会计判断中的人机协作：认知编码方案与基于规则的自适应协作协议

Yulu Huang[1,2] Lijun Xiao[1] Gang Wang[2,†]

[1]江西财经大学（JUFE），中国南昌[2]字节跳动

[†]通讯作者：Gang Wang， wanggang@bytedance.com

第一作者邮箱： yuluhuang324@gmail.com

预印本，编制于2026年11月22日

arXiv:2611.29442v1 [cs.CE] 22 Nov 2026

摘要

大语言模型带来了会计信息系统设计问题：应如何在模型与会计师之间分配证据生成、专业评估和责任？我们提出会计判断认知分解（CDAJ）编码方案，将四个流程阶段与信息密集型需求和评价型需求相交叉。自适应协作协议（ACP）将分配规则与独立人工估计、关联来源的模型输出、结构化核验及有记录的人工结论相衔接。两份分别提交的编码表涵盖20个构造案例。公开说明的仅采用主要答复的分析保留了尚未解决的辅助答复格式问题。单位划分的原始一致率为19/20，阶段为20/20，需求为18/20。在14个双方均判定为单一且可编码的案例中，阶段的名义 Cohen's κ 为1.000，I/E/M 需求概况为0.752；20个案例中仅有12个经双方共同判定为 I 或 E。公开披露诊断检索了三家银行年度报告的1,406页，以查找18张数值参考卡。在查询信息完全相同的情况下，向 BM25 添加数值／期间／单位共现门槛，并未提升指定来源找回率（排名第一或前三）：两种方法在排名第一时均找回13/18个来源，在排名前三时均找回18/18个来源。事后审计区分了页面找回与所请求表格单元格的核验。合成检验考察事件排序和记录完整性；高斯规划实验则考察未来有效性评估的规划。另有一项来源答复审计记录显示，36份返回记录中有16份状态不利或不明确。这些范围有限的发现涉及编码可重复性和实施特性。它们并不能证明存在最优分配、经验证的分类体系或专业判断能力的提升。

关键词： 人机协作；会计判断；会计信息系统；自动化偏见；认知强制；系统设计。

1. 引言

专业会计判断结合了证据获取、分析工作和在不确定性与问责要求下进行的评估（Solomon 和 Shields，1995；Nelson，2009）。大语言模型可以协助完成这项工作中的部分任务，但未必能提升最终的专业结论。正确的信息提取、有说服力的备忘录以及依据充分的判断，是系统产出的不同形式。因此，会计信息系统（AIS）的设计问题在于，如何组织产出证据的技术与负责审查并承担决策责任的专业人士之间的互动。

既有会计信息系统研究将技术采纳与组织、监管和工作实践之间的互动联系起来（Austin 等人，2021年；Krieger 等人，2021年）。关于人工智能在审计中应用的研究也指出了伦理和问责方面的关切（Munoko 等人，2020年）。Stratopoulos and Wang（2025年）提出的最新框架，将人工智能置于更广泛的会计研究背景中。我们的关注点更为具体：专业人士评估一项特定的人工智能辅助会计判断时所使用的界面和记录结构。专业人士接受人工智能建议未必意味着过度依赖。会计实验记录了对人工智能来源建议的排斥，而用户控制的影响取决于参与者的特征（Commerford 等人，2022年；2024年）。因此，一概接受或一概抵制，都不是合理的设计目标。

我们提出会计判断认知分解（CDAJ）编码方案和自适应协作协议（ACP）。CDAJ将工作的四个阶段与两种主要认知需求交叉组合。ACP将由此得出的分配规则转化为可供检查的一系列步骤：记录独立估计，展示与来源关联的人工智能输出，核验证据和推理，并记录人工结论。此处的“自适应”是指，分配方式以及迭代式质疑是否可用，会随任务需求而变化。这并不意味着已实施的系统能够学习最优策略，也不意味着实时自适应策略已接受评估。

该提案保留了规范性约束与绩效预测之间的区别。保留一名负责任的人类签署人是一项设计约束；独立估计或检查点是否能够改善判断，则是一个实证问题。这一区别可以防止将专业问责误认为人类主导的配置始终更为准确的证据。

1.1 研究问题与贡献

本文提出三个问题。RQ1：流程阶段和认知需求如何组织针对会计的分配规则？RQ2：如何将这些规则落实为可检查的界面要求和决策记录？RQ3：哪些证据能够区分有用的协议增量与不必要的负担？我们从概念层面回答前两个问题，并为第三个问题提供公开来源歧义诊断、计算检查和人工评估议程。

这项研究的贡献分为三部分。第一，CDAJ 将一个包含八个单元的分配矩阵与 25 项示例性会计任务相联系，指出哪些信息生产工作可以委派，同时明确保留评估和签核职责。第二，ACP 将每项保障措施与可观察的界面状态和记录字段相连接。其会计专属内容在于贯穿整个工作底稿链，保留来源溯源、受质疑的假设、可接受的判断范围和责任归属，而非仅仅为人工智能建议添加解释。第三，一项已执行的公开披露歧义诊断揭示了定位页面与核实会计主张之间的差异；可复现计算和合成检验则用于考察其他运行特性。另有一项练习使用两种返回的响应形式，考察了 20 个构造案例的初始编码可重复性

，并披露了有限的执行情况，同时辅助响应格式仍未解决。计算检查和编码一致性均不能证明会计师在 ACP 下的推理能力有所提升。

本文依据设计科学原则（Hevner 等人，2004），提出一套会计信息系统编码方案以及记录与工作流规范，并进行了范围有限的技术评估和编码练习。本文并非对该制品在专业实践中的效用所作的完整评估，也不是注册报告或预注册的人类研究。形式化状态与记录规范是可供查验的设计制品。三个案例仅用于展示设计；完整的面向参与者的证据文件和经裁定的评分标准仍有待后续工作。本文明确保留了评估议程，因为可审计流程仍可能导致糟糕的判断，其成本也可能超过收益。

2. 理论基础

2.1 会计判断与任务结构

会计判断研究表明，必须综合考量任务、证据、决策者和复核环境（Libby，1976；Bonner，1999；Solomon 和 Shields，1995）。职业怀疑既包括怀疑性判断，也包括相应行动，因此，记录下疑虑并不等同于采取了具有实质影响的核查步骤（Nelson，2009）。早期审计研究也考察了概率推断中的锚定与调整（Joyce 和 Biddle，1981）。这些发现为记录初始立场及后续证据使用情况的工作流提供了依据，但并不能证明预先承诺总是有益：人类最初的错误本身也可能成为锚点。

CDAJ 将信息密集型工作与评价性工作区分为一种主导需求编码，而不是将其视为彼此排斥的心理能力。提取契约条款可能是机械上可核查的；但要判断一项不确定的再融资计划是否会改变持续经营结论，则需要结合上下文进行评价。一项完整的业务委托同时包含这两类工作。因此，会计领域、流程阶段和认知需求应分别进行编码。预期信用损失案例并不会仅仅因为包含数值计算，就本质上属于信息密集型工作。

2.2 分配、互补性与依赖

自动化等级和混合主动交互为在人类与系统之间分配职能提供了一般基础（Parasuraman 等人，2000；Horvitz，1999）。信任应根据系统能力和上下文进行校准（Lee 和 See，2004）；依赖本身并不是质量衡量标准。人类–人工智能互补性要求考察二者组成的整体系统，而不能仅根据任一智能体单独推断团队绩效（Hemmer 等人，2025）。会计领域之外的实地证据发现，人工智能辅助对知识型任务的影响并不一致（Dell’Acqua 等人，2026），这进一步凸显了任务依存型评估的必要性，但并不能证明会计师也会产生相同的影响。

会计证据增加了一种重要的反向作用机制。Commerford 等人（2022年）研究了审计师在复杂估计中对人工智能的依赖，而 Commerford 等人（2024年）考察了提供输入如何影响依赖，其中控制效应取决于外控点。这些研究使算法厌恶成为自动化偏见的一个可信替代解释。检查点可以纠正过度顺从，但也可能促使人们不必要地拒绝合理建议。相关结果是由此产生的判断的质量与可辩护性

，而依赖行为应据此加以解释。

2.3 核验与认知强制

自动化偏见指不恰当地依赖自动化辅助（Parasuraman 和 Manzey，2010年；Goddard 等人（2012年））。解释可能改变依赖程度，但未必能提升人机团队的表现（Bansal 等人，2021年）。Buçinca 等人（2021年）探讨了认知强制作为减少过度依赖的一种方法及其代价。ACP 借鉴了这种设计思路，但并不假定其效果在专业会计领域也能原样实现。强制核验可能占用宝贵时间，流于形式，或强化用户对最初错误看法的执着。

专业审查既要求提供可核查证据，也要求理由清晰易懂。流畅的理由可能包含缺乏依据的主张；词元似然并不代表会计结论正确的概率。因此，ACP 将来源核验、不确定性披露与任何数值校准主张区分开来。国际审计与鉴证准则理事会关于技术过度依赖的支持材料提供了实践背景（国际审计与鉴证准则理事会，2021年）；但该材料并未证明本协议符合审计准则。

2.4 从一般原则到会计设计

所提出的扩展方案是一条可追溯链条，将四项设计要素连接起来：按阶段和需求进行分配；独立的人类证据评估；主张层面的来源比较；以及解释所采纳和拒绝的建议的签署的理由说明。一般性的人类—人工智能原则已经为人类控制和解释提供了动因（Amershi 等，2019年）。CDAJ/ACP明确了这些原则如何与区间值会计估计值、相互竞争的基于准则的解释以及业务审查相互作用。该清单和案例展示了这种映射关系，而非证明编码方案的完备性或优越性。其他分解方式以及现有的更强工作流程同样可能达到或超过其效果；评估必须保留这种可能性。

3. 会计判断的认知编码方案

3.1 动机：从人工智能能力到任务分配

在会计领域采用大语言模型部署的能力优先方法会提出这样的问题：“大语言模型能做好什么？”并据此分配任务。这种能力优先视角并不完整，因为它忽视了第二层问题：人类应继续做什么？能力基准测试可以回答大语言模型是否能够起草备忘录或估计备抵，但无法回答：允许其无条件执行这些任务，是否会改善事务所的审计质量、监管立场或利益相关者信任。

本节提出一种互补视角：一种以**判断优先**为导向的分解方法，沿流程阶段和认知需求两个维度对会计决策任务进行分类，并对人类与人工智能代理之间的任务分配作出有原则的预测。

3.2 流程阶段分解

借鉴 Solomon 和 Shields（1995）以及 Parasuraman 等人（2000）的研究，我们将典型的会计判断分解为四个连续的流程阶段：

1. **数据获取** ——从业务委托文件、监管数据库、以往工作底稿和外部来源中定位、提取并规范化证据。

2. 分析综合 ——计算比率和趋势，与基准进行比较，进行统计异常检测，并构建连贯的证据包。

3. 评价性判断 — 在不确定性和重要性条件下形成专业意见（例如，持续经营假设是适当的；备抵是充足的；第三级估值是合理的）。

4. 专业沟通 — 将判断记录在备忘录、脚注或审计报告中；向利益相关者解释结论的依据；并对该意见承担个人问责。

这四个阶段在所需认知工作的类型、输出的可验证性以及与结果相关的问责方面各不相同。这促使我们沿着认知需求这一维度进行第二次分解。

3.3 认知需求分解

我们沿单一维度对每个阶段进行分类：其工作主要是信息密集型（需要回忆或处理大量结构化证据）还是评价性 （需要权衡相互竞争的考量、运用职业怀疑，并在存在不确定性时对判断承担问责）。

信息密集型工作重视一致性、完备性和计算准确性。评价性工作则重视对上下文的理解、伦理推理，以及质疑先例的意愿。

大语言模型可以协助检索、提取和起草，但必须依据针对具体任务的参考资料评估其表现。确定性计算应使用经过验证的计算工具执行，而不应依赖流畅的文本生成。评估任务还涉及尚未解决的诠释问题、职业怀疑和责任。这些特性说明有必要采用条件分配规则，但并不能证明任何一方都普遍具备能力优势。

3.4 分配矩阵

将四个流程阶段与两种认知需求交叉组合，便得到表1中的 4×2 矩阵，该矩阵将每个单元格对应到一个有待评估的建议分配假设。

表1. CDAJ 分配矩阵：按流程阶段和认知需求划分的建议分配假设 .

流程阶段	信息密集型任务	评估任务
数据获取	大语言模型协助提取；人类检查覆盖率	人类主导（策展重-需要上下文）
分析性综合	大语言模型主导，人类进行核验	迭代式：大语言模型生成备选方案；由人类选择并为其辩护
评价性判断——ment	迭代式：大语言模型生成结构化分析；人类在怀疑态度下发表意见——cism	人类主导；大语言模型作为对抗性审阅者
专业沟通	大语言模型主导起草；人类主导语气和签核	由人类主导；大语言模型协助进行格式编排和引文处tion

注：“由大语言模型主导”指由人工智能代理主要负责生成内容，并由人类审核；“迭代式”指双方轮流贡献，并设有明确的检查点；“由人类主导”指由人类撰写，并可按需获得大语言模型的协助。这些是假设性的分配方案，尚未经过测试，并非默认的操作方式，也不代表授权委派。编码一致性并不能证明这些方案在质量、效率或安全性方面孰优孰劣。

该矩阵得出了若干不显而易见的预测。首先，即使在大语言模型协助信息提取的阶段（数据获取）中，人类仍承担核验职责：大语言模型的检索覆盖率和事实准确性尚不确定，而证据收集中的遗漏会影响整个下游分析。其次，对于归类为评价性的任务，该矩阵明确排除了在评价性判断和专业沟通阶段采用“仅大语言模型”的部署方式：正是在这些阶段，专业问责责任得以落实。第三，该矩阵指出，分析综合阶段是最适合协作的环节：信息密集型要素与评价性要素在此并存，而协议必须同时支持这两种模式。

3.5 编码规则与边界条件

CDAJ 是一种拟议的编码启发式，而非经验证的分类体系。它对工作单元进行分类，工作单元定义为一项具有指定输入、独立输出和验收标准的行动。它不根据职称对整个业务委托进行分类。编码流程分为四个步骤。

1. 每当输出或验收标准发生变化时，就对工作进行分段。提取贷款余额与选择叠加调整属于不同的工作单元，即使两项工作由同一人完成也是如此。

2. 根据单位的直接产出确定阶段：已定位或规范化的来源属于数据获取；推导出的比较或情景属于分析综合；选定的结论属于评价性判断；为审查或披露而呈现该结论，则属于专业沟通。

3. 当适用标准和假设已经确定，且无需解决存在争议的解释即可通过追踪核查或重新计算来检查输出结果时，将需求标记为信息密集型。

当接受结果需要在合理的假设之间进行选择、评估来源可信度或权衡相互竞争的解释时，将其标记为评价性。阶段名称指向输出结果；需求标签则指向决定是否接受结果的方式。因此，应用一个已经获批准的契约条款阈值属于判断阶段的信息密集型单位，而选择哪个管理层假设可信，即使发生在数据获取期间，也属于评价性需求。

4. 当两种需求仍然都具有实质重要性时，若该单位的输出可以分离，则拆分该单位；否则保留两种候选编码，将未解决的选择提交给由人类主导的检查点，并记录这种模糊性。不要仅仅为了填入矩阵而强行采用多数意见标签。

我们依据这些规则，对20个构造的案例描述分别收集了两位编码者提交的结果。第8节区分了可用的主要编码分析与尚未解决的辅助响应格式。保留独立标签；计算一致意见时不使用协商后的标签。附录A提供了25个由作者选取的工作单元示例及其拟定的主导编码，这些示例并非经过验证的总体，也不构成可靠性估计。以下分解使各主要领域中的编码过程均可检查。

表2. 依据明确编码规则进行的工作单元分解示例。

Case	阶段	信息密集型单位	评价性单位
持续经营-cern	收购	提取注明日期的契约条款	评估融资证据的可信度-据
	综合	重新计算指定的覆盖率	选择合理且可辩护的预测压力假设
	判断	应用 经批准的阈值于已检查的输入	在结论中权衡相反证据论
	沟通	汇总来源交叉引用	为结论及其注意事项提供依据
《国际财务报告准则第9号》预期信用损失	收购	提取贷款余额和日期	评估前瞻性内容的完备性-查看证据
	综合	执行已批准的情景计算—tion	选择情景权重或叠加调整假—设
	判断	将估计值与已批准的约束条件	判断定性调整是否合理
	沟通	核对披露数据	解释重大判断和不-确定性
第三级价值	收购	提取交易条款	评估候选对象的可比性交易
	综合	执行指定的贴现现金流计算	选择有充分依据的不可观察输入值puts
	判断	将值与经批准的调整项进行比较可接受范围	从备选项中选择一个值并为其辩护项
	沟通	汇总输入数据和敏感性分析表ules	解释估值不确定性和局限引文

该矩阵为在这些编码单元上进行测试提供候选分配。证据获取和经独立核查的组件能力被提出作为必要条件，但并不构成该分配有益的证据。它既不能替代专业标准，也不能证明大语言模型在信息密集型单元中更具优势。

人类能力、模型质量、时间压力、错误后果以及现有审查控制可能使另一种分配更为合理。该设计的独特承诺在于，改变生成主导方不能抹去从来源到判断的记录，也不能将最终责任转移给生成者。

3.6 边界案例详解与委派关卡

三个边界案例阐明了编码规则所规定的内容。这些是经过推演的概念示例，并非来自编码信度研究的观察结果。

信用损失计算与叠加调整选择。 假设输入是一个包含经批准的阶段、情景、权重和模型参数的贷款组合。按照固定规范执行并生成准备金明细表，属于分析综合／信息密集型；验收时需根据经批准的输入进行独立重新计算。如果请求的输出还包括因管理层对宏观经济预测存在异议而选择叠加调整，那么这就是一个独立的评价性工作单元。验收时需说明为何选择某项假设而非另一项。应将明细表的执行与叠加调整选择拆分，并关联相关记录。如果两者的影响无法区分，则保留混合编码，并禁止由大语言模型主导完成。

融资证据与持续经营结论。 在来源身份已确认的情况下，提取已签署融资协议的日期、金额和条件，属于数据获取／信息密集型。判断管理层电子邮件是否构成有关即将进行的再融资的可信证据，属于数据获取／评价性：直接产出是经认可的证据，而非最终意见。将这些证据与现金流短缺一并权衡以得出结论，则属于评价性判断／评价性。这一区分可避免用一个“持续经营”标签混淆检索、可信度评估和专业结论。

估值计算与输入可接受性。在固定折现率和现金流的条件下执行现金流折现计算，属于分析综合／信息密集型工作。判断一项交易是否具有可比性，或判断某个不可观察输入值是否具有可辩护性，则属于评价性工作。一旦授权审查确定了可接受范围，根据基于输出的阶段规则，检查计算出的值是否落在该范围内，属于评价性判断／信息密集型工作。选择或论证该范围则不属于此类：它仍然是评价性的。因此，要求模型同时设定范围并认证自身计算值的工作单元，不符合固定验收标准条件，必须拆分或保留由人工主导。通过识别那些在上述边界下会产生不相容分类的工作单元，可以对该方案提出质疑。

矩阵中由大语言模型主导的单元格仅提议待评估的分配，并不授予委派权限。在评估此类分配之前，我们建议记录五项必要前提条件：(i) 验收标准和重要假设已确定；(ii) 已具备授权来源访问权限和重要证据清单；(iii) 针对特定组件和版本的本地能力检查通过；(iv) 重要来源链接和计算过程可以独立核查；以及 (v) 有负责审查人能够解决例外情况。未知的门槛应视为未满足，而非视为获得许可。在评价性或尚未解决的混合工作单元中，无论模型的语言流畅度如何，仍然必须由人工主导解决。

对于确定性算术，门槛（iii）要求在输入、单位和声明的数值容差相同的情况下，与独立重新计算的结果一致；若存在差异，则不得由模型主导完成。对于检索，负责人必须在测试前，针对相关的本地验证集，界定重要参考项以及可接受的覆盖率/错误率阈值。每一项指定的关键遗漏都必须得到解决；通过有限

样本集并不能证明总体召回率。本文未提出通用数值阈值，也未主张任何特定大语言模型已通过这些门槛。若缺少阈值或相关验证集，则仍不批准委派。能力／版本检查未通过时，应将工作转交给经审核的确定性工具，或采用人工主导的处理；重要证据缺失时，应升级处理或暂不下结论，而不能允许任何一方凭空编造。门槛测试结果、受测版本、例外情况以及所选分配方式的理由，都应记录在判断记录中。满足这些建议的前置条件，并不足以证明委派是安全或有益的：比较评估以及适用的专业授权仍是独立要求。

3.7 AIS设计的重要性

拟议的CDAJ方案为设计人工智能辅助会计信息系统提出了三项设计原则。第一，按流程阶段设计，而非按职位头衔设计：人类与人工智能之间的工作分配，应针对每项重复性任务的各个阶段逐一明确，而不是笼统地按“审计师”或“分析师”的身份来确定。第二，保留问责边界：根据拟议的问责约束，评估性判断和专业人员签字确认仍由人类主导，因为问责不能委托给不承担专业责任的系统。第三，以协作为导向，而非以替代为目标：成功的衡量标准是团队产出的质量，而不是由人工智能完成的工作比例。

4. 自适应协作协议

4.1 设计目标

第 3 节的 CDAJ 分配矩阵规定了谁应做什么，但没有说明协作应如何进行。本节将该矩阵转化为**自适应协作协议**（ACP），这是一套可验证的工作流程，其中嵌入了三项旨在防止不当依赖的保障措施，其准确性和时间成本仍有待评估。

ACP 有三个设计目标：

1. 通过确保每项记录的结论均由具备资质的人类负责，并且该人员能够在专业审查中为其立场辩护，**维护专业问责**。2. 通过**抵御自动化偏见**，在界面中加入旨在阻止过早认同 AI 建议的功能。3. 通过要求每项由人工智能提出的重要主张都附有来源引用和明确的不确定性披露，**支持可验证性**。

4.2 协议架构

ACP 为重要判断工作单元规定了四阶段工作流程。第 7 节中的事件测试框架实现了选定的事件顺序检查，这些检查与记录验证器及公开来源实验相互独立；它们均未实现面向参与者的平台。

4.2.1 阶段A：人类独立估计

在查看任何大语言模型输出之前，人类专业人士会记录对结论的独立估计值、所考量的关键支持性证据，以及置信度（0–100 percent）。这样做旨在保留可观察的建议前立场，并鼓励独立考量证据（Buçinca 等，2021；Joyce 和 Biddle，1981）。它究竟能减少有害依赖，还是反而强化错误的人类锚定，仍是一个开放的行为问题。初始立场并非真实值。

拟议的界面会在记录独立估计之前隐藏大语言模型面板；视觉标记用于指示流程状态（“估计待记录”与“已记录估计”）。

4.2.2 阶段B：来源关联输出与不确定性披露

模型输出包含拟议的结论、逐项主张的理由、证据文件中的来源位置，以及一个不确定性面板，用于指出缺失信息、未解决的假设，以及可能改变建议的条件。人类必须能够打开引用来源，并核查其是否支持该主张。模型提供的参考资料本身并非经核实的证据；失效或不支持主张的参考资料会被标记以供审查。

数值置信度是可选的；只有在预测目标经过独立评估、具备留出参考案例、校准程序以及校准性能的不确定性估计时，才可将其标记为已校准。模型生成的自信度和词元概率并不满足这一要求。因此，示例工作流程和拟议的首次人类评估采用定性不确定性披露，而非未经验证的百分比。日后可基于规模足够且相关的案例集，逐步积累校准证据；三个固定情境输出无法提供此类证据。

4.2.3 阶段C：结构化验证检查点

人类专业人士根据三个验证问题审查大语言模型的输出，并以自由文本回答和李克特量表评分的形式记录：

问题1. 一致意见：您是否认同大语言模型的建议？（李克特量表： 1 = 强烈不同意， 7 =强烈同意）

问题2. 证据补充：您查阅了哪些大语言模型未查阅的证据？反之亦然？

问题3。推理分歧：如果有的话，你的推理在哪些方面偏离了大语言模型陈述的理由？

这些问题旨在引出一致意见、遗漏的证据以及有理有据的异议。这些是拟议的界面诊断，并非经过验证的自动化偏见机制分类体系。

4.2.4 阶段D：记录的结论与审计追踪

最终输出是一份结构化记录，包含：(i) 人类在阶段A给出的独立估计；(ii) 阶段B提供的关联来源的模型输出和不确定性披露；(iii) 阶段C的核验回应；以及 (iv) 一项可能与大语言模型的建议一致、也可能不一致的最终结论，并由具备资质的人类专业人士签署。

该记录有助于审查哪些证据和协助影响了结论。但该记录本身并不能证明符合任何审计准则、审计证据充分，或记录内容真实

可信。访问控制、保留规则、签名、来源授权和防篡改存储需要单独实现并进行保证。初始答案、模型答案和最终答案之间的差异体现了修订和依赖；若没有独立质量标准，便无法确定其价值。

4.3 正式事件与记录规范

ACP由两个可分离的构件组成：事件顺序接受器和判断记录模式。事件接受器用于确定是否遵循了规定的顺序；记录模式用于确定是否包含所需的对象和参考资料。两者都不负责判断证据是否真实，也不负责判断结论在专业上是否可靠。这种区分使系统设计者能够明确规定失败条件，而不将程序完整性等同于质量。

表格 3. 单个工作单元的参考事件状态规范。

当前状态	事件	守卫条件	下一状态
开始	独立事件	初始置信度字段在范围内；第一项	A：初始记录
A	有效载荷	仅限 C2–C4；版本固定及冻结的内容哈希匹配	B：建议已呈现
B	核验	仅限 C3–C4；一致意见在范围内，且必需文本已提供	V：已记录检查点
V	质疑	C4 仅限；no 待处理挑战；非空文本	问：响应待处理
Q	响应	匹配挑战 ID；版本和响应内容哈希有效	V：已记录检查点
A / B / V	最终	来自 A 的 C1；来自 B 的 C2；来自 V；无待处理交换	D：已定稿
D	任何事件	不允许进一步的状态转换	拒绝轨迹

所有未列出的状态转换、未知事件以及缺失的定稿均会被拒绝。已实现的事件检查器将 Q 编码为待响应标志，同时保留 V；该表格是等效的显式状态表示。哈希值用于检查其与预期载荷的一致性，而非验证其来源的真实性。参考实现假设时间戳严格排序；即使时钟时间相同，部署时也需要唯一的事件序列。拒绝意味着该轨迹不符合规范。运维修复、异常升级、访问控制以及对已拒绝事件的保存，都需要通过平台实现，事件检查器本身并未实现这些功能。

表4. 核心判断记录对象与会计专属完整性规则。

对象	必需内容	可检查的关系或失效
上下文	记录标识符、框架、计量日期，用于测试样例的合成标记	不得脱离其明确规定的框架/日期
证据登记表	来源 ID、版本、内容哈希、来源位置 tor	每个引用的来源 ID 均可解析；字节发生变化时存储的哈希值失效
主张和假设（续）	主张 ID、陈述、来源/假设 ID，声明的支持状态	结论的输入仍可追溯；未——支持／有争议的主张需要例外——tion
初始估计	值、引用的证据、理由、置信度	建议前立场被单独保留——独立于模型建议
核验	一致意见、新增证据、推理分歧、例外清单	每项重大例外都指出了受到质疑的主张
交流记录	交流记录 ID、质疑与回应文本	可选的迭代对话会予以保留，而不是取代初始建议
最终记录	值、结论、理由、例外处置	每个列出的例外均恰好获得一个在定稿前记录的处置
	处置、签核	

该模式将主张的声明的 支持状态与经核实的真实性区分开来。对于每项存在争议或缺乏支持的主张，核验时必须记录一项例外。定稿时必须给出有理据的处置方式：已处理、有理由地接受，或暂缓决策。有理由地接受并不会使薄弱证据变成充分证据；最终审阅人仍须对处置负责。该实现会检查这些字段及其参考资料是否存在，但不会判断说明是否充分，也不会核验签署人的真实性。在生产环境中，必须通过授权身份和受保护的存储，将记录的签核与具体人员绑定；测试实现使用的是虚构的签署人 ID。

扩展记录契约使会计背景成为类型化信息，而非单一的描述性标签。它要求提供框架、会计认定、计量日期、明确的三字母货币代码、单位和指引引用。指引引用要么标识权威机构和段落，要么记录适用性尚未解决及其理由。可接受范围对象记录已确立、未确立或不适用状态，并注明边界依据；只有已确立的范围才具有有限且有序的边界。因此，专业依据缺失会被表示为知识缺失，而不是填入虚构的准则引文或数值范围。

对于表格类主张，扩展契约要求提供请求的参考值绑定和类型化的来源绑定：来源 ID、文档/版本、从 1 开始计数的页码、表格、行、列、会计计量项目、期间、货币、单位、印刷原值和规范化数值。每个绑定都必须与单独提供的来源目录条目匹配，并解析至所引用的证据登记表。检查器将货币保留为明确字段，而不是从去除货币符号的值中推断货币。它会检查数值规范化，以及绑定的来源文本中是否出现该值和期间。这在请求的会计计量项目与其来源位置之间增加了可检查的关系；它不会从自由文本中推断该关系，也不会自动重建表格。

目录是一项受信任的输入假设，而非由验证器创造的事实。在实际部署中，必须独立于待核查的建议来确立并审查该目录。匹配目录可以防止已明确规定的绑定被不一致地替换，但无法证明目录本身正确，也无法证明自由文本与其结构化字段相矛盾的主张正确。下面执行的扩展明确包含一个

揭示这一局限的案例。主张蕴含关系、准则适用性、来源真实性和专业判断，均不同于类型化记录一致性。

一个明确为虚构的贷款组合记录展示了这一差异。假设风险敞口为 1,000,000 个货币单位，违约概率为 0.03，违约损失率为 0.40，则算术乘积为 12,000。该记录将此标注为计算示例，而非完整的国际财务报告准则第9号备抵。记录还指出，阶段划分、折现及前瞻性情景充分性尚未确定。最终处置是不作出专业损失准备结论，同时保留算术值以及暂不作出结论的理由。本示例中的所有来源文本、初始值以及名义上的质疑/响应均为虚构的测试输入，并非观察结果、专家判断或大语言模型输出。该示例展示了该模式的一项具体会计用途：计算完成与专业结论属于不同的知识状态，记录可以保留前者，而不声称已得出后者。

因此，运营适配采用基于规则的方式：依据任务编码确定生成主导方；若解释仍未达成一致，则要求人工主导解决；并在任何获准的完善过程中保留相同的证据/例外链条。ACP 不会根据观察到的结果优化该路由策略。其拟议贡献在于明确建立工作单元分配与可审查的会计判断记录之间的联系，而一般性建议解释本身并未说明这种联系。

4.4 三种应用示例：审计、报告与估值

我们通过三个会计情景展示 ACP，涵盖审计、财务报告和估值领域。

4.4.1 持续经营评估（审计领域）

在一个示例性的美国上市公司持续经营评估中，审计高级职员依据适用的美国公众公司会计监督委员会框架，考虑客户持续经营能力是否存在重大疑虑。ACP 规定：

- **数据获取**：大语言模型建议从业务档案中提取债务契约遵循证据、近期管理层预测和监管往来函件。人类负责筛选相关内容。
- **分析综合**：大语言模型调用经核验的计算工具，根据预测计算偿债覆盖率并进行压力情景分析。人类评估管理层假设的合理性。
- **评价性判断**：由人类提出意见，并按照ACP四阶段工作流程组织。大语言模型充当对抗性审阅者，提出反驳论点。
- **专业沟通**：由人类撰写持续经营备忘录。大语言模型协助处理脚注交叉引用和PCAOB引文格式。

该协议旨在应对模糊性下的评价性判断任务；其明确要求进行对抗性审查，以防范自动化偏见文献中所记录的锚定风险（Skitka 等人（2000）；Goddard 等人（2012））。

4.4.2 预期信用损失（ECL）估计（报告领域）

示例性预期信用损失案例采用国际财务报告准则第9号下一般方法的贷款组合场景。违约概率、违约损失率和违约风险暴露支持相关计算；对信用风险是否显著增加的评估

决定适用的阶段划分。美国公认会计原则下的信用损失要求必须单独具体化，而不能将国际财务报告准则第9号的阶段划分视为会计准则编纂第326号的要求。ACP规定：

• **数据获取**：大语言模型提取贷款层面属性和历史违约数据。人类审查其完备性。

• **分析综合**：大语言模型借助经核验的工具，协助执行和解读指定的定量模型。人类评估模型适当性，并识别定性调整（即实践中常见的“叠加调整”）。

• **评价性判断**：人类为每笔贷款撰写信用风险显著增加（SICR）的判断及叠加调整的理由。大语言模型质疑不同贷款之间叠加调整应用的一致性。

• **专业沟通**：人类撰写国际财务报告准则第7号的信用风险披露。大语言模型协助交叉核查内部一致性。

4.4.3 第三级公允值计量（估值领域）

对于第三级（不可观察输入值）公允值估计，ACP 规定：

• **数据获取**：大语言模型收集市场可比数据、以往交易数据和相关宏观变量。

• **分析综合**：大语言模型协助执行经过核查的估值模型（贴现现金流、市场倍数、以往交易），并呈现估计值范围。

• **评价性判断**：人类选择点估计，并为输入值权重进行辩护。大语言模型针对替代权重提出对抗性论证。

•**专业沟通**：人类撰写第三级披露，包括适用的不可观察输入值敏感性披露。

4.5 失效模式覆盖

ACP 旨在防范文献中记载的五种特定失效模式。表5将每种失效模式映射至旨在应对该失效模式的协议要素。覆盖率属于设计映射，而非已测量的缓解效果。

表5。 人机协作中已记录失效模式的ACP覆盖范围。

失效模式	描述	拟议的ACP响应
锚定偏差	过早趋同于AI建议	阶段A独立估计人工智能接触之前
自动化偏见	接受AI输出而未经核验	阶段C结构化核验问题
失准的信任	过度或不足依赖，因……转向未经校准的信心	阶段B明确假设+ 来源链接
隐藏的推理依据	无法为人工智能的结论辩护——在合作伙伴审查中	阶段B结构化推理依据专家组
遗漏的证据	基于以下内容的AI建议不完整的证据	C阶段第二季度证据-添加提示

5. 命题与竞争性解释

分配矩阵为条件性命题提供依据；它并不能从逻辑上证明绩效优势。以下主张将所提出的机制与设计约束区分开来。在本文中，所有这些内容均未在会计领域经过检验。若足够精确的估计值显示相反方向，则会对相关命题提出质疑。仅有不显著的估计值并不能证伪该命题，也不能证明两者等价。

P1：经核验的信息综合。 当模型能够获取相关证据、其输出可独立核查，且人类确实核验了重要主张时，预期辅助综合相比无辅助综合能够减少参考范围偏差。来源覆盖度不足、计算错误或核验成本可能使这一预测反转。这是一个取决于能力的预测，并非主张所有数值任务都更适合大语言模型。

P2：未解决的评估下的有害依赖。 当建议有误或缺乏充分依据，且证据留有解释空间时，预期无结构的接触会使判断朝着加剧参考值不一致的建议方向移动。算法厌恶和有充分依据的拒绝则给出了相互竞争的预测。朝着合理建议移动属于恰当的更新，而非有害锚定；仅凭移动方向无法判定 P2。

P3：强制设置检查点的净效应。 与其他条件相同但不设置检查点的人工智能使用情形相比，要求明确比较来源证据与推理的检查点，预期会提高最终的参考值一致性。其作用机制是对重要主张进行额外审查。流于形式地完成检查点、逆反心理以及对最初错误的人类估计的固守，都可能抵消或逆转这一益处。P3 关注的是强制设置检查点的总体效应，包括其注意力和时间成本，而非某个孤立的心理机制。

P4：提供完善的总效应。预计在核验之后提供结构化质疑和模型回应，相较于单独核验，能够进一步提高参考值一致性。这种益处

在假设确实仍存在争议的情况下可能更大，但额外的说服性错误或时间压力也可能使其适得其反。提供完善与实际使用完善是两种不同的干预；拟议的随机对比估计的是提供完善的效果。

P5：任务相关成本。与非结构化访问相比，所需验证预计会增加专业人员耗时，而准确性与时间之间权衡的价值则随任务可验证性和模糊性的不同而变化。更快地获取证据可能抵消新增步骤，因此不一定会导致总耗时增加。必须先分别报告准确性和耗时；当增量耗时接近于零或为负时，增量每分钟准确度比率便无法解释。

校准是独立的计量项目和实现要求，而不是功效命题。保持不确定性披露不变的研究无法识别该披露的因果效应。同样，拟议研究保持独立估计不变，因此无法估计阶段A本身的益处。检验这些组件需要增加随机化条件。

以下候选人类评估仅在满足测量与可行性门槛后，才对 P3 和 P4 进行验证性推断。由于三案例设计无法清晰区分领域、需求、建议质量和案例特有因素，P1、P2 和 P5 仍属于探索性命题。要有说服力地检验 CDAJ 的分配预测，最终需要每个单元包含多个独立抽样的任务，并对需求和建议质量进行析因变化。

6. 人工评估议程

这些概念性命题需要未来的人类证据。公开来源实验和合成验证器无法证明专业判断有所改善，而下文的规划计算也不能验证行为终点指标。技术补充材料保留了一个候选四条件设计：无辅助的人类判断（C1）、固定模型建议（C2）、所需验证（C3），以及验证加可选优化（C4）。所有初始估计值都将在分组和接触实验条件之前收集。一种可能的验证性检验系列为 K2=C3−C2 和 K3=C4−C3，其解释为界面的总体效应，包括注意力、时间和对话的影响，而非孤立的认知机制。其他对比和案例层面的效应仍属探索性分析。

任何验证性注册都必须先通过若干道关卡。拟议中的独立五人评审组将对每个案例至少 12 个证据／假设／论证条目进行评分，并记录可辩护范围的边界及分歧。序数一致性分析（Krippendorff 2011）针对的是这些条目评分，而非单一判断或虚构的信度系数。随后，独立试点必须考察分布的地板效应、量表稳定性、建议质量，以及终点指标是否反映了预期判断。一个候选终点指标是经标准化的辩护范围外距离，用于衡量可接受性一致，并刻意不对范围内的答案进行排序。我们不假定它服从高斯分布或已得到验证。补充材料另行说明了推理质量指标和合理中心估计敏感性指标。

在确定样本量之前，必须先明确专业领域最小重要改进、可接受的额外时间、联合效能要求、招募渠道、负担、报酬和预算。每组 225 名完整案例等效参与者这一数值情景仅为假设性示例，并不代表已获得经费支持或具备可行性的招募承诺。在招募人类受试者之前，必须先完成本地伦理审查认定、知情同意、隐私保护及第三方处理安排。本研究并未声称已设立参考评审组、开展有效性试点、获得伦理

批准、完成预注册，或确定专业领域有效性研究的样本量。如果无法满足这些测量与可行性门槛，下一项研究应开展可行性调查，而非名义上的验证性检验。完整的候选程序、评分选项、缺失情况假设和注册前提均保留在补充材料中，以便审查这一限制。

7. 公共来源模糊性诊断与计算验证

7.1 范围与可复现性

我们开展了公共披露来源模糊性诊断、事件顺序与记录完整性检查，以及前瞻性高斯规划实验。`gpt-6-astra-max`（通过已配置服务访问的 OpenAI 模型所使用的模型显示名称）协助准备和检查 Python 代码；所报告的检索、检查和模拟指标由确定性脚本生成，而非依据模型断言。本节中的计算结果不同于拟议的人类有效性研究，以及下文单独报告的编码练习。随附的可复现性档案记录了来源 URL 和哈希值、有限的源材料衍生卡片、脚本、环境、随机种子、派生检索评分、模拟抽样、机器可读结果和校验和。公开报告的财务披露是真实来源数据；合成测试样例和高斯抽样则是生成的测试输入。两者均不是参与者响应数据集。在这些计算检查中，未使用实验性大语言模型运行或从业者有效性评估所产生的任何会计建议。相关程序检查所提供的模式/事件和记录契约；它们并未实现自动 CDAJ 任务分配器。这些检查依据既定假设确立了选定的实现和统计属性；但并未验证 CDAJ 的分配规则或 ACP 的行为有效性。

7.2 公开披露来源追溯与模糊性诊断

7.2.1 问题与语料库构建

这项诊断要回答的问题是：特定的页面级共现规则是否会改变指定来源页面的找回率，以及页面找回率与来源中的印刷表格标签之间还存在哪些模糊性。这并非对结构化上下文要求的一般性测试。它检验的是一项与 ACP 相关的证据定位操作，而非该协议对人的影响，也不评估会计解读在专业上的正确性。主要比较向常规字段袋检索和确定性的上下文存在门控提供完全相同的查询信息。仅值搜索是次要的信息移除消融，并非现实分析人员或大语言模型的基线。

我们于2026年11月22日下载了汇丰控股、巴克莱和渣打银行在发行方网站发布的完整2024年年度报告（汇丰控股 2025年；巴克莱 2025年；渣打银行 2025年）。在检索评分之前，已记录指定的发行方/年份选择、符合条件的单元格、评分方法和修订内容。三份报告均成功下载，且未作替换。报告分别包含460、536和410页 PDF，共计1,406页；语料库中仍有三页为空白文本。我们保留了来源网址、下载时间戳、PDF 哈希、提取版本信息以及失败/排除记录。这是一个有目的选取、仅用于评估的可行性语料库，并非随机样本、训练集或对拟合模型的留出测试集。

每份报告均选取两张合并主题表格：信用减值/预期信用损失和第三层次金融资-

产。在每张指定表格中，按行优先顺序保留符合条件的 2024年数值单元格中的前三个，且仅限于所述主题列。破折号不属于数值单元格；其中两个已跳过并记录。重复金额仍作为不同的行/列观测值保留。最终得到 18 张卡片，集中在六个来源页和三家发行人。每张卡片均保留报告、PDF 和印刷页码、表格、行、列、期间、单位及印刷值。 `gpt-6-astra-max` 辅助提取并对六页进行视觉检查；另一位代理核查了相同的来源单元格。初始标签由模型辅助生成，并非独立生成的人工标注或从业者判断。之后，人类请求者确认已使用来源材料包完成对全部 18 个来源数值、行/列标签、期间和单位的核查；此项确认也涵盖下文报告的五项事后比较。未要求进行更正，冻结标签和结果均未改变。未另行提供实际核查者的身份、专业资质及是否采用盲法的信息。这是有记录的人为来源核查，并非独立专家裁定。随后，使用另一款 PDF 提取库对相同的经哈希校验的报告进行技术交叉核查，版本为 `pdfplumber` 0.11.9：在指定页面上均找到了全部 18 个印刷值及完全一致的单位字符串，其中八张卡片的数值出现了多次。坐标佐证和模糊性均予以保留。此项独立实施的提取核实了文本的存在，但并未将数值对应到特定行/列，也不构成独立人工裁定。

表6. 公开来源语料库及指定参考页。

发行人 / 2024年年度报告	PDF 总页数	预期信用损失起源	第三级来源	卡片
汇丰控股	460	144 (142)	389 (387)	6
巴克莱	536	308 (307)	479 (478)	6
渣打银行	410	212 (210)	325 (323)	6

注：起始页码采用从 1 开始计数的 PDF 页，括号内为印刷页码。每个主题表格包含三张卡片；这些是聚类单元格，而不是 18 份相互独立的报告或来源页。

7.2.2 冻结检索方法和指标

所有方法都在完整的合并页面语料库中进行检索。查询仅包含印刷值、根据行/列标签推导出的指标短语、期间和单位。查询中不提供确切的来源 ID、页面定位符和发行方身份。由于这些字段是根据来源本身构建的，因此这是一项有利的受控查询测试，而非对自然发生请求的抽样，也不是对生成正确上下文能力的测试。

方法 B 是采用固定 $k_1 = 1.2$ 和 $b = 0.75$ 的常规 BM25（Robertson 和 Zaragoza，2009）。它对每个不同的查询词元仅使用一次，并采用正值 IDF 形式 $\log(1 + (N - df + 0.5)/(df + 0.5))$。方法 C 使用与方法 B 完全相同的分数和查询，但仅保留包含规范值以及所有非空期间和单位词元的页面；若没有符合条件的页面，则弃答。作为次要消融的 方法 A，仅根据数值词元的出现频率对页面排序。这里不使用学习模型、人工重排序、依赖结果的调优或源页面提示。

固定分词器会将 NFKC 规范化文本转换为小写、规范化逗号分组数字和小数，并将括号括起的数字字面量视为负数。它会保留带有不同前缀的 ASCII 字母和数字词元，但会移除货币符号：`$m` 和 `£m` 都会生成词元 `m`。这明确限制了货币区分。它不会重建表格坐标，也不会验证年份/单位是否适用于匹配值，而非同一页上其他单元格中的值。现已提供精确的预处理、代码/输入哈希值，

评分前修订内容和数值检查。

主要描述性指标是指定来源页召回率@1和Recall@3。倒数排名和候选页面数量属于次要指标。固定的报告/页面顺序用于打破显示排名中的分数并列；针对并列组的平均指标则对并列组内所有位置进行平均。得分为零或被 C 排除的来源，即使确定性的零分排序会将其排在较前位置，也会获得零找回率。所有卡片均保留在结果中。不报告总体置信区间或显著性检验：这 18 个有目的选取的聚类单元格并非相互独立的抽样企业。

7.2.3 结果：门控机制未带来额外的 Recall@1 或 Recall@3 收益

表 7. 18 张公开披露卡片的指定来源页找回率。

方法	Recall@1	Recall@3	平均倒数排名	平局平均 Recall@1
A：仅数值消融	7/18	11/18	0.5200	0.2545
B：全字段 BM25	13/18	18/18	0.8333	0.7222
C：同信息门控	13/18	18/18	0.8426	0.7222

注：B 与 C 是主要比较；A 会移除信息。在此处，B 和 C 的确定性 Recall@1/@3 及按平局平均的 Recall@1/@3 完全相同。精确来源恢复不等同于语义会计准确性。

方法 C 只能删除候选项；不会改变其保留页面的相对顺序。只有移除排名较高的干扰项，找回率才会提升。因此，此项比较针对的是这一有限的共现规则，而非所有结构化证据核验的价值。

新增的门控机制在任何卡片上都未提升 Recall@1 或 Recall@3：18 项比较均无变化。有一个来源从第三名升至第二名，这解释了倒数排名的微小变化；但这并不能证明检索能力普遍更优。没有任何查询导致门控机制弃答或排除其来源。因此，实验发现，仅要求值／期间／单位共现，并未在全字段 BM25 之外带来 Recall@1 或 Recall@3 的额外收益。

B 的 Recall@1 在汇丰银行为 5/6，在巴克莱为 5/6，在渣打银行为 3/6；按主题划分，预期信用损失为 8/9，第三级为 5/9。与 A 相比，B 在十张卡片上的按平局平均 Recall@1 有所提升，五张保持不变，三张有所下降；这三项下降均来自渣打银行的第三级卡片。包含该数值的页面平均数量总体为 35.83，其中预期信用损失为 68.11，第三级为 3.56。这些计数描述的是本次选取的小额重复数值和较大金额，并不能证明不同会计任务的研究人群存在差异。全部 18 个来源数值均在其记录的页面上找到。档案保留了全部 25,308 条卡片逐页评分行，以及每张卡片的排名、平局情况、符合条件的页面和来源词元核查记录。

7.2.4 五项来源漏判的事后标签比较

评分冻结后，我们对排名最高的五个非原始来源候选项进行了说明性视觉审计，涉及四个不同页面。其中三个明显重复了报告中的期末余额：汇丰银行衍生资产为 2,149；渣打银行客户贷款为 1,937；渣打银行债务证券为 1,593，且替代表格中的类别标签略有扩展。第四个候选项在一张标注为净敞口的敏感性分析表中重复列出了渣打银行 3,239 的逆回购金额，而非原始来源中的资产公允价值层级。这种呈现差异使其无法自动视为语义等价。

第五个候选项的表格标签和列标签与请求的源单元格在字面上不同。目标是巴克莱 2024年英国零售抵押贷款第一阶段减值准备，金额为 3,600 万英镑（PDF 第 308 页）。检索到的第 324 页包含一张零售抵押贷款经济不确定性调整表：其总额为 3,600 万英镑，而第一阶段金额为 700 万英镑。因此，值、年份、单位和行文字可能同时出现，但该值实际属于不同的计量项目和列。冻结的门控机制未能解决这一不匹配。

这些比较针对报告中印出的金额和标签，并不裁定发行人的会计处理是否正确、估计值是否恰当，也不涉及专业人士应如何使用证据。报告未说明核查者的资历，也不会根据确认结果推断其专业能力。

这项事后审计既不会重新标注来源目标，也不会生成新的语义准确率。本次仅检查了五个来源错误；并不存在预先指定且经独立裁定的语义参考集。原有的 13/18 和 18/18 指标保持不变。相反，这项审计揭示了两个不同的局限：非来源页面未必有误，而词元匹配的页面未必能支持预期的会计主张。研究未构建持续经营实证案例，也未测量参与者或大语言模型的会计绩效结果。

一个关于信息损失的基本观察。设 $R(s)$ 仅保留某来源页面的汇总词元计数和长度，并设 q 为以下主张：投资组合 A 显示的备抵为 36。考虑两张表格：它们的投资组合标签 A 和 B 以及备抵表头完全相同，但数值 36 和 7 在两行之间互换。保持所有标题、期间／单位标签及其余语料不变。此时，$R(s_1) = R(s_2)$，但对于其中一页，q 为真；对于另一页则为假。只观察 $(R(s)\ q)$ 的验证器，在两种情况下必须作出相同的决策，或产生相同的随机决策分布。因此，对于任意此类印刷单元格主张，它无法以概率 1 保证既不发生错误接受，又接受每一个真实主张；在随机情形下也无法作出这种保证。弃答可以通过牺牲完备性来保障可靠性；但这并不排除实现有用的近似性能。这个基本观察针对的是行值绑定的损失，而非会计处理是否正确、实际检索率的上限，或人类或 ACP 能解决该问题的证据。观察到的语料结果仍是上述描述性结果。

这项设计启示要求实现主张级表格绑定：对于表格证据，定位信息应保留文档/版本、页、表格、行、列、期间、单位和字面值。公开卡片提供了这些字段。基础记录检查器会核验标识符和定位信息是否存在。后续的类型化契约扩展还会检查这些信息是否与提供的单元格绑定目录及数值/期间出现情况一致，但不会解析表格，也不会确立语义蕴含。源标签诊断促使我们增加这一验证义务，并界定了较弱的实现，而不是核验 ACP 的人工检查点。

7.3 选定的事件顺序和负载检查

新建的参考验证器会检查数量有限的单案例 ACP 事件轨迹。五条有效的合成轨迹和 21 条经过刻意修改的轨迹，用于测试多种情况，例如阶段缺失、与条件不符的事件、冻结输出哈希值不匹配、未匹配的细化响应，以及待处理交换期间定稿。验证器对全部 26 个测试样例均给出了预期判定。这是作者设计的软件正确性测试，并非基于真实会话进行的独立验证。

验证器不会检查实际初始估计、初始证据、最终结论或签名是否存在，也不会核验解释或核验回应的质量。它

不实现浏览器界面、身份控制、仅追加存储、隐私保护，也不确保参与者与案例之间的隔离。其严格的时间戳顺序是一种简化假设；生产系统应使用可靠的事件序列，因为现实中的毫秒级时间戳可能相同。通过这些测试表明，选定的工作流约束可以编码实现，但并不意味着ACP系统已准备好部署，也不意味着用户会有意义地核验建议。

7.4 已定稿记录完整性检查

另有一个可执行验证器，用于实现表4中的记录规则。它在33个合成测试样例上得出了预期决策：两个可接受的记录（分别包含和不包含可选交换），以及31个刻意构造的无效变更。这些变更包括：缺失初始值或最终结论、来源内容已变更但哈希值未变、证据ID或假设ID悬空、未追踪有争议的主张、缺失或重复的例外处置，以及缺失或未经确认的签核。档案中提供了完整的输入、决策和拒绝理由，以及上述明确虚构的会计样本。重新运行后，四项JSON结果/测试样例输出逐字节完全一致。

新的检查器验证的是是否存在及交叉引用，而非真实性或专业充分性。如果攻击者同时控制文档及其哈希值，来源哈希值就无法验证文档的真实性。即使记录符合模式，标注为“有证据支持”的标签、标记为“已处理”的例外，以及模拟签核标记，也都可能不属实。该检查器不验证事件顺序；事件检查器不验证记录内容。不得将二者的测试数量合并起来，作为集成式ACP平台已获认证的依据。实质性结论更为有限：所述证据—例外—结论关系是可执行的要求，而不只是叙述性建议。

7.5 类型化契约扩展与有意设置的语义反例

单独版本化的扩展保留了原有的33个测试样例验证器，不作任何更改，并增加了类型化会计上下文和来源绑定规则。其42个确定性测试样例产生了指定的结构性判定：40项拒绝、2项接受。变异测试涵盖缺失字段、无效范围、悬空的目录/来源引用、与未变更的请求绑定不符的计量项目或列、字面值规范化不一致，以及尽管来源哈希值匹配，数值/期间出现情况仍然缺失等情形。在记录的运行环境中，5个 JSON 输出均可逐字节复现。这些是软件测试样例，并非抽样得到的会计错误，也不能据此推断专业工作的错误检测率。

其中一个被接受的测试样例是一条标签齐全的伪造记录，但其指引尚未明确，可接受范围也尚未确立；因此不提供专业结论。另一个是有意设置的语义反例：主张的自由文本仍然请求第一阶段减值金额，但结构化请求绑定和提供的绑定均已改为调整总额。尽管自由文本的含义与之冲突，自洽的类型化契约仍然通过。相比之下，如果仅更改提供的绑定，同时保留原始请求标识，则会被拒绝。该测试样例的灵感来自真实的同一36披露冲突，但其目录和记录文本均为伪造内容，并非额外的报告观察结果或专家标注。

这一结果既展现了形式化工件的价值，也揭示了其边界。相对于固定的请求，类型化契约可以拒绝不一致的来源替换；但它无法判断哪个请求、目录条目或会计解释在实质上正确。被接受的反例被报告为语义保证方面的局限，而不计作正确的会计结果。此次扩展未更改任何原始公开来源标签、检索

评分或第一版固定装置。

7.6 假设性连续终点下的规划示例

此示例针对假设性连续终点；它并非针对未经试点的范围距离评分进行的效能验证。假设每位参与者有三个权重相等的观测值、单案例方差相等σ^2，且相关性为复合对称相关 ρ。参与者均值的方差为

$$\mathrm{Var}(\bar{Y}_i) = \sigma^2 \frac{1 + 2\rho}{3}.$$

对于每组样本量相等的情况 n，且标准化差异以单案例标准差为单位，标准误为

$$SE_d = \sqrt{\frac{2(1 + 2\rho)}{3n}}.$$

在 $n = 225$ 和 $\rho = 0.30$时， $SE_d = 0.068853$。在 Holm 阈值中最保守的第一个阈值（Holm 1979）下，双侧 $\alpha = 0.025$，正态近似下的 80% 最小可检测差异为

$$d_{80} = (z_{0.9875} + z_{0.80}) SE_d = 0.212276.$$

对于 $d = 0.20$，0。30、0。40，固定阈值效能分别为 0.7464、0.9828 和 0.9998。这些是依赖于假设的规划数值，并非实际达到的精度。相应的单案例最小可检测效应为 0.2907，说明合并效能不能套用于案例特异性结论。将 ρ 从 0 提高到 0.60，会使合并最小可检测效应从 0.1678 增至 0.2489。更灵活的协方差模型并不保证能提高效能。

保留固定的 $n = 225$ 目标作为规划情景，而不是将其视为具有普遍意义的 0.40 效应，或视为已证明的招募能力。在未来研究确定样本量之前，应与实践者在原始评分量表上商定具有实际意义的改进幅度。这些公式假设方差相同；基于试点的标准化并不能保证不同组别或新样本也满足这一假设。

7.7 设定随机种子的蒙特卡罗模拟结果

七种情景各进行了 100,000 次重复模拟，随机种子为 20261122–20261128。每次重复都模拟 C2、C3 和 C4 的高斯组均值，其方差为 $v(1 + 2\rho)/(3n)$，其中 $v = 1$，但方差加倍的压力情景除外；随后使用已知方差计算两个对比统计量，并应用 Holm 校正。对于这两项检验，C1 并非必需。由于 C3 以相反符号共用，对比误差在该模型下的相关性为 -0.5；这是人为设定的代数性质，并非行为发现。

表8. 声明的高斯规划模型下的运行特性。

情景	数量	估计值 [95% 蒙特卡洛区间]
两个增量均为零	任何错误拒绝	0.04662 [0.04533，0.04794]
一个零假设；另一个 -0.40	拒绝真实零假设	0.05039 [0.04905，0.05176]
两者均增加 -0.20	两者均被拒绝。	0.65671 [0.65376，0.65965]
两者均增加 -0.30	两者均被拒绝	0.98296 [0.98214，0.98374]
两者均为 -0.30；$\rho = 0.60$	两者均被拒绝	0.92212 [0.92044，0.92376]
两者均为原始增量；方差加倍	两者均被拒绝	0.73691 [0.73417，0.73963]

注：威尔逊区间（威尔逊，1927）量化的是蒙特卡洛抽样不确定性，而非对真实会计师的不确定性。默认值为 $n = 225$, $m = 3$, $\rho = 0.30$，单案例方差为 1。最后一行在保持原始效应不变的同时增加方差，因此标准化效应会降低。第七种情景、两个增量 -0.40 以及所有原始抽样值均已存入档案。

估计的全局零假设族错误率低于名义水平 0.05。当两个真实增量均为 -0.20 时，经 Holm 校正的边际拒绝概率分别为 0.82215 和 0.81585（不同于固定为 0.025 时 0.7464 的检验效能），但联合拒绝概率仅为 0.65671。如果理论要求核验和完善都能增加值，这一区别就很重要。部分零假设结果 0.05039 在蒙特卡洛不确定性范围内与 0.05 相符；它并不能证明错误率控制超过了其名义上限。

完整重跑后，16 个归档输出文件均逐字节复现。另一个独立的验证脚本重新读取模拟抽样值，并根据解析模型核验抽样方差、对比相关性及固定阈值拒绝概率。这些检查验证了指定规划实验的实现情况，但并未对估计方差、非正态范围距离、地板效应、脱落、不依从，或所提出混合模型的精确有限样本行为进行建模。在招募之前，仍需针对这些特征开展基于试点研究的模拟。

8. 主要编码练习与执行偏差

8.1 材料、回报与范围

针对20个固定不变、构造的假设性会计案例描述，共收到两份分别填写的答复工作簿。这些案例有意包含复合型、边界模糊型和信息不充分型描述；它们既不是随机抽取的任务样本，也不是对专业业务项目的观察。它们不同于附录A中的25项示例性分配。程序说明采用中文；定义和案例描述采用中英双语，财务报告来源则为英语。20个案例均未提供首选标签。编码员可以将某项描述标记为SINGLE、请求SPLIT，或将其边界标记为UNCERTAIN。阶段代码指CDAJ的四个流程阶段，而非ACP的四个界面阶段。需求M同时保留信息密集型需求和评价型需求；U是行政性未编码标记，并非第五个阶段或新型需求。

两位编码员均声明同意将其资料用于研究，彼此为不同的个体，独立完成任务，未参与此前的研究设计，也未接触过论文或拟议测试条目的标签，未与另一位编码员协商，且未借助生成式人工智能辅助作出判断。一位自述目前担任会计师，拥有会计学学士学位；另一位自述担任审计经理和会计学博士生，具备中国注册会计师资格，并拥有五年四大会计师事务所审计经验。这些是自述的职位和资格，并非经过核实的资历。两位均表示与作者不存在关联，也未获得支持或报酬。

据报告，使用的语言为英语和中英双语。未单独评估对中文说明的理解情况。完成日期和主要锁定日期均为2026年11月22日。作者报告称，江西财经大学于2026年11月22日确认无需进行伦理审查；未提供参考编号。具体负责办公室、回复形式及依据均未提供。尚未对任何机构文件进行独立核实，也不据此推断此前获得过批准。个人回应和身份识别元数据均予以保密；仅报告汇总统计。

原始工作簿已计算哈希值并予以保存。经审计的工作副本规范化了一处仅大小写不同的人类来源声明，并移除了两行仅含 ID、其他内容为空且重复的占位行（存在已填写的行）。这些行政性修正不会改变单元划分、阶段或需求响应。所有初始主要标签均保持不变；分析中未纳入共识标签或裁定标签。

8.2 尚未解决的辅助格式与有限执行

冻结的完整方案读取器因格式错误而停止。两份返回结果均包含数值型真实性回答，而随附指南和工作簿规定的选项为可信、不可信或不明确。原始工作簿设有类别下拉菜单，但禁用了输入和错误警报，这是回答界面的一个缺陷。返回文件中的验证设置与原始设置不同；现有文件无法说明原因。没有依据将数值转换为类别。另有一份来源比较回答在一个规定只能填入一个类别的字段中包含了两个类别。这些辅助回答均予以保留，未作重新编码，也未指定首选解释。因此，完整的冻结分析仍未完成。

现实性与来源比较的回答不纳入预先声明的主要编码估计目标，也不纳入编码员声明的资格规则。对于本次仅基于主要数据的稿件修订，我们于2026年11月22日08:33:52（协调世界时）记录了单独的执行范围，早于08:34:40（协调世界时）的首次一致性系数计算。该范围使用冻结的一致性计算和可比性规则，对同两名编码员的主要字段及全部20个案例进行验证。这是在收到回答后披露的执行偏差，并非公开预注册，也不代表完整的原始流程已通过。不会因辅助格式而移除任何编码员或案例，也不会对缺失或未解决的回答进行插补。

8.3 报告规则

强制性的描述性层保留全部20个案例，包括拆分请求、不确定边界、U、M和缺失回答。精确原始一致率包含明确标记为未解决的标签，因此单凭这一指标并不能说明可分类性是否实用。仅当两名编码员都选择 SINGLE，且都提供非缺失、非 U 的代码时，才针对每个维度分别计算次要名义尺度 Cohen’s κ 。需求系数比较 I/E/M 回答概况；它不是排他性二元 I/E 系数，也不能验证八格分类体系。因此，已解决 I/E 案例的覆盖率需单独报告。未对齐子单元不合并计算一致性。本文不提供总体置信区间、显著性检验、通过阈值或共识系数。

8.4 汇总主要结果

全部20个案例均有配对单元划分、阶段和需求响应，无一缺失。两份匿名响应档案分别包含14和15个单一决策、3个和2个拆分决策，以及

每份档案均有3个不确定决策。对于完整的单元划分—阶段—需求三元组，完全一致率为17/20。表9报告了全部案例及条件结果。

表9. 20个构造案例的主要编码一致性与覆盖率。

数量	单元划分	阶段	需求
完整的配对回答	20/20	20/20	20/20
所有案例完全一致	19/20	20/20	18/20
全案例一致率	0.950	1.000	0.900
条件可比较案例	—	14/20	14/20
条件精确一致率	—	14/14	12/14
条件名义值 κ	—	1.000	0.752

注：全案例一致率包括 U 以及在非单一父级描述中提供的标签。条件结果仅保留双方均为单一、且轴代码非缺失、非 U 的案例。需求系数针对 I/E/M 配置，而非排他性 I/E 分类。未指定单元化系数。

对于两个维度，有六个案例因并非同时为单一而被排除在条件系数计算之外；其中三个案例还带有明确的 U 代码。这些是相互重叠的排除原因，并非九个不同的排除项。在 14 个可比案例中，根据观测边际分布计算得出的阶段和需求的偶然一致率分别为 0.265306 和 0.423469。在全部 20 条描述中，每个配置的阶段标签频次均为：S1 五次、S2 四次、S3 四次、S4 四次，U 三次。两个匿名配置的需求频次分别为 I/E/M/U = 5/7/5/3 和 5/8/4/3。一些原始阶段或 M 标签出现在非单一父项上；这些标签予以保留以作描述，但不纳入条件计算。冻结的结构检查还标记出，一个配置中的三条描述和另一个配置中的两条描述包含非单一父项维度标签。对于两条拆分描述，其中一个配置提供的子单元行不足两行。这些不利诊断结果均予以保留，未对标签进行修正、未对齐子单元，也未增加其他排除项；六个并非同时为单一的案例已根据原有条件规则排除。

个别来看，12/20和13/20条描述被标记为 SINGLE，且带有 I 或 E 需求；每份编码资料中还有两条被标记为 M 的 SINGLE 描述。共同解决的 I/E 覆盖率为 12/20（60%）。有一个案例中，两位编码员均将其标记为 SINGLE 和 M；另有两个共同标记为 SINGLE 的案例，至少一位编码员将其标记为 M。因此，阶段条件一致性完美，但单元边界仍未解决，排他性需求覆盖率也不完整。这些发现表明，在这些材料和声明条件下，部分编码响应具有可重复性。但这并不能证明标签正确、自然任务覆盖率、人机分配最优，或 ACP 对判断的影响。未提供可选分配响应。下文将单独分析来源核查响应；它们不构成来源准确性评分或专家裁定。

8.5 单独的描述性来源-响应练习

两次返回均包含对全部 18 个来源核查项目的回答，共形成 36 条记录。2026年11月22日 09:27:45 UTC 记录了一项单独的执行范围，早于 09:28:42 UTC 首次汇总来源状态的时间。该范围保留了两位受访者和全部 18 个项目，并按照原有模式说明了五个组成字段及总体判断。它既不更改主要编码规则，也不将原有的完整计划错误视为已解决。40 个数值现实性答案以及尚未解决的第五项比较仍不属于此次描述性执行；没有引入新的系数或通过阈值。

表10. 36 条响应记录中汇总记录的来源核查状态。

字段	匹配	差异	不明确	缺失
值	28	4	4	0
Row	30	4	2	0
列	32	4	0	0
期间	32	4	0	0
Unit	32	4	0	0
总体	20	12	4	0

注：分母为 18 个项目乘以两次返回。这些是记录的判断，并非针对独立裁定参考的正确性评分。未披露项目级或与受访者关联的回答。

36 条记录中有 16 条至少存在一种不利或不明确状态。这不能解读为有 16 张不同的错误来源卡片：这些计数指的是回答，而不是经独立确认的缺陷。组成状态与总体状态分别保留；总体回答不会用于填补或覆盖任何组成部分。私人审计保留了每一项潜在的跨字段不一致。汇总前记录的一项披露规则会抑制边界计数，因为这些计数与其他表格结合后，可能会泄露受保护的完整回答标签。本文披露了这一隐私限制，而没有以声称来源工具已通过来取代这一说明。

前四个比较项目均具有完整且符合架构规范的关系。在这八个回答中，四种类别各出现两次：印刷值和标签相同；值不同；值相同但标签不同；以及标签相关但范围不明确。聚合计数无法确定每个标签对应的是哪项比较或哪位受访者。第五个项目仍未解决，因为其中一个回答合并了多个类别；明确将其排除在这些计数之外，并不意味着将一项五题工具变成一项已完成的四题研究。根据这些计数，没有任何关系类别被裁定为正确。

这些回答补充了最初的请求者确认，其中包括另行返回的来源判断，包括不利和不确定的判断。它们并未解决所有来源标签问题，也无法验证专业知识。来源位置审计必须将年份、列、子组和上级表头与所显示的值对应起来：同一页内可能出现重复值或重复的叶级标签。这类文档核查不同于对受访者回答的判断，不得覆盖原始响应类别。原始检索目标和所有检索结果均保持不变，包括 Recall@1/Recall@3 未提升，以及唯一一次从第三位升至第二位的变动。

9. 对人工智能设计与研究的启示

9.1 系统设计与业务委托审查

公开来源诊断具体揭示了这一薄弱实现的局限：在排名第一和第三的位置，强制加入词元共现并未带来任何恢复收益，而且检索到的数值可能出现在与请求不同的表格或列标签下。这支持将页面检索视为候选证据步骤，随后进行明确的表格单元格核验和主张核验；但这并不能证明所提出的人类检查点确实能够很好地执行此类核验。

这项直接的设计贡献，是明确人工智能辅助判断记录应当让哪些信息可供检查：谁提供了估计值、每项重要主张由哪个来源支持、人类

质疑了什么，以及为何得出最终结论。系统可以要求填写这些字段并规定事件顺序，但这并不能证明它们能够改善判断。因此，管理者应将语义质量和专业人员耗时与完成率一并评估。检查清单即使填写完整，也可能掩盖审查不充分的问题。

CDAJ 还建议将计算服务、检索、生成式推理依据和专业签核分开。不同组件需要不同的检查。确定性计算结果正确，并不能证明缺乏依据的假设成立；引文看似可信，也不能证明来源支持该结论。异常处理应允许专业人员在证据缺失时暂停人工智能工作流程，而不是强行生成名义上完整的记录。

如果所提出的 K2 对比显示出明确的益处，就能支持在所研究的案例和输出中设置必要检查点。如果 K3 并未带来经精确估计的实际益处，则可能需要重新设计或更有针对性地使用；若区间较宽，其价值仍无法确定。无论哪种结果，都不能据此为所有会计工作确立普遍规则。任何部署决策都必须考虑耗时、错误严重程度、证据获取情况，以及现有审查程序的质量。

9.2 专业监督与教育

该协议提出了一种候选文档结构，供事务所、教育工作者和准则制定者讨论，并非对监管合规性的认定。证据留存和问责性签字确认与审查相关，但适用的法律和专业要求取决于司法管辖区和具体业务。计算检查不能成为强制使用 ACP 或改变专业责任的依据。

在教育中，这些案例可用于设计练习，比较初始判断、人工智能建议、来源证据和修订。重要的教学对象是接受或拒绝建议的理由。学习效果提升或锚定偏差减轻仍需单独评估；要求学生解释分歧，并不能证明他们已培养出职业怀疑。

10. 局限性与未来研究

主要局限在于缺乏人类有效性证据。公开披露诊断的规模也较小，且采用目的性抽样：来自三家银行六张表格的18个来源单元格，查询标签则取自来源内容。这并非具有代表性的会计基准，也不是对大语言模型的比较。来源信息提取和事后审计借助人工智能完成，之后又通过一条人类请求者针对来源材料包的回复予以确认。这项初步检查并非盲法专家裁定。后续的36条来源—响应记录中，有16条至少存在一种不利或不明确状态，且无法确立独立裁定的参考标准。这两项检查均不足以支持修订后的语义成功率。CDAJ是作者构建的分类体系，分配清单仅供示例，案例文件和评分小组也尚未最终确定。这些限制使本文的贡献仅限于概念规范、范围有限的技术评估，以及所报告的编码可重复性练习。后者使用构造案例和条件系数，辅助格式尚未解决，且没有经过独立认证的专业资质。它并未验证分配建议或专业实用性。未来有效性研究所用的案例文件和评分小组仍未最终确定。

未来研究设计本身也有局限。它估计的是在三个案例和固定初始建议条件下的效应。它无法区分认知需求与领域或案例难度，无法确立校准后的不确定性，无法分离独立估计效应，也无法证明相较于人工智能单独使用的互补性。A

要支持更广泛的分配命题，还需要开展多案例因子复制研究，改变建议质量和认知需求。跨案例迁移效应和专业异质性也需要明确处理。

与参考范围的距离是一个易于操作但并不完整的判断质量代理指标。它可能奖励与错误专家组的一致意见，忽略薄弱的推理，并产生大量零分。独立的评分标准核验和单独的推理质量指标至关重要。更丰富的计算评估应采用通过试点估计的结果分布和缺失情况，而不是假设高斯分布的统计功效计算可以直接适用于有界观测值或零膨胀观测值。

最后，协议本身所带来的负担可能反过来削弱其目标。要求陈述独立估计可能会固化早期的错误观念；核验可能沦为表演；而完善可能使错误模型更具说服力。这些都是对所提机制的实质性替代解释。未来的证据应能够否定或缩小 ACP 的分配建议，而不是将所有额外的人的参与都视为一种收益。

11. 结论

CDAJ 和 ACP 提供了一种会计特定的方法，将任务分配、来源核验、专业评估和可供审查的决策记录联系起来。本文将人的责任保留为设计约束，同时将绩效收益视为有待检验的命题。一项公开来源诊断发现，简单的情境门控在排名第一和第三的位置上没有带来收益，并说明页面级匹配为何无法核验所请求的表格单元格引用。可复现的记录／事件检查和规划实验确立了其他有限的技术属性，并揭示了未来评估的精度要求。下一项实质性检验是：所规定的增量是否足以改善可辩护的专业判断，从而证明其成本合理。该结果仍有待确立。

数据与代码可用性

作者已准备并维护一份独立的可复现性档案，其中包含官方来源网址、获取／提取代码、PDF 哈希、有限的源材料衍生卡片、选择／失败日志、检索代码与评分、合成测试样例、模拟抽样值和数值检查。计算档案以及一份单独的匿名初级编码与来源回应汇总数据补充材料，均可通过同一私有代码库供审稿人查阅。保密审阅链接另行载于编辑材料中。该档案并未公开存档，也未声称拥有已注册DOI或许可证授权。发布方报告和完整的提取文本留存以供审计，但不在此处再分发，因为尚未取得开放再分发许可。文中引用的官方报告网址和版本哈希值有助于识别来源。本版本还报告了来自两份已返还表格的初级编码与来源回应汇总结果。原始工作簿和衍生的个人回应均由作者私下留存；两位编码员均未授权行级数据共享。尚未生成任何人类效能数据集。

资助

截至目前，本研究未获得专项资助。

利益冲突声明

作者声明不存在利益冲突。

手稿准备过程中关于生成式人工智能和人工智能辅助技术的声明

gpt-6-astra-max （通过已配置服务访问的 OpenAI 模型所使用的模型显示名称）被用于协助手稿修订、来源卡片提取和视觉检查、方法学检查，以及计算代码和材料的准备。DeepSeek-v4-pro（API 响应模型标识符 deepseek-v4-pro，DeepSeek）被用于对修订后的手稿及其配套材料进行独立的自动化评议。这些系统不是研究参与者，也不是所观察到的会计行为的来源。对于已配置的 GPT 部署，无法获得公开权重哈希值或提供商训练快照标识符。印刷的模型名称是面向读者的显示名称；它不代表已单独验证的公开发布版本，也不代表确切的 API 路由标识符。作者负责核实来源、代码、解释及最终内容。最初由人工智能辅助生成的来源标签随后经过人工确认，确认已进行来源检查；本文不声称存在独立的专家标注或专业裁定。

参考文献

Amershi, S., D. Weld, M. Vorvoreanu, A. Fourney, B.Nushi, P.Collisson, J.Suh, S.Iqbal, P.Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, and E. Horvitz. 2019年. 人机交互指南。载于 *2019*年计算机系统中的人为因素*CHI*会议论文集，1–13. https://doi.org/10.1145/3290605.3300233

Austin, A. A., T. D. Carpenter, M. H. Christ, and C. S. Nielson. 2021年. 数据分析之旅：审计师、管理者、监管与技术之间的互动。《当代会计研究》38 (3): 1888– 1924. https://doi.org/10.1111/1911-3846.12680

Bansal, G., T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, 和 D. Weld。1999年。整体是否大于部分之和？人工智能解释对互补型团队绩效的影响。载于计算机系统中的人为因素*CHI*会议*2021*年会议论文集，1–16。 https://doi.org/10.1145/3411764.3445717

巴克莱集团。2025年。*2024*年年度报告。访问日期：2026年11月22日。 https://home.barclays/content/dam/home-barclays/documents/investor-relations/reports-and-events/annual-reports/2024/Barclays-PLC-Annual-Report-2024.pdf

Bonner, S. E. 1999年。会计中的判断与决策研究。《会计视野》13（4）：385–398。 https://doi.org/10.2308/acch.1999.13.4.385

Buçinca, Z., M. B. Malaya, 和 K. Z. Gajos。2021年。信任还是思考：认知强制功能可以降低人工智能辅助决策中对人工智能的过度依赖。《*ACM*人机交互会议论文集》5 (CSCW1)：1–21。 https://doi.org/10.1145/3449287

Commerford, B., S. Dennis, J. Joe, 和 J. Ulla。2022年。人机对抗：复杂估计值与审计师对人工智能的依赖。《会计研究杂志》60 (1)：171–201。 https://doi.org/10.1111/1475-679X.12407

Commerford, B., A. Eilifsen, R. Hatfield, K. Holmstrom, and F. Kinserdal. 2024年. 控制问题：提供输入如何影响审计师对人工智能的依赖。《当代会计研究》41 (4): 2134–2162. https://doi.org/10.1111/1911-3846.12974

Dell'Acqua, F., E. McFowland, E. Mollick, H. Lifshitz, K. C. Kellogg, S. Rajendran, L. Krayer, F. Candelon, and K. R. Lakhani. 2026年. 驾驭崎岖的技术前沿：人工智能对知识工作者生产力和质量影响的现场实验实证。《组织科学》37 (2): 403–423. https://doi.org/10.1287/orsc.2025.21838

Goddard, K., A. Roudsari, and J. C. Wyatt. 2012年。《自动化偏见：频率、效应中介因素及缓解措施的系统综述》。《美国医学信息学会杂志》19 (1): 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Hemmer, P., M. Schemmer, N. Kühl, M. Vössing, and G. Satzger. 2025年。人机协作中的互补性：概念、来源与证据。《欧洲信息系统杂志》34 (6): 979– 1002. https://doi.org/10.1080/0960085X.2025.2475962

Hevner, A. R., S. T. March, J. Park, and S. Ram. 2004年。信息系统研究中的设计科学。《管理信息系统季刊》28 (1): 75–106. https://doi.org/10.2307/25148625

Holm, S. 1979年. 一种简单的序贯拒绝式多重检验程序. 《斯堪的纳维亚统计学杂志》6 (2): 65–70. https://www.jstor.org/stable/4615733

Horvitz, E. 1999年. 混合主动式用户界面的原则. 载于*SIGCHI*计算系统人为因素会议论文集, 159–166. https://doi.org/10.1145/302979.303030

汇丰控股有限公司. 2025年. *2024*年度报告及账目. 访问于2026年11月22日. https://www.hsbc.com/-/files/hsbc/investors/hsbc-results/2024/annual/pdfs/hsbc-holdings-plc/250219-annual-report-and-accounts-2024.pdf

国际审计与鉴证准则理事会. 2021年. 应对因使用自动化工具和技术以及实体系统生成的信息而产生的过度依赖技术的风险. 纽约：国际会计师联合会. https://www.iaasb.org/publications/addressing-risk-overreliance-technology-arising-use-automated-tools-and-techniques-and-information

Joyce, E. J. 和 G. C. Biddle。1981年。《审计中的概率推断中的锚定与调整》。《会计研究杂志》19（1）：120–145。 https://doi.org/10.2307/2490965

Krieger, F.、P. Drews 和 P. Velte。2021年。解释审计中高级数据分析的（非）采用：过程理论。《国际会计信息系统杂志》41：100511。 https://doi.org/10.1016/j.accinf.2021.100511

Krippendorff, K. 2011年。编码可靠性中的一致性与信息。传播方法与测量5（2）：93–112。 https://doi.org/10.1080/19312458.2011.568376

Lee, J. D. 和 K. A. See。2004年。对自动化的信任：为适当依赖而设计。人为因素46 (1): 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Libby, R. 1976. 人与人的模型之争：一些相互冲突的证据。组织行为与人类绩效16 (1): 1–12. https://doi.org/10.1016/0030-5073(76)90002-7

Munoko, I., H. L. Brown-Liburd, 和 M. A. Vasarhelyi. 2020. 在审计中使用人工智能的伦理影响。商业伦理学期刊167 (2): 209–234. https://doi.org/10.1007/s10551-019-04407-1

Nelson, M. W. 2009. 审计中职业怀疑的模型与文献综述。《审计：实践与理论期刊》28 (2): 1–34. https://doi.org/10.2308/aud.2009.28.2.1

Parasuraman, R. 和 D. H. Manzey。2010。《人类对自动化的使用中的自满与偏差：注意整合》。《人为因素》52（3）：381–410。 https://doi.org/10.1177/0018720810376055

Parasuraman, R.、T. B. Sheridan 和 C. D.Wickens。2000。《人机自动化交互的类型与层级模型》。《*IEEE*系统、人类与控制论汇刊——*A*辑：系统与人类》30（3）：286–297。 https://doi.org/10.1109/3468.844354

Robertson, S. 和 H. Zaragoza。2009。《概率相关性框架：BM25及其后续发展》。《信息检索基础与趋势》3（4）：333–389。 https://doi.org/10.1561/1500000019

Skitka, L. J.、K. L. Mosier 和 M. D. Burdick。2000。《问责与自动化偏见》。《国际人机研究期刊》52（4）：701–717。 https://doi.org/10.1006/ijhc.1999.0349

Solomon, I., and M. D. Shields. 1995. 审计中的判断与决策研究。载于会计与审计中的判断与决策研究，R. H. Ashton 和 A. H. Ashton 编，137–175。剑桥：剑桥大学出版社。 https://doi.org/10.1017/cbo9780511720420.008渣打集团。2025年。*2024*年年度报告。访问日期：2026年11月22日。 https://www.sc.com/en/uploads/sites/66/content/docs/standard-chartered-plc-full-year-2024-report.pdf

Stratopoulos, T. C., and V. X. Wang. 2025年。人工智能与会计研究：框架与议程。《国际会计信息系统杂志》56：100760。 https://doi.org/10.1016/j.accinf.2025.100760

Wilson, E. B. 1927. 概率推断、继承律与统计推断。美国统计学会期刊 22（158）：209–212。 https://doi.org/10.1080/01621459.1927.10502953

附录A：分配矩阵详情

表格A.1和A.2列举了25个会计工作单元示例，以及它们在CDAJ编码方案下的建议分配假设。这些任务是作者从审计、报告、税务和估值领域选取的示例。它们是对所命名工作单元的主导编码建议，并非对整个业务项目经从业者验证的分类。更广泛的任务必须使用第3.5节进行分解；混合单位应保留替代编码，而不是强行赋予标签。

表A.1。 25个示例性工作单元的CDAJ分配详情（第1部分，共2部分：任务1–12） .

工作单元示例	主导阶段 / 去-mand	建议分配
权衡再融资证据—据（持续经营）	评估 / 评价	人类主导；大语言模型作为对抗性审阅者
选择重要性水平基准	评估 / 评价	人类主导；大语言模型提供历史基准
汇总已批准的风险指标	评估 / 信息	迭代式：大语言模型生成结构化分析；人类发表意见
评估补偿性控制措施的充分性	评估 / 评价	人类主导；大语言模型作为证据聚合器
选择实质性程序程序	分析 / 评价	迭代式：大语言模型提出方案；人类选择
审计证据提取	数据 / 信息	大语言模型主导；人类审核相关性
主表编制	数据 / 信息	大语言模型主导，由人类核验
比率和趋势计算-tion	分析类 / 信息类	大语言模型主导；人类审查异常标记
审计备忘录草稿-ing	沟通/评价	人类主导；大语言模型协助格式编排
审计报告起草	沟通/评价	人类主导；大语言模型支持交叉引用
选择备抵叠lay	分析 / 评价	迭代式：大语言模型基线 +人类叠加调整
执行已批准的违约概率规范	分析类 / 信息类	大语言模型主导；人类审核模型适当性

注：第 1 部分，共 2 部分；第 2 部分续见表 A.2。

表 A.2。 25 个示例性工作单元的 CDAJ 分配详情（第 2 部分，共 2 部分：任务 13–25）。

工作单元示例	主导阶段 / 判-mand	建议分配
显著增加的信用风险（SICR）判定	评估 / 评价	人类主导；大语言模型质疑一致性
预期信用损失披露确定性起草	沟通/评价	人类主导；大语言模型检查内部一致性
报价公允价值计量项目	分析类 / 信息类	大语言模型主导，人工核验
不可观察输入的估值-值分析	分析 / 评价性	迭代式：大语言模型生成候选项；人类进行选择
公允价值选择分析	评估 / 评价	人类主导；大语言模型提供情景扇形图
收入确认政策选择（ASC 606 / IFRS 15)	评估 / 评价	人类主导；大语言模型支持合同挖掘
合同条款提取	数据 / 信息	大语言模型主导；人类负责整理
租赁分类测试	分析 / 评价	迭代式：大语言模型提出方案；人类签字批准
不确定的税务立场评估——估	评估 / 评价性	人类主导；大语言模型担任对抗性审阅者
填写经核实的税务——返回输入字段	数据 / 信息	大语言模型主导，由人类核验
ESG 披露数据收集tion	数据 / 信息	大语言模型主导；人类检查完备性
计算一项已获批准的ESG指标	分析类 / 信息类	大语言模型主导，人工核验
内部审计报告草稿-ing	沟通/评价	人类主导；大语言模型辅助起草

附录B：持续经营记录示例

本示例规定的是记录结构，而非最终定稿的案例、实际模型输出或观察到的参与者回应。在未来的案例中，证据文件将包含注明日期的预测、债务条款、违约通知以及管理层的再融资陈述。案例中必须说明适用的持续经营框架和相关评估期限；任何通用的二元答案都无法取代这一规范。

阶段A。 在接触人工智能建议之前，专业人士会记录初步评估、支持该评估的证据、其他解释以及置信度。记录会区分有文档佐证的再融资承诺与管理层关于将获得再融资的缺乏支持的预期。

阶段B。 示例性建议可以指出，违约通知是支持存在持续经营疑虑的证据，将主张与其来源关联起来，并说明豁免或具有约束力的融资协议为何可能改变结论。不确定性面板会指出缺少对融资安排的确认。本示意稿不赋予数值置信度。

阶段C。 专业人士记录一致意见或分歧，指出被遗漏或误解的证据，并解释推理上的任何偏差。诸如“管理层预计进行再融资”这样的响应需要进行来源评估；它并不会自动成为比正式违约通知更有力的证据。在C4中，可以要求模型在“再融资尚未作出承诺”这一反事实情境下重新考虑结论。任何回复都必须依据同一份证据文件进行审查。

阶段D。 最终记录保留初始估计、确切的建议及来源、核验响应、可选交流记录，以及专业人士经过论证的结论。若不参考支持性证据和独立制定的评分标准，初始估计的改变既不能视为成功，也不能视为失败。

附录C：前瞻性注册检查清单

在拟开展的人类效能评估开始之前，研究团队需要取得有文档记录的伦理审查决定、经批准的知情同意和数据处理安排，并确保招募切实可行。应单独开展一项试点，以完善案例文件、评分范围、标准化量表、模型刺激材料和平台。在招募验证性样本之前，应通过带时间戳的注册确定以下事项：

1. 两项验证性合并对比、结果定义、多重性处理程序以及联合结论的判定标准。2. 四种界面、共同基线阶段、分配和案例排序程序、冻结的初始建议、版本信息以及日志字段。3. 资格条件、固定的招募上限和结束日期、缺失数据处理、分配后标记、敏感性分析以及停止规则。4. 评审组流程、原始参考范围、独立试点量表、评分代码，以及证明选定的结果具有足够信息量的证据。

5. 分析实现、协方差和分布假设、收敛备用方案，以及由试点数据提供信息的运行特性模拟。

6. 确证性分析与探索性分析之间的区别，包括特定案例、依赖、时间和置信度调节分析。

后续修订内容应保留原始规范，记录日期和原因，并区分依据结果数据所作的变更。此检查清单仅为一项提议，并不代表已进行任何注册。

附录D：技术细节与未来评估规范

本附录将技术补充材料与预印本合并。文中提及的独立计算档案，是指作者准备的文件，并非声称已完成公共存储库存档。新的主要编码分析在第8节单独报告；原始计算档案不包含编码者私人响应。

D.1 证据状态

本补充材料记录了主稿件中报告的公开披露检索实验、合成记录/事件检查以及高斯规划计算。三份真实发行人年度报告提供了18个源衍生参考单元格，这些数据并非人类受试者结果，也不是经专业裁定的基准。合成输入和高斯抽样是独立的测试数据。研究未开展人类有效性实验或会计模型绩效基准测试。单独的主要编码练习见主稿件；其私人回答不属于本计算档案。主稿件还对两次返回中的全部18个源项目作了描述性报告，其中包括不利状态和不明确状态。单独发布的仅含汇总数据的版本保留了仅主要数据和源描述性数据的执行偏差；两者均非专家黄金标准验证，完整的原始人类响应计划也尚未完成。任务清单和形式状态/记录规范见主稿件。

D.2 汇总编码列联表

以下表格再现了已发布的聚合计数。行和列分别表示两个匿名响应档案，不含项目标识符或传记链接。它们比较的是响应类别，而不是将预测与黄金标签进行比较。全案例表保留了明确的 U 和非单一父标签。两个条件表仅保留 14 个共同为单一且轴代码非缺失、非 U 的案例；U 不计入这些条件系数。6 个不共同为单一的案例仍在全案例层中可见。未引入任何额外统计量、事后子集、裁定或个体响应。

原始完整计划读取器在数值现实性答案不符合分类模式时，以验证错误告终；本文不声称已获得完整计划结果。独立的主要范围记录于2026年11月22日08:33:52（协调世界时），早于08:34:40（协调世界时）的首次系数调用。独立的源范围记录于09:27:45（协调世界时），早于09:28:42（协调世界时）的汇总。公开执行说明保留了这些偏离情况，但未复现私密的错误输出行。四十个现实性答案及合并后的第五次比较仍未解决。原始回答既未重新编码，也未以共识取代。

表D.1。 全案例单元化（20 个案例）。

曲线行 / 列	单一	拆分	不确定	总计
单一	14	0	0	14
拆分	1	2	0	3
不确定	0	0	3	3
总计	15	2	3	20

表 D.2. 全案例阶段（20 个案例）。

曲线行 / 列	S1	S2	S3	S4	U	总计
S1	5	0	0	0	0	5
S2	0	4	0	0	0	4
S3	0	0	4	0	0	4
S4	0	0	0	4	0	4
U	0	0	0	0	3	3
总计	5	4	4	4	3	20

表 D.3。 全案例需求响应曲线（20 个案例）。

曲线行 / 列	E	I	M	U	总计
E	7	0	0	0	7
I	1	4	0	0	5
M	0	1	4	0	5
U	0	0	0	3	3
总计	8	5	4	3	20

表 D.4。 条件阶段（14 个可比案例）。

曲线行 / 列	S1	S2	S3	S4	总计
S1	5	0	0	0	5
S2	0	3	0	0	3
S3	0	0	3	0	3
S4	0	0	0	3	3
总计	5	3	3	3	14

表格 D.5. 条件 I/E/M 需求概况（14 个可比案例）。

曲线行 / 列	E	I	M	总计
E	7	0	0	7
I	1	4	0	5
M	0	1	1	2
总计	8	5	1	14

D.3 档案内容与执行

另行编制的可复现性档案包含 `run_diagnostics.py`、 `protocol_harness.py`、`verify_results.py`、分析和蒙特卡洛 JSON 结果，以及七个压缩的原始抽样 CSV 文件，

包含判定结果的合成轨迹测试样例，以及环境信息。三个脚本均仅使用 Python 标准库。解压档案，按以下顺序运行脚本，并保留其输出：

```
python3 run_diagnostics.py
python3 protocol_harness.py
python3 verify_results.py
```

脚本会自行定位所在目录，因此 Shell 工作目录不会影响输出位置。第一条命令会重新生成模拟和解析结果；第二条命令会重新生成合成轨迹结果；第三条命令会检查原始模拟文件。不同计算机上的环境元数据可能有所不同。已在记录的 Python 环境中验证带种子的数值可复现性和原始抽样可复现性；比较字节级哈希值时，应考虑跨版本伪随机数实现的变化。

输出 `legacy_design_audit.json` 包含辅助计算，用以说明为何非劣效性和调节效应需要不同的统计功效模型。它不属于当前验证性检验族，也不是人体研究注册。原始稿件快照、原始投稿材料和内部审稿人往来通信不包含在公开计算档案中。

D.4 公开披露模糊性诊断

`public_corpus/` 子目录提供官方 URL、PDF 哈希值、有限的来源卡片、收购和提取代码、评分前设计及修订内容、检索代码、派生分数和验证脚本。该目录不包含完整报告、完整提取文本或页面图像；目前尚未确立开放再分发许可。访问该档案时，请查阅其中的 README，了解固定的 `pypdf==6.21.4` 以及准确的重新获取/重建命令。主要的高斯分布和合成脚本仍仅使用标准库。

固定语料为汇丰控股、巴克莱和渣打银行的2024年年度报告，共含1,406页完整 PDF 页面，其中三页没有文本。18张卡片取自六张指定的预期信用损失和第三级表格，这些表格在评分前选定；重复金额仍予保留。卡片字段标明报告、PDF 页码/印刷页码、表格、行、列、期间、单位和印刷值。人工智能辅助了来源信息提取和视觉检查；这些标签未经人类会计师独立裁定。之后，一名人工请求者对照一份11页的来源核验资料包，确认了全部18个来源单元格和五项事后比较，未作更正。资料包哈希值、范围和确认记录均予保留；不声称核查者具备特定资历或采用了盲法，且冻结标签和分数均未改变。输入冻结和来源核查均有保存。之后进行的 `pdfplumber` 0.11.9交叉核查发现全部18个字面值和完全一致的单位字符串，并保留了八个多重出现情形。有限的坐标佐证和重建脚本随公开数据子目录一并提供；完整的页面词语输出仍保存在本地。这属于不同提取器的佐证，而非语义表格绑定或人工核验。

主要比较的是全字段 BM25 与采用相同 BM25 分数、但添加值/期间/单位词元存在性筛选的方案。两者使用相同的值、计量项目、期间和单位查询；发行人 ID 和来源定位符不是查询提示。固定的 BM25 参数为 $k_1 = 1.2$ 和 $b = 0.75$。仅按数字频率排序是次要的信息移除消融。分词器不会重建表格，并会剥除货币符号；因此，词元上下文相同并不能证明货币、行、列或会计计量项目正确。

两种全字段方法均在排名第一时找回 13/18 个指定来源，并在排名前三内找回 18/18 个。筛选仅将一个来源的排名从第三提升至第二，使平均倒数排名从 0.83333 提高至 0.84259，但 Recall@1 或 Recall@3 均未提升。没有来源被排除，也没有查询被弃答。所有卡片及全部 25,308 条卡片-页分数记录均予以保留。独立检查器根据保存的分数重建排名、并列情况、筛选后的分数一致性及汇总结果。重新获取全部三个源 PDF 后，其 PDF 哈希值及逐文档提取文本哈希值均与原先一致。该结果仅描述十八个聚类单元格，并非对会计领域的推断性推广，也不能作为 ACP 有效性的证据。

对四个不同页面上排名靠前的五个非原始候选项进行事后审计后发现，其中三项是重复余额，一项是附带计量范围说明的重复金额，还有一项是计量项目/列匹配错误。最后一项涉及巴克莱目标第一阶段抵押贷款备抵 3,600 万英镑，与检索到的总额 为 3,600 万英镑的调整项不匹配；该调整项的第一阶段金额为 700 万英镑。原有的溯源标签和 13/18 的结果均未更改。这项解释性审计并非新的语义准确性基准。完整的方法局限性及有限的来源定位证据见子目录。

D.5 定稿记录架构检查

`record_schema/` 子目录包含一个独立的定稿记录验证器、其字段规范、一份虚构的贷款组合样本、33 个完整测试样例以及记录在案的决策。请从档案根目录执行 `python3 record_schema/record_schema_validator.py` 。两个可接受的记录和 31 项刻意设置的无效修改均得到了预期判定；重新运行后，四项 JSON 输出均可复现。验证器要求提供实际的初始/最终字段、可解析的来源 ID 和假设 ID、匹配的来源内容哈希值、受跟踪的争议主张、每项所列例外恰有一项有文档记录的例外处置，以及模拟签署。

这些属于语法和引用完整性检查。它们无法证明来源真实性、语义支持、专业充分性、签名真实、执行顺序或平台已部署。来源定位符是一个字符串，而不是语义表格单元格解析器。样本中的 12,000 只是算术数值，并明确不代表完整的国际财务报告准则第9号备抵；其专业结论暂不作出。不要将 33 个记录测试样例和 26 个事件测试样例合并，进而声称已获得集成系统认证。

D.6 类型化契约扩展（版本二）

`record_schema_v2/` 目录保留了原始验证器的冻结副本，并补充了必需的会计背景和类型化来源绑定。运行 `python3record_schema_v2/record_schema_validator_v2.py`。随附的 README 和源代码规定了指引/范围状态、明确的货币和单位、文档/版本/页/表格/行/列/计量项目字段、打印值/规范化值、目录一致性以及数值/期间出现情况。若缺少专业依据，则标记为未解决或尚未确立，而非填入虚构权威依据或参考界限。

42 个测试样例中，40 个被结构性拒绝，2 个被结构性接受。其中一个接受结果是预期的伪造记录。另一个则在语义不匹配的情况下被有意放行：自由文本要求提供第一阶段减值金额，而共同变更后的结构化目标和绑定却指向调整总额。这一结果表明语义保证存在不足，并不代表会计结果正确。单独更改提供的绑定、但保持请求的参考值不变，会遭到拒绝。提供的目录是伪造的

，并被视为可信输入，用于一致性检查；它本身并非经过独立确认的会计事实。5 项 JSON 输出均可确定性地复现。原有的 33 个记录案例、26 个事件案例，以及公开的检索标签和结果均保持不变；本文不声称已对集成系统进行汇总验证。

D.7 统计算法

每个情景中，每组有 $n = 225$ 名参与者和 $m = 3$ 个观测值。这些是假设的设计值，而非观测到的样本量。给定相关性 ρ 和单案例方差乘数 v，独立抽样各组均值

$$\bar{Y}_2 \sim N(0, V), \quad \bar{Y}_3 \sim N(d_2, V), \quad \bar{Y}_4 \sim N(d_2 + d_3, V), \qquad V = \frac{v(1+2\rho)}{3n}.$$

在已知标准误 $\sqrt{2V}$ 的情况下，计算 $K2 = \bar{Y}_3 - \bar{Y}_2$ 和 $K3 = \bar{Y}_4 - \bar{Y}_3$。它们的协方差为 $-V$，相关性为 $-1/2$。双侧正态 p 值采用 Holm 双重检验规则进行调整：较小的 p 必须不超过 0.025；只有这样，较大的值才可与 0.05 进行比较。原始 CSV 列记录重复编号、两个估计值、两个 p 值以及两个拒绝指标。

对于每个情景，100,000 次独立重复试验均使用专用种子。七个情景依次为：全局零假设；一个零假设和一个 -0.40 增量；两个增量分别为 -0.20、-0.30 或 -0.40；两个 -0.30，增量为 $\rho = 0.60$；以及两个原始增量均为 -0.30、方差乘数为 2。默认 $\rho = 0.30$ 和 $v = 1$。种子依次为 20261122 至 20261128。Wilson 二项分布区间用于量化蒙特卡洛抽样不确定性，但不用于量化现实世界效应量或协方差假设的不确定性。

正文将固定阈值正态检验效能和 MDE 与联合 Holm 拒绝分别报告。该模拟未拟合混合模型、未估计方差、未生成真实的参与者级行为，也未模拟缺失情况。在原始效应固定时，方差均匀增加会降低标准化效应；这并非在保持标准化效应 d 不变的条件下进行的模拟。

D.8 合成验证器

该验证器是对选定事件排序和载荷完整性检查的参考实现。5 条可接受的轨迹和 21 条经修改的轨迹构成了 26 个测试夹具。输入以及预期/观察到的决策保存在 `synthetic_trace_results.json` 中。修改内容包括阶段缺失、与条件不符的事件、无效的置信度/一致意见字段、模型或冻结输出标识符不匹配、验证字段为空、完善消息未配对或不匹配，以及重复定稿或过早定稿。

这些检查针对编码字段和排序，而非语义质量。事件夹具模式不验证是否存在实际的初始估计值、证据列表、最终结论或签名，也不验证其具体内容。它无法确立源信息真实性、界面可见性、存储不可变性、用户身份、隐私或跨案例隔离。严格递增的时间戳是测试中的简化处理；在生产环境中，即使时间戳相同，事件也必须具备可靠的排序机制。C4 精炼交互是可选的；若已开始，则在此参考检查器中，必须先收到响应，才能定稿。

D.9 独立数值检查

独立的验证脚本会重新读取全部七个原始数据集。它会核验估计的对比相关性是否与-0.5一致、抽样方差是否与声明值一致，以及固定为 0.025 的拒绝率是否与分析正态

功率一致。容差界限在代码中指定。完整重跑在开发环境中复现了 16 个归档输出文件；归档的内部记录包含清单，这些清单不属于科学产出，也未作为公开源数据发布。已发布档案的校验和清单描述了其实际提供的文件。

D.10 前瞻性研究与已执行代码

档案中不包含已部署的参与者平台、最终证据文件、裁定后的评分范围、试点规模估计值、实际冻结的阶段B建议，也不包含拟合后的人类结果分析。执行前瞻性评估需要这些内容。候选的未来主要假设族恰好是跨案例合并的K2和K3，但须满足主论文中的测量与可行性门槛；其他组别比较以及所有非劣效性或调节效应主张，均按正文所述列为次要或探索性分析。

D.11 人工智能辅助

`gpt-6-astra-max`借助通过已配置服务调用的OpenAI模型，协助准备代码、构建公开来源卡片、进行视觉比较和检查。所有数值均由Python执行得出。DeepSeek-v4-pro用于评议修订后的手稿和资料包，而非生成观测数据或运行计算。对外提交前，仍需人类作者核验并批准。

D.12 候选人类评估议程：完整规范

D.12.1 状态、范围与估计目标

尚未开展任何人类有效性实验。以下内容是一份包含候选对比结构的评估议程，而非最终确定的验证性方案、注册研究或可执行设计。招募协议、伦理审查决定、报酬、完整的案例文件、从业者裁定、试点测量、软件部署以及冻结的模型刺激材料均尚未确立。这些是开展人类评估的前提，而不是计算结果所暗示的事实。

主要估计目标是：分配至所需验证检查点（C3 相对于 C2）以及提供可选优化的界面（C4 相对于 C3）所产生的意向性治疗效应，并在所指定的三个案例上进行等权平均。这些效应包括由所分配界面引起的任何额外时间、注意力和信息输入。它们无法识别纯粹的认知机制、实际遵从所产生的效应，也无法识别跨会计任务或大语言模型输出的一般效应。

D.12.2 案例、参与者与基线测量

拟议案例涵盖持续经营、《国际财务报告准则第9号》预期信用损失和第三层级公允价值。每个案例都需要一份限定范围的证据文件、明确的报告与审计框架及计量日期、与来源关联的模型建议，以及独立的参考评分量规。正文第4节中的三个示例规定了工作角色；它们并非已定稿的业务档案。未来样本应由具有相关会计经验的专业人士组成，并在筛选前明确资历和经验方面的资格条件。学生或注册会计师候选人不应被称为持证专业人士；学生重复研究应作为单独的研究人群和分析。

拟组建一个由五人组成的独立专家组，成员具备相关领域专业知识，且未参与案例撰写；该小组将评估

案例真实性和评分标准。每个案例至少设置 12 个预先编写的条目，用于评估证据、假设和论证依据，并由专家组成员依据锚定式五点评分量表独立评分。专家组还将分别给出合理的数值范围，并为每个界限提供书面理由。条目级一致性分析将报告条目—评分者矩阵中的序数型克里彭多夫 α 系数（Krippendorff 2011）、完整的分歧模式，以及逐一剔除评分者的敏感性分析。条目数量较少时，仅凭点估计无法确立信度。若初步 α 系数低于 0.80，或对评分界限存在重大且尚未解决的分歧，则须进行修订；该阈值是拟议的设计门槛，并非观测结果，也不是普遍适用且充分的标准。修订后，受影响条目将再次由成员独立评分，并保留原始评分。若分歧仍然存在，该案例将不纳入确证性研究。审议不得通过强求人为共识来掩盖合理解释之间真实存在的范围。此程序仅为提案；尚未招募专家组成员，也未收集评分。

所有参与者都将在随机分配之前且接触任何人工智能之前，完成全部三个案例的初始估计、证据清单和信心评分。这样可以获得真正的分配前测量值，供后续探索性调节效应分析使用。随后，将在各经验分层内平衡分配至四个组别。最终阶段的案例顺序将另行随机分配，并在描述性分析中予以记录和调整。在该阶段内，先前接触仍可能产生学习效应和残留效应；这会限制对后续案例效应的解读。该设计估计的是会话层面的工作流程效应，而非相互独立的单次试验效应。

D.12.3 条件与对比隔离

表 D.6。共同的分配前基线阶段之后拟采用的条件。

	C1	C2	C3	C4
模型输出	缺失	冻结输出	相同输出	相同输出
不确定性面板	缺失	Same 定性 专家组	同一专家组	同一专家组
所需验证	No	No	三个问题	相同的问题
模型质疑界面	缺失	缺失	缺失	可选优化
最终回答	人类	人类	人类	人类

C1 是无辅助的人类判断；C2 是人类在获取固定模型建议的情况下做出的判断。两者都不是“仅大语言模型”组。信息来源、初始估计、最终回答字段、一般任务说明以及最长会话时间必须保持一致。C3 增加了阶段C的提示词；C4 则进一步提供质疑和模型响应。额外对话是 C4 处理的一部分，因此不能将 K3 解读为独立于额外信息或时间的迭代。不存在单独的人工智能单独表现基准，也没有证据表明团队表现优于这两个智能体。

初始阶段B输出应针对每个案例生成一次，审查其是否适用于研究，随后冻结，并在所有人工智能辅助组中以完全相同的方式提供。必须记录模型标识符、提供商版本信息、提示词、解码设置、完整输出和哈希值。模型标识符或较低的温度设置并不能保证应用程序编程接口的行为具有确定性。交互式 C4回复需要完整记录提示词／回答日志；可复现性是指重建观察到的交流记录，而非承诺未来生成的内容完全相同。此处不假设进行微调或会计领域适配。

在知情同意、合同控制措施和机构数据治理审查允许之前，不得将参与者信息发送至第三方应用程序编程接口。

D.12.4 结果定义与准确性的含义

一个候选主要终点是参与者的估计值与经独立裁定的可辩护范围之间的标准化距离。对于案例 j，令 x_{ij} 为参与者的估计值，$[L_j, U_j]$ 为参考范围。定义

$$D_{ij} = \max\{L_j - x_{ij},\ 0,\ x_{ij} - U_j\}, \qquad Y_{ij} = D_{ij}/s_j,$$

其中，$s_j > 0$ 是在收集验证性结果之前，根据独立无辅助试点中 D 的标准差确定的固定尺度。Y越低，表示与参考范围越一致。若试点标准差为零或不稳定，则必须在注册前修订案例或尺度；不能在主样本中通过依赖数据的分母来补救。必须报告试点尺度的不确定性，并使用预先固定的其他领域尺度进行检验。

持续经营案例将要求对重大疑虑进行数值评估，并另行作出分类决策；预期信用损失案例将要求针对特定风险敞口或投资组合评估备抵，而非仅评估违约概率；公允价值案例将要求以指定货币单位给出值。参考范围必须附有记录在案的理由，包括允许采用的假设以及对相互冲突证据的处理方式。必须检查其宽度和稳定性，以排查地板效应。这些分数本身并非处于共同的 0–100 量表上。标准化使合并对比能够解释为以试点标准差为单位的等案例平均值；但这并不意味着不同领域的专业质量完全相同。

命题中的“准确性”在此操作性定义为与参考值的一致性，而非客观真相、监管合规性或事后预测成功。专家组可能出错，也可能存在合理分歧。证据完备性和论证质量的盲法评分应分别报告；不能将其与距离评分静默合并。在采用这一终点指标之前，未来的案例验证必须证明，数值任务和评分标准能够捕捉所关注的专业判断。

次要结果包括分类的持续经营决策、主动用时和总耗时、证据覆盖率、报告的信心程度以及检查点使用情况。对于探索性依赖分析，令 A_{ij} 和 B_j 表示初始的人类估计值和冻结模型估计值。报告向建议值的移动幅度（$x_{ij} - A_{ij}$）$\mathrm{sign}(B_j - A_{ij})$，以及与参考值的距离变化，$D(x_{ij}) - D(A_{ij})$。初始分歧为零的情况应单独报告。向建议值移动并伴随与参考值的距离增加，是潜在有害依赖的描述性指标；但其本身并不能证明存在锚定机制。对建议错误敏感性进行因果检验，需要随机化建议质量和更多刺激材料。

D.12.5 分析、多重性和缺失数据

如果下述测量门槛支持采用范围距离终点，一个候选的主要分析方案是使用重复测量模型分析标准化的与参考值的距离，其中组别、案例以及组别与案例交互项为固定效应，并纳入参与者随机截距。估计边际均值对三个案例赋予相同权重。限制性最大似然结合小样本自由度调整，是一种可供选择的实现方式，最终方案将在试点中确定并检验。由于三个目的性选取的案例不足以支持对会计任务总体进行推断，因此将案例设为固定效应。正文第 7 节中的规划计算，并非针对这一拟合模型进行的精确功效计算。

候选验证性检验族恰好包含两项双侧检验：K2=C3−C2 和 K3=C4−C3，二者均按案例等权汇总。Holm程序（Holm 1979）将检验族的错误率控制在 0.05。负的估计值表示新增元素更有利。报告未经调整和调整后的 p 值、估计对比、常规 95% 置信区间，以及针对两项比较的保守同时 97.5% 置信区间。不得将常规置信区间表述为族系置信区间。要检测出两项增量，必须两项调整后检验均在有利方向上拒绝原假设；单独的边际检验效能并不能证明联合结论具有足够的检验效能。

K1=C2−C1、K4=C4−C1、所有案例特定对比、初始置信度交互作用、依赖结果和时间结果均属于次要或探索性分析。本文不提出确认性非劣效性主张。尤其是，C4−C1 对比未达到统计显著，并不能证明人类水平表现已恢复。未来的非劣效性研究需要有专业依据的界值，并需单独进行样本量计算。置信度并非客观任务模糊性，而且每个领域仅有一个案例，无法识别一般性的认知需求交互作用。

主要分析将所有符合资格的随机分配参与者保留在其分配组别中。资格条件、同意和重复项筛查应在随机分配前完成。分配后出现的注意力失败、速度标记、外部协助的使用以及方案偏离，应按组别报告，并仅用于敏感性分析，不得悄然从意向性治疗人群中删除。还必须报告组别分配完整性和缺失结果。退出和部分完成可能导致无法观察到结果；意向性治疗描述的是估计目标，并不能消除对缺失数据假设的依赖。

该模型在随机缺失假设下使用可用观测值。预先规定的插补敏感性分析应纳入分配、案例、分配前估计值、置信度、经验和熟悉程度。会话时长是在分配后测量的，不得称为基线协变量。应针对预先规定的可信偏移开展模式混合敏感性分析，以评估偏离随机缺失假设的情况。必须检查地板效应、异方差性和模型收敛情况；参与者层级重抽样和其他预先固定的量表可用于稳健性检验。分析方案和备用方案必须根据独立试点数据最终确定，而不能依据有利的验证性结果。

D.12.6 招募规划与治理

计算考察了每组 225 名完整案例等效参与者、总计 900 名参与者的情景。假设固定招募上限为 1,200 名符合资格的随机分配参与者，即每组 300 名，按算术计算相当于完整病例信息损失 25%。这两个数字都不构成招募承诺，也不代表建议启动研究。该损失率是一项假设，而非经验估计值。达到固定上限或预先确定的截止日期时即停止入组；不得根据有利结果或实际各组表现有条件地继续入组。各组缺失情况对信息量的影响可能不同，因此招募 1,200 人并不能保证每组都有 225 个有效观测值。验证性注册前，试点必须为可行性和协方差假设提供依据。任何调整后的招募上限都必须在招募该样本之前确定。

试点招募前，必须先完成适用的机构伦理审查，或取得有文档记录的豁免认定。知情同意、公平补偿、退出权利、假名化、访问控制、数据保留和第三方模型处理，均须遵循经机构批准的程序。本文不声称已获得伦理批准、完成注册或取得招募权限。试点可用于完善材料；随后，应在验证性实地研究阶段开始前注册最终的验证性方案、刺激材料、评分规则、样本上限、分析代码和修订政策。概念性论文和拟议的注册清单并不能

替代该项注册。

D.12.7 计量与可行性门槛

候选终点指标有意将落在可辩护范围内的值视为在参考标准可接受性方面不可区分，因此不能据此对其实际质量进行排序。采用与评审组中点的距离虽能恢复范围内的差异，却会引入另一种假设——即中点比所有可辩护的边界值都更正确。我们不会默默采纳这一假设。因此，独立试点必须将范围距离终点指标与两个单独报告的敏感性终点指标进行比较：在有充分依据时，采用与评审组论证的中心估计值之间的绝对距离；以及采用有锚定标准的证据使用和推理序数评级。若地板比例过高、量表不稳定，或与独立评定的论证依据对应不佳，则范围距离不应作为主要终点指标。

在选择最终终点指标之前，研究团队应公布试点得分分布、零值比例、评审组边界敏感性和量表不确定性。随后，基于试点研究的模拟必须检验所选结果的分布、估计协方差、缺失情况及计划采用的模型。当前的高斯计算尚未通过这一门槛。只有在验证性注册之前，才可以在试点结束后改用其他主要终点指标；这需要重新进行功效计算，并按照该量表重新表述假设。

预先固定的建议需要单独进行适用性检查。独立评审员应在不参考后续参与者结果的情况下，将每项实质性建议主张编码为有依据、存在争议或缺乏依据。试点必须表明，案例中存在拟议检查点发挥作用的机会，同时也存在有意义的可能性，即该检查点会带来成本或促使参与者拒绝有用的建议。如果建议质量未经随机化，任何估计的 K2 或 K3 效应都仍取决于特定的质量组合。由此不能得出关于对建议错误的敏感性或一般锚定机制的因果主张。

在确定样本量之前，必须就专业领域最小重要改进、可接受的额外时间、目标联合统计功效和招募预算达成一致。可行性研究应记录符合条件者的招募率、两个案例阶段各自的完成负担、预期报酬，以及获取目标专业人群的途径。目前不假定能够获得稳定的参与渠道或预算。如果这些关卡无法支持切实可行的样本量，团队应重新设计评估，或将其报告为可行性研究，而不是把统计功效不足的研究称为验证性研究。这些尚未解决的决策反映了人工评估议程仍处于初步阶段。