人工智能金融应用评价体系研究报告

北京国家金融科技认证中心
National Financial Technology Certification Center (Beijing)

北京国家金融科技认证中心有限公司

二零二一年十月

每日免费获取报告

1、每日微信群内分享**7+**最新重磅报告；

2、每日分享当日**华尔街日报**、金融时报；

3、每周分享**经济学人**

4、行研报告均为公开版，权利归原作者所有，起点财经仅分发做内部学习。

扫一扫二维码
关注公号
回复：**研究报告**
加入“起点财经”微信群。。

◆ 编写单位（排名不分先后）：

北京国家金融科技认证中心

中国网络安全审查技术与认证中心

中国金融认证中心

国家金融科技测评中心（银行卡检测中心）

中国软件评测中心（工业和信息化部软件与集成电路促进中心）

中国互联网金融协会

北京金融科技产业联盟

北京百度网讯科技有限公司

华为技术有限公司

深圳市腾讯计算机系统有限公司

腾讯云计算（北京）有限责任公司

金杜律师事务所

◆ 编写组成员（排名不分先后）：

北京国家金融科技认证中心 张海燕

北京国家金融科技认证中心 刘力慷

北京国家金融科技认证中心 温昱晖

北京国家金融科技认证中心 李冬妮

北京国家金融科技认证中心 窦永金

北京国家金融科技认证中心 何一江

中国网络安全审查技术与认证中心 彭琳赓

中国网络安全审查技术与认证中心 张玉朋

中国金融认证中心 胡莹

中国金融认证中心 王飞宇

中国金融认证中心 吴宝民

国家金融科技测评中心（银行卡检测中心） 杨波

国家金融科技测评中心（银行卡检测中心） 邱晓慧

中国软件评测中心（工业和信息化部软件与集成电路促进中心）赵亮

中国软件评测中心（工业和信息化部软件与集成电路促进中心）翟云

中国互联网金融协会 朱勇

中国互联网金融协会 李健

中国互联网金融协会 单剑锋

中国互联网金融协会 于圆

中国互联网金融协会 肖翔

中国互联网金融协会 田然

中国互联网金融协会 靳亚茹

北京金融科技产业联盟 聂丽琴

北京金融科技产业联盟 黄本涛

北京金融科技产业联盟 刘宝龙

华为技术有限公司 符海芳

华为技术有限公司 曹晓琦

深圳市腾讯计算机系统有限公司 蒋增增

深圳市腾讯计算机系统有限公司 刘海涛

深圳市腾讯计算机系统有限公司 杨晓光

腾讯云计算（北京）有限责任公司 周磊

北京百度网讯科技有限公司 董晋利

北京百度网讯科技有限公司 冯博豪

北京百度网讯科技有限公司 张萌

北京百度网讯科技有限公司 王云菲

北京市金杜律师事务所 吴涵

北京市金杜律师事务所 张子谦

目　录

◆ 研究背景

金融业在场景、数据、技术、人才等方面沉淀了大量资源，在国家行业政策的大力支持下，人工智能金融应用发展迅猛，与此同时人工智能应用所带来的安全和伦理等问题也引起国家社会的广泛关注，向善、公正、安全、可信成为人工智能金融应用行业发展的底线。如何评价金融业人工智能技术应用符合“科技向善，安全可控”的社会需求，是否满足国家行业对人工智能金融应用的监管治理要求，是当前金融业人工智能技术发展的共性问题。

为此北京国家金融科技认证中心有限公司联合行业内认证机构、检测机构、商业银行、第三方支付机构、金融科技企业等相关单位，在金融业内开展人工智能技术的政策法规、应用场景、评测技术等方面的调查和研究工作，致力于构建符合金融行业特点的评价信任体系，形成多方共治的人工智能金融应用生态。

一、 人工智能金融应用现状

人工智能作为第四次工业革命的引擎，产业发展已成为国家和区域经济转型升级的关键驱动力。国家鼓励并大力支持人工智能技术的创新与应用，出台一系列政策引导人工智能产业规范化发展，以促进人工智能应用适应现代化发展需求。2017 至 2019 年，连续三年的政府工作报告中均提出加快人工智能产业发展的要求；2020 年，人工智能更是与 5G 基站、大数据中心、工业互联网等一起被列入新基建范围。近期，政策的出台更具针对性，更强调技术的落地效应，更为关注人工智能与产业的融合。同时，相关立法和行业伦理规范正在日趋完善，行业标准体系也在逐步建立，为人工智能产业的健康发展提供了良好环境。当前，拥有丰富场景资源、高质量数据资产、并具备业务创新迫切需求的金融行业急需人工智能技术的落地与良好实践。

（一）人工智能金融应用政策环境

从政策环境看，2017 年国务院印发《新一代人工智能发展规划》提出了面向 2030 年我国新一代人工智能发展的指导思想及战略目标。在国家纲领的指引下，金融行业持续推进人工智能金融应用相关政策措施的制定。2019 年 9 月，中国人民银行发布《金融科技发展规划》，提出探索相对成熟的人工智能技术在资产管理、授信融资、客户服务、精准营销、身份识别、风险防控等领域的应用路径和方法，强化智能化金融工具安全认证。2021 年

9 月，网信办等 9 部委联合印发《关于加强互联网信息服务算法综合治理的指导意见》，提出算法安全治理机制、完善监管体系、规范算法生态的建设目标，积极开展算法安全评估，有序推进算法备案。2021 年 3 月，中国人民银行发布金融行业标准《人工智能算法金融应用评价规范》规范了人工智能算法在金融领域应用的基本要求、评价算法、判定准则。2021 年 9 月，科技部发布《新一代人工智能伦理规范》旨在将伦理道德融入人工智能全生命周期，为从事人工智能相关活动的自然人、法人和其他相关机构等提供伦理指引。通过以上政策，我们可以看到国家高度重视人工智能在金融领域的应用与发展，力图构建安全可信、公正向善的人工智能技术应用。

（二）人工智能金融应用要素映射

人工智能技术发展的三大要素是算力、算法、数据，而人工智能金融应用的发展还需要场景要素的支撑。目前，金融行业主要在智能风控、智能投顾、智能支付、智能客服、智能营销、智能监管、智能运营七大场景中应用人工智能技术，涉及用户、行为、业务等多种数据，如下表所示。

表 1 人工智能金融应用场景

序号	应用场景	数 据	算 法
1	智能风控	客户数据、外部数据、交互数据、规则库	LR、XGBoost、GBDT、DeepFM 、 RNN 、 LSTM 、 Dijkstra 、PageRank、LPA 等
2	智能投顾	股票基本面、技术面数据、舆情数据	多因子模型、集成学习模型、LSTM 模型、强化学习模型
3	智能支付	客户生物特征数据	GMM-HMM、DeepSpeech Insightface 、 facenet 、 MTCNN 、DenseBox 等
4	智能客服	金融行业语料、历史对话数据、规则库	GMM-HMM、DeepSpeech WaveNet 信息检索、语义理解、图算法、语义特征计算等
5	智能营销	客户画像、产品画像、触达交互数据	因子分解机算法、集成学习类算法等
6	智能运营	业务流程数据、用户行为数据	GBDT、LR、XGBoost 等
7	智能监管	监管规则、财务数据、流水数据	LR、XGBoost、GBDT、DeepFM、RNN、LSTM、Dijkstra、PageRank、LPA、图神经网络

（三）人工智能金融应用调研

为深入了解人工智能算法在金融领域应用情况以及发展水平，我们在北京金融科技产业联盟人工智能专委会内开展了人工智能算法金融应用调研工作。调研范围包含 56 家金融机构、72 家科技企业，3 家检测认证机构，共计 131 家单位参与其中。调研内容分别从算法以及场景两个维度梳理行业现状及难点痛点，调研反馈结果汇总如下。

从算法应用维度看，人工智能技术金融应用的背后有一系列多种类算法组合的支撑，当前各调研反馈机构使用的人工智能算法已多达 40 余种，传统机器学习类应用最为广泛，深度学习类也在部分场景有所应用，其中 94%的机构使用决策树类算法，87%的机构应用了逻辑回归类算法，42%的机构应用了语言处理类算法，35%的机构应用了图神经网络算法，29%的机构应用了聚类算法。通过调查可以看到金融机构对算法的选择还是以安全发展、风险可控为前提，本着成熟算法先行试点的基本原则，稳步推进人工智能算法的应用。

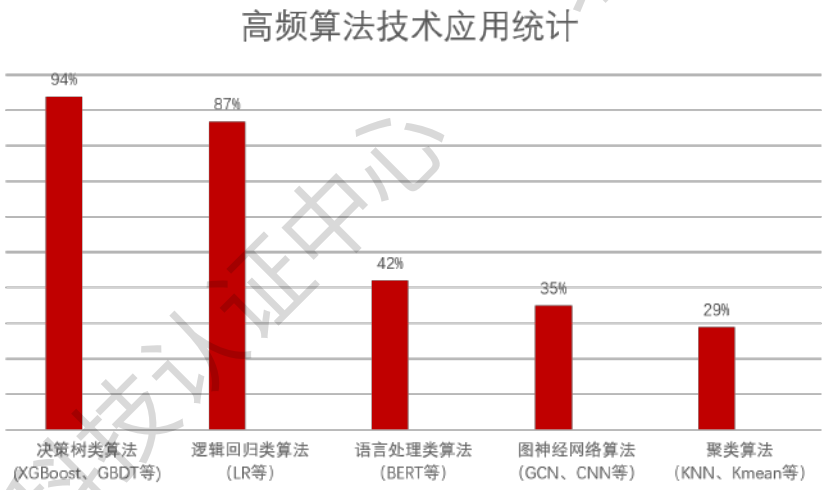

图 1 高频算法技术应用统计

从场景应用维度看，当前智能风控、智能客服、智能营销三大应用场景热度最高，其中 100%机构构建智能风控、90%机构部署智能客服及智能营销、76%机构升级智能运营，52%机构探索智能支付及智能顾投，还有 5%将人工智能技术应用到智能监管。

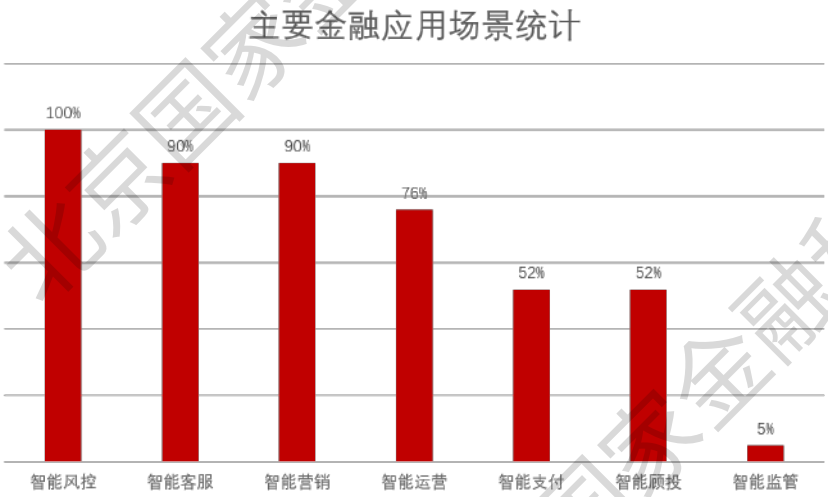

图 2 主要金融应用场景统计

从算法构建难点看，当前遇到的困难主要集中在算法可解释

性提高、算法风险安全评估以及数据样本不足三个层面。在算法可解释性提高方面，目前缺少成熟和体系化的标准指导算法开发者和使用者判断可解释性高低，同时部分算法因可解释性差无法有效指导使用者做出行为决策。在算法安全风险评估方面，评估依据以及评估指标还有待完善，评价基准还未形成。在数据样本不足方面，缺少符合金融特性的开源数据集以及安全可靠的数据共享机制，导致数据集筹建成本偏高。

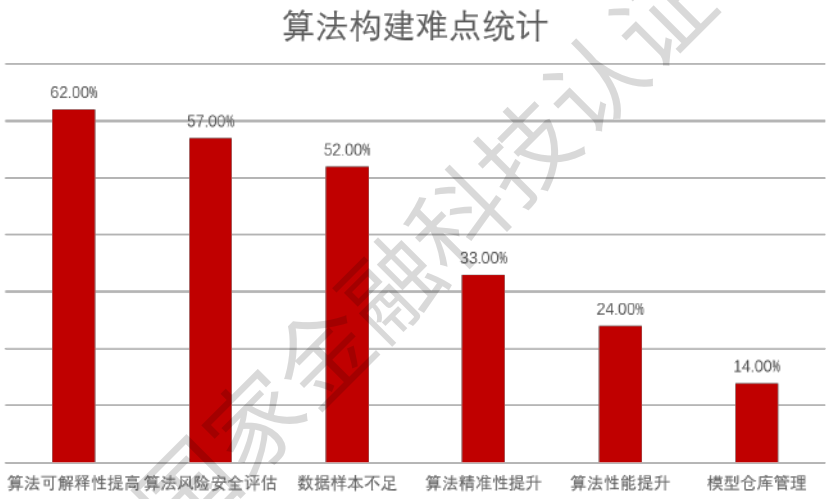

图 3 算法构建难点统计

从场景应用难点看，当前遇到的困难主要集中在场景碎片化、隐私保护、风险责任界定模糊三个方面。场景碎片化暴露出模型泛化能力不足，影响模型推广。人工智能应用涉及海量客户隐私数据的获取与加工，隐私信息界定以及数据权益归属仍需进一步明确。同时，人工智能算法通过结果输出辅助涉及多个行为主体的金融决策，各主体间的行为不易认定和划分，责任难于追溯。

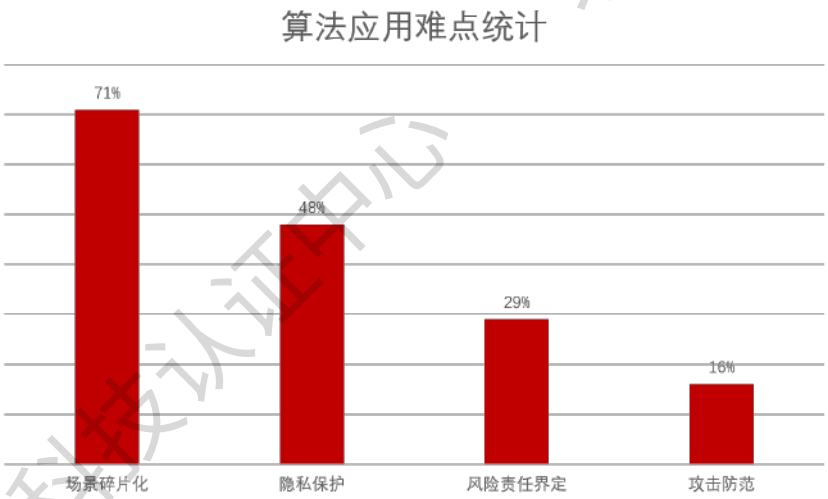

图 4 算法应用难点统计

通过本次调研，我们可以看到当前人工智能金融应用发展与《金融科技发展规划》中提出的思路保持高度一致，金融行业正在稳步推进成熟人工智能技术在重点金融场景的应用，打造安全可信、透明可解释、可控可靠的智能金融新生态。

（四）人工智能金融应用评价

随着人工智能技术的纵深发展，人们对于人工智能自动化便利的依赖越来越强，同时也引发社会对于“一切自动化”的多种担忧。特别在金融、医疗等高风险领域，智能应用的结果可能对财产、生命带来不可逆的重大影响，对社会治理带来极大挑战。

因此，积极做好人工智能应用评价工作，对实现人工智能健康持续发展显得尤为重要。人工智能金融应用的评价是一项较为复杂的系统性工程，不同应用场景间需要评价标准及评价指标的精准适配，更需要技术创新型评价工具的有力支撑。

从目前人工智能金融应用评价内容看，可总体分为“数据安全”、“算法应用评价”和“服务能力评价”三个层次。其中，数据安全评价是指对数据安全生命周期内的安全保护机制的作用，保护措施的效果进行评价，严防数据误用、滥用，切实保障金融数据和个人隐私安全；算法应用评价是指金融应用场景下算法表现的评价，更专注于算法本身；服务能力评价是指人工智能金融应用作为产品服务的能力，是算法应用评价的延伸与拓展，包括环境、系统、管理能力等。

1、数据安全评价

（1）数据来源的合法合规性

AI 金融应用开发者或使用者应按照《个人信息保护法》、《个人金融信息保护技术规范》等法律规范的要求，确保 AI 金融应用所使用和处理数据来源的合法性；应以正当、成比例的方式进行数据采集，需以理性人的视角判断数据处理可被接受的程度，确保其“可被预期、可被接受”。

（2）数据质量管控

为确保人工智能金融应用输出结果的可预期性以及可用性，应当要求开发者或使用者对人工智能金融应用所使用和处理数据进行质量管控，确保数据的可靠性、关联性、准确性、完整性。

（3）数据使用过程中的数据保护

人工智能金融应用中对数据的使用、加工、存储等多环节的处理行为也可能导致应用结果的负面效应，应当确保数据在研发、测试以及投入市场过程中其用途均严格限制于个人信息主体或数据提供方的授权范围之内。

（4）数据治理人员的管控

应根据人员岗位、职责、级别等确定其数据治理权限，包括对数据的访问、读写、复制等，并明确相应的奖惩机制；其次，应当在权限设置的基础之上建立严格的权限审批流程和制度；再次，当数据治理人员需超范围访问或修改数据时，建议建立风险评估制度，针对访问场景、目的、数据范围、人员等多方面进行风险评估；从次，为后续审计或自证合规目的，应当对数据访问、修改等处理操作进行记录并妥善留存；最后，需对相关的数据治理人员开展教育培训以确保其可以理解并落实相关的管控制度。

2、算法应用评价

（1）算法安全性

开源框架及依赖库的安全：人工智能应用过程中，为避免开源框架或者依赖库出现底层缺陷,进而产生误判问题,应当对算法中的各种依赖库和开源框架进行评估。

算法训练安全：为避免算法训练过程中出现安全问题，应对人工智能算法训练步骤进行安全评估。

（2）算法可解释性

可解释性按照模型本身具有可解释性和模型本身不具有可解释性方法处理。涉及下面几方面内容。

模型评估指标可解释：训练模型和测试模型的过程中，评估指标需要可解释。常用的模型评估指标为 AUC 值、精确率、召回率、F1-Score、MicroF1、Macro F1 等，通过测试进行指标统计验证评估指标的准确性。

模型特征重要性可解释：模型特征重要性可解释通过评估设计文档中是否有特征重要性的生成方式，是否有决策依据；检测算法实现方式与设计文档是否一致；通过测试，查看特征排序是否符合业务逻辑。

模型预测结果分布可解释：模型预测结果可解释性的评估需要利用全量测试集，查看模型预测结果是否符合正态分布或者业务常见的分布形式；抽样进行测试，查看模型预测结果是否会出现极端分布的情况。

（3）算法性能

算法性能评估的指标包括四个方面：模型平均响应时间、模

型并发能力、模型吞吐量、模型稳定性，算法性能指标评估需要查看设计文档，是否包含有模型服务平均响应时间、模型吞吐量、模型稳定性等的指标说明；抽取一定样本，进行服务压测，测试结果与设计文档的说明进行对比，判断模型性能是否达标。

（4）算法健壮性

极端和对抗样本防范能力：评价模型在极端样本测试情况下，是否拥有容错性；评价模型在对抗样本攻击下，是否具有防范能力。算法健壮性的评估需要查看设计文档，是否包含有正常样本和异常样本的定义，否有根据业务应用场景需求制定的对抗样本攻击的防范措施；抽取一定异常样本，检查服务返回是否正常；采用基于 FGSM、DeepFool、C&W、JSMA 等算法生成对抗样本进行攻击测试，检查算法服务是否有针对对抗样本的防范能力。

对噪声的过滤能力：训练数据或者测试数据中加入噪声后，可能导致模型效果出现下降。评估算法对噪声的过滤能力需要检查规划方案或实施方案中对过滤检测数据集中的噪音和异常值是否明确规定了采用的方法；核查系统中采用的方法与规划方案或实施方案的一致性；执行测试，核查系统中采用的方法能有效检测并过滤数据集中的噪声和异常值。

（5）算法精准性

算法准召率：对于不同的业务场景，应该选择不同的评估方

式。通常的指标评估需要根据业务场景选择模型的评估方式，比如在风险欺诈领域，召回率应该是一个更加重要的指标；准备合理的测试数据进行模型的推理测试获得模型效果的评估。模型在不同的场景下的通过标准也不同，例如在反欺诈场景下，模型识别欺诈者的准确性需要定得非常高，以防止现实中的欺诈行为发生。

算法效果稳定性：模型预测能力的稳定性体现在模型的预测能力在时间维度上是一致的，即模型在测试集、时间外样本集、线上测试和正式使用的时候有同样的区分度。评估设计文档中对于模型效果稳定性是否有相应的保障措施；检测模型效果稳定性保障实现是否与说明文档一致；准备合理的测试数据进行模型的推理测试获得模型效果的评估随时间的变化，是否能够保持稳定。

算法泛化能力：算法对新样本的适应能力评价，具有对于未知样本也可以做到很好预测的能力。评估需要检查设计文档中对于模型是否有相应的保障措施；检测模型稳定性保障实现是否与设计文档一致；准备新的测试数据测试，获得模型效果的评估，是否存在欠拟合或者过拟合的情况。

（6）算法可追溯性

训练数据可追溯：人工智能算法的发展离不开数据，但是训练数据来源需要合理合法，需要具有可追溯性。评估需要检查算

法训练系统是否对训练数据获取时间进行了记录，设计文档中是否有记录训练数据来源的相关内容，是否有记录训练数据量的相关内容，是否有数据采样方法的记录；检查算法训练系统是否有训练数据采样方法记录。

建模过程可追溯性：外部刺激和内部技术因素等都会引发人工智能失控。为了尽可能大降低误判和失控风险，那么需确保算法过程可追溯、可分析。评估需要检查设计文档中是否有详细的建模过程描述，是否记录了建模过程中使用的软硬件环境和建模过程操作者，是否按要求记录建模的起止时间戳和迭代次数的相关内容；检查算法训练系统是否记录了建模的起止时间戳和迭代次数。

算法部署可追溯性：算法部署可追溯是指算法部署的操作者、操作时间、部署脚本等都有记录。评估需要检查部署文档中是否有要求记录人工智能算法部署操作者的相关内容，是否记录了部署的软硬件环境配置信息；检查系统是否标识了部署时间及相关结果，是否有对保存人工智能算法模型部署相关脚本的说明。

（7）算法公平性

算法效果公平性：人工智能算法都是基于数据驱动的，数据采集过程中可能存在不同群体数据占比权重不均衡。基于不均衡的数据进行训练，算法可能在无意间编码人类的偏见而对个人或

群体产生偏见或歧视，造成了算法决策的不公平。算法公平性需要评估不同群体之间算法的识别效果是否存在显著差异，以此来保障算法决策的公平性。

3、服务能力评价

（1）基础设施

运行时隔离评价： 人工智能服务运行时需要隔离，即两个或两个以上的服务或网络在断开连接的基础上，实现信息交换和资源共享，也就是说，应通过运行隔离技术既可以使两个网络实现物理上的隔离，又能在安全的网络环境下进行数据交换。

运行时保护评价：运行时的保护需要制定健全的管理制度和严格管理相结合。评价制度的制定和落实情况是否能保障计算、存储、网络的安全运行，使其成为一个具有良好的安全性、可扩充性和易管理性的信息网络。

（2）系统安全性

人工智能系统安全性问题与传统计算机安全领域中的安全问题相似，威胁着人工智能技术的保密性、完整性和可用性，评价的方式方法相同。人工智能系统安全问题主要分为硬件设备安全问题以及系统与软件安全问题两类。硬件设备安全问题，主要指数据采集存储、信息处理、应用运行相关的计算机硬件设备被攻

击者攻击破解,例如芯片、存储媒介等。系统与软件安全问题,主要指承载人工智能技术的各类计算机软件中存在的漏洞和缺陷,如承载技术的操作系统、软件框架和第三方库等。

（3）系统应用性能

训练性能测试方法及指标：系统的性能取决于系统硬件计算能力、软件框架和模型及其实现技术。单独的算法或模型，一般利用复杂度（complexity）从理论上表征算法运行可能消耗的时间或/和空间。给定复杂度的算法/模型，运行于不同的计算系统，性能则可能不同。人工智能训练过程的性能，一般是指以特定人工智能计算系统训练某特定模型时，所使用的时间、能量以及获得数据吞吐率等。这里的时间，可描摹整个训练过程（如端到端）及其细部时间（某个 epoch 耗时，数据预处理时间，分布式训练参数同步时间等），能量可使用训练过程的系统整体平均功率来表征，也可分离系统空载功率，来衡量系统真正用于训练的那部分功率。数据吞吐率一般是指单位时间内消耗的训练数据样本数，如 images/s，元组数/s 等。它表征计算系统的有效计算能力。

推理性能测试方法及指标：系统推理的性能取决于硬件、软件框架和模型。在某些情况下，可在无框架下直接运行模型，即将模型“压”入硬件内存，并在此过程中，实施模型优化（如剪枝、合并算子、内存使用优化等）。推理指标与训练指标相似，也以所使用的时间、能量以及获得数据吞吐率等表征。其中推理用

时，可指推理整个测试集的用时，也可指某些特定样本的用时。在不同的作业达到模式和压力下，系统的推理用时可能有所变化，推理的准确率与模型本身的特性有关。推理的能量消耗可使用推理过程中的平均功率来衡量。数据吞吐率一般是指特定系统在单位时间内处理的样本数量，表征计算系统的有效计算能力。

（4）运维管理能力

需要对模型系统所依赖的基础设施、基础服务、线上业务进行稳定性加强，发现人工智能服务可能存在的隐患，对整体架构进行优化以屏蔽常见的运行故障，多数据中心接入提高业务的容灾能力。评价通过监控、日志分析等技术手段，是否能及时发现和响应服务故障，减少服务中断的时间。同时需要关注业务运行所涉及的各个层面，确保用户能够安全、完整地访问在线业务。对业务进行各方面优化，确保公司业务数据和用户隐私数据的安全，并保证服务具备抵御各种恶意攻击的能力。

二、问题挑战与解决思路

（一）人工智能金融应用的问题挑战

人工智能金融应用面临的挑战可概括为人工智能数据安全、人工智能算法安全、人工智能伦理治理、以及人工智能应用评价四类挑战。

在人工智能数据安全方面，数据是人工智能算法的源泉，应从源头防范化解人工智能算法安全风险。随着技术应用的深入，数据非法盗用、数据恶意攻击、数据泄露以及数据滥用等风险事件频发，暴露出人工智能应用在数据隐私保护、全生命周期数据渠道管理等方面的不足。

在人工智能算法安全方面，出于技术的复杂性以及商业竞争力的敏感性，部分算法缺少对决策过程的合理解释以及关键信息的公开披露，引发算法黑箱化，极大程度降低算法应用的可信度，直接影响最终使用者做出正确决策，甚至造成不可逆的经济损失。

在人工智能应用评价方面，《人工智能金融应用算法评价规范》为金融行业提供了首个通用性评价依据，解决了行业“从无到有”的问题，但在金融场景的细分需求上，仍需建立细化评价标准及评价指标以满足差异化需求。同时，算法技术的更新迭代对评价工具、评价方法的有效性提出更高要求，需要紧跟技术前沿不断深化评价手段的创新。

在人工智能伦理治理方面，近年来人工智能技术的滥用引发大数据杀熟、智能推荐信息误导以及人工智能偏见歧视等问题，影响正常的市场秩序和社会秩序，给维护意识形态安全、社会公平公正和用户合法权益带来挑战。急需在伦理治理实践指南、伦理风险识别工具以及道德伦理评价等方面进一步完善。

（二）人工智能应用评价的问题挑战

目前金融行业不同机构间人工智能测试方面较为封闭，单一的检测技术无法实际应用到多类不同的应用场景上，检测技术的复杂度较高，检测技术方法需要根据场景应用相关的特性进行细化，检测工具和检测环境还不健全。因此，目前部分可解释性、安全性的技术评价以自声明结合评价为主。

考虑现阶段标准和技术手段不足，从评价对象来看，对人工智能金融应用评价主要以算法评价为主，场景应用相关的特性评价指标有部分覆盖但仍需细化完善；从评价方法来看，可解释性、安全性的技术评价以企业自声明结合评价为主，后续随着技术进步、系统化的评价体系建立，测评工作中的技术水平将得到提高，并进一步完善自声明审核的标准及相关依据。

通过技术手段对人工智能金融应用开展评价，是一套复杂的系统性工程活动，需基于安全性、可解释性、精准性、性能乃至道德伦理等多个维度，包括数据、算法、算力、场景等多种要素，涉及产业方、应用方、独立第三方、监管方及自律组织等多方力量。此外，人工智能金融应用评价工作需要相关工具以及方法不断地更新迭代，而对于精准性和性能虽有部分技术检测手段支撑，但还需在数据、评价基准等基础建设方面予以储备。

（三）解决思路与方法

为有效应对上述问题和挑战，秉持“公正向善、安全可控”为目标的解决思路，通过建立多元化评价体系、提升检测认证专业化能力、探索算法备案及信息披露自律机制，全面贯彻落实人工智能金融应用发展规划和治理要求。

建立多元化评价体系：根据国家建立算法安全治理机制、完善监管体系、规范算法生态的治理目标，金融业有必要建立一个符合人工智能技术发展需要、满足金融服务特点的多元化评价体系，来配套国家行业人工智能安全治理机制的建立，按照行业管理的要求引导人工智能金融应用朝着公正向善、安全可信方向健康发展。

提升检测认证专业化能力：金融业需要深入研究人工智能的测评技术，建立科学的方法论，强调公正向善，持续优化和丰富贴合金融应用场景的评价规则、评价方法、评价技术、评价工具等，依照技术和应用水平分阶段有效地支撑人工智能金融应用评价体系的构建。

探索算法备案及信息披露自律机制：一是依托人工智能金融应用规则、标准规范及行业公约，适时建立监管部门、行业协会、从业机构、检测机构、认证机构多方行业主体协同联动的人工智能金融应用相关算法备案机制。二是研究建立适当的人工智能信

息披露制度，考虑算法设计、研发、运行中可能存在的偏见和漏洞、数据来源合法合规性问题以及可能对个人和社会造成的潜在危害，针对不同业务场景及可能的风险程度明确相应的算法披露要求。三是探索建立人工智能伦理审查机制，将人工智能伦理原则要求细化为针对产品和服务的技术标准及相关自评估清单、风险评估表，引导机构发挥主动性，把以人为本、公平公正、权责清晰等伦理要求贯彻到业务规划、技术应用、产品研发等金融科技活动全过程，坚决杜绝伦理失范现象。

三、建立多元化评价体系

为支撑金融行业人工智能应用安全治理机制，建议通过标准化评价指标体系、统一测试评价方法、规范测试测评技术，打造形成政府监管、企业履责、行业自律、质量监督的多元化评价体系，即政府通过顶层设计制定人工智能金融应用的国家行业政策并引导行业发展，企业开展自查自纠履行社会责任，行业协会发挥自律约束的作用，检测认证机构充分发挥权威第三方作用开展质量认证审查活动。

（一）评价体系阶段建设

现阶段金融行业依据《人工智能算法金融应用评价规范》从安全性、可解释性、精准性和性能方面开展人工智能算法评价工作，人工智能金融应用评价体系已初具雏形。然而，评价体系的

建设不是一蹴而就的，随着技术的发展、应用场景的创新以及标准体系的完善，评价体系的建设也应是一个循序渐进的过程。因此，具体建议按探索、扩展和成熟三个阶段分布实施，具体内容如下表所示。

表 2 评价体系阶段建设

发展阶段	评价规则	评价方法	评价技术	评价工具
探索阶段	国家和行业政策要求，行业通用标准规范	自声明与验证相结合； 技术测试与技术评估相结合	基于目前技术应用水平，充分采用已有安全攻防、性能和精准性测试技术手段； 基于模型优化等实践经验应用于可解释性的评价技术	工具场景耦合度高； 工具类型单一； 专用工具缺乏
扩展阶段	部分成熟应用或高风险应用场景的评价标准； 实用性评价工作实施指南等	以标准符合性验证为主； 技术测试为主，测评结合	基于长期积累的最佳实践形成的安全攻防、性能和精准性测试技术手段； 逐步丰富的漏洞扫描技术； 规范统一的多纬度可解释性评价技术	专用工具丰富； 场景无关的通用工具； 可比较性的测试基

				准
成熟阶段	覆盖各种金融应用场景的评价指标体系；完善的评价实施指南等	基于统一规范的技术测试要求，自动化的测试技术支撑系统化的评价方法	基于风险控制的安全攻防技术评价体系；具有可比性且基准一致的性能和精准性测试技术手段；系统性支撑的漏洞扫描技术；基于理论支持的多纬度可解释性评价技术等	体系化成套工具；体系化的数据、工具集、测试基准等基础设施。

探索阶段主要是起步推广，即在方法、技术手段相对不足，贴近场景的评价规则还不完善等情况下，基于业界技术应用现实情况，通过调研访谈、技术检测与企业自声明相结合的方式开展综合评价工作。

扩展阶段主要是持续优化，这是在技术进步、系统化的评价体系逐步建立，检测工作技术手段日趋成熟的情况下，通过优化自声明审核的标准及相关依据，依托最佳技术实践的测试工作，测试与评估相结合的评价方式。

成熟阶段主要是系统化普及，这是在金融应用场景全覆盖的评价指标体系基础上，结合风险控制的思路，采用具有成熟理论支撑的体系化技术手段，以自动化和系统化的检测方法开展评价

工作。

（二）评价标准体系建设

图 5 金融行业人工智能评价标准体系

由于人工智能技术的赋能属性使其与场景结合更加紧密，建议在国标委人工智能标准体系框架的基础上规划金融行业人工智能评价标准体系框架，贴近金融应用场景，引导科技向善控制创新风险，建立从科技伦理风险控制、技术风险控制到行业应用风险控制不同层次，覆盖数据、模型、系统且能体现金融服务特色的应用评价标准体系。建议依据金融行业人工智能技术应用风险开展分类分级管理，优先或重点围绕“科技向善，安全可控”等配套人工智能安全治理需要的标准建设，重视高风险的金融应用场景，支持行业人工智能技术应用痛点难点的解决。

（三）检测能力全面提升

完善人工智能算法测评方案

当前人工智能算法测评工作展开的范围和深度还比较有限，行业经验积累也比较有限；另一方面人工智能算法的发展也极为迅猛，更新迭代速度快。因此，基于《人工智能算法金融应用评价规范》JR/T 0221-2021 衍生的当前测评方案，还有较大的完善空间，在落地层面仍有很多地方值得探索。

在未来的工作中，金融业内认证机构与检测机构应联合技术厂商、金融机构、行业内外部人工算法应用单位一同，根据试点和前期定制测试积累的实际经验，进一步完善人工智能算法测评方案。方案完善的方向主要有测评项的完备性、测评项的必要性、测评项的分类分级、测评通过标准的公平性与正确性、测评方法的可实施性和准确性等等，这些方向的完善需与人工智能算法演进的方向一致。

重点检测技术攻关

测评的准确性极大的依赖重点检测技术，技术的完备性和有效性决定了测评的公平程度和推广程度。目前看来，人工智能测评在一些检测技术方面还有待提升。

在安全方面，对算法的系列攻防对抗技术和漏洞扫描技术需

要进一步攻关。例如窃取攻击、药饵攻击、闪避攻击、模仿攻击、逆向攻击、供应链攻击、后门攻击的攻击技术方法，有的需要解决从无到有的问题，有的需要解决由浅入深的问题。对于广泛的漏洞扫描能力，其不同于传统的软件漏扫能力，需要结合算法特性定制研发能力足够的漏洞技术。目前安全角度的检测技术能力提升是具有挑战性的，有很长的道路要走。

在可解释性方面，在完善测评方案的基础上，需要有更好的测评维度来评价算法的可解释性问题，这一方面需要有理论层面的突破，另一方面实现从理论到实践的技术方法跨越。

在精准性与性能方面，虽然相关测评理论相对成熟，但需要用完善的技术能力对人工智能不同算法类型、不同场景的公平测评技术方法，一方面需要自动化的检测工具，一方面需要标准化成熟数据集。

统一提升检测工具能力水平

检测工具是实施测评工作的重要辅助手段，一方面能够提高检测效率，一方面可以减少人为误差导致的测试主观不公开问题。

虽然一些机构具备了某些点上的测试工具，但由于人工智能算法测评还处于探索阶段，该领域尚无成熟公认的统一检测工具。

未来，需根据成熟的检测方案，研发统一的检测工具，尤其

针对安全性攻击工具、精准性与性能测试工具上需下大力度推进研发工作，这些领域工具的成熟也将标志人工智能算法测评工作进入一个相对成熟稳定的阶段。

建设完善标准化测试数据集

数据样本是测试人工智能算法必不可少的环节，数据样本的质量直接关系到测评的有效性和准确性。

目前金融领域人工智能算法的测试还未能建立起统一的标准化测试数据集，影响着测评工作的大范围开展。未来可通过检测认证机构与技术厂商联合的方式，在合理合法合规的情况下建设标准化测试数据集，覆盖不同算法类型、不同应用场景、不同测评目标的高质量、大体量数据集，促进人工智能算法测评工作迈上新的台阶。

四、发挥认证价值 助力人工智能治理

目前，我国已经开展了信息安全产品认证、服务认证、管理体系认证和信息安全从业人员认证，而关于人工智能金融应用相关认证，存在迫切而明确的社会需要。因此，服务于金融业的主要检测认证机构在人民银行科技司的指导下，结合人工智能金融应用各方需求，已启动人工智能金融应用认证认可制度的建立和试点工作。

（一）建立“产品+服务”双认证体系

前期，北京国金认证牵头组织成立了人工智能算法金融应用认证工作组，来自认证机构、检测机构、商业银行、第三方支付机构、金融科技企业等相关单位的二十多位专家参与。结合问卷调研和实地走访情况，工作组经过多次讨论，初步设计出“产品+服务”的双认证模式。其中，产品认证重点关注算法本身的算法建模准备、算法建模过程、算法建模应用等全生命周期，服务认证聚焦通过智能信息系统向金融客户提供的服务质量，侧重于客户的感知和体验，保障金融服务的有效供给且无偏见、无歧视。检测认证实施采用查阅材料、查看系统、访谈人员、系统测试、攻击测试、算法测试和查看算法等方式进行，共计 166 个审查项，从安全性、可解释性、精准性和性能等方面开展人工智能算法评价，确保认证范围和认证对象的全覆盖。检测认证使用专业工具，通过对目标系统的扫描、探测等操作确认响应结果，并利用专业攻击方法对 AI 算法模型进行攻击，同时基于业务样本数据对目标变量进行预测，通过结合自声明验证的方式确认算法是否满足评价指标。体系建设方面，工作组已初步形成人工智能算法金融应用认证实施细则、检测方案、审查列表和评估准则等支撑文件体系。

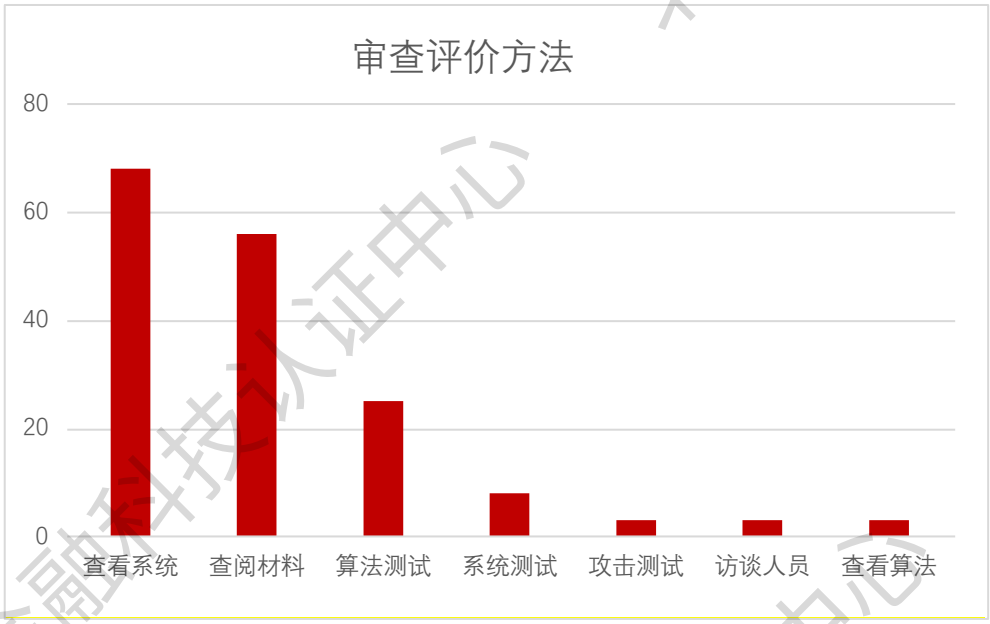

图 6 审查评价方法统计

（二）开展人工智能金融应用认证试点

在人民银行上海总部、北京营管部、杭州及宁波中支的大力支持下，我们调研了金融机构正在应用的十余个人工智能算法项目。在经过对各项目的机构代表性、算法典型性、场景覆盖广泛性、金融产品创新性等多重指标的全面考量后，最终在国有大型商业银行、城市商业银行、第三方支付机构中各选取一个具有代表性的机构开展人工智能算法金融应用项目试点，其中两个试点同时也属于金融科技创新监管试点。北京国金认证、CCRC 作为认证机构，BCTC、CFCA、中国软件测评中心作为检测机构，华为、百度、腾讯、银联商务作为技术支撑机构参与工作。

通过前期的调研和试点的开展，我们发现金融领域人工智能算法应用最多的是智能风控和贷前反欺诈两大金融业场景，具体

算法包括 LGBM、XGBoost 等。每个金融场景往往应用多个算法，每个算法又同时支撑多个场景实现，检测认证需要抽丝剥茧，层层深入逐个审查点，应用专业能力和测试工具验证算法的内在逻辑、实现路径、决策过程、预期目标等。在此过程中同时发现，各金融机构目前对《JR/T 0221—2021 人工智能算法金融应用评价规范》的对标达标尚不到位，多数机构设计文档中缺乏避免偏见歧视的可解释性说明。此外，人工智能算法的发展速度快、迭代周期短，对检测认证手段的先进性和证后监督的及时性都带来了一定挑战。

当前，各试点已进入中期实施及逐步收尾阶段。试点的开展首先促进了相关机构对标达标的有效性，推动金融机构增加训练数据集的分布验证，加强对目标函数和选择特征避免偏见歧视的研究；其次对金融场景应用广泛的算法（如 XGBoost、LightGBM 等）进行了重点检测和审查，对标准落地的验证形成了经验和实践积累；再次，提升了检测认证工具水平和方法，储备和提高了检测认证能力以匹配人工智能算法等前沿技术的迅猛发展；最后，聚焦了算法科技伦理问题，对机构入模参数特征分布等进行了重点分析，以应对算法歧视等问题。

（三）推动认证结果多方采信

作为一种权威第三方评价活动，认证具有传递信任的重要使命，对于认证结果的采信彰显了认证的社会价值，有助于行业的

深化治理。因此，建议充分运用认证行为与认证结果采信两种手段，共同建立基于信任的人工智能金融应用生态。一方面，建立金融行业人工智能产品及服务准入机制。另一方面，采用认证认可手段进行准入资质管理。再者，形成政府引领、产业参与、多方采信的共建共治共享格局。

综上所述，依托双认证体系和多方采信机制发挥认证价值，能够有效提高金融行业人工智能治理效能。对于监管部门，通过认证采信机制推动政策落地及标准应用；对于行业自律组织，探索基于算法备案和信息披露的伦理审查自律机制；对于金融机构，提供服务质量安全保障；对于产业机构，借助认证采信提高市场竞争力；对于金融消费者，借助认证采信增强安全感、获得感和幸福感。

参考文献

[1]曹和平.坚持科技向善，实现包容性增长[N].中工网,2021-09-28

[2]李兴锋.金融业数字化转型四项修炼[N].金融电子化杂志,2021-09-18

[3]范一飞.中国（北京）数字金融论坛讲话稿.中国人民银行,2021-09-10

[4]JR/T 0221-2021 人工智能算法金融应用评价规范