

作者：赵恒、李之行、张怀玉

邮箱：

## 人工智能技术在信用评级领域的应用研究

## 摘要

随着人工智能技术的飞速发展，其深度赋能效应正加速渗透至制造、医疗、交通、金融等几乎所有行业，推动着全球经济和社会形态发生前所未有的系统性变革。信用评级作为金融体系风险管理和资源配置的重要基石，也正处在这场智能化浪潮的重要地带。近年来，部分信用分析师已在日常工作流程中尝试应用人工智能工具，行业智能化转型初见端倪。然而，这种探索性应用的广泛兴起，也暴露出一系列亟待破解的现实难题。一方面，AI 应用的合规性边界模糊，现有评级流程和监管框架尚未充分应对 AI 的深度介入；另一方面，行业整体应用效能还有较大提升空间，目前 AI 技术主要停留在基础的数据处理和文本生成阶段，尚未充分释放其在风险洞察与前瞻性分析等高价值环节的潜能。

随着 AI 技术渗透日益加深，全球各国监管机构也在加快应对步伐，对数据安全、算法透明度与模型伦理等提出了更高的治理要求。以中国《生成式人工智能服务管理暂行办法》（2023 年）、欧盟《人工智能法案》（2024 年）等为代表的新一轮政策落地，使得信用评级行业在智能化转型的同时，也面临着更加严格和清晰的合规基线。这一技术创新加速与监管规范升级并行的行业变局，使得如何在合规框架下最大化释放 AI 价值、同时有效管控其潜在风险，成为金融行业必须直面的核心课题。

本研究将首先描绘人工智能在信用评级全流程中的广阔应用图景，并直面其伴随而生的核心挑战；进而，本文将构建并提出一套以可信 AI 治理为核心的系统性解决方案。在此基础上，研究将进一步展望未来行业的技术演进与格局重塑，最终凝练核心结论与政策建议，力求为监管机构、评级从业者及技术开发者，提供一份在合规浪潮下实现负责任创新的行动蓝图与实践指南。

## 一、研究背景与意义

正如引言所述，人工智能正驱动着一场深刻的产业变革。要理解这场变革的本质，首先需要厘清其核心的技术构成与能力边界。当前的人工智能技术体系已高度复杂化，其中，大语言模型（Large Language Model, LLM）以其强大的自然语言生成与理解能力，成为人机交互和内容创作的关键；机器学习（ML）特别是深度学习，则通过从海量数据中学习复杂模式，为预测与决策提供了核心算法支持；而图神经网络（GNN）与知识图谱（KG）则善于挖掘实体间的关联关系，为深度风险传导和关系网络分析提供了可能。这些技术共同构成了从感知智能（如图像识别）向认知智能（如逻辑推理）与决策智能（如自主优化）跃迁的关键驱动力，为重塑传统行业提供了前所未有的技术工具箱。

在众多应用领域中，金融业作为数据密集型行业，已成为人工智能应用的天然沃土，信用评级领域的融合更是大势所趋。这种融合体现为一种深刻的双向赋能关系。一方面，信用评级为 AI 技术提供了高质量的垂直应用场景和海量结构化、非结构化数据，为 AI 模型的训练与优化提供了重要的数据基础。信用分析师在日常工作中对 AI 工具的应用与反馈，也在不断推动着人工智能技术在复杂金融风险识别这一垂直领域的加速成熟与迭代。另一方面，人工智能正从三个维度系统性地重塑信用评级行业。首先，它深刻地重塑了分析师的工作范式。AI 正成为信用分析师进行高效分析的辅助工具，全球领先的评级机构如穆迪、标普均已试点部署智能评级辅助系统。这些系统利用 AI 自动采集数据、处理非结构化信息、识别风险因子并辅助报告生成，从而将分析师从繁重的重复性劳动中解放出来，使其能够专注于更具价值的复杂风险判断。其次，人工智能正在重塑整个行业的数据生态。其强大的多源数据处理能力，打破了传统评级对历史财务报表的过度依赖，能够将舆情、供应链乃至各类另类数据有效纳入评估体系，推动行业数据生态向更全面、更动态、更实时的方向演进。最终，这些变革共同重塑了行业的核心风控能力。通过深度学习与预测建模，AI 显著提升了风险识别的精度与前瞻性，使信用评估从事后的回顾式诊断，升级为事前的前瞻式预警，从根本上提升了信用风险管理的有效性。

尽管人工智能带来了巨大的发展机遇，但其深度应用也伴随着一系列系统性的挑战。行业在实践中普遍面临数据隐私保护与合规边界模糊、模型因其复杂性导致可解释性不足、算法可能存在的伦理偏见等严峻问题。这些挑战不仅制约了 AI 潜能的完全释放，更对评级结果的公信力与金融市场的稳定性构成了潜在威胁。因此，如何在拥抱技术红利的同时，构建稳健的“人机协同”模式与“可信 AI 治理”框架，已成为行业亟待破解的核心矛盾。

为系统性地回应上述矛盾，本文旨在为人工智能在信用评级领域的负责任创新提供理论框架与实践路径。第二节将首先描绘 AI 在评级全流程中的应用图景。第三节将深入剖析其面临的核心挑战。第四节将针对性地提出以“可信 AI 治理”为核心的系统性解决方案。第五节将前瞻性地展望未来发展趋势。最终，第六节将在首尾呼应研究背景的基础上，凝练核心结论与政策建议。本文力求通过这一完整的研究路径，为行业提供一份兼具理论深度与实践指导价值的参考。

## 二、人工智能技术在信用评级中的应用场景

针对前述的传统信用评级瓶颈，人工智能技术，特别是大语言模型（LLM），凭借其强大的自然语言处理（NLP）、知识整合与自动化生成能力，正驱动信用评级流程向智能化、自动化转型。基于业务流程，人工智能正系统性地重塑信用评级从客户接洽、信息获取、风险研判到价值交付的全流程价值链。大语言模型已逐步渗透至业务流程的多个核心环节，展现出其在提升效率、规范流程与辅助决策方面的重要价值。如表 1 所示，人工智能技术已全面渗透信用评级的多个核心环节，不仅提升了各环节的自动化和规范化水平，也为信用评级流程注入了数据驱动和智能决策的新动能。

表 1：人工智能技术在信用评级中的应用场景及作业环节

| 应用场景 | 作业环节      | 描述                             |
|------|-----------|--------------------------------|
| 客服沟通 | 商务沟通      | 利用大模型提供业务、产品、政策等智能答疑，提升客户服务效率。 |
|      | 征求意见      | 利用大模型解答受评主体对报告及意见的疑问并收集反馈。     |
| 尽职调查 | 调查提纲设计    | 利用大模型生成和优化访谈提纲，辅助调查人员设计问题。     |
|      | 资料收集与整理   | 利用大模型辅助整理公开信息和访谈记录，提升资料收集效率。   |
| 分析评估 | 宏观及行业环境分析 | 利用大模型整合宏观和行业数据，辅助识别环境趋势与风险。    |
|      | 定性及定量指标评估 | 利用大模型辅助进行定性和定量分析，生成指标解读与结论。    |
| 报告撰写 | 辅助报告撰写    | 利用大模型辅助起草评级报告，提升撰写效率和质量。       |
|      | 辅助审核      | 利用大模型辅助报告内容审核，自动识别问题并给出改进建议。   |

资料来源：公开信息，远东资信整理

（一）客户沟通中的人工智能应用

人工智能可用于商务沟通环节。在传统信用评级业务中，大量的客户咨询集中在流程指引、材料清单、政策依据等重复性较高的事务上，人工客服往往需要投入大量时间与精力。大语言模型通过自然语言处理和知识整合能力，可承担基础性业务咨询与流程指引的任务。例如，当客户咨询某类政府债券的评级流程时，AI 系统能够即时生成涵盖政策摘要、所需材料清单及标准化流程说明的回复。这种智能化交互不仅显著减轻了人工客服的重复劳动，还能够提高响应效率与客户满意度。然而，由于信用评级业务涉及合规性和政策敏感性，AI 生成内容必须建立严格的准确性与合规性监控机制，避免因错误引导而产生声誉与法律风险。

同时，人工智能可用于征求意见环节。在评级结果披露与报告发布后，客户往往会对部分指标、风险评估或结论提出异议与反馈。人工处理这些反馈不仅耗时，而且容易出现遗漏。人工智能可通过自动化文本分析与语义分类，对客户意见进行收集、归类与结构化整理。例如，当客户对某一风险评级结果表示质疑时，系统可自动将其归类并汇总至“异议反馈表”，形成可追溯的意见数据库。这不仅节约了人工整理成本，还为评级模型的迭代优化提供了数据支持，有助于增强评级体系的透明度与客户信任度。

AI 在客服沟通环节的应用能够有效提升客户响应速度与服务体验，同时在意见收集与反馈整理方面实现流程自动化。但其核心挑战在于如何建立完善的质量控制与合规审查机制，确保输出结果的准确性与权威性。未来，随着 AI 技术与信用评级流程的进一步融合，客服沟通环节有望成为信用评级机构提升客户关系管理与数据积累的重要入口。在这一模式下，人类分析师的角色也从重复性的事务应答，转变为处理复杂异议与维护关键客户关系的高级顾问。

## （二）尽职调查环节的智能化辅助

人工智能可辅助调查提纲设计。尽职调查是信用评级的基石，其核心在于对受评主体的财务状况、治理结构、行业环境及潜在风险进行全面、系统的考察。然而，传统尽调准备往往依赖分析师的经验与人工梳理，耗时长、覆盖面有限。借助人工智能，评级机构能够实现调查提纲的自动化生成。模型可基于受评主体的基础信息（如行业特征、企业规模、业务模式）进行语义分析，并智能生成涵盖财务指标、治理风险、运营效率和外部依赖度的访谈问题清单与资料目录。例如，在新能源企业的尽调中，AI 可自动识别“供应链稳定性”“政策补贴依赖度”等关键议题，并将其纳入提纲。这种智能化方法显著提升了调查准备的全面性和针对性，但其生成结果仍需人类专家进行专业审查，以确保问题设计的合理性与适用性。

人工智能可辅助资料收集与整理。传统尽调资料收集过程面临信息来源分散、数据结构不一及信息冗余等挑战。人工智能在此环节的价值主要体现在三方面：人工智能问答界面，引导客户按需提交所需材料，避免遗漏与重复；自动化检索和抓取受评主体在公开渠道的信息资源，如工商登记、涉诉与仲裁记录、招投标数据及行业研究报告；对多源异构数据进行初步清洗、归类与摘要提炼，形成结构化、标准化的数据包供分析师进一步处理。例如，在整理企业公开披露的财务信息时，AI 能够自动识别并标注“缺失的分部门利润数据”，从而提示分析师关注潜在的信息缺口。这不仅提高了资料处理效率，也增强了数据完整性与风险识别能力。

人工智能在尽职调查环节展现出显著的赋能潜力：一方面，它通过调查提纲设计实现了问题覆盖的系统化与针对化；另一方面，它在资料收集与整理中有效解决了多源信息碎片化的难题。虽然其应用能够提升效率并扩展分析视角，但由于信息来源的可靠性与 AI 模型的判断局限，人工专家的交叉验证仍是确保调查结果真实可靠的关键。未来，随着语义搜索、知识图谱等技术与评级业务的深度结合，尽调环节的智能化程度有望进一步提升。相应地，分析师的工作重心从繁琐的资料搜集升级为基于 AI 提供的全面信息进行深度访谈与关键风险验证。

## （三）分析评估环节的深度赋能

宏观及行业环境评估是信用评级的重要环节，其中外部环境分析直接决定了对受评主体风险状况的整体把握。传统方法依赖分析师逐一查阅宏观经济数据、产业政策文件及行业研究报告，费时费力且难以保证全面性。人工智能能够快速整合多源权威信息，包括宏观经济指标、政策导向、行业周期与竞争格局，并生成连贯、逻辑清晰的分析文本。例如，在光伏产业评级中，人工智能可整合能源转型政策、全球供需数据及主要厂商竞争态势，自动生成“行业景气度与政策风险评估报告”，帮助分析师在短时间内获得外部环境的整体图景。这种能力不仅提

升了分析效率，也有助于实现对系统性风险的前瞻性识别，但其结论的可靠性仍高度依赖于所接入数据的权威性与时效性。

同时，在具体的信用评级过程中，分析师需要对受评主体的管理能力、技术优势、财务状况等多维度指标进行综合评估。人工智能能够依据评级方法论的指标定义，对定性与定量指标进行初步处理与分析。一方面，其非结构化文本信息（如企业年报、舆情数据、管理层访谈记录）转化为半定量指标，从而为管理层能力、战略执行力等难以量化的要素提供数据化支撑；另一方面，人工智能可结合财务报表与历史数据自动计算关键财务比率，并与行业平均水平进行对比，生成分析草稿和初步结论。例如，人工智能可基于企业的资产负债率和现金流数据生成偿债能力分析草稿，并对照行业基准提出风险提示。这不仅提高了评级的一致性和客观性，也为分析师最终打分提供了逻辑参考。然而，由于模型在指标权重分配与算法透明性方面存在潜在不足，其生成的结论仍需由人工专家进行审查与修正。

人工智能在分析评估环节展现出从宏观到微观的全方位辅助能力：既能快速整合外部环境信息，又能在定性与定量指标处理上提供初步结论。这一过程显著提升了评级效率与分析覆盖面，有助于减少主观偏差。但由于其结论高度依赖数据源质量与算法透明度，人工校验与专家审议仍是保证评级结果科学性与公信力的关键。这使得分析师能够从数据处理和初步计算中解放出来，专注于融合隐性经验和前瞻性判断的核心风险决策者。

#### （四）报告撰写与审核的自动化提升

信用评级报告是评级结论的直接载体，其撰写过程不仅涉及对公司基本情况、业务模式、财务表现和行业风险的系统性阐述，还要求具备高度的逻辑性与专业性。传统报告撰写往往耗费分析师大量时间在材料整合和文案组织上，容易分散其在核心风险判断上的精力。人工智能可基于前期输入的结构化数据与分析要点，自动生成信用评级报告的各章节草稿，如“公司概况”“业务与治理分析”“财务与偿债能力分析”等。以某能源企业评级为例，人工智能可自动生成一份涵盖“行业政策环境、财务杠杆水平、现金流压力”三方面的分析草稿，帮助分析师快速搭建报告框架。这不仅显著减轻了文案负担，还提升了报告撰写的规范化程度。然而，其生成的内容在事实准确性、深度解读与风险敏感性上仍存在局限，必须由人类分析师进行复核与修正，以确保其专业性与权威性。

人工智能还可辅助报告审核。在报告审核环节，人工智能可作为质量控制的“第一道防线”，通过自动化审查机制识别报告中的逻辑不一致、数据前后矛盾、格式错误及表述不当等问题。例如，当文本叙述中“净利润同比增长 5%”与表格中“净利润同比增长 2%”出现冲突时，能够即时标注并生成修订批注。此外，工智能还能针对格式标准化、专业术语使用及行文逻辑提出修改建议，从而提升报告审核的效率与准确性。这种智能化预审机制有效降低了人工初审的工作量，但最终审定权仍需掌握在人类专家手中，以确保报告结论的稳健性和对潜在风险的敏锐识别。

人工智能在报告撰写与审核环节的应用极大优化了评级业务流程：既能通过自动生成草稿提升撰写效率，也能通过智能校对强化审核质量。但 AI 在事实判断与风险敏感度上的不足，决定了其角色更适合作为辅助工具而非

决策主体。未来，随着 AI 与知识图谱、智能审计技术的结合，其在报告生成与质量控制中的作用有望进一步增强，为信用评级机构实现标准化、智能化与高效化的运作提供新路径。在此流程中，分析师不再是报告的撰写员，而是利用 AI 草稿专注于核心逻辑、风险揭示和文字精准性的最终质量把控人。

综上，本节系统论述了大语言模型在企业信用评级全流程中的多元化应用场景，涵盖了从客户沟通、尽职调查、分析评估到报告撰写的各个环节。分析表明，人工智能的应用能显著提升评级作业的效率 and 数据处理能力，是实现行业智能化转型的关键驱动力。然而，必须清醒认识到，其在专业深度、风险预见性和可解释性方面仍存在局限。因此，未来的发展并非简单的技术替代，而是走向人机协同的深度融合模式：由人工智能负责处理海量信息、执行标准化任务和提供初步洞察，而人类分析师则聚焦于复杂判断、风险解读和最终决策。这种协同机制将是提升信用评级质量与效率的必然路径。要实现这一理想的人机协同模式，必须首先正视并解决其背后潜藏的一系列严峻挑战，这正是下一节将要深入探讨的核心议题。

三、人工智能技术在信用评级中的挑战

人工智能技术，特别是大语言模型（LLM），正在深度渗透信用评级流程。然而，作为一种新兴技术，其在实际应用中仍面临一系列挑战。这些挑战不仅涉及技术层面，还包括法律、伦理与合规问题。如果缺乏有效的风险防控措施，AI 在信用评级中的应用可能削弱评级结论的科学性与公信力。这些挑战可被归入三个层面。首先是关乎数据与算法本身的基础技术风险，包括数据隐私与安全、AI 幻觉风险。其次是关乎模型内在缺陷的核心模型风险，包括模型透明度与可解释性、道德与合规风险、技术复杂性。最后是关乎应用与监管的应用治理风险，包括任务适配性不足、证据可追溯性缺失。表 2 针对人工智能在信用评级领域应用过程中面临的主要挑战，系统梳理了其潜在影响及对应的治理对策。

表 2：人工智能技术在信用评级中的挑战、影响与对策

| 挑战         | 影响          | 对策                      |
|------------|-------------|-------------------------|
| 数据隐私与安全性   | 法律风险、信誉损失   | 全流程加密、权限管理、合规审计         |
| 模型透明度与可解释性 | 客户信任下降、监管质疑 | 可解释 AI、决策溯源机制           |
| 技术复杂性与维护难度 | 系统不稳定、迭代滞后  | 建立专职技术团队、自动化运维 pipeline |
| 道德与合规风险    | 歧视性评级、法律诉讼  | 公平性检测、偏差校正、多样性数据集       |
| AI 幻觉风险    | 评级结论错误、市场误导 | 交叉验证、权威数据优先、人工复核        |
| 任务适配性与边界不足 | 自动化失败、合规漏洞  | 明确 AI 适用边界、设置人工审核点      |
| 证据可追溯性缺失   | 无法应对审计或诉讼   | 全流程日志记录、审计追踪系统          |

资料来源：公开信息，远东资信整理

## （一）数据隐私与安全性

信用评级涉及大量企业与个人的敏感信息，包括财务报表、合同文本、负债结构、诉讼记录等。一旦数据在采集、存储或传输环节被泄露，不仅会带来严重的法律风险，还可能损害评级机构的信誉与客户信任。例如，若某客户的财务报表在未加密的情况下通过网络传输，被第三方截获并泄露，将直接导致客户商业机密外泄，并可能引发合规诉讼。此外，据 IBM《2023 年数据泄露成本报告》，金融行业平均每次数据泄露成本为 585 万美元。因此，评级机构必须实施全流程加密机制（包括传输加密与存储加密），同时建立严格的权限管理与访问控制系统，以确保敏感信息的机密性与完整性。

## （二）模型透明度与可解释性

深度学习模型常被批评为“黑箱”，这一挑战的根源在于 AI 模型，特别是深度神经网络，固有的“黑箱”性质。这导致其核心问题表现为决策过程的不可解读性，即我们难以用直观的方式理解模型为何会得出特定的评级结论。当信用评级依赖模型结论时，这种不透明性会引发一系列严重的风险与影响。首先，它会削弱客户与市场的信任。例如，AI 系统可能给出某企业 BBB 的信用评级，但无法明确说明是因为“现金流不足”还是“行业风险过高”，导致客户与监管者无法判断结论的合理性。其次，它给评级机构的内部模型验证带来了巨大困难，验证团队难以对模型的稳健性和逻辑一致性进行有效评估。最后，它使机构面临日益严格的监管审查与合规风险。例如，欧盟《人工智能法案》已明确要求高风险 AI 系统必须具备可解释性，以保障用户的知情权与决策的公平性。从根源到影响的完整分析路径，以及对应的解决方案框架，如图 1 所示。

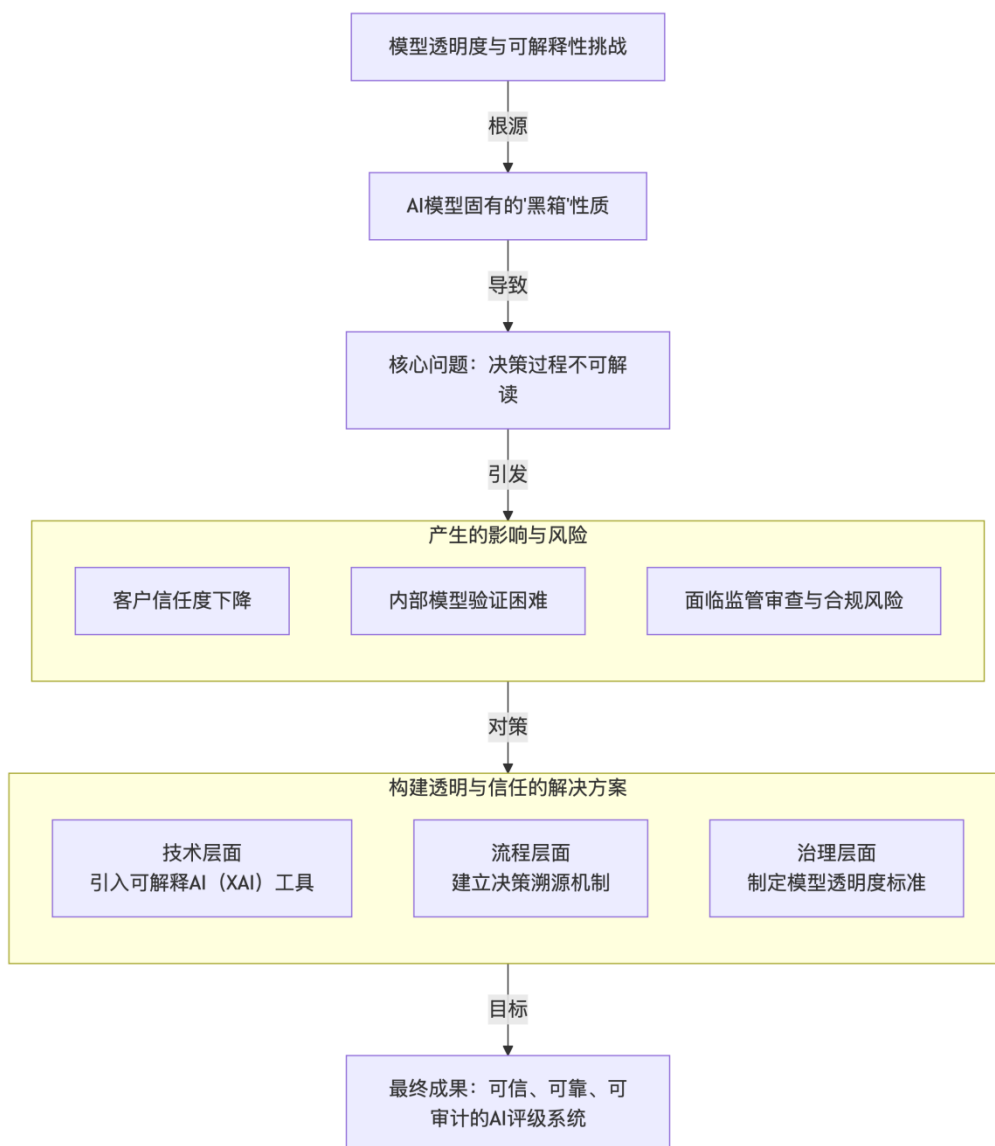


图 1 模型透明度与可解释性挑战

资料来源：公开信息，远东资信整理

### （三）技术复杂性与维护难度

AI 模型需要不断迭代以适应新兴行业与动态风险因子。模型的开发、调优与更新要求配备具备高水平知识与跨学科背景的技术团队，这对传统评级机构是巨大挑战。例如，若新能源行业因政策调整出现新的风险因子，评级模型必须在短期内完成结构调整与参数更新，否则将出现预测偏差甚至评级失效。因此，评级机构需建立长期的技术运维机制，包括持续的算法升级、跨部门协作和专业团队建设，以保持模型的有效性与前瞻性。

### （四）道德与合规风险

AI 模型的训练依赖大量历史数据，若数据本身存在偏差，将可能导致歧视性或不公平的评级结果。例如，若历史数据集中缺乏中小企业的正面样本，模型可能系统性地对中小企业给出偏低评级，进一步限制其融资机会，形成“马太效应”。

### （五）人工智能幻觉风险

大模型在生成文本时可能出现“幻觉”，即输出与事实不符的信息。一旦幻觉出现在评级报告或结论中，可能直接误导市场。例如，AI 在撰写评级报告时可能引用“并不存在的财务数据”或“虚构的新闻报道”，导致评级结论失真。

这一风险的成因是多方面的。其一是训练数据中可能存在的噪声与偏差，模型可能会学习并“复现”数据中存在的错误信息。其二是模型训练过程中可能出现的过拟合，导致模型过度记忆训练样本的细节，而不是学习通用的规律，从而在面对新情况时做出错误的推断。其三是生成式 AI 语言机制本身的不可控性，其概率性的文本生成方式有时会为了追求语言的流畅性而“创造”事实。

这些因素共同导致模型生成与事实不符的信息，最终对信用评级业务产生评级结论失真的严重影响。例如，AI 在撰写评级报告时可能引用“并不存在的财务数据”或“虚构的新闻报道”，从根本上动摇评级结论的可靠性。

图 2 系统梳理了关于 AI 幻觉的成因、影响以及应对策略。

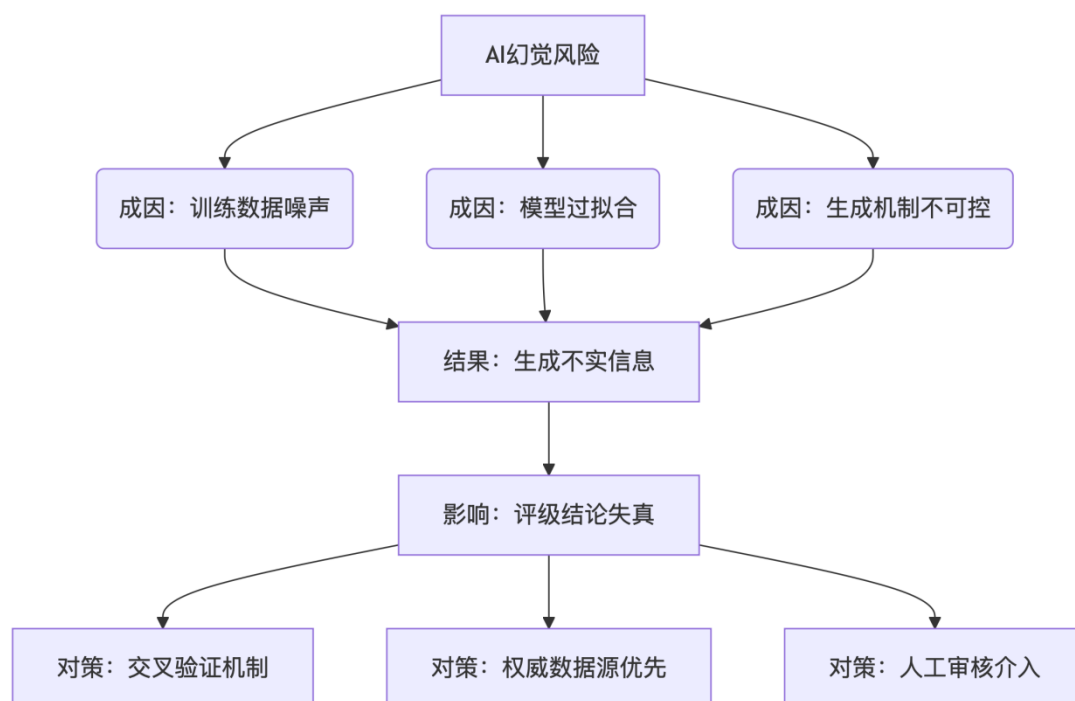


图 2 人工智能幻觉风险成因与对策

资料来源：公开信息，远东资信整理

## （六）任务适配性与边界不足

信用评级业务中部分环节对实时性与合规性要求极高，AI 并不适合全自动执行，尤其在需要法律解释或政策解读时，人工判断不可或缺。例如，涉及法律意见的评级说明，如果完全由 AI 生成而未经过专业律师审核，可能因对法规理解错误而触发合规风险。因此，评级机构需明确 AI 的适用范围与人工介入点，建立“人机协同”模式，而非完全替代人工，以确保评级结论的合法合规性。

## （七）证据可追溯性缺失

信用评级活动常需接受监管检查或面对客户争议，若缺乏完整的证据链，将难以证明评级结论的合理性与合规性。例如，在监管审查中，若机构无法提供 AI 模型在给出评级时所使用的原始数据、参数设置与中间过程记录，将削弱评级结论的公信力。

总体来看，AI 在信用评级中的应用，正面临着从基础技术、核心模型到应用治理三个层面的系统性挑战。这些挑战若得不到妥善解决，将严重削弱评级结论的科学性与权威性，甚至引发法律与声誉风险。因此，构建一套全面而有效的应对策略，是推动人工智能在信用评级领域健康发展的必然要求。

# 四、解决思路与方案

为系统性应对人工智能在信用评级中面临的多维度挑战，评级机构需构建一个全面可信的 AI 治理框架。该框架并非零散对策的集合，而是建立在夯实技术基石、建立流程规范和优化人机协同这三大核心支柱之上。本节将围绕这三大支柱，针对前述挑战提出具体解决方案。

## （一）数据隐私与安全性

为有效应对数据隐私与安全性的挑战，评级机构必须建立覆盖数据采集、存储、处理和使用全生命周期的防护体系。关键举措之一是采用先进的数据加密技术。通过采用 AES 或 RSA 等成熟加密算法，可以将原始数据转换为只有授权方才能解密的密文，从而有效防止数据在传输和存储过程中被窃取或篡改。机构应制定详细的数据加密实施流程，明确各环节的操作步骤与责任人，确保数据全链路安全，其标准化流程如图 3 所示。

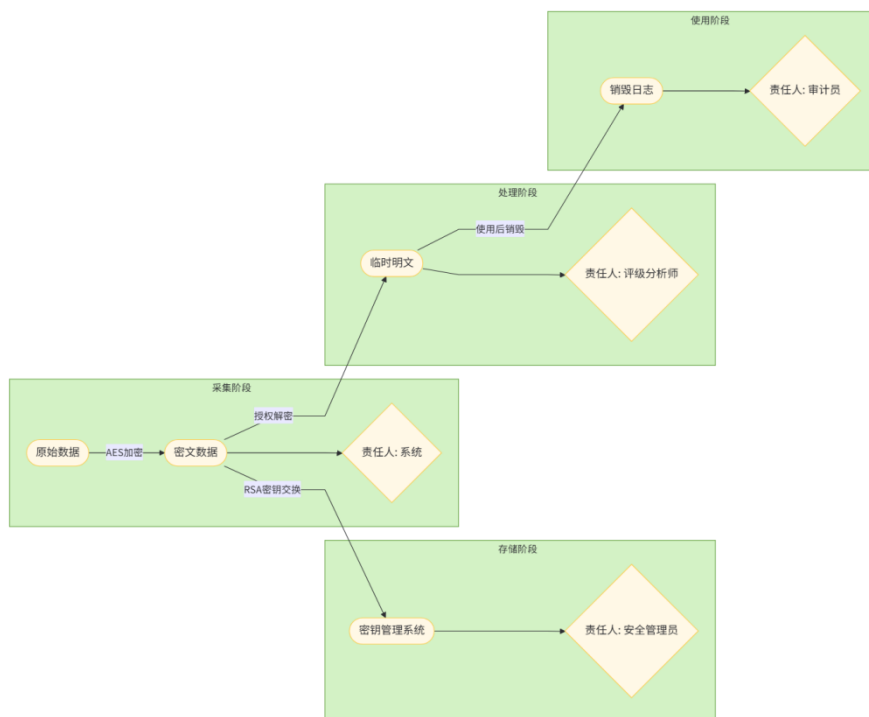


图 3 数据加密实施流程图

资料来源：公开信息，远东资信整理

另一关键举措是建立严格的数据访问权限管理制度，从而确保只有经过授权的人员才能访问和使用信用评级数据。访问权限方面，可以根据信用评级工作中不同岗位和人员的职责需求，对数据访问权限进行分级，例如，区分为管理层、项目作业分析师、其他人员三个角色。审批流程方面，评级机构需建立严格的权限审批流程，任何人员申请数据访问权限都需要经过多级审批。并明确审批环节和审批人员，确保权限授予的合理性和安全性。同时，IT 部门应制定定期审查计划，如每月或每季度对所有用户的数据访问权限进行全面审查。审查内容包括权限是否与用户的岗位职责相匹配、是否存在异常访问行为等。对于可能存在数据泄露风险的行为，IT 部门应及时将情况移交给内控部门，进行调查和处理。对于离职人员、岗位变动人员或权限过期的人员，也要及时清理其数据访问权限。

## （二）模型透明度与可解释性

为解决模型透明度与可解释性不足这一核心挑战，评级机构需从模型选择、技术工具和结果呈现三个方面入手，揭开模型决策的“黑箱”，以重建监管与市场的信任。在模型与工具层面，机构可以优先选择逻辑回归、决策树等本身具有高透明度的模型。对于神经网络等复杂模型，则可通过“知识蒸馏”等技术将其决策逻辑简化并迁移至一个可解释的代理模型上。通过整合专业人才，可以将这些方法开发成与现有评级系统无缝对接的 AI 可解释性工具，并为开发全过程编写详细文档，方便后续审计。见表 3，对比不同 AI 模型的透明度与适用场景，可见

逻辑回归、决策树等方法在可解释性上具有天然优势，而深度学习模型则需辅以知识蒸馏等手段以提升可解释性，为实际应用选择提供了参考依据。

表 3：可解释性模型选择对比表

| 模型类型       | 透明度等级 | 技术特点               | 适用场景                | 局限性              |
|------------|-------|--------------------|---------------------|------------------|
| 逻辑回归       | 高     | 基于线性公式，特征权重直接可解释   | 个人信用评分（如 FICO 基础版）  | 难以捕捉非线性关系        |
| 决策树        | 中     | 规则树结构可视化，决策路径可追溯   | 中小企业贷前风险初筛          | 深度过深时易出现“枝叶冗余”   |
| 知识蒸馏代理模型   | 一般    | 用简单模型（如线性模型）拟合复杂模型 | 替代深度学习模型（如 GBDT、NN） | 可能损失部分预测精度       |
| 深度学习（原始模型） | 低     | 多层神经网络，特征交互黑箱化     | 大规模复杂数据预测（如供应链金融）   | 无法解释中间层特征权重及决策逻辑 |

资料来源：公开信息，远东资信整理

在结果呈现层面，信用评级机构也可以通过可视化技术展示模型的决策过程，使评级结果更具可信度和可接受性。例如利用交互式仪表盘展示模型决策逻辑，支持用户逐层拆解评级结果，通过拖动滑块调整评级指标，实时观察评级结果的变化过程及关键影响因素，从而使评级结果更具可信度和可接受性。

（三）技术复杂性与维护难度

应对技术复杂性与维护难度高的挑战，关键在于采用现代化的系统架构设计。核心思路是采用模块化设计，将复杂的评级系统拆分为数据采集、数据处理、评级模型等多个独立且可通过清晰接口交互的功能模块。这种设计能够显著简化系统架构，降低开发、测试和维护的难度。同时，应用自动化部署工具可实现各模块的快速、准确部署，减少人为错误，提高系统迭代的效率和稳定性。

随着业务的不断发展和市场环境的变化，系统还需要不断扩展和升级。因此，在系统设计和开发过程中，应选择具有良好可扩展性的技术和框架，并在系统架构中预留扩展接口，方便后续功能的扩展和升级，以满足新的业务需求和监管要求。

（四）道德和合规风险

为系统性应对道德与合规风险及证据可追溯性缺失的挑战，必须构建一个覆盖事前预防与事后追溯的闭环问责体系。这不仅要求在模型设计阶段就根除歧视与偏见，还要求建立一套完整的证据追溯机制，以保障全流程的透明、可审计，并满足监管要求。

1.加强模型公平性评估，从源头防范歧视风险

训练数据中的历史歧视（如地域、性别关联的信用差异），或模型设计中的特征选择偏差，最终可能导致特定群体被系统性低估信用。因此，需要加强模型公平性评估，避免偏见，确保决策的公正性。可以采用多种评估

方法对模型的公平性进行全面评估，并定期对模型信用评级模型进行公平性评估。结合表 4 所列的多维公平性评估方法，可以实现对信用评级模型公平性的系统检验，确保模型输出在群体和个体层面均符合公正性要求。

表 4：AI 信用评级模型公平性评估方法对比表

| 评估维度  | 核心方法     | 技术实现                   | 适用场景                 | 优势              |
|-------|----------|------------------------|----------------------|-----------------|
| 统计公平性 | 人口学平等    | 按敏感属性（如性别、地域）分组计算通过率差异 | 检测群体间系统性偏差（如地域歧视）    | 直观反映宏观层面的分配公平性  |
| 个体公平性 | 相似个体相似结果 | 计算相似样本对的预测差异（如欧氏距离）    | 避免对特征相似个体的差异化对待      | 关注微观层面的个体权利平等   |
| 机会平等性 | 平等机会     | 按真实信用状态分层，计算不同群体的假阴性率  | 确保高信用群体不因敏感属性被错误拒绝   | 平衡公平性与模型预测准确性   |
| 校准公平性 | 群体校准     | 比较不同群体的预测违约概率与实际违约率    | 避免对特定群体的风险预测系统性偏高/偏低 | 保障模型概率输出的跨群体一致性 |

资料来源：公开信息，远东资信整理

同时，需要建立严格的模型评估和审核机制，确保人工智能技术的应用符合道德和法律规范。组建由信用评级专家、数据科学家、法律合规人员等组成的模型评估团队，确保评估团队具备多方面的专业知识和技能。制定详细的模型评估和审核流程，明确评估的各个环节和责任人。将模型评估和审核结果及时反馈给模型开发团队和管理层，对不合理的结果进行整改，以提升信用评级的公正性。

2.建立全流程证据追溯体系，保障决策可审计性

证据的可追溯性是人工智能在评级行业的合规性保障。当前，可以通过构建全链路数据与决策日志系统，实现评级过程透明化追溯。设计覆盖“数据采集—特征提取—模型计算—结果输出”的全流程日志框架，记录每个环节的关键信息：数据层面包括数据源、采集时间、预处理规则；模型层面包括算法版本、参数配置、中间变量；决策层面包括人工调整记录、审批意见等。采用区块链技术存储核心日志数据，利用其不可篡改特性确保追溯信息的真实性。

其次，制定标准化的证据归档与审计规范，满足合规要求。要明确评级证据的归档范围（如包括目录清单、评级流程记录表、投标文件、业务承揽材料、利益冲突审查情况、原始资料、内部研究报告、外部参考资料、现场调查记录、访谈确认单等）、保存格式（纸质版、电子签名版）及存储期限（如保存至评级协议期满后 5 年、评级对象存续期满后 5 年或者评级对象违约后 5 年，且不得少于 10 年）。建立定期审计机制，由内部合规部门或第三方机构对评级证据的完整性、准确性进行抽查，例如随机抽取 20%的评级案例，核查其数据来源与模型计算逻辑的一致性。同时，将证据追溯能力纳入模型验收标准，未通过追溯测试的模型不得上线应用。

（五）人工智能幻觉

针对第三节所述的人工智能“幻觉”风险，解决方案必须是一个集技术预防与流程保障于一体的综合体系。在技术层面，需要从数据源头和模型训练两方面入手建立质量控制流程；在流程层面，则必须引入有效的人工审核作为最后一道防线。

首先，需要构建覆盖金融、替代及第三方权威数据的多源数据交叉验证机制，通过设计数据校验规则库，对不同来源数据进行自动核验，这一交叉验证体系如图4所示。

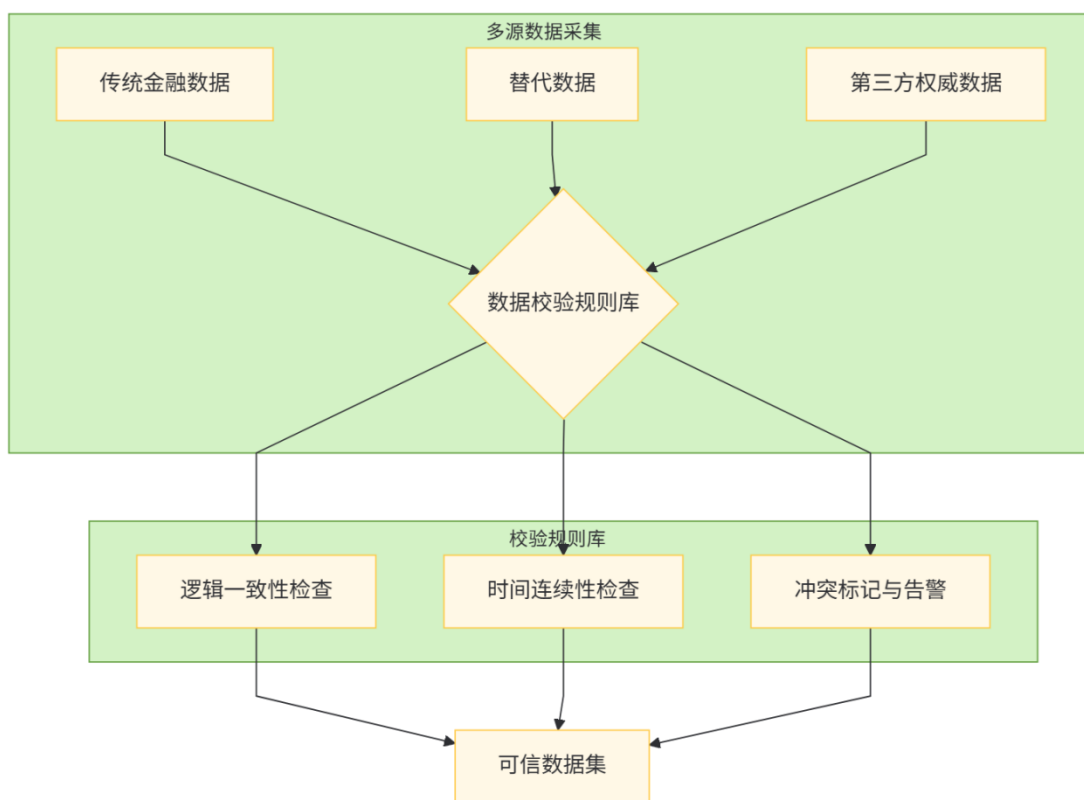


图4 多源数据交叉验证体系图

资料来源：公开信息，远东资信整理

同时，优化模型训练数据质量，提升抗干扰能力。建立数据清洗与预处理标准化流程，包括缺失值处理、异常值剔除、噪声过滤等环节，例如采用 IQR（四分位距）法识别并处理极端值。在模型训练中引入对抗性训练技术，通过生成“噪声样本”模拟真实场景中的虚假信息，增强模型对异常数据的辨别能力。定期对训练数据分布进行监控与更新，确保数据与实际业务场景的动态适配，避免因数据陈旧导致模型产生过时或错误的判断。图5直观地展示了这一从原始多源数据采集，经过缺失值填补、异常值剔除和噪声过滤等清洗环节，最终生成干净训练集的标准化流程。

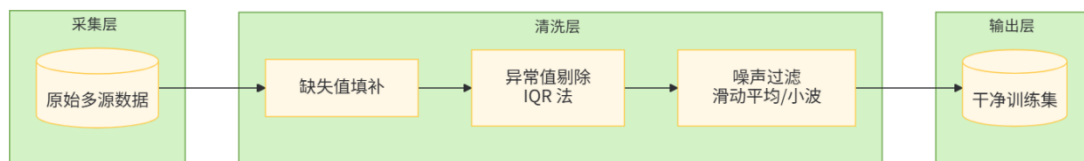


图5 数据清洗标准化流程图

资料来源：公开信息，远东资信整理

然而，仅有上述技术层面的防控尚不足以完全消除“幻觉”风险。技术手段主要用于最大化地保证输入信息与生成内容的客观性与一致性。但对于最终评级结论的商业逻辑、风险敏感性与合规性判断，必须结合下节所述的“人机协同”模式，将人工审核介入作为不可或缺的最后一道防线。技术验证与人工审核的有机结合，才能共同构筑起抵御人工智能幻觉风险的坚实壁垒。

## （六）任务适配性与边界不足

为解决任务适配性不足与应用边界模糊的挑战，根本出路在于优化人机协同模式。这要求从明确职责边界与开发动态适配模型两个维度入手，构建高效、安全的评级机制。

首先，明确 AI 与人工的职责边界。通过业务流程梳理，应清晰 AI 负责数据处理、指标计算、初步风险预警等标准化任务，例如自动计算企业的流动比率、资产负债率等财务指标，而人工则聚焦于非结构化信息解读（如管理层访谈、行业政策分析）、特殊案例评估（如新兴行业企业评级）及复杂风险判断。建立“AI 初评—人工复核—最终审批”的评级流程，确保关键决策环节保留人工干预权。

其次是开发动态任务适配模型以提升场景化应用能力。机构应基于个人信贷、企业债券等不同评级场景构建优化的子模型，并引入迁移学习技术，将成熟场景的模型经验迁移至新兴场景（如绿色金融评级），缩短模型开发周期。同时，建立模型性能监控体系，实时追踪不同场景下的评级准确率等指标，一旦指标低于阈值时自动触发模型调优或人工介入流程。图 6 展示了信用评级场景化子模型的总体架构，不同应用场景通过定制化特征工程和算法参数实现更高的精准度和适应性。

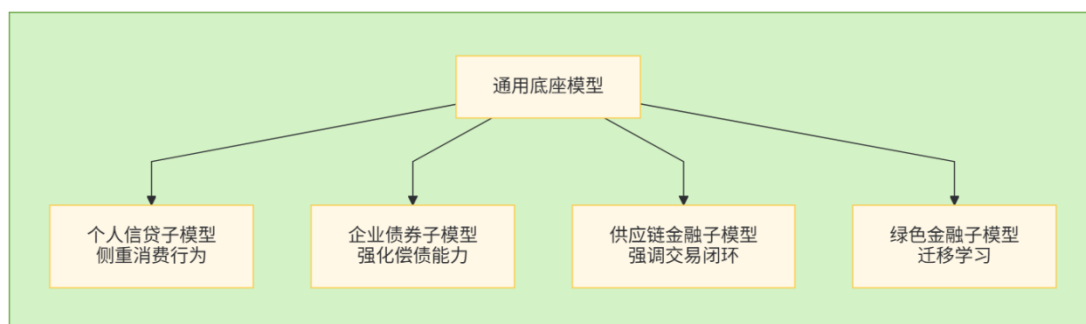


图6 场景化子模型架构图

资料来源：公开信息，远东资信整理

实施上述策略，通过构建以技术基石、流程规范、人机协同为三大支柱的可信 AI 治理框架，信用评级机构不仅能够系统地化解人工智能应用中的七大核心挑战，更能从根本上推动信用评级运作模式的变革。这套方案的实施，将驱动信用评级从传统的、基于静态数据的回顾式评估，升级为现代化的、基于动态数据的前瞻式风险洞察体系。最终，它不仅是风险控制的工具，更是提升金融市场资源配置效率、服务实体经济高质量发展的强大引擎。

### 五、未来发展趋势与展望

展望未来，人工智能技术于信用评级领域的应用前景，正驱动着一场从技术体系、治理生态到行业格局的系统性变革。这场变革的路径将遵循技术前沿的突破、治理生态的进化与行业终局的重塑三大演进脉络。本节将循此路径，描绘一幅由 AI 驱动的未来信用评级新蓝图。

#### （一）技术前沿：从增强辅助到逻辑重塑

从技术发展的内在趋势而言，深度学习与生成式人工智能（AI）的深度融合，将持续推动信用评级的核心逻辑及其能力边界的重塑。传统信用评级模型依赖人工特征工程与静态数据输入，然而，深度学习借助神经网络的多层非线性映射，能够自动挖掘数据中的复杂关联与动态模式，促使信用评估从“表面特征匹配”迈向“深层风险洞察”。生成式 AI 的引入，进一步突破了传统模型的局限，其不仅能够基于历史数据生成符合信用风险分布规律的模拟样本，拓宽训练数据的广度并覆盖极端场景，还能通过构建假设情景开展压力测试，模拟宏观经济波动、行业政策调整或供应链断裂等突发事件对企业信用状况的连锁影响，进而增强评级模型的前瞻性与风险预警能力。

表 5：传统信用评级 vs 深度学习评级 vs 生成式 AI 评级对比表

| 维度     | 传统模型  | 深度学习模型   | 生成式 AI 模型   |
|--------|-------|----------|-------------|
| 特征处理方式 | 人工选取  | 自动提取     | 自动生成与扩展     |
| 数据类型支持 | 结构化为主 | 结构化+非结构化 | 多模态（文本、图像等） |
| 风险识别能力 | 静态评估  | 动态识别     | 压力测试、极端预测   |
| 模型解释性  | 较高    | 中等       | 需特别设计       |

资料来源：公开信息，远东资信整理

根据表 5 对比可知，深度学习与生成式 AI 模型在数据处理、风险识别等方面相较传统模型具备显著优势，但也对模型可解释性和监管合规性提出了更高要求。随着这些技术的持续迭代与优化，信用评级模型的精准度和运行效率将得到显著提升。一方面，深度学习能够开展更为精细的特征分析，涵盖对非结构化数据（如企业年报文本情感、社交媒体舆情、卫星图像等）的深度解析，以及对时序数据中长期依赖关系的捕捉，从而精准识别传统

指标难以覆盖的隐性信用风险因素，降低误判率并提高对高风险客户的识别能力。另一方面，深度学习和生成式 AI 的自动化特性大幅缩短了模型训练、参数调优和效果评估的周期，使信用评级能够更迅速地响应市场变化和个体信用状态的动态演进，显著提高评级结果的时效性，这在瞬息万变的金融市场和高频交易场景中尤为关键。

此外，多技术的融合已成为推动信用评级向“更全面、更可信、更实时”方向演进的关键驱动力。单一的人工智能技术在解决数据孤岛、确保模型透明性或应对动态风险等方面往往存在局限，而不同技术的协同作用将开启新的可能性。例如，区块链技术与人工智能的结合，能够利用区块链的分布式账本与不可篡改特性，确保信用数据在采集、传输及使用过程中的全程可追溯性与安全性，并通过智能合约实现数据共享的权限管理，在保障数据隐私的前提下打破信息壁垒，解决中小企业信用信息缺失、数据造假等问题。物联网技术的普及为信用评级提供了更为丰富、实时的底层数据来源，从传感器监测制造企业的生产设备运行状态，到农业贷款中的农田环境和作物生长数据，乃至物流企业的车辆轨迹与运营效率，这些动态数据的引入将使信用评估从依赖滞后的财务报表转变为更具预测性的“行为信用”分析，进一步提升评级的准确性与前瞻性。这一多技术融合的趋势，不仅是技术层面的创新，更是对传统信用评级范式的系统性重构，将推动信用服务从单一的风险评估工具，向赋能企业及个人信用管理、优化资源配置的综合金融基础设施转型。

表 6：AI 在不同评级场景中的应用矩阵

| 应用场景    | 数据来源        | 使用技术       | 输出结果           |
|---------|-------------|------------|----------------|
| 小微企业评级  | 实时交易、物流轨迹   | 深度学习、图神经网络 | 动态信用评分         |
| 个人信用评估  | 社交媒体、消费记录   | 情感分析、GAN   | 行为信用等级、舆情风险预警  |
| 企业级压力测试 | 宏观经济指标、政策变动 | 生成式 AI     | 不同场景下的违约概率变化预测 |

资料来源：公开信息，远东资信整理

从表 6 可以看出，AI 技术已在小微企业、个人信贷、企业压力测试等多场景下实现差异化应用，极大丰富了信用评级的数据基础和评估维度。

（二）治理生态：从适应到引领的监管变化

在人工智能技术重塑信用评级行业的进程中，政策与监管环境的适应性和动态优化起着至关重要的作用。政策监管不仅是技术创新的“护航者”，更是防范系统性风险的“压舱石”。监管机构需秉持更为前瞻和包容的态度，密切关注人工智能技术在信用评级应用中的演进路径与潜在影响，构建与技术发展相适配的监管框架。这要求传统以“事后审批”和“规则导向”为主的监管模式，向“事中监测”与“原则导向”相结合的方向转变。监管机构不仅要确保人工智能应用符合金融市场的整体稳定性要求，防范算法缺陷或数据偏见引发的系统性风险，还应坚守公平性原则，避免算法歧视导致特定群体或企业在信用评估中受到不公平对待。例如，针对人工智能模型可能存在的“黑箱”问题，监管机构可推动建立模型可解释性的最低标准，要求评级机构能够清晰阐释关键决策逻辑和影响因素，并加强对模型输出结果的持续监测与回溯检验，及时发现并纠正可能出现的偏差。

此外，积极的政策支持与战略引导对于激发行业创新活力、加速技术落地同样不可或缺。政府可通过出台专项激励政策，鼓励信用评级机构、商业银行、科技企业等市场主体加大在人工智能信用评级技术研发上的投入。例如，给予相关研发费用税收优惠，或为采用人工智能技术提升评级效率与准确性的机构提供业务试点机会。通过设立国家级或行业性科研基金，重点支持信用评级领域的基础性与前沿性技术攻关，如多模态数据融合算法、隐私计算与 AI 协同技术、极端风险预测模型等，能够突破关键技术瓶颈。更进一步，应推动构建“政产学研用”协同创新体系，支持高校、科研机构与信用评级机构、金融机构开展深度合作，联合培养专业人才、共建实验室或创新中心，将学术研究成果与行业实际需求紧密对接，加速技术成果向商业化应用的转化，从而整体提升我国信用评级行业在智能时代的核心竞争力。

在全球化背景下，推动国际监管协调同样是完善人工智能信用评级治理体系的重要组成部分。人工智能技术的应用具有跨境数据流动与算法输出全球化的特点，不同国家和地区在 AI 技术的金融应用上的监管思路与合规要求存在差异，这可能引发监管套利风险，并增加跨国金融机构的合规成本。因此，我国监管部门应积极关注国际金融领域对人工智能技术应用的监管趋势与政策变化，参与国际金融组织及标准制定机构关于 AI 信用评级的规则讨论与标准制定，推动国际监管框架的建立，加强与主要经济体金融监管机构和国际信用评级机构的常态化沟通与协调，特别是在数据隐私保护、模型风险管理、跨境评级结果互认等领域寻求共识。综合表 7 的国际监管政策对比，可以发现中国在数据安全、算法透明度和监管机制上与欧美保持高度同步，同时展现出以安全可控为核心的本土特色。

表 7：中美欧在 AI 信用评级监管方面的政策对比

| 维度      | 美国                    | 欧盟            | 中国                                 |
|---------|-----------------------|---------------|------------------------------------|
| 监管框架    | 分散式监管（多机构协同）          | 统一立法+行业监管     | 多部门协同监管                            |
| 核心原则    | 市场驱动+风险导向             | 权利本位+风险分级     | 安全可控+数据主权                          |
| 数据治理    | 州级法律差异大（如 CCPA）       | GDPR 统一规范     | 《中华人民共和国个人信息保护法》                   |
| 算法透明度要求 | 披露“关键因素”（如 FICO 评分要素） | 高风险 AI 需全流程透明 | 《互联网信息服务算法推荐管理规定》《生成式人工智能服务管理暂行办法》 |
| 执法与处罚机制 | 案例式执法                 | 分级处罚          | 阶梯式处罚                              |

资料来源：公开信息，远东资信整理

（三）行业未来：迈向实时、动态与普惠的信用新基建

从行业发展的长远视角来看，人工智能技术的应用正推动信用评级行业经历一场前所未有的变革。未来的信用评级将逐步摆脱对传统评级模型所需专家经验的过度依赖以及有限样本数据的束缚，向基于大数据驱动、复杂算法支撑的科学化、精准化评估模式转变。人工智能的引入，首先将实现评级流程的全面自动化与标准化升级——从数据采集、清洗、特征提取到模型训练、评级模型输出的全链条，均可通过算法完成，这不仅大幅提升评级效率，更能有效减少人为干预和主观判断可能带来的偏差，使评级结果更具客观性和一致性。

更为关键的是，人工智能展现出强大的多源数据整合与深度挖掘能力。它能够突破传统财务数据的局限，将企业的结构化数据与非结构化数据均纳入评估体系，构建更为立体、全面的信用画像。例如，基于实时交易数据和物流信息，AI模型可更精准地评估小微企业的经营状况和现金流健康度，为传统评级中难以覆盖的小微企业或“信用白户”提供合理的信用评估。同时，AI的实时监测能力意味着评级不再是静态的“一次性报告”，而是能够对评级主体的信用状况进行24小时不间断的动态追踪，一旦出现异常波动（如财务指标大幅下降、负面舆情集中爆发等），系统能够迅速发出预警信号，为相关金融机构和投资者提供警示，从而显著提升风险预警的前瞻性和有效性。

人工智能在信用评级领域的应用将为整个金融市场生态带来深远影响。利用人工智能技术生产真实反映评级主体信用状况的评级结果，将有效缓解金融市场上信息不对称问题，从源头上降低信用风险的发生概率。对于金融机构而言，更为科学、准确的评级结果也有助于帮助自身快速识别优质借款人，优化信贷资源配置，将资金更高效地导向有真实需求和还款能力的企业或个人，提高资金使用效率。同时，人工智能的应用还可显著降低金融机构在信息搜集、尽职调查和风险评估等环节的人工和时间成本。通过这些变革改变，人工智能技术将推动金融市场朝着更高效、更稳定、更具包容性的方向健康发展，为实体经济高质量发展注入更强大动力。

## 六、结论与建议

本研究系统地探讨了人工智能技术在信用评级领域的应用、挑战、解决方案及未来趋势。通过对技术应用场景的深入剖析、潜在风险的系统性梳理、应对策略的框架性构建以及未来前景的前瞻性展望，我们得出以下核心结论，并提出相应的政策建议与未来研究方向。

### （一）研究结论

第一，人工智能正从辅助工具升级为重塑信用评级业务模式的核心引擎。研究发现，人工智能的应用已全面渗透至客户沟通、尽职调查、分析评估和报告撰写的评级全流程。其价值不再局限于提升单一环节的效率，而是通过数据驱动和智能分析，系统性地重塑着整个行业的价值链。当前及未来的最佳实践是构建高效的“人机协同”模式，由AI处理标准化、数据密集型任务，人类分析师则专注于复杂、高附加值的风险判断与决策。

第二，技术应用伴生了贯穿技术、模型与治理三个层面的系统性风险。本研究识别出，AI在信用评级领域的应用挑战具有多层次和系统性的特点。基础技术层面存在数据隐私安全与人工智能幻觉等风险；核心模型层面则面临因其复杂性导致可解释性不足的问题以及算法偏见等内在缺陷；应用治理层面则暴露出人机权责边界模糊、决策证据难以追溯等治理难题。这些风险相互交织，共同对评级结果的科学性、公正性与合规性构成潜在威胁。

第三，构建系统性的“可信AI”治理框架是应对挑战的必然路径。面对复杂的系统性风险，零散的技术性修复措施难以奏效。研究表明，必须构建一个全面的“可信AI”治理框架。该框架应建立在三大核心支柱之上，即夯实数据与算法安全的技术基石、建立覆盖全流程的透明与合规的流程规范，以及优化权责清晰的人机协同模式。这一框架是确保AI技术在信用评级领域实现稳健、可信与可持续发展的核心保障。

第四，未来行业发展将呈现技术深度融合、监管动态适应与格局全面重塑的趋势。展望未来，人工智能技术将与区块链、物联网等技术进一步深度融合，驱动信用评级向实时化、动态化和普惠化方向发展。与此同时，全球监管环境将更趋严格与动态，要求评级机构在技术创新与合规风控之间取得精妙平衡。最终，这些力量将共同推动信用评级行业的格局重塑，使其从传统的风险评估工具，进化为服务于整个金融市场资源配置效率提升的新一代信用基础设施。

## （二）政策建议

第一，针对 AI 重塑业务模式的趋势，建议引导并规范“人机协同”的应用实践。监管机构应鼓励行业协会制定“人机协同”评级作业的最佳实践指南，明确 AI 与人类分析师在评级流程中各自的职责边界与问责机制。同时，应将复合型人才培养纳入行业发展规划，鼓励企业与高校合作，培养既懂信用评级业务又精通人工智能技术的专业人才，以适应新的业务模式需求。

第二，针对多层次的系统性风险，建议建立风险导向的分类分级监管体系。监管部门应借鉴欧盟《人工智能法案》的思路，对不同场景下的 AI 评级应用进行风险分级。对于风险较高的应用（如对系统重要性机构的评级），应提出更严格的监管要求，包括强制性的算法审计、压力测试和应急预案。对于风险较低的辅助性应用，则可给予更大的创新空间，从而实现精准监管。

第三，为推动“可信 AI”治理框架的落地，建议加快制定关键技术与流程标准。政府与行业协会应牵头制定针对模型透明度与可解释性、算法公平性以及数据安全的行业标准或团体标准。例如，可推广“模型卡片”等机制，要求评级机构以标准化格式披露其 AI 模型的关键信息，并建立独立的第三方评估机制，对模型的合规性与稳健性进行认证。

第四，为适应未来的融合与变革趋势，建议构建前瞻性、适应性的产业发展与监管协调机制。建议设立“监管沙盒”，为新技术（如 AI 与区块链的融合应用）提供风险可控的创新测试环境。同时，我国监管部门应积极参与全球 AI 治理规则的制定与协调，特别是在跨境数据流动、算法监管互认等领域加强国际合作，为我国信用评级行业的全球化发展营造有利的政策环境。

## （三）未来研究方向

本研究虽构建了较为完整的分析框架，但人工智能在信用评级领域的应用仍处于快速演进中，若干深层次难题当前尚无法完全解决，值得后续持续关注与研究。

其一是生成式 AI 事实性偏差的根本性治理难题。当前通过交叉验证和人工审核等手段仅能做到事后修正，但如何在模型层面从根本上消除或量化此类偏差的产生概率，仍是基础理论与技术上的一大挑战。这关系到 AI 能否在核心判断环节被真正信任。

其二是模型可解释性与性能的平衡问题。虽然已有可解释性 AI 工具，但在最先进、性能最强的深度学习模型上实现真正意义上的、符合监管要求的完全可解释性，往往伴随着性能的牺牲。如何在两者之间找到最优平衡点，是技术、业务与监管都需持续探索的难题。

其三是“数据孤岛”与隐私保护的内在矛盾。AI 模型的性能依赖于海量、多样化的数据，但这与日益严格的数据隐私保护法规之间存在天然的矛盾。联邦学习等技术虽提供了思路，但其在金融领域的规模化、商业化落地仍面临效率、成本与安全性的多重制约，是未来需要重点攻关的方向。

### 【作者简介】

赵恒，远东资信研究院创新业务部高级研究员。

李之行，远东资信研究院创新业务部研究员。

张怀玉，远东资信研究院创新业务部研究员。

### 【关于远东】

远东资信评估有限公司（简称“远东资信”）成立于1988年2月15日，是中国第一家社会化专业资信评估机构。作为中国评级行业的开创者和拓荒人，远东资信开辟了信用评级领域多个第一和多项创新业务，为中国评级行业培养了大量专业人才，并多次参与中国人民银行、国家发改委会和中国证监会等部门的监管文件起草工作。

远东资信资质完备，拥有中国人民银行、国家发改委会、中国证监会、中国银行间市场交易商协会和中国保险资产管理业协会等政府监管部门和行业自律机构认定的全部信用评级资质。



### 远东资信评估有限公司

网址: [www.sfecr.com](http://www.sfecr.com)

#### 北京总部

地址: 北京市东城区东直门南大街11号中汇广场B座11层

电话: 010-5727 7666

#### 上海总部

地址: 上海市杨浦区大连路990号海上海新城9层

电话: 021-6510 0651

### 【免责声明】

本报告由远东资信提供。报告引用的相关资料均为已公开信息，远东资信进行了合理审慎的核查，但不应视为远东资信对引用资料的真实性及完整性提供了保证。

远东资信对报告内容保持客观中立态度。报告中的任何表述，均应严格从经济学意义上去理解，并不含有任何道德偏见、政治偏见或其他偏见，远东资信对任何基于这些偏见角度理解所可能引起的后果不承担任何责任。报告内容仅供读者参考，但并不构成投资建议。

本报告版权归远东资信所有，未经许可，任何机构或个人不得以任何形式进行修改、复制、销售和发表。如需转载或引用，需注明出处，且不得篡改或歪曲。

我司对于本声明条款具有修改和最终解释权。