中国信息协会投融资信息分会
CHINA ASSOCIATION OF INFORMATION AND FINANCING INFORMATION BRANCH

TSINGHUA PBCSF FSRC
清华五道口 金融安全研究中心

金融人工智能发展与安全白皮书（2025）

、

中国信息协会投融资信息分会

清华大学五道口金融学院金融安全研究中心

课题组简介

编撰单位：

中国信息协会投融资信息分会

清华大学五道口金融学院金融安全研究中心

组长：

谢荣全　中国信息协会投融资信息分会会长

周道许　清华大学五道口金融学院金融安全研究中心主任

课题组参与人员：

中国信息协会投融资信息分会：

周　波　中国信息协会投融资信息分会专职副会长兼秘书长

沈国华　中国信息协会投融资信息分会专家委员会副主任委员

陈克军　中国信息协会投融资信息分会专家委员会副主任委员

周嘉仪　慕尼黑工业大学社会科学与技术学院人工智能与社会专业硕士研究生

清华大学五道口金融学院金融安全研究中心：

周　京　清华大学五道口金融学院金融安全研究中心总监

黄旌沛　清华大学五道口金融学院金融安全研究中心研究专员

沈乐阳　清华大学五道口金融学院金融安全研究中心研究专员

丁浩洋　清华大学五道口金融学院技术转移专业硕士研究生

徐禔泽　纽约大学经济学专业博士研究生

刘铠硕　清华大学五道口金融学院金融安全研究中心实习生

序言一：擘画智能时代新蓝图，筑牢国家金融安全屏障

当今世界，新一轮科技革命与产业变革的进程加速推进，人工智能作为引领性的战略性技术，正以前所未有的广度与深度重塑全球经济社会格局。金融业作为现代经济的核心，是人工智能技术应用最为活跃、影响最为深远的关键领域之一。推动金融与人工智能的健康发展，不仅是把握未来发展的主动权、培育产业新质生产力的战略抉择，更是维护国家金融安全、筑牢网络强国根基的时代要求。

实践表明，金融人工智能在提升服务效率、创新产品模式、优化风险控制等方面展现出巨大潜力，为金融高质量发展注入了强劲动能。然而，技术应用的背后，新的风险与挑战也随之浮现。从数据安全、算法公平，到模型可靠性，再到技术供应链自主可控，每个环节都潜藏着可能冲击金融体系稳定的安全隐患。必须秉持总体国家安全观，坚持发展与安全并重，以高水平安全护航金融高质量发展。

本白皮书的撰写，旨在全面梳理全球及我国金融人工智能的发展现状与产业格局，系统剖析其核心技术应用与潜在安全风险，并探索构建兼具前瞻性与适应性的治理框架。我们期望通过这份报告，凝聚行业共识，为主管部门制定科学决策提供参考，为金融机构部署 AI 战略提供指引，为全社会共同构建安全、可信、普惠的金融智能生态贡献智慧与力量。

谢荣全

中国信息协会投融资信息分会

2025 年 10 月

序言二：洞察技术变革深层逻辑，探索金融安全治理新范式

人工智能，特别是以大语言模型为代表的生成式 AI，正对金融理论与实践构成颠覆性冲击。它不仅是提升效率的工具，更是一种重构金融服务价值链、改变风险生成与传导机制的底层力量。理解并驾驭这股力量，需要我们超越对技术应用的浅层观察，深入其内在逻辑，以严谨的学术精神和深厚的行业洞察，系统性地审视其带来的机遇与挑战。

金融的本质是经营风险，而金融安全是金融稳定与发展的基石。人工智能的“黑箱”特性、决策的不可解释性、以及潜在的算法偏见和模型脆弱性，都给传统的金融风险管理和监管体系带来了前所未有的考验。如何确保 AI 在金融领域的应用“可知、可信、可控、可用”，如何构建与技术创新相匹配的“敏捷治理”框架，是摆在我们面前亟待破解的重大课题。

作为国家重要的金融智库，清华大学五道口金融学院金融安全研究中心联合业界权威机构，共同发起此项研究。我们致力于搭建一个跨学科、跨领域的对话平台，融合技术前沿、金融实务与公共政策，对金融人工智能的安全与治理问题进行体系化、学理化的深入剖析。本报告力求以客观的数据、审慎的分析和前瞻的思考，为探索中国特色的金融人工智能治理新范式、维护国家金融长治久安贡献学术界的绵薄之力。

周道许

清华大学五道口金融学院金融安全研究中心

2025 年 10 月

目录

前言

我们正处在一个由人工智能（AI）驱动的深刻变革时代。作为现代经济的核心，金融业正以前所未有的速度和广度拥抱这一变革，开启了新一轮的科技革命与产业重塑。从重塑客户体验到革新风险管理，从优化交易决策到提升运营效率，人工智能正从金融业的外围辅助工具，迅速演进为驱动业态发展的核心生产力。特别是以金融大模型为代表的生成式 AI，其强大的认知、理解与创造能力，正在叩开一个前所未见的智能金融新纪元的大门。

然而，机遇与挑战并存，创新与风险相伴。在金融 AI 领域，这种对立统一的关系尤为显著：一方面，它凭借数据处理效率的提升、风险识别精度的优化以及服务场景的拓展，为金融行业注入了强大动能——从智能投顾实现个性化资产配置，到 AI 风控实时拦截欺诈交易，再到智能客服覆盖 7×24 小时服务需求，其价值创造能力已在支付、信贷、资管等多个细分领域充分显现；另一方面，随着技术深度融入金融业务流程，前所未有的复杂风险也随之滋生，成为行业发展必须直面的课题

在此关键发展阶段，我们深刻认识到，推进智能金融建设的过程中，创新发展与安全治理二者缺一不可、不能偏废。安全不再是制约发展的“刹车片”，也不应被视为单纯的成本中心，而是支撑金融 AI 高质量发展的内在要求和核心竞争力。一个缺乏安全保障、无法确保可信性与可控性的 AI 系统，即便短期内能带来一定的效率提升或收益增长，其创造的价值也必然是脆弱且不可持续的。唯有将安全理念贯穿于金融 AI 从技术研发、模型训练到业务应用的全生命周期，在创新中筑牢安全防线，在治理中为创新保驾护航，才能推动智能金融行稳致远。

基于这一共识，《金融人工智能发展与安全白皮书（2025）》由清华大学五道口金融学院金融安全研究中心与中国信息协会投融资信息分会联合编撰。我们旨在融合两种视角：一方面，是清华大学五道口金融学院立足于宏观经济、金融稳定与公共政策的学术前瞻性洞察；另一方面，中国信息协会投融资信息分会深耕行业实践，依托服务广大金融机构过程中积累的经验、产业资源整合能力及实务案例储备，提供贴合真实金融业务场景的实践视角。

我们期望通过这双重视角的深度融合，系统性地梳理金融人工智能的发展脉

络、应用场景与潜在风险，并提出一套兼具战略高度与实践价值的安全治理框架。我们的目标是为中国金融机构的 AI 转型提供清晰的路线图，为科技企业的技术创新指明方向，也为监管部门的政策制定提供有益的参考，共同擘画并推动我国金融人工智能生态的健康、稳定与繁荣。

摘要

金融 AI 已进入规模化应用与价值兑现的关键时期，其渗透正从外围辅助业务转向核心决策系统。以金融大模型为代表的生成式 AI 技术，正以其强大的内容生成与逻辑推理能力，深刻变革着智慧营销、智能投研、风险控制和运营管理等各个环节，成为金融业新一轮增长的核心引擎。然而，这种深度融合也催生了前所未有的复合型风险：模型自身的脆弱性（如对抗性攻击、后门攻击）成为新的攻击平面；算法的“黑箱”与“幻觉”问题挑战着决策的可靠性与可信度；海量数据的集中处理放大了隐私泄露与数据污染的风险；而算法偏见可能固化甚至加剧社会不公，对金融普惠构成挑战。此外，当多家机构采用相似的 AI 模型时，可能引发“模型趋同”效应，构成新型系统性金融风险。

面对机遇与挑战并存的局面，本白皮书提出，金融 AI 的发展必须超越单纯的技术视角，构建“以业务价值为导向、以数据安全为基石、以可信 AI 为治理框架、以智能防御为技术手段”的四位一体发展与安全新范式。在此范式下，安全不再是发展的制约因素，而是保障其商业价值实现、赢得市场信任、构筑长期竞争优势的内在核心能力。

白皮书对各方提出具体行动建议：对金融机构，要把 AI 安全与治理提升到公司战略层面，设立由高管层领导的跨部门 AI 治理委员会；技术上，构建覆盖 AI 全生命周期的安全防护体系，将安全能力“左移”至开发阶段，积极采用隐私增强计算、零信任架构等新一代安全技术，还要大力培养兼具金融业务、AI 技术与安全知识的复合型人才。对科技企业，需将“安全设计”理念融入 AI 产品与服务的设计研发全过程，提供安全、透明、可解释的 AI 解决方案，加强供应链安全管理，与金融机构深度合作，共建开放、协同的金融 AI 安全生态。对监管部门，建议加快完善与 AI 技术发展相适应的顶层法规体系，研究出台全国统一的《金融人工智能管理办法》；探索“监管沙盒”等敏捷治理模式，在鼓励创新时严守风险底线；推动建立行业级的 AI 风险案例库、安全威胁情报共享平台和第三方 AI 测评认证机构，提升整个金融体系的风险抵御能力。展望未来，随着多模态融合、AI Agent 等技术进一步成熟，金融 AI 会有更巨大的潜能。我

们坚信，一个安全、可信、负责任的金融 AI 生态，不只是防范风险的“防火墙”，更是驱动中国金融业迈向更高质量、更具韧性未来的“加速器”。

展望未来，随着多模态融合、AI Agent 等技术的进一步成熟，金融 AI 将展现出更为巨大的潜能。我们坚信，一个安全、可信、负责任的金融 AI 生态，不仅是防范风险的“防火墙”，更是驱动中国金融业迈向更高质量、更具韧性未来的“加速器”。

一、绪论：新范式下的金融变革与安全共生

(一)研究背景与意义

进入21世纪的第三个十年，人工智能不再是遥远的概念，而是触手可及、深刻重塑世界的现实力量。金融，作为现代经济中数据最密集、知识最密集的行业之一，天然地成为了AI技术落地应用、创造价值的核心场域。本白皮书选择在2025年这一关键节点发布，是基于我们对技术、市场与风险三重要素交汇演进的深刻洞察。

1. 技术奇点：人工智能成为重塑金融业态的核心驱动力

以深度学习,特别是自注意力机制和Transformer架构为基础的大模型技术，在近年来取得了突破性进展。这不仅是量变，更是质变。AI的能力从感知、识别，跃升至理解、推理与创造，其通用性与赋能潜力使其成为引领新一轮科技革命和产业变革的战略性技术。从华尔街的量化交易基金到中国大型商业银行的智能风控平台，AI正以前所未有的深度和广度，从根本上改变着金融服务的提供方式、金融产品的设计逻辑、金融市场的运行机制和金融机构的运营模式。它不再仅仅是的工具，而是驱动金融业迈向未来的核心引擎。

2. 应用发展：2025年成为规模化落地与价值兑现的交汇点

2025年被普遍视为金融AI发展的一个关键分水岭。其一，成本效益的“剪刀差”正在形成。以OpenAI、Google、Anthropic等为代表的头部模型公司掀起价格战，API调用成本在2024年出现90%以上的降幅，同时模型性能却在持续提升。这使得AI应用的门槛被前所未有地拉低。其二，应用场景从“点”到“面”。AI应用正从智能客服等外围环节，加速渗透至信贷审批、投资决策、合规审查等核心业务地带，从少数头部机构的“试验性”项目，向广大中小金融机构的“规模化”部署迈进。据统计，超过30%的中国金融机构已实际应用AI技术，一个万亿级的金融AI市场正加速形成。

3.风险显现：安全成为金融AI行稳致远的核心议题

当AI系统开始承担核心决策功能，其潜在的风险便直接关联到金融消费者的切身利益乃至整个金融系统的稳定。模型的脆弱性如同潜在隐患，在实际应用中一旦被放大，极易引发决策灾难，给相关领域带来严重损失；数据的滥用可能引发大规模的隐私危机；算法的偏见可能固化甚至加剧社会的不平等。这些前所

未有的安全风险，随着应用的深化而集中暴露，使“AI 安全”这一议题，从技术人员的讨论范畴，上升为金融机构董事会和监管机构必须直面的战略性问题。

3. 范式转换：安全从成本中心到金融人工智能核心竞争力的跃升

在传统 IT 时代，信息安全往往被视为业务发展的制约因素，是一个需要控制投入的成本中心。但在智能时代，安全的内涵和外延都发生了根本性变化。AI 系统的安全、可信与稳健，直接决定了其业务价值能否充分、可靠地实现。一个存在严重漏洞的智能投顾模型，不仅无法创造收益，反而可能带来巨额亏损；一个带有歧视的信贷审批模型，不仅无法提升效率，反而会引发合规风险和声誉危机。因此，AI 安全不再是可有可无的附加选项，而是内嵌于 AI 系统需求、设计、开发、运营全生命周期的核心属性，它正在成为金融机构在智能化浪潮中赢得客户信任、规避重大风险、实现可持续发展的核心竞争力。

(二)核心概念界定

为确保论述的清晰与严谨，本白皮书对以下核心概念进行界定：

1. 金融人工智能（Financial AI）

金融人工智能是一个综合性的技术应用体系，它运用包括机器学习（特别是深度学习）、自然语言处理（NLP）、计算机视觉、知识图谱、生成式 AI 等在内的人工智能技术，对海量、多模态的金融数据进行处理、分析、理解与决策，以赋能或重塑金融产品设计、客户服务、投资交易、风险管理、运营支持等各类金融活动。其技术栈涵盖了从底层的 AI 芯片与算力设施，到中间层的数据处理与 AI 框架，再到上层的模型算法与应用场景的全链条。

2.生成式 AI（Generative AI）在金融领域的特定内涵

生成式 AI 模型能够学习金融领域的专业知识、数据分布和复杂模式，并据此生成全新的、原创性的、符合特定要求的内容的 AI 模型。在金融领域，其生成内容可以是文本（如市场分析报告、营销文案、客服对话）、代码（如 Python 量化策略、SQL 查询语句）、结构化数据（如模拟财务报表）、乃至投资策略组合等。以金融行业大模型（Financial Large Language Model,Fin LLM）为代表，它不仅能进行交互和分析，更能进行归纳、演绎、推理和创造，是当前推动金融 AI 变革最具颠覆性的力量。

3.金融 AI 安全（Financial AI Security）

金融 AI 安全是一个贯穿 AI 系统全生命周期的、多层次、多维度的综合性安全概念。它是一个远超传统网络安全范畴的拓展集合，其内涵至少包括：

（1）基础架构安全：承载 AI 运行的软硬件基础设施（网络、服务器、云平台）的安全。

（2）数据安全：AI 全生命周期中（采集、传输、存储、处理、销毁）所涉及数据的保密性、完整性、可用性与合规性，特别是个人金融信息的保护。

（3）模型安全：AI 模型自身的内在安全，包括抵御对抗性攻击、后门攻击、模型窃取等恶意攻击的鲁棒性，以及模型自身的稳定性和可靠性。

（4）应用安全：AI 作为服务或产品被调用和使用时的安全，包括 API 安全、权限控制、防范提示注入攻击等。

（5）治理与合规：确保 AI 系统的设计与应用符合法律法规、伦理规范和社会价值观，涵盖算法的公平性、透明度、可解释性、问责制以及对 AI 产生内容的管理等。

当前人工智能已深度融入金融行业，成为重构行业生态的关键力量。以深度学习为核心，人工智能从感知识别延伸至理解推理、自主创造，全面渗透金融服务、产品设计、市场运行及机构运营等领域，推动金融业向智能化跃迁。AI 场景从智能客服等外围环节，向信贷审批、投资决策等核心业务渗透。需警惕的是，AI 承担核心决策职能后，模型鲁棒性不足、数据管理不当、算法设计缺陷等风险显现，直接关联金融消费者权益与系统稳定，安全防护至关重要。

二、全球与中国金融人工智能发展态势

金融 AI 的浪潮已席卷全球，但不同经济体在政策导向、市场应用和产业生态上展现出各具特色的发展路径。深入理解全球格局，并在此坐标系下审视中国的发展现状与特色，对于把握未来趋势至关重要。

(一)全球市场格局与多元化发展趋势

1.市场规模与投资热点

全球金融 AI 市场正经历爆炸式增长。据麦肯锡预测，在生成式 AI 的推动下，AI 技术每年可为全球金融业（银行、保险、投资管理）带来 2500 亿至 4100 亿美元的巨大增量价值。这一巨大的商业前景正吸引着资本的大量涌入。投资热点主要集中在三大领域：

一是生成式 AI 的垂直应用，生成式 AI 在金融领域已进入“场景渗透+专业深化”的新阶段，典型落地场景包括智能投研、智能财富管理、代码生成与辅助开发三大方向；

二是 AI 驱动的风险管理与合规科技（RegTech），通过 AI 技术重构风险识别与合规管理流程，显著提升反欺诈、反洗钱及信贷风控三大核心场景的运行效率与判断精度；

三是金融基础设施的智能化升级，以NVIDIA等厂商为代表的AI算力提供商，其高性能 GPU 已成为支撑这轮 AI 浪潮的关键基石，其市值飙升也反映了市场的巨大预期。

2.区域对比：美、欧、中引领，各具特色的发展路径

全球金融 AI 的发展并非铁板一块，而是呈现出鲜明的区域特色：

（1）美国：“市场驱动，技术引领”

美国凭借其强大的科技巨头（如 Google,Microsoft,OpenAI）、全球领先的金融市场和活跃的风险投资生态，在金融 AI 的应用深度和广度上保持全球领先。摩根大通（JPMorgan Chase）、高盛等顶级投行不仅是 AI 技术的重度使用者，更是积极的研发者。例如，摩根大通已将 AI 技术深度整合到交易执行、风险建模、客户服务等核心流程中，并开发了自有大模型用于文档解析。根据 KPMG 的数据，美国金融机构的 AI 渗透率高达 88%。其发展模式的特点是，由市场需求和技术创新双轮驱动，监管则相对保持灵活性和适应性，鼓励合理的创新。

（2）欧盟：“监管先行，人本导向”

欧盟在全球范围内率先举起了AI强监管的旗帜。其于2024年正式通过的《人工智能法案》（AI Act）是全球首部具有法律约束力的横向AI监管框架。该法案的核心是基于风险的分级管理，对高风险金融应用（如信贷评分、保险定价、员工招聘）施加了极其严格的规制，要求其在数据质量、透明度、人类监督和网络安全等方面满足一系列强制性要求。欧盟的模式体现了其保护基本权利的核心价值观，其“监管先行”的策略虽然可能在短期内增加创新成本，但旨在为AI的长期健康发展奠定信任基础。

（3）中国：“政策引导，场景制胜”

中国的发展特色在于“顶层政策设计”与“海量市场规模”的双轮驱动。下文将详细展开。

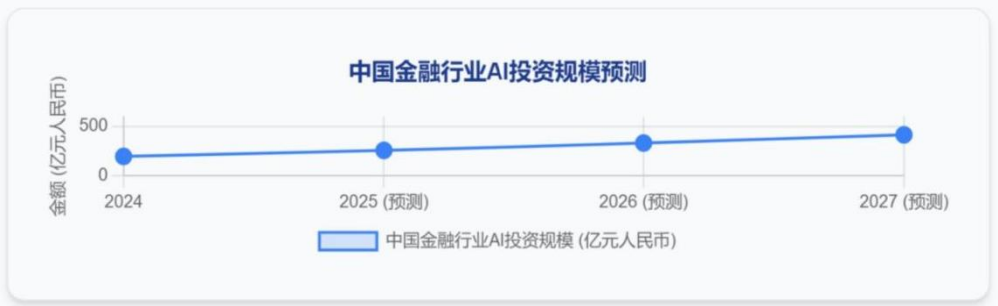
3.产业生态：多元竞合，巨头与“小巨人”共舞

全球金融AI的产业生态呈现出多元化、多层次的竞合关系。头部金融机构凭借其深厚的领域知识、海量的数据资产和雄厚的资本，积极自研或与科技巨头深度合作，打造专属的AI能力。科技巨头（如Google,Microsoft,AWS）则提供底层的通用大模型和强大的云计算平台，成为AI时代新的“基础设施”提供商。与此同时，大量充满活力的初创公司在各个细分赛道（如智能投顾、合规科技、AI安全）不断涌现，凭借其技术专长和敏捷性，成为推动金融AI创新的重要力量。这种多方参与、互为补充的生态系统，共同推动着金融AI技术的快速迭代和应用的深化。

(二)中国金融人工智能发展现状：政策与市场双轮驱动

1.市场规模与增长预测：驶入高速增长快车道

中国金融AI市场正以惊人的速度扩张。根据IDC数据显示，2024年中国金融行业AI投资规模已达到196.94亿元人民币，预计到2027年将翻番增至415.48亿元，增幅将达111%，这一数据远超同期金融科技整体市场的增速。其中，2024年，中国金融大模型市场迎来了爆发式增长，市场规模达到28.66亿元，同比增长80%。中国金融AI的应用目前已经度过了早期的概念验证（PoC）阶段，正迈向规模化落地和价值创造的新时期。



2.政策驱动与“信创”背景：国家战略下的历史性机遇

中国政府将人工智能上升到国家战略高度，自 2017 年发布《新一代人工智能发展规划》以来，一系列支持政策密集出台，为产业发展提供了强有力的引导和支持。与此同时，在金融信息技术应用创新的大背景下，金融机构对安全、自主、可控的 IT 基础设施和解决方案的需求空前高涨。这一宏观战略要求金融机构在核心系统、数据库、中间件乃至上层应用中，逐步实现国产化替代。AI 作为新一代信息技术的皇冠，其自主可控的重要性不言而喻。这为拥有自主知识产权的国产 AI 芯片、AI 框架和 AI 应用厂商，提供了前所未有的历史性发展机遇。

3.应用渗透：从“外援”到“主力”，从“辅助”到“决策”

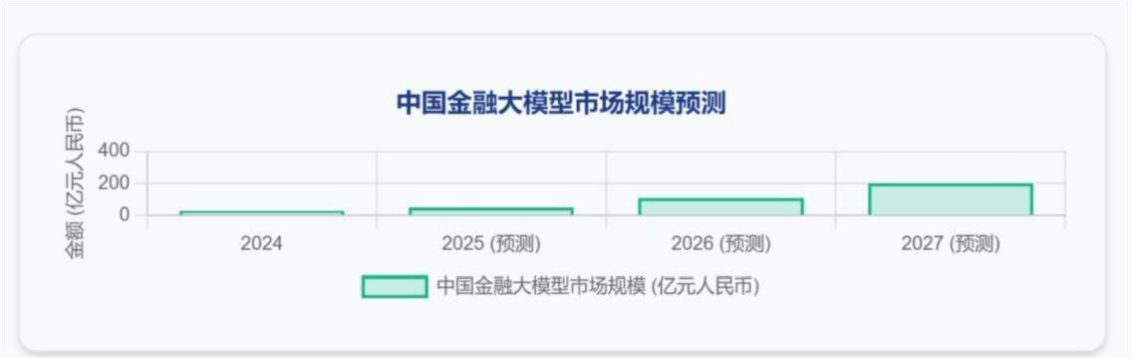
中国金融 AI 的应用广度和深度正在发生质的飞跃。在早期阶段，AI 多被用于智能客服、OCR 识别等外围、辅助性环节，扮演着“降本增效”的角色。如今，AI 正加速渗透至银行的信贷审批、证券的投资决策、保险的精算定价等核心业务领域，开始直接参与价值创造，从“辅助者”转变为“决策者”之一。根据 NVIDIA 的调查报告，超过 30%的中国金融机构已在实际业务中应用 AI 技术，尤其是在国有大行和头部股份制银行中，AI 已成为其数字化转型的核心引擎和关键战略抓手。

(三)核心技术突破与演进路径：更懂金融，更易获得

1.大模型技术演进：从“通用全才”到“金融专才”

大模型技术正沿着一条从通用到专有的路径快速演进。直接使用 GPT-4 等通用大模型虽然能力强大，但在处理高度专业化的金融任务时，常存在语义理解不深、数据时效性不足等问题。因此，业界的主流路径是“通用基座模型+行业数据微调”。通过采用更高效的微调技术，以及更先进的模型架构，金融机构能够

在通用大模型的基础上，高效、低成本地训练出更懂金融业务术语、遵循金融监管逻辑、响应更精准可靠的金融专属模型。



2. 成本效益分析：“价格战”成为 AI 普及的催化剂

技术进步和市场竞争共同带来了 AI 使用成本的急剧下降。以 OpenAI 的 GPT-4o 和 Google 的 Gemini 系列模型为例，其 API 调用价格在 2024 年均出现了“腰斩式”的大幅下调，降幅普遍在 50%以上。这意味着，过去需要高昂成本才能实现的 AI 功能，如今的开发和部署成本可能仅为原来的几分之一。成本的下降极大地降低了 AI 应用的门槛，使得更多中小金融机构能够负担得起 AI 技术，从而极大地加速了 AI 在整个金融行业的普及和深化。

3. 本地化与私有化部署：安全与合规驱动下的必然选择

鉴于金融数据的极端敏感性以及《中华人民共和国网络安全法》、《中华人民共和国数据安全法》、《中华人民共和国个人信息保护法》等法律法规的严格要求，将 AI 模型（特别是处理核心数据的大模型）部署在公有云上，面临着巨大的数据安全和合规挑战。因此，本地化与私有化部署成为中国金融机构应用 AI 的必然选择。将模型部署在机构自有的私有云或本地数据中心，可以最大限度地保证核心业务数据不出安全域，从而在源头上满足数据安全和监管合规的要求。近年来，以 DeepSeek、智谱 AI 等为代表的国产开源或商用模型的崛起，及其对本地化部署的良好支持和优化，进一步推动和加速了这一趋势。

三、金融人工智能核心应用场景与实践

金融 AI 的应用已全面渗透至金融机构的前、中、后台业务流程，形成高效协同的数据链路，正在系统性地重塑金融服务的价值链。本章节将从业务流程视角，深入剖析各环节的核心应用场景、价值创造逻辑及典型实践案例。

(一)前台业务智能化：重塑客户体验与价值创造

前台业务是金融机构与客户直接交互的界面，AI 的应用核心在于提供极致的个性化体验，从而提升客户获取、活跃度和终身价值。

1.智慧营销与个性化推荐

在流量红利见顶的时代，粗放式的营销模式已失效，精准触达是关键。

通过整合客户的人口统计学信息、交易行为、APP 使用习惯、社交数据等多方面信息，AI 可构建起全面的 360 度全景用户画像。基于这一画像，利用机器学习模型对客户在理财、贷款、信用卡等金融产品上的潜在需求和兴趣度进行预测，再选择客户发薪日、大额消费后等恰当节点，通过 APP 推送、短信、电话等数字化渠道，为客户提供贴合其需求的个性化产品与服务建议。

行业实践表明，AI 驱动的精准营销模式，在提升客户响应率、降低营销成本方面效果突出，进而推动营销转化率与客户粘性实现显著提升。以某股份制银行为例，其运用 AI 推荐模型后，理财产品交叉销售的成功情况得到了有效改善。

2.智能客服与数字人应用

客户服务是劳动密集型部门，成本高昂，且服务质量难以标准化。

基于自然语言处理（NLP）和金融大模型的智能客服，能够 7x24 小时处理大部分常见、重复性客户咨询（如账户查询、业务办理指引）。更进一步，通过多模态技术，将智能客服与虚拟数字人形象相结合，在手机银行或营业网点大屏上，提供更具亲和力、更接近真人交互的沉浸式服务体验。

极大提升服务效率，释放人力从事更复杂的客户关怀工作。几乎所有银行和券商均已部署智能客服。部分领先机构的 AI 客服表现亮眼，凭借成熟的技术支撑，已能有效解决客户的绝大多数问题，服务效能处于较高水准。

3.智能投顾与财富管理

专业的财富管理服务过去门槛极高，广大长尾客户的需求未被满足。

AI 系统首先通过线上问卷评估客户的风险偏好、财务状况和投资目标。然后，

基于现代投资组合理论，结合对宏观经济和市场走势的AI分析，为客户自动生成并动态调整一个全球化的、跨资产类别的投资组合建议。生成式AI能辅助理财经理进行高效决策与客户服务，快速总结市场动态、分析产品，辅助其为高净值客户提供更高效、更专业的服务。

降低财富管理服务的门槛，推动普惠金融。例如，摩根士丹利与OpenAI合作，为其超过16000名财富顾问配备了基于GPT-4的AI助手，赋能其从海量研究报告中快速获取洞察，更好地服务客户。

(二)中台能力中心化：构筑智慧风控与决策中枢系统

中台是金融机构的中枢系统，负责风险控制、产品定价和决策支持。AI的应用核心在于从海量数据中挖掘风险、洞察规律，提升决策的科学性与时效性。

1.智能风控与反欺诈

这是AI应用最成熟、价值最显著的领域之一。金融欺诈手段日益隐蔽化、团伙化、产业化，传统基于规则的系统防不胜防。

AI通过实时分析海量交易数据、设备指纹、用户行为、关系网络等，构建复杂的风险模型。例如，采用图神经网络（GNN）技术，可以挖掘出看似孤立账户背后的隐秘关联，精准识别洗钱、组团骗贷等复杂欺诈模式。

AI模型提升迭代速度比传统方法提50%，推升不良贷款识别准确率。在不影响用户体验的前提下，实现对信用卡盗刷、信贷欺诈、虚假交易等风险的毫秒级识别与阻断。某大型银行利用AI构建的信用卡分期智能工作流，能够对申请进行实时风险评估并自动决策，有效降低了坏账率。

2.智能信贷审批与贷后管理

传统信贷审批依赖人工，流程长、成本高，且对小微企业等客群的覆盖不足。

AI模型能够整合央行征信、工商、税务、司法以及用户的交易流水等多维度信息，对借款人进行更全面、动态的信用风险评估，实现秒级自动授信审批。在贷后管理阶段，AI能持续监控借款人的经营状况和舆情变化，动态调整风险预警，并根据风险等级进行智能化的催收提醒或人工介入。

极大提升普惠金融服务的可得性和效率，让更多小微企业和个体经营者获得信贷支持。

3.智能投研与量化交易

投资研究员和基金经理需要处理的信息量呈指数级增长，人脑处理能力已达极限。

生成式 AI 能够扮演一位初级分析师，7x24 小时监控全球新闻舆情、快速阅读并总结海量的研究报告、公司财报和业绩发布会纪要，从中自动提取关键财务指标、管理层观点和市场情绪等关键信息。AI 还可用于构建更复杂的量化交易策略，通过深度学习模型挖掘市场中非线性的、高维度的价格规律。

将投资决策从“经验驱动”转向“数据驱动”。例如，东方财富、同花顺等国内领先的金融信息服务商，均已推出基于大模型的 AI 产品，可实现财报智能分析、业绩预告解读和 AI 问答，大幅提升了投资者的投研效率。

(三)后台运营自动化：迈向极致效率与智能合规

后台是金融机构稳定运行的基石，涉及大量重复性、规则性的操作。AI 的应用核心在于实现流程自动化和智能化，最终达成“降本增效”与“合规保障”的双重目标。

1.智能运维（AIOps）与流程自动化（RPA）

金融业务高度依赖 IT 系统的稳定性，后台运营中存在大量人工操作的“断点”。AIOps 利用 AI 技术对 IT 系统的性能指标（CPU、内存、网络流量）进行实时监控和异常检测，能够提前预测潜在故障并辅助进行根因分析，保障金融业务的连续性。

RPA（机器人流程自动化）与 AI 结合，形成的“智能自动化”（IA），能够模拟人工操作，自动执行报表生成、数据录入、跨系统对账等重复性、规则性的后台任务。实现 7x24 小时不间断运营，减少人为差错，将后台员工从繁琐的重复劳动中解放出来。

2.合同/文档智能审核与分析

法务和合规人员需要审阅大量法律文件，耗时耗力且容易出错。利用 NLP 和知识图谱技术，AI 可以“阅读”并“理解”贷款合同、招股说明书、信托计划等复杂的非结构化法律文件。系统能够自动识别其中的关键条款（如担保、管辖权）、发现潜在风险（如缺失条款、矛盾表述）和不合规项，并与标准模板进

行比对，高亮差异。

将法务和合规人员的审查效率提升数倍，并提高风险识别的准确性。摩根大通开发的合同解析 AI 工具，能在几秒钟内完成过去律师需要 36 万小时才能完成的商业贷款协议审查工作。

3.智能合规与监管科技（RegTech）

金融领域监管要求正呈现日趋严格复杂的特征，推动机构合规运营成本持续攀升。AI 能够帮助金融机构自动监控海量交易行为，以识别是否符合反洗钱（AML）、了解你的客户（KYC）等监管规定。例如，通过分析资金流转网络，发现异常的、不符合客户身份的交易模式。AI 还能自动追踪全球监管政策的更新，并将其转化为内部可执行的合规规则，自动生成满足监管要求的报告。

在降低合规成本的同时，提升合规的准确性、覆盖面和时效性，助力金融机构更高效地应对各类监管压力。

四、金融人工智能的安全风险与严峻挑战

金融 AI 在创造巨大价值的同时，也如一枚硬币的两面，不可避免地引入了前所未有的、系统性的安全风险。这些风险不再局限于传统的网络攻防范畴，而是贯穿 AI 系统的整个生命周期，从数据源头到模型核心，再到应用终端，对金融机构的风险管理能力提出了前所未有的严峻挑战。从产业安全视角看，金融 AI 正面临一个由技术、数据、业务、管理共同交织的复杂风险矩阵。

(一)技术脆弱性：模型与算法自身成为新的“攻击平面”

在 AI 时代，算法和模型本身从分析工具，变成了可被直接攻击和操纵的对象。这种针对 AI 核心的攻击，更为隐蔽，后果也更为严重。

1.模型安全风险：看不见的“安全隐患”

AI 模型本身可能成为新的攻击平面，其面临的威胁迥异于传统软件漏洞。主要威胁包括：

（1）对抗性攻击（Adversarial Attacks）：这是对 AI 模型最典型也最具威胁的攻击之一。攻击者通过对输入数据（如一张用于身份验证的人脸图片，或一笔交易的文本描述）添加人眼或常规检测手段难以察觉的微小扰动（即“对抗性噪声”），就能诱导模型做出完全错误的分类或判断。例如，在信贷审批场景，攻击者可能通过微调申请材料中的某些字符，使得一个高风险的贷款申请被 AI 模型误判为优质客户，从而骗取贷款。这种攻击的危害在于，模型在绝大多数正常情况下表现优异，在面对精心构造的对抗样本时，给信贷业务埋下风险隐患。

（2）后门攻击（Backdoor Attacks）与数据源头性威胁：在模型训练阶段，存在一种需要重点防范的情况。在此过程中，若有人在庞大的训练数据集中暗中植入少量经特殊设计的样本，可能导致模型被预设特定“后门”。这种带有特殊设定的模型，在常规测试与日常使用中，表现与未受影响的正常模型并无差异；但当遇到包含特定触发器的输入时，就会按照预设逻辑执行特定操作，例如将所有包含该触发器的交易判定为合法交易。

（3）模型窃取与隐私泄露：AI 模型，特别是作为核心资产的商业模型，也面临被非法获取的风险。攻击者并不需要访问模型的源代码，仅通过大量、反复地查询模型的 API 接口，观察输入与输出之间的关系，就能够利用模型提取、逆向工程等技术，近似地复制出一个功能相似的替代模型。更严重的是，成员推断

攻击甚至可以通过查询，判断出某一个具体的用户数据（例如某位特定客户的交易记录）是否曾被用于训练该模型，从而造成严重的数据隐私泄露。

2.算法“黑箱”：可解释性缺失引发的信任危机与决策风险

许多性能卓越的AI模型，尤其是结构极其复杂的深度学习模型，其内部决策过程对人类而言是不透明的，形成了所谓的“黑箱”。我们知道它给出了某个决策，但很难确切地知道它是基于哪些特征、通过何种逻辑得出这个结论的。

这种可解释性的缺失，带来了两大核心问题：其一，信任危机。当客户、合作伙伴甚至监管机构询问决策依据时，“模型如此判断” 这类模糊回应完全无法令人接受，这一问题已成为AI在关键决策领域落地应用的重要阻碍。其二，风险排查困难。当AI决策失误并造成损失时，由于无法理解其内在逻辑，我们很难进行有效的归因、修复和迭代优化，使得风险排查和模型纠错变得异常困难。

3.生成式AI特有风险：强大能力伴生的新型威胁

以金融大模型为代表的生成式AI，在展现惊人能力的同时，也带来了全新的、不容忽视的风险：

（1）提示注入与越狱：这是针对大语言模型最主要的攻击方式之一。攻击者通过构造恶意的、欺骗性的提示词（Prompt），可以绕过开发者设置的安全护栏，诱导模型忽略其原始指令，转而执行攻击者下达的恶意指令。

（2）有害与虚假内容生成：如果缺乏有效的安全过滤机制，大模型可能被滥用或被诱导生成涉及歧视、暴力、仇恨言论的内容，或用于生成高度逼真的金融钓鱼邮件、诈骗短信，对金融机构的声誉和社会稳定造成负面影响。

（3）模型“幻觉”： 这是生成式AI的一个内在缺陷。当模型在其知识范围之外或数据不充分的情况下被要求进行推理时，它可能会编造出看似合理但与事实完全不符的信息。在智能投研场景，如果模型“幻觉”出一家公司不存在的巨额盈利或虚假的并购新闻，并以此为依据生成投资建议，可能对投资者造成灾难性的误导。

(二)数据安全与隐私保护：人工智能的基础要素，安全的核心环节

数据是驱动AI模型的基础要素，其安全与质量直接决定了上层AI应用的成败。在金融领域，数据安全更是关乎机构生命线和客户信任的基石。

1.训练数据污染：源头上的威胁

数据的质量和安全是AI模型的根基。如前所述，攻击者如果在模型训练数据中注入少量精心构造的恶意样本，就可能从源头上破坏整个模型的性能，或为其植入难以察觉的“后门”。在金融领域，训练数据的来源复杂，可能包括第三方数据提供商，这无疑增加了数据被污染的风险。

2.用户隐私数据在全生命周期中的泄露风险

金融AI应用不可避免地需要处理海量的个人敏感信息，包括身份信息、财产信息、交易信息、生物识别信息等。从数据采集环节的授权是否充分合规，到数据传输和存储过程中的加密与访问控制是否到位，再到模型训练和推理使用中如何防止隐私信息被无意识地“记忆”和泄露，任何一个环节的疏忽都可能导致大规模的数据泄露事件，引发严重的法律责任、监管处罚和声誉危机。

3.数据跨境流动的合规性难题

对于在中国运营的跨国金融机构，或使用了海外云服务、开源模型、第三方数据接口的本土金融机构而言，训练数据、模型参数乃至用户请求的跨境流动，都必须严格遵守《中华人民共和国网络安全法》、《中华人民共和国数据安全法》、《中华人民共和国个人信息保护法》以及相关的数据出境安全评估办法。如何在利用全球先进AI技术的同时，确保数据流动的合规性，已成为一个极其复杂且具有挑战性的法律与技术问题。

(三)业务应用风险：决策偏见与系统性隐患

当AI模型从后台走向前台，其决策直接影响业务结果，新的业务风险也随之产生。

1.算法偏见与歧视：技术的“无心之失”可能固化社会不公

人工智能模型本身并不具备价值观，其仅从数据之中学习规律。如果用于训练模型的数据本身就包含了现实世界中存在的、历史形成的偏见，那么AI模型在学习后会忠实地、甚至放大这些偏见。这可能导致在信贷审批、保险定价、招聘筛选等场景中，AI系统做出对特定地域、性别、种族或社会阶层的歧视性决策。这不仅严重违背了金融服务应有的公平性原则，也可能为金融机构招致严厉的监管处罚和旷日持久的法律诉讼。

2.风险集中度与系统性风险：“模型趋同”的潜在危机

随着AI应用的普及，一个值得高度警惕的新现象是“模型趋同”。如果多家关键的金融机构过度依赖于同一个或少数几个外部供应商提供的AI模型或技术平台（例如，使用同一家公司的风控模型），那么该平台的任何技术漏洞、模型偏差或安全事件,其风险都可能通过这些机构迅速传导、放大至整个金融体系，形成一种由算法驱动的新型系统性风险。在市场出现极端行情时，如果这些趋同的模型做出相似的、程序化的反应（例如，同时抛售某一类资产），可能会加剧市场波动，造成流动性危机。

(四)合规与管理风险：责任的模糊地带

1.责任边界模糊：当AI犯错时，谁来负责？

当一个由AI系统自主做出的决策导致了客户的重大损失或引发了市场混乱时，责任归属成为一个极其复杂的法律与伦理难题。责任应该由AI算法的设计者、模型的训练者、数据的提供方、使用该AI系统的金融机构，还是批准其上线的管理人员来承担？在现有的法律框架下,这一责任归属尚未形成清晰统一的界定，直接导致风险最终承担主体模糊不清。

2.现有法律法规的滞后性：创新在“无人区”探索

技术发展的速度往往领先于法律法规制定与完善的速度。尽管各国和地区都在加紧制定AI相关的监管规则，但当前针对金融AI的专门性、系统性法律法规仍然不足。许多金融机构在探索和应用AI时，感觉像是在一个规则尚不明确的“无人区”中行进，既渴望创新，又需要警惕触碰未知的合规红线。这种合规的不确定性，在一定程度上也制约了更大胆、更具突破性的创新步伐。

五、金融人工智能安全治理与防护体系构建

面对金融 AI 带来的复杂、多维的风险挑战，被动、零散的修补式防护已然失效。必须采取系统性的、前瞻性的应对策略。我们主张，金融机构必须构建一个“治理”与“技术”双轮驱动、深度融合的安全防护体系，将安全的基因深度融入 AI 战略规划、组织架构、技术研发和业务运营的全过程，实现从被动响应到主动防御的根本性转变。

(一)顶层设计：构建可信人工智能治理框架

技术防护需要正确的方向指引，而这个方向就来自于顶层的治理框架。我们建议建立一个健全的 AI 治理框架，是确保 AI 应用始终朝着负责任、可持续方向发展的“压舱石”。

1.确立治理原则：可信人工智能的六大基石

金融机构应在董事会和高管层面，正式确立并推行符合国际共识和中国国情的可信 AI 核心原则，并将其作为 AI 战略的根本遵循。这六大原则是：

（1）公平（Fairness）：致力于在 AI 系统的设计、开发和部署中，识别、评估和消减不公平的偏见，确保 AI 决策不会对任何个人或群体产生歧视。

（2）透明（Transparency）：努力提升 AI 决策过程的透明度和可解释性，使其决策逻辑能够被相关方（如客户、员工、监管机构）所理解、审查和质询。

（3）问责（Accountability）：建立清晰的内部问责机制，明确 AI 系统的开发、部署、运营等各个环节的责任主体，确保当 AI 系统出现问题时，能够进行有效的追溯和问责。

（4）安全（Security）：采取强有力的技术和管理措施，保护 AI 系统及其处理的数据免受恶意攻击和非授权访问，确保系统的保密性、完整性和可用性。

（5）可靠（Reliability）：确保 AI 系统在各种正常、异常甚至恶意攻击的环境下，都能保持稳定、可靠的性能表现，避免因外部扰动而产生灾难性故障。

(6)隐私保护(Privacy):将隐私保护作为 AI 系统设计的核心要素(Privacy by Design），在数据全生命周期中严格遵守相关法律法规，充分保护个人信息主体的合法权益。

2.完善组织架构：从“三道防线”到人工智能治理委员会

有效的治理需要有力的组织保障。我们建议金融机构应设立一个由高层管理

人员（如首席风险官、首席信息官或首席科技官）牵头的跨部门的“AI 伦理与治理委员会”。该委员会成员应包括来自业务、技术、风险管理、合规、法律、内审等核心部门的专家，共同负责制定 AI 战略、审查高风险 AI 应用、监督治理原则的落地。同时，需要将 AI 风险管理明确地嵌入到传统的业务、风控、内审“三道防线”中，明确每一道防线在 AI 风险识别、评估、监控和审计中的具体职责，从而形成一个权责清晰、协同运作的治理体系。

3.实施风险评估与分级：将有限的资源投入到最关键的领域

并非所有 AI 应用都带来同等级别的风险。借鉴欧盟《人工智能法案》的思路，金融机构应建立一套动态的、与自身业务相结合的人工智能风险评估与分级机制。根据 AI 应用对客户核心权益、机构财务稳健和金融市场稳定的潜在影响程度，将其划分为不可接受风险（如利用潜意识技术操纵用户）、高风险（如信贷审批、保险定价、关键岗位招聘）、有限风险（如智能客服）和低风险（如后台流程自动化）等不同等级，并针对不同等级的 AI 应用，实施差异化的管理要求、测试标准和审批流程。这有助于将最严格的治理和技术资源，集中投入到风险最高的应用上。

(二)技术防护体系：从开发到运营的全生命周期安全

为筑牢金融人工智能应用的技术安全屏障，我们建议构建一套覆盖人工智能开发、模型、运行三大阶段的全生命周期技术防护体系：

1.开发安全（Development Security）

（1）数据安全与质量保证：在数据准备阶段，建立严格的数据来源审查和质量评估机制。对敏感数据进行脱敏、匿名化处理，并对训练数据集进行扫描，以发现潜在的数据偏见或误导迹象。

（2）源代码安全：在 AI 应用及算法模型的编码阶段，需提前嵌入安全防控措施。可借助静态应用安全测试（SAST）工具开展代码安全审计，从源头阻断注入、越权等传统软件漏洞的产生。

（3）安全组件管理：现代 AI 开发大量依赖第三方库（如 TensorFlow）和开源框架（如 PyTorch）。必须建立严格的软件成分分析（SCA）机制，对开发过程中引入的所有第三方组件进行持续的安全审查和漏洞扫描，防止因上游组件

的漏洞而引发供应链攻击。

2.模型安全（Model Security）

（1）模型鲁棒性测试与加固：在模型上线前进行系统的、严格的压力测试。这包括利用自动化工具生成对抗性样本，测试模型在攻击下的表现；模拟各种异常输入，评估模型的稳定性。基于测试结果，可以采用对抗性训练等技术对模型进行加固，提升其“免疫力”。

（2）模型可解释性增强：采用 LIME、SHAP 等主流的可解释性 AI 技术，对模型的决策进行分析，为关键决策提供归因分析报告。这不仅满足合规要求，也有助于发现模型内在的逻辑缺陷。

（3）隐私增强技术（PETs）：在模型训练和发布时，应积极探索和应用差分隐私、同态加密等隐私增强技术。差分隐私通过在计算过程中添加可控的噪声，使得即使模型被攻破，也无法逆向推断出任何单个用户的具体信息，从而在数学上为隐私保护提供了严格的保障。

3.运行安全（Secure Operation）

（1）零信任访问控制：AI 模型和应用一旦上线，其 API 接口就成为核心的暴露面。必须摒弃传统基于边界的信任模型，全面转向“零信任（Zero Trust）”安全架构。其核心原则是“永不信任，始终验证”，对所有访问 AI 模型 API 的主体（无论是用户、设备还是其他应用）进行严格的身份验证和动态授权，确保每一次调用都是经过授权和可信的，从而有效防止未经授权的访问和滥用。

（2）API 安全防护：在 AI 应用前部署专业的 API 安全网关，对所有流入模型的 API 流量进行深度检测、实时监控和访问控制。这能够有效防范针对大模型的提示注入、拒绝服务（DoS）以及其他利用 API 漏洞的攻击。

（3）模型运行状态持续监控：建立对线上模型性能的持续监控机制，不仅监控其业务指标（如准确率），更要监控其安全指标（如输入数据的分布是否发生漂移、是否出现大量异常预测结果等），以便及时发现模型性能衰退或遭受未知攻击的迹象。

(三)数据安全纵深防御：保护人工智能时代的核心资产

1.数据分类分级与全生命周期加密

数据分类分级与全生命周期加密是数据安全的基石。金融机构必须建立一套清晰、完善的数据分类分级制度，根据数据的敏感性和重要性（如公开数据、内部数据、敏感数据、核心数据），实施差异化的保护策略。在此基础上，对静态存储（在数据库、文件系统中）和动态传输（在网络中）的数据进行全程、高强度的加密，确保数据在任何状态下的机密性。

2.隐私增强技术（PETs）的战略性应用

面对日益突出的“数据孤岛”和隐私保护的矛盾，金融机构应将隐私增强技术（PETs）作为战略性技术进行布局和应用。其中，联邦学习最具代表性。它允许多个金融机构在不共享各自原始客户数据的情况下，联合训练一个性能更优的AI 模型，从而在合规的前提下，安全地汇聚多方数据智慧，破解数据协作与隐私保护的两难困境。

3.数据防泄漏（DLP）与数据流动监控

金融机构应部署先进的数据防泄漏（DLP）系统，对机构网络出口、终端设备、电子邮件等所有可能的泄露渠道进行内容级的监控。当发现有敏感数据（如客户名单、模型参数）尝试通过非授权渠道外发时，系统能进行实时告警和阻断。同时，结合数据水印等技术，对数据的流动轨迹进行全链路追踪，确保任何数据的使用行为都是可审计、可追溯的。

(四)人工智能赋能安全运营：以“智能”应对“智能威胁”

未来的网络安全对抗，必然是“智能”与“智能”的对抗。面对由 AI 驱动的、自动化、高效率的攻击手段，传统依赖人力和静态规则的安全运营模式已捉襟见肘。必须利用 AI 技术来赋能安全运营，提升威胁发现、分析和响应的效率与能力。

1.智能威胁检测与响应

传统的基于病毒特征库和已知攻击规则的检测手段，对于未知的、高级的、经过伪装的威胁（如 APT 攻击）几乎无能为力。利用大数据平台和 AI 引擎，可以从金融机构全网海量的日志、网络流量和终端行为数据中，通过无监督学习进行异常行为分析，从而精准地发现那些偏离正常基线的、微小的、可疑的活动痕迹，进而发现高级威胁。

2.AI 驱动的安全态势感知

构建一个金融 AI 应用的全景安全视图，将来自网络设备、安全产品、服务器、终端、业务应用、AI 模型等各个层面的海量、异构的安全数据进行汇聚、清洗和关联分析。利用 AI 技术自动识别出真正有价值的威胁告警，并将复杂的攻击链条以可视化的方式呈现出来，帮助安全团队实现对整体安全态势的实时感知、风险预警和快速决策。

3.安全编排自动化与响应（SOAR）

通过预先设定的剧本（Playbook），将大量重复性的安全告警分析、威胁研判、调查取证和应急处置等流程，由 SOAR 平台自动化、协同化地完成。例如，当 EDR 发现一个可疑进程时，SOAR 可以自动触发防火墙阻断其外部通信，同时从威胁情报平台查询其信誉，并自动创建工单给安全分析师。这能够实现对安全事件的秒级响应，将安全运营团队从海量的告警和重复性工作中解放出来，专注于处理更高级的威胁。

六、政策法规、伦理规范与国际发展

有效的治理离不开清晰的规则。技术的发展必须在规范的轨道上运行，才能行稳致远。随着金融 AI 应用的日益深化和普及，其潜在的巨大影响已引发全球范围内的监管机构、国际组织和社会各界的高度关注。一个多层次、动态的、旨在平衡创新与风险的治理框架正在全球范围内加速形成。

(一)全球主要经济体监管框架对比分析：趋同与分歧

尽管目标都是促进负责任的 AI 创新，但全球主要经济体在监管路径和哲学上展现出明显的分野，形成了以欧盟、美国、中国、英国为代表的几种典型模式。

1.欧盟《人工智能法案》：基于风险分级的“硬法”强监管模式

欧盟无疑走在了全球 AI 监管的最前沿。其于 2024 年正式生效的《人工智能法案》（AI Act）是全球首部具有横向法律约束力的、全面的 AI 法规。该法案的核心是采取了基于风险的分级方法，将 AI 应用划分为四个等级：

（1）不可接受风险：如利用潜意识技术操纵个人行为、社会评分系统等，被完全禁止。

（2）高风险：这是法案监管的重点。金融领域的信贷评分、信用评估、保险风险评估与定价等被明确列为高风险应用。这类应用在进入市场前，必须满足一系列严格的合规要求，包括高质量的数据治理、详细的技术文档、高水平的透明度和可解释性、强有力的人类监督、以及高度的网络安全和稳健性。

（3）有限风险：如与人交互的聊天机器人（Chatbot），需履行透明度义务，明确告知用户正在与 AI 系统互动。

（4）最低或无风险：大多数 AI 应用属于此列，不受额外法规约束。

欧盟的模式是一种典型的自上而下的“硬法”监管，强调事前规制和对基本权利的保护，旨在为 AI 的发展建立一个可信赖的法律环境。

2.美国：自愿原则与行政命令并行的“软硬兼施”引导式监管

与欧盟不同，美国采取了更为灵活和鼓励创新的监管方式，避免过早地用统一的、僵化的法律来束缚技术发展。其监管路径呈现出“自愿指引先行，行政命令跟进，垂直领域立法”的特点。早期，白宫发布的《人工智能权利法案蓝图》、国家标准与技术研究院（NIST）发布的《人工智能风险管理框架》（AI RMF 1.0）等，都是不具强制性的自愿性指南，重在引导行业自律。近期，拜登政府签署的

关于 AI 安全的行政命令，则开始加强对 AI 安全测试、标准制定和政府采购的引导。其监管特色是联邦与州级立法协同，更侧重于在具体领域（如消费者金融保护局 CFPB 关注算法偏见）解决实际问题，而非欧盟式的全面市场准入监管。

3.英国：以创新为导向的“上下文相关”治理模式

英国模式被称为“亲创新”的“软法”治理。它没有制定一部统一的、横向的 AI 法案，而是选择了一条“上下文相关”（Context-specific）的路径。由各行业现有的监管机构（如金融行为监管局 FCA、信息专员办公室 ICO）在其各自的职责范围内，应用现有法律框架并发布新的指导原则，来管理本领域的 AI 风险。这种模式的优点是灵活性强，能够更好地适应不同行业的特殊性，其核心目标是在不扼杀创新的前提下，建立一个敏捷、适应性强的监管环境。

4.国际协调：在分歧中寻求共识

面对 AI 这一全球性技术，国际合作与协调至关重要。G7 集团发布的《广岛 AI 进程》、在英国召开的全球 AI 安全峰会通过的《布莱切利宣言》、以及联合国、OECD 等国际组织的工作，都在积极推动形成关于 AI 风险、透明度和治理的国际共识，努力避免全球 AI 治理的碎片化。

(二)中国金融人工智能监管政策演进与前瞻

中国对 AI 的治理始终坚持“发展与安全并重、鼓励创新与依法规范相结合”的原则，政策体系正在从宏观到微观、从通用到专用，逐步细化和完善。

1.政策演进脉络：从战略规划到落地规范

中国的 AI 治理政策体系可以梳理为一条清晰的演进脉络。始于 2017 年国务院《新一代人工智能发展规划》的顶层宏观布局；其后，2019 年国家新一代人工智能治理专业委员会发布的《新一代人工智能治理原则》和 2021 年发布的《新一代人工智能伦理规范》，确立了 AI 发展的伦理基调和基本准则；再到近年来，针对算法推荐、深度合成、生成式 AI 等具体技术和应用，密集出台了一系列管理规定，标志着监管进入深水区。

2.关键金融监管文件解读：从“原则”到“标准”

在金融领域，中国人民银行发布的一系列文件具有里程碑意义。特别是《人工智能算法金融应用评价规范》（JR/T0221—2021），它首次从国家金融行业标

准的层面，对 AI 算法的安全性、可解释性、精准度和性能等四大方面，提出了具体的、可操作的评价指标和要求。例如，在可解释性方面，要求对信贷审批等影响重大的决策结果提供解释。这份文件标志着中国的金融 AI 监管，已经从宏观的原则性指导，迈向了可量化、可操作、可评估的标准化管理新阶段。此外，《生成式人工智能服务管理暂行办法》也为金融机构应用大模型提供了基础的合规指引。

3.未来监管趋势预判：迈向系统化、动态化、协同化

展望未来，我们预判中国金融 AI 监管将呈现三大核心趋势：

（1）系统化与专门化：随着应用的深化，零散的、针对特定技术的“打补丁式”规定，将逐步整合。我们强烈建议并预判，在总结现有规范和试点经验的基础上，由国家金融监督管理总局、中国人民银行等多部门联合，加快研究并出台一部全国统一的、更高层级的《金融人工智能管理办法》，系统性地规范 AI 在金融领域的准入、应用、风控、外包、问责等全链条活动。

（2）敏捷化与动态化：监管将从事前的一次性备案，更多地转向事中、事后的持续性风险监测与动态评估。灵活运用“监管沙盒”、“合规科技”等敏捷治理工具，为负责任的 AI 创新提供一个风险可控的“试验田”，在严守安全底线的前提下，为新技术、新模式的探索留出空间。

（3）协同化与生态化：未来的监管将不再是监管机构的“独角戏”。监管机构、产业界（金融机构与科技公司）、学术界和第三方评测机构之间的合作将更加紧密，共同推动技术标准、测评体系、最佳实践和风险案例库的形成，构建一个多方参与、协同共治的监管生态。

(三)人工智能伦理原则与社会责任：技术向善的罗盘

法律法规是底线，而伦理道德是更高的追求。金融 AI 的发展不仅是技术问题，更是深刻的社会问题。金融机构作为技术的应用者，必须肩负起相应的社会责任，确保技术始终是“向善”的力量。

1.“以人为本”原则的落地路径：确保人的最终控制权

“以人为本”是所有人工智能发展的最高原则。在金融领域，这一原则的落地意味着，无论 AI 系统多么智能和自动化，其应用都必须始终服务于增进人类

福祉和金融消费者的利益。金融机构在追求效率和利润的同时，必须在系统设计中建立起有意义的人工干预和监督机制（Human-in-the-loop），对于信贷拒绝、保险拒赔等重大决策，必须保障用户的知情权、解释请求权和申诉权，并提供有效的人工复核和救济渠道，确保技术不会成为无法挑战的“数字权威”。

2.促进金融普惠与防止“数字鸿沟”：AI 的双重效应

AI 技术无疑为金融普惠带来了巨大机遇。它能够以更低的成本、更广的覆盖面，为小微企业、农民、个体工商户等传统金融服务难以触及的群体，提供信贷、理财和保险服务。但与此同时，我们也必须警惕其潜在的负面效应。如果算法设计不当或训练数据存在偏差，AI 可能会加剧对老年人、低收入群体、教育程度较低者等弱势群体的排斥，使他们更难获得金融服务，从而形成或加深新的“数字鸿沟”。金融机构有责任在 AI 系统设计中，充分考虑并测试其对不同群体的包容性，确保技术的普惠价值真正得以实现。

3.构建负责任的金融 AI 生态：超越单一机构的努力

一个健康、可持续的金融 AI 生态，需要所有参与方的共同努力和责任担当。金融机构应将社会责任内化为企业文化和行为准则。科技公司在追求技术突破的同时，应主动思考和减轻其技术可能带来的社会风险。学术界应加强对 AI 伦理和社会影响的跨学科研究。而公众和媒体也应加强对 AI 的理解和监督。只有通过开放的对话、紧密的协作和共同的责任感，我们才能共同构建一个值得信赖的金融 AI 未来，确保金融 AI 的发展行稳致远，真正造福于整个社会。

七、结论与建议

历经数十年发展，人工智能正以前所未有的力量，开启金融业的深刻变革。本白皮书通过融合学术界的前瞻洞察与产业界的一线实践，力图为这一波澜壮阔的变革提供一个清晰的认知框架和行动指南。站在 2025 年的关键节点，我们得出以下核心结论，并提出具体行动建议。

(一)主要结论总结

1.范式转移：金融 AI 已进入规模化应用与价值兑现的关键时期。以大模型为代表的技术突破与 API 成本的持续下降，正在共同推动 AI 从过去的辅助工具向未来的核心生产力转变。它不再是 IT 部门的探索，而是关乎业务增长、客户体验和竞争优势的“一把手工程”，正在系统性地重塑金融业的竞争格局与价值链。

2.安全重构：安全不再是 AI 发展的制动器，而是高质量发展的核心竞争力与加速器。伴随应用的深化，由模型脆弱性、数据隐私、算法偏见和责任模糊等交织而成的新型安全风险日益显现。我们必须深刻认识到，一个安全、可信、稳健的 AI 系统是其商业价值得以实现和持续的根本前提。AI 安全能力，正迅速从成本中心，转变为金融机构赢得客户信任、防范重大风险、构筑护城河的核心竞争力。

3.应对之道：“治理+技术”双轮驱动是应对金融 AI 风险的核心路径。面对复杂的 AI 风险矩阵，单纯依靠技术防护或仅凭管理制度都无法完全应对。金融机构必须建立一个从董事会到执行层、自上而下的 AI 治理框架，并将其与一个覆盖数据、模型、应用全生命周期的、以“零信任”和“安全左移”为代表的新一代技术防护体系紧密结合，形成系统性、全局性的解决方案。

4.生态共建：金融 AI 的未来在于开放、协同与共赢的生态系统。没有任何一个单一实体能够独立解决金融 AI 发展中的所有问题。未来的创新和安全保障，将越来越依赖于监管机构、金融机构、科技公司、学术界和第三方服务机构之间的高效协同与良性互动。构建一个开放、合作、共担责任的生态系统，将是未来竞争与发展的主要源泉。

(二)对金融机构的行动建议

1.战略层面：将AI安全与治理纳入企业顶层设计。应由董事会或战略委员会牵头，制定清晰的、与机构价值观和长期目标相一致的AI发展战略、伦理准则和风险偏好。AI战略必须与业务发展战略同规划、同部署、同考核，确保其成为企业发展的内在驱动力，而非孤立的技术项目。

2.组织层面：建立权责清晰的治理架构与专业化的人才队伍。建议设立由高管层领导、跨部门协作的AI风险管理委员会，并明确AI风险在业务、风控、内审“三道防线”中的具体职责。同时，加大投入，通过内部培养和外部引进相结合的方式，着力建立一支深刻理解金融业务、精通AI技术、并具备安全思维的复合型AI人才队伍。

3.技术层面：构建主动、智能、纵深防御的技术防护体系。投入资源，构建覆盖AI数据、模型、应用全生命周期的AI-DevSecOps安全体系。积极评估和采用零信任架构、隐私增强计算（如联邦学习）、模型鲁棒性测试、AI赋能安全运营（AIOps/SOAR）等新技术，打造一个能够动态适应威胁、主动发现风险的智能防御能力。

4.流程层面：将AI风险评估嵌入关键业务流程。建立标准化的AI应用上线前审查和风险评估流程。对于高风险AI应用，应进行独立的、严格的技术和伦理审查。同时，建立对线上运行AI模型的持续监控、定期重估和应急处置机制，确保风险可控。

(三)对监管部门的政策建议

1.加快立法进程，提供稳定预期：建议在总结现有规范和试点经验的基础上，加快研究并出台全国统一的《金融人工智能管理办法》，为行业发展提供清晰、稳定、可预期的顶层制度环境，明确各方权利、义务与责任。

2.拥抱敏捷治理，鼓励负责任创新：灵活运用“监管沙盒”、“试点计划”、“技术白名单”等敏捷治理工具，为负责任的AI创新提供一个风险可控的测试环境，在严守安全底线的前提下，鼓励金融机构和科技公司对新技术、新模式进行探索，实现“监管”与“发展”的良性互动。

3.推动行业协同，共建安全生态：牵头或支持行业协会、科研机构，推动建立行业级的AI风险案例库、安全威胁情报共享平台、以及权威的第三方AI算法

测评认证机构。通过制定统一的技术标准、推广最佳实践、共享风险信息，提升整个金融行业应对 AI 风险的“集体免疫力”。

参考文献

1. 毕马威金融行业研究中心. 人工智能（AI）国际金融监管初探[R]. 北京: 毕马威, 2024.
2. NVIDIA. 2024 年金融服务业中国 AI 现状与趋势调查报告[R]. 美国: NVIDIA Corporation, 2024.
3. 国信证券. 人工智能专题：行业 AI 落地在即，金融领域快速渗透[R]. 深圳: 国信证券研究所, 2025.
4. 中国信息通信研究院. 金融人工智能研究报告[R]. 北京: 中国信息通信研究院, 2022.
5. 中伦律师事务所. 2023 数据合规与人工智能监管的回顾与展望[EB/OL]. (2024-01-23) [2025-10-13]. https://www.zhonglun.com/research/articles/52552.html.
6. 南方都市报. 监管就 AI 风险发声！专家建议尽快出台金融人工智能管理办法[N]. 南方都市报, 2025-01-01).
7. 普华永道中国. 构建信任，谨控风险——人工智能时代[R]. 北京: 普华永道, 2023.
8. 中国人民银行. 人工智能算法金融应用评价规范: JR/T 0221—2021[S]. 北京: 中国人民银行, 2021.
9. 欧洲议会与欧盟理事会. 欧盟人工智能法案: Regulation (EU) 2024/1689[Z]. 布鲁塞尔: 欧盟官方公报, 2024.
10. 美国国家标准与技术研究院. 人工智能风险管理框架: AI RMF 1.0[S]. 美国: NIST, 2023.

附录

对抗性攻击(Adversarial Attack)

一种针对机器学习模型的攻击技术，通过向输入数据添加微小的、人难以察觉的扰动，使得模型做出错误的输出。例如，微调图片像素让模型将“猫”识别为“狗”。

后门攻击(Backdoor Attack)

在模型训练阶段，通过向训练数据中注入带有特定触发器（Trigger）的恶意样本，在模型中植入“后门”。正常情况下模型表现正常，一旦输入包含触发器，就会执行恶意任务。

提示注入(Prompt Injection)

针对大语言模型的一种攻击，通过构造特殊的、欺骗性的输入文本（提示词），绕过模型的安全限制，使其执行攻击者意图的指令，而非其预设任务。

模型幻觉(Hallucination)

指生成式 AI 模型（特别是大语言模型）产生看似合理但实际上是虚假的、与事实不符或与输入源不一致的输出的现象。俗称“一本正经地胡说八道”。

联邦学习(Federated Learning)

一种分布式机器学习技术，允许多个参与方在不共享各自原始数据的情况下，联合训练一个模型。其核心思想是“数据不动模型动，数据可用不可见”，在保护数据隐私的同时实现知识共享。

差分隐私(Differential Privacy)

一种提供数据隐私保护的数学框架。它通过在算法输出中添加受控的随机噪声，使得对数据库的任何单次查询结果，都不会暴露数据库中任何单个个体的信息。

可信 AI(Trustworthy AI)

一个综合性框架，旨在确保 AI 系统的开发和运作方式是安全的、符合伦理的、并值得社会信赖的。通常包括公平、透明、可解释、稳健、安全、问责、隐私保护等多个维度。

零信任(Zero Trust)

一种网络安全模型，其核心思想是不自动信任网络内部或外部的任何人/设备/系统，而是对任何访问请求都进行严格的身份验证、授权和加密，遵循“永不信任，始终验证”的原则。

金融信创(Financial Innovation in Information Technology Application) 指在金融行业推动信息技术应用的创新，特别是推动核心技术、基础软硬件的安全、自主、可控，实现国产化替代。