KPMG
毕马威

于混沌处，求可控之道

银行业生成式人工智能模型验证

白皮书

kpmg.com/cn

简介

摘要

越来越多的银行正在部署生成式人工智能，将其用于汇总信息、辅助专家判断和自动执行知识密集型任务。虽然这项新技术能够显著提高生产力，但也从根本上改变了模型风险的性质。与产生确定性数值输出的传统模型不同，生成式人工智能系统会基于概率输出看似可信的描述、建议和解释 ，即使这些信息并不完整、带有偏见或有违事实。

对于**董事会和高级管理层**而言，这种转变带来了全新的问责挑战：生成式人工智能的输出可能会影响客户交互、合规决策和管理报告，且其运作方式无法完美适配现有的风控体系。

生成式人工智能要求**模型所有者和开发者**重新审视系统的设计、文档记录和测试流程，因为其风险不仅来自数据和算法，还来自提示词、检索机制和业务运行场景。

这个挑战在**模型验证和模型风险管理（MRM）职能中**最为严峻。目前的模型验证框架是为输出明确、可解释性强的模型设计的。相比之下，生成式人工智能系统具有动态性，缺乏透明度，并且经常依赖于第三方供应商。因此，验证工作的重心将会从核验数学层面的正确性，转向**持续确认模型**的行为是否始终符合银行的风险偏好。

本文为银行业生成式人工智能模型的验证提供了一套结构化、可落地的实操方法。具体包括：

识别形成生成式人工智能模型验证监管原则的合规与伦理驱动因素；

阐释为何传统验证技术不适用于生成式人工智能模型的验证；

提出了毕马威对生成式人工智能模型验证的框架，该框架以现有模型验证实践为基础进行扩展，充分考虑生成式人工智能自身的特定风险，如幻觉问题、提示敏感性、以及第三方依赖性和不当使用等风险。

我们的目的并不是阻碍创新，而是通过将稳健的验证机制作为生成式人工智能治理的核心要素，使银行能够负责任地应用这一技术。

背景与动因

1

生成式人工智能代表了人工智能的领域一次范式演变，它能够根据从海量数据集中习得的特征自动生成文本、图像、代码和其他复杂内容。这一演变给金融机构的模型风险、治理需求和验证要求带来了全新维度的挑战。

大语言模型（LLM）：以业界顶尖的自研大语言模型为代表，这些模型基于大量文本语料训练而成，能够像人类一样流畅地处理、总结和生成自然语言信息。

小语言模型（SLM）：这类模型是轻量级、面相特定领域的专有模型，针对效率、隐私保护和本地部署进行了优化，因此在强监管行业的重要性日益显现。

生成式人工智能包含多种模型架构，目前对于银行业应用最广的几类架构包括：

代理式人工智能和自主系统：除了简单生成内容之外，这类“智能体”还能够自主执行多步骤工作流程，例如处理争议或执行纠正任务，使生成式人工智能不再只是被动开展分析，还可以主动执行任务 。

扩散模型和多模态架构：这类模型将生成能力从文本扩展到包括视频、音频和数据模态，使银行内的各个职能得以实现高级自动化。

金融机构正利用这些模型实现自动化分析，提升报告质量并辅助决策制定。其应用范围广泛，涵盖所有关键领域，包括：

- **自动信贷决策：**将生成式人工智能模型集成到决策引擎中，从而可以通过非结构化信息（如财务报表和市场披露信息）确定信用等级。
- **软件工程和数据迁移：**通过代码生成和调试助手加快开发周期，同时促进既有代码库的转译和现代化，以减少技术债务。
- **交易监控与合规管理：**检测异常交易行为，并自动起草可疑活动报告（SAR）或生成关于疑点的解释说明。
- **预警系统：**总结大量市场新闻和内部报告，提前识别潜在风险的信号。
- **数据质量和元数据管理：**使用大语言模型来检测数据质量问题，推断元数据，并对数据进行分类以便后续建模。
- **知识管理和政策查询：**部署基于检索增强生成技术（RAG）的系统，使员工能够查询海量内部监管规则和法律文档，实现及时精准的信息检索。
- **环境、社会和治理（ESG）和声誉风险评分：**分析披露和社交媒体内容，生成风险评分，为制定交易对手与行业相关限额提供依据。

这些应用可通过以前难以量化的方式从非结构化数据中挖掘洞察信息，这表明生成式人工智能在推动银行转型方面极富潜力。

生成式人工智能给模型风险带来了诸多挑战

然而，复杂的生成式人工智能模型也带来了比传统模型更为显著的风险。核心挑战包括数据隐私和保密问题，尤其是敏感数据的外部处理导致的风险，以及员工使用未经批准的公开工具带来的“影子人工智能”风险。模型偏见引起的伦理问题，加上“幻觉”和编造事实等问题，会造成重大的声誉和合规风险，而可解释性和可追溯性不足也影响了结果的可复现性。

尤为关键的是，依赖少数外部模型供应商会产生第三方集中度风险。在这种情况下，单一供应商出现问题很可能会影响银行内多个关键职能。此外，包括《欧盟人工智能法案》和《数字运营韧性法案》（DORA）在内的新兴监管框架对已部署的人工智能系统形成了更严格的问责标准。

鉴于生成式人工智能的潜在影响范围，全面的模型验证成为了生成式人工智能治理框架的核心控制要素。它能确保模型在技术上文件可靠，符合伦理和监管期望，并能有效融入现有的模型风险管理体系。系统性的验证方法可为监管机构和利益相关提供透明度，降低运营和声誉风险，并帮助银行负责任地使用生成式人工智能。

监管机构对生成式人工智能模型验证的要求

2

与生成式人工智能相关的风险已推动监管讨论的重心从宏观的伦理原则转移到具体的验证要求。因此对于金融机构而言，生成式人工智能验证正从一项侧重技术的工作，演变为一个与合规相关的控制过程。尽管法律要求仍因司法管辖区不同而有所差异，但监管思路已达成共识，即围绕对生成式人工智能的概率性输出行为特征来进行管控、追溯和持续监督。

关键监管框架

欧盟（欧盟人工智能法案和监管期望）：欧盟已采用基于风险导向的法律框架。对于银行业的场景应用，如信用评估或风险评分等，通常被归类为“高风险”用例。这要求银行开展严格的数据治理、记录留存和人工监督。此外，欧洲银行业管理局 (EBA)和欧洲央行（ECB）等监管机构要求银行将现有的模型风险管理原则应用于生成式人工智能，要求开展与模型的复杂性和重要性相匹配的独立验证。

美国（美国国家标准与技术研究院（NIST）和模型风险管理监督指南（SR 11-7））：美国的监管方式结合了NIST人工智能风险管理框架以及现有的银行业监管指引（如SR 11-7）。在实践中，并非所有的生成式人工智能应用都被纳入模型风险管理的范畴，即使这些应用从技术角度来说也属于模型。监管对模型的分类越来越多倾向由具体应用场景驱动，而不是“一刀切”，从而使其管理方式与用例实际产生的风险相匹配，例如，避免将辅助性工具纳入模型风险管理。在生成式人工智能用例被视为模型的情况下，验证重点是具体实施方式的**适当性和有效性**（例如RAG架构和提示词策略），而非对第三方提供的基础模型进行深入的概念性评估。独立验证应该以风险为导向，并与用例的重要性匹配，而更广泛的人工智能治理则沿用NIST原则。

中国香港（香港金融管理局（HKMA）和香港证券及期货事务监察委员会（SFC））：香港在人工智能创新方面采取审慎与鼓励并重的态度。其中，HKMA发布了《应用人工智能高层原则》及《生成式人工智能沙盒实务洞见报告》，并要求人工智能应用的可解释性与其重要性相匹配，同时提供了生成式人工智能模型验证的实践案例。此外，SFC在《生成式AI语言模型的使用》中要求持牌法团在受规管活动中（例如提供投资意见）应用生成式人工智能模型时进行模型风险管理。HKMA与SFC均明确金融机构需落实独立验证、持续审查和监控机制，以确保模型的准确性。

国际标准（国际标准化组织（ISO）和经合组织（OECD））：包括ISO 42001在内的国际标准为人工智能管理体系的制定提供了基础框架。这些框架强调，验证工作不能仅局限在技术性能，还需延伸到伦理问责层面，以确保模型符合社会公平和反歧视等社会规范。

虽然不同司法管辖区的具体法律文本有所不同，但对银行业的整体影响是一致的，都要求从纯粹的技术测试转向更广泛的由治理驱动的验证程序。为满足此类监管期望，金融机构必须围绕四个关键点来搭建其验证体系。

2.1 向风险导向型验证转变

监管机构，尤其是欧盟的监管机构，承认并非所有的生成式人工智能应用都具有同等的风险。用于内部信息技术支持的聊天机器人不会像自动信贷决策引擎那样对金融稳定或消费者权利构成威胁。因此，监管机构要求采取风险导向型的方法，验证深度应与用例的重要性和风险等级相匹配。验证的资源需要相应做动态的分配，对高风险应用应该进行严格的对抗性测试（Adversarial Testing）和深度可解释性分析，而对于低风险的、面向内部使用的工具可以采用简化的治理流程。

2.2 管理第三方和供应商风险

正如上文关于“基础模型”相关的说明，银行很少从零开始训练大语言模型，他们通常会选择合适的第三方模型（如通过API或云平台接入）。由此产生了独特的监管难题：在基础训练数据和模型权重闭源，银行无法获取的情况下，如何遵守SR 11-7等监管要求？监管机构要求，无法获知此类信息不等于银行可以免除问责。因此，验证方式必须从检查源代码向**质疑供应商**转变。包括开展严格的基于输出结果的测试，审查供应商文件（如模型卡），并验证第三方模型外围的“封装层（wrapper）”或管控措施，以确保外部模型在银行风险偏好范围内运行。

2.3 落实可解释性和透明度要求

生成式人工智能的非确定性和“黑盒”性质与现有的消费者保护法（如《通用数据保护条例》（GDPR）对“解释权”的规定或美国对“不利行动通知”的要求）存在冲突。监管机构要求，如果人工智能参与银行业务决策，那么该决策必须可追溯、可理解。由于生成式人工智能模型无法始终解释其内部的神经网络路径，因此验证工作必须聚焦在**功能层面的可解释性**。

所以，需要测试模型是否能够引述信息来源（例如，从RAG架构中的引用溯源），并确保输出内容基于事实数据，而非幻觉编造的信息验证工作如同一道控制闸门，确保生成式人工智能的“魔法” 不会侵犯客户的透明度权利。

2.4 运营韧性和持续监控

与不经调整就不会改变的传统统计模型不同，生成式人工智能系统会动态变化，并且容易因用户提示词变化或基础模型更新导致“漂移”问题。欧盟的《数字运营韧性法案》和NIST的《人工智能风险管理框架》等规范均已着重指出部署前测试不足的问题。必须将验证重新视为一项**全生命周期活动**。这就需要对生产环境中的监控智能体进行验证，以确保能够实时发现错误输出、性能下降或数据泄露等问题。监管机构明确要求：必须实施连续控制，以确保模型在首次上线后长时间保持安全运行。

除了正式监管之外，金融机构还应确保在人工智能系统的设计和验证中引入伦理标准、社会价值观和规范。《经合组织人工智能原则》和欧盟委员会**《可信赖人工智能伦理准则》**等框架中均已提出“公平、透明、问责和人类监督”等核心原则。对于模型验证，这些原则衍生出具体的控制要求：必须证明输出不存在系统偏见，决策路径具有可解释性，且关键用例具备适当的人机协同机制。

“毕马威可信人工智能框架”综合了上述原则，可用于在人工智能全生命周期内负责任地设计、构建、部署和使用人工智能，并通过以下核心支柱落地实施：人工监督和问责、公平性、透明度/披露、可解释性、数据质量、隐私、网络安全和稳健性、应用安全性和可持续性等。

[1] **参见： https://kpmg.com/xx/en/what-we-do/services/ai/trusted-ai-framework.html**

中国香港生成式人工智能模型监管要求

在全球多地监管框架不断完善的背景下，中国香港的生成式人工智能模型的监管要求在理念与实践上更贴近国际标准，为中国内地市场的落地提供重要的借鉴。随着金融机构将生成式人工智能模型从后台运营自动化迅速拓展至高风险、面向客户的投资服务领域，其在通过自然语言交互显著提升运营效率的同时，也放大了传统模型风险。针对这一趋势， SFC的**《生成式AI语言模型的使用》**（2024） 、HKMA的**《应用人工智能的高层次原则》**（2019）和**《生成式人工智能沙盒实务洞见报告》**（2025），为香港金融行业内规范管理和审慎应用生成式人工智能模型提供了重要指引。

治理与高级管理层职责

HKMA与SFC对生成式人工智能模型的治理，以及董事会与高级管理层的相关职责提出了明确要求。董事会及高级管理层应对生成式人工智能模型驱动的决策和输出结果负有责任，并建立完善的治理框架、政策与程序。

金融机构的治理框架应涵盖模型开发、验证、使用到停用的全生命周期，并识别高风险场景（例如自动生成投资建议或财务报告），以防止不准确或具偏见的输出内容造成不当后果。同时，金融机构应确保“三道防线”配备数据科学与模型风险管理的专业人员。

生成式人工智能模型验证的核心原则

生成式人工智能模型验证应遵循独立性、全面性和持续性的核心原则，确保模型在全生命周期内稳健运行。

在将生成式人工智能模型部署至生产环境前，金融机构应开展独立的验证，以检验模型的准确性、适用性及在网络安全与数据治理方面的稳健性。模型验证单位应与开发单位分离，并可由二、三道防线或外部专家执行。

生成式人工智能模型验证应涵盖架构、假设、输入、计算、输出等方面。验证生成式人工智能模型应将评估范式从传统的统计回溯测试，转向对“可解释性” 的聚焦。测试覆盖从输入到输出的全流程及相关模块，如RAG、内容过滤与提示管理等。

生成式人工智能模型投入运作后应持续监测模型表现，及时应对市场环境变化或新数据集引入等情形。

沙盒实践项目中的模型验证实操

HKMA在2024年推出生成式人工智能模型沙盒实践项目，让金融机构在风险可控的环境中探索并测试生成式人工智能模型在风险管理、反欺诈及客户体验等领域的应用。在试验过程中，参与的金融机构进行了各种测试，验证了生成式人工智能模型多种配置的适用性，并完成相关模型验证与评估。

在沙盒实践项目中，金融机构通过不同的验证方法，对生成式人工智能模型的性能进行验证。这些方法包括：

- 提示词结构验证：构建包含常规业务指令、连续追问、故意误导等测试。
- 超参数验证：在不同超参数组合下测试模型表现。
- 基准测试：在多个基础大模型间进行并行评估。

在生成式人工智能模型验证中，模型评估是确保生成式人工智能模型在可靠性与准确性方面达到预期标准的环节。参与沙盒实践项目的金融机构普遍采用结构化的方法和多维度指标对生成式人工智能模型进行评估，包括常见分类指标（如准确率、召回率等）以及面向特定任务的专用指标，例如摘要生成、机器翻译、问答语义匹配等场景的度量标准。部分金融机构还引入了与业务场景更紧密相关的度量标准，例如聊天机器人应用的相关性准确率、幻觉率及响应延迟等，并结合量化与质化分析从多角度全面评估性能。

网络安全与数据隐私

生成式人工智能模型的使用引入了复杂的网络安全威胁及数据风险，金融机构应实施有效的网络安全措施以管理相关风险。

鉴于对抗性攻击可能窃取训练数据中的机密信息、诱导模型输出错误或不一致结果、甚至远程执行恶意代码，防护措施应同时涵盖模型本身及其用于训练或微调的数据。金融机构应在可行范围内定期进行对抗性测试，以提升模型的抗攻击能力并强化安全防护。

在数据管理方面，金融机构应确保对静态与传输中的非公开数据进行加密，并建立机密信息保护机制，以防止敏感客户及业务数据泄露。针对训练数据提取攻击，金融机构应实施监控机制，防止用户输入敏感机密数据或此类数据被灌输进生成式人工智能模型。在可行情况下，应优先使用已去除敏感信息的数据，以降低隐私风险并满足相关监管要求。

中国香港模型风险管理监管要求

在生成式人工智能模型验证的核心要素基础上，金融机构还需将验证环节纳入更高层次的整体模型风险管理框架中。2025年，SFC发布**《适用于模型风险管理的一般原则》**咨询稿。这是香港首份在整体层面确立模型风险管理监管要求的文件，奠定了香港对模型风险管理进行全面监管的基础。该原则要求理论基础，输入数据、定量估计等内容都要受到严谨的模型生命周期管理、独立验证以及定期监控。

治理、政策与文件记录

模型风险管理的最终责任由金融机构的董事会承担，其应建立全面的的模型风险管理框架，并通过适当的政策与程序规范模型风险管理活动。模型风险管理政策与程序应包括模型定义、开发标准、实施与变更流程、验证、重新验证及定期审查的标准，以及相应的治理与质询机制。同时，金融机构应明确不同职能在模型生命周期各环节中的角色与职责，包括模型所有者、开发者、使用者，验证单位、及内部审计。

严谨的文件记录是该原则的核心要求——模型开发的各个环节、验证报告、独立测试结果及已识别的缺陷都必须被全面记录。这一全面的文件记录标准保证第三方审阅人员或内部审计团队能够独立评估模型风险管理实践是否保持全面性与严谨性，并符合监管原则。

模型识别与模型清单维护

金融机构应建立统一且正式的 “模型”定义，该定义应由三部分构成：信息输入部分（提供数据和基准假设）、处理部分（通过数学或金融理论将输入转化为定量估算）、以及报告部分（输出定量结果）。

金融机构应建立完整及准确的模型清单，记录所有当前使用、正在开发或已退役的模型。。该清单在治理框架中可作为蓝图，通过在模型生命周期中明确分配责任来识别具体的模型所有者、开发者、使用者以及验证单位。

健全的模型开发管理

在模型开发阶段，SFC要求模型的基础理论、数学设计及核心逻辑在概念上具备合理性，并与其预期业务目标保持一致。开发阶段中应进行充分测试，以确保模型稳健性与适用性，例如进行输入敏感性分析、在不同经济压力情景下的情景测试、基准测试等。

在模型使用阶段，金融机构应执行持续性控制机制，包括评估市场及业务的变化，以确保模型符合性能与风险要求。这一过程需要模型所有者、开发者与使用者保持积极协作，以便及时识别并纠正潜在问题。

独立模型验证、重新验证与定期审查

SFC要求模型在投入生产使用前应进行全面的、独立的模型验证，并且在模型设计、假设、输入或输出发生重大变化时亦需重新验证。独立性是该原则的核心要求，确保验证过程对模型的三大核心部分——输入、处理和报告——进行客观的评估，以确认其在特定运营环境中的适用性。

除了部署前的初始验证外，SFC还规定持续开展模型重新验证，并每年至少进行一次定期审查，从而系统跟踪已知模型局限、发现新出现的漏洞，并评估随时间推移的性能衰退。验证单位人员应具备先进的技术能力、领域专业知识与深入的业务理解，并拥有明确权力向开发者提出挑战、实施严格的操作限制或直接拒绝不合格模型。

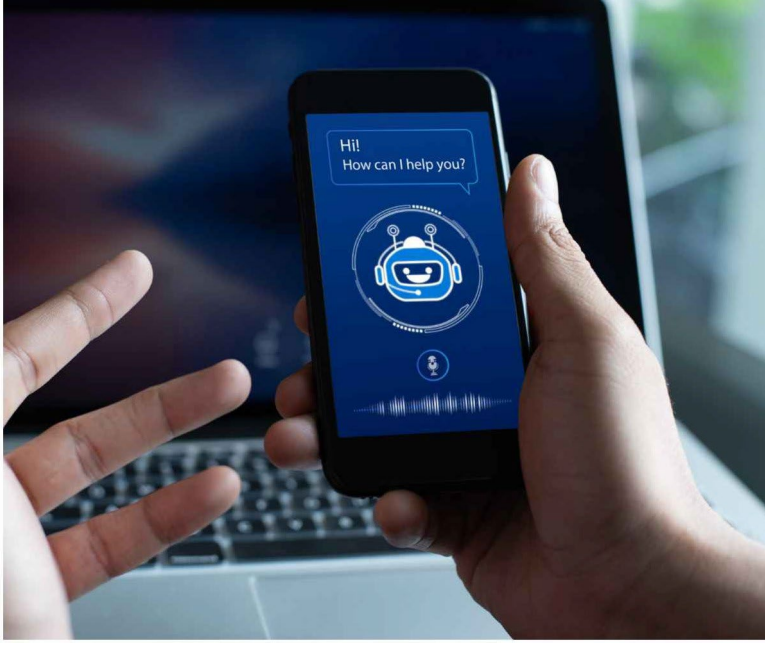
供应商模型管理

金融机构应在其模型风险管理框架内对供应商模型实施审慎管理与监督。金融机构应建立严格的供应商选择流程，以确保模型的适用性。金融机构应在选择供应商模型过程中获取供应商模型的资料，并将相关信息纳入模型清单，以便统一管理和后续审计。在使用供应商模型时，金融机构应开展持续性能监测，确保其在运行期间符合既定的质量与风险控制要求。

与模型验证相关的生成式人工智能特征

3

尽管生成式人工智能模型的验证建立在原有统计和机器学习验证实践的基础上，但由于这些系统可自动生成内容、输出具有不确定性且运作复杂，因此这类模型的验证需要根本性的范式转变。与传统预测模型不同，验证方式不能“一刀切”。必须针对具体的部署模式（云API或本地部署）、模型规模（大语言模型或小语言模型）和系统架构（对话、RAG或智能体）进行定制。



3.1 部署与模型控制：黑盒 vs. 白盒

验证的主要挑战取决于机构对模型的控制程度。

- 云端部署的大语言模型（API集成）：使用通过超大规模云平台部署的专有云托管基础模型时，金融机构会面临“黑盒”场景。训练数据缺乏透明度，“人工干预”的干预机制不明确，以及因未通知的模型更新导致模型行为迅速改变，都对验证构成了阻碍。因此，在这种情况下，验证必须聚焦在行为测试和漂移检测上，以识别外部变化。

- 小语言模型与本地部署：在验证方面，小语言模型具有明显的优势。由于它们可在本地运行且对硬件要求较低，因此可以进行“白盒”验证。金融机构可检查完整的技术栈，控制微调数据，并以远低于调用大模型的成本，开展网络活动追踪、全面敏感性分析等需要大量计算的验证工作。

3.2 核心模型验证领域

无论采用哪种部署模式，概率性输出的特性都使生成式人工智能需要特殊的测试方法，以确保其可靠性和推理能力。

- 功能正确性和推理能力：为了应对非确定性问题，验证必须采用黄金数据集比对（根据经专家验证的答案进行基准测试）和任务单元测试。自我一致性和多数投票等技术对于验证模型推理的稳健性（而非偶然性结果）至关重要，对于专家混合（MoE）架构而言尤为如此。

- 语境敏感性：必须测试模型的确定性和敏感性，以确保提示词中与语义无关的微小差异不会导致截然不同的输出。变形测试 (Metamorphic testing) 可以验证输入和输出之间的逻辑关系不会受模型变换影响。

- 信息检索限制：验证时，必须对模型在其上下文窗口中保留和处理信息的能力进行量化，利用“大海捞针”(Needle-in-the-Haystack tests) 测试来识别长文档中“丢失”的信息。

3.3 架构特定验证：RAG与智能体

随着模型架构的复杂程度已远超简单的聊天机器人，模型验证的范围也在不断扩大。

- 检索增强生成 (RAG)：对于知识系统，检索机制和生成过程必须分开验证。应进行检索效果评估（文档精确度和文档回忆）、归因和事实校验（验证答案是否真实存在于源文本中）以及数据污染测试，以确保模型不是在背诵已记住的公共数据，而是遵循内部策略生成内容。

- 智能体与函数调用：将模型接入工具使用时，验证成为一项软件工程工作。 关键测试领域包括针对工具使用进行契约测试（验证API模式的合规性）、通过计划和行动评估以验证多步骤逻辑的合理性，以及通过状态/内存损坏测试确保智能体不会陷入死循环。还必须进行成本和延迟验证，以防止过度消耗词元（token）。

3.4 安全、误用和稳健性问题

生成式人工智能会使受攻击面扩大，传统验证方式很少能够应对这一问题。安全性验证则可以兼顾模型风险和网络安全风险。

- 对抗性红队测试 (Adversarial Red-Teaming)：核心要求是让模型通过越狱套件和提示词注入挑战，来验证安全护栏无法被绕过。

- 数据泄露防护：验证过程中必须主动尝试提取训练数据或提示上下文。包括在提示词中使用诱饵标记词元（泄漏陷阱）验证上述数据会否出现在输出内容之中，并执行特定的个人身份信息/保密信息泄漏测试。

- 沙盒技术：对于智能体系统，应使用工具沙盒隔离和模糊测试来确保模型即使在收到恶意提示词的情况下也无法执行破坏性命令。

3.5 组织价值观、运营与文档

模型验证框架作为一项控制职能，应与组织的价值观和运营标准保持一致。

- 价值观一致性与公平性：必须对模型进行偏见/公平性的定量测试和毒性基准测试，以衡量其是否符合机构的行为准则和核心价值观。在作为银行业基本价值观的隐私方面，如果机构使用了不同的隐私技术，应验证隐私预算 (privacy budget)声明。

- 运营遥测：为确保有效验证，需具备用于端到端遥测和连续漂移检测的基础设施。在高风险用例全面发布之前，应该根据生产流量验证影子部署 (shadow deployment)。

- 可追溯的证据：为了满足监管审查要求，上述所有验证活动都必须记录在可追溯的证据包中，其中需包含将风险点与具体测试项一一对应的覆盖矩阵。这些记录应作为模型最终发布的条件。

3.6 战略整合：监管要求与模型风险管理

实施这些技术控制措施需要从根本上调整现有的模型风险管理流程。其中一个关键挑战在于将高层级的监管原则转化为具体的验证程序，例如将“稳健性”和“透明度”原则转化为具体的验证程序。

传统模型风险管理框架针对静态模型的预测准确性验证而设计，通常不适用于具有动态和开放特性的生成式人工智能。针对不同的监管制度协调开展验证，需要持续了解监管发展，合规和验证团队之间也必须密切配合。**如欲详细了解这些战略转变，请参阅毕马威发布的《生成式人工智能时代模型风险管理》[2]和《人工智能模型的现代风险管理》[3]白皮书。**

[2] 参见：https://hub.kpmg.de/en/model-risk-management?utm_campaign=FS%20-%20Whitepaper%20-%20Model%20Risk%20Management&utm_source-aem

[3] 参见：https://hub.kpmg.de/modern-risk-management-for-ai-models

毕马威
生成式人工智能
模型验证框架

4

基于前文已经识别出的关于生成式人工智能模型的特有考量，毕马威提出了全新的模型验证框架，并将相关要求转化为了可落地的验证活动。

4.1 模型及应用分析

首先，是对模型的预期能力、以及输入和输出的概念性理解。对于生成式人工智能模型，应阐明底层架构（例如大语言模型、小语言模型、多模态模型等）、其处理的数据类型及其输出的性质。应对模型文档记录进行详细审查，对开发人员进行访谈，并开展独立分析。模型的适用范围定义了模型能可靠运行的边界，包括其设计所覆盖数据域、语言类型或提示词类型。应对模型的理论框架进行分析，确定假设条件、简化处理和潜在概念缺陷问题，以估计这些因素对验证过程中对已记录场景和未覆盖场景的影响程度。适用的监管和伦理要求通常取决于模型用途、司法管辖区以及依赖模型输出的组织部门、职能和产品。同样，验证范围由模型的相关性和重要性决定，如模型的风险等级或输出结果的重要程度。

4.2 模型与输入质量

模型适用范围分析

验证过程的第一步是对模型的预期能力、以及输入和输出建立全面的概念层面理解。对于生成式人工智能而言，这需要跳出算法本身，来分析整个系统架构、数据关联关系和模型运行的特定业务环境。初始阶段应确模型适用范围，即模型能可靠运行的具体边界。鉴于生成式人工智能的开放特性，定义此类边界对于防止误用至关重要，例如避免误将创意十足的营销模型用于高精度的财务计算场景。

架构解构与系统设计

与传统预测模型不同，生成式人工智能解决方案很少仅包含单一组件。它们通常是综合系统，需要进行全面的架构审查。为了评估系统设计的理论可靠性，必须将解决方案分解到每个组件层面以进行验证。还应考量组件交互效应与系统涌现行为，以捕捉复杂的系统变化产生的风险。这个流程从**基础模型**本身开始。验证人员必须评估所选的基础模型（例如，领先的专有或开放式大语言模型）是否具有适配业务用例的参数数量和训练重点。例如，验证人员必须确保针对编程进行优化的模型不会被误用到创意写作任务之中。

分析范围还从基础模型扩展到RAG系统的**检索层**。这一层的理论缺陷，如语义搜索算法不佳或文档分块不正确，可能会导致“垃圾进、垃圾出”，使得即便是最强大的大语言模型也因此失效。验证过程中还需检查**提示词工程和系统指令**。这些指令是模型的超参数，必须对其稳健性进行审查，以确保它们能有效约束模型，避免其采用不当角色设定或或违反银行政策。

数据评估：预训练 vs 上下文

生成式人工智能在概念上的合理性取决于模型训练数据与银行实际运营环境的适配度。这个分析的关键之处在于确定基础模型的**知识截断时间**。例如，如果一个模型的用途是分析当前的市场风险，但其训练数据的截止时间在数年前，那么除非通过外部数据进行适当补充，否则这个模型就会存在基本的概念缺陷。

此外，验证人员还须评估**上下文窗口利用率**，即评估在推理过程中数据如何被输入模型。如果一个应用试图输入的词元数量超过模型能够可靠处理的词元数量，则可能会出现“中间丢失”问题，使位于长文档中间的指令或数据遭到忽略。因此，验证需要确认数据摄入策略是否与模型注意力相符。

定义模型的适用范围

这一阶段的一个关键点是首先明确界定 “不适用范围”即模型不能做什么。大语言模型的开放性可能会导致误以为其具备普遍能力，因此定义明确的**领域边界**至关重要。例如，用于“Python代码生成”的模型必须明确禁止用于提供“法律建议”。

这也包括**语言和文化**方面的限制。如果一个模型主要基于英语数据进行训练，那么必须对它能否有效解释德国当地银行法规进行深度质疑，以识别由翻译层引入的潜在细微差错。此外，分析必须明确区分确定性与创造性任务。将概率模型应用于确定性问题，例如使用标准大语言模型来计算利率，在概念上是存在缺陷的。在这种情况下，验证应使用确定性工具（例如Python计算器），而不应依赖大语言模型的next-token预测机制。

风险分级与监管映射

最后，作为风险分级环节的一部分，技术分析会与监管要求，以及银行内部的模型分级框架对齐。因此，验证范围需要根据模型的**重要性**进行校准，以反映模型的财务影响、声誉风险和对客户的可见性等。

通过将人工智能应用场景与监管框架如《欧盟人工智能法案》进行对标映射，这一评估流程将直接服务于监管分级处置工作。例如，在信用评估或生物识别场景中，如果某一应用被归为高风险应用，则验证计划需要调整为包含强制性合规评估、强化数据治理审查、并核验其具备严格的人工监督机制。这种对标方式确保了验证的深度与技术的复杂度、监管责任、以及具体用例的风险等级相匹配。

模型与输入质量

在完成模型的概念定义后，模型验证的重点将转移到系统的结构完整性上。这一阶段是对用于开发生成式人工智能解决方案的“原料”（数据）和“配方”（配置）进行严格检查。

在模型内部逻辑通常是非线性且难以解释的情况下，确保高质量的输入和稳健的配置是防止幻觉输出、一异常行为以及模型漂移的主要保障。

模型假设与局限性评估

每一个量化模型都会依赖简化假设来拟合现实场景。在传统的风险模型中，假设通常是明确的（例如，具体统计分布），而生成式人工智能的假设通常是隐性的，隐藏在模型架构和训练语料之中。因此，验证中必须通过理论分析来识别和严格评估这些**隐性假设**。例如，如果某解决方案假设大语言模型能够进行“逻辑推理”，那么验证人员必须确定模型是否真正执行多个逻辑步骤，还是仅仅进行概率模式的匹配。

当这些假设与观察结果不符时（例如由于训练数据时间过早导致模型未能反映最新的监管框架），这些差异都需要被认定为关键局限性。验证团队应随后量化这类局限性的**重要性水平**，以界定模型的可靠性边界。例如，如果模型隐性假设所有输入均为英文，那么验证时必须确定其在输入非英文文本情况下属于失效模式。这部分模型验证的结论应该包括模型在哪些条件下运行是可靠的，同时要明确指出在什么情况下模型的输出结果应该被拦截或者标注预警。

数据质量：训练、验证与检索

数据构成了任何生成式人工智能系统的功能核心，因此有必要对其进行严格细致的审查，其程度远超标准的质量检测。对于金融机构自己**微调**的模型，验证必须评估数据集代表性，确保数据集能够包含各类极端边缘场景，例如一些罕见的欺诈类型。重要的是，验证必须仔细排查**数据污染**问题，核查用于模型测评问题没有错误地包含在训练集中，否则将会造成性能指标虚高，并对模型的实际正确性产生误导。

在检索增强生成（**RAG**）系统中，“输入”这一概念扩展为包含检索到的上下文信息。验证工作需要对向量数据库中存储的非结构化数据（如PDF、政策文件）质量进行测试。文档解析残缺或文件版本过时不可避免地会导致“垃圾进，垃圾出”现象，即模型会生成语法正确但有悖事实的答案。此外，还应对数据集存在的**偏见和伦理风险进行**评估。如果使用历史贷款申请数据对信用评分模型进行了微调，那么验证必须确保过往业务中针对受保护群体的决策偏见不会被纳入新的模型参数权重中。

模型配置和超参数：模型配置的精细化决定了模型稳定性和创造性之间的平衡。验证时应审核这些配置，以确保它们与业务目的紧密挂钩。对于**“白盒”**模型（例如，本地部署的或微调模型），验证时应审查训练的严格程度，验证损失函数、优化算法和收敛指标的适用性。

这包括检查**“灾难性遗忘”**问题，用来确认模型在针对特定任务做微调时，不会造成其通用推理能力的退回。

对于“黑盒”模型，比如通过API调用的闭源基础模型，验证重点应该是银行可管控的推理参数。这包括评估各种设置，如**温度系数**（随机性）、Top-P（核采样）以及**频率惩罚系数**。例如，较高的温度系数配置可能适用于起草营销材料的工具，但对于需要确定性、可复现输出的监管报告工具而言，则会被视为一个高风险缺陷。然而，模型行为在很大程度上可能取决于受供应商控制的配置和隐藏在底层的实施逻辑（例如，量化压缩、多租户隔离、后端工具调用、路由和负载均衡）。这就导致“黑盒”部署模式的透明度受限。因此，不存在100%可靠的验证方法。只有依靠常态化管控和周期性复测，才能持续保障人工智能应用可靠可控。

生成式人工智能专项：提示词工程与模型对齐

归根到底，验证所要检查的是用于引导模型输出的各类技术。对于生成式人工智能，**提示词工程**本质上属于一种软件开发工作。系统提示词需要作为软件产物进行检查，并针对逻辑性严谨性、语义清晰度以及抵御注入攻击的稳健性进行测试。验证环节需要判断：应确定提示词能否有效约束模型行为，还是依赖模糊、基于主观感觉的指令，从而导致输出结果不一致。

此外，若模型利用**基于人类反馈的强化学习（RLHF），**则必须在验证时审阅其对齐标准。如果奖励模型将“有用性”置于“真实性”之上，系统可能会倾向于**阿谀附和：**即使用户给出错误信息，模型也会同意用户观点。验证时应确保对齐策略将事实准确性和安全性置于用户满意度之上，确认模型的构建目的是输出正确信息，而非取悦用户。

4.3 落地实施与输出质量

在完成模型原理合理性和输入数据质量核验之后，验证流程进入**实施完整性和输出表现**检查环节。这一阶段是衔接理论设计和实际投产的关键纽带，确保模型已从设计方案落地为稳健的、可用于生产的系统，能够在银行自有的技术架构环境中可靠运行。

技术实施与基础设施验证

在对模型实施进行评估时，首先应承认生成式人工智能解决方案很少仅包含一个孤立的模型，它通常是一个综合系统，依赖于一系列由检索机制、API和安全封装模块构成的复杂链路。因此，验证必须检查整个**端到端基础设施**，以确保技术实现内容完全符合已审批的概念设计。这涉及到对连接组件的“粘合代码”进行仔细审查，特别是用于个人身份信息脱敏和提示词模板的前置处理逻辑，以及用于过滤有害输出的后置风控规则。

对于使用RAG的架构，验证范围延伸至**向量数据库或图数据库**实施落地。验证人员应检查索引逻辑和文本分片策略，以确保检索机制能够在高负载下正常运转，在获取相关上下文时不截断关键信息。检查还应包括API集成层，以验证系统能否以可控的方式处理接口延迟、调用延迟和服务超时，并确保在出现问题时能够“安全失效”而非“故障开放”。这一代码级检查旨在确认模型算法可靠，且整套技术栈的集成落地符合机构安全和开发规范。

独立测试与概率性复现

生成式人工智能模型的正确性验证，需要从传统的确定性复现思路转变为概率性复现。传统风险模型可在验证过程中分布复现运算结果，而生成式人工智能因随机采样机制，每次的输出结果存在差异。为了解决这一问题，独立测试依托**“黄金数据集”**，即使用由专家筛选核验的输入和预期输出作为基准，来评估逻辑是否一致，而非强求输出文本完全一致。

验证团队通常采用**影子测试**或“模型对模型”的基准测试，将待验证模型的输出结果与已经过验证的可信原型或者能力更强的参考模型（例如，用业内顶尖的主流基座大模型对标评测本地部署的小模型）的输出结果进行比对。适用这种方法，验证人员能够对推理能力的偏差幅度进行量化。如果实施的模型始终违反负面约束（如“不得提供投资建议”），则应对此类问题进行分析，以确定它们是源于提示词脆弱性还是存在根本性的实施缺陷。也要通过多轮重复试验评估模型推理逻辑是否在具体措辞不同的情况下也可以再现。

解锁黑盒：注意力分析与反事实测试

想要确保模型可信赖，需验证模型生成特定输出的内在原因。为此，这就需要借助**可解释人工智能（XAI）**技术，以应对系统的“黑盒”问题。对于内部权重可知的模型，验证人员应利用**注意力权重分析**揭示模型如何在其上下文窗口内对信息进行优先级排序。这对于验证RAG系统尤其重要；通过映射注意力头，验证人员可以确认生成的答案是否真正基于检索到的文档，还是无视上下文，仅凭预训练知识凭空生成内容（模型幻觉）。

对于闭源模型，则依赖**反事实测试**来评估其稳定性和公平性。这种方法的操作逻辑是：在保留原有语义意图不变的前提下，系统性的更改提示词中的特定属性，如变更人口统计标签、位置或币种等。如果某个无关紧要的微小变化就造成输出的前后矛盾（“蝴蝶效应”），则模型应视为不稳定。

评估输出性能和稳健性

最后，还要对照模型本来的业务目来评估其输出的质量。传统的回溯测试方法被**语义性能指标**所取代。句法得分（如BLEU或ROUGE）可用于衡量单词重叠，而银行业则更重视语义一致性。包括使用基于嵌入的指标，如BERTScore或**“LLM即评审（LLM-as-a-Judge）”**框架。此类框架利用经过校准的高级模型对目标模型响应的相关性、准确性和基调进行评级。

除功能正确性外，模型还必须经过严格的**稳健性测试和压力测试**。验证团队应模拟各类对抗场景，包括旨在绕过安全过滤器的“越狱”攻击、用于测试模型上下文记忆上限的长文本输入，以及含噪或格式错误的异常数据。验证团队还要开展专项测试来评估模型的**幻觉发生率**。这些测试会使用知识库中未包含的事实来向模型提问，以验证它是否会承认自己的知识局限性，而不是编造答案。对于高风险应用，自动化量化指标永远需要搭配**人工干预**做补充校验，以最终证实模型的输出在统计逻辑上成立，在运行上具有安全性，在专业上具有合理性。

4.4 模型缺陷与局限性

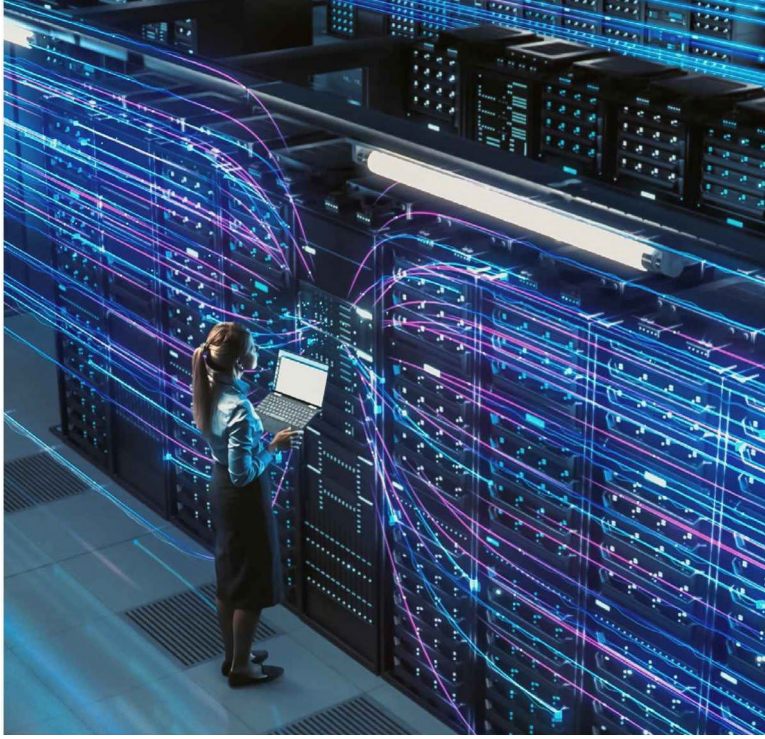
验证流程的重点不仅局限于一份技术评分表，还包括对模型剩余风险的全面评估。在最后这个阶段，应将架构分析、数据审查和性能测试的结果综合起来，以全面反映模型的适用性。

问题汇总与风险严重性评估

验证人员应将所有观察结果（从概念设计缺陷和数据缺陷，到幻觉等具体输出问题）整合到结构化的风险清单中。对于传统确定性模型，可以很容易将结论归类为对或者错，而生成式人工智能模型的缺陷分级需要更加精细化。所有问题应该依照**严重程度和影响重要性**进行评级。例如，聊天机器人中的轻微行文风格不一致可能被评为“低严重性”问题，而“越狱”成功或信用决策辅助工具存在的明显偏见，则可能被评为会导致模型不能被部署的“高严重性”问题。

补偿性控制措施评估

鉴于生成式人工智能的概率性输出特性，在技术上通常无法实现“零缺陷”。因此，此阶段的一个关键点是评估**补偿性控制措施**。验证时应确定模型外围的“封装层”（如严格内容过滤、置信度阈值或强制性人工干预工作流）是否能有效缓释已识别到的风险。然后，验证人员应评估**剩余风险**，确定运行控制措施是否足以在故障影响客户之前提前发现故障，从而确定模型能否在存在局限性的情况下安全部署。



治理层沟通与审批决策

验证结果共同构成**有效质疑**的基础。验证人员应就所有识别出的缺陷及其对合规的影响与模型开发人员和业务负责人进行透明的讨论。通过讨论，各方对模型缺陷补救的优先级达成一致，确定哪些问题必须在模型上线之前解决（关键问题），哪些问题可以在部署后进行监控。

最后，验证人员应撰写一份正式的**验证报告**，以作为决策者的的权威参考依据。该报告要强调模型的适用范围和使用局限，并明确该模型不能被使用的业务场景。在这一全面评估的基础上，验证职能部门可给出三类审批建议：完全批准使用该模型、附带加强监督和限制的有条件批准、以及驳回使用该模型。

模型验证的实际应用和适用范围

以上关于模型验证的整套框架可针对不同部署场景和系统架构，在确保风险适当的前提下以结构化的方式，对生成式人工智能模型开展验证。针对生成式人工智能的特性，这个框架清晰定义了标准化的验证步骤，同时与现有惯例（如模型风险管理和监管期望）保持一致。此框架为反映模型假设、局限性和控制措施提供了统一的方法基础。这有助于开展适当的模型审批和后续审查，并为依据现有监管和机构要求以受控的方式使用生成式人工智能模型奠定了坚实的基础。

展望

5

依托现有的模型风险管理体系和陆续出台的人工智能专项法规，金融行业生成式人工智能的监管框架已基本搭建完成。后续监管部门会通过发布指引细则、行业标准、以及监管实操细则还持续细化监管要求，但底层监管原则在短期内不会再发生根本性变化。因此，金融机构面临的主要挑战不是应对频繁的监管变化，而是在各类新兴技术场景下，对现行监管要求形成统一的落地解读。

相比之下，生成式人工智能系统的技术发展速度远超监管要求变化的速度。新的模型架构、部署模式和系统集成方案不断涌现，往往在未改动其监管分类的情况下，就改变了生成式人工智能应用的风险特征。这种法规落地节奏与技术发展速度的错位，凸显了模型验证工作的核心思路：即聚焦根本性的风险驱动因素，如透明度、可管控性、稳健性和适用性等，而非拘泥在具体模型的落地形态。

在上述背景下，一套具备技术适配韧性的人工智能验证方案，可在一个稳定的，原则导向的分层框架下搭建，能够适应不断迭代变化的各类生成式人工智能架构和部署模式。

1. 模型适用范围分析

通过明确模型用途、架构、输入/输出，来定义清晰的模型**适用范围**。

- 识别关键假设和模型缺陷，并结合实际用例、司法管辖区、重要性水平和风险等级等因素映射相应的监管和伦理要求。

2. 模型与输入质验证

对整个生成式人工智能系统（包括基座大模型+RAG/检索+提示词/系统指令）进行验证，而不仅仅检查算法本身。

- 验证训练/微调+检索上下文数据的完整性，排查偏见、污染和漂移风险（“垃圾进、垃圾出”）
- 检查模型配置/推理参数（如温度系数）和提示词稳健性（包括防注入抵抗能力），以确保能模型能够根据业务目的稳定输出。

3. 模型实施与输出质量验证

验证模型的端到端实施是否与设计（数据流、安全性/隐私控制措施、安全围栏机制）匹配，确保系统出现故障时可安全降级运行。

- 使用黄金数据集、重复试验、压力/越狱攻防测试以及溯源/可解释性检查来测试模型的概率化输出表现。

4. 缺陷与局限性评估

将所有验证结果整合到风险清单中，对严重程度/风险重要性水平进行评级，并确定补偿性控制措施（人工干预、内容过滤、使用范围约束）

· 最终出具验证报告，并完成治理层审批决议（完全批准/有条件批准/驳回使用），并明确该模型不能被使用的业务边界。

从理解生成式人工智能的风险到积极应对

生成式人工智能已不再是一项试验性技术工具，它正迅速融入包括风险评估、合规分析、报告编写和决策支持在内的银行业核心业务流程。尽管人工智能的监管框架仍在不断完善之中，但核心监管要求已经明确：即**银行应对其部署的人工智能系统的行为、输出结果及其产生的影响承担全部责任。**

因此，对于高级管理层而言，其战略要务不再是能否安全地使用生成式人工智能，而在于**是否具备充分的控制措施，从而有足够的信心来规模化部署生成式人工智能。**

为此，在等待监管要求完全明晰的同时，银行应该立即采取行动，将现有成熟的模型风险管理和模型验证原则适配到生成式人工智能模型的特性上，搭建生成式人工智能的模型验证体系并开展验证工作。

三大关键举措

1

首先，将生成式人工智能模型验证工作纳入现有风险管理体系

董事会和高级管理层应该对每个生成式人工智能的用例都明确责任主体，完成风险分级，并根据其重要性水平开展相应的独立验证。生成式人工智能不能仅被当做一个“技术问题”，它属于业务和风险管理范畴，应该和其他对银行有重要影响的模型一样收到严格管控。

2

第二，动态调整验证频率和范围

与传统模型不同，生成式人工智能系统的行为可能会由于提示词漂移、第三方模型修改或知识库（如RAG）更新而发生改变。因此，金融机构需要动态调整验证的频率和范围来应对相应的变化。

3

第三，确保验证结论能服务于业务实操

模型验证不能仅局限于发现技术层面的问题。验证结果必须能够直接用于确定模型应用边界、人工干预规则、设置监控阈值和异常上报机制。

因此，将生成式人工智能模型纳入整体的模型风险管理框架中至关重要。

如果金融机构采取果断行动应对上述问题，将获得实实在在的竞争优势。金融机构可以更快的落地生成式人工智能模型，从容应对监管检查，且无需因为技术的迭代升级反复重构治理架构。最重要的是，金融机构可借此确保监管机构、客户和内部利益相关方相信他们会以负责任的方式寻求创新。

从这个角度而言，生成式人工智能的模型验证不仅仅是一种防御行动，更是一项**战略赋能抓手：**依托这个稳健的管控体系，银行能够在不影响安全性、合规性或机构诚信的情况下持续创新。

长远来看，模型验证、常态化的模型监测、和实操风控体系这三者的协同将持续加深。模型验证结果在定义模型使用范围、监控阈值和异常上报机制等方面将发挥越来越重要的作用。如果能成功地将生成式人工智能的模型验证纳入其整体治理框架，那么金融机构将会从容地应对这项新技术的持续迭代，而无需从根本上重构风控体系。

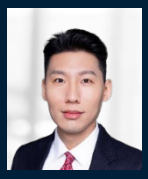
联系我们

毕马威中国

张楚东
金融业主管合伙人
毕马威亚太区及中国
tony.cheung@kpmg.com

史剑
银行业主管合伙人
毕马威中国
sam.shi@kpmg.com

杨权林
银行业咨询主管合伙人（内地）
毕马威中国
david.yang@kpmg.com



赵鹏
香港金融风险咨询主管合伙人
香港中资金融机构咨询联席主管
毕马威中国
robert.zhao@kpmg.com

李砾
金融行业研究中心主管合伙人
毕马威中国
raymond.li@kpmg.com

kpmg.com/cn/socialmedia

如需获取毕马威中国各办公室信息，请扫描二维码或登陆我们的网站：
https://kpmg.com/cn/zh/about/office-locations.html

刊物编号：1781493645559