有金融领域定制的模型库及算法，能够快速响应市场变化，并支持在线学习或增量训练
应用开发平台与工具链	**开发工具与SDK** 考察厂商是否提供完善的开发工具包、SDK和API文档，支持模型二次开发、定制化功能扩展与集成商	**平台生态构建** 考察厂商是否能与主流的开发、监控、日志等工具链无缝对接，为金融机构构建一个全栈开放的平台
基础硬件与云服务	**硬件制造与云服务能力** 考察厂商是否拥有自研或自主采购的高性能计算集群，且这些集群是否支持异构芯片。评估厂商是否能提供公有云、私有云或混合云部署方案	**分布式与冗余设计** 考察厂商是否能够支持多区域部署、自动故障切换、容灾备份等机制，确保系统稳定运行

- **考虑到金融大模型部署的稳定性，金融机构在选择金融大模型厂商时，通常关注厂商是否具备全栈AI解决方案能力**

随着金融行业数字化转型需求的不断增长，金融机构在部署金融大模型时，越来越倾向于选择能够提供全栈AI解决方案的平台。这类平台不仅具备独特的异构资源调度能力，能够灵活支持国内外主流芯片架构的混合部署，还能实现算力资源池的智能动态编排与负载均衡，确保高效利用不同计算资源，从而提升计算能力和系统响应速度。此外，在自主可控性方面，这些平台构建了一个完整的技术体系，涵盖了芯片指令集适配层、分布式训练框架以及模型微调工具链，确保了全栈技术的无缝协同和高度兼容。这一体系不仅使平台能够满足金融行业对高效、安全、可定制化解决方案的需求，还能为金融机构提供从底层算力调度到上层AI应用的全生命周期支持。通过这种自主可控且具备灵活调度能力的技术架构，金融机构可以更好地应对快速变化的市场需求，同时保障数据安全性与系统稳定性，最大化发挥AI技术在金融领域的潜力。

来源：沙利文、头豹研究院

部署选择（接上页）

- **金融机构在选择大模型部署时，通常关注厂商是否能提供全链路保障体系能帮助其构建稳定性与安全性**

全链路保障体系是通过贯穿从基础硬件、网络设施、云平台到应用层的全程监控和管理，实现端到端的稳定性保障。在这一体系下，供应商不仅提供高可靠性的服务器、存储设备和冗余网络架构，还采用统一的监控与自动故障恢复机制，确保系统在出现异常时能够迅速切换并恢复正常运行。此外，全链路安全措施涵盖物理隔离、数据加密、访问控制、日志审计等多个层面，形成多重防护屏障，有效应对金融领域严苛的安全与合规要求。

- **金融机构在选择大模型部署时，通常关注厂商是否能提供端到端流程协同，以提升准确性及低延时高并发性能**

金融大模型的价值在于其决策的精准性以及对实时交易和风险管理场景的快速响应能力。全链路整合能力能实现数据采集、预处理、模型训练到在线推理各环节的无缝对接与深度优化。通过统一平台的调度与管理，不仅确保了数据在各阶段的一致性和高质量，也通过硬件加速（如GPU、TPU、FPGA）和边缘计算等技术，实现了极低的延时和高并发处理能力。这样的端到端协同，使得金融机构能够在瞬息万变的市场中迅速响应并作出精准判断，进而获得竞争优势。

- **金融机构在大模型部署中，开始关注新型数据平台产品，这些平台专为AI应用的优化而设计，相比传统的大数据平台能够显著提升AI应用的效果**

传统的大数据平台主要针对数据存储和处理，尽管在大规模数据处理上有其优势，但对于AI应用的支持却存在局限。随着AI技术的不断发展，特别是在金融大模型的应用中，传统数据平台难以满足高效的数据处理、实时反馈和智能决策等需求。为此，市场上逐渐出现了新型的数据平台，这些平台专门为AI应用进行优化，具备更强的计算能力、更低的延迟以及更高的可扩展性。与传统平台相比，这些新型平台能够更好地支持AI模型的训练、推理以及数据处理，尤其在数据适配、算法优化和应用场景支持上表现出更强的优势。金融机构在选择金融大模型厂商时，应重视这一新兴趋势，选择能够提供AI优化数据平台的解决方案，以确保AI应用的高效性和精准度。

来源：沙利文、头豹研究院

中国金融大模型市场追踪，2024H1

关键发现

沙利文联合头豹深入洞察中国金融大模型的商业化进展，重点关注中国金融大模型整体市场、MaaS（模型即服务）、标准化产品的市场份额。同时，分析了中国金融机构在部署金融大模型时的核心考量因素

报告完整版登录www.leadleo.com《中国金融大模型市场追踪报告，2024H1》

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

报告完整版登录www.leadleo.com
搜索《中国金融大模型市场报告,2024H1》

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

头豹 LeadLeo

FROST & SULLIVAN 沙利文