FinRule-Bench：财务表格与原则联合推理基准

Arun Vignesh Malarkkan
arun.malarkkan@asu.edu
亚利桑那州立大学 坦佩，亚利桑那州，美国

Manan Roy Choudhury
mroycho1@asu.edu 亚利桑那州立大学 坦佩，亚利桑那州，美国

Guangwei Zhang
changhu1633@gmail.com Pine AI 新加坡，新加坡

Vivek Gupta
vgupt140@asu.edu 亚利桑那州立大学 坦佩，亚利桑那州，美国

Qingyun Wang
qwang16@wm.edu 威廉与玛丽学院 威廉斯堡，弗吉尼亚州，美国

Yanjie Fu
yanjie.fu@asu.edu 亚利桑那州立大学 坦佩，亚利桑那州，美国

Denghui Zhang
dzhang42@stevens.edu 史蒂文斯理工学院 新泽西州，美国

arXiv:2603.11339v1 [cs.AI] 11 Mar 2026

摘要

大语言模型（LLMs）日益广泛地应用于财务分析，但它们在明确的会计原则约束下审计结构化财务报表的能力仍缺乏充分研究。现有基准主要评估问答、数值推理，或对合成篡改数据进行异常检测，因此模型能否在正确的财务报表上可靠地验证规则合规性或定位违规之处，尚不明确。我们推出**FinRule-Bench**，这是一个用于评估模型在真实财务表格上进行基于规则的财务推理时诊断完备性的基准。FinRule-Bench 将真实财务报表与明确且由人工整理的会计原则配对，涵盖四种典型报表类型：资产负债表、现金流量表、利润表和权益变动表。该基准定义了三项审计任务，对推理能力的要求逐步提高：（i）**规则验证**，检验是否符合单条原则；（ii）**规则识别**，要求从给定规则集中选出被违反的原则；（iii）**联合规则诊断**，要求在记录级检测并定位多个同时发生的违规。我们在零样本和少样本提示设置下评估大语言模型，并提出一种因果反事实推理协议，以确保判断、解释与反事实判断之间的一致性。在各项任务和报表类型中，我们发现，模型虽然能够较好地完成孤立的规则验证任务，但在规则判别和多违规诊断任务中的性能却显著下降。错误主要表现为违规覆盖不全和错误定位，揭示了现有财务或表格基准未能捕捉到的系统性失效。FinRule-Bench 为研究高风险财务分析中大语言模型受规则约束的推理、诊断覆盖和失败模式，提供了有原则且可复现的测试平台。我们在 https://anonymous.4open.science/r/FinRule-Bench-6723/ 发布 FinRule-Bench 基准及完整的评估代码库，其中包括适用于所有财务报表类型的三项任务专用大语言模型提示脚本。

1 引言

大语言模型（LLMs）越来越多地应用于财务分析和决策支持，包括数值问答、报告摘要和异常检测等应用。然而，从大语言模型评估的角度来看，财务审计中的一项核心任务仍未得到充分研究：判断结构化财务报表是否符合明确的会计原则；若不符合，则识别并定位不合规的确切来源。这一要求不止于给出正确答案或检测孤立的不一致之处；它还要求具备诊断完备性，即能够在结构化且相互依存的财务表格上，全面执行形式规则系统。

财务报告是现代资本市场的基石，也是企业向投资者和监督机构传达业绩、风险及监管合规情况的主要方式。这些报告的可靠性通过严格的审计流程加以保障，审计由 PCAOB [34] 等监管机构监督，并以美国公认会计原则 [12] 和国际财务报告准则 [20]等正式会计框架为依据。审计要求核实财务报表是否满足大量明确且相互依存的会计原则，包括算术恒等式、层级汇总约束和条件适用规则。因此，有效审计需要可评估其完备性的结构化推理，要求模型识别违规的原则，并将每项违规定位到具体记录，而不是仅给出二元判断或孤立答案。

现有基准只能部分涵盖这些要求。通用表格问答数据集（如 TAT-QA [44],）主要评估从单个表格中提取信息以及针对单个表格进行数值推理，而不要求依据外部规则体系进行验证。金融问答基准（如 FinQA [2] 和 ConvFinQA [3],）侧重于检索和算术计算，但不评估模型能否执行会计原则，或区分不同的规则违规。因此，模型可以在现有基准上取得很高的准确率



图 1：FinRule-Bench 基准概览。FinRule-Bench 评估大语言模型在财务审计任务上的表现，这些任务要求在明确的会计规则下进行表内推理，以及跨多个财务报表的联合推理。

但仍不适用于审计，因为审计中的正确性取决于全面执行规则并进行精确归因，而非孤立的答案。

这一差距促使我们采用一种不同的评估设置。基于真实财务报表验证会计合规，与问答或异常检测有着根本区别。在真实的审计场景中，财务表格通常在构建时就是正确的，问题源于推理错误，而非数据损坏或噪声。因此，在这一设置下评估大语言模型，需要采用能够将推理保真度与数据伪影区分开来的基准，并直接检验模型能否穷尽地识别被违反的规则，以及能否将每项违规归因于结构化财务报告中的相应记录。

为满足这一需求，我们推出 **FinRule-Bench**，这是一个旨在评估基于规则的真实世界财务报表推理能力的基准。FinRule-Bench 将结构化财务表格与明确且由人工整理的会计原则相结合，并定义了三项对推理能力要求逐步提高的审计任务：规则验证，用于测试是否符合单条原则；规则识别，要求从给定规则集中选出唯一被违反的原则；联合规则诊断，要求在记录级检测并定位多个同时发生的违规。这些任务共同评估模型在真实审计条件下组合运用会计规则、区分被违反的原则，以及将违规定位到特定记录的能力。

除了任务形式，FinRule-Bench 还为结构化财务分析引入了一套因果—反事实推理协议。这套协议要求模型不仅要输出决策，还要通过明确的因果解释为决策提供依据，并评估对财务表进行最小限度的反事实修改是否会改变这些决策。该协议并非旨在全面提升准确率，而是作为一种诊断工具：通过比较决策与反事实判断，它可以揭示仅评估最终答案时无法发现的不一致、覆盖不全和错误定位。

总体而言，FinRule-Bench 为评估大语言模型在高风险财务分析中的规则约束推理、诊断覆盖和失败模式提供了有原则的测试平台。通过关注规则遵循情况，而非财务报表是否正确，它与现有表格类和财务基准形成互补，并支持精准分析当前模型在哪些方面表现良好，以及在哪些方面存在根本性不足。我们的贡献总结如下：

(1) **基准**。我们推出**FinRule-Bench**，这是一个基于真实世界财务报表构建的基准，涵盖资产负债表、现金流量表、利润表和权益变动表。该基准将真实值表格与明确的会计原则配对，从而能够在问答或异常检测之外，对基于规则的财务推理进行受控评估。

(2) **任务套件**。我们将三项互补的审计任务形式化，用以评估逐步增强的推理能力：规则验证，用于检验是否符合单项会计原则；规则识别，要求从给定的规则集中选出被违反的原则；以及联合规则诊断，要求进行多违规检测和记录级定位。

(3) **推理与评估框架**。我们提出一种因果反事实提示协议，用于结构化财务推理，并引入诊断性评估分析，包括规则复杂度细分、错误类型分类以及性能—成本可扩展性研究，从而揭示仅凭总体准确率无法反映的诊断完备性方面的系统性失效。

2 相关工作

用于表格理解的大语言模型。将大语言模型应用于表格数据面临着与文本推理存在根本差异的挑战，包括对二维结构进行建模、执行精确的数值推理，以及保留跨行列的语义关系 [11]。早期研究采用特定任务的编码器模型来应对这些挑战，例如 TaPas [17],TaBERT [41], TURL [7], 和 TUTA [37], ，这些模型对表格结构和单元格级交互进行了显式建模。近期方法则利用指令微调大语言模型和基于提示的方法。Table-GPT [26]和 TableLlama [42]等以表格为中心的架构针对表格推理对大语言模型进行了适配，而 StructGPT [21] 和 DATER [40] 等基于智能体的框架则引入了外部工具或结构化表示。在这些方法中，序列化策略和提示技术（包括思维链推理 [38], ）已被证明能够显著提升性能。综合综述总结了这一领域的任务、数据集和方法趋势 [11, 32]。尽管已有这些研究，表格推理研究中的评估仍主要集中于问答和单元格级预测。这些评估设置无法衡量模型能否在复杂的真实世界表格中全面识别并定位规则违规，而这正是审计和合规导向应用的核心要求。

用于金融数据处理与分析的大语言模型。大语言模型已广泛应用于金融数据处理任务 [25, 43],，包括算法交易 [19], 、投资分析 [13], 、市场预测、情感分析 [27], 、欺诈检测 [23], 和风险评估 [24]。另一条并行研究路线聚焦于适应领域需求的金融语言模型，例如基于大规模金融语料库训练的 FinGPT [31], ，并经过调整以提升性能。近期研究还探索了检索增强生成（RAG），以提升财务分析中的事实依据和上下文感知能力 [35]。FinDABench [30]和 AuditBench [36] 等基准，为涉及检索、摘要生成和预测的金融任务提供了标准化评估设置。然而，这些基准主要评估信息获取能力或描述准确性，并未检验模型能否运用明确的会计原则来验证财务报表的内部一致性或诊断其中的违规。因此，在现有财务基准上表现出色，并不意味着模型适用于审计或合规验证。

结构化数据上的基于规则的推理与逻辑推理。 人工智能领域对结构化数据进行基于规则的推理与逻辑推理由来已久，近期研究则转向神经符号方法 [18], 。一些方法通过为大语言模型加入符号推理模块、约束求解器或逻辑引导解码，来提升逻辑正确性和推理忠实度 [14]。相关研究还探讨了复杂查询问答 [6, 33]。多跳推理 [22], 和逻辑引导的神经网络架构 [29]。Logic Game 等基准数据集 [15, 28]有助于了解抽象推理能力，但依赖领域语义有限的合成规则。相比之下，FinRule-Bench 评估基于真实财务报表和正式会计原则的基于规则的推理；其中的违规对应现实中的审计错误，并带来规则优先级排序、归因和多重违规覆盖等现有基准未能涵盖的挑战。

3 FinRule-Bench：数据集与评估框架

鉴于现有表格财务基准的局限性，我们推出 **FinRule-Bench**，这是一个用于评估在明确会计原则约束下对财务表格进行基于规则推理的基准。FinRule-Bench 体现了财务审计的核心要求：正确性取决于能否验证多项相互依赖的规则与结构化财务报表之间的合规性。不同于以往侧重于问答或检测合成受损数据中的异常的基准，FinRule-Bench 评估基于真实标签的财务数据上的推理，并将基于规则的推理保真度作为首要目标。

FinRule-Bench 使用从公司申报文件中提取的真实财务报表构建。这确保评估聚焦于真实的表格数据，而非合成数据集。数据正确性由明确且经人工整理的会计原则验证，并通过确定性验证器实现，从而确保真实标签可靠且可复现。为评估大语言模型在该基准上的表现，我们针对这些干净财务报表，设计了受控且考虑规则的记录级错误注入流程。以干净的表格财务报表作为真实基础数据集，通过受控的最小幅度编辑创建注入错误的数据集实例，使其违反特定会计原则。任务定义对应于针对这一扰动数据集的不同受限视角，彼此仅在违规数量、所需模型输出和复杂度方面有所不同。本节将介绍采用报表特定规则和验证器的数据集构建、任务形式、采用既定提示技术的评估协议，以及数据集统计。

3.1 数据集构建：基础财务报表

FinRule-Bench 基于真实世界的公司财务披露构建，主要数据来源为2024年10-K表格申报文件（图7）。这些文件是标准化且经监管机构批准的财务申报文件，遵循正式的会计框架和原则。我们从申报文件中提取财务报表，并将其转换为结构化的*Markdown*表格。我们使用 Qwen-2.5-VL-72B 视觉语言模型 [1] 协助将图像中的财务报表转录为 Markdown表格结构，同时保留行层级、列语义和数值（表6）。之所以采用 Markdown 作为规范表示形式，是因为它既便于人类阅读，也能直接供大语言模型使用。Markdown 文件中的每条记录对应某一财政年度内一家公司的单份财务报表。报表类型包括资产负债表、现金流量表、利润表和权益变动表。当某家公司某一财政年度的报表跨越多页或多张图像时

我们会将所有片段合并为一条完整连贯的 Markdown 记录。每张最终定稿的表格均作为单独文件存储，并按报表类型编制索引。FinRule-Bench 中包含的所有表格均对应真实财务报表。在数据集构建过程中，不会修改、扰动或人为破坏任何数值。因此，检测到的任何违规都完全源于依据明确的会计原则对表格进行评估，而非数据噪声或提取过程中的伪影。

3.2 数据集构建：会计规则与确定性验证器

我们的基准为每种报表类型整理了明确的会计原则。我们定义了四套贴近实际的会计原则集，分别对应资产负债表、现金流量表、权益变动表和利润表（表22）。每套规则集都涵盖与相应报表相关的算术一致性、层级汇总、跨部分依赖关系以及条件适用性。这些规则集互不重叠且针对特定报表，而非单一的全局规则池。这些原则/规则集反映了相应报表各自不同的语义与结构。尽管部分规则要求在一条记录内跨多行或多章节进行推理，FinRule-Bench并不强制执行不同报表类型之间的对账约束。这一设计将基于规则的推理限定在单一报表类型之内，从而避免将报表内部推理与跨报表一致性相互混淆。

每条规则 $r \in R$ 都关联一个确定性验证器函数 $v_r(T) \in \{\text{True}, \text{False}\}$。验证器以编程方式实现会计逻辑，仅用于生成真实标签和评估模型预测。在所有实验中，验证器的输出都被视为模型评估的唯一事实依据。

3.3 数据集构建：受控错误注入

为对模型进行受控评估，我们通过确定性地违反预定义的会计原则，构建基础数据集的注入错误的变体。错误注入由各陈述类型对应的会计规则定义指导，从而确保产生真实的会计错误，而非一些近期研究论文 [4, 5]中采用的任意扰动。干净财务报表与受控且遵循规则注入错误的报表共同构成一个统一数据集，旨在评估模型的联合推理能力。

错误注入围绕两种任务形式进行 - 情形下。对于单一违规情形，每条记录恰好注入一项会计原则违规。这一限制确保规则违规验证和识别具有明确性，从而将预期的推理要求单独区分出来。对于多违规情形，则在同一份财务报表中注入多项会计原则违规。这些违规可能影响不同的行、章节或规则类型，并需要进行多步联合推理，以模拟真实的审计场景。在这两种情形下，错误注入都会保留原有的表格结构、模式和格式。只有与目标原则直接相关的数值或派生字段会被修改。因此，性能反映的是对会计规则的推理能力，而非对合成格式变化的敏感度。

3.4 FinRule-Bench 上的任务形式

FinRule-Bench 定义了三项审计任务，用于评估模型在结构化财务表格上的推理能力，且这些任务的难度逐步提升。每项任务对应一种独特的审计问题，并评估模型能否理解会计原则、一致地应用这些原则，并在存在违规时定位违规。这些任务针对同一统一数据集采用不同的任务形式，并且在违规数量和所需模型输出方面各不相同。

规则验证。 给定一个财务Markdown表格和一条会计规则，模型预测该表格是否符合指定规则。此任务通过二元决策评估基本规则合规性。

规则识别。 给定一个财务表和一组会计规则，模型在记录级别识别被违反的规则。在此设置中，报表专用表格中的每条记录恰好包含一条被违反的规则，从而确保模型评估的明确性，并要求模型在相互竞争的会计原则之间进行区分。联合规则诊断。给定一个财务表和一组会计规则，模型首先按报表类型、逐条记录确定是否存在任何违规；如果存在，则通过记录级归因识别所有被违反的规则。该任务评估多步骤诊断式联合推理及召回率，模拟多个违规可能同时出现的真实审计情境。

所有任务均在单独的财务报表内进行，不假设明确的跨报表一致性约束。正式的任务定义见附录A。

3.5 提示与推理协议

我们比较了三种提示协议，这些协议在结构和模型监督方面有所不同。

零样本。模型接收任务指令、序列化财务表格 T 以及相关规则或规则集。此协议不提供示例。

少样本。 模型接收一小套与任务相同输入/输出格式的示例，随后是测试实例。示例在可能的情况下被选择以覆盖多样的规则类型，并在所评估的模型之间保持固定，以确保公平性。

带有因果与反事实推理的少样本。 在标准零样本与少样本提示之外，我们引入一种因果-反事实提示协议，以探究基于规则的推理保真度。我们通过上下文学习实现该提示协议，并使其在三种任务形式化中系统变化。在所有设定中，因果与反事实推理结构仅通过少样本示例引入，而非通过架构变更或额外监督。在因果-反事实提示中，示例被构建为使因果依赖显式化，而非隐含。它们同时说明导致规则违规的条件，以及能够消除该违规的最小反事实修改。

规则验证任务 在此设置中，因果-反事实提示针对特定规则，并包含简洁的示例，展示一种违规模式及相应的最小反事实修复，例如更正一个算术值或替换一个非标准标签。这项评估旨在检验接触因果与反事实

结构是否能提升对明显违规的检测能力，而无需在多条规则之间进行区分。

规则识别任务 在此，因果反事实示例旨在通过阐明哪些表格属性与每条规则具有因果相关性，以及对其他被违反规则和表面异常情况进行相应的反事实诊断，来支持规则判别。

联合规则诊断任务 在这一联合诊断推理任务中，示例包含详细的因果推理和反事实测试。然而，模型在推理时的输出被限制为规则代码的结构化列表；无需提供解释。该设置评估模型是否理解每条记录内表格结构中的因果结构，以及这些结构与规则合规之间的关系，从而足以在多个违规行为通过联合组合推理相互作用时实现广泛覆盖。

因果反事实提示并不强制保证推理一致性，也不保证准确率提升。相反，它充当一种条件化机制，在上下文学习过程中揭示因果结构。因此，不同难度级别和提示策略下的性能差异，反映了模型在推理会计规则违规时对因果相关性和最小干预的内化程度，尤其是在需要规则判别和多规则覆盖的场景中。我们强调，因果反事实推理完全通过示例设计引入，并不依赖思维链提示或事后解释评分。

3.6 模型评估与验证协议

模型评估通过与既定会计原则相对应的确定性规则验证器进行。模型输出会被解析为结构化表示，并与真实标注直接比较。对于规则验证和规则识别任务，评估采用精确匹配协议。对于联合规则诊断，多步正确性要求在记录级准确检测违规并准确识别被违反的规则。该协议确保评估客观、公平且可复现。

在评估协议方面，FinRule-Bench 使用通过与规则验证器进行精确匹配，计算与任务对齐的确定性评估指标。规则验证。 我们报告准确率、精确率、召回率和F1分数，以衡量总体正确性，以及误报与漏检违规之间的权衡。

规则识别。 我们报告Top-1准确率作为主要指标。为考虑规则类别不平衡，我们还报告平衡准确率、宏平均F1和加权F1。

联合规则诊断。 对于违规检测，我们报告二元准确率和 F1分数。对于正确性，我们报告精确匹配和微平均 F1，以量化多违规情形中的部分覆盖。

效率指标。 对于所有任务，我们报告总词元数、每个实例的平均词元数、每个实例的成本和总成本，以分析性能与效率之间的权衡。

3.7 数据集统计

我们总结了 FinRule-Bench 在不同报表类型和任务形式化下的构成，以对数据集进行定量概览。FinRule-Bench 涵盖四种财务报表类型：资产负债表（BS）、现金流量表（CF）、权益变动表（SE）以及利润表（SI）。对于每种报表类型，我们分别定义特定于该报表的会计规则集，并构建包含相应干净、单一违规和多重违规记录的 Markdown 表格。每条记录

表 1：FinRule-Bench 的数据集统计。每条记录对应一家公司的财务报表。单项违规记录恰好包含一条被违反的规则，而多违规记录则包含多条同时发生的违规。

陈述类型	# 规则	真实 记录	单项违规 记录	多重违规 记录
资产负债表（资产负债表）	5	212	1060	100
现金流量表（现金流量）	3	318	954	100
权益变动表（SE）	3	287	871	100
利润表（SI）	5	300	1500	100
总计	16	1117	4385	400

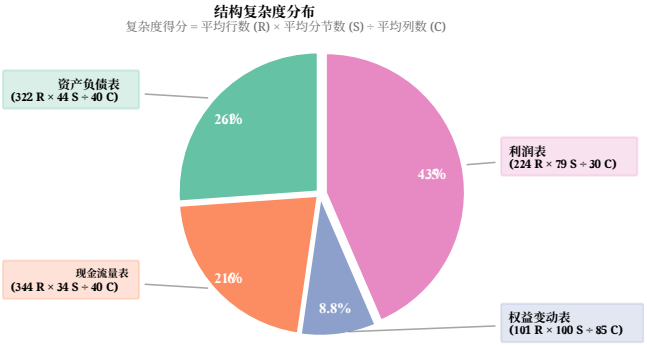


图 2：不同财务报表类型的结构复杂度分布。

对应给定会计期间的一份公司财务报表。干净实例由完全合规的财务报表构成，而注入错误的实例则包含以确定性方式生成的规则违规。在规则验证和规则识别任务中，每个实例至多包含一项违规。在联合规则诊断任务中，实例可能包含影响不同记录或原则的多项违规。在所有报表类型中，该数据集涵盖了多样化的会计原则，包括算术一致性规则、结构与分类规则、条件适用规则以及多部分对账规则。这种多样性确保评估能够全面衡量基于财务规则的推理能力。表 1 和图 2 总结了数据集统计、规则集以及结构复杂度分布。

3.8 可复现性与制品

所有验证器、错误注入提示、示例提示和评估脚本都是确定性的，并包含在发布制品中。模型输出会由确定性解析器解析为用于精确匹配的结构化形式。我们提供规则定义、验证器代码、示例提示和注入脚本，以确保所报告结果完全可复现。

4 实验

4.1 实验设置

我们介绍实验中使用的模型、数据、提示协议、推理配置和评估流程。

模型。 我们评估了四个具有代表性的大语言模型：GPT-4o [10], Gemini 2.5 Pro（思考） [9], Gemini 2.0 Flash [16], 以及 LLaMA 3.3 [8] 这些模型兼具以推理为重点和以效率为导向的架构。

数据与划分。 用于所有实验的数据集在第3节中进行了描述。干净的Markdown表格在实例化之前经过验证，注入错误的表则根据目标任务使用（第3.4节）。测试集与用于少样本提示的任何示例互不相交。所有按任务统计的实例数量和划分统计均报告于表1。

提示协议。 所有模型均在三种提示设置下进行评估：零样本、少样本，以及结合因果反事实推理的少样本。提示包含任务说明、序列化的财务表格 T（Markdown格式）以及相关规则或规则集。

上下文示例。 在两种少样本策略中，我们会在测试实例之前提供固定数量的带标签示例。示例采用与目标任务相同的输入—输出格式，并经过挑选以涵盖多种规则类型。所有模型使用相同的示例集，以确保比较公平。

解码与推理。 所有模型均采用确定性解码进行评估，温度设为零，输出长度固定。除确定性输出解析外，不使用任何外部工具、检索机制或后处理启发式方法。

评估协议。 每个测试实例均独立评估。对于生成过程中涉及随机性成分的任务，在设置相同的多次运行中对结果取平均。真实标签完全由确定性规则验证器生成。

成本与词元核算。我们记录每次模型调用的提示词元、补全词元和总词元数。货币成本根据评估时各模型公开的定价计算。记录的词元统计数据有助于分析性能与效率之间的权衡。

4.2 实验结果

4.2.1 各项任务的总体性能。表 2 报告了 GPT-4o、Gemini 2.5 Pro（思考）、Gemini 2.0 Flash 和 LLaMA 3.3 在 Finrule-Bench 上进行规则验证、规则识别和联合规则诊断的模型性能。这些任务对推理能力的要求逐步提高，从单独的规则检查扩展到多规则诊断推理。在规则验证任务中，所有模型的准确率都相对较高，表明现代大语言模型基本能够胜任基本的合规性检查。在这一设置下，模型之间的性能差异不大，这说明现代大语言模型基本能够处理针对单一会计原则的基本合规性检查。在规则识别任务中，性能大幅下降，因为模型必须从所有会计原则中找出被违反的原则。所有模型的准确率均有所下降，且推理型模型与效率型模型之间的性能差距扩大。在联合规则诊断任务中，性能进一步下降。虽然步骤一违规检测的表现仍相对较高，但所有模型在步骤二的诊断精确匹配表现都较低。

表 2：FinRule-Bench 上的总体任务性能。

模型	基准提示词	Rule 验证	Rule 识别	诊断 步骤一	诊断 步骤二
		Acc	Acc	Acc	精确匹配
GPT-4o	零样本	0.640	0.592	0.713	0.234
	少样本	0.736	**0.689**	0.786	0.297
	少样本 +反事实推理	**0.738**	0.661	**0.801**	**0.342**
Gemini 2.5 Pro （思考）	零样本	0.614	0.547	0.695	0.216
	少样本	0.629	**0.612**	0.728	0.259
	少样本 +反事实推理	**0.643**	0.601	**0.742**	**0.304**
Gemini 2.0 Flash	零样本	0.564	0.524	0.631	0.152
	少样本	0.566	0.558	0.667	0.189
	少样本 +反事实推理	**0.587**	**0.569**	**0.684**	**0.223**
LLaMA 3.3	零样本	0.553	0.496	0.619	0.147
	少样本	**0.692**	0.532	0.654	0.181
	少样本 +反事实推理	0.681	**0.541**	**0.672**	**0.215**

即使是表现最强的系统，在大多数情况下也无法定位所有被违反的规则。然而，因果反事实少样本提示在此展现出良好效果，使不同模型都获得了最大幅度的相对提升。这些结果凸显了在多个交互约束下进行穷尽式记录级推理的挑战。此外，我们观察到，因果反事实提示对 Gemini 2.0 Flash 和 LLaMA 3.3 等轻量级模型的益处更为明显，能够持续提升其在验证、检测和诊断定位任务上的性能。相比之下，GPT-4o 和 Gemini 2.5 Pro（思考）等推理型模型的提升并不稳定，在某些情况下性能甚至下降，尤其是在规则识别任务中。这表明，对于已经进行大量内部推理的模型，显式因果框架可能引入冗余，或干扰其判别决策。总体而言，随着任务要求从验证逐步转向判别和诊断，模型性能呈非线性下降，因此有必要在后续章节中进行详细的诊断分析。

4.2.2 规则复杂度与错误分析。 在此分析中，我们评估模型在不同规则复杂度分层、错误类型以及各任务分布方面的性能。本研究剖析了导致模型在各项任务中出错的规则语义属性，这些属性无法仅从总体准确率中看出。

规则复杂度分析。 表 3、4 和 5 分别报告了规则验证、规则识别和联合规则诊断任务按规则类别划分的性能。在所有任务中，算术规则始终最容易，在零样本提示下就能取得较高性能，额外的推理结构带来的提升有限。结构规则难度居中，从少样本和因果-反事实提示策略中获益最大，这表明借助示例能够改善对模式和章节层面的解读。条件规则和多记录规则在各种设置下最具挑战性。零样本性能较低，尽管少样本提示能够提升结果，但准确率仍显著低于那些更简单、直接的规则类型。我们还观察到，因果-反事实提示对多记录约束的益处最为明显，尤其是在联合

表3：规则验证的规则复杂度分析。

类别	规则数	实例数	Acc		
			零样本	少样本	少样本 +反事实推理
算术	2	499	0.916	0.657	0.759
结构	5	424	0.735	0.794	0.791
条件	6	1357	0.597	0.978	0.991
多记录	3	510	0.916	0.822	0.790

表 4：规则识别的规则复杂度分析。

规则类别	规则数	实例数	准确率		
			零样本	少样本	少样本 +反事实推理
算术	2	499	0.922	0.910	0.896
结构	5	424	0.503	0.559	0.540
条件	6	1357	0.590	0.703	0.590
多记录	3	610	0.159	0.184	0.257

表 5：联合规则诊断的规则复杂度分析。

规则类别	#规则数	#实例数	指标	
			步骤一（准确率）	步骤二（微平均 F1）
算术	2	79	0.797	0.545
结构	5	136	0.824	0.586
条件	6	285	0.961	0.910
多记录	3	144	0.576	0.591

注：结果按各规则类别中的实例进行微平均，并在 GPT-4o 模型的所有财务报表类型上汇总。由于多规则聚合，未报告类别层面的精确匹配。

诊断设置，其中推理必须跨多个记录进行组合。然而，这些收益伴随着计算成本的增加。

错误类型分析：联合规则诊断。这里，我们分析联合规则诊断任务中的失败模式。结果显示，在所有规则类别中，步骤一违规检测与步骤二诊断定位之间始终存在明显差距。尽管模型经常能够检测到违规的存在，但它们无法识别所有被违反规则，也无法正确定位这些规则，尤其是在条件约束和多记录约束方面。错误类型分析进一步表明，在各种报表类型中，部分检测和错误定位错误占主导地位。数值约束中常见假阳性，而依赖上下文的规则中则持续出现假阴性。这些模式表明，联合诊断失败主要源于覆盖不完整和错误归因，而非无法检测异常。图 3 进一步展示了在联合规则诊断中，不同报表类型和提示协议下错误类型的分布。在所有设置中，部分检测和混合错误占主导地位，表明模型经常能够识别出部分被违反规则，但

未能给出完整或定位正确的诊断结果，这证实联合诊断失败主要源于覆盖不完整和错误归因，而非未能检测到异常。这些结果表明，模型在 FinRule-Bench 上的失败具有结构性，并且取决于规则。随着任务要求对上下文适用性和跨部分依赖关系进行推理，性能会系统性下降，这证实该基准揭示了诊断推理中的根本局限性。

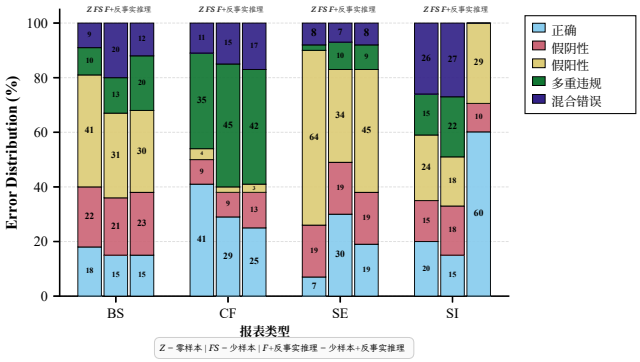


图 3：不同提示方法下的错误分布。

4.2.3 性能-成本权衡。我们分析了所有提示策略在各项任务中的性能与 Token 消耗之间的权衡，以评估结构化推理在 FinRule-Bench 中的成本效益。

规则验证。 图 4 显示，从零样本提示转向少样本提示，准确率会稳定提升，而令牌使用量仅小幅增加。然而，结合因果反事实推理的少样本提示会大幅增加令牌消耗，同时对准确率的提升有限，且提升幅度取决于陈述类型。对于资产负债表和现金流量表等复杂陈述，因果反事实提示可能具有较高的成本效益；而对于较简单的陈述，这种方法会增加词元开销，却无法带来显著收益。规则识别。 在这项任务中，我们观察到，与~~零样本提示相比，~~ 少样本提示在大多数报表类型上都能提高识别准确率；但加入因果反事实推理会显著增加令牌使用量，却未能提升判别性能，甚至往往会使其下降 4。这一规律在资产负债表、权益变动表和利润表评估中均成立，表明冗长的推理并不适用于需要从给定规则集中识别单条被违反规则的任务。联合规则诊断。图 5 展示了另一种权衡。因果反事实提示对步骤一违规检测的提升有限，但对于步骤二精确匹配定位，则能带来更大的相对提升，尤其是对于包含多个交互约束的报表。然而，这些提升伴随着高昂的令牌成本：与标准少样本提示相比，少样本+反事实推理会增加令牌使用量。这些结果表明，当任务要求穷尽定位多条规则时，结构化因果推理最为有益。报表层面也呈现出一致的规律。因果反事实提示在利润表（SI）评估中的收益最显著且最稳定；在 FinRule-Bench 的各类报表中，利润表的结构复杂度最高。利润表陈述包含小计、条件性行项目和跨章节汇总之间错综复杂的依赖关系，因此交互约束出现的可能性更高 2。在这种情况下，明确的因果示例和精简的反事实修复能更有效地引导模型穷尽定位被违反的规则。这清楚地表明，其效果与结构复杂度同步提升，而不仅仅取决于任务难度。

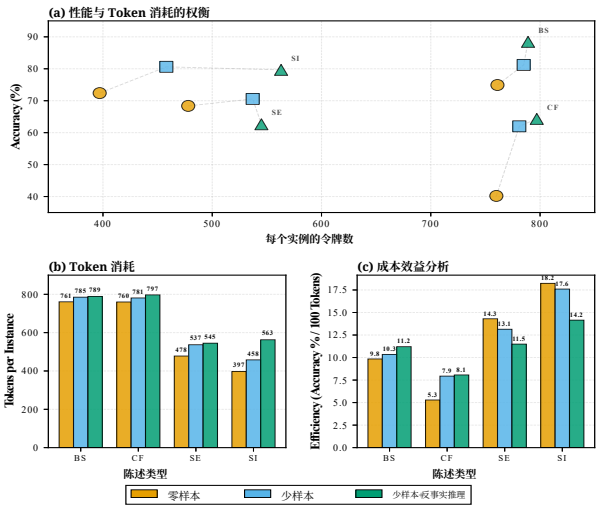


图4：规则验证的性能-成本分析。

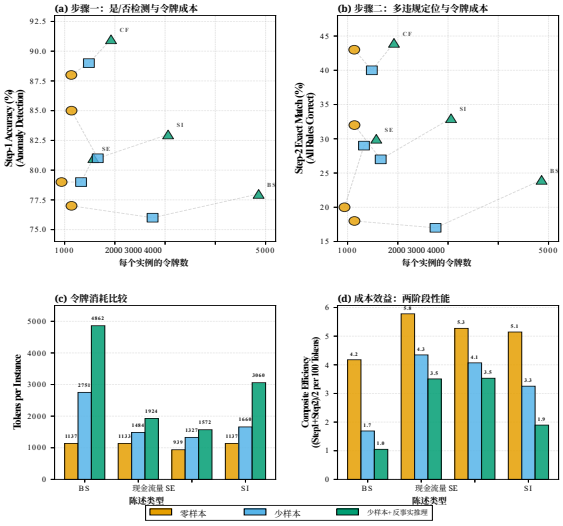


图5：联合规则诊断任务的性能-成本分析——步骤一异常检测与步骤二定位。

总体而言，我们对性能与效率的分析表明，结构化提示并非总是具有成本效益。额外的推理开销对验证和判别任务的帮助有限，但在联合规则诊断中的诊断定位任务上可能有所助益。这凸显了制定任务感知型提示策略的必要性，以平衡准确率与计算效率。

5 局限性

FinRule-Bench 评估规则约束推理，其前提是已提供相关会计原则，且财务表格完整并在内部保持一致。该基准不涉及规则发现、在含噪或存在缺失数据时的推理、跨报表勾稽或领域知识检索

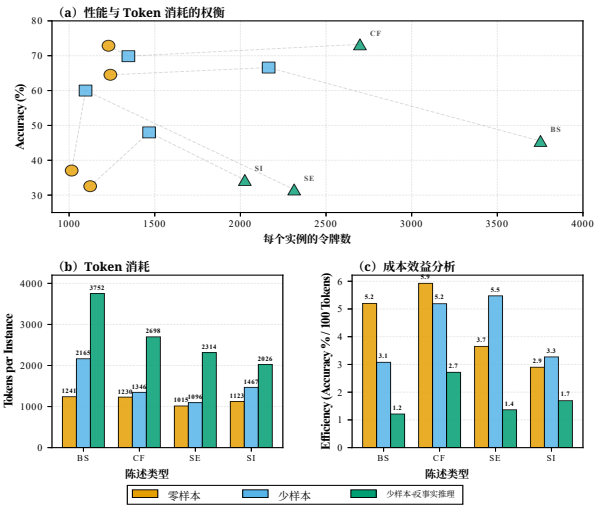


图6：规则识别的性能-成本分析。

所提供规则之外的内容。因此，它旨在对联合推理忠实度进行有针对性的诊断，而非作为通用的财务分析基准。我们也不研究监督式微调或强制生成思维链。我们的目标是在固定参数和受控任务条件下，评估预训练基础模型的行为，而不是通过额外监督或解释约束来优化性能。虽然微调或显式推理轨迹可能提高准确率，但它们也会引入混杂因素、优化选择以及不属于本基准范围的解释冗长性。

6 结论

我们推出了**FinRule-Bench**，这是一个用于评估在明确会计原则下，针对财务表格进行基于规则的合规推理的基准。不同于以问答或对合成扰动数据进行异常检测为重点的以往财务和表格基准，FinRule-Bench 评估的是在真实财务报表上执行正式会计规则的能力。该基准定义了三项审计任务：单规则验证、违规规则识别和联合多错误诊断，所需推理能力逐步增强。对不同模型和提示策略的评估结果显示，大语言模型虽然能够可靠地处理孤立的算术检查，但在结构推理、条件适用性以及多个相互作用规则的组合方面，性能会急剧下降。尤其是联合诊断揭示了模型在检测违规与实现完整的记录级定位之间存在差距。这些失败具有系统性，且因规则而异：当推理需求从表层数值一致性转向全面的上下文及跨记录执行时，性能会随之下降。通过以明确的会计原则和真实的财务报表为评估基础，FinRule-Bench 为评估高风险表格领域中的联合诊断推理提供了有原则的框架，并推动未来研究迈向可靠的端到端合规验证。

伦理声明

本研究推出*FinRule-Bench*，这是一个研究基准，用于评估大语言模型在结构化财务报表上的基于规则的推理和诊断能力。该基准仅用于研究和评估目的，旨在分析明确会计原则下的推理保真度、覆盖范围和失败模式。它并非为部署于实际审计、监管执法或金融决策系统而设计。本基准使用的所有财务报表均来自公开可得的公司披露，主要为根据监管要求发布的 10-K 表格备案文件。该数据集仅处理汇总财务报表表格和会计科目，不会收集、推断或重构任何个人身份信息、个人级财务记录或机密数据。因此，该数据集不会带来超出原始公开备案文件本身所固有风险之外的新增隐私风险。

为实现受控评估，FinRule-Bench 纳入了以规则为依据、通过确定性方式生成的注入错误变体，这些变体基于原本正确的财务报表。这些合成违规仅用于检验模型推理和诊断的完整性，并明确标记为评估产物。不应将模型在该基准上的性能视为其适用于现实世界审计或合规验证的证据，除非有具备资质的人员进行监督。我们承认，该基准反映了其源数据所处的制度和监管背景，包括美国财务披露中常用的会计准则和报告惯例。因此，在 FinRule-Bench 上评估的模型可能表现出特定领域或司法管辖区的行为特征。本文记录了这些局限性，并鼓励未来将基准扩展至更多会计框架和司法管辖区。人工参与数据集构建的工作仅限于整理会计原则、实施并验证确定性规则验证器，以及核实任务定义。大型语言模型仅用于受限的辅助用途（例如协助转录表格或编辑稿件），未用于生成真实标签。所有评估标签均由确定性验证器生成，以确保结果可复现且客观。

总体而言，该基准旨在通过支持对结构化金融领域中的推理失败进行透明分析，并避免语言模型在高风险金融场景中被过早或不加甄别地部署，来促进负责任的数据挖掘和 AI 研究。

参考文献

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yi-heng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-VL 技术报告。arXiv:2502.13923 [cs.CV] https://arxiv.org/abs/2502.13923[2] Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA：一个关于金融数据数值推理的数据集。*2021* 年 *EMNLP* 论文集 (2021)。[3] Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. ConvFinQA：探索对话式金融问答中的数值推理链。载于*2022* 年自然语言处理经验方法会议论文集，Yoav Goldberg、Zornitsa Kozareva 和 Yue Zhang（编）。协会计算语言学协会，阿布扎比，阿拉伯联合酋长国，6279–6292。doi: 10 18653 1 2022 421 。/v / .emnlp-main。Manan Roy Choudhury、[4] Adithya Chandramouli、Mannan Anand 和 Vivek Gupta。2026。Better Call CLAUSE：用于审计大语言模型法律推理能力的差异基准。arXiv:2511.00340 [cs.AI] https://arxiv.org/abs/2511. 00340[5] Manan Roy Choudhury、Anirudh Iyengar Kaniyar Narayana Iyengar、Shikhhar Siingh、Sugeeth Puranam 和 Vivek Gupta。2025。TABARD：用于表格异常分析、推理与检测的新型基准。载于计算语言学协会论文集：*EMNLP 2025*，Christos Christodoulopoulos、Tanmoy Chakraborty、Carolyn Rose 和 Violet Peng（编）。计算语言学协会，中国苏州，21783–21817。doi:10.18653/v1/2025. findings-emnlp.[6] Antonia Creswell、Murray Shanahan 和 Irina Higgins。2022。选择-推理：利用大语言模型实现可解释逻辑推理。arXiv: 2205 09712 [ ] 2205 09712 . cs.AI https://arxiv.org/abs/ 。[7] Xiang Deng、Huan Sun、Alyssa Lees、You Wu 和 Cong Yu。2022。TURL：通过表征学习实现表格理解。*SIGMOD Rec.* 51，1（2022年6月），33–40。doi:10.1145/3542700.3542709[8] Aaron Grattafiori 等。2024。Llama 3 模型家族。arXiv: 2407 21783 [ ] 2407 21783 . cs.AI [9] https://arxiv.org/abs/ 。Gheorghe Comanici 等。2024。Gemini 1.5：凭借高级推理、多模态、长上下文和新一代智能体能力推动前沿发展。arXiv: 2507 06261 [ ] 2507 06261 . cs.CL [10] https://arxiv.org/abs/ 。OpenAI 等。2024。GPT-o 系统卡。arXiv: 。cs.CL https: // arxiv.org/abs/2410.21276[11] Xi Fang、Weijie Xu、Fiona Anting Tan、Jiani Zhang、Ziqing Hu、Yanjun Qi、Scott Nickleach、Diego Socolinsky、Srinivasan Sengamedu 和 Christos Faloutsos。2024。表格数据中的大语言模型（ LLM ）：预测、生成与理解 —— 综述。arXiv:2402.17944 [cs.CL] https://arxiv.org/abs/ 2402.17944[12] 美国财务会计准则委员会。无日期。财务会计准则委员会。https://www.fasb.org。[13] George Fatouros、Kostas Metaxas、John Soldatos 和 Manos Karathanassis。2025。MarketSenseAI 2.0：利用大语言模型智能体增强股票分析。arXiv: 2502 00415 [ ] 2502 00415 . q-fin.CP [14] https://arxiv.org/abs/ 。Artur d'Avila Garcez 和 Luís C. Lamb。2023。神经符号人工智能：第三次浪潮。人工智能评论56，11（2023年3月），12387–12406。doi:10.1007/s10462-023- 10448-w[15] Jiayi Gui、Yiming Liu、Jiale Cheng、Xiaotao Gu、Xiao Liu、Hongning Wang、Yuxiao Dong、Jie Tang 和 Minlie Huang。2024。LogicGame：大语言模型规则推理能力基准测试。arXiv:2408.15778 [cs.AI] https://arxiv.org/abs/ 2408 15778 。Demis Hassabis 和 Koray Kavukcuoglu。2024。[16] 推出 *Gemini 2.0*：面向智能体时代的全新人工智能模型。https://blog.google/innovation-and-ai/models-and-research/google-deepmind/google-gemini-ai-update-december-2024/[17] Jonathan Herzig、Pawel Krzysztof Nowak、Thomas Müller、Francesco Piccinno 和 Julian Eisenschlos。2020。TaPas：通过预训练实现弱监督表格解析。载于第*58*届计算语言学协会年会论文集，Dan Jurafsky、Joyce Chai、Natalie Schluter 和 Joel Tetreault（编）。计算语言学协会，线上，4320–4333。doi:10.18653/v1/2020.acl-main.398[18] Aidan Hogan、Eva Blomqvist、Michael Cochez、Claudia D'amato、Gerard De Melo、Claudio Gutierrez、Sabrina Kirrane、José Emilio Labra Gayo、Roberto Nav- igli、Sebastian Neumaier、Axel-Cyrille Ngonga Ngomo、Axel Polleres、Sabbir M. Rashid、Anisa Rula、Lukas Schmelzeisen、Juan Sequeda、Steffen Staab 和 An- toine Zi mmermann。2021。知识图谱。*ACM*计算调查54，4，文章 71 2021 37 10 1145 3447772 （7月），页码。doi: 。/ Giorgos [19] Iacovides、Thanos Konstantinidis、Mingxue Xu 和 Danilo Mandic。2024。FinLlama：用于算法交易的基于大语言模型的金融情感分析。载于第*5*届美国计算机协会人工智能金融国际会议（美国纽约州布鲁克林）（*ICAIF '24*）。美国计算机协会，美国纽约州纽约，134 141 10 1145 3677052 3698696 – 。doi: 。/ 。IASB。无日期。关于国际会计准则理事会。https://www. ifrs.org。[21] Jinhao Jiang、Kun Zhou、Zican Dong、Keming Ye、Xin Zhao 和 Ji-Rong Wen。2023。StructGPT：大语言模型对结构化数据进行推理的通用框架。载于*2023*年自然语言处理实证方法会议论文集，Houda Bouamor、Juan Pino 和 Kalika Bali（编）。计算语言学协会，新加坡，9237–9251。doi: 10 18653 1 2023 574 . /v / .emnlp-main。Jeonghoon Kim、Heesoo [22] Jung、Hyeju Jang 和 Hogun Park。2024。利用上下文感知查询表示学习增强知识图谱中的多跳逻辑推理。载于计算语言学协会论文集：*ACL 2024*，Lun-Wei Ku、Andre Martins 和 Vivek Srikumar（编）。计算语言学协会，泰国曼谷，15978–15991。doi: 10 18653 1 2024 946 . /v / .findings-acl。Sukanth Korkanti。2024。利[23] 2024 用大语言模型和高级数据分析增强金融欺诈检测。载于2024年第2届自我

可持续人工智能系统（*ICSSAS*）。1328–1334。doi:10.1109/ICSSAS64001.2024.10760895[24] Andrei Lazarev 和 Dmitrii Sedov。2024。利用现代大语言模型（LLM）进行金融趋势分析与摘要生成。载于*2024*年第*6*届控制系统、数学建模、自动化与能源效率国际会议（*SUMMA*）。317–321。doi:10.1109/SUMMA64428.2024.10803746[25] Jean Lee、Nicholas Stevens 和 Soyeon Caren Han。2025。金融领域的大语言模型（FinLLMs）。《神经计算与应用》（2025），1–15。[26] Peng Li、Yeye He、Dror Yashar、Weiwei Cui、Song Ge、Haidong Zhang、Danielle Rifinski Fainman、Dongmei Zhang 和 Surajit Chaudhuri。2024。Table-GPT：针对多种表格任务进行微调的GPT。*ACM*管理数据论文集2，3，文章176（2024年5月），28页。doi:10.1145/3654979[27] Chenghao Liu、Arunkumar Arulappan、Ranesh Naha、Aniket Mahanti、Joarder Kamruzzaman 和 In-Ho Ra。2024。大语言模型与金融市场情感分析：综述、数据集及案例研究。*IEEE Access* 12 2024 134041 134061 10 1109 2024 3445413 ( ), – 。doi: 。/ACCESS. 。[28] Hanmeng Liu、Ruoxi Ning、Zhiyang Teng、Jian Liu、Qiji Zhou 和 Yue Zhang。2023。评估 ChatGPT 和 GPT-4 的逻辑推理能力。arXiv: 2304 03439 [ ] 2304 03439 . cs.CL https://arxiv.org/abs/ 。[29] Lihui Liu、Zihao Wang 和 Hanghang Tong。2025。知识图谱上的神经符号推理：从查询视角出发的综述。*SIGKDD*探索通讯27，1（2025年7月），124–136。doi:10.1145/3748239.3748249[30] Shu Liu、Shangqing Zhao、Chenghao Jia、Xinlin Zhuang、Zhaoguang Long、Jie Zhou、Aimin Zhou、Man Lan 和 Yang Chong。2025。FinDABench：大语言模型金融数据分析能力基准测试。载于第*31*届国际计算语言学会议论文集，Owen Rambow、Leo Wanner、Marianna Apidianaki、Hend Al-Khalifa、Barbara Di Eugenio 和 Steven Schockaert（编）。计算语言学协会，阿布扎比，阿拉伯联合酋长国，710–725。https://aclanthology.org/2025.coling-main.48/[31] Xiao-Yang Liu、Guoxuan Wang、Hongyang Yang 和 Daochen Zha。2023。Fin-GPT：让互联网规模数据惠及金融大语言模型。arXiv: 2307 10485 [ ] 2307 10485 . cs.CL https://arxiv.org/abs/ 。[32] Weizheng Lu、Jing Zhang、Ju Fan、Zihao Fu、Yueguo Chen 和 Xiaoyong Du。2025。用于表格处理的大语言模型：综述。《计算机科学前沿》19, 2 2025 10 1007 11704 024 40763 6（1月）。doi: 。/s - - - Qing[33] Lyu、Shreya Havaldar、Adam Stein、Li Zhang、Delip Rao、Eric Wong、Marianna Apidianaki 和 Chris Callison-Burch。2023。忠实的思维链推理。arXiv: 2301 13379 [ ] 2301 13379 . cs.CL[34] https://arxiv.org/abs/ 。美国公众公司会计监督委员会。无日期。美国公众公司会计监督委员会。https://pcaobus.org。[35] Jingru Wang、Wen Ding 和 Xiaotong Zhu。2025。财务分析：基于大语言模型检索增强生成的智能金融数据分析系统。arXiv:2504.06279 [q-fin.ST] 2504 06279 https://arxiv.org/abs/ 。Rushi Wang、Jiateng[36] Liu、Weijie Zhao、Shenglan Li 和 Denghui Zhang。2025。AuditBench：用于财务报表审计的大语言模型基准。载于人工智能交通运输国际研讨会。Springer，59–81。[37] Zhiruo Wang、Haoyu Dong、Ran Jia、Jia Li、Zhiyi Fu、Shi Han 和 Dongmei Zhang。2021。TUTA：用于通用结构化表格预训练的基于树的Transformer。载于*KDD '21*：第*27*届美国计算机协会知识发现与数据挖掘会议，线上活动，新加坡，*2021*年*8*月*14*日至*18*日，Feida Zhu、Beng Chin Ooi 和 Chunyan Miao（编）。美国计算机协会，1780–1790。doi:10.1145/3447548.3467434[38] Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Brian Ichter、Fei Xia、Ed H. Chi、Quoc V. Le 和 Denny Zhou。2022。思维链提示可激发大语言模型的推理能力。载于第*36*届神经信息处理系统大会论文集（美国路易斯安那州新奥尔良）（*NIPS '22*）。Curran Associates Inc.，美国纽约州雷德胡克，文章1800，14页。[39] Changyi Xiao 和 Yixin Cao。2024。通过校准知识图谱补全模型实现复杂逻辑查询回答。载于计算语言学协会论文集：*ACL 2024*，Lun-Wei Ku、Andre Martins 和 Vivek Srikumar（编）。计算语言学协会，泰国曼谷，13792 13803 10 18653 1 2024 819 – 。doi: 。/v / .findings-acl。[40] Yunhu Ye、Binyuan Hui、Min Yang、Binhua Li、Fei Huang 和 Yongbin Li。2023。大语言模型是多面手式分解器：为基于表格的推理分解证据与问题。载于第*46*届美国计算机协会*SIGIR*信息检索研究与发展会议论文集（中国台北）（*SIGIR '23*）。美国计算机协会，美国纽约州纽约，174 184 10 1145 3539618 3591708 – 。doi: 。/ 。[41] Pengcheng Yin、Graham Neubig、Wen-tau Yih 和 Sebastian Riedel。2020。TaBERT：文本与表格数据联合理解的预训练。载于第*58*届计算语言学协会年会论文集，Dan Jurafsky、Joyce Chai、Natalie Schluter 和 Joel Tetreault（编）。计算语言学协会，线上，8413–8426。doi:10.18653/v1/2020.acl- main.745[42] Tianshu Zhang、Xiang Yue、Yifei Li 和 Huan Sun。2024。TableLlama：迈向面向表格的开放式大型通用模型。载于*2024*年计算语言学协会北美分会人类语言技术会议论文集（第*1*卷：长篇论文），Kevin Duh、Helena Gomez与Steven Bethard（编）。计算语言学协会，墨西哥城，墨西哥，6024–6044。doi:10.18653/v1/2024.naacl-long.335[43] Huaqin Zhao、Zheng Liu、Zihao Wu、Yiwei Li、Tianze Yang、Peng Shu、Shaochen Xu、Haixing Dai、Lin Zhao、Gengchen Mai、Ninghao Liu与Tianming Liu。2024。用大语言模型革新金融：应用与洞见综述。*ArXiv*abs/2401.11641 (2024)。https://api.semanticscholar.org/CorpusID:267069342[44] Fengbin Zhu、Wenqiang Lei、Youcheng Huang、Chao Wang、Shuo Zhang、Jiancheng Lv、Fuli Feng与Tat-Seng Chua。2021。TAT-QA：面向金融领域表格与文本内容混合数据的问答基准。载于第*59*届计算语言学协会年会暨第*11*届自然语言处理国际联合会议论文集（第*1*卷：长文）。计算语言学协会，线上，3277–3287。doi:10.18653/v1/2021.acl-long.254

A 形式化任务定义

本附录给出FinRule-Bench所用任务的形式化定义。收录这些定义是为了保证完整性与可复现性。正文以面向评测的表述呈现任务描述。

A.1 预备知识

令 T 表示一个由 m 条记录组成的结构化财务表 t_1，……，t_m。令 $R = \{r_1, \ldots\ldots, r_k\}$ 表示适用于 T 的会计规则集合。每条规则 $r \in R$ 都关联一个确定性验证函数

$$v_r(T) \in \{\texttt{True}, \texttt{False}\},$$

该函数用于评估 T 是否满足规则 r。验证器的输出被视为模型评估的真实标签。

A.2 规则验证

在规则验证任务中，模型会获得一张财务表 T和一条会计规则 $r \in R$。目标是确定 T是否满足指定规则。

形式化地说，该任务要求生成一个二元输出

$$y \in \{\texttt{True}, \texttt{False}\}$$

使得

$$y = v_r(T).$$

该任务对应于二元分类，用于评估模型在结构化财务表上验证明确会计原则的能力。

A.3 规则识别

在规则识别任务中，模型会获得一张财务表 T和一组会计规则 R。其中恰好有一条规则 $r^\star \in R$ 被违反。

目标是输出一个规则标识符 $\hat{r}$，使得 t

$$\hat{r} = r^\star,$$

其中约束条件

$$|\{r \in R \mid v_r(T) = \texttt{False}\}| = 1$$

对本任务中所有报表类型的记录均成立。

评估时，将预测的规则标识符与被违反的规则进行精确匹配。

A.4 联合规则诊断

在联合规则诊断任务中，模型会获得一张财务表 T和一组会计规则 R。其中任意子集的规则都可能被违反。

该任务包含两个阶段。

违规检测。 模型会生成一个二元决策

$$d \in \{\texttt{Yes}, \texttt{No}\},$$

其中

$$d = \texttt{Yes} \iff \exists r \in R \text{ such that } v_r(T) = \texttt{False}.$$

违规识别。 如果 d = 是，则模型会生成一张映射

$$M : \{1, \ldots, m\} \to 2^R,$$

其中， $M(i)$ 表示模型针对记录 t_i 判定为被违反的会计规则集合。

若要使预测正确，以下两个条件必须同时成立：

(1) 违规检测决策 d 与真实标签一致。(2) 预测规则集 $M(i)$ 与验证器定义的记录级违规规则一致。

此任务评估在多个同时发生的违规情况下进行诊断推理的能力，并要求准确检测和完整定位。

B 会计规则与验证器规范

在此，我们介绍 FinRule-Bench 中使用的会计规则，以及用于生成真实标签的确定性验证器。验证器为每条规则提供可执行规范，下面展示了具有代表性的错误检测算法。

B.1 验证器接口

每条会计规则 r 都关联一个确定性验证函数：

$$v_r : T \to \{\texttt{True}, \texttt{False}\},$$

其中， T是一个结构化财务表，以有序的记录集合表示，包含数值和层次元数据。若该表满足规则，验证器则返回 True；否则返回 False。

所有任务的真实标签均通过对原始表或注入错误的表评估相关验证器来计算。

B.2 代表性规则类别

我们根据验证所需的主导推理操作，将会计规则分为四类。

算术规则。 算术规则强制执行数值一致性关系。例如，资产负债表必须满足会计恒等式：

资产总额 = 负债总额 +总权益。

算法 1： 算术验证器（BS01）

```
def validate_BS01(表格):
    assets ← table["资产总额"]
    liabilities ← table["负债总额"]
    equity ← table["总权益"]
    // 检查会计恒等式
    return assets = liabilities + equity
```

结构规则。 结构规则约束记录的存在、位置或标注。这些规则不需要进行算术计算，但依赖于在表格层级中进行正确分类。

算法 2： 结构验证器 (SI03)

```
def validate_si03(table):
    // 非控制性权益不应在
        负债
    返回 "非控制性权益" ∈ table.labels()
```

条件规则。 条件规则仅在满足特定条件时适用。验证需要同时检查其适用性和正确性。

算法 3： 条件验证器 (SI05)

```
定义validate_si05(表格):
    如果 table.contains("利息费用") 则
        // 如果存在，则不得为 None
        返回 table.value("利息费用") ≠ None
    返回 True
```

多节规则。 多节规则用于捕捉跨多个行或时间段的依赖关系，例如滚动结转或对账约束。

算法 4： 多节验证器 (SE01)

```
def validate_se01(table):
对于 year ∈ table.years() 执行
        begin ← table.value(year, “期初余额”)
        change ← table.value(year, "净变动")
        end ← table.value(year, "期末余额")
        // 滚动结转检查
        if begin + change ≠ end then
            return False
    return True
```

B.3 通过验证器反转进行错误注入

错误注入实例通过进行最小编辑生成，这些编辑会反转目标验证器的判定结果，同时保留表格结构。对于规则 r，我们构造一个修改后的表格 T'，使得：

$$v_r(T) = \texttt{True} \quad \text{and} \quad v_r(T') = \texttt{False}.$$

对于算术规则，这通常涉及扰动单个数值。对于结构规则或条件规则，这涉及删除、重命名或移动单条记录。对于多记录规则，则修改依赖链中的单个组成部分。所有注入都是确定且可逆的。

B.4 面向联合诊断的验证器组合

对于联合规则诊断任务，通过在同一张表上评估所有验证器来计算真实标签：

$$V(T) = \{r \in R \mid v_r(T) = \texttt{False}\}.$$

当且仅当模型预测准确识别出集合 $V(T)$，并将每条被违反的规则与正确的记录相关联时，才视为正确。这一严格标准确保将部分检测和错误定位视为不同的失败模式。

B.5 可复现性保证

对于相同的输入表，验证器始终会生成相同的输出。这种设计确保所有报告的结果都完全可复现，并且任何性能差异都仅源于模型行为，而非标注噪声或评估器歧义。会计规则见表 22

C 错误注入与数据集实例化

本附录介绍了如何从干净财务报表的基础语料中创建特定任务实例。所有流程均为确定性流程、可逆，并已在发布的制品集中实现。

C.1 符号与设置

设 T 为一张干净的基础表，$R = \{r_1, \ldots\ldots, r_k\}$ 为 T（特定于报表）的规则集。设 $v_r(T)$ 为规则 r 的确定性验证器；对于干净表，该验证器满足：对所有 $r \in R$，$v_r(T)$ = 为 True。针对规则 r 的注入会构造一张扰动表 T_r，使得 $v_r(T_r)$ = 为 False，同时保留 T 的结构属性。

在我们的数据集中：

- 资产负债表（BS）：5 条规则。
- 现金流量表（CF）：3 条规则。
- 权益变动表（SE）：3 条规则。
- 利润表（SI）：5 条规则。

C.2 单一违规注入

对于规则验证和规则识别任务，我们按如下方式生成单一违规实例。该过程具有确定性，并且能够感知规则：

算法5： 单一违规注入

```
def注入_single(T, r):
    // 前置条件： v_r(T) == 真
    识别验证器检查和最小依赖字段
      针对规则 r
    确定一个最小编辑（数值扰动、标签重命名，
      或行重定位），使 v_r 返回 False
    对 T 应用编辑，得到 T′
    记录编辑脚本 δ = diff(T, T′)，以确保可逆性
    返回 (T′ δ)
```

C.3 多违规构造（困难任务）

对于联合诊断任务，我们创建了包含多个同时被违反规则的多违规实例。这些实例按照以下方式从干净表中确定性地构造：

C.4 数据集实例化策略

任务视图从同一基础集合中的干净表中确定性地派生：

算法 6： 多违规构造

定义 inject_multi(T、 $\{r_a$、 r_b、 .....$\}$)：
  // 前置条件：对于所有 r， $v_r(T)$ == 为 True
  初始化 $T_0 \leftarrow T$
  针对预定义序列中的每条目标规则 r_i **执行**
    计算 $(T_i,\ \delta_i) \leftarrow$ 注入_单个$(T_{i-1},\ r_i)$
    对所有规则验证 $v_{r_J}(T_i)\, r_j$；如果注入意外地翻转了非目标验证器，则应用次级最小编辑以仅恢复所需的翻转（同时保持确定性）
  记录组合编辑脚本 $\Delta = \{\delta_1,\ \delta_2,\ .....\}$
  返回 $(T_{\text{final}},\ \Delta)$

- **干净实例：** 原始提取的 Markdown 表格，未经任何编辑。
- **规则验证 / 违规识别实例：** 针对每条规则，确定性地生成单一违规实例；同时确保每个实例中被违反的规则唯一。
- **联合多步诊断实例：** 按照上述描述生成多违规记录；在我们的实验中，每种陈述类型生成了 100 条多违规记录。

该策略确保所有任务视图具有可比性、可复现性，并且均源自相同的真实世界披露信息。

D 财务报表样本与结构特征

本附录列举了 FinRule-Bench 中使用的财务报表的代表性示例。我们同时纳入了原始源文档及其抽取后的表格表示，以展示数据来源、抽取保真度和结构复杂度。这些示例旨在提高透明度，并非详尽无遗。

D.1 原始财务报表来源

图7展示了公司申报文件中报告的代表性财务报表快照。这些图像对应于提取 FinRule-Bench 表格所依据的原始源文档。

D.2 提取后的 Markdown 表格表示

表6展示了一个财务报表示例，该报表经提取后转换为用作模型输入的 Markdown 格式。提取后的表格保留了原始申报快照7中的章节层级、记录顺序、标签和数值。

D.3 结构特征

FinRule-Bench 中的财务报表在长度、层级和内部依赖方面存在显著差异。表格从结构相对扁平的紧凑报表，到包含多个章节、嵌套子节、小计行和跨记录约束的复杂报告，不一而足。

E 数据集验证

本节报告用于评估标注一致性和可靠性的标注者间一致性指标。我们还将介绍提供给两位标注者的标注指南，其中界定了人工审查期间遵循的验证标准和流程。

E.1 人工验证

为评估数据集的可靠性以及受控规则违规的正确性，我们对随机抽取的15%扰动财务表格子集进行了人工验证。该子集被划分为两个互不相交的样本，记为 α_1 和 α_2，各覆盖数据集的7.5%，以独立评估标注一致性。结果如表8所示。

表8：人工验证的标注者间一致性结果

样本	科恩卡帕系数（%）	杰卡德系数（%）
α_1	93.3	97.5
α_2	95.3	95.8

两名具备自然语言处理专业知识且拥有金融领域经验的标注者，独立审阅了抽样表格。我们使用科恩卡帕系数衡量违规规则识别的一致性，并使用杰卡德系数衡量记录级违规定位的重合程度，以量化标注者间的一致性。我们分别针对 α_1 和 α_2 计算一致性，以评估不同独立样本之间的稳定性。这些一致性指标既验证了数据集构建流程，也为用于大规模标注的确定性规则验证器的可靠性提供了额外检验。

E.2 标注指南

在开始人工验证之前，我们向标注者提供了一套详细的书面指南，以确保不同标注之间的一致性、清晰度和可复现性。指南旨在统一对会计规则违规的解读、验证方式，以及在结构化财务表中的定位方法。

我们要求每位标注者独立审阅经过扰动的财务表及其相关背景信息，包括所陈述的会计原则，以及通过受控错误注入引入的预期违规。标注者需要评估注入的违规是否（*i*）按照指定的会计规则正确实施，（*ii*）根据财务报表的结构和语义在上下文中合理，且（*iii*）准确定位到表格中的相关行或记录。我们明确要求标注者不得推断出除已引入违规之外的其他违规，并且只需核验所提供违规的正确性和范围。

指南强调，标注应以每个实例所提供的明确会计原则为依据，而不应依赖隐含假设、外部知识或其他会计解释。如果某项违规存在歧义，则要求标注者根据其

表6：提取的资产负债表示例（万事达卡，2023年12月31日）。该表展示了 FinRule-Bench 中使用的层级结构、小计、跨区段合计和符号约定。为简洁起见，省略了各个行项目。

章节	小节	Item	值
资产	流动资产	现金及现金等价物	8,588
		...	...
		流动资产总额	**18,961**
	非流动资产	商誉	7,660
		...	...
资产总额			**42,448**
负债	流动负债	应付账款	834
		...	...
		流动负债总额	**16,264**
	非流动负债	长期债务	14,344
		...	...
负债总额			**35,451**
夹层权益		可赎回非控股权益	**22**
权益	股东权益	A类库存股，按成本	-60,429
		...	...
		总权益	**6,975**
负债和权益总额			**42,448**

表 7：结构复杂度统计。

陈述类型	# 记录	Avg Rows	Min Rows	Max Rows	Avg 章节	比例 含小节	比例。 含小计	平均列数 (最小值-最大值)
资产负债表（资产负债表）	212	32.2	7	49	4.4	1.00	1.00	4.0（4-4）
现金流量表（现金流量）	318	34.4	7	60	3.4	1.00	1.00	4.0（4-4）
权益变动表（SE）	287	10.1	1	29	10.0	1.00	0.98	8.5（5-16）
利润表（SI）	300	22.4	2	40	7.9	1.00	0.99	3.0（3-3）

判断进行标注，即依据所述规则定义，判断该违规是否有合理依据，而不试图通过主观解释消除歧义。

此外，还向标注者提供了说明性示例，以展示常见的规则类别，包括算术一致性规则、结构分类规则、条件适用规则和多记录约束。这些示例旨在使标注者的判断与数据集构建中使用的确定性规则验证器保持一致，同时保留独立的人工作评估。

标注过程完全自愿。两位标注者均为自然语言处理研究人员，之前曾从事金融领域数据和会计相关任务；他们因对该研究问题及其在评估高风险金融场景中的推理可靠性方面的潜在影响感兴趣而选择参与。标注者未接触任何敏感信息或个人身份信息，也未被要求执行超出验证合成研究型数据伪影范围的任务。

标注期间未进行裁定或共识构建。分歧均予以保留，仅用于计算标注者间一致性指标；这些指标是标注一致性的经验性度量，而非一种机制用于纠正标签。该设计确保所报告的一致性反映的是真正的独立判断，而非事后对齐。

F 实验结果

本节展示了补充实验结果，以补充第4节中的主要发现。

表9：规则合规性验证性能（GPT-4o）

陈述	方法	Acc	精确率	召回率	F1分数	词元（K）
BS	零样本	0.749	0.824	0.890	0.855	967.7
	少样本	0.812	0.882	0.894	0.888	998.0
	少样本 + 反事实推理	**0.884**	**0.962**	**0.897**	**0.928**	1003.1
CF	零样本	0.402	0.696	0.361	0.476	982.5
	少样本	0.620	0.819	0.634	0.714	1009.2
	少样本 + 反事实推理	**0.643**	**0.849**	**0.638**	**0.728**	1030.8
SE	零样本	0.684	0.734	**0.908**	**0.812**	548.3
	少样本	**0.706**	**0.887**	0.696	0.780	616.6
	少样本 + 反事实推理	0.626	0.753	0.746	0.750	616.1
SI	零样本	0.724	0.834	0.835	0.834	713.9
	少样本	**0.806**	**0.882**	**0.885**	**0.884**	825.3
	少样本 + 反事实推理	0.797	0.881	0.874	0.878	1014.2
总体	零样本	0.640	0.772	0.749	0.744	3212.4
	少样本	0.736	**0.868**	0.777	0.816	3449.1
	少样本 + 反事实推理	**0.738**	0.861	**0.789**	**0.821**	3664.2

表 10：规则合规性验证上的性能（Gem- ini2.5 Pro）

陈述方法		Acc	精确率	召回率	F1分数	词元（K）
BS	零样本	**0.739**	**0.835**	**0.857**	**0.845**	1272.0
	少样本	0.676	0.810	0.798	0.804	2617.9
	少样本 + 反事实推理	0.473	0.768	0.526	0.624	3316.7
CF	零样本	0.619	0.803	0.652	0.720	2067.7
	少样本	0.606	0.800	**0.707**	**0.751**	1595.3
	少样本 + 反事实推理	**0.628**	**0.886**	0.578	0.700	2415.5
SE	零样本	0.634	0.833	0.641	0.724	2114.6
	少样本	0.585	0.795	0.602	0.685	2411.0
	少样本 + 反事实推理	**0.651**	**0.837**	**0.663**	**0.740**	2820.8
SI	零样本	0.459	0.757	0.517	0.615	2504.3
	少样本	0.647	0.806	**0.759**	0.782	2637.8
	少样本 + 反事实推理	**0.658**	**0.833**	0.738	**0.783**	3491.9
总体	零样本	0.614	0.792	0.615	0.726	7958.6
	少样本	0.629	0.803	**0.717**	**0.756**	9262.0
	少样本 + 反事实推理	**0.643**	**0.832**	0.676	0.746	12044.9

表 11：规则合规性验证性能（Gemini 2.0）

陈述方法		Acc	精确率	召回率	F1分数	词元（K）
BS	零样本	**0.759**	0.870	0.836	0.853	1109.4
	少样本	0.519	**0.900**	**0.862**	**0.881**	1142.3
	少样本 +反事实推理	0.297	0.840	0.739	0.786	2257.7
CF	零样本	0.805	0.803	**0.652**	0.417	1045.6
	少样本	0.858	0.919	0.540	0.680	1070.7
	少样本 + 反事实推理	**0.907**	**0.953**	0.632	**0.760**	2522.6
SE	零样本	0.105	0.705	0.714	0.710	639.8
	少样本	0.150	0.717	0.719	0.718	761.8
	少样本 + 反事实推理	**0.190**	**0.815**	**0.752**	**0.782**	819.9
SI	零样本	0.130	0.792	0.661	0.720	791.2
	少样本	**0.210**	**0.843**	**0.851**	**0.847**	906.0
	少样本 + 反事实推理	0.087	0.808	0.771	0.789	3769.9
总体	零样本	**0.450**	0.793	0.716	0.675	3586.1
	少样本	0.434	0.845	**0.743**	**0.781**	3880.8
	少样本 + 反事实推理	0.370	**0.854**	0.724	0.779	9370.1

表12：规则合规性验证上的性能（LLaMA）

陈述方法		Acc	精确率	召回率	F1分数	词元（K）
BS	零样本	0.606	0.801	0.709	0.748	1006.1
	少样本	0.749	0.873	**0.818**	**0.845**	1036.6
	少样本 +反事实推理	**0.759**	**0.959**	0.743	0.837	2016.7
CF	零样本	0.458	**0.808**	0.364	0.502	1027.8
	少样本	0.567	0.765	0.609	0.678	1053.9
	少样本 +反事实推理	**0.579**	0.762	**0.638**	**0.695**	2435.7
SE	零样本	0.628	0.806	0.663	0.728	584.3
	少样本	**0.745**	**0.915**	**0.727**	**0.810**	653.2
	少样本 + 反事实推理	0.688	0.899	0.657	0.760	692.1
SI	零样本	0.520	0.791	0.577	0.667	770.2
	少样本	**0.706**	0.839	**0.801**	**0.819**	883.3
	少样本 + 反事实推理	0.698	**0.851**	0.774	0.811	3376.8
总体	零样本	0.553	0.802	0.578	0.661	3388.4
	少样本	**0.692**	0.848	**0.739**	**0.788**	3627.0
	少样本 + 反事实推理	0.681	**0.868**	0.703	0.776	8521.3

表13：单一被违反规则识别的性能（GPT-4o）

陈述方法		准确率	平衡准确率	宏平均精确率	宏平均召回率	宏F1	加权 F1
BS	零样本	0.663	0.663	**0.751**	0.663	0.641	0.641
	少样本	**0.724**	**0.724**	0.750	**0.724**	**0.702**	**0.702**
	少样本 +反事实推理	**0.724**	**0.724**	0.750	**0.724**	**0.702**	**0.702**
CF	零样本	0.841	0.842	0.510	0.505	0.504	0.843
	少样本	**0.868**	**0.861**	0.516	0.516	0.516	**0.868**
	少样本 +反事实推理	0.782	0.782	**0.645**	**0.563**	**0.587**	0.816
SE	零样本	0.408	0.409	0.294	0.245	0.200	0.333
	少样本	**0.420**	**0.421**	0.348	0.252	0.205	**0.341**
	少样本 +反事实推理	0.416	0.416	**0.435**	**0.312**	**0.255**	0.340
SI	零样本	0.384	0.377	**0.571**	0.377	0.338	0.345
	少样本	**0.414**	**0.409**	0.535	**0.409**	**0.368**	**0.375**
	少样本 +反事实推理	0.389	0.389	0.446	0.324	0.300	0.360

表 14：单条被违反规则识别的性能（Gemini2.5Pro）

陈述方法		准确率	平衡准确率	宏平均精确率	宏平均召回率	宏F1 加权 F1
BS	零样本	0.472	0.611	0.730	0.611	0.616 **0.691**
	少样本	0.428	0.674	0.708	0.674	0.652 0.671
	少样本 +反事实推理	**0.491**	**0.729**	**0.759**	**0.729**	**0.666** 0.668
CF	零样本	0.937	0.938	**1.000**	0.938	**0.968 0.968**
	少样本	**0.943**	**0.943**	0.926	**0.943**	0.944 0.944
	少样本 +反事实推理	0.664	0.743	0.520	0.743	0.726 0.818
SE	零样本	0.340	0.549	0.590	0.549	0.552 0.566
	少样本	0.326	0.526	0.458	0.526	0.476 0.488
	少样本 +反事实推理	**0.346**	**0.603**	**0.642**	**0.603**	**0.608 0.612**
SI	零样本	0.339	0.339	0.191	0.339	0.232 0.232
	少样本	**0.342**	0.345	0.183	0.345	0.226 0.246
	少样本 +反事实推理	0.333	**0.372**	**0.245**	**0.372**	**0.298 0.325**

表 15：Gemini 2.0 单条被违反规则识别的性能

陈述方法		准确率	平衡准确率	宏平均精确率	宏平均召回率	宏F1 加权 F1
BS	零样本	0.658	0.659	**0.819**	0.659	0.620 0.621
	少样本	0.678	0.678	0.694	0.678	0.618 0.618
	少样本 +反事实推理	**0.704**	**0.708**	0.782	**0.708**	**0.665 0.664**
CF	零样本	0.862	0.862	0.883	0.862	0.864 0.864
	少样本	0.861	0.861	0.890	0.861	0.860 0.860
	少样本 +反事实推理	**0.873**	**0.873**	**0.912**	**0.873**	**0.887 0.887**
SE	零样本	0.360	0.360	**0.667**	0.360	0.232 0.232
	少样本	0.345	0.345	0.518	0.345	0.200 0.200
	少样本 +反事实推理	**0.362**	**0.362**	0.608	**0.362**	**0.240 0.240**
SI	零样本	0.378	0.378	0.213	0.378	0.259 0.259
	少样本	0.381	0.382	0.204	0.382	0.252 0.275
	少样本 +反事实推理	**0.407**	**0.409**	**0.240**	**0.409**	**0.280 0.310**

表 16：单条被违反规则识别的性能（LLaMA-3）

陈述方法		准确率	平衡准确率	宏平均精确率	宏平均召回率	宏F1 加权 F1
BS	零样本	0.540	0.702	0.781	0.702	0.707 0.748
	少样本	**0.617**	**0.763**	**0.804**	**0.763**	**0.768 0.789**
	少样本 +反事实推理	0.475	0.644	0.733	0.644	0.658 0.671
CF	零样本	**0.809**	**0.894**	**0.923**	**0.894**	**0.902 0.904**
	少样本	0.687	0.814	0.842	0.814	0.819 0.821
	少样本 +反事实推理	0.699	0.823	0.861	0.823	0.830 0.833
SE	零样本	0.331	0.497	0.541	0.497	0.503 0.518
	少样本	0.332	0.499	0.547	0.499	0.505 0.520
	少样本 +反事实推理	**0.333**	**0.500**	**0.552**	**0.500**	**0.506 0.522**
SI	零样本	**0.396**	**0.567**	**0.601**	**0.567**	**0.571 0.584**
	少样本	0.391	0.562	0.596	0.562	0.566 0.579
	少样本 +反事实推理	0.377	0.547	0.583	0.547	0.552 0.565

表 17：GPT-4o 在财务报表联合多步诊断上的性能比较

任务	方法	步骤 1 违规检测（是/否） 准确率	精确率	召回率	F1	违规分类（多标签） 精确匹配	精确率$_\mu$	Rec$_\mu$	F1$_\mu$
BS	零样本	77.0	79.4	**96.0**	86.9	1	5.0	49.3	**84.7** 62.3
	少样本	76.0	80.1	93.5	86.3	1	8.0	52.1	78.3 **62.6**
	少样本 +反事实推理	**78.0**	**82.3**	92.7	**87.2**	**24.0**	**59.3**		66.1 62.5
CF	零样本	88.0	90.2	**96.0**	**93.0**	**43.0**		89.4	**72.1 79.8**
	少样本	89.0	**98.2**	88.3	**92.9**	40.0		91.2	64.5 75.5
	少样本 +反事实推理	**91.0**	97.1	91.9	**94.4**	**43.0**		**95.6**	65.8 77.9
SE	零样本	79.0	81.4	**97.5**	88.7	20.0		60.6	**93.2** 73.4
	少样本	**81.0**	**83.2**	96.0	**89.1**	29.0		**66.5**	85.2 **74.7**
	少样本 +反事实推理	**81.0**	81.7	**98.1**	**89.1**	**30.0**		65.6	86.4 74.5
SI	零样本	**85.0**	**90.0**	**91.1**	**90.5**	**33.0**		**69.2**	**65.1 67.1**
	少样本	81.0	87.4	88.1	87.7	26.0		66.5	62.1 64.2
	少样本 +反事实推理	83.0	89.6	90.5	90.1	**33.0**		66.9	67.0 66.9

表18：Gemini2.5 Pro 在财务报表联合多步诊断上的性能比较

任务	方法	步骤 1 违规检测（是/否） 准确率	精确率	Rec	F1	违规分类（多标签） 精确匹配	精确率$_\mu$	召回率$_\mu$	F1$_\mu$
BS	零样本	81.0	80.8	**100.0**	89.4	26.0		63.6	**85.0** 72.7
	少样本	**83.0**	**82.5**	**100.0**	**90.4**	**35.0**		**66.5**	84.4 **74.4**
	少样本 +反事实推理	81.0	80.8	**100.0**	89.4	28.0		59.3	83.8 69.4
CF	零样本	84.0	83.8	**100.0**	91.2	**57.0**		81.5	**90.4 85.7**
	少样本	87.0	89.8	95.2	92.4	49.0		89.9	74.7 81.6
	少样本 +反事实推理	**91.0**	**96.3**	92.8	**94.5**	39.0		**96.3**	62.0 75.5
SE	零样本	86.0	86.8	**97.5**	91.9	55.0		82.2	82.7 82.5
	少样本	**90.0**	**90.8**	**97.5**	**94.0**	**62.0**		**85.2**	**85.2 85.2**
	少样本 +反事实推理	88.0	88.8	**97.5**	92.9	56.0		82.6	82.1 82.4
SI	零样本	**83.0**	**85.1**	94.9	89.7	**37.0**	**65.9**	75.6	**70.4**
	少样本	81.0	83.9	93.6	88.5	36.0	64.2		76.9 70.0
	少样本 +反事实推理	**83.0**	84.3	**96.2**	**89.8**	34.0		61.7	**81.4** 70.2

注：所有指标均以百分比报告。**反事实推理** = 反事实推理。**精确匹配** = 精确匹配（所有违规代码的严格匹配）。精确率$_\mu$/召回率$_\mu$/F1$_\mu$ 表示所有违规类型的微平均指标。

表 19：Gemini 2.0flash 在财务报表联合多步诊断中的性能比较

任务	方法	步骤 1 违规检测（是/否）				违规分类（多标签）		
		准确率	精确率	召回率	F1 分数	精确匹配	精确率$_\mu$	Rec$_\mu$ F1$_\mu$
BS	零样本	78.0	**82.2**	92.5	87.1	21.0	59.8	69.8 63.9
	少样本	**81.0**	82.1	**97.5**	**89.1**	**29.0**	**65.1**	**81.9 72.5**
	少样本 + 反事实推理	77.0	80.6	93.7	86.7	20.0	55.9	70.6 62.4
CF	零样本	90.0	93.9	**93.9**	93.9	40.0	93.8	63.8 75.9
	少样本	**91.0**	**95.1**	**93.9**	**94.5**	**47.0**	94.1	**68.0 79.0**
	少样本 + 反事实推理	85.0	94.7	86.7	90.5	41.0	**94.2**	59.6 73.0
SE	零样本	68.0	79.5	81.5	80.5	26.0	62.4	79.5 68.9
	少样本	**81.0**	**82.2**	**97.5**	82.2	**34.0**	65.9	**93.2 77.2**
	少样本 + 反事实推理	76.0	81.3	91.3	**86.0**	30.0	**69.2**	72.2 70.7
SI	零样本	82.0	**89.4**	87.1	88.3	**25.0**	**69.0**	55.7 61.7
	少样本	**83.0**	85.8	**93.5**	**89.5**	23.0	62.1	67.3 **64.6**
	少样本 + 反事实推理	81.0	83.9	**93.5**	88.5	22.0	55.7	**69.2** 61.7

注： 所有指标均以百分比表示。 **反事实推理** =反事实推理。 **精确匹配** = 精确匹配（严格匹配所有违规代码）。精确率$_\mu$/召回率$_\mu$/F1$_\mu$ 表示所有违规类型的微平均指标。

表 20：LLaMA-3 在财务报表联合多步诊断中的性能比较

任务	方法	步骤 1 违规检测（是/否）				违规分类（多标签）		
		准确率	精确率	Rec	F1 分数	精确匹配	精确率$_\mu$	Rec$_\mu$ F1$_\mu$
BS	零样本	80.0	80.0	**100.0**	88.9	8.0	45.6	84.4 59.2
	少样本	**84.0**	**84.8**	97.5	**90.7**	**16.0**	45.7	**91.9 61.0**
	少样本 + 反事实推理	73.0	82.7	83.8	83.2	14.0	**47.9**	48.8 48.3
CF	零样本	**83.0**	83.0	**100.0**	**90.7**	**53.0**	79.0	**92.8 85.3**
	少样本	81.0	83.5	95.0	88.9	11.0	45.7	89.4 60.5
	少样本 + 反事实推理	82.0	**100.0**	78.3	87.8	28.0	**76.3**	63.9 69.5
SE	零样本	34.0	74.2	28.4	41.1	11.0	58.5	19.1 28.8
	少样本	**55.0**	**92.9**	**48.1**	**63.4**	**20.0**	**62.4**	**35.8 45.5**
	少样本 + 反事实推理	31.0	87.5	17.3	28.9	17.0	56.5	8.0 14.1
SI	零样本	**81.0**	**88.3**	**87.2**	**87.7**	**16.0**	48.2	**69.2 56.8**
	少样本	79.0	87.1	86.2	**87.7**	15.0	49.0	66.0 56.3
	少样本 + 反事实推理	79.0	88.0	84.6	86.3	14.0	**58.2**	52.6 55.2

注： 所有指标均以百分比表示。 **反事实推理** = 反事实推理。 **精确匹配** = 精确匹配（严格匹配所有违规代码）。精确率$_\mu$/召回率$_\mu$/F1$_\mu$ 表示所有违规类型的微平均指标。

表 21：少样本 + 反事实提示下联合规则诊断的错误类型分布（GPT-4o）。

陈述正确		错误分解			
		第一步错误	部分检测	部分幻觉	错误定位
BS	18%	22%	10%	41%	9%
CF	41%	9%	35%	4%	11%
SE	29%	19%	10%	34%	8%
SI	20%	18%	22%	31%	9%

注：结果针对联合规则诊断任务，在少样本 + 因果反事实提示条件下按实例层面计算。

3M Company and Subsidiaries
Consolidated Balance Sheet
At December 31

(Dollars in millions, except per share amount)	2023	2022
Assets		
Current assets		
Cash and cash equivalents	$ 5,933	$ 3,655
Marketable securities — current	53	238
Accounts receivable — net of allowances of $141 and $174	4,750	4,532
Inventories		
Finished goods	2,293	2,497
Work in process	1,424	1,606
Raw materials and supplies	1,105	1,269
Total inventories	4,822	5,372
Prepaids	485	435
Other current assets	336	456
Total current assets	16,379	14,688
Property, plant and equipment	26,870	25,998
Less: Accumulated depreciation	(17,711)	(16,820)
Property, plant and equipment — net	9,159	9,178
Operating lease right of use assets	759	829
Goodwill	12,927	12,790
Intangible assets — net	4,226	4,699
Other assets	7,130	4,271
Total assets	$ 50,580	$ 46,455
Liabilities		
Current liabilities		
Short-term borrowings and current portion of long-term debt	$ 2,947	$ 1,938
Accounts payable	3,245	3,183
Accrued payroll	904	692
Accrued income taxes	365	259
Operating lease liabilities — current	225	261
Other current liabilities	7,611	3,190
Total current liabilities	15,297	9,523
Long-term debt	13,088	14,001
Pension and postretirement benefits	2,471	1,966
Operating lease liabilities	534	580
Other liabilities	14,322	5,615
Total liabilities	45,712	31,685
Commitments and contingencies (Note 18)		
Equity		
3M Company shareholders' equity:		
Common stock par value, $.01 par value; 944,033,056 shares issued	9	9
Shares outstanding - December 31, 2023: **552,581,136**		
Shares outstanding - December 31, 2022: 549,245,105		
Additional paid-in capital	6,956	6,691
Retained earnings	37,479	47,950
Treasury stock, at cost:	(32,859)	(33,255)
Accumulated other comprehensive income (loss)	(6,778)	(6,673)
Total 3M Company shareholders' equity	4,807	14,722
Noncontrolling interest	61	48
Total equity	4,868	14,770
Total liabilities and equity	$ 50,580	$ 46,455

(a) 资产负债表

3M Company and Subsidiaries
Consolidated Statement of Cash Flows
Years ended December 31

(Millions)	2023	2022	2021
Cash Flows from Operating Activities			
Net income (loss) including noncontrolling interest	$ (6,979)	$ 5,791	$ 5,929
Adjustments to reconcile net income (loss) including noncontrolling interest to net cash provided by operating activities			
Depreciation and amortization	1,987	1,831	1,915
Long-lived and indefinite-lived asset impairment expense	—	618	—
Goodwill impairment expense	—	271	—
Company pension and postretirement contributions	(152)	(158)	(180)
Company pension and postretirement expense	145	178	206
Stock-based compensation expense	274	263	274
Gain on business divestitures	(36)	(2,724)	—
Deferred income taxes	(3,855)	(663)	(166)
Changes in assets and liabilities			
Accounts receivable	(170)	(105)	(122)
Inventories	567	(629)	(903)
Accounts payable	138	111	518
Accrued income taxes (current and long-term)	(218)	(47)	(244)
Other — net	14,979	854	227
Net cash provided by (used in) operating activities	6,680	5,591	7,454
Cash Flows from Investing Activities			
Purchases of property, plant and equipment (PP&E)	(1,615)	(1,749)	(1,603)
Proceeds from sale of PP&E and other assets	119	200	51
Purchases of marketable securities and investments	(1,466)	(1,250)	(2,202)
Proceeds from maturities and sale of marketable securities and investments	1,660	1,261	2,406
Proceeds from sale of businesses, net of cash sold	60	13	—
Cash payment from Food Safety business split-off, net of divested cash	—	478	—
Other — net	35	1	31
Net cash provided by (used in) investing activities	(1,207)	(1,046)	(1,317)
Cash Flows from Financing Activities			
Change in short-term debt — net	205	340	(2)
Repayment of debt (maturities greater than 90 days)	(3,086)	(1,179)	(1,144)
Proceeds from debt (maturities greater than 90 days)	2,835	1	1
Purchases of treasury stock	(33)	(1,464)	(2,199)
Proceeds from issuance of treasury stock pursuant to stock option and benefit plans	264	381	639
Dividends paid to shareholders	(3,311)	(3,369)	(3,420)
Other — net	(21)	(60)	(20)
Net cash provided by (used in) financing activities	(3,147)	(5,350)	(6,145)
Effect of exchange rate changes on cash and cash equivalents	(48)	(104)	(62)
Net increase (decrease) in cash and cash equivalents	2,278	(909)	(70)
Cash and cash equivalents at beginning of year	3,655	4,564	4,634
Cash and cash equivalents at end of period	$ 5,933	$ 3,655	$ 4,564

(b) 现金流量表

3M Company and Subsidiaries
Consolidated Statement of Changes in Equity
Years ended December 31

		3M Company Shareholders				
(Dollars in millions, except per share amounts)	Total	Common Stock and Additional Paid-in Capital	Retained Earnings	Treasury Stock	Accumulated Other Comprehensive Income (Loss)	Non-controlling Interest
Balance at December 31, 2020	$12,931	$ 6,171	$43,821	$(29,404)	$ (7,721)	$ 64
Net income	5,929		5,921			8
Other comprehensive income (loss), net of tax	970				971	(1)
Dividends declared ($5.92 per share, Note 8)	(3,420)		(3,420)			
Stock-based compensation	267	267				
Reacquired stock	(2,199)			(2,199)		
Issuances pursuant to stock options and benefit plans	639		(501)	1,140		
Balance at December 31, 2021	15,117	6,438	45,821	(30,463)	(6,750)	71
Net income	5,791		5,777			14
Other comprehensive income (loss), net of tax	69				77	(8)
Dividends declared ($5.96 per share, Note 8)	(3,369)		(3,369)			
Stock-based compensation	262	262				
Reacquired stock	(1,464)			(1,464)		
Dividend to noncontrolling interest	(29)					(29)
Split-off of Food Safety business	(1,988)			(1,988)		
Issuances pursuant to stock options and benefit plans	381		(279)	660		
Balance at December 31, 2022	**14,770**	**6,700**	**47,950**	**(33,255)**	**(6,673)**	**48**
Net income (loss)	**(6,979)**		**(6,995)**			**16**
Other comprehensive income (loss), net of tax	**(105)**				**(105)**	**—**
Dividends declared ($6.00 per share, Note 8)	**(3,311)**		**(3,311)**			
Stock-based compensation	**265**	**265**				
Reacquired stock	**(33)**			**(33)**		
Dividend to noncontrolling interest	**(3)**					**(3)**
Issuances pursuant to stock options and benefit plans	**264**		**(165)**	**429**		
Balance at December 31, 2023	**$ 4,868**	**$ 6,965**	**$37,479**	**$(32,859)**	**$ (6,778)**	**$ 61**

Supplemental share information	2023	2022	2021
Treasury stock			
Beginning balance	**394,787,951**	372,187,578	366,283,418
Reacquired stock	**290,379**	10,865,635	11,834,681
Split-off of Food Safety business	—	15,989,536	—
Issuances pursuant to stock options and benefit plans	**(3,626,410)**	(4,254,798)	(5,930,521)
Ending balance	**391,451,920**	394,787,951	372,187,578

(c) 所有者权益变动表

3M Company and Subsidiaries
Consolidated Statement of Income (Loss)
Years ended December 31

(Millions, except per share amounts)	2023	2022	2021
Net sales	$ 32,681	$ 34,229	$ 35,355
Operating expenses			
Cost of sales	18,477	19,232	18,795
Selling, general and administrative expenses	21,526	9,049	7,197
Research, development and related expenses	1,842	1,862	1,994
Gain on business divestitures	(36)	(2,724)	—
Goodwill impairment expense	—	271	—
Total operating expenses	41,809	27,690	27,986
Operating income (loss)	(9,128)	6,539	7,369
Other expense (income), net	560	147	165
Income (loss) before income taxes	(9,688)	6,392	7,204
Provision (benefit) for income taxes	(2,691)	612	1,285
Income (loss) of consolidated group	(6,997)	5,780	5,919
Income (loss) from unconsolidated subsidiaries, net of taxes	18	11	10
Net income (loss) including noncontrolling interest	(6,979)	5,791	5,929
Less: Net income (loss) attributable to noncontrolling interest	16	14	8
Net income (loss) attributable to 3M	$ (6,995)	$ 5,777	$ 5,921
Weighted average 3M common shares outstanding — basic	553.9	566.0	579.0
Earnings (loss) per share attributable to 3M common shareholders — basic	$ (12.63)	$ 10.21	$ 10.23
Weighted average 3M common shares outstanding — diluted	553.9	567.6	585.3
Earnings (loss) per share attributable to 3M common shareholders — diluted	$ (12.63)	$ 10.18	$ 10.12

(d) 利润表

图7：原始财务报表屏幕截图

表 22：财务报表规则示例

类别	规则编号	规则描述	FASB 参考文献	美国财务会计准则委员会指南 文本
I. 资产负债表规则	BS-01	资产总额 = 负债总额-负债 +所有者合计-权益（包括母公司普通及优先股权益、非-控制性权益，以及任何分类为权益的夹层-zanine项目)。	FASB 概念陈述第6号（CON 6）	权益或净资产是.....中的剩余权益主体扣除其负债后剩余的资产。在企业中，权益是所有者权益。
	BS-02	资产和负债必须 be 分类为 cur-流动和非流动类别别。	210-10-45-5/210-10-05-4	应在分类资产负债表中列示流动负债总额。/ 大多数资产负债表实体会分别列示流动资产和流动负债（通常称为分类资产负债表），以便于确定营运资本。
	BS-03	现金流量表流量必须使用描述性术语，例如“现金”或“现金及现金等价物”。	230-10-45-4	陈述应使用描述性术语，例如现金或现金及现金等价物，而非 am-义术语，例如资金。
	BS-04	Any 分配 of 留存 收益 must 应单独列报在股东权益部分内。	505-10-45-3	允许进行留存收益分配，前提是该分配列示于股东权益部分，并且清楚标明于资产负债表的股东权益部分作为留存收益分配。
	资产负债表-05	库存股应作为扣减项列报从股东权益总额中扣除。	505-10-05-1	权益主题包括以下子主题：a. 总体 b. 股票股利和股票分割 c. 库存股 d. 本款已被《会计准则更新》第2018-07号取代。计数《会计准则更新》第2018-07号。e. 分拆和反向分拆。
II. 利润表规则	SI-01	结果 of 经营活-动必须与.....分开列报非经营活动。	255-10-50-3	主体应披露以下所有信息，适用于最近五个年度中的每一个年度：a. 净销售额及其他经营收入 b. 来源于按现行成本基础计算的持续经营业务 c. .....
	SI-02	本期折旧费用本期 必须计入经营费用中-列入经营费-费用。	360-10-35-4	生产性设施的成本，是其在经济使用年限内提供服务所产生的成本之一。公认会计原则（GAAP）要求将这项成本分摊至预期的设施的使用寿命为.....
	SI-03	每股收益（EPS）基本和稀释股份必须列示于利润表表面陈述。	260-10-45-2	具有简单资本结构的实体，即那些仅有普通股发行在外的实体，应在利润表正面列示持续经营所得的基本每股金额以及净利润的基本每股金额。利润表。.....

类别	规则编号	规则描述	FASB 参考文献	美国财务会计准则委员会指南 文本
	SI-04	在利润表中 ，来自一项主要业务运营之外的交易的重大收益必须列示为一个独立的单列项目，并且该交易的性质应明确披露。	220-10-S99-2	列入杂项 其他收入的重大金额应单独列示于综合收益表或其附注中，并明确说明这些项目所源自的交易的性质。
	SI-05	一项重大事项或交易且在性质上非同寻常，或在发生频率上不常见，发生频率较低必须作为一项来自持续经营的单独组成部分，且不得单独列报，并且不得在利润表正文中列报，扣除所得税后的利润所得税。	220-20-45-1	主体认定为性质异常或属于某种类型的重大事项或交易，该类型表明其发生频率低，或兼具两者特征，应作为所得中的单独组成部分予以报告持续经营.....
III. 现金流量表规则	CF-01	经营活动产生的现金流量净额-营活动 +投资-融资活动 + 融资活动 =净变动现金及现金等价物中的价物（注：考虑汇率变动的影响）。	230-10-45-24	某一期间的现金流量表应报告经营、投资活动提供或使用的现金净额，以及融资活动和这些现金流对现金、现金等价物总额的净影响，.....
	现金流量-02	财务报表不得列报每股现金流量金额。	230-10-45-3	财务报表不得列报每股现金流量金额。现金流量或任何 com-其中一部分是净利润的替代指标，可作为主体业绩的指标，因为报告每股金额可能意味着.....
	CF-03	清晰展示以下项目的变动现金、现金等价物以及现金流量表中的受限制现金，使用准确的术语（而非“资金”）。	230-10-45-4	现金流量表应说明以下项目总额的变动在本期内，现金、现金等价物——额，以及通常被描述为受限制的现金或受限现金等价物。该陈述应使用.....
IV. 权益变动表规则	SE-01	权益变动表必须遵循某些规则，例如以下等式-关于构成的-总权益的构成：总-权益（年度） = Com-普通股和额外实收资本 + 留存收益 +累计其他综合收入 + 非控股权益 + 库存股。	505-10-50-2	如果同时列报财务状况和经营成果，应披露构成股东权益的单独账户（除留存收益外）的变动，以及留存收益的变动，并披露至少期间内权益证券股份数量的变动最近的年度会计期间以及任何后续中期所列信息，是使财务报表具有充分信息性的必要条件。披露此类变更的方式可以是单独编制报表，也可以在基本财务报表或其附注中予以披露。

类别	规则编号	规则描述	FASB 参考文献	美国财务会计准则委员会指南 文本
	SE-02	本期净利润（或亏损）必须列示为留存收益的增加（或扣除）转入留存收益）。	215-10-50-1	如果同时列报财务状况和经营成果则应披露构成单独账户的股东权益（除留存收益外）的变动，以及权益证券股份数量在至少最近的年度会计期间以及任何随后列报的中期均须足够使财务报表提供充分信息。披露此类变更可采用单独报表的形式，也可列入基本财务报表或其附注中。
	SE-03	其他综合收益（OCI）项目必须列示为对累计其他综合收益（AOCI）的变动。	220-10-45-14	某一会计期间的其他综合收益总额应转入权益的一个组成部分该组成部分在列报时应与留存收益以及财务状况陈述中的额外实收资本分别列示，该财务状况陈述列报于会计期间结束时。A诸如累计其他综合收益等描述性标题应用于该权益组成部分。

G 因果反事实提示详情

本附录提供了在 FinRule-Bench 中进行因果反事实提示时使用的准确提示模板和代表性截图。这些材料旨在确保可复现性和透明度。正文在第3节中概述了提示协议。

G.1 提示模板概述

因果-反事实提示通过上下文示例实现，这些示例明确指出：(i) 规则被违反的因果来源，以及 (ii) 最小限度的反事实修改，使其能够恢复合规。以下模板展示了各项任务所使用的确切提示结构。每个示例都包含一张财务表、一个目标决策、一段指出被违反原则的因果解释，以及一项能够改变验证器判定结果的最小反事实编辑。所有模型和任务使用的示例均保持固定，以确保可比性。

G.2 代表性提示

用于**规则验证**任务（单条违反规则识别）的代表性因果-反事实提示。

系统指令： 你是一名严格的财务审计师，负责评估以下会计规则。
规则定义（BS-01）： 资产总额必须等于（负债总额 + 总权益 + 夹层权益）。
示例：
- **合规 →T**
  资产：$42,448$ = 负债：$35,451$ + 权益：$6,975$ + 夹层资本：22
  原因：会计恒等式完全成立。
  反事实：如果资产变为42,449，该恒等式就会被破坏。
- **不合规 →F**
  资产：$42,500 \neq$ 负债：$35,451$ + 权益：$6,975$ + 夹层资本：22
  原因： 资产总额不等于负债、权益和夹层资本之和。
  反事实： 如果资产改为 42,448，该恒等式便会成立。

评估任务： 评估以下资产负债表：
资产负债表数据： <在此插入 Markdown 表格>
合规请回答"T"，不合规请回答"F"。
答案：

用于**规则识别**任务（单条违反规则识别）的代表性因果-反事实提示。

系统指令：你是一名严谨的财务审计师。**任务：**判断资产负债表是否违反任何会计规则。如存在违规，请报告所有被违反的规则。**规则优先级：**先检查 BS01，再检查 BS02 → BS03 → BS04 → BS05。**规则：**

- **BS01：** 资产总额必须等于（负债总额 + 总权益）。
- **BS02：** 所有资产项目必须归类为"流动资产"或"非流动资产"。_ _ _
- **BS03：** 现金必须使用精确术语"现金及现金等价物"。
- **BS04：** 必须将留存收益标记为"留存收益"。
- **BS05：** 库存股（如有）必须列入权益，并以负值表示。

参考示例：

- **不合规 →是：[BS03、BS04]**
  数据上下文： 资产 |流动_资产 |资金及资金等价物 | 5000权益 | 股东_权益 | 累计利润 | 10000 原因： - 由于使用了非公认会计原则现金术语（"资金....."），BS03 违规。- 由于使用了"累计利润"而非"留存收益"（非标准标签），BS04 违规。反事实： - 将"资金"重命名为"现金及现金等价物"即可解决 BS03。- 将"累计利润"重命名为"留存收益"即可解决 BS04。

- **合规 →否**
  条件： BS01： 18000 = 8000 + 10000 （恒等式成立）。
  BS02：流动_资产/非_流动_资产均存在。
  BS05：不存在库存股。

输出格式： YES: [规则1、规则2.....]
NO

多规则诊断任务的代表性因果-反事实提示。

系统指令： 你是一名严格的财务审计师。**任务：** 判断资产负债表是否违反任何会计规则。如果存在违规，请报告所有被违反规则。**规则优先级：** 先检查 BS01，然后检查 BS02 → BS03 → BS04 → BS05。**规则：**

- **BS01:** 资产总额必须等于（负债总额 + 总权益）。
- **BS02:** 所有资产项目都必须归类为"流动资产"或"非流动资产"。_ _ _
- **BS03:** 现金必须使用准确术语"现金及现金等价物"。
- **BS04:** 留存收益必须标注为"留存收益"。
- **BS05:** 库存股（如有）必须列入权益，且以负值列示。**参考示例：**

- **情景1：发现违规**输入数据片段：资产 | 流动_资产 | 资金及资金等价物 | 5000权益 | 股东_权益 | 累计利润 | 10000输出： 是： [BS03、BS04]原因： - 由于使用了非公认会计原则现金术语，违反了 BS03。- 使用"累计利润"而不是"留存收益"（非标准标签）。反事实： - 将"资金"重命名为"现金及现金等价物"即可解决 BS03。- 将"累计利润"重命名为"留存收益"即可解决 BS04。

- **场景 2：合规**条件：BS01： 18000 = 8000 + 10000 （恒等式成立）。BS02：存在流动_资产/非_流动_资产。BS05：无库存股。输出：否**输出格式：**是： [规则1、规则2.....]否