

2025中国金融大模型洞察

企业竞争分析

金融大模型，铸就企业核心竞争力

（阿里云·百度云·华为云·商汤科技）

企业标签：阿里云、华为云、商汤科技

AI变革行业创新发展

China Financial Large Model Industry
中国金融モデル産業

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。

名词解释

- ◆ **金融大模型**：指应用于金融领域的大型语言模型，它拥有大量参数和复杂结构，通常基于机器学习和人工智能技术。这些模型通过分析金融相关数据，并基于历史数据和主流的金融理论进行训练，从而能够识别和预测市场趋势，制定相关策略，提高金融决策的精度和效率。
- ◆ **金融级AI原生**：描述专为满足金融行业最严格需求而设计和优化的AI系统，包含安全性、可靠性、可扩展性、合规性等方面，适用于高复杂性和高风险的金融场景。
- ◆ **生成式AI**：一种人工智能模型，关注学习输入数据的分布规律，并生成与输入数据类似的新数据，广泛应用于图像生成、文本生成等领域。
- ◆ **判别式AI**：一种通过分析输入和输出之间的关系来进行分类或回归的AI模型，常用于推荐系统、风险控制等任务。
- ◆ **RAG（检索增强生成技术）**：结合信息检索与生成技术的系统，用于应对复杂查询和生成任务，如问答系统和内容创作。
- ◆ **大模型规模定律（Scaling Law）**：描述大模型参数数量与性能提升的关系，强调模型规模扩大会显著提升其学习能力和任务处理性能。
- ◆ **低延时与高并发**：大模型处理实时任务（如交易监控）和高并发场景（如大规模用户同时请求）的关键能力，依赖模型优化和分布式架构实现。
- ◆ **模型微调**：在预训练模型的基础上，利用特定领域的数据对模型进行调整，以提升其在特定任务中的表现。
- ◆ **云原生架构**：通过容器化、微服务等技术实现资源弹性和高效管理，为大模型训练和推理提供灵活的基础设施。
- ◆ **金融云专属VPC模式**：一种部署形式，将大模型应用和知识库部署在金融云客户的专属虚拟私有云中，确保数据隐私和安全。

Chapter 1

金融大模型

企业竞争优势分析

- ☐ 企业选择
- ☐ 阿里云
- ☐ 百度智能云
- ☐ 华为云
- ☐ 科大讯飞
- ☐ 商汤科技
- ☐ 火山引擎
- ☐ 智谱AI

中国金融大模型企业评价——企业选择

- 在金融大模型的部署过程中，供应商必须确保在稳定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



■ 全链路保障体系构建稳定性与安全性

全链路整合能力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，实现了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。

■ 端到端流程协同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。全链路整合能力实现了数据采集、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。

来源：头豹研究院

中国金融大模型代表性企业——阿里云

- 阿里云在金融大模型领域核心竞争力体现在其强大的技术平台、深厚的AI能力、全面的解决方案和严格的数据安全与合规保障，同时通过产品标准化和灵活的部署模式，满足不同规模金融机构的需求

阿里云金融大模型服务框架



✓ 企业亮点

01

全栈技术整合

提供全链路解决方案，涵盖硬件、云服务、机器学习框架及开发工具链，确保金融大模型稳定、准确、安全

02

Cloud+AI深度融合

金融大模型与云计算结合，利用其可扩展资源加速训练部署，提升迭代效率，为金融行业智能化升级奠定基础

03

产品标准化

旗下众多以大模型为基底的标准产品能以MaaS的部署模式，满足追求快速，轻量化上线大模型产品的中小型金融机构

■ 阿里云具有强大的技术平台与深厚的AI能力

阿里云在金融大模型领域的核心竞争力之一是其强大的技术平台和深厚的AI能力。阿里云通过其“飞天”云计算平台提供了高性能的计算资源，并结合大规模的分布式计算框架，能够支持复杂的金融大模型训练和推理。这一平台可以有效处理大规模的数据流和计算需求，保障模型的高效运行。与此同时，阿里云在机器学习、深度学习等AI领域的技术积累，使其在金融领域的应用场景中具有显著优势。例如，阿里云的金融大模型能够支持复杂的信用评估、风险管理和智能投顾等业务，提升金融机构的决策精度与业务效率。

■ 阿里云为金融行业提供的全面解决方案和服务体系

阿里云不仅提供标准化的产品和服务，还能根据客户的特定需求提供高度定制化的解决方案。这种灵活性确保了不同规模和类型的金融机构都能找到适合自己的服务模式。同时，阿里云注重数据安全性与合规性，为金融客户提供了一系列的安全保障措施，包括但不限于数据加密、隐私保护等，满足了金融行业严格的安全标准。此外，阿里云还建立了一支由行业专家和技术精英组成的团队，他们不仅具备丰富的行业经验，而且能够及时响应客户需求，提供专业的技术支持和服务。

来源：Fabarta，头豹研究院

中国金融大模型代表性企业——百度智能云

- 百度智能云凭借千帆大模型开发与服务平台，结合“百舸”异构算力平台，为金融机构提供了强大的定制化模型构建能力和高效的算力管理，显著降低了技术成本，推动了业务创新和风险管理提升

百度智能云金融大模型服务框架

千帆AppBuilder						
千帆ModelBuilder						
模型广场	大模型	文心大模型	第三方大模型			
	通用大模型	ERNIE Bot	ChatGLM	baichuan2	Llama 2	HuggingFace Transformers
	行业大模型	ERNIE-ViLG	RWKV	Stable Diffusion	元象Xverse	
大模型工具链	数据管理	模型精调	模型评估&优化	推理服务部署	Prompt工程	
	数据集管理	Post-pretrain	模型管理	体验中心	预置Prompt模版	
	数据标注	SFT-全量更新	模型评估	预置服务调用	自制Prompt模版	
	数据清洗	SFT-LoRA	模型压缩	定制服务部署	Prompt优化	
	数据增强	RLHF		调用监控	Prompt评估	
	数据分析	增量训练		插件中心		
BML•AI开发平台						
百度百舸 AI异构计算平台						

01

一站式企业级大模型平台
“千帆”平台为企业客户提供从数据管理、模型训练、评估与优化到预测服务和Prompt工程等全流程的工具链和支持

02

行业深耕的定制化解决方案
百度智能云针对金融行业的需求，提供了专门的AI解决方案，涵盖智能投顾、风险控制、智能客服等多个领域

03

高效的算力管理和成本控制
该平台能够统一管理不同地点、规模和集群的智能算力，并通过芯片性能优化、自动芯片选型等技术提升效能

- 百度智能云凭借其千帆大模型开发与服务平台，为金融机构提供了强大的定制化模型构建能力
- 百度智能云在高效算力管理方面具有显著优势，大幅降低了金融机构使用大模型技术的成本和门槛

通过千帆平台，百度智能云能够支持金融机构利用私域数据进行深度挖掘和分析，进而构建出高度定制化的金融大模型。该平台不仅提供丰富的算法模型和数据处理工具，还拥有灵活的部署选项，如云端、本地或混合部署等，使得金融机构可以根据自身需求选择最适合的方案。此外，千帆平台还能实时监控模型性能并提供优化建议，确保模型在实际应用中的准确性和稳定性，从而帮助金融机构提升服务效率和风险管理水平。

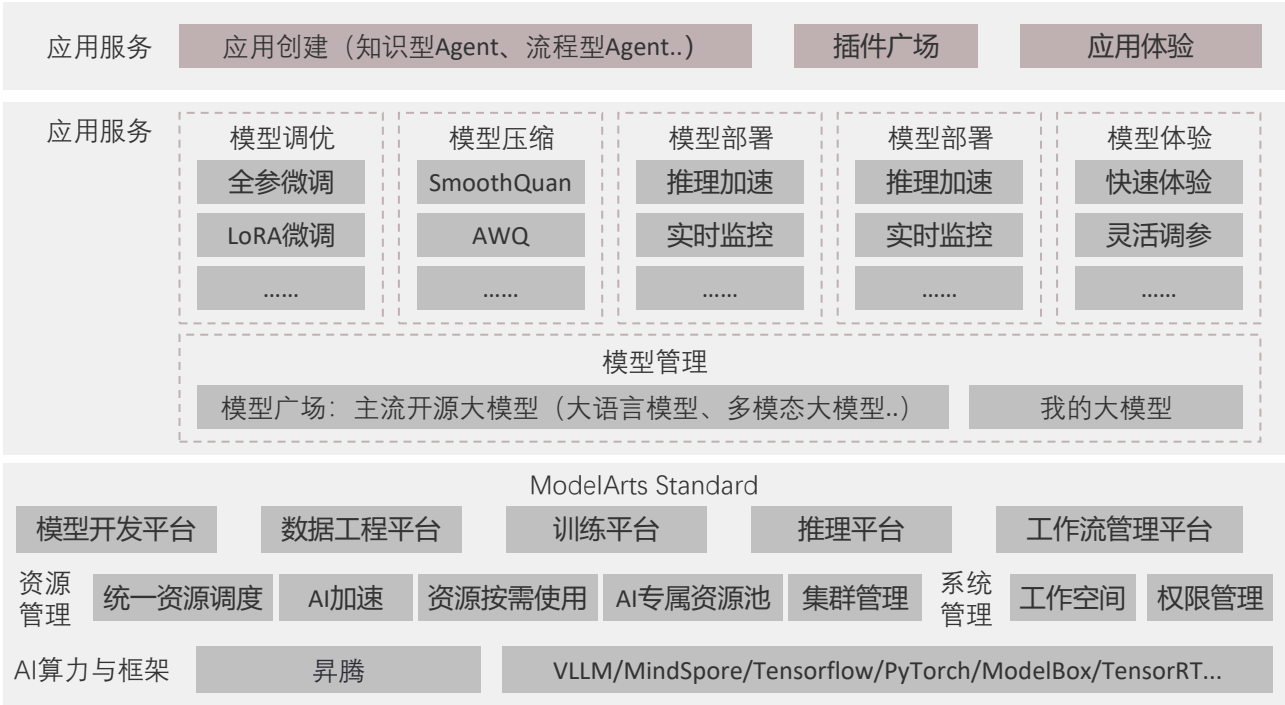
面对大模型训练所需的庞大计算资源和高昂成本，百度智能云的“百舸”异构算力平台提供了一站式的解决方案。它不仅能够对不同厂商的AI芯片进行统一管理和优化，还采用了潮汐混部等先进技术提高计算效能。这使得金融机构可以在享受高性能计算的同时，有效降低算力成本。例如，对于相同的大模型任务，百舸可以将资源成本减少一半，极大地提升了投资回报率，并让更多的金融机构有能力采用先进的大模型技术来驱动业务创新和发展。

来源：头豹研究院

中国金融大模型代表性企业——华为云

- 华为云通过其卓越的软硬一体化解决方案，结合昇腾AI处理器与鲲鹏服务器，为金融机构提供高效稳定的计算能力，能够满足复杂模型训练与数据处理的需求

华为云金融大模型服务框架



01

自研AI芯片昇腾

华为云依托自研的昇腾AI芯片，提供强大的计算能力，尤其在大模型训练和推理过程中，能够提供更高的运算效率和成本效益

02

全栈技术方案与灵活云部署

华为云提供从硬件、操作系统到算法框架的全栈技术方案，帮助金融机构根据具体需求进行灵活的云服务部署

03

软硬一体化的AI解决方案

华为云通过自研的昇腾AI芯片和高性能计算硬件，结合强大的AI框架和算法，提供软硬一体化的解决方案

- 华为云通过其卓越的软硬一体化解决方案，为金融机构提供了高效稳定的计算能力，满足了复杂模型训练与数据处理的需求

华为云利用昇腾AI处理器和鲲鹏服务器等硬件设施，结合其优化的软件生态系统，如华为云Stack及AI开发平台ModelArts，构建了一个强大的计算平台。这一平台不仅能够提供行业领先的AI算力支持，还通过软硬件协同优化技术，实现了资源使用的最大化效率。对于金融机构而言，这意味着可以更快速地进行大规模数据处理、复杂的金融大模型训练以及实时风险分析等关键任务，同时保证了系统的稳定性和安全性。

- 华为云凭借在金融行业的深厚积累与广泛合作，提供定制化的金融大模型解决方案

华为云已经服务于超过500家金融机构，包括中国的主要银行和保险企业，这表明其在金融领域拥有丰富的实践经验和深刻的理解。基于这种背景，华为云能够针对金融机构的具体需求提供定制化的解决方案，例如使用盘古大模型来预训练银行文档并自动生成操作指导。此外，华为云还致力于推动技术创新与业务场景的深度融合，推出了诸如主机上云、金融多活容灾等针对性服务，有效帮助金融机构应对市场变化和技术挑战，加快其数字化转型的步伐。

来源：Fabarta，头豹研究院

中国金融大模型代表性企业——科大讯飞

- 科大讯飞凭借自主可控的技术平台和深度结合行业应用的能力，在金融大模型领域不仅提供高效、安全的解决方案，还成功推动了AI技术在金融行业的广泛落地

科大讯飞金融大模型服务框架



01

强大的技术研发实力

科大讯飞拥有自建的一流AI数据中心，并实现了百亿参数大模型推理效率的显著提升

02

应用场景广泛

科大讯飞金融大模型被广泛应用于多个金融场景，包括但不限于智能投顾、风险管理、客户服务等

03

持续创新能力

科大讯飞不断投入研发资源，推动金融大模型的迭代升级。公司紧跟市场趋势和技术发展，不断优化模型算法，提升模型的准确性

科大讯飞在金融大模型领域技术自主可控与性能领先

科大讯飞金融大模型的核心竞争力之一在于其技术和平台的自主可控性以及在性能上的领先地位。作为中国少数几家能够提供从底层算力到上层应用全面解决方案的人工智能企业，科大讯飞不仅构建了全国产化的训练和推理平台“飞星一号”，还通过与华为的合作实现了对国产算力的深度优化。这使得科大讯飞能够在保障数据安全的前提下，为金融机构提供高效稳定的服务。这一自主可控的技术架构使得科大讯飞在应对金融行业日益严格的合规性要求时具备明显优势，同时能够有效降低对国际技术供应商的依赖。

科大讯飞金融大模型实现行业理解和应用场景深度融合

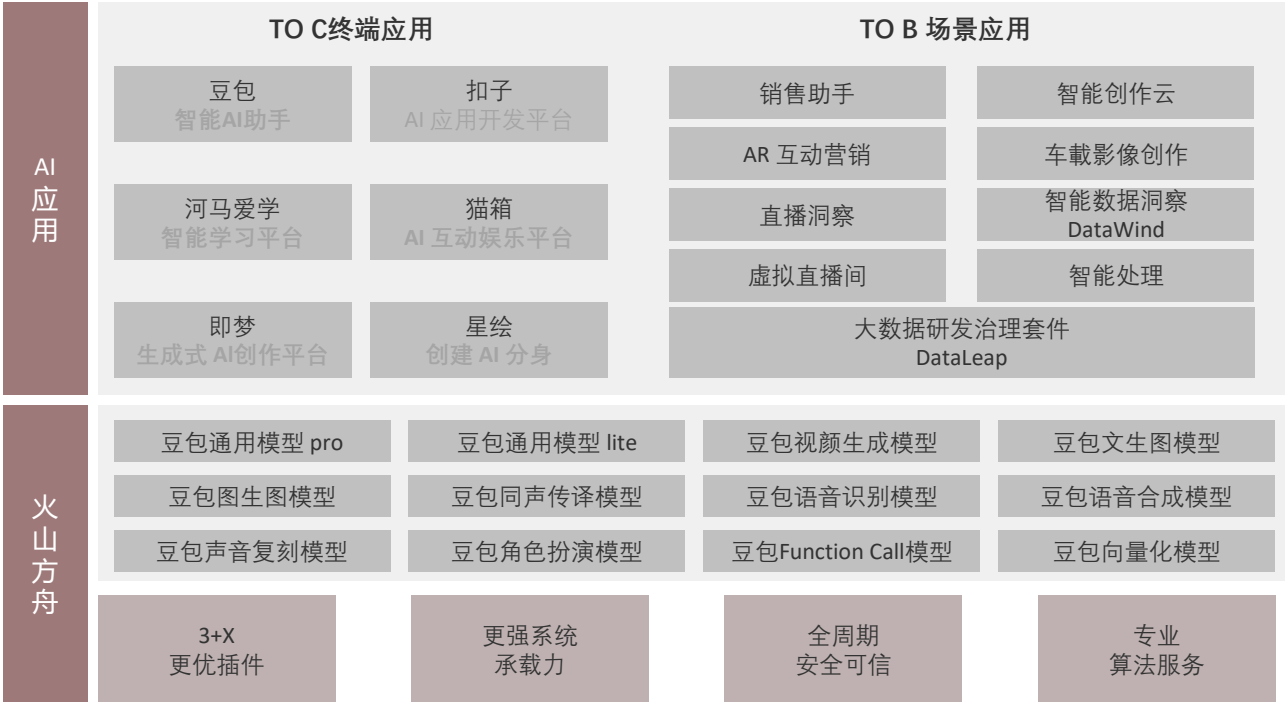
科大讯飞凭借对金融行业的深刻理解以及将AI技术与实际业务场景紧密结合的能力，在金融领域展现出强大的竞争力。尽管金融行业是最适合大模型落地应用的领域之一，但监管合规和数据安全等问题使得金融机构在采用新技术时往往保持谨慎态度。面对这些挑战，科大讯飞依托多年积累的经验和雄厚的技术实力，成功推出了多款面向金融行业的AI助手产品，如星火展业助手等，全面覆盖了客户经理展业、尽调报告撰写、客户服务等多个关键业务环节。

来源：Fabarta，头豹研究院

中国金融大模型代表性企业——火山引擎

- 火山引擎金融大模型通过深度融合“模型-应用-数据”飞轮机制，推动技术与实际业务场景紧密结合，提供高效且定制化的解决方案，同时借助强大的技术和灵活服务体系，满足不同金融机构多样化需求

火山引擎金融大模型服务框架



01

数据安全与合规保障

高度重视数据安全和隐私保护，采用多种安全技术和措施确保客户数据的安全存储和传输

02

持续创新与优化能力

字节跳动不断加大在AI领域的研发投入，推动豆包大模型等技术的持续创新和优化

03

强大的算力支持与工程化能力

混合云 veStack 智算版拥有万卡集群组网和3.2T 高性能无损网络的优势，能实现97.78%训练加速比，并具备分钟级故障发现和自愈功能

- 火山引擎金融大模型深度融合业务场景的模型-应用-数据飞轮
- 火山引擎金融大模型的核心竞争力之一在于其构建了“以模型-应用-数据为核心、算力为支撑”的大模型飞轮机制。这一机制不仅强调了模型的重要性，还突出了模型与应用场景之间的紧密结合以及数据反馈对模型迭代的支持作用。具体来说，火山引擎与金融机构合作，共同探索和实践了多个金融领域的应用场景，如智能客服、精准营销、风险管理等，并在这些领域内取得了显著成效。例如，在海尔消费金融的实际案例中，利用大模型优化坐席助手功能，使得通话记录的归纳摘要更加准确，从而提高了工作人员的工作效率和服务质量。
- 火山引擎金融大模型具备强大的技术支持与灵活的服务体系
- 火山引擎另一大核心竞争力体现在其强大的技术支持和灵活的服务体系上。火山引擎提供了包括混合云 veStack 智算版在内的先进算力支持，以及针对不同客户需求定制化的模型服务。特别是其“1+1+N”模型生态策略，结合了字节跳动自主研发的豆包大模型、与智谱AI联合打造的火山引擎-智谱AI金融行业大模型以及其他主流开源或商业化模型，为企业提供了多样化的选择。

来源：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

头豹深圳研究院

广东省深圳市南山区粤海街道华润置地大厦E座
4105室

头豹上海研究院

上海市静安区南京西路1717号会德丰国际广场
2504室

头豹南京研究院

江苏省南京市栖霞区经济开发区兴智科技园B栋
401

报告作者



袁栩聪
首席分析师
oliver.yuan@leadleo.com



陈庆民
行业分析师
qingmin.chen@leadleo.com

业务咨询

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

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。