Journal of Frontier in Economic and Management Research

DOI: https://doi.org/10.63944/sjmx.JFEMR
Vol.2 No.7

超越“黑箱”：审计大模型的因果可解释性框架

罗洋皓[1] 莫 磊[1*]

（1.广西财经学院，广西 南宁 530003）

[摘 要] 大模型的“黑箱”特性与审计结论必须可追溯、可验证、可复核的刚性要求之间存在根本张力，已成为制约审计大模型规模化应用的瓶颈。本文借鉴可解释人工智能理论与因果推断方法，构建基于因果链的内部审计大模型可解释性框架：第一阶段由自然语言处理模型抽取凭证实体，并以 SHAP 值量化各特征对风险评分的贡献；第二阶段以因果图刻画特征间的因果路径、区分直接原因与伪相关，并由大语言模型生成反事实解释；第三阶段在审计报告中呈现风险等级、关键特征置信度与完整因果链。以应收账款审计场景为例的比较表明，该设计能使不透明的评分转变为可追溯、可验证的审计证据链。本文进一步讨论幻觉风险与数据治理成本，并提出人机协同复核与“T 型”人才培养等对策。

[关键词] 内部审计；审计大模型；可解释性；因果链；反事实推理

Beyond the “Black Box”: A Causal Explainability Framework for Audit Large Models

Yanghao Luo[1], Lei Mo[1*]

[1] Guangxi University of Finance and Economics, Nanning, China

Abstract: The “black box” nature of large models stands in fundamental tension with the rigid requirement that audit conclusions be traceable, verifiable and reviewable, and this tension has become a bottleneck constraining the large-scale application of audit large models. Drawing on explainable artificial intelligence (XAI) and causal inference, this paper constructs a causal chain-based explainability framework for audit large models, which proceeds in three stages: in the first stage, a natural language processing model extracts voucher entities, and SHAP values quantify the contribution of each feature to the risk score; in the second stage, a causal graph depicts the causal paths among features and distinguishes direct causes from spurious correlations, while a large language model generates counterfactual explanations; in the third stage, the audit report presents the risk level, the confidence of key features and the complete causal chain. A comparison based on an accounts receivable audit scenario shows that this design can turn opaque scores into traceable and verifiable audit evidence chains. The paper further discusses hallucination risks and data governance costs, and proposes countermeasures such as human-in-the-loop review and “T-shaped” talent cultivation.

Keywords: internal audit; audit large models; explainability; causal chain; counterfactual reasoning

一、引言

国有企业是中国特色社会主义的重要物质基础和政治基础。内部审计作为公司治理的核心环节，正被期望

作者简介：罗洋皓（2000—），男，硕士研究生，研究方向：内部审计；
莫 磊（1983—），男，博士，教授，研究方向：审计、内部控制、会计史。
通信作者：莫 磊

从传统的合规导向模式转向具有事前预警、事中赋能与事后优化特征的价值增值型管理审计。然而，国有企业集团普遍呈现管理层级多、业务板块跨行业、资金流量巨大的特征，传统的抽样式、事后型审计模式已难以在有限资源下及时覆盖风险、识别隐蔽问题。

大模型技术为上述挑战提供了可行解法。二十届中央审计委员会第一次会议明确提出“坚持科技强审”，《审计署关于加快审计数字化转型的指导意见》进一步要求在审计实践中深化人工智能、区块链等前沿技术的应用[1]。大模型凭借全样本分析、跨域关联与持续学习的能力，被普遍视为审计智能化转型的核心引擎；但其“黑箱”特性与审计结论必须可追溯、可验证、可复核的刚性要求之间存在根本张力[2]，这一张力正是本文所讨论的可解释性问题。

二、文献回顾与理论基础

（一）相关文献回顾

已有研究大致沿三条线索展开。其一，人工智能赋能内部审计。文献[3]分析了审计人员与人工智能在各阶段的比较优势，文献[4]指出交易性、重复性工作可由智能审计软件按审计人员的思维自动完成，文献[5]通过建立智能审计信息系统提升了审计工作的自动化与智能化水平；但上述研究均未讨论智能系统生成的结论如何被解释、被验证。其二，审计大模型的构建与实践。文献[1]探讨了 DeepSeek 等大模型在审计中的应用路径，文献[2]构建了企业审计画像数据治理体系，文献[6]分析了智能审计模型的技术架构与实现模式，文献[7]梳理了大语言模型在审计信息采集与数据分析中的应用；总体偏重“如何建、如何用”，对“如何使输出取信于审计人员、可被监管接受”关注不足。其三，可解释人工智能。文献[8]提出统一特征归因框架 SHAP，基于 Shapley 值公平分配模型预测并保证归因的全局一致性与局部准确性；文献[9]主张通过因果图、干预与反事实推理实现从统计相关到因果判断的跃迁；文献[10]强调高风险场景不应依赖对“黑箱”模型的事后解释，而应直接构建透明可解释的模型。上述理论提供了方法论基础，但审计场景要求结论由完整、可追溯的证据链支撑并经受外部复核，需面向领域设计而非直接移植通用 XAI 方法。

（二）核心概念与理论基础

内部审计大模型是大语言模型、深度学习算法与内部审计专业知识体系深度融合所构建的智能应用系统，具备数据处理、风险识别与审计证据核验等能力[2]。此处的可解释性，指模型能够以人类可理解的方式说明其决策依据与推理过程，其实质是将模型输出转化为“结论—理由—证据”链条。本文的理论基础有三：审计证据理论要求结论由充分、适当、可追溯的证据支持，界定了可解释性的目标与验收标准；因果推断理论以因果图刻画变量间的因果关系，为审计风险归因提供方法论支撑[9]；数字孪生审计理论通过整合全域实时数据构建“业务数字孪生”，使预测性结论与回溯性结论适用同样的证据要求。

三、审计大模型可解释性的现实困境

（一）审计证据适当性的刚性要求

审计工作遵循“结论—证据”的逻辑，结论必须说明形成所依据的基础、路径与判断标准，并接受复核与验证。在传统审计模式下，审计工作底稿完整记录了证据来源、分析过程与判断依据，链条可由主管人员、质量复核人员与外部审计师重建。然而，大模型的输出往往只提供结论与评分，仅表述“模型内部计算评分超过阈值”无法回答审计的基本证据问题：哪些具体特征驱动了该结论？某笔交易为何构成高风险？有何证据支持该判断？由此，此类输出不能构成符合证据规范的审计证据，这种“黑箱”特性与审计的专业属性直接冲突。

（二）监管与合规要求

审计大模型的应用必须在严格的法律法规框架内运行。《中华人民共和国数据安全法》确立了数据分类分级保护制度，《中华人民共和国个人信息保护法》强调个人信息处理的“告知同意”与“最小必要”原则；在国有资产监管领域，《中央企业合规管理办法》与持续推进的“AI+”专项行动则体现了“鼓励创新”与“严控风险”并重的取向。若审计大模型无法说明其决策依据，将难以通过合规审查、责任认定与监管检查，其生成的审计意见也可能受到质疑或被否弃。

（三）应用实践中的软约束缺失

在实践中，审计大模型的应用多从技术导向推进，重视算法准确率等硬指标，而忽视国有资产监管要求、内部控制制度与企业文化等软约束，形成“技术上可用、业务上不可接受”的困境：审计人员无法核验模型判断，审计委员会与管理层难以采纳模型结论，外部审计师与监管机构对模型生成的证据持保留态度。从制度信任的角度看，审计结论的权威性不仅来自技术上的正确，也来自可被公开解释、检查与质询的推理过程；当推理仍不透明时，审计工作所依赖的制度信任即被削弱，即便预测准确，模型的价值也大打折扣。可解释性由此成为审计大模型规模化部署的关键前提。

四、基于因果链的可解释性框架构建

（一）框架总体设计

针对审计实践对合规性、结论可解释性与系统适应性的要求，本文构建基于因果链的可解释性审计框架，其核心是在输出审计结论的同时，同步生成一条可追溯、逻辑严密的因果证据链。设计遵循三项原则：证据导向，即每一项结论都必须伴随可重建的证据链；领域适配，即框架须契合审计的专业规范与工作程序；操作可嵌入，即框架应以较低成本嵌入现有审计流程。据此，审计过程被划分为输入与特征提取、因果推理、可解释输出三个阶段（如图 1 所示）。



图 1 内部审计大模型三阶段可解释性框架

（二）第一阶段：输入与特征提取

审计系统接收发票、合同、银行流水等原始凭证，借助自然语言处理模型识别并抽取交易金额、合同条款、日期等关键实体，将非结构化或半结构化文档转化为结构化数据，并处理跨业务系统的异构格式与语义对齐问题。同时，系统采用 SHAP 方法量化分析各特征维度对风险评分的具体贡献[8]；基于 Shapley 值思想，SHAP 将模型预测公平地分配至各特征，保证归因的全局一致性与局部准确性，其输出的特征重要性既识别了最具影响力的因素，也为因果推理阶段提供可靠输入。

（三）第二阶段：因果推理

本阶段引入因果图模型刻画特征间的因果路径，判断“金额巨大”是否构成“高风险评级”的直接原因而非受其他变量混淆的统计关联，并识别混淆变量、中介变量与直接效应，避免将伪相关误判为因果。在此基础上，系统融合大语言模型生成反事实解释，构建“若交易金额低于预设阈值 X，风险评级预计将降至较低等级”的假设情景[9]，使审计结论不仅能回答“是什么”，还能回答“为什么”与“若何而变”，并为对推理过程本身的审计提供机制。

（四）第三阶段：可解释输出

最终审计报告除风险评级结果外，还系统呈现三项要素：风险等级的具体分类，使审计人员直观把握风险程度；影响风险决策的关键特征及其置信度指标，明确哪些因素主导了结论；从原始输入数据经特征提取与因果推理直至结论的完整因果链，使整个决策路径可被重建。审计人员据此可定位决定性特征、评估证据充分性并判断是否需进一步调查。

（五）框架特征

本框架具有三个特征：全流程可追溯，各阶段的输入、方法与输出均可被重新审视，契合审计质量控制对工作底稿的要求；兼顾合规与可嵌入性，通过标准化接口嵌入现有审计作业流程，降低学习成本与采纳阻力；以“证据链”而非“模型评分”支撑结论，将技术上的可解释性转化为管理层、监管机构与外部审计师可接受的制度接口。

五、应用示例：应收账款审计场景

本文以应收账款风险识别为例。应收账款审计是财务审计中风险最敏感的领域之一：虚增应收账款可能掩盖虚构的收入，坏账风险直接影响财务报表的真实性，舞弊风险评估则高度依赖交易金额、客户信用历史与合同条款等多维证据。这些特征使其成为展示可解释性框架如何将不透明的模型输出转化为透明证据链的适宜场景（见表 1）。

表 1 传统模型与可解释模型输出比较

项目	传统模型输出	可解释模型输出
结论	该笔应收账款存在极高舞弊风险，建议全额计提坏账准备。	该笔应收账款存在极高舞弊风险，建议全额计提坏账准备。
理由	模型内部计算评分超过阈值。	特征 A（金额）：500 万元（权重+40%），账龄远超行业平均信用期，触发高敏感规则；特征 B（信用历史）：客户存在逾期付款历史（权重+30%），还款意愿较低；特征 C（合同条款）：合同约定付款期限 60 天，但信用期已超出 120 天（权重+30%），构成严重违约。
反事实验证	无	若金额低于 50 万元，该笔交易的风险评分将降至“关注”等级。

注：表中数据为研究项目的示例性数据。

在传统模型下，模型得出该笔应收账款“存在极高舞弊风险、建议全额计提坏账准备”的结论，理由仅为“模型内部计算评分超过阈值”。审计人员无从了解判断依据，难以验证结论的合理性，也无法向管理层与监管机构解释风险成因，模型输出因而难以作为正式审计证据，整个审计发现建立在一项无法验证的断言之上。

在可解释模型下，同一笔交易输出完整证据链：特征 A（金额 500 万元，权重+40%）金额巨大且账龄远超行业平均信用期，触发高敏感规则；特征 B（逾期付款历史，+30%）表明客户还款意愿较低；特征 C（合同条款，+30%）约定付款期限 60 天而信用期已超出 120 天，构成严重违约。三项特征共同支撑风险结论；反事实验证进一步显示，若金额低于 50 万元，风险评分将降至“关注”等级，有助于评估结论对关键变量的敏感程度。比较表明，在保持结论一致性的前提下，可解释模型显著提升了特征归因的透明度与证据链的完整性，使审计大模型由“结果工具”转变为“决策伙伴”。

需说明的是，本示例仅作示意，实际应用中的特征、权重与阈值须经业务数据校准与试点验证。

六、挑战、对策与结论

（一）主要挑战

本框架的部署仍面临四方面挑战：一是大语言模型存在“幻觉”风险，自动生成的解释可能与事实不符，

而貌似合理但事实上错误的解释比没有解释更危险；二是因果图的构建依赖高质量数据与领域知识，在数据治理薄弱、标准不一的企业中实施成本较高；三是审计人员的数字素养与接受度决定框架能否真正落地；四是可解释性与模型性能存在权衡，实时生成解释的计算成本可观。

（二）应对策略

相应对策有四：坚持“人在回路”机制，由审计人员复核修正模型输出并承担最终决策责任，反事实解释须与独立证据交叉验证；强化数据治理与模型全生命周期管理，将可解释性评估纳入模型测试与验证环节[2]；培育“T 型”复合型人才，提升审计人员的人工智能素养、提示工程能力与输出甄别能力；实施渐进式部署，从低风险、高价值场景切入，试点后逐步扩围。还需强调，框架落地是一项组织变革，领导层承诺、跨部门协同与激励对齐不可或缺。

（三）结论

可解释性是审计大模型部署的前提。本文构建的框架通过“输入与特征提取—因果推理—可解释输出”的三阶段递进，将模型输出由“评分”转化为可追溯、可验证的“因果证据链”，回应了审计证据适当性、监管合规与软约束的三重要求。未来研究可进一步探索因果图模型与审计知识图谱的融合、多模态审计证据链的构建，以及可解释性框架在工程项目审计、物资采购审计、经济责任审计等场景中的适配与推广。

参考文献：

[1] 陈伟. DeepSeek 人工智能大模型在审计中的应用研究[J]. 中国注册会计师. 2025(3): 49-54.

[2] 刘政, 郭旭, 邱实. 基于人工智能大模型的企业审计画像数据治理体系探索[J]. 中国内部审计. 2025(4): 28-35.

[3] 张莉, 魏华. 人工智能赋能内部审计：深度融合与创新发展[J]. 中国内部审计. 2025(4): 36-39.

[4] 赵鑫. 人工智能与大数据在内部审计中的应用[J]. 中国乡镇企业会计. 2024(12): 55-57.

[5] 胡云, 姚文. 新质生产力背景下企业智能内部审计路径探索[J]. 中国农业会计. 2024, 34(23): 72-74.

[6] 罗斌, 曾静. 基于人工智能大模型的智能审计模式研究[J]. 中国注册会计师. 2025(9): 79-83.

[7] 张莉. 大语言模型在审计领域的潜在应用与技术路径[J]. 会计之友. 2024(24): 2-9.

[8] Lundberg S M, Lee S I. *A unified approach to interpreting model predictions*[C]. Advances in Neural Information Processing Systems. 2017, 30: 4765-4774.

[9] Pearl J. *Causality: Models, reasoning, and inference*[M]. 2nd ed. Cambridge: Cambridge University Press. 2009.

[10] Rudin C. *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*[J]. Nature Machine Intelligence. 2019, 1(5): 206-215.