AI大模型驱动的智能博弈财务舞弊识别系统构建

——基于深交所监管数智化转型实践

深圳证券交易所财务舞弊监管AI大模型课题组

(深圳证券交易所，广东 深圳 518038)

摘要：运用好人工智能等新兴技术手段高效识别违法违规线索和风险隐患，提升资本市场监管科学性、有效性，是落实“十五五”规划建议要求和中央金融工作会议精神的重要举措。本文基于深交所监管数智化转型实践，深入探讨如何应用大模型识别财务舞弊问题，创新提出“舞弊识别思维链提示词+结构化多维信息工作底稿+多智能体博弈对抗”的智能化舞弊识别理论范式，开发构建大模型驱动的智能博弈财务舞弊识别系统，针对性解决了当前应用大模型识别财务舞弊的障碍，有效运用大模型对上市公司财务舞弊风险进行“拟人化”智能推理分析，并基于分析结果向监管人员提示上市公司可能存在的舞弊风险以及监管应对建议。相关实测结果表明，该系统的舞弊识别精准度较高，漏报与误报得到较好控制，有效弥补了机器学习模型识别舞弊的短板，以及利用专家规则模式下舞弊识别指标孤立、缺乏综合推理分析的问题，切实发挥对监管人员的智能辅助作用。深交所构建AI大模型驱动的智能博弈财务舞弊识别系统是响应国务院“人工智能+”行动意见、推动金融监管数智化转型的重要探索。

关键词：大模型；财务舞弊；舞弊识别思维链；多维信息工作底稿；智能体；博弈对抗

Abstract: The effective utilization of emerging technologies such as artificial intelligence to efficiently identify illegal violations and potential risks, and enhance the scientific rigor and effectiveness of capital market supervision, represents a crucial measure for implementing the requirements of the 15th Five-Year Plan proposal and the spirit of the Central Financial Work Conference. Based on the practice of digital and intelligent transformation of supervision at the Shenzhen Stock Exchange (SZSE), this paper thoroughly explores the application of large language models in identifying financial fraud issues. It innovatively proposes an intelligent fraud detection theoretical paradigm of “fraud detection chain-of-thought prompting + structured multi-dimensional information working papers + multi-agent game-theoretic confrontation”, and develops a large language model-driven intelligent game-theoretic financial fraud detection system. This system specifically addresses current obstacles in applying large language models for financial fraud detection, effectively employing large language models to conduct “human-like” intelligent reasoning and analysis of financial fraud risks in listed companies, and provides regulatory personnel with alerts regarding potential fraud risks and regulatory response recommendations based on the analysis results. Relevant empirical testing results demonstrate that the system achieves high precision in fraud detection, with false negatives and false positives being well controlled. It effectively compensates for the shortcomings of machine learning models in fraud detection, as well as the problems of isolated fraud detection indicators and lack of comprehensive reasoning analysis under expert rule-based approaches, genuinely fulfilling its role as an intelligent assistant to regulatory personnel. The SZSE’s construction of an AI large language model-driven intelligent game-theoretic financial fraud detection system represents an important exploration in responding to the State Council’s “AI+” action initiative and promoting the digital and intelligent transformation of financial supervision.

Key words: large language models, financial fraud, chain-of-thought for fraud detection, multi-dimensional information working papers, intelligent agents, game-theoretic confrontation

作者简介：深圳证券交易所财务舞弊监管AI大模型课题组。课题负责人：陈文新，会计硕士，深圳证券交易所监管执行部副总监，研究方向：现场监管、会计准则；叶茂，女，经济学硕士，深圳证券交易所创业板公司管理部执行经理，研究方向：公司监管、舞弊监管；许明峰，工程硕士，深圳证券交易所信息科技二部执行经理，研究方向：人工

智能、大数据。课题组成员：邱晓明、兰茹婷、黄锦鸿、薛傲、李铁照、郭旺振、邓媛尹、王佳奇、张慧慧、单益峰等，均任职于深圳证券交易所，研究方向：会计监管、现场监管、科技监管、公司监管等。

中图分类号：F234.4　　**文献标识码：**A

一、引言

党的二十届四中全会通过的《中共中央关于制定国民经济和社会发展第十五个五年规划的建议》(以下简称“十五五”规划建议)明确提出，要全面加强金融监管，丰富风险处置资源和手段，构建风险防范化解体系。中央金融工作会议强调，坚持把防控风险作为金融工作的永恒主题，对风险早识别、早预警、早暴露、早处置。财务造假是资本市场重要的风险源，近年来，其隐蔽性和复杂性显著增强，对风险监测预警的及时性和精准性提出了更高要求。充分运用人工智能等现代信息技术手段高效识别违法违规线索和风险隐患，提升资本市场监管的科学性、有效性，是落实“十五五”规划建议要求和中央金融工作会议精神、防范化解金融风险、维护资本市场稳定和保护投资者合法权益的重要举措。

有效识别财务舞弊风险是监管实践中打击财务造假的首要环节。在大语言模型(以下简称大模型)技术取得突破以前，学术界已有不少关于机器学习模型特别是深度学习模型在财务舞弊识别领域应用的研究。例如，从上市公司财务会计信息角度切入，构建机器学习模型，实现对舞弊风险的识别与预警(Cecchini et al.，2010；Bao et al.，2020；周卫华等，2022)；基于舞弊三角理论，将公开信息转化为反映压力、机会、借口等方面的指标用于舞弊识别(Lin et al.，2015)；利用年报“管理层讨论与分析”等叙述性披露文本提取语言信号或信息，辅助识别舞弊风险(Purda and Skillicorn，2015；Craja et al.，2020；赵纳晖和张天洋，2022)等。在此基础上，有研究将财务信息与股评文本等市场信息一并纳入模型(曹策等，2024)，或者以业绩电话会内容为基础，引入多模态学习框架(Kaikaus et al.，2022)，以进一步提升舞弊识别能力。此外，机器学习等人工智能技术也被用于企业财务困境预测或财务分析等场景(吴世农等，2021；汪春华，2023)。

然而，在财务舞弊识别方面，机器学习模型存在舞弊判别结果仅简单关注“是否舞弊”这一笼统的“二元分类”问题、舞弊判别过程对非财务和非结构化数据的应用较少，以及舞弊判别结果可解释性不足等三大短板(田莉等，2025)。2025年以来，以DeepSeek-R1为代表的推理大模型发布，国产大模型推理能力取得长足进展。相较于传统大模型，DeepSeek-R1的强化学习框架促进了高级推理模式的自主涌现，包括自我反思、多轮验证和动态策略调整。经此训练的模型在数学、编程、工程领域等可验证任务中表现卓越，超越了依赖人工示范数据进行传统监督学习的同类模型(Guo et al.，2025)。因此，深度推理大模型更适用于财务舞弊识别等需要综合分析、多维判断的应用场景。相关研究也表明，大模型能够应用于数据采集、跨多文档与非结构化文档解析、证据链构建、人机交互以及优化现有规则模型等工作，进而提升财务舞弊识别效能(叶钦华和黄世忠，2025；Bai et al.，2025)。

为深入落实“十五五”规划建议要求和中央金融工作会议精神，积极响应国务院“人工智能+”行动意见，更好履行新形势下监管职责，推动监管数智化转型，2025年2月深圳证券交易所(以下简称深交所)启动了财务舞弊监管AI大模型的研究和项目建设工作，以人工智能赋能上市公司财务舞弊识别。本文的贡献具体包括以下三个方面：

(1)对国内主流通用大模型自身的会计专业能力和舞弊识别效果进行了测试，发现国内主流通用大模型在直接使用的情况下存在四大障碍，即并不具备识别舞弊的成熟系统方法论、无法完整读取并分析图表文混排长文本、难以全面准确获取舞弊识别所需多维信息，以及难以合理平衡防止漏报与控制误报。

(2)将财务舞弊动因、识别特征理论与大模型思维链、注意力机制、博弈对抗理论相结合，创新提出“舞弊识别思维链提示词+结构化多维信息工作底稿+多智能体博弈对抗”的智能化舞弊识别理论范式。一是为大模型提供清晰的舞弊识别方法论指引，即构建了一套融

合专家经验、系统性、流程化且易于大模型理解执行的舞弊识别思维链提示词体系，引导大模型以“拟人化”“专家化”方式识别舞弊；二是建立多维信息处理机制，即设计了一套可被大模型完整阅读且有效使用的结构化多维信息工作底稿，确保大模型有效读取识别舞弊所需的多维信息；三是建立可有效平衡舞弊漏报与误报的舞弊识别智能体结构体系，即通过引入“博弈对抗”理念，构建了一种对舞弊识别思维链提示词和结构化多维信息工作底稿进行优化编排的多智能体结构，有效提升舞弊识别精度与执行效率，并在防止漏报和控制误报之间实现合理平衡。三项核心创新要素逐一解决了大模型识别舞弊过程中“怎么看”“看什么”“如何更高效、有效地看”等关键性问题，针对性破解了当前大模型识别舞弊时面临的四大障碍，有效缓释了大模型幻觉[1]、随机性[2]等问题对舞弊识别结果的负面影响，进一步提升了大模型舞弊识别分析过程的稳定性和结果的置信度。

(3)在此基础上，开发了大模型驱动的智能博弈财务舞弊识别系统(以下简称舞弊识别大模型系统)，在应用大模型识别财务舞弊领域取得新突破。该系统通过充分利用大模型深度推理能力，在上市公司年报披露后，自动对上市公司财务舞弊风险进行“拟人化”智能推理分析，实现“生成判断结论”和“提出工作建议”两大目标，向监管人员提示上市公司可能存在的具体舞弊风险以及分析依据，并提出监管应对建议。此外，相比机器学习模型、单纯利用专家规则模式，舞弊识别大模型系统充分使用结构化财务数据和各类非财务数据、非结构化数据，以及对所触及的舞弊识别特征进行综合分析和逻辑推理，可以更好识别第三方配合造假等复杂、新型财务舞弊行为，识别结果明确提示整体舞弊风险等级，明确指向具体舞弊领域、舞弊模式以及舞弊手法，并提供分析依据和针对性监管应对建议，大幅提升了舞弊识别结果的可解释性和可使用性。

二、通用大模型直接分析年报识别舞弊的初步尝试

在对信息科技、金融、高校等多领域机构进行走访调研的基础上，本文通过注册会计师考试试题以及直接提供部分样本公司年报全文的方式，对国内主流通用大模型自身的会计专业能力和舞弊识别效果进行了初步测试。测试结果表明，国内主流通用大模型在直接使用的情况下，财务舞弊识别结果存在较明显漏报和误报，且分析过程中幻觉较多、随机性较强，难以满足监管工作需要。经测试分析，应用通用大模型直接分析年报识别舞弊风险存在以下四个方面的障碍。

(一)通用大模型不具备识别舞弊风险的成熟系统方法论

经测试，通用大模型基于深度推理能力，能够在标准化、封闭式注册会计师考试中取得较好成绩，对于会计准则执行有较好的判断力，为识别舞弊奠定了一定会计专业基础。但近年来财务造假复杂性、系统性、隐蔽性进一步增强，舞弊公司披露的年报在形式上并不一定存在明显违反会计准则的会计差错，从年报中直接发现会计处理疑点继而识别舞弊风险的模式已难以适应新形势，识别舞弊须有成熟系统的舞弊识别方法论支撑。例如，上市公司可能在客户、供应商等第三方配合下，通过虚构交易、体外资金循环方式虚增收入，其会计处理在形式上并不存在明显违反会计准则的情形，必须结合舞弊动机、虚构交易类舞弊典型识别特征予以综合研判，很难通过简单判断年报中是否存在会计处理差错即识别出相关舞弊风险。因此，在未经财务舞弊识别领域专门训练的前提下，通用大模型并不具备识别舞弊风险的系统成熟方法论，难以精准有效识别非标准化、开放式的各类复杂财务舞弊风险。

(二)通用大模型无法完整读取并分析图表文混排长文本

直接上传年报全文的测试表明，受可读取记忆的上下文词元(Token)数量限制、注意力机制等因素影响，通用大模型难以完整读取和理解图表文混排的长篇幅上市公司年报全文，且传统年报格式(Pdf或Word)也非大模型擅长识别的格式类型，导致其对年报信息的识别、获取存在明显遗漏且读取质量低下，进而无法对年报信息进行有效的舞弊风险分析。例如，如无法有效获取年报中的营业收入或营业成本数据，则无法计算应收账款周转率、存货周转率等财务指标进行基础财务分析，更无法有效执行舞弊识别工作。

(三)通用大模型难以全面准确获取舞弊识别所需多维信息

实践中，监管人员为有效识别上市公司舞弊风险，除需阅读当年披露的年报信息外，通常需要结合历史财

务信息、临时公告信息、可比公司信息以及客户、供应商、关联方等相关主体信息等多维信息进行综合判断、交叉验证。但测试发现，通用大模型难以全面准确获取前述多维信息，亦无法有效甄别相关信息的专业性、权威性，严重限制了大模型推理能力的发挥，导致舞弊识别结果的精准性和全面性存在不足。

(四)通用大模型难以合理平衡防止漏报与控制误报

测试发现，在执行舞弊识别任务时，大模型较易形成“天然”的舞弊指控倾向，舞弊识别结果存在广泛报警，难以在防止漏报与控制误报之间取得合理平衡。虽然舞弊识别的主要目的是充分有效识别出存在舞弊风险的公司(即防止漏报)，但在实践中，如为降低漏报率而容忍较高误报率，监管人员将耗费大量时间精力甄别错误预警信息，难以发挥人工智能辅助提高监管效率的作用。因此，能有效服务于监管实践的舞弊识别系统，须合理平衡防止漏报与控制误报。

三、舞弊识别大模型系统舞弊识别的理论范式

前述尝试发现，通用大模型尚不能直接有效识别舞弊风险，但考虑到DeepSeek-R1的会计专业能力和舞弊识别效果综合表现相对较优，具有开源属性，故可被确定为本系统的基座大模型。相关研究指出，尽管大模型在专业考试中表现尚可，但在实际审计任务中可靠性较低(Street et al.，2023)，财务舞弊识别仍需专家经验协同(叶钦华和黄世忠，2025)。构建成熟系统舞弊识别方法论、汇聚全面准确的舞弊识别所需多维信息，并结合通用大模型能力特点，将前述方法论与多维信息以适当方式提供给通用大模型，进而引导其“拟人化”“专家化”识别舞弊，成为解决通用大模型舞弊识别问题的可行思路。因此，针对初次尝试中发现的障碍，基于财务舞弊及大模型领域相关理论，以及深交所长期以来所积累的舞弊风险监测经验和丰富的结构化数据，本文研究提出“舞弊识别思维链提示词+结构化多维信息工作底稿+多智能体博弈对抗”的舞弊识别理论范式。

(一)构建舞弊识别大模型系统的理论基础

1. 财务舞弊相关理论及研究

财务舞弊研究理论成果丰硕，主要包括对财务舞弊动因的研究和对财务舞弊识别特征的研究。

(1)财务舞弊动因相关理论及研究

Albrecht et al.(1995)提出的舞弊三角理论认为，财务舞弊的产生由压力、机会和借口三要素构成，其中业绩考核等催生舞弊压力，内控缺陷、信息不对称及监督不力等形成舞弊机会，借口要素则是舞弊者为自己的行为找到的合理化理由。Bologna et al.(1993)提出的GONE理论，对舞弊动因进行了更全面的分析，增加的“暴露”因子包括舞弊行为被发现的概率，以及被发现后相应惩罚的性质与力度，认为舞弊是由贪婪(Greed)、机会(Opportunity)、需要(Need)和暴露(Exposure)四种因素综合作用的结果。Rezaee(2005)提出的CRIME理论，则从舞弊行为人(Cooks)、舞弊手段(Recipes)、动机(Incentives)、监管机制(Monitoring)以及舞弊结果(End-Results)五个方面，进一步分析了财务舞弊行为的产生机制与影响。CRIME理论除了强调关注舞弊动机、监督机制外，还关注到大部分财务舞弊行为由公司高级管理人员等“关键少数”实施，涉及较多舞弊手段，且最终造成严重负面后果。韦琳等(2011)和洪荭等(2012)分别基于舞弊三角理论与GONE理论对我国上市公司舞弊行为进行了本土化验证，证实了前述理论在不同制度环境下的适用性。

上述理论从不同角度揭示了舞弊发生的动因，回答了“为什么会产生舞弊”以及“在什么条件下容易产生舞弊”的问题。这说明财务舞弊是在特定条件下产生的可预测行为，而非完全随机的事件，通过系统考察与财务舞弊相关的动因，既可确定形成财务舞弊的关键因素，也可为梳理舞弊识别特征提供方向性指引。例如，在识别舞弊风险时，既需关注公司是否面临舞弊压力，是否存在内控缺陷等舞弊机会，也需要关注董监高等“关键少数”的行为与道德情况、所涉及的具体舞弊手段等多种因素。

(2)财务舞弊识别特征相关理论及研究

财务舞弊识别特征包括财务信息和非财务信息识别特征两类。在财务信息识别特征方面，Persons(1995)通过财务杠杆、资本周转率等关键财务指标构建了舞弊预测模型。Lee et al.(1999)从现金流视角出发，揭示了盈余与经营现金流差异在识别财务舞弊中的预警作用。Beneish(1999)基于营业收入指数、应收账款指数、毛利率指数等八个财务信息提出M-Score模型，以及以此为基础，

Dechow et al.(2011)提出F-score模型，钱苹和罗玫(2015)结合我国市场特点提出C-score模型，成为使用财务信息识别舞弊的代表性成果。F-score模型和C-score模型中均包含非财务信息，如是否发生再融资、股权集中度等。

就舞弊识别逻辑而言，财务信息是舞弊结果的最终呈现，若要实现对财务舞弊的更精准研判，需要结合更多非财务信息。Albercht et al.(1986)提炼出的可用于预测财务舞弊的预警信号(即“红旗标志”)包括大量非财务信息。Dechow et al.(1996)研究发现公司治理结构缺陷与财务舞弊存在显著关联。吴革和叶陈刚(2008)也从财务报表与公司治理两个方面，对相关指标与财务舞弊之间的关系进行了实证分析，构建了舞弊识别指标体系。陈彬和刘会军(2012)研究做空机构的报告，系统总结了与财务舞弊相关的特征，其中包括销售依赖代理商、更换过审计机构等。胡丹(2018)对相关上市公司利润调节的风险表征进行了研究，认为业绩承诺或股权激励精准达标、突击更换审计机构、年底突击交易以及管理层发生重大变化等表征和利润调节具有强正相关性。叶康涛和刘金洋(2021)验证了销售量、生产量、库存量等经营信息在舞弊识别中的价值。叶钦华等(2022)基于复式簿记和会计信息系统论构建了涵盖财务税务、行业业务、公司治理、内部控制与数字特征的五维度财务舞弊识别框架，推动财务舞弊识别向多维立体范式转变。

上述研究回答了“如何识别舞弊”以及“什么是舞弊危险信号”的问题，为梳理具体舞弊识别特征提供了方法论支持。这表明通过充分挖掘有效的各类舞弊识别特征，充分使用包含财务信息、非财务信息在内的多维信息，有助于更精准地识别财务舞弊风险。

2. 大模型相关理论及研究

在构建成熟系统舞弊识别方法论的同时，还需结合大模型的能力特点与固有技术局限，进一步考虑以何种方式将舞弊识别方法论以及舞弊识别所需的多维信息提供给大模型。

主流大模型的基石Transformer架构，具备自注意力机制与强大的序列上下文建模能力(Vaswani et al.，2017)。随着模型参数规模与训练数据量跨越关键阈值，大模型表现出显著的“涌现能力”，即在小型模型中不存在但在大型模型中存在的能力(Wei et al.，2022a)。同时，引入指令微调能帮助大模型执行许多训练时未见过的任务，对跨任务泛化能力有着积极影响(Wei et al.，2022b)。强化学习等技术被用于激励与增强模型能力，如InstructGPT(GPT3改进版本)在真实性方面有所改进，并且减少了有害内容的生成(Ouyang et al.，2022)；DeepSeek-R1版本则进一步提升了推理能力(Guo et al.，2025)等。但大模型在分词过程也经常会出现错误和歧义(Sgantzos et al.，2023)，并且存在幻觉、随机性等固有技术局限(Bender et al.，2021；Ji et al.，2023)。

(1)思维链相关理论及研究

思维链相关理论研究通过引入序列化推导机制，首次系统性地提出并验证了“思维链提示”方法的有效性，即当大模型经过恰当引导，被要求“分步骤说明推理过程”时，其在复杂判断任务中的准确性和一致性将显著提升(Wei et al.，2022c)。审计领域的大模型应用研究表明，通过融合审计专家知识与风险先验的推理引导，配合智能体编排，大模型能够有效克服单点推理的局限性，不仅在舞弊风险识别的召回率上取得显著提升，还能实现更清晰的证据定位、更强的可解释性(Bai et al.，2025)。

(2)注意力机制相关理论及研究

由于注意力机制的计算复杂度，大模型在长文本推理时会受限于计算资源(Vaswani et al.，2017)。同时，当改变相关信息的位置时，大模型性能亦会明显受到影响，特别是当相关信息出现在输入上下文的开头或结尾时，性能往往最高；而当模型必须在长上下文中间获取相关信息时，性能则会明显下降，即使是明确的长上下文模型也是如此，这表明当前的语言模型不能稳健地利用长输入语境中的信息(Liu et al.，2024)。相关基准测试进一步揭示，由于注意力分配的非均匀性，大多数声称支持长上下文的模型在处理复杂多跳任务时的“有效长度”远低于宣称值(Hsieh et al.，2024)。

(3)博弈对抗相关理论及研究

Goodfellow et al.(2020)提出了对抗网络框架，即生成模型与判别模型对抗。生成模型基于模型产出内容，判别模型则识别确定相关内容是来自模型生成还是来自真实世界。通过二者博弈对抗，进而有助于提升模型生成逼真数据的能力。该研究为提升大模型推理分析精准性提供了重要思路借鉴。

(二)舞弊识别大模型舞弊识别理论范式的形成与内涵

基于上述理论研究，为了更好地解决通用大模型识别舞弊存在的问题，本文提出以舞弊识别思维链提示词向大模型提供成熟系统舞弊识别方法论，通过开发文本长度与位置受控的结构化多维信息工作底稿，向大模型提供舞弊识别所需的多维信息，通过多智能体博弈对抗，在将舞弊识别思维链提示词和结构化多维信息工作底稿进行优化编排构建指控分析智能体基础上，引入辩护分析智能体进行博弈对抗，更有效实现防止漏报与控制误报间的平衡，进而最终形成“舞弊识别思维链提示词+结构化多维信息工作底稿+多智能体博弈对抗”的智能化舞弊识别理论范式。进一步的，将舞弊识别理论范式工程化落地，开发形成舞弊识别大模型系统。

从具体内涵来看，(1)舞弊识别思维链提示词指在系统研究常见舞弊动机、高风险舞弊领域典型舞弊模式和手法，提炼总结系列舞弊识别特征后，构造出一套系统性、流程化且易被大模型理解和执行的舞弊识别步骤，再将该舞弊识别步骤通过提示词工程输入大模型，引导其按照所设定的方法对舞弊风险予以识别。(2)结构化多维信息工作底稿指基于舞弊识别需求，整合历史与当期财务信息、临时公告信息、可比公司信息以及客户、供应商、关联方等相关主体信息等多维信息，形成一份内容丰富、层次分明的“浓缩整合版年报”。(3)多智能体博弈对抗指在将舞弊识别思维链提示词和结构化多维信息工作底稿进行优化编排基础上，在系统中设置三类智能体分别执行不同任务：指控分析智能体专职负责寻找舞弊迹象与线索，预警舞弊风险；辩护分析智能体专职负责就部分指定的舞弊迹象与线索进行合理性解释，与指控分析智能体进行博弈对抗，旨在对不合理的指控予以纠偏；裁决分析智能体则在充分评估指控分析智能体和辩护分析智能体意见的基础上，经综合研判后就公司舞弊风险等级作出最终裁决，并提示具体舞弊领域、舞弊模式、舞弊手法以及监管应对建议。

四、舞弊识别思维链提示词工程

本文在借鉴前述财务舞弊相关理论成果、吸收各类财务舞弊识别实践经验的基础上，经充分考虑相关舞弊识别特征在监管实践中的适用性与所涉及底层信息可获取性，对各类舞弊识别特征进行了系统化、类型化、层次化梳理，形成以“舞弊动机+舞弊特征→舞弊领域+舞弊模式+舞弊手法”为路径的舞弊识别思维链，并通过提示词工程提供给通用大模型，以实现舞弊识别过程“拟人化”“专家化”，有效提升通用大模型舞弊识别专业能力，从而针对性解决通用大模型“不具备识别舞弊风险的成熟系统方法论”问题。此外，通过舞弊识别思维链的引导约束，一定程度有助于解决通用大模型“难以在防止漏报与控制误报之间取得合理平衡”问题，也能有效控制大模型因幻觉、随机性问题可能引发的负面影响，增强舞弊识别分析过程的可靠性。

(一)舞弊识别思维链的研究与构造思路

1. 舞弊识别思维链研究思路

舞弊识别思维链研究主要分为两方面：一方面，构造基础舞弊风险识别思维链，涵盖通用型舞弊风险识别特征和舞弊动机识别特征。另一方面，构造特定舞弊领域舞弊识别思维链，即通过研究高风险舞弊领域的舞弊模式、舞弊手法，提炼总结对应的舞弊识别特征及其组合，形成具体舞弊领域指向型舞弊风险识别特征。详见图1。

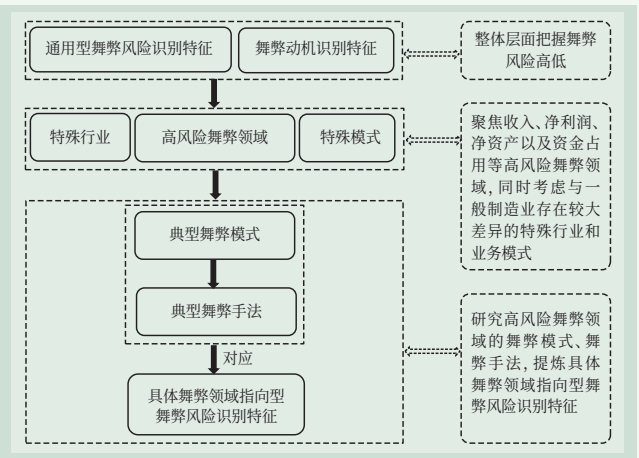

图1 舞弊识别思维链研究思路示意图

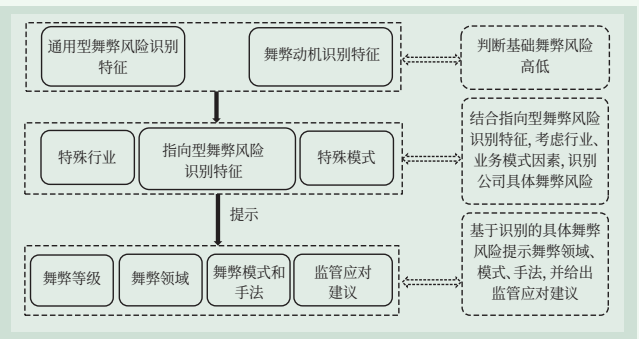

图2 舞弊识别思维链构造思路示意图

2. 舞弊识别思维链构造思路

在上述对思维链进行研究基础上，本文构造了舞弊识别思维链：一方面，通过通用型舞弊风险识别特征和舞弊动机识别特征判断上市公司基础舞弊风险高低。另一方面，结合上市公司所触发的指向型舞弊风险识别特征，考虑特殊行业、业务模式等因素，识别具体舞弊风险，研判舞弊风险等级，提示可能涉及的舞弊领域、舞弊模式以及舞弊手法，并给出监管应对建议。详见图2。

基于舞弊识别思维链研究和构造思路，本文从通用型舞弊风险识别特征、舞弊动机识别特征以及指向型舞弊风险识别特征三个维度系统构造了数百条舞弊风险识别特征，并据此形成了舞弊识别思维链。同时，结合监管经验，本文还将通用型与指向型舞弊风险识别特征进一步区分为高风险类特征、一般风险类特征，通过赋予不同识别特征差异化风险权重，更精准反映公司舞弊风险高低。

(二)构造系统化舞弊识别思维链提示词

提示词是大模型有效执行思维链的重要基础。在工程化实施环节，为便于大模型精准运行舞弊识别思维链，必须将思维链内容转换为便于大模型理解执行的提示词语句。为提高舞弊识别结果的全面性、准确性，有效控制对舞弊风险的漏报与误报，系统开发过程中对舞弊识别思维链提示词进行了大量、系统的测试调优，以构造结构清晰、表述精准、易于理解的舞弊识别思维链提示词体系。

1. 构造基于思维链识别特征的核心业务提示词

思维链识别特征转换为提示词时的表达方式、关联方式、结构层次以及所阐述的应用逻辑是否合理、背景信息是否清晰，会对舞弊识别结果产生显著影响。

(1)优化识别特征表述方式

思维链识别特征表述是否便于大模型理解、执行，直接影响大模型舞弊识别效果。例如，大模型对包含或然选择连词的复合型识别特征容易理解错误，会在仅满足部分条件时即认为触及特征。为此，将此类复合型识别特征拆解为多条独立识别特征，通过分别列示形成更清晰的判断路径，有效提高大模型执行思维链准确性。又如，对于识别特征中“大幅上升、明显下降、显著超过”等程度副词，大模型的判断与监管人员存在差异，可能导致误报或漏报，有必要基于专家经验和数据统计情况明确相关识别特征中程度副词具体阈值。

(2)强化识别特征指标关联

财务舞弊行为对相关会计科目或财务指标的影响通常具有联动性。为此，在构造提示词时，本文对识别特征强化指标关联，形成“组合型识别特征”，以更精准定位舞弊行为，降低误报风险。例如，在识别资金占用风险时，将相关会计科目进行组合，设置关联识别特征，更精准定位相关异常行为。

(3)完善识别特征结构层次

公司触及的识别特征数量是大模型确定舞弊风险高低的重要依据之一，部分识别特征实质指向同一类风险，如不对此类识别特征从结构层面予以归类整合，可能导致大模型误认为触发多项识别特征，错误提升舞弊风险等级。[3]例如，多条识别特征实质均为同一指向时，如同时触发，则通过提示词引导大模型按照所触及风险大类判断风险等级，避免重复考虑同类识别特征。

同时，在监管实践中，触发某些识别特征并不必然由舞弊引发。因此，除了高风险类特征、一般风险类特征外，在思维链中还新增“关注类识别特征”涵盖此类“中性特征”，要求大模型结合其他舞弊识别特征触发情况综合判断是否将“关注类识别特征”用于判断舞弊风险，避免错误认定存在舞弊风险。

(4)明确识别特征应用逻辑

思维链中不同舞弊识别特征具有不同的定位和风险权重，大模型能否合理应用不同识别特征的识别结果，能否准确把握不同识别特征之间的关系，直接影响舞弊识别效果。因此，有必要准确说明不同识别特征的应用逻辑，使大模型分析结果更契合监管人员的分析逻辑。

例如，基础舞弊风险对大模型判断公司整体舞弊风险的影响较大，但是基础舞弊风险仅表明具有初步的舞弊迹象和动机，并不直接指向公司所实施的具体舞弊行为。为使大模型合理应用基础舞弊风险识别结果，通过构造提示词，要求大模型必须综合考虑基础舞弊风险与指向型舞弊风险识别结果，并明确基础舞弊风险识别特征中与舞弊风险存在更直接关联的识别特征类型，避免产生误判。

又如，大模型执行资金占用舞弊思维链时，存在不当高估舞弊动机识别特征权重的倾向。根据监管实践，在判断公司是否存在资金占用风险时，还应考虑指向型

舞弊风险识别特征，不能简单因未识别到舞弊动机识别特征，即认为公司资金占用风险较低。因此，通过构造提示词，进一步明确舞弊动机识别特征属于舞弊风险识别的“加强项”而非“必要条件”。

(5)补充识别特征具体含义

考虑到大模型具有较强的“拟人化”能力，可以通过对部分识别特征补充说明其具体含义，帮助大模型更好理解识别特征内涵，从而更准确地分析相关识别特征可能指向的风险，而非仅机械地判断是否触及相关识别特征。同时，因大模型具有“举一反三”泛化能力，对于部分需进行主观判断的识别特征，可以通过举例方式进行补充说明，对大模型提升分析准确性具有正向价值。

2. 构造规范大模型分析输出的关键约束提示词

测试发现，仅运用基于思维链识别特征构造的核心业务提示词，在分析相关案例时仍存在较多误报与漏报，并且所输出的舞弊分析报告格式规范性较差。据此，本文从多维度构造了进一步防范误报与漏报风险、规范舞弊分析报告输出格式的关键约束提示词。

(1)引导准确判断舞弊风险领域

在监管实践中，虚构交易类收入舞弊和资金占用舞弊识别特征天然相互重叠，即公司将资金套取到体外后，可能用于虚构销售回款虚增收入，也可能用于资金占用。若不加以引导，大模型在判断相关识别特征指向的舞弊风险领域时可能存在混淆风险。

对此，通过在思维链中加入引导提示词，一方面要求大模型复核虚构交易类收入舞弊和资金占用舞弊相重叠识别特征的判断结论是否一致，降低幻觉与随机性影响；另一方面要求大模型在研判时综合考虑基础舞弊风险、所触及的虚构交易类收入舞弊识别特征和资金占用舞弊识别特征数量，以及所触及的具体舞弊识别特征内容，以更精准定位是虚构交易类收入舞弊还是资金占用舞弊。

(2)引导合理评定舞弊风险等级

因大模型对于“高舞弊风险”“低舞弊风险”内涵的认知并非每次均与监管人员保持一致，导致大模型在准确识别部分公司具体舞弊风险点的情况下，最终评定的舞弊风险等级仍可能与监管人员的判断存在差异。

对此，通过在思维链中明确各类舞弊风险等级的内涵，引导大模型更精准评定相关公司舞弊风险等级。例如，“高舞弊风险”意味着上市公司舞弊风险极高，舞弊迹象非常明显，监管机构可能需对其实施现场检查或立案调查；而对于评定为“低舞弊风险”的公司，监管机构通常需了解公司所触及的识别特征是否具有合理性。

(3)引导有效分析使用定量数据

在对部分思维链识别特征明确设定定量阈值后，必须引导大模型准确分析相关数据，并合理使用分析结果。具体如下：

一是统一数据比较方法。对于绝对金额类和百分比类数据，明确要求大模型应分别关注相对变动幅度和绝对百分点差异，并统一不同类型数据比较方法。

二是关注数据波动烈度与时长。为避免思维链对单项财务指标长期、剧烈异常波动“视而不见”，明确要求大模型考虑相关财务指标异常变动的剧烈程度、持续时间等因素。例如，当公司仅触及少量识别特征但相关指标波动程度非常剧烈时，大模型将充分考虑该事项对舞弊风险的影响。

(4)引导控制高频误报漏报因素

一是要求大模型把握重要性标准。若公司某一报表项目金额极低时，因低基数效应，其增长率、周转率等指标波动较易出现异常，可能导致大模型“小题大作”误报舞弊风险。为此，在思维链中对报表项目金额、与相关主体交易金额或往来科目余额等均设置了执行相关识别特征的重要性标准。

二是要求大模型关注识别特征之间协同性。为避免大模型孤立判断相关识别特征的分析结果引发广泛报警，在输出分析结果的提示词中明确要求大模型充分关注识别特征之间是否具有协同性，如果协同考虑后均指向某一舞弊风险，则误报可能更低。同时，也明确提示，如果不同识别特征之间不存在明显协同性，并非表明不存在舞弊风险，仍要求大模型基于识别情况予以综合分析。

三是要求大模型不得预设立场。监管实践表明，大型企业也存在舞弊风险。但在某些案例中，大模型思考过程中却认为“作为大型知名企业集团，其部分财务异常可能具备合理解释”，从而降低了对舞弊风险判断的敏感性，存在明显预设立场问题。为此，通过在提示词中加入约束性指令，引导大模型严格基于事实依据进行分析，不得预设立场。

四是纠正部分常识性误报。例如，根据识别特征要求，大模型在识别隐性关联方时，会比对相关主体部分工商信息是否存在重叠或相似。但对于上市公司已公开披露的关联方，因上市公司已如实披露双方具有关联关系，无需再识别此类主体是否为隐性关联方。又如，在提示词中明确引导，当工作底稿中缺失部分数据时，通常是客观原因导致的数据获取失败，而非公司故意隐瞒未披露相关信息。

(5)引导统一结果输出呈现体例

因大模型输出具有随机性，如不加引导约束，则其每次识别输出呈现的分析结果可能存在较大差异。为进一步标准化输出结果呈现形式，通过提示词明确了输出结果的体例要求。例如，明确设定了舞弊识别综合分析报告输出的框架结构以及各部分应呈现的具体内容，并要求大模型必须输出所触及识别特征的分析依据及其所涉及的财务数据、财务指标情况等。

五、结构化多维信息工作底稿工程

本文通过结构化多维信息工作底稿工程，为通用大模型提供了用以识别舞弊的多维信息。该工作底稿不仅在长度上充分考虑大模型的阅读能力，并将相关具体信息在工作底稿中出现的位置与舞弊识别思维链顺序尽量对应，确保提供的识别舞弊所需信息便于大模型阅读分析，针对性解决通用大模型“无法完整读取并分析图表文混排长文本”“难以全面准确获取舞弊识别所需多维信息”问题。

(一)结构化多维信息工作底稿构造思路

在工作底稿构造过程中，需对舞弊识别思维链逐条拆解，并根据每条思维链识别舞弊所需的信息，编制对应的工作底稿模板，继而从深交所上市公司相关数据库中提取相关信息填充至工作底稿模板中，最终形成一份结构化多维信息工作底稿。大模型在执行舞弊识别思维链时可直接获取工作底稿信息，无需再从公司年报等长文本中寻找所需信息，既可有效减少输入大模型信息的文本长度，还可降低大模型定位信息的难度，有效提升大模型信息获取效率与准确性。

同时，由于部分上市公司在披露年报后又对财务数据进行会计差错更正，在工作底稿获取此类公司财务数据时，面临选用会计差错更正前数据(最早披露)还是会计差错更正后数据(最新披露)问题。相关研究表明，从识别舞弊动机角度看，更正后财务数据还原了公司真实财务状况，更能如实展示公司前期存在的舞弊动机；而从识别虚增收入、利润等舞弊手法角度看，更正前财务数据更能反映公司存在的异常特征，有利于更好识别具体舞弊风险点，并不能一概而论(吴溪等，2025)。但从监管目标出发，研判上市公司舞弊风险的主要目的是揭示尚未被识别的舞弊。会计差错更正后，意味着上市公司相关舞弊风险已在一定程度上被识别并得到纠正，舞弊风险已部分“显性化”。若对此类公司的分析仍沿用更正前数据，实质上是将监管注意力集中于过去已暴露的舞弊问题，这与监管应聚焦于“潜在未发现风险”的目标不符。因此，综合考虑取数逻辑统一性、便利性以及监管目标，如公司存在会计差错更正情形的，工作底稿构造中统一选用会计差错更正后的最新披露数据。

(二)结构化多维信息工作底稿设计开发

通过开发预置多维信息块和利用大模型自动抽取信息，设计开发了聚合多维信息的工作底稿，并支持灵活高效配置底稿，大幅提升了舞弊识别大模型系统的迭代效率和舞弊识别分析的精准度。具体如下：

1. 设计开发多维信息块

根据思维链舞弊识别所需信息，设计开发了由历史与当期财务信息、临时公告信息、可比公司信息以及客户、供应商、关联方等相关主体信息等多维信息块组成的工作底稿模板。

2. 利用大模型自动抽取信息

通过将定期报告进行篇章拆解，对表格等进行特殊化处理，构建定期报告Rag(检索增强生成)系统。在生成多维信息工作底稿时，通过指定位置、指定提示词等方式，准确提取所需数据。

3. 增强工作底稿可比公司信息精准性

根据舞弊识别思维链要求，舞弊识别大模型需分析相关财务指标与同行业可比公司相比是否存在明显偏离。工作底稿中提供的可比公司是否具备较高可比性，对分析结论准确性有较大影响。而实践中常见的行业分类标准并非完全基于舞弊识别目的开发，所划定的同行业公司，其产品、服务的可比性不一定能精准满足舞弊

分析需要。为此，需在对公司产品、服务名称进行标准化处理的基础上，综合考虑产品、服务的性质以及在主营业务中的占比，定量计算不同公司之间产品、服务的相似度，进而对相关公司的可比性进行排名，以更精准锁定可比公司范围，进一步增强大模型舞弊识别精准性。

六、舞弊风险识别智能体工程

舞弊识别大模型系统在实际运行中，需同时获取舞弊识别思维链信息与工作底稿信息。通过舞弊风险识别智能体工程，优化编排舞弊识别思维链和结构化多维信息工作底稿，并引入辩护分析思维链实施“博弈对抗”，全面提升大模型识别舞弊效能，进一步解决通用大模型难以合理平衡防止漏报与控制误报的问题。

(一)测试思维链与工作底稿最优编排方式

在舞弊识别思维链方面，需确定是将思维链“一次性输入”大模型，还是“分模块、分层次输入”；在结构化多维信息工作底稿方面，也需考虑是将多维信息“一次性提供”给大模型，还是比照思维链对应拆分后“分模块、分层次逐一对应提供”。

通用型舞弊风险识别特征与舞弊动机组成基础舞弊风险识别智能体，用于初步反映公司整体舞弊风险情况，但并不指向具体舞弊领域、舞弊模式以及舞弊手法；指向型舞弊风险识别特征根据舞弊的具体类型和特征，细化区分为虚构交易类收入舞弊、会计操纵类收入舞弊、净利润类舞弊、净资产类舞弊、资金占用舞弊以及特殊行业与业务模式类舞弊，分别构建形成六类指向型舞弊风险识别智能体，用于反映公司舞弊风险指向的具体舞弊领域、舞弊模式以及舞弊手法。每类舞弊风险识别智能体涉及的思维链与工作底稿被定义为单项思维链与单项工作底稿。例如，虚构交易类收入舞弊思维链及其对应的工作底稿，即属于单项思维链与单项工作底稿范畴。测试发现，上述七类舞弊风险识别智能体涉及的单项思维链与单项工作底稿的编排方式，对大模型的分析效果具有重要影响。

本文实测了三种编排方式，方式一系在单项思维链运行时，对应提供全部七类单项工作底稿供其分析输出结论使用，例如虚构交易类单项思维链运行时，提供包含虚构交易类单项工作底稿在内的全部工作底稿信息；方式二系在单项思维链运行时，对应提供基础舞弊风险类单项工作底稿和自身领域单项工作底稿供其分析输出结论使用，例如虚构交易类单项思维链运行时，提供基础舞弊风险类单项工作底稿、虚构交易类单项工作底稿；方式三系在单项思维链运行时，仅对应提供自身领域单项工作底稿供其分析输出结论使用，例如虚构交易类单项思维链运行时，仅提供虚构交易类单项工作底稿。

经多案例对比实测，方式三的编排方式识别效果最佳，通过“分模块输入、分智能体处理”，整体逻辑清晰、可追溯性强，既避免信息冗余引发大模型注意力分散问题，提升单项识别分析精度与执行效率，也可以通过逐一对应保障系统整体处理流程的稳定性和可扩展性，降低幻觉发生率。详见图3。

(二)引入辩护分析思维链实施“博弈对抗”

测试发现，虽然通过构造规范大模型分析输出的关键约束提示词，有效降低了大模型的误报率，但因舞弊识别思维链有“天然”舞弊指控倾向，舞弊识别大模型系统仍存在一定程度的广泛报警问题。同时，在单一指控舞弊智能体下，大模型在实现防止漏报与控制误报的双重目标时，面临“左右互搏”的天然困难，导致注意力分散，也影响分析质量。

监管实践表明，对于某些特殊行业、特殊业务模式公司或公司业务结构发生重大调整的情况，公司触及某些具体舞弊识别特征具有商业合理性，并非表明其存在舞弊风险。例如，对于客户、供应商重叠识别特征，部分业务多元的特大型企业集团，会通过旗下不同业务单元子公司分别开展不同业务，从而出现同一集团既是公司客户、又是公司供应商的情形，此时公司客户、供应商重叠或具有关联关系通常具有商业合理性。在应对该类误报问题时，如在思维链舞弊风险识别特征处强调特定情形下不适用该识别特征，反而容易引发大模型幻

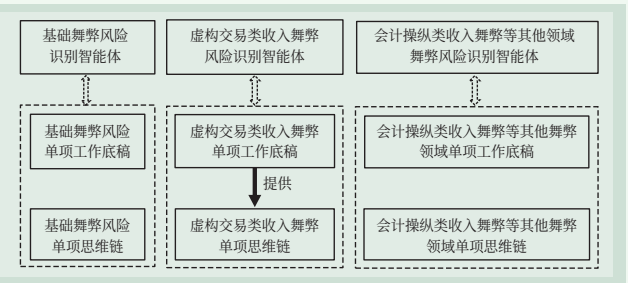

图3 编排方式三示意图

觉，从而降低整体分析质量。

因此，本文充分借鉴“博弈对抗”思路，创新性地构造了相关辩护思维链与辩护分析智能体，即将公司触及的所有舞弊识别特征信息均提供给辩护分析智能体，由其专职辩护分析触及的舞弊识别特征是否具有合理性，在不增加大模型“幻觉”情况下，进一步减少误报，有效提升了舞弊风险识别精准度。

(三)舞弊识别智能体结构体系

基于基础舞弊风险识别智能体、六类指向型舞弊风险识别智能体、辩护分析智能体，形成了舞弊识别大模型系统的核心分析功能。通过裁决分析智能体，在统筹考虑上述智能体全部分析结果的基础上，最终评定公司舞弊风险等级，并提示具体舞弊领域、舞弊模式、舞弊手法以及监管应对建议。舞弊识别大模型系统完整架构如图4所示。

七、舞弊识别大模型系统建设成效分析

经过实测，舞弊识别大模型系统在舞弊识别方面应用效果较好，有效弥补了机器学习模型以及单纯利用专家规则识别舞弊的能力短板。

(一)舞弊风险识别智能体功能实现情况

1. 应用功能实现情况

目前，舞弊识别大模型系统已实现如下核心应用功能：一是覆盖虚构交易类收入舞弊、会计操纵类收入舞弊、净资产类舞弊、净利润类舞弊以及资金占用舞弊等典型高风险舞弊领域的舞弊风险识别。二是初步支持多智能体博弈对抗，引入辩护思维链，充分考虑行业和规模特殊性，合理平衡防止漏报和控制误报。三是支持舞弊识别思维链提示词在线修改，快速应用舞弊识别最新研究成果。四是支持底稿章节配置，以及通过预置信息块和大模型自动抽取信息两种方式提供工作底稿信息，便于根据思维链的最新修改情况快速进行底稿调整。

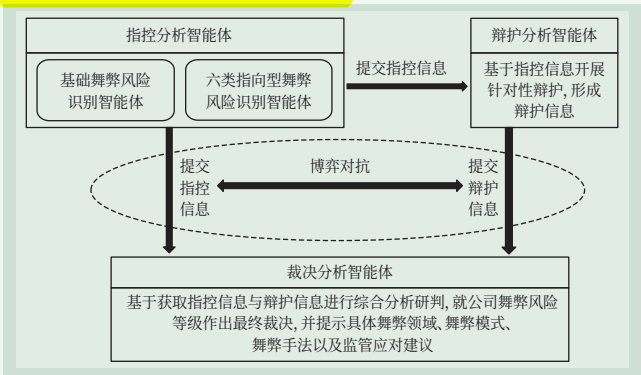

图4 舞弊识别大模型系统完整架构图

2. 综合分析结果展示

形成综合分析结果展示页面，用于展示裁决分析智能体生成的舞弊风险综合分析结果。该页面采用简洁明了的布局设计，方便监管人员清晰查阅舞弊识别分析的总体结论、识别的具体舞弊风险点、识别舞弊风险点的分析过程和分析依据。此外，页面还支持报告搜索、标记和注释等辅助功能。

3. 智链轨迹视图

智链轨迹视图功能支持监管人员自行下载各智能体(如虚构交易类收入舞弊分析智能体)获取的单项工作底稿以及所生成的单项舞弊识别分析报告。监管人员在阅读综合分析结果后，可通过此功能查阅各舞弊识别领域更详实的单项舞弊识别分析报告，进而获取相关舞弊风险点涉及的具体数据以及更细化的分析过程，有助于监管人员进一步验证舞弊识别大模型系统分析结果的合理性。

(二)舞弊识别大模型系统舞弊识别效果实测表现

1. 测评方法

为真实验证舞弊识别大模型系统识别效果是否能满足监管工作需要，本文将舞弊识别大模型系统直接应用于真实生产环境进行实测。本文以2024年深市上市公司年报为测试基准，构建了覆盖不同日常监管风险等级、不同行业、不同规模公司的测试样本集，并且样本公司均尚未被明确认定存在财务舞弊情形，同时邀请负责监管相关样本公司的监管人员对舞弊识别大模型系统识别结果进行多维度详细测评。测评时，监管人员需在阅读舞弊识别大模型系统生成的分析报告后，从舞弊识别精准性、分析逻辑性以及对监管工作帮助程度和使用意愿度等多个维度，按照0分(极差)–10分(极好)进行评分。其中，5分为中间基准，8.0分–10.0分为优秀分段，6.0分–7.9分为良好分段，5.0分–5.9分为一般分段，3.0分–4.9分为较差分段，2.9分以下为极差分段。

需要说明的是，测试机器学习模型财务舞弊识别效果时，通常采取历史数据回溯，即选取一定历史跨度内上市公司的历史财务数据作为测试样本，关注是否可以精准识别已公开的被查实财务舞弊案例，并计算召回

率、精准率等指标。但本文在测试时未采用历史数据回溯，而是直接将舞弊识别大模型系统应用于真实生产环境，选择尚未被明确认定存在财务舞弊情形的样本公司进行测试，主要考虑如下：一是避免大模型预训练语料库干扰。大模型预训练语料中很可能已包含公开的历史财务舞弊案例，在对这些舞弊案例进行分析时，大模型可能会预先调高案例公司的舞弊风险等级，并精准识别公司舞弊风险。二是避免过拟合问题干扰。部分机器学习模型基于历史数据的测评结果显示，其实现了较高的召回率和精准率，但此类模型往往仅在测试训练数据时表现较好，对样本外公司的识别效果不佳，即容易出现过拟合问题(田莉等，2025)。因此，考虑到舞弊识别思维链中部分识别特征、量化阈值等的设定本身即参考借鉴了过往财务舞弊案例，历史数据回溯也可能会一定程度受到过拟合问题影响，这不利于测试舞弊识别大模型系统在真实生产环境下的泛化能力与舞弊识别效果。

2. 测评结果

经综合多维度评价指标得分，舞弊识别大模型系统对测试样本集的识别结果中，良好以上等级的评分占比为77.2%，监管人员对舞弊识别大模型系统整体认可程度较好。

(1)舞弊识别精准性方面

就评定的整体舞弊风险等级而言，监管人员基于相关信息，结合监管经验和专业判断，认为存在舞弊迹象的测试样本公司中，有87%被舞弊识别大模型系统识别为高或中高风险，即对高舞弊风险公司的漏报率为13%；认为不存在舞弊迹象的测试样本公司中，有74%被舞弊识别大模型系统识别为低风险，即对低舞弊风险公司的误报率为26%。就识别的具体舞弊风险点而言，在漏报控制方面，评分达到良好以上等级的比例为76.1%，即识别的舞弊风险点相较于监管人员日常监管的关注点遗漏较少；在误报控制方面，评分达到良好以上等级的比例为73.9%，即识别的舞弊风险点与监管人员日常监管中关注的风险点较为匹配。因此，就舞弊识别精准性而言，舞弊识别大模型在防止漏报与控制误报之间实现合理平衡，未出现大面积漏报或广泛误报情形，具备在监管实践中部署应用的性能条件。

需要说明的是，因本文选择测试的样本公司均尚未被明确认定存在财务舞弊情形，监管人员仅系结合自身判断情况，评价舞弊识别大模型系统的识别结果是否合理。舞弊识别大模型系统提示的财务舞弊风险仅为初筛排查，能否正式形成财务舞弊线索，仍需监管人员进一步采取发函问询等措施进行验证核实，对于相关公司是否确实存在财务舞弊情形，还需以证券监管机构的检查或调查结果为准。

(2)辅助监管实践有用性方面

监管人员对舞弊识别大模型系统的有用性评分达到良好以上等级的比例为79.3%，即监管人员认为，舞弊识别大模型系统有助于其提升年报审查效率、提高舞弊识别质量。此外，监管人员使用舞弊识别大模型系统意愿度评分达到良好以上等级的比例为88.2%，即监管人员使用舞弊识别大模型系统意愿程度较高。

(三)舞弊识别大模型系统舞弊识别能力对比分析

1. 与机器学习模型相比

与基于机器学习的财务舞弊识别模型相比，舞弊识别大模型系统借助通用大模型的深度推理能力，融合基于专家经验的舞弊识别思维链以及多维信息工作底稿，在舞弊识别能力方面有效弥补了机器学习模型存在的三大短板。首先，舞弊识别结果既提示整体舞弊风险等级高低，也明确提示舞弊风险所指向的具体舞弊领域、舞弊模式以及舞弊手法；其次，舞弊识别中既使用结构化财务数据，也充分使用各类非财务数据、非结构化数据，有助于更好识别第三方配合造假等复杂、新型财务舞弊行为；最后，舞弊识别结果具有高度可解释性，既提示识别到的具体舞弊风险点，也提示认定相关舞弊风险点的分析依据，并输出针对性监管应对建议，可为实际监管业务提供直观的舞弊线索与决策支持。

2. 与单纯利用专家规则模式相比

目前，深交所上市公司监管系统中部署了一套基于专家规则的“风险标签”系统，可以自动识别上市公司是否触及相关风险预警指标，供监管人员参考，该系统已稳定运行多年。

本文将舞弊识别大模型系统与基于专家规则的“风险标签”系统的舞弊识别效果进行了对比测试。如果将“风险标签”系统基准得分设置为5分，对比评价下舞弊识别大模型系统有用性评分达到良好以上(6分及以上)的

比例为88%，且无低于5分的评分。监管人员普遍认可舞弊识别大模型系统对上市公司触及异常特征的推理分析更为综合、深入，能够提供较多增量信息，而“风险标签”系统的指标预警相对较为孤立、缺乏综合推理分析。

八、结论与建议

本文通过舞弊识别思维链提示词工程、结构化多维信息工作底稿工程以及舞弊风险识别智能体工程，有效解决了通用大模型不具备识别舞弊风险的成熟系统方法论、无法完整读取并分析图表文混排长文本、难以全面准确获取舞弊识别所需多维信息，以及难以合理平衡防止漏报与控制误报等关键问题。舞弊识别大模型系统1.0版本将应用于上市公司监管实践之中。在上市公司年报披露后，舞弊识别大模型系统将自动抓取公司年报信息和其他相关信息，对公司是否存在舞弊风险进行“拟人化”智能推理分析，并向监管人员提示上市公司可能存在的舞弊风险以及监管应对建议，为监管人员识别上市公司舞弊风险提供智能化辅助。

但现阶段应用大模型识别财务舞弊仍然存在一定局限，在舞弊识别思维链全面性、工作底稿多维信息丰富度与精准度、大模型分析过程详实度以及漏报和误报控制水平等方面均有提升空间，并且大模型仍有少量幻觉与随机性难以彻底消除等。未来拟从三个方面对舞弊识别大模型系统进行持续优化。一是密切跟进通用大模型更新迭代进展，持续提升大模型分析推理能力、上下文记忆读取能力以及幻觉控制水平；二是根据最新学术研究成果和监管实践经验，对思维链识别特征、具体阈值、关联组合方式、结构层次以及应用逻辑等进行动态更新与完善，进一步增强舞弊识别思维链的全面性、精准性、有效性；三是在增强工作底稿已有信息准确度的基础上，探索提供维度更丰富的信息内容，助力进一步提升大模型舞弊识别质效。

基于结论，本文提出以下政策建议：一是金融监管机构切实加强科技监管能力建设，提升人工智能在金融监管领域的应用广度与深度，优化相关信息披露与报备要求，强化与工商、司法、税务以及海关等部门的数据共享，增强可用数据信息的标准化、结构化水平以及丰富度、精准性，为人工智能赋能金融监管提供优质、充沛的数据层“源头活水”。同时，进一步明晰人工智能应用的监管规则与伦理准则，筑牢应用人工智能的安全边界，让人工智能用得好、管得住。

二是中介机构积极适应人工智能时代发展趋势，稳步推进相关业务流程数智化转型，探索在专业人员把关复核前提下，合理应用人工智能等新技术提升执业质量与核查质效，更好地承担资本市场“看门人”责任。同时，适应性调整完善内部治理体系与组织结构，加快具备信息化专业知识的复合型人才队伍建设，持续做好行业专业知识库搭建、内部业务数据信息的整理与清洗等基础性工作，并采取有效配套措施防范数据泄漏风险。

三是理论界深入开展人工智能与会计、审计交叉领域研究，有效探索人工智能技术在会计、审计实务中的多元应用场景，为会计、审计领域的数智化转型提供理论支撑与智力支持。其中，在财务舞弊研究领域，建议紧密围绕我国会计、审计实务以及监管实践，结合财务舞弊案例的新情况、新特点以及新趋势，就如何完善会计准则、审计准则以及信息披露规则，如何构建更全面、精准、有效的舞弊识别框架，如何更有效从源头防范财务舞弊等，深入开展具有实践指导价值的理论研究，推动形成具有中国特色的“财务舞弊识别与防范知识体系”，为构建中国自主会计知识体系添砖加瓦。 ■

注释

1. 幻觉指大模型在推理过程中，因训练数据局限、算法机制缺陷等，可能生成与事实不符的结论，可能对用户产生误导，包括事实性幻觉与忠实性幻觉。其中，事实性幻觉是指生成内容与可验证的现实事实不符，表现为事实矛盾或捏造；忠实性幻觉是指生成内容偏离用户输入指令，甚至不响应用户指令。

2. 随机性指大模型在完全相同的信息输入下，每次生成的输出结果并不完全相同。

3. 当前舞弊识别思维链中，舞弊风险等级分为低风险、中高风险和高风险。

参考文献：

[1] 曹策, 陈焰, 周兰江. 基于深度学习和文本情感的上市公司财务舞弊识别方法[J]. 计算机工程与应用, 2024, 60(4): 338-346.

[2] 陈彬, 刘会军. 什么样的公司有财务造假嫌疑?——来自香橼公司和浑水公司的启示[J]. 证券市场导报, 2012, (7): 66-71.

[3] 洪荭, 胡华夏, 郭春飞. 基于GONE理论的上市公司财务报告舞弊识别研究[J]. 会计研究, 2012, (8): 84-90+97.

[4] 胡丹. 上市公司利润调节问题监管研究[J]. 证券市场导报, 2018, (11): 50-59.

[5] 钱苹, 罗玫. 中国上市公司财务造假预测模型[J]. 会计研究, 2015, (7): 18-25+96.

[6] 田莉, 吴思思, 邸超伦. 财务舞弊量化判别模型发展趋势及监管应用[J]. 证券市场导报, 2025, (6): 50-56+79.

[7] 汪春华. 基于深度学习的企业财务困境预测方法研究[J]. 工程经济, 2023, (3): 19-40.

[8] 韦琳, 徐立文, 刘佳. 上市公司财务报告舞弊的识别——基于三角形理论的实证研究[J]. 审计研究, 2011, (2): 98-106.

[9] 吴革, 叶陈刚. 财务报告舞弊的特征指标研究：来自A股上市公司的经验数据[J]. 审计研究, 2008, (6): 34-41.

[10] 吴世农, 林晓辉, 李柏宏, 王举明. 智能财务分析与诊断机器人的开发及实证检验——来自我国A股上市公司的经验证据[J]. 证券市场导报, 2021, (2): 62-71+78.

[11] 吴溪, 付荣, 耿春晓. 虚假会计数据影响财务舞弊模型的识别效力吗[J]. 会计研究, 2025, (5): 3-17.

[12] 叶康涛, 刘金洋. 非财务信息与企业财务舞弊行为识别[J]. 会计研究, 2021, (9): 35-47.

[13] 叶钦华, 黄世忠. AI大模型赋能财务舞弊识别的实践探索[J]. 中国注册会计师, 2025, (3): 36-40.

[14] 叶钦华, 叶凡, 黄世忠. 财务舞弊识别框架构建——基于会计信息系统论及大数据视角[J]. 会计研究, 2022, (3): 3-16.

[15] 赵纳晖, 张天洋. 基于MD&A文本和深度学习模型的财务报告舞弊识别[J]. 会计之友, 2022, (8): 140-149.

[16] 周卫华, 翟晓凤, 谭皓威. 基于XGBoost的上市公司财务舞弊预测模型研究[J]. 数量经济技术经济研究, 2022, 39(7): 176-196.

[17] Albrecht W S, Romney M B. Red-flagging management fraud: A Validation[J]. Advances in Accounting, 1986, (3): 323-333.

[18] Albrecht W S, Wernz G W, Williams T L. Fraud: Bring the light to the dark side of business[M]. New York: Irwin Inc, 1995.

[19] Bai S, Wu B, Zhang Y, et al. AuditAgent: Expert-guided multi-agent reasoning for cross-document fraudulent evidence discovery[C]// Proceedings of the 6th ACM International Conference on AI in Finance, 2025: 1-9.

[20] Bao Y, Ke B, Li B, Yu J, Zhang J. Detecting accounting fraud in publicly traded U.S. firms using a machine learning approach[J]. Journal of Accounting Research, 2020, 58(1): 199-235.

[21] Bender E M, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big?[C]// Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021: 610-623.

[22] Beneish M D. The detection of earnings manipulation[J]. Financial Analysts Journal, 1999, 55(5): 24-36.

[23] Bologna J, Lindquist R J, Wells J T. The accountant's handbook of fraud and commercial crime[M]. New York: Wiley, 1993.

[24] Cecchini M, Aytug H, Koehler G J, Pathak P. Detecting management fraud in public companies[J]. Management Science, 2010, 56(7): 1146-1160.

[25] Craja P, Kim A, Lessmann S. Deep learning for detecting financial statement fraud[J]. Decision Support System, 2020, 139: 113421.

[26] Dechow P M, Sloan R G, Sweeney A P. Causes and consequences of earnings manipulation: An analysis of firms subject to enforcement actions by the SEC[J]. Contemporary Accounting Research, 1996, 13: 1-36.

[27] Dechow P M, Weili G E, Larson C R, Sloan R G. Predicting material accounting misstatements[J]. Contemporary Accounting Research, 2011, 28(1): 17-82.

[28] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[29] Guo D, Yang D, Zhang H, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning[J]. Nature, 2025, 645(8081): 633-638.

[30] Hsieh C P, Sun S, Ginsburg B, et al. Ruler: What's the real context size of your long-context language models?[C]// First Conference on Language Modeling(COLM), 2024, https://openreview.net/forum?id=kIoBbc76Sy.

[31] Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation[J]. ACM Computing Surveys, 2023, 55(12): 1-38.

[32] Kaikaus J, Hobson J L, Brunner R J. Truth or fiction: multimodal learning applied to earnings calls[R]. Osaka: IEEE International Conference on Big Data, 2022.

[33] Lee T A, Ingram R W, Howard T P. The difference between earnings and operating cash flow as an indicator of financial reporting fraud[J]. Contemporary Accounting Research, 1999, 16(4): 749-786.

[34] Lin C C, Chiu A A, Huang S Y, Yen D C. Detecting the financial statement fraud: the analysis of the differences between data mining techniques and experts' judgments[J]. Knowledge-Based Systems, 2015, 89: 459-470.

[35] Liu N F, Lin K, Hewitt J, et al. Lost in the middle: How language models use long contexts[J]. Transactions of the Association for Computational Linguistics, 2024, 12: 157-173.

[36] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[J]. Advances in Neural Information Processing Systems, 2022, 35: 27730-27744.

[37] Persons O S. Using financial statement data to identify factors associated with fraudulent financial reporting[J]. Journal of Applied Business Research, 1995, 11(3): 38-46.

[38] Purda L, Skillicorn D. Accounting variables, deception, and a bag of words: assessing the tools of fraud detection[J]. Contemporary Accounting Research, 2015, 32(3): 1193-1223.

[39] Rezaee Z. Causes, consequences, and deterence of financial statement fraud[J]. Critical Perspectives on Accounting, 2005, 16(3): 277-298.

[40] Sgantzos K, Hemairy M A, Tzavaras P, et al. Triple-entry accounting as a means of auditing large language models[J]. Journal of Risk and Financial Management, 2023, 16(9): 383.

[41] Street D, Wilck J, Chism Z. Six principles for the effective use of artificial intelligence large language models: how to leverage ChatGPT, Bard, and Bing Chat in accounting work[J]. The CPA Journal, 2023, 93(11-12): 50-57.

[42] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]// Advances in neural information processing systems, 2017, 30: 5998-6008.

[43] Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models[J]. Transactions on Machine Learning Research, 2022a, https://openreview.net/forum?id=yzkSU5zdwD.

[44] Wei J, Bosma M, Zhao V, et al. Finetuned language models are zero-shot learners[C]// International Conference on Learning Representations, 2022b, https://openreview.net/forum?id=gEZrGCozdqR.

[45] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[C]// Advances in Neural Information Processing Systems, 2022c, 35: 24824-24837.

(责任编辑：卢一宣)