流程重复、人效低：AI 流程自动化与员工工具普及

作者：SVSAC 硅谷战略顾问委员会

核心判断

AI 普及应从员工真正受困的重复工作开始，先把一条流程做顺，再谈全员提效。对企业而言，值得优先推进的是：按频次、耗时、差错和责任风险筛选场景，明确人工接管与复核规则，评估端到端周期而非单步速度。本篇结合公开资料与分析，讨论企业可以如何识别机会、组织能力并安排下一步投入。

关键词：生成式人工智能；流程自动化；劳动生产率；端到端流程；质量审核；组织实施

SVSAC - sv-sac.org

AI 普及应从员工真正受困的重复工作开始，先把一条流程做顺，再谈全员提效。

我们的建议是：按频次、耗时、差错和责任风险筛选场景，明确人工接管与复核规则，评估端到端周期而非单步速度。企业可以先据此梳理自己的产品、客户、能力与投入边界，把讨论推进到可以分工、可以验证的具体任务。

对管理层来说，这个议题值得放进经营议程。围绕“先做什么、凭什么做、做到哪一步再扩大投入”形成共识，往往比同时铺开多个方向更有推进价值。下文保留相关资料、比较和推演，供企业判断自身条件与行动顺序。

1. 核心问题：从“做得更快”到“经营结果改善”

企业讨论 AI 生产率时，常从一个易观察的变化开始：员工生成初稿、检索资料、归类请求或填写表单所需的分钟数下降。这项变化值得测量，却只是结果链的起点。员工若省下十分钟而必须额外花十二分钟核对事实、修复格式或向系统补录，单项任务提速并未减少实际劳动。若这些时间确实净省，但订单、客户问题或内部请求的到达量没有增加，组织得到的可能是闲置容量，而非同额现金节省。若需求随服务变快而上升，产量可能增加，单位成本却因新审核岗位、模型费用或更高服务标准而不降。由此，本篇不把“个体节省分钟”直接称为“人效提升”，而将其视为需要验证的中间结果。

核心核心问题是：AI 工具如何改变一项工作从进入队列到形成可接受交付的完整生产过程？分析需要同时回答三个子问题。第一，工具改变了哪个工作环节：起草、检索、分类、决策、交接，还是后续修改？第二，局部输出以何种错误和不确定性进入下一环节，谁承担审查与例外处理？第三，释放的时间最终转成了什么：更多合格产出、较短等待、较低加班、员工转向更高价值任务，或可验证的预算减少？只有在这三个层次有可追溯关系时，任务级收益才可能解释为经营层收益。

本篇所说的“AI 流程自动化”包括生成式助手、分类/检索系统和自动化规则对工作步骤的改变；“员工工具普及”强调名义可用、实际使用和稳定嵌入日常操作之间的差异。工具访问权限并非使用，使用也并非正确使用；正确完成单个任务也不等于流程成功。分析单位须随估计对象变化：单个任务可估算完成时间或盲评质量，员工–周可估产出率和加班，团队或流程实例可估从受理至完成的周期与返工，财务期间则考察实际现金支出或可避免成本。将这些单位混写，容易把局部百分比移植成企业整体生产率。

SVSAC | sv-sac.org

一个简单的逻辑链可用来说明差别。假设一个团队每天收到一百项请求，员工处理初稿、主管复核、再由业务部门确认。助手使初稿从平均二十分钟降到十二分钟，但主管每份需增加六分钟复核，且十分之一请求因不准确而退回重做。初稿工时下降并不足以推出总工时下降；还须知道退回率是否改变、主管复核是否原本就发生、返工由谁承担、队列是否跨过了瓶颈点。即使总工时下降，如果业务部门只能每天确认七十项，流程吞吐仍受该环节限制。这个例子只是待检验的机制示意，不是对任何企业的观察。

因此，本篇比较的不是“有无 AI”这个抽象标签，而是明确的工作包及其处理方式。工具版本、提示或规则、员工训练、任务分派、复核要求、权限边界和计量窗口共同构成干预。它们可能产生互补：标准化知识库使助手引用可查，培训使员工识别低置信答案；也可能产生替代或摩擦：过度自动草拟令复核者负担增加，或员工绕过系统后形成双重记录。可识别的结论应指向一个具体实施包，而不是泛称 AI 总有效或总无效。

2. 概念边界与结果变量

2.1 四类生产率结果不能互换

第一类是**任务效率**：对相同或可比任务，计算从开始到交付初稿所花的主动时间、等待时间或返工时间。时间起止要预先界定。工具打开时间不是完整劳动时间；任务被后台生成也不是零成本，因为描述问题、等待、阅读、验证与修正都可能由人承担。任务成功率必须与速度并列，否则把提前交出错误答案视为效率改善。

第二类是**合格产出率**：单位时间内形成并通过既定质量验收的交付数量。以付薪小时为分母才是按劳动时间的产出率；合格数除以分派任务数或流程实例数，则是固定窗口合格比例，不能都称每单位时间产出率。若采用“每个完成任务耗时”，未完成、撤销或超时任务会被排除，容易产生幸存者偏差。更可解释的做法是同时报告固定总体的完成率、质量合格率及预定窗口内合格交付量。质量评审需要在不知处理组别的条件下按既定标准执行；若质量由 AI 参与的同一团队自评，评分标准随干预变化便可能把表面改善误作实质改善。

第三类是**流程性能**：流程实例从触发到最终完成的日历时间、阶段等待、队列长度、例外率、一次通过率与返工路径。端到端周期由主动加工与等待共同构成。主动劳动时间下降而队列等待不变，可能降低成本却不缩短客户等待；反过来，队列被缓解即使总工时下降有限，也可能提高按期完成率。流程实例必须能

SVSAC | sv-sac.org

跨系统关联，且要记录暂停、撤回、重新打开、补充资料与转交，否则日志只看得到系统内的一段操作。

第四类是组织与财务结果：经质量调整的产出、实际预算、加班、外包支出、编制或单位服务成本。节省的员工小时只有在时间可被重新安排、产量需求存在且不需要等量新增审核时，才是可释放容量。它可能改善服务水平或腾出专业员工做复杂工作，却未必降低薪酬支出。若工资、排班、合同和人数均不变，名义小时下降不等于现金流出减少。反之，现金节省也可能来自削减服务或把工作转给客户，不代表真实生产效率提高。本篇把这类兑现结果单独定义，避免用一项小时率替代所有结果。

这些结果之间可以建立一个概念性桥接，但不能假定其参数已知。令每周期进入流程的任务数为 A，按预先确定标准完成并通过验收的数量为 Q，周期总付薪工时为 H，则合格产出劳动率可定义为 Q/H；端到端周期时长另以每个实例的触发–完成时间分布衡量；现金成本则需以实际工资、外包、软件和运营支出等账目计算。Q/H 提升与等待时间下降并非同一命题，Q/H 提升也不必然使现金成本下降。如果单位服务的需求增加、任务复杂度变化或标准提高，两个期间的简单比率甚至可能比较了不同产品。

2.2 流程边界与瓶颈转移

对实际流程进行建模，至少应把一项实例表示为“进入–执行–检查–例外分流–返工或批准–关闭”。各环节可并行、串行或条件触发。把平均阶段时间相加只在串行路径且没有并行与队列影响时才近似周期；现实里，多个任务共享主管、系统或专业人员，等待由到达率和服务能力共同决定。若某环节接近满负荷，即使上游平均处理更快，新增输出也可能积压在下游，甚至因批量审核而延长等待。AI 带来的负荷并非只通过处理时间影响系统，还可能改变到达量：当提案、报告或客户答复变得便宜，组织可能要求更多产出。

这类“需求诱发”是一项机制假说，不能仅凭产量上升判断效率。需要区分原有任务更快完成、未满足需求被处理、标准提高、重复产出增多，及团队把其他工作移入同一流程。任务数增长若伴随合格率不变、等待稳定或下降，可能体现容量释放；若例外和返工同步攀升，则可能是把审核成本推迟到后端。评估时应预先定义一个稳定的任务总体和服务边界，同时另行报告干预后新增任务，避免把流程扩张的分母变化掩盖掉。

角色间的相互依赖也决定测量单位。员工独立处理的标准件适合随机分配任务；共享资料库、主管审查、排班或互相学习的团队，则可能发生干预污染。把

SVSAC | sv-sac.org

同一主管下个体当成完全独立样本，可能低估共同冲击和知识传播。反之，若只在部门层级比较，试点部门可能同时获得更好的管理关注或更易任务。可行设计需要与实际分派单位对齐，记录资源共享边界，并清楚说明估计的是“提供工具的团队政策效果”还是“实际使用者的效果”。后者常受主动采用和任务选择影响，不能直接当作随机分配效果。

一个最小的流程账本应把每个实例的任务类别、进入时间、分派时间、开始与结束、工具调用、人工修改、复核结论、返工、责任人、最终验收及关闭状态连在一起。时间戳需区分自动系统时间和人类主动工作时间，防止把等待模型的墙钟时长都误作员工工时，也防止只计算点击时间而漏掉离线核验。对于质量指标，还应保存标准版本、评审人及盲态信息。否则，“平均处理时间下降”可能仅意味着简单任务占比增加，或者低质量样本更早被标记关闭。

3. 原始证据：任务实验与工作现场各自回答什么

3.1 专业写作的随机实验

Noy 与 Zhang 的 Science 论文保留为专业写作与生成式工具的待核文献线索。本篇没有取得支持正式版本具体样本、预注册、激励、阶段行为及方向结果的可定位原段，因此不将这些细节或正向效应纳入已核证据。正式论文与早稿应按各自版本取证，不能互相移植。本篇的企业流程问题仍需分别观察排队、跨职能复核、客户接受、返工和财务结果，不能由未核的短任务结果代答。

把潜在写作收益归于“初稿替代”还是“构思、编辑互补”，需要相应正式版本的阶段计时与行为证据，不能仅凭总时间下降判断。该文的这些原段未能重新取得，因此旧记录中关于直接提交或编辑行为的解释不作新的核对结论。机制上，若工作只要求可用初稿，替代重复起草可能提高任务生产率；若依赖组织特有事实、法律承诺或工程判断，核验劳动可能占主要成本。这是待验证的外推条件，而非该写作实验已经证明的企业流程机制。

该正式原文访问受限，本篇不依据旧记录确定其任务满意度、样本、评分或设计。即使未来核实短任务时间或评分，也须另检验其与客户验收、组织流程、长期满意度和技能积累的关系；这些是我们提出的外推问题，不是该研究已经核实的结果。[2]

3.2 客服工作中的现场部署

Brynjolfsson、Li 与 Raymond 的 QJE 正式论文研究一家财富 500 强企业的客服人员，使用 5,172 名代理人的工作数据。研究对象不是自动客服取代人员，而

SVSAC | sv-sac.org

是一个基于 GPT-3、实时向人工客服提供建议和内部文档链接的辅助系统；人员保留是否采纳或编辑建议的决定权。作者利用不同地点与团队逐步上线、资源限制导致的分批接入，分析每位员工的月度业绩。其首要产出指标是每小时解决问题数，估计平均提高约 15%。论文另分别考察单次客服聊天处理时间、每小时对话量与解决率，显示生产率变化不能简化为“每次对话少花几分钟”：并行处理能力与解决概率也可能变化。[1]

识别策略是分批部署的差分中的差分，纳入员工和月份固定效应，并控制任职月数；论文描述工具在同团队、同办公室内按员工逐步上线，且经理安排培训时间，因此分派并非纯粹随机。作者讨论潜在排序偏差，使用更稳健的分期处理估计量、事件研究及以员工所在公司、办公室地点及团队的最早接入日期作为个体接入日期的工具变量等分析。读者应把结果理解为这一企业、这一客服流程和当时模型实施条件下的现场证据，而非任意单位可复制的平均处理效应。具体识别仍依赖相应的平行趋势、无提前反应和分期处理条件；稳健性检查不能证明所有未观测混杂均已消失。[1]

论文报告的异质性提示了工具为何可能改变团队能力分布：经验较少、基线技能较低的员工平均获益更大，建议有助于转移经验丰富员工的做法。作者也报告工具支持学习与英语熟练度变化。重要的是，这不等于“组织可减少培训或用员工平均效率替代岗位能力”。相同生产率增幅在工作内容稳定、知识库可用、员工可拒绝建议、质量错误成本可控的客服环境中形成；换到须跨部门定责、更新频繁或错误代价很高的流程，复制条件不成立。指标 RPH 是客服吞吐代理，不能自动换算成企业现金节约，尤其当工资、需求、客服复杂度或服务质量目标没有同步纳入时。

3.3 同一工作流中的能力边界

Dell’Acqua 等人在 Organization Science 正式发表的研究，基于 BCG 咨询人员的预注册随机实验，758 名顾问被分配到 GPT-4、GPT-4 加简要使用概览或无 AI 条件。受试者分入互不重叠的任务臂，而非所有人都完成两类任务。设计包含：18 项研究者预先判断在当时模型能力范围内的咨询任务，以及一项刻意设置为模型能力范围外、需要人类引导的战略问题。前一组内，论文报告处理组完成任务数量提高约 12.2%、完成速度更快，并在特定评分指标上提高质量；后一项模型范围外任务中，AI 组正确解决的概率比对照组低 19 个百分点。这里的“19 个百分点”不能误写成相对下降 19%。该研究以同一职业和工作流里任务能力边界不均匀为核心，说明工作表面相近，并不意味着模型适用性相同。[3]

SVSAC | sv-sac.org

这项研究的贡献与边界都应保持清晰。研究将参与者随机分配至处理条件，任务由专业人士设计，并对输出进行评分；它比无对照的厂商试点更能支持工具访问的因果解释。可是它仍是限时、结构化、模拟真实工作内容的任务实验，不观察公司长期项目交付、多人接力、客户复购、人员预算或完整月度账本。文中18项“范围内”任务和一项范围外任务的组合，也不构成所有组织任务的代表性抽样。若将结果用于部署，首先应检验自家任务是否处于相似能力范围，不能把平均加速外推给未验证任务。

3.4 证据的交集与分歧

已核的客服与咨询证据支持较窄的判断：在特定任务或工作场景下，生成式AI可能提高部分劳动者的处理速度或合格产出，并且效果取决于任务、使用方式和人群。它们的设计与终点不同，不能平均成一个通用“AI生产率系数”。写作文献仍待核，未进入这一经验判断；BCG实验提供随机分配但受限于实验任务；QJE研究追踪实际现场业务，但由企业分批部署而非完全随机分派。对外部效度而言，这种差异不应被消除，而应成为后续设计的结构：先在具体流程里明确工作单元、标准、审核、资源与任务适用边界，再判断哪种既有证据最接近目标环境。

更广泛的解释还须考虑三种反例。其一，工具提高员工产出但审核人满负荷，组织总吞吐不变；其二，模型生成大量初稿却扩大差错与返工，合格产出率下降；其三，合格产出和吞吐均提高，但需求不足，腾出的工时并未转为现金节约。反过来，现金成本可能下降但服务质量恶化或员工负担上升，也不能简单称为全面生产率改进。每种情形需要独立指标和明确的价值判断，不能由一个处理时间均值解决。

4. 分析框架：把任务结果连接到可验收的流程结果

4.1 单项任务的产出账本

我们建议将一项可比工作包定义为包含输入、操作边界、输出规格和验收规则的最小单位。测量该单位的总人工时间时，应记录实际执行者与复核者在工作包上的全部主动分钟，并分开记录纯等待、阻塞与系统耗时。质量结果不应只按是否“有输出”计算，而至少包含一次验收通过、需要修改后通过、未通过或在观察期内未决。这样，任务端的初级量可写为：固定分派任务的完成比例、首次验收通过比例、最终验收通过比例、通过件的人时、以及从分派到最终验收的日历

SVSAC | sv-sac.org

时间。固定分派总体是关键，因为只计算已完成任务会把处理后未完成的困难件排除。

以一个虚构流程为例，假设团队分派 100 件可比任务。传统方法完成 90 件，其中 80 件首次通过，10 件返工后通过；AI 辅助组完成 96 件，其中 70 件首次通过、20 件返工后通过、6 件到观察窗结束仍未过。若只比较初稿平均时长，AI 组可能领先；若比较固定窗验收率，传统组为 90/100，AI 组为 90/100，但及时通过、返工和未决构成不同。仅看通过件工时会排除 6 个未决任务，造成“完成者更快”的偏差。此账本数字仅为演示口径，未代表实测结果。真正项目需事先定义观察窗、质量等级、重开规则与失联处理，已知截止窗未验收者在固定窗 Y 上记 0，另保留排队、补件等原因，不等于断定最终业务质量失败；只有窗口内是否验收本身未知者才作缺失及界限分析。

总人工时间也需要防止重复计量。若员工起草、主管抽审、专家复核和业务签核各花一段时间，应以每人实际投入的增量工时合计，并记录工作是否原本就存在。若审核人同时审核多项，按有依据的工时日志或预先定义的抽样测时分摊，不能将整个会议时长全计给某一单项；也不能因系统自动生成就把提示输入、核验、重写视为零。除总工时外，应分别报告员工类别工时，便于识别成本是否从低成本岗位转移到稀缺专业人员。

4.2 工作流结果、队列与资源边界

对于流程层评估，应预先画出状态转换和责任交接：进入、排队、执行、复核、返工、拒绝、关闭。每一个状态转换要定义日志来源和时钟口径。处理时间只计算活动中的工时；日历周期则从统一起点到终点，包括夜间、等待和外部依赖。两者回答不同管理问题。工作流实例的流程产出，是在固定窗口和既定验收规则下完成的合格实例数；可以同时报告期末积压、老化任务、重开次数及服务等级达成率。只报“平均时间”会让长尾积压被少数快速案例遮盖。

队列的机制可用简化关系理解：若平均到达速度接近某环节可处理速度，少量波动就会造成等待迅速增加；但真实流程是否满足某个排队模型的假设，需依据到达分布、服务时间和并发结构验证。本篇不把任何简单队列公式当作企业实证。测量上，先记录每环节到达、服务、等待和容量，再观察工具在哪个环节改变输入量或服务率。如果 AI 提高上游产出却让复核队列变长，瓶颈很可能只是迁移；如果复核者能借助同一工具，或流程重新设计减少非必要签核，端到端变化则可能不同。工具和流程设计的贡献要通过比较臂或分阶段实施区分，而不是事后归功于模型。

SVSAC | sv-sac.org

资源边界还需覆盖模型订阅、集成与维护、数据清理、培训、提示与模板维护、审核质量、异常升级、信息安全与合规工作。成本账本要将经常性现金支出、一次性实施投入、内部人员机会成本和已释放但未变现的容量分栏。工资成本只有在岗位、加班、外包或招聘计划实际改变时才成为可避免现金支出。若一个团队把节省的工时投入未完成的积压，组织可能获得真实产能和服务改善，但不是预算下降。反之，预算削减造成服务量减少，单位成本下降也不能代表产出效率提升。

4.3 比较设计和可证伪判断

可信的比较应将任务的可比性、处理分派、质量标准和观察期先行固定。对于独立、短周期任务，可在员工或任务层随机分派工具权限；同一员工跨处理任务时需预先考虑学习和残留效应，避免将后处理方法学会后仍把对照任务视作未处理。对于团队共同使用知识库、模板或主管审核的流程，应考虑团队或小组层随机分派、分阶段上线或配对比较，并把员工间信息传播列为可能的干扰。若无法随机，需选择可比的同期对照并公开识别假设、前趋势与任务构成变化；单纯前后比较不能排除季节、管理、人员经验和需求结构改变。

预注册分析应至少确定：固定分母下主要终点、合格标准、任务难度分层、复核者盲态、工具版本、意向处理效应、失访和未决处理、时间及质量权衡、成本口径、最小值得关注的改善范围，以及安全或质量停止规则。无真实基线时，不应虚构样本量或效应门槛；先用历史任务估计波动和聚类程度，再制定有功效的设计。主要估计先回答“提供工具这一政策包是否改进固定任务总体”，随后才考虑“实际使用工具者是否获益”的依从者效应。后者需要额外假设，不能将自愿使用者与未使用者直接比较。

本篇框架导出若干可反驳判断。若 AI 只缩短初稿时间，却没有增加窗口内合格交付且未减少等待，端到端能力改善的主张不成立；若首次通过下降、后续审核与返工工时上升，局部提速可能由质量成本抵消；若固定任务总体中合格吞吐上升、每件全角色投入下降且未决和长尾无恶化，则证据更符合流程生产率提升；若这些变化存在但现金支出不变，只能说产能或服务水平改善，不能宣称已实现现金节约。反证标准需在试验前确定，避免成功故事只挑选最受益的任务、团队或时间窗口。

SVSAC | sv-sac.org

5. 研究方法、范围与本篇结构

本篇采用定向原始研究综述与机制分析，而非估计跨行业平均效应。纳入来源优先为可核对的方法与原始结果：随机实验用于说明任务级处理效应和任务边界；现场部署用于说明实际采纳、运营过程及现场结果；原始作者版本与正式发表版本需分开登记。研究不把平台厂商宣传、模型榜单、新闻报道或机构调查摘要当作效果因果证据。对每项证据，分别记录研究人群、任务、处理分配、主要结果、时间窗口和外推限制。若只读摘要，就仅将摘要明确报告的内容作为事实，不推断未报告的方法细节。

为使综述回答机制问题，本篇以五段结果链组织材料：任务投入与工具调用；初稿或建议的质量与速度；检查、审核与返工；流程实例是否按期完成；人员、预算和服务结果是否兑现。贯穿各段的问题是处理效应有没有经过可观测中介传递，以及是否出现了新的约束。中介变量描述机制，未经独立识别时不当作因果解释。例如，工具使用强度与产出相关并不能说明使用导致产出，因为高能力员工可能更常使用、也更有效；审核时间变多也可能是任务难度不同，而非 AI 质量降低。

现有原始研究不能填补企业流程的全部证据缺口。QJE 客服研究提供实际系统上线与代理人业务数据，但来自一家企业和一个可观察到每小时解决数的场景；Noy 与 Zhang 文献在本篇仍是待核线索，不提供已核样本、设计或效果；Dell’Acqua 等人的顾问实验有明确的处理随机化和任务能力边界，但并非长期企业财务数据。已核证据适合界定有哪些条件性正向证据、有哪些需要测量的异质性，不能证明其他公司采用同类工具会获得相同比率，也不能推断岗位或组织规模会如何变化。[1][2][3]

后文将依次讨论工作单元与任务标准如何定义、不同流程结构中的瓶颈与质量外部性、员工采纳和工作重组、可复现的现场比较协议、资源与收益的分解，并在最后讨论替代解释、适用边界和研究限制。这个安排并不意味着本篇已经实施企业试验；实证设计是供未来研究或组织评估使用的可检验方案。后续案例如属构造分析，将标注“假设情境”；公开研究则按其真实样本、估计对象和发表版本介绍。全文不作成功保证，也不对减少岗位、提升薪酬或节省预算作超出证据的结论。

SVSAC | sv-sac.org

6. 工作结构决定局部提速能否传递

6.1 任务不是同质的“重复步骤”

把工作说成“重复”往往把不同生产逻辑混在一起。第一类是规则稳定、输入结构化、例外可枚举的事务步骤，例如字段提取、格式核验、重复记录筛查。其困难不一定在单次操作，而在大量边界异常、来源质量和下游责任。第二类是信息组织与起草：从多个来源找材料、按模板归纳、生成初稿。输出通常可读，但“表面完整”不等于事实正确，组织语境和来源适用性需要复核。第三类是判断与承诺：确定是否满足政策、选取行动、对客户给出交期或责任承诺。此类任务既依赖未明说的上下文，也可能产生不可逆后果。把三者笼统称作“流程自动化”，容易以第一类的规则稳定性推断第三类也可安全直通。

任务结构应沿至少五个维度描述：输入规范程度；处理规则在观察期内是否稳定；案例间是否存在实质异质性；输出的可观察与可验证程度；错误后果及修复成本。对常规事务，单位成本可能由处理速度主导；对起草任务，返工和来源核对重要；对判断任务，错误责任、申诉和异常处置可能远大于执行分钟。还有一个时间维度：规则或数据源更替频繁时，今天正确的自动流程可能在下次字段改版后静默失效。若评估仅抽取当前成功样本，可能错过这种漂移风险。

这些维度相互作用。结构化输入并不保证规则稳定；例如表格列固定，但“有效客户”定义随退款或重复账户规则而改变。可读输出并不意味着可验证；一段貌似合乎政策的回答可能引用已过期条款。规则易编码也不意味着自动执行合理；将提取结果写入下游系统可能改变付款或权限。反过来，高度依赖判断的流程也可能有可自动化的子步骤，例如将证据按来源归档，但不把最终决定交给模型。因此，合适的对象不是职位或部门整体，而是带有边界、责任和验收的工作包。

流程映射还要把“技术可执行”与“组织可接受”分开。某一步能被界面机器人点击，或模型能生成貌似合格的文本，只说明输出在技术上可形成。组织是否允许该输出触发付款、拒绝、对外承诺、工单关闭或客户通知，是授权与责任问题。技术完成与制度完成之间的落差会生成隐藏步骤：员工先在 AI 工具完成，再复制到正式系统；复核者另开资料核对；员工在私下表格记下错误以避免正式系统遗漏。若只分析软件日志，隐藏劳动会被系统性漏掉。

6.2 标准化收益与例外成本的联合分布

低频例外并不必然对总收益无关紧要。设流程中标准件占比为 p，标准件单件净省时间为 s，例外件占比为 $1-p$，单件新增处理时间为 e，额外抽审成本按每件平均 r 计，则总节省的工时可写成 $N[p\cdot s-(1-p)\cdot e-r]$。这个表达式只是线性账

SVSAC | sv-sac.org

本，隐含每类平均成本稳定、无排队互动、无错误串联且数量 N 固定。它可用于暴露一个常见误读：对大量标准件小幅提速，可能被少量高成本例外吞掉；也可能反过来，即使例外率略升，标准件容量增加仍足以改善总体交付。实际决策还要把错误严重性、风险厌恶和尾部损失单独考虑，不能只优化平均分钟。

例如，一个受理团队的常见请求可以由固定资料自动归类，但少量请求涉及资料缺失、账户合并或规则冲突。若系统为提高“自动处理率”而把边界案例一并自动关闭，短期处理量会变好看，后续客户申诉和账目纠正却可能集中出现。更合理的测量同时记录标准件和例外件的基线数量、分类转移、误分率、人工接管与最终处理；抽样质量审核要覆盖自动关闭项目，而不能只看模型主动拒绝的难例。例外抽样比例还须与发生率和后果严重性相匹配。否则，表面稀少的错放事件无法仅靠总准确率揭示。

例外的“发现位置”同样重要。上游发现一份缺字段资料，可能只需发回补齐；若自动化已将其写入客户账户、生成应付金额、安排发货，再发现错误则修复成本高得多。相同预测正确率因此可能对应完全不同的净收益。流程可将动作分为可逆与不可逆、内部与外部、可延期与立即触发，并在每类动作中记录发现延迟和恢复耗时。把错误发现率与错误发生率分开，有助于区分模型质量、流程防护与审核能力；单独提升某个“模型准确率”无法说明哪一环节减少了真实损失。

7. 自动化权限不是单一开关

7.1 从建议到执行的权限阶梯

工具可以只提供候选答案，也可以填充草稿、提交待审批动作、经授权写入系统，或在特定条件下完全自动闭环。这个阶梯改变了错误暴露和人工负担。建议型工具通常留下采纳/拒绝的人工环节，但若建议大量出现，员工可能因注意力负担而默认采纳；待审批执行仍保留一道正式签核，但签核可能流于快速点击；自动执行减少显式人工步骤，却把验证责任前移到规则设计、监控和异常回滚。使用界面中“人还在环路里”不等于存在有效监督，监督质量取决于信息是否可检查、复核时间是否充分、不同意见是否能安全提出。

更可操作的做法是按行动风险定义授权，而非按模型置信度统一设一个阈值。举例而言，系统可自动整理可撤销的内部草稿；对低金额、可冲回且规则明确的事务生成待审批变更；对外部客户承诺、付款、权限开通或安全影响事项必须经具名审批；对输入冲突、数据越界或不在验证范围的项目直接拒答或转人

SVSAC | sv-sac.org

工。这是作者提出的设计框架，不是法律认证标准。动作级分层应由业务责任人确认可接受后果，并形成可追溯的审批链。

权限扩大带来的预期净收益应包含错误暴露概率。对动作 a，可定义每期预期损失项为“该动作错误概率×错误后果”，并与减少的人工处理、等待和复核成本比较。这个乘法要求概率与损失能合理估计；罕见严重事件、依赖关系和声誉成本通常不适合简单平均值。更重要的是，有些边界是约束而非可交易的成本：即使平均节省更高，也不应让未经授权的自动动作突破职责或外部承诺。风险评估要能说明什么情况下模型不应做决定，而不只是报告一个平均准确率。

7.2 观察性“使用率”与实际因果效果

员工使用数据很容易被误解。高使用者可能本来就更擅长拆解任务，也可能面对更多适合工具的工作；低使用者可能处理高风险案例，或因系统不适配而转用其他办法。因而，使用频率和绩效的相关性混合了工具选择、任务配置与人的能力。将高使用组和低使用组直接比较，不能回答若同一类工作被提供工具会怎样。随机提供权限的意向处理效应能回答“提供工具的政策包”的平均效果；如果有人不使用或对照组使用替代工具，实际使用效应需要额外识别条件，不能用简单按使用筛选来替代。

员工采纳并非单纯的态度变量。工具嵌入需要账号、权限、知识源、界面位置、任务模板、培训和适当的时间安排。一个人可能在小任务上试用，却在正式流程完全不敢使用；也可能将输出复制到未授权渠道，出现安全与版本控制问题。采纳的可观察层次至少有可用权限、首次试用、重复使用、任务类型覆盖、建议采纳/改写/拒绝、以及最终流程状态。每个指标都应与任务类别和工具版本相连。仅统计打开率、prompt 数或 tokens 不能衡量合格工作产出。

对采纳机制的因果判断尤其要防止“成功者更常使用”的反向因果。某员工被经理挑选参加培训，常同时得到额外辅导和更适配任务；上线后团队又会分享提示和技巧。可以把工具许可、培训、知识库和流程变更设为不同处理臂，以识别组件或互补关系；若样本和团队规模不足以分臂，结论就只能归于整套实施包。观察记录要覆盖未使用原因和替代做法，且不能因员工没有采纳便从分母移除。把实际使用者效果作为补充时，应披露依从、任务选择与干扰假设。

7.3 人机协作中的验证成本

模型输出具有说服力时，复核者的难点不只是花多少时间，还包括能否发现错误。审核者面对大量常见正确输出，可能形成自动化偏误；但为了防止偶发错误而逐字从头核对，则会使节省时间消失。更高效的复核设计可提供与输出逐项

SVSAC | sv-sac.org

对应的源证据、数据血缘、变化差异和不确定事项，使审核者检验决定性断言，而非重新执行全部工作。若模型无法给出可靠溯源，某些任务仍可能需要全量复核。这一机制的成效不能只看平均审核时间，还要同时测审核后漏错、误拒、升级和返工。

训练、提示说明或额外监督可能改变输出质量。Dell'Acqua 等人的随机咨询实验比较了 GPT-4 与 GPT-4 加使用概览，说明"是否有工具"与"有没有适当操作指导"可以分别成为处理条件；论文报告两种 AI 条件在任务表现中均与对照不同，但加概览并非对所有结果都显示显著优于单独 GPT-4。[3] 组织评估可用多臂设计识别简短培训是否提升任务匹配、核验行为或边界识别，而不能把"多培训总会更有效"当作先验事实。培训也有机会成本，且仅增加自信而未改善判断时，可能加大错误采纳。

8. 生产结构的比较：单件处理、批量流转与专业复核

8.1 单件流转的重复事务

假设一个共享服务组每周处理一批内部费用单。系统可自动读取金额、供应商和成本中心，员工核对票据后提交。这个情境的关键不是模型能否识别字符，而是企业定义的匹配规则能否把撤销、重复单、币种转换和缺失审批正确处理。对于规则稳定、可追溯来源且单据错误易撤回的项目，自动填表可能减少录入；若供应商主档变更频繁，或同一费用存在多种可接受的会计处理，则不完整上下文会把"识别准确"误当作"会计判定正确"。

比较时可以以固定期间、固定类型和来源完备度分层的单据为单位，随机分派由传统界面处理或提供 AI 填表建议；接收人和最终审批人保持既定权限。终点包括：字段级正确率、一次通过、返工、错付/重复支付、每角色工时、结案周期和未决余额。若只对已完成单据抽样，可能漏掉拒绝处理或等待审批的难单。若 AI 组将更多不确定单据拒答，完成量下降可能是合理的风险控制，而非生产力失败；但需要同时报告未处理率与风险，避免把拒绝当作成功自动化。

8.2 批量起草与审核队列

再考虑每周数百份内部运营报告需要汇总多个系统数据并形成摘要。工具让每位分析员更快生成初稿，然而所有报告必须由领域负责人确认数字和措辞。此时平均任务工时可能由写作环节转向核验环节。若复核集中在少数专家，批量输出反而可能把他们的队列推高。快速生成还可能诱发"每个团队都要求周报"的需求增长，使原来只做月报的流程扩展为周报，员工总工作量不降反升。

SVSAC | sv-sac.org

用于分析的单位应把报告任务与其源数据快照锁定，避免输入版本变化混入模型效果；以报告实例为单位记录生成、事实核验、修改、批准、发布和撤回，并保存报告周期。评价需有独立盲评，判断摘要是否准确表达数据，而不是比较文字流畅度；对于用来决策的报告，还要跟踪是否被使用及其决策后果，但“报告被打开”不等于决策改善。新增报告数量应与固定原有报告篮子分开分析，避免效率提升通过服务范围膨胀被隐藏，也避免将新增任务误当成同一任务集的产能提升。

8.3 专业判断与低容错审批

第三类是合规、风控或技术审批：分析员汇集证据、适用规则、提出意见，最终由授权人员决定。AI 可协助定位旧案例、抽取事实、生成核查清单，但自动给出“通过/拒绝”会改变责任、上诉、记录与后续追踪。此类工作总流程常有不同等级：机械核对、原则适用、例外裁量和外部承诺。对前两类可设计可验证的助手；对例外裁量则需明确何种证据必须由具名人员判断，模型给出结论后是否会锚定人的判断，以及人员是否有足够时间看原始来源。

评估单位应按申请或审批案卷而非模型回答条数。无 AI 组和 AI 组都用相同的政策版本和最终验收标准；盲评者检查事实遗漏、规则适用、正确拒绝、申诉改判及升级时间。若高严重度事件罕见，普通小样本的“没有发现事故”不能说明风险为零，应扩大观察期、采用后续事件追踪并报告不确定性。不能为了统计方便随机撤掉法律、安全或客户要求的审批。可随机比较的是辅助手段、分流顺序或培训包，而非越权执行。

三种结构的核心差异在于错误何时变得可见、谁能纠正以及修正成本是否可逆。低风险单件字段错误可能局部修复；一份错误报告可能传到多个管理决策；一项错误审批则可能触发外部损失。把平均效率改善与高严重度损失压成同一个“净效益分”会藏掉选择责任。组织可以报告多个并列指标或约束边界，但边界阈值应由实际责任人根据已知后果和适用规则设定，不应由研究者编造为通用行业数值。

8.4 构造容量账本：单位工时下降也可能伴随吞吐下降

为使瓶颈转移可以复算，构造一个每周稳定收到 80 件的事务流程。基线中，事务员每件操作 12 分钟，复核员每件 3 分钟；事务员每周可用 20 小时，复核员可用 8 小时。故每周事务员投入为 80×12=960 分钟，即 16 小时；复核为 80×3=240 分钟，即 4 小时。两个环节都有容量余量，80 件均可在周内通过。

SVSAC | sv-sac.org

现在假设 AI 把事务员的例行处理降到每件 4 分钟，但将每件复核要求提高到 6 分钟，且 10%的案件需额外 12 分钟查证。若仍要求 80 件全在本周完成，事务员投入为 80×4=320 分钟（5 小时 20 分）；复核员投入为 80×6+8×12=576 分钟（9 小时 36 分）。总人工从 20 小时降至 896 分钟，即 14 小时 56 分，净省约 5 小时 4 分；然而复核需要超过其 8 小时可用容量 96 分钟。若不加班且病例结构近似稳定，复核的期望处理需求为每案 7.2 分钟，8 小时容量只能支撑约 480÷7.2=66.7 案，约 13.3 件积压到下周。这里的平均分件仅是流量容量计算，不表示实物任务可以分数完成；实际周吞吐取决于离散任务、到达顺序和例外聚集。

因此，同一个构造账本同时允许“每个通过件总劳动更低”和“当周合格吞吐更低”成立。若主管加班完成 80 件，现金是否上升取决于加班支付和固定薪酬安排；若把例外交给另一名专家，瓶颈会转至专家；若复核通过辅助工具也缩短到 3 分钟且例外没有增加，结果又会不同。该例的用途是要求在试验中记录环节容量、例外分布与队列，而不是证明 AI 必然造成瓶颈。它还说明比较时应并列报告每件劳动率、周内完成量、期末积压和加班，单一“平均工时节省”会漏掉经营上最重要的约束。

9. 反向证据与效应异质性：工具不保证加速

正向实验结果不应构成筛选文献的规则。METR 针对早期 2025 年 AI 工具开展过开发者任务随机实验，16 名经验丰富的开源软件开发者在熟悉的成熟代码库中提交 246 项真实问题与开发任务；任务层随机分派为允许或不允许 AI，参与者记录屏幕并自报完成时间。作者报告该研究条件下允许使用 AI 的任务平均耗时增加 19%，而参与者事前预测和事后主观判断倾向认为 AI 提速。作者也明确警告，这一人群、工具时期、任务选择和代码仓库不代表多数软件工作，更不代表其他职业。[4]

这个结果并非 AI 普遍拖慢工作的一般证明。它提示，资深开发者对熟悉代码库已有高效心智模型，工具输出带来的检验、整合、运行测试和解释成本可能大于节省的键入；对刚接触领域的初级员工或新代码库，替代机制可能不同。不同任务成功标准也会改变方向：如果“完成”意味着代码必须通过风格、测试和文档审查，模型展示能生成代码并不能代表合格提交。METR 强调自身任务是现实 PR 并与模型基准及用户轶事的估计对象有别，这为本次分析的关键论点提供了反例：采用率、主观有用性、能力基准得分与净劳动生产率不是同一构念。[4]

SVSAC | sv-sac.org

METR 在 2026 年发表的后续作者更新更进一步指出，后期研究信号可能不可靠：工具普及后，一些开发者不愿在实验中被分配到无 AI 任务，部分任务在随机化前就不提交；开发者也报告随机到无 AI 后较少完成任务。与此同时，报酬由每小时 150 美元降到 50 美元，多 agent 并行使单任务人时难以计量。作者据此认为选择会影响估计，但其对偏差方向和规模的判断依赖参与者反馈，更新并非一个无选择偏差的校正估计。[5] 该更新的价值是揭示实验设计面临的实际难题，而不是提供一个可以替代原估计的后期通用效应。

从方法上看，这形成了一个非显然的矛盾：AI 越普及，使用工具与不用工具的随机比较可能越难执行；而员工越认为工具不可或缺，任务和参与者在随机前的选择就越受预期收益影响。将任务先固定再随机，可以减少任务选择偏差，却可能降低任务代表性；允许员工自己选真实任务更贴近工作，却需要独立记录全部候选、被排除理由和拒绝参与者。随机后退出、未完成和额外并行工作必须进入分母与时间口径，而不能只留可完成任务。METR 提出短时强制任务、开发者层随机、固定任务基准及观察数据等后续方向，各自也有权衡：短任务内在效度较高但外推较弱；开发者层随机容易污染少但样本需求高、招募选择更重；日志观察覆盖实际使用但处理采用内生。[5]

9.1 何时对照任务不再是对照

工具知识可在员工之间传播。对照组可能从同事获取提示、模板或复核建议；控制任务因此并非“没有 AI 知识”。个体层随机估计的是当前共享环境下的许可差异，未必等于全团队上线效果；正向知识外溢可能缩小组间差异，共享审核拥堵或负向外溢则可能改变其他方向，不能预定偏差必向下。反过来，团队层分派可能因为团队经理、客群、排班或领导关注差异产生偏差。研究报告必须明确主要估计量：同一时期提供使用权限相对不提供的员工层效应，还是团队整套上线政策的差异；不能在看到数据后将两种含义切换。

对共享资源更合适的是划定干扰簇：员工共享主管、知识库、工作台、培训或任务队列时，若目标是隔离团队政策，优先以可隔离的簇为单位分派并追踪跨簇接触；簇内个体分派则须另定义包含外溢的估计对象和相应设计，不能当作已隔离。若每个簇数量很少，估计精度受限，不能以任务条数很多为由忽略簇相关性。可采用分阶段上线，在每个阶段保留同期尚未上线簇，记录上线准备和培训，事先确定时间固定效应和比较窗口；但若管理者将最容易受益的簇先上线，仍须在设计中处理选择。分阶段比较并不因“轮流上线”自动成为随机试验。

SVSAC | sv-sac.org

9.2 时间、质量和任务组合的三角验证

当计时来自员工回忆或简短日记，AI 生成、等待模型、后台 agent 运行与员工切换任务之间会混淆。较稳健的测量可以把系统时间戳、抽样观察、任务记录与付薪工时对齐，并明确并行工作中每个人的主动时长、墙钟时长、机器运行时长和多任务重叠。对于任务级随机实验，不能把同时运行的 agent 时长当成人力节省，也不能把人等待 agent 时全部算作无产出：等待期间员工可能做其他任务，但这要求按队列与资源整体测量。

质量评估需跨越工具当下所生成文本或代码。任务输出可用盲评标准、客观验证、返工率和一段延迟后的错误暴露进行验证；如不同工作类型需要不同标尺，应报告评审者一致性和具体维度。团队可把高风险事实错误、延迟、客户重开、代码回退等列为独立结果，而非用一个容易受流畅性影响的总体分数。即使评审质量提升，也需审查样本构成：AI 组若更愿意接收容易任务或放弃高难任务，已完成工作的平均质量会提高，但固定总体上的合格产出可能没有改善。

10. 将工具能力与员工优势匹配，而非寻找“最易自动化职位”

10.1 任务频率、员工经验与知识可用性

常见任务未必是 AI 最能增加员工生产率的任务。若员工早已熟练处理密码重置、常规退款或固定格式报告，模型虽能快速生成建议，却可能只替代员工已经很快完成的步骤。另一方面，低频但有充分内部知识库支持的问题，可能让缺少经验的员工避免长时间搜索；过于罕见且训练或资料不足的任务，则可能连建议都不生成或输出不可靠。这种相对能力机制不同于模型在孤立样本上的绝对准确率：工具的边际价值取决于模型与人各自的基线能力、可用信息、验证成本和任务发生频率。

QJE 客服现场研究发现，AI 的时长收益在问题常见程度上呈非单调关系：系统在最常见问题上的处理时间改善不最大，在中度稀有主题上的改善更大，而最稀有问题因可用训练资料不足又出现较小收益；当资料不足时系统不提供建议。[1] 这一结果支持的是特定客服系统、支持主题和人类能力组合下的关系，不是“中度稀有任务普遍最适合自动化”的通用规则。它提示企业应同时测量员工的先验熟练度和系统资料覆盖，避免仅按任务频次或模型得分排优先级。

任务筛选可采用二维定位：一轴是人员当前独立处理能力，另一轴是 AI 输出在当地数据与规则下的可验证程度。人员熟练且 AI 容易验证时，收益可能表现为减少低价值步骤或提高并行处理；人员不熟练但 AI 有充分依据时，收益可

SVSAC | sv-sac.org

能通过快速辅导或减少搜索实现；人员不熟练且系统缺乏依据时，风险最高。人员能力高但 AI 难核验时，系统即便生成得快，也可能增加警惕性复核。二维图并非在试点前就能精确分类，任务位置应通过盲评、错误与工作日志逐步估计，并将新例外、制度变更或工具版本更新视为重新验证的触发条件。

10.2 互补与替代可以同时发生

“增强员工”与“替代劳动”不一定互斥。工具可能替代最初起草，却增强员工在复杂沟通中的表达；也可能缩短新人摸索时间，同时使高级员工减少投入原创解决方案的时间。一个任务的组成环节若方向相反，仅报总平均难以看清组织效果。将过程拆成信息查找、初稿生成、事实核对、判断、沟通与修正，逐项记录谁做、用了多久、输出质量如何，才能识别替代发生在哪里。真正的劳动替代需要看该任务人工需求是否减少、工作是否从该角色移出以及其他岗位是否接回；不应从模型生成占比推断岗位减少。

在 QJE 研究里，客服助手的建议采纳率约为 38%，且不同员工差异较大；作者发现初始采纳更高的员工估计生产率回报也更高，但明确承认这可能由选择或“对某些人回报更大”造成。[1] 因而，采纳率既是实施指标，也是内生结果：更适配的建议更容易被采纳，能力较强的员工也可能更善于判断何时采纳。直接回归产出对采纳率，不能证明强制提高采纳率会提高业绩，更不能以“团队采纳率未达标”为由惩罚合理拒绝。

组织可能希望通过管理目标鼓励采用，但指标会改变行为。若只按调用量、接受率或每小时件数考核，员工可能接受不适配建议、将复杂请求拆成更多短工单，或把难例转交他人；若单纯惩罚错误，则可能不愿尝试边界内的新方法。这里是需实际检验的激励反应假说，不是跨企业既定事实。试点评价应把采用率与质量、拒绝理由、任务复杂度和人员负担并列，确保不把“拒绝不可靠建议”视为不配合。

11. 采用过程、学习与过度依赖的双向机制

11.1 从可用到稳定实践的多步转化

真实采用不是单一开关，而是连续转化：具备权限、知道如何访问、愿意在适当任务上尝试、能够判断输出边界、能将其接入正式流程，并持续在质量可接受时使用。任何一步都可能中断。账号已开放但入口难找，属于可用性阻塞；员工尝试后遇到错误后无修复路径，属于反馈阻塞；输出可以用但不符合审批格式，属于流程接口阻塞；员工理解其边界却因没有时间核验而不采纳，属于容量

SVSAC | sv-sac.org

阻塞。把全部障碍归因于“员工抗拒变化”，会掩盖系统设计缺陷与合法风险顾虑。

评估采用漏斗时，应对固定、合格的任务集合记录每一阶段的分子和分母，例如：符合规则的任务、工具可用任务、收到建议、员工尝试、建议被修改/采纳/拒绝、独立复核通过、最终完成。不同阶段不能沿用同一个总体分母。若某类任务系统从不返回建议，不能把它列入“建议采纳率”的分母；若管理者只把容易任务配给工具，完成者中的使用率会高估普遍覆盖能力。建议采纳率也不能替代工具可用率和任务覆盖率。

QJE 论文将 adherence 定义为系统实际给出建议时员工直接复制或输入高度相似内容的比例，排除了系统未给建议的消息；其图 VI 按员工–月计算采纳率分布，并用该员工–月建议数的对数加权，报告均值 38%、四分位区间 23%至 50%；不是未经加权的所有建议合并比例。[1] 这说明在一个现场部署里，员工经常会忽略建议，而研究仍发现提供助手的总体效应。该比例不是“38%的客服使用 AI”或“38%的全部对话被自动化”，也不能外推到其他工具。对企业自己的指标，定义需要说明是否计入编辑后采用、未展示建议的任务、重复调用和错误撤回。

11.2 学习证据及其替代解释

工具可能同时是执行器与反馈媒介。若助手展示一个可验证的诊断路径，员工尝试并观察客户响应，经验可能留存在员工身上；若只把答案复制到系统，短期完成更快但技能不一定增长。QJE 研究利用 AI 服务发生的系统中断观察曾接触助手的员工在缺少建议时表现，并报告较高初始采纳员工在中断期仍有处理时间改善；作者将其视为与工作学习一致的证据，同时指出中断稀少、估计较噪，遇到中断的对话构成可能与非中断期间不同。[1] 这支持一个有吸引力的机制，却不等于随机切断助手后的干净学习实验：员工中断经历并非由研究者随机安排，且主要结果在更细颗粒度上是对话时长而非问题最终解决。

学习测量应区分“工具在线时更快”与“离线也更会做”。可在许可前后安排相同难度、无 AI 的校准任务，使用独立评分；或利用预先计划的短时暂停工具作为学习保持测试，但不能突然撤销必要服务。测试应固定工具外知识来源，避免员工用其他生成式助手替代；也要避免只在上线初期比较，以区分熟练度曲线和团队选择变化。若处理组在离线任务中更快但质量下降，不能称为完整技能提升；若速度与质量保持，才更符合可迁移学习的解释。所有离线测试结果也受练习和记忆影响，需设对照顺序或匹配题目。

SVSAC | sv-sac.org

学习的替代解释至少有三种：高采纳者本来就更熟练；员工在 AI 使用期间接触了额外指导；处理组随着时间获得不同任务或主管反馈。即使同一员工固定效应不能自动消除随时间变化的任务构成。评估可记录培训剂量、任务难度、主管辅导、助手建议采纳、工具宕机与换版，并用预先设计的随机培训或提示指导分臂，而非只将学习改善归因于模型。QJE 关于采纳的分组估计是异质性关联，作者明确讨论选择和选择获益，不能当作“鼓励强采纳”的因果证据。[1]

11.3 过度依赖与专业输入的长期价值

短期表现和长期知识生产可能存在张力。若资深员工不断提供高质量案例、纠正错误和解释例外，内部知识库与模型训练材料可能依赖这些贡献；若工具把高质量建议直接呈现，专家可能减少原创诊断或细致反馈。与此同时，要求专家逐条审查低质量生成结果也可能挤占其本职工作，降低知识更新。两种机制方向相反，不能只根据上线后的答案质量判断长期知识资产是否改善。

QJE 作者提出，高技能客服员工的建议采纳比例增长虽不低，但其生产率获益较小且某些质量指标存在小幅下降，可能意味着过度依赖或对高质量输入贡献减少；论文同时说明短期样本中未发现模型质量随时间下降。[1] 这是一项机制警示和可检验问题，不是已经观测到知识库恶化。要识别长期影响，需要记录专家主动贡献、知识条目更新与撤回、错误传播、不同技能员工的复核时间，以及模型学习数据的使用边界，并在模型更新与数据冻结下追踪。短期试验通常没有足够长的时间跨度证明此类积累或侵蚀。

奖励制度同样可能影响结果。若节省时间被转为更高目标、更多案件或更短处理时长，员工可能得到奖金，也可能出现劳动强度上升。QJE 论文讨论客服人员绩效奖金与目标可能被调整的情形，但作者缺少个体薪酬数据，指出目标上调可能带来 ratchet effect，长期劳动需求均衡也不可由短中期结果推出。[1] 因此，企业不能把一个短期效率指标直接翻译成工资、岗位或永久编制决策。至少要分别记录每班工作量、加班、目标调整、拒绝与返工，以及员工对工作质量和控制感的反馈。

12. 将局部收益转成经营变化的因果链

12.1 产能释放、需求吸收与排班

对固定任务总体，若每件合格工作所需人工下降，组织得到的是潜在产能。其价值取决于现有需求是否超过当前产能、需求是否在释放的时间段出现、员工是否具有相应技能，以及排班能否把时间转到积压任务。若一个月末结账流程节

SVSAC | sv-sac.org

省的时间发生在低峰，而真正瓶颈出现在季度末，理论小时无法直接变成提前完成。若需求存在但任务必须由具有不同授权或语言能力的人完成，释放的普通工时也不能替代稀缺专家小时。任务之间的可替代性应从能力、时点、权限和地理位置拆分，而不能用总人时相互抵扣。

若单件人时只针对合格件，“合格数量×合格件平均人时”仅计合格件投入；总人时还须加入未通过、未决、退出任务及共用维护的投入。只有将全部这些成本明确分摊到合格件后，才可用合格数量乘全成本单位人时重建总额。在固定需求下，减少主动人时可能允许减少加班或外包；在积压需求下，可能增加关闭量；在需求随价格或服务质量响应的环境中，可能带来新请求；在固定编制与固定薪酬下，则可能只形成空闲容量或转岗价值。上述四种路径对个人、团队和企业财务的影响不同。评估者应在预注册时明确主要决策终点：处理积压、减少加班、改善 SLA、重新配置专业人才，还是削减预算。事后再挑最有利终点会带来多重选择偏差。

需求增长也会改变比较分母。工具让服务更快后，更多员工可能愿意提需求，或管理者把原来不做的文书、分析与客户跟进纳入工作。总产出上升并不代表单件效率提高；单件效率提高也不表示组织层面的社会收益必然变大。应把基线任务篮子与干预后新生成任务分成两张账，并为新增任务标注提出者、价值假设、是否替代既有工作及是否经过验收。没有需求与价值测量，“更多输出”可能只是数量膨胀。

12.2 目标上调与生产率基线的变化

一种常见但不一定显而易见的动态是目标上调：员工达到原目标后，管理者把标准提高，员工未必获得更轻松的工作或更高收入。若只看每小时处理量，目标提高可能与 AI 的初期收益同时发生，后续对照就变成 AI 加目标变化的组合效果。固定队列的随机处理可以估计短期任务速度；要估计新目标制度下的持续结果，则应把目标调整作为显式政策干预，设置匹配比较组并记录工资激励、劳动强度和服务质量。前后对比中目标与工具同时变化时，无法将后续差异单独归给 AI。

QJE 作者指出他们观察的是短至中期的局部均衡效果，未获得员工薪酬资料；经理访谈提到奖金可能随绩效目标改善，但管理者也认为目标可能因许多人达到指标而上调，形成目标棘轮。[1] 论文还无法观察较长期的客服劳动力供需均衡，且生产率提升既可能降低所需劳动，也可能因更好服务诱发更多需求。论文用简化假设作了约 12%减少工时的反事实计算，但明确不构成长远就业因果估

SVSAC | sv-sac.org

计。[1] 本篇不将该反事实直接移植到企业；它提示需求弹性、目标、激励和员工供给会共同决定最终资源结果。

12.3 服务质量和员工体验是并行终点

一些工具改善可能先体现为服务质量或员工体验，而非员工处理时间。例如减少升级请求、提高首次解决或改善客户对话，可能长期降低重复联系；但若客户更愿意使用服务，也可能使总接触量上升。员工对工作的体验可能改善，因为工具缓解复杂问题；也可能变差，因为绩效监控更密、重复任务增加或自主性减弱。上述长期路径需要独立的客户、员工和流程测量，不从处理时间推断。

QJE 研究报告客服对话的情绪和升级请求变化，并观察到与较低离职相关联；不过离职每人只观察一次，无法纳入代理人固定效应，作者提醒接入选择可能高估离职下降。[1] 客户满意度、升级、复联与离职属于不同结果变量。论文的对话情绪指标来自 SiBERT 文本评分，与客户满意度/NPS、购买、续费或利润不是同一结果；不能混称。企业若把服务质量作为主要价值来源，应在固定期内跟踪经独立测量的复联、投诉、升级与合格解决，而不只报告表面响应速度。

13. 透明的阶段性资源账本

13.1 构造性例子：固定需求下如何报告，不如何承诺 ROI

设想一个假设的内部请求队列：每月固定基线请求 1,000 件，传统流程最终合格 850 件，未决或退回 150 件；每件合格结果消耗执行者 24 分钟和复核者 6 分钟。AI 辅助流程下，分配总体仍固定为 1,000 件，最终合格 900 件，未决或退回 100 件；执行者每件实际投入从 24 分钟降至 16 分钟，但复核每件增加到 9 分钟。假设只有为简化账面而不计一次性部署投入、返工以外的异常成本和固定软件合同费用，这组账本下执行者投入为 AI 方案 900×16=14,400 分钟（240 小时），复核投入为 900×9=8,100 分钟（135 小时），合计 375 小时；传统已合格件的执行投入为 850×24=20,400 分钟（340 小时），复核为 850×6=5,100 分钟（85 小时），合计 425 小时。以已合格件为分母，每件投入为 25 分钟与 30 分钟，改善约 16.7%；合格件相差 50 件，未决数相差−50 件。

这个比较并不足以宣布“每月节省 50 小时”或“成本下降”。首先，执行与复核工作对未完成任务投入了多少时间还没有记录；仅按合格件核算对组间总劳动差的偏差方向未知：还取决于两组未合格任务各投入多少，不能先断言一定有利于 AI。其次，未计 AI 组是否新增人工修改、超时任务是否退出观察、处理期是否服务范围不同；再次，一次性知识准备、工具配置和员工学习可能相当可观。最

SVSAC | sv-sac.org

后，小时变化是否被排班兑现为加班减少、积压处理或岗位预算变化，账本没有给出。此例只展示为何要在固定总体下分开报告合格产出和角色工时。

若再假设每名执行者月薪成本仅在可避免加班时兑现，复核专家薪资固定且无加班安排，那么 375 与 425 小时的差额不是同额现金节省。假如 AI 方案产生的时间用于处理原来每月未完成的需求，组织或客户可能获得价值，但需追踪额外完成任务是否达标及其业务价值；如果新增需求没有经营价值或任务只是格式化扩产，则不应将其算作利润。费用表应分列工具订阅、集成、培训、管理员维护、数据准备、审核、例外、返工和实际加班/外包变动，不能将一次性时间乘工资率假装为已实现现金。

13.2 试验终点和估计量应按决策分层

在上述假设队列中，可以建立三类互补的估计量。第一，按随机分派的固定总体估计意向处理下合格率差异，例如 AI 组最终合格率 900/1,000=90%，对照组 85%；第二，估计全角色劳动时间对固定分派总体的差异，分母仍包括失败、未决和退出项目，避免只比较合格任务；第三，针对经营选择估算现金项，要求工资、加班、外包、合同与编制记录证明支出实际变化。第一和第二可由任务日志与盲评得到；第三需要财务与人事的兑现记录。不能把中间一项外推为第三项。

若随机化采用团队为单位，应按原团队分派估计 ITT，不能将同队任务当独立随机单位。可在团队×任务类别层事先分层，并预定聚合权重，让主要估计按组织稳定的任务篮子加权；若处理改变任务类别或取消某类任务，还需报告该处理导致的分母构成改变。对工具实际使用的依从性、任务选择和分享污染，可以另作次要分析，但不得改写主 ITT。质量标准由独立评审者在不知道分组的条件下执行；同一任务多次调用、重试和版本升级都要保存在日志，以判断处理定义是否一致。

商业决策中还存在异质性。低风险标准件或许可以先扩大部署；稀有高后果任务则可能采取较高人工审查。这样的分层能提升有限审核资源的配置，但应设置独立抽样复核，避免规则把困难但重要的案例永久排除；需要估计总体错误时，应记录每层抽样概率并按概率加权，且对未回传或不可查样本说明处理。此处建议是作者拟议协议，不是已运行的部署试验。若没有历史数据估计任务构成、初次通过率、返工和出错成本，需把它们作为未知量进行敏感性分析，而非编造一个 ROI 百分比。

要让算例可以推翻，应列出哪些观测会改变结论：如果固定总体合格率没有提升或未决比例更高，不能以合格者工时改善宣称全部任务更好；如果额外复核

SVSAC | sv-sac.org

集中在资深员工且发生拥堵，阶段性净节省不能代表流程吞吐；若请求需求新增后出现有效的增量业务结果，则有理由报告服务扩张价值；若工资或外包支出没有变动，只可报告释放小时及其重新部署用途。研究应将这些分歧结果并列呈现，而不是汇总成单一效果值。

14. 现场比较协议：从任务池到估计对象

14.1 在随机化之前锁定任务总体

现场试验最容易受到的偏差，未必来自随机算法，而可能发生在任务进入随机化以前。经理可能把容易、资料齐全或期望更快的任务先交给 AI；员工也可能只提交自己愿意在无 AI 情况下做的事项。要使处理组和对照组有可解释的比较，应先定义目标流程、合格任务的进入条件、观察期、任务类别和排除原因，并在任务进入队列时赋予不可重复的标识。任务清单最好由业务系统或到达日志自动形成；若需人工提交，需保留所有候选及未提交/排除的理由、提交时间、难度和责任人。不能用结果出来后才被整理的“典型任务集”代替随机化前的总体。

任务复杂度和材料完备度应在分派前测量，因为它们可能同时影响是否调用工具和最终完成速度。可用历史数据定义少数稳定分层，如新/熟悉类型、资料齐/缺、低/高风险，而不以模型自己对任务的置信度作为唯一分层变量。如此分层不是为了事后挑出 AI 最有利的组，而是使总体构成和处理效应异质性可以复核。任务资格应依赖企业原有规则或独立业务判定；如果 AI 在干预后把部分任务重新定义为“不适用”并删除，仍须按照原分派进入意向处理分母。

另需预先说明目标总体是现有任务还是因工具而新增的任务。若试验中管理者扩大可请求事项，新增任务属于处理可能改变的结果，但并非原先固定队列内的任务效应。主要分析可以锁定上线前的类别篮子或同期实际到达的预先资格任务，再将新增服务另作扩展效应报告。对所有任务均应保存进入、合并、撤回、重复、补件、拒绝和重开的事件，以免同一问题被分拆后重复计数，或因重开后编号改变而从跟踪中消失。

14.2 分派层级要与共享资源吻合

若员工完全独立处理、任务不会跨人流转，任务层或员工层随机化可能较有效。若同一主管审核两组、共用知识库或模板、员工互相学习、任务会在组间调度，则个体层随机化受到干扰：AI 组的改进经验可能进入对照流程，或对照组的审核队列被 AI 任务挤占。此时可以先划定共享资源簇，再在簇层分派；代价是

SVSAC | sv-sac.org

可随机化的单位变少、统计不确定性增大。不能靠增加同一团队内部任务条数假装增加独立分派单位。

随机单位、分析单位和实际资源单位应逐项画出映射。例如，五个地区各有多个小组，但所有地区共用一支中央复核队伍。若按小组随机，中央复核队列是共同资源，AI 处理可能改变对照队列等待；若按地区随机仍共用中央队伍，干扰也未完全隔离。可以选择中央队伍所在的完整业务单元作为簇、分阶段随机上线，或把全流程政策效应定义为主要估计量并测量跨臂资源竞争。若无法在实际网络上隔离，试验应明确其估计的是当前共享环境下的增量，而非无干扰条件下的技术效应。

个体或团队无法盲于是否获得工具，但可以盲化最终评分人员和数据分析规则。业务主管在试验期间不应根据早期结果改变目标、排班或审查尺度；若不得不改变，应记录时间和原因，将这些变化视为干预组成部分或偏离方案。预注册要写明分层随机、簇大小、失衡处置和主要模型，并使标准误反映分派层级。若仅有少数簇，需通过设计阶段的功效计算讨论精度，不应以员工或任务的总数替代簇数量。

14.3 明确处理、版本和依从

“AI 处理”不是一项稳定不变的东西。试验方案应列出模型和接口版本、检索库快照、提示或工作流模板、权限、培训内容、允许的外部工具、使用记录方式、失败重试和人工接管条件。若版本在试验中升级，应有变更日志、时间戳、影响面和回滚规则。若每个员工可随意选不同模型，试验比较的就是可变工具组合政策，不是单一模型效果；这可以是有意义的组织问题，但必须如实定义。

主要估计应依原始随机分派计算意向处理效应：无论分派后的员工是否使用、使用几次、是否中途放弃，都保留在原组。另可报告收到建议比例、实际调用、接受或编辑比例，以及协议依从。但按真实使用筛出的比较会重新引入选择偏差。对照组若越过隔离机制使用了外部 AI，也记录为污染，而不是事后把其改分为处理组。任务跨人移交时，保留最初随机组和后续责任链；任务更换处理者不是删除任务的理由。

应用流程上线与模型更新最好不要在同一时间发生，除非目标就是估计整体更新包。若模型更新不可避免，应预先规定“试验时期”或按版本分层，并在变更前后检查输入和结果分布。版本前后出现的性能差异可能来自工具能力，也可能来自任务内容、员工学习、季节需求、知识库变化或复核规则调整。对版本变化

SVSAC | sv-sac.org

建立校准任务集有帮助，但该集需独立于工具开发与调参，避免用其既调试又宣布泛化成功。

14.4 在原分派总体里处理退出、未决与失败

任务可能在试验中取消、等待客户资料、超出观察期或交给第三方。每种状态并不等价。取消可能是需求方撤回，也可能因 AI 组输出不可靠而放弃；等待资料可能不属于执行人控制，但仍是端到端周期；第三方接手是资源转移，不是任务无成本结束。方案应为每种原因定义状态码、触发时点、可重启性，并保留全部初始分派任务作为主要比较总体。主要的合格完成率按所有已随机任务计算，同时报告原因分层；已知截止窗未验收者 Y=0 并保留原因；仅验收状态本身未知者保留缺失并作界限/敏感性分析，不假定完全随机缺失。

对于任务周期，可从随机化开始算日历时间，并把等待与主动工作分解。观察期未完成属于右删失但并非自动“无事件”；对任务完成时间的比较需定义估计目标和删失假设。若项目被取消、需求改变或移交，不能不加区分当作普通删失，因为它们可能对应不同终点或竞争事件。若不用生存分析，应直接报告固定窗口完成比例、未决比例和观察窗限制。所有方法参数需依据任务流定义，而非套用医学随访名词后不检查业务机制。

错误在被发现前可能进入下游，故发生时间、发现时间、纠正时间和影响范围要分别记录。返工不应覆盖初次输出；保留首版、修订轨迹、工具版本、人工审阅者与验收决定。最终结果可为首次通过、修正后通过、纠正失败、客户/业务撤回或窗口末未决。通过后发生的后续撤回和投诉应在预先设定的追踪窗纳入质量结果。未触发投诉不等于没有错误，特别是客户难以识别或反馈的场景。

15. 独立质量评审与多维验收

15.1 评审标准必须来自任务，而非输出形式

合格定义应在看到处理组结果以前，由对业务负责的角色针对每类任务确定。对字段录入，可检查指定字段准确、完整、来源可追溯；对摘要，可核对关键信息、例外、数值和出处；对审批意见，可检验适用规则、理由和责任授权是否完备。语言流畅、格式统一或响应快速可作为体验指标，却不能替代实质正确性。任务类别若具有不同合格标准，应分别报告，而不能只给一个总体质量均分。

评分表要包含重大错误、一般可修错误、非必要重写和风格偏好等可区别等级，并说明何种问题使输出不能放行。至少在样本期开始时，两个独立评审者应

SVSAC | sv-sac.org

对同一批输出盲评；不一致需由第三方裁决或依据预先订立规则解决。若评审者知道任务组别，可能更愿意宽容 AI 输出或更严苛检查；若同一员工既使用工具又评价自身产出，会产生利益冲突。盲态虽不能消除所有主观判断，但能减少结果偏向。

评审还应检查工具引用的材料是否真实支持输出，而不是引用链接存在就判定正确。保留资料版本、字段血缘和适用日期；当源系统发生修订，应判断该输出是否依据当时可用快照。若员工拥有 AI 所不具备的业务上下文，可在验收中记录该信息来自哪里，避免将额外人工搜索成果归给模型。对争议记录由业务责任方裁决，但争议处理时间也要进入人时账本。

15.2 质量指标需与风险类型对应

平均正确率可能被大量简单项目主导，因此报告应分任务类别、严重度和首次/最终验收。错放、错拒、遗漏、错误承诺、重复执行和迟延属于不同损失。对可纠正错误，审核和返工成本更直接；对不可逆动作或客户伤害，应设定不可接受的约束或停止标准，而不以少数节省分钟抵消。风险等级、阈值和停止条件必须来自企业实际责任、历史事故和适用规则，不由本篇给出通用数值。

抽检若只抽 AI 主动拒答或低置信案例，会遗漏系统最自信却错误的输出。复核样本应包括随机抽取的“自动通过”任务、人工修改任务、拒答、重开和末端未决，并记录抽样概率。有限资源下可对严重风险层过度抽样，再用概率权重恢复到总体，前提是各层抽样框完整且抽样机会已知。若没有抽到某类任务或资料无法取得，该层错误率不能由其他层推断；需要明确报告不可估范围。审核人员发现问题的比例不是总体错误率，除非抽样机制和标签验证允许这种转换。

15.3 审核本身也是生产步骤

如果 AI 输出需要检查，评估者的审核时间和质量能力就是流程的一部分。可以用抽样时间研究和系统日志结合，区别主动检查、等待资料、争议升级和复核后的修复。反复打开同一来源、复制错误内容和向原作者询问的时间应按规则计入。由模型生成的解释或置信提示可能缩短阅读，但也可能让审核者过度依赖。可在子样本中随机提供/隐藏来源证据或变更摘要，以识别审计辅助本身的作用，但不得改变必要的安全或法律审查要求。

15.4 评审者一致并不等于质量真值已识别

对工作输出进行双人盲评，首先解决的是两个评审是否按同一规则给出相同标签；它本身不能证明共同标签与业务事实一致。若两名评审都使用了同一份模

SVSAC | sv-sac.org

糊标准、看不到必要源材料，或把格式完成误当成语义正确，他们可能高度一致地漏掉同一种错误。反过来，评审者依据不同经验产生分歧，也不意味着业务质量必然模糊；分歧可能来自量表、任务材料、可接受例外条件或培训差异。必须把重现性、对独立参考标准的准确性和最终业务适用性作为不同问题处理。

一个构造例可显示高一致率的局限：100 件输出中，业务规则所定义的严重错误有 10 件，另 90 件没有该错误。两位评审均未能看到关键源字段，因此都把这 10 件误标为合格；其余 90 件也都标为合格。原始一致率为 100%，但严重错误敏感度为 0%。这个例子并不是说真实评审会如此失败，而是说明仅凭两人一致不能识别他们共享的盲区。若把第一位评审的结果当作“真值”来计算第二位的准确率，也会把共同错误隐藏起来。业务参考标签需要独立于被评估工具和首轮评分，并能访问决定任务结论所需的原始请求、源数据、适用规则和时间截面。

参考标准并非总能以全自动方式获得。订单是否正确、投诉是否可归责、文件是否足以支持决策，可能需要具备责任授权的专家判断。此时研究协议应将标注问题缩窄：评审者必须判断哪些字段、证据和决策后果；允许哪些不确定状态；哪些事实在当时不可观测；有争议时由哪一责任角色按什么规则裁定。可建立“确认合格、确认不合格、证据不足/无法裁定”三类结果，并将无法裁定单独报告。把不确定项强行并入合格或不合格，会把参考标准的缺失伪装成模型性能。

若所有样本都由两人评分，报告原始一致率、混淆表及按标签分别计算的正/负一致率，而不只给一个 Kappa。少数严重错误很罕见时，Kappa 会受到标签边际分布影响；相同的实际分歧可能在不同患有率的样本中得出不同系数。因此应呈现每类样本数、评审边际分布、分歧方向，以及评分是在同一任务/版本/审查资料条件下完成。Kappa、Krippendorff alpha 或其他系数都不是“质量正确”的概率，也不能替代独立业务裁定。

15.5 有限复核资源下如何估计共同错标

若团队无法对全部任务进行专家裁决，最关键的设计不是事后只复核两人有分歧的案例，而是从完整任务框中随机抽取一部分“两人一致”案例做第三方盲态参考审核。只裁决分歧项能解释分歧来源，却无法估计一致标签中共同漏错的比例。样本应覆盖两组、不同任务类别、风险等级、工具是否采用、自动通过、人工修改及未完成状态。对于稀有严重错误可在审计样本中有意多抽，但必须保存各层抽样概率，再用设计权重还原总体比例；若某类任务被完全漏出抽样框，不能由其他类推断它的错误率。

SVSAC | sv-sac.org

可复算的抽样方案示例：某固定任务总体含 800 件低风险输出和 200 件高风险输出。复核经费只能提供 200 次第三方裁决。研究者按两层各抽 100 件简单随机样本，而不是按总体比例抽 160/40，以提高高风险层精度。若在低风险层发现 2 个共同误标，在高风险层发现 8 个，则未加权的审计样本共同误标率为 5%。但恢复到业务总体的估计是 0.8×(2/100)+0.2×(8/100)=3.2%，不是 5%。若第三方裁决只在自动通过输出里做、却未抽到人工接管或未决事项，这个 3.2%只代表该框，不能声称覆盖所有分派任务。

设计权重只能纠正已知的抽样比例不同，不能弥补无法接触某一群体、参考标准缺失或评分流程因处理组而异。若第三方审核对 AI 文本更费时或无法理解其来源，审核缺失可能不是随机的；应记录未完成裁决的原因，给出缺失率，并对未裁决项目在可辩护的最坏/最好错误状态下计算边界。对样本量有限的严重事件，置信区间可很宽，不能由点估计低就宣称风险低。资源安排应先保证风险层和一致标签的随机审计，再用富余预算研究评分分歧的细分机制。

第三方裁决也不是无误差“金标准”。实际操作可以由一名业务责任专家在盲态下裁决，并对随机抽样的裁决结果再由另一名专家复核；若两位专家再次分歧，由预先指定的责任人按证据规则裁定，并保留原始意见。最后的裁定是用于经营质量口径的参考结论，不意味着客观真理。报告中需同时给出首轮评分、一致情况、第三方审核覆盖、最终参考标签和无法裁定比例。若复核标准更新，重新标注旧输出和新输出并行样本，或将标准版本变化作为结果解释的一部分，不能把标准更新造成的标签改变误认为模型质量变化。

一组精确的标签结果表至少应区分：双人一致合格、双人一致不合格、双人分歧、仅一人标记严重问题、参考审核发现共同漏错、参考审核推翻共同错误报警、证据不足。对质量敏感的主分析可以按参考标签估计“分派任务中最终业务合格率”，并另报告初审标签相对参考标签的假放行和假拦截率。若仅在一致项抽样做参考审核，则需将抽样概率用于校正；在没有对一致项的已知概率抽样时，共同错标率不可识别。

15.6 重复评分的独立性与盲态边界

双人评分只有在信息和执行足够独立时，才可能揭示不同误判。若两人共享前一位评语、模型置信分或主管建议，第二次评分不再是独立复核。可将输出去除处理组标识，随机排列，隐藏对方结论和后续业务结果；评分顺序与工作量也应平衡。若某些输出特征暴露了 AI 身份，盲态无法完全实现，应记录破盲比例与识别信心，并对可盲和不可盲子集做敏感性说明，而不能声称“评审双盲”。

SVSAC | sv-sac.org

评分者内部的稳定性也可能影响组间比较。同一评审者在疲劳时变宽松、在月底时加快判定，或先看到某组大量格式问题后形成预期，都会造成时间与组别相关的打分漂移。研究可把任务在评审者间交叉分配、随机插入固定校准样本，并保留评分时间、撤回/重评轨迹。校准样本只能检查评分规则执行是否稳定，不能替代独立真值；若定期培训改变了尺度，需记录版本并对适用期间分层报告。

在责任风险较高的流程中，复核不应仅追求双人一致率。可规定严重错误分歧一律升级，常规差异按第三方抽审，无法裁定项不进入“通过”分子但留在分母并单列。若工具政策导致更多拒答或更多人工转交，业务质量提升可能伴随等待与专家工时上升；应同时报告最终合格、未决、返工、风险事件与总角色工时。评分本身也是流程成本，不能免费假设。

15.7 参考标签自身不确定时，错误率只能给敏感性范围

第三方审查可降低共同误标，但如果审查员仍会漏掉某些严重缺陷，抽审结果不等于潜在真值。设参考审查对真正错误的敏感度为 Se，对真正合格项的特异度为 Sp，参考审查标成错误的观察率为 q，潜在真实错误率为 p；在简化的二分类模型下，$q=Se\cdot p+(1-Sp)\cdot(1-p)$。仅观察 q 不能同时识别 Se、Sp 和 p。若没有第二种独立证据来验证参考审查，就不能把参照标签准确率视为 100%。

以下代数校正将 Se、Sp 限定在[0,1]，并限于 $Se+Sp>1$ 的正向信息性参考标签；分母 $Se+Sp-1$ 必须非零。将 q 视为模型中的总体判错概率时，其可行范围为 $[1-Sp, Se]$，求得的 p 须在[0,1]。若 $Se+Sp=1$，则 $q=1-Sp$ 与 p 无关，不能由该标签识别真实错误率。若样本判错率代入后落在可行范围之外，当前点校正与所设参数不相容，但有限样本波动也可能造成这种现象；应报告抽样不确定性并检查参考参数或模型失配，不能将负解或大于 1 的解机械截成 0 或 1 后宣称已校准。

一个完全构造的敏感性计算：假设参考审计观察到 3.2%的任务被判为严重错误。若暂定 Se=90%、Sp=99%，则校正后的 $p=(q+Sp-1)/(Se+Sp-1)=(0.032-0.01)/0.89\approx2.47\%$。若敏感度只有 70%、特异度仍为 99%，同一 q 对应 $p\approx3.19\%$。这两个输入均未从真实资料估计，只展示参照质量假设会改变风险估计。实际报告可列出 Se/Sp 的可辩护区间，计算对应真实错误率范围；若该范围跨越企业停止或扩展标准，结论应是证据不足，而不是挑选最乐观参照假设。

敏感度和特异度可以通过第二层验证获得有限信息：在初轮第三方裁决中随机抽取一部分样本交由具备不同职责或信息来源的专家复审；对临床式/机械式“金标准”不存在的业务任务，可分别保留多名专家观点与证据链，而非多数票直接抹去分歧。若第二层复审也不独立，或共享同一不完整源材料，共同偏差仍

SVSAC | sv-sac.org

可能存在。允许的结论应按证据等级表述：对可由原始数据完全验证的字段，可以给出较强正确性判断；对涉及未来结果或专业裁量的项目，报告评审一致程度、未决比例与合理误差边界。

若只想估计“组织当前放行规则造成的观测严重事件”，参考标签误差的校正可能不是首要端点；可直接定义真实业务事件并通过独立记录追踪。但对于“工具输出的真实语义准确率”，标签构造本身就是测量过程。两种估计对象不能混用。一个实用决策表将原始评审结果、第三方复核、最终放行、下游纠正、客户损害分别保留；财务或人事决策应依其需要的对象使用对应结果，不能将某个评分表合格率自动解释成真实无风险率。

16. 从试验到规模化：容量、拥堵与回退条件

16.1 先确定有限瓶颈而非扩大工具覆盖

当试点成功后，常见决策是立即把权限扩到全部人员。然而试点中的审核由项目团队、临时专家和密集管理关注承担，规模化后这些支持资源不会按比例增加。上线前应列出每个阶段的容量：工具服务并发、资料更新、专业复核人时、异常响应时长、申诉处理、数据治理和培训时段，并在峰值到达量下做压力推演。若所有资源只按试点均值规划，需求波动、低概率异常或上游导入任务增长就可能使队列失控。

一个实际的扩展准则应关注瓶颈余量与质量而非简单设定“使用率达到某比例”。例如，只有当标准任务在独立审计中保持合格，异常分流有明确负责人，且复核容量对高峰负荷仍有余量时，才将自动权限扩大一级。若质量不确定，先维持影子模式：系统生成建议但不触发正式动作，比较其与实际处理并记录潜在差异。影子模式能发现输入/版本问题，却不能完全观察真正使用后员工的采纳、监督和行为反馈，因此不能替代有限的受控上线。

16.2 回退和停机需进入处理定义

工具故障、数据陈旧或外部系统接口错误时，员工如何继续工作会影响平均效果。试验应明确切换到手动流程的判据、恢复旧版本的权限、任务去重和幂等机制、人工接管队列、未完成动作核对，以及故障期间是否仍属于处理组。无回退方案时，短期提速可能以单点失效为代价。每次回退都需记录发生原因、影响任务数、恢复时间、是否有重复执行及客户影响，并按意向处理分组保留。

自动执行比辅助建议对监控更敏感。对外部通知、付款、账户变更等操作，除模型输出外还应验证身份、交易状态、目标系统确认和执行后对账；指令已发

SVSAC | sv-sac.org

送但无回执的情形不能盲目重试，否则会重复扣款或重复通知。对不可逆或高影响事项应使用单独审批和最小权限；模型无法核验必要前提时，应停止并升级。此处为通用工程控制建议，需结合具体系统架构和业务责任另行验证。

16.3 规模化会改变流程构成与学习条件

试点可能主要吸引热情员工和易支持任务；全员上线后，使用者构成、任务类型与主管支持会变化。QJE 研究在单一企业客服环境中观察到不同技能与资历员工收益不同，同时讨论了较长期均衡未知；METR 后续研究则表明，在开发者广泛使用工具后，参与和任务提交可能受“不愿无 AI 工作”的偏好影响。[1][5] 两项证据对应不同人群和工具时期，不应合并成普遍规律，却提醒规模化评估需跟踪试点以外的拒绝参与者、边缘任务及不同技能岗位。

若试点通过后，组织改变绩效目标、人员排班、复核制度或工作分配，规模化效应已成为新的实施包。对流程复制，应保留一组同期对照、交错上线或明确的阶梯设计，追踪工具版本和组织规则变化。已经被处理的团队无法充当未处理对照；员工转岗也会改变任务构成。每次扩大范围都应记录当时的部署决策、预期收益、风险边界和真实观察，而不是把成熟系统阶段的成功倒归因于首个模型试点。

17. 可解释的资源兑现：工时、服务和现金分别结账

17.1 内部工时先成为“可重新安排的时间”

员工在任务上少花的时间要先经过可兑现链条：是否在排班内真正释放，是否与其他工作时间重叠，是否具备承接其他任务的技能，是否有足够需求，是否能由管理者授权重新分配。若员工在 AI 等待期间本就处理其他案件，工作台计时显示的墙钟缩短不等于释放同等员工时；若节省发生在班次尾部但不能提前下班或调班，时长也没有自动变成现金。内部工时可以作为容量、工作体验或服务改善指标报告，但不应直接乘平均工资推算利益。

需要区分四项账：每件主动工时、按班次/周统计的付薪工时、释放时间的实际去向和最终财务账。工时数据来自任务日志、抽样观察与排班系统的交叉核验；转岗时间应记录到新任务；未被吸收的空闲容量单独报告。若组织以标准工时而非真实工时核算，则应同时给出两套口径以及标准是否因试点后调整。释放后交给主管审核或专家处理的时间需从总收益中扣除。

SVSAC | sv-sac.org

17.2 现金收益要有可观察的决策证据

可验证的现金变化通常需要具体管理决策：取消加班、降低外包量、推迟招聘、减少供应商合同或避免某项费用。每一项都要记录决策生效日期、数量、单价、相关性和可避免条件。若同时发生人员冻结或需求下降，费用变化可能由其他原因造成；若岗位未裁撤且薪酬不变，称“节省了工资”不准确。推迟招聘也仅在有获批计划、需求未被放弃且由 AI 容量承接时，才能视为相对于明确反事实的避免成本。

工具总成本还包括集成和维护、知识库整理、权限配置、模型费、审查与纠错、员工训练、停机备用、法务或安全检查和退出/迁移。一次性成本与持续成本应分开，且区分内部工时的机会成本和真实现金支出。若所有成本均无法获得，财务净收益就不可估；可以报告生产时间和结果指标，但不应填上行业平均成本或编造现值。成本分析可以使用区间敏感性，列出在不同需求、失败率和审查能力假设下盈亏平衡的条件。

17.3 机会成本、分配和可持续服务水平

不以现金形式体现的收益仍可能重要：员工把释放时间用于更复杂案件，客户等待更短，减少重复联系或改善工作体验。此类收益需对应具体受益人和服务终点，不同人群可能分配不均。高熟练员工可能被调去审核 AI，低熟练员工处理更多例外；平均产量上升时，专业人员负担或员工自主性可能恶化。报告除平均效果外，也应按技能、班次、任务风险和客户群检查分布结果，并审视岗位变动与工作质量。

“释放时间”也可能随绩效目标提高而被重新吸收，长远价值便取决于管理方式。组织可以把容量投入积压任务、服务扩大、培训或减少加班，分别形成不同的经营收益。若试点的主要目标是降低员工重复劳动，应直接测重复工时和工作体验；若目标是缩短客户等待，应测日历时长、按期完成和服务体验；若目标是预算，则需财政或人事决策证据。结果不一致时应报告这种权衡，而不能仅以某一指标最大化代表整体成功。

18. 识别工具、培训与流程重设计各自的贡献

18.1 一个上线包包含多个可变组件

企业落地通常同时变更多项：开放工具权限、统一任务模板、清理知识库、培训员工、增加复核、重划职责、调整排班。若所有变化同日发生，试验可以回

SVSAC | sv-sac.org

答“这个上线包相对于现行流程怎样”，不能说模型单独造成了全部变化。更进一步，组件之间可能互补：知识源可追溯时，模型起草才有可核查依据；培训让员工学会如何向模型提供问题；工作台集成减少复制成本。也可能出现替代：标准模板降低重复录入，因而 AI 价值变小；额外审核抵消生成提速。要决定扩哪一项，需要设计可识别各组件而不是仅记录工具使用。

原则上可采用多臂比较，例如现行流程、仅提供工具、工具加训练、工具加流程改造、工具与两者都加。其统计和组织成本随臂数迅速增加；样本不足时，估计会变得宽，且主管更难维持不同操作条件。另一种是先在明确可比的子流程中做序贯实验：第一阶段测基础工具政策，第二阶段固定工具后比较训练或复核方案。后续阶段使用的是经过试点学习后的工具包，因此结果条件于既往经验，不能当作一次性的总体交互效应。

实施组件不一定可以单独随机。若知识库由一个部门统一更新，部门层是更自然的分派单位；若培训材料会在员工间流传，个体培训组与对照组难以隔离。可记录自然发生的非依从与交叉接触，但结果须解释为被污染环境下的政策效应。未经随机或可信准实验识别的组件比较，应标为探索性或描述性。不能根据谁实际参加培训、谁遵守新流程进行事后配对，暗示这些差异是该组件造成的。

18.2 测量“处理”而不把版本变化掩盖成噪声

模型、界面、系统提示、搜索索引和企业文件会一起变化。员工可能使用浏览器助手、企业聊天插件或嵌入业务系统的模型；对同一类任务，各版本可见上下文和拒答策略并不一样。上线日志应区分生成版本、上下文来源、员工操作、人工补充资料和外部工具调用。若模型提供的建议每次都不同，处理并不是固定文本，而是提供特定时点可用系统的政策；应通过版本和输入记录使处理可复现，而不是假设一次回答能代表整个产品。

数据冻结与生产使用的目标不同。为比较模型可先用冻结任务集验证稳定性，但日常流程需要应对最新订单、政策和客户信息。冻结集表现只能测试该快照；在生产数据上的观察又混有需求和人员变化。较稳妥的路径是：用历史冻结样本作调试与盲态验收；再以影子模式接入实时输入；最后在风险允许的实际流程中分阶段授权。每阶段的估计对象都不同，不能把模型基准成绩当成正式生产流程效果。

版本更新应预设变化检测规则，如输出类别突然改变、拒答率或人工覆写上升、某一客户群错误聚集等；检测后先确认数据输入和规则变更，再决定暂停、回滚或重新验证。在线指标只捕捉暴露的错误，未被使用或未被抽查的项目不可

SVSAC | sv-sac.org

见。规则也可能形成反馈回路：员工在早期纠错后把结果写入知识库，新版模型因此看起来更准确，但增量价值包含员工积累的资料投入。若要分别估计，可冻结知识快照或将知识库更新作为独立处理组件；若做不到，须将其纳入上线包。

18.3 版本更新也会改变知识环境

18.2 已说明知识库投入须与模型一起登记，此处只强调反馈改变数据生成：早期自动输出可能改变员工行为，员工行为再改变业务记录，后续训练或检索又使用这些记录。若不保留输出、人工纠正和来源版本，模型可能把自己先前的错误重新读作组织事实。该循环是待检验机制，不是本篇观察到的知识库退化。

因此，版本监控不应只记录模型名称。相关检索索引、源文件及生效日期、模板版本、审批阈值和员工指引均需保存。更新后表面改善可能来自规则纠正而非模型推理进步；表面下降也可能来自任务混合或评审尺度变化。运行环境改变时，应重新核验输入和输出假设，并保留回滚到最近验收配置的路径。

19. 比率、分母与任务组合的变化

19.1 每小时产出和单位时间的统计陷阱

“每小时处理数”是产出与时间的比率。若处理组的员工加班、并行处理、任务间切换、接手他人任务或未完成工单未记入分子/分母，比例会失真。假设一组通过减少每件沟通时间而在同一班次接手更多简单任务，另一组则负责少量复杂案；直接比较每小时件数会把任务组合差异当成效率。理想报告至少有原始总量和分项：任务到达、完成、通过、重开、转移、付薪时长、加班与人员构成。时间既有以员工小时为分母的速度指标，也有以固定任务数为分母的总投入指标，二者可能方向不同。

一例说明加权问题。若流程原有类别 A 占 80%，每件用时 10 分钟；类别 B 占 20%，每件用时 50 分钟，平均每件时间为 18 分钟。AI 运行后，任务篮子变成 A 占 90%、B 占 10%，即使两类的条件用时均未改变，观察到平均时间会降为 14 分钟。这看起来快了约 22%，其实只是困难类别占比下降。反过来，AI 若使更多难任务进入系统，平均耗时上升也不等于条件内效率恶化。报告应给出基线固定权重下的标准化估计和实际新组合下的运营结果，并将两者区别开。所有数字仅是用来演示构成效应的算术示例。

合格产出率也需要固定口径。例如任务关闭数上升可能来自把未解决的问题分成多个短工单；客户或内部请求的单位实体却没有变化。应定义独立请求 ID 和最终问题解决条件，再以事件关联合并重开。若不同任务可以合理拆分，生产

SVSAC | sv-sac.org

率单位需说明是案件、活动步骤、决策还是服务需求。选择更细单位会自然增加数量，不能自动解释为业务价值增长。

19.2 质量调整比率的权重选择

组织可能将质量分数乘以产出得到“质量调整产量”，但质量量表权重带有业务价值判断。把一次关键遗漏等价于几次格式瑕疵，或将客户安全与体验以同一分数相加，都会改变结论。对多维质量，较透明的做法是先报告并列维度：重大错误率、一次通过、返工、延迟、拒绝/升级、客户结果；若需要决策合成，再明示权重来源并进行敏感性分析。不能仅靠构造一个综合指数解决不同价值取舍。

质量评审还会因样本权重而变化。若高风险类被过度抽样以提高发现能力，原始错误总数不等于总体错误率；需按抽样概率加权，并同时报告各层结果。若以业务损失加权，需说明损失值来自哪些真实账目、历史记录或责任方估值。罕见严重风险不能因平均权重低而被隐藏，应作为约束或单独的风险曲线呈现。任何无法测量的损失需列为不确定性，而非默认为零。

20. 从短期试用到持续经营效果

20.1 试用期学习、稳定运营和模型更新

新员工第一次使用工具时要花时间理解界面、学习输出边界并建立信任；同一工具在几周后可能更熟练。反过来，初期项目团队往往获得更多专家支持，后续普通员工的支持水平下降。因此，平均处理效应随时间变化并不奇怪。试验应将上线准备期、学习期、稳定使用期和版本变更期分开报告，而不能把它们平均为一个永久效果。停止在试用后的第一个月可能低估熟练度，也可能高估密集监督下的效果。

QJE 研究考察了上线后的分批接入和员工在不同任职期的变化，指出助手可能缩短新人爬坡时间，同时影响已有员工；作者也明确将结果限定于单一企业和中短期。[1] 该研究不能证明所有企业都能减少入职培训。较长期员工技能、工具依赖和知识库更新仍需不同设计追踪。METR 报告的早期 2025 年开发任务变慢与后期研究选择问题进一步表明，不同模型版本、任务组合、参与者预期和计时方式可能共同改变表面效果。[4][5] 版本或时期差异要作为研究变量，而不是把不一致结果简单平均。

SVSAC | sv-sac.org

20.2 需求和目标反馈形成动态平衡

AI 让服务成本降低可能带来更多需求，提升用户体验也可能增加使用；与此同时，管理目标可能上调，员工会把节省时间转给更复杂的工作，流程量与任务构成随之变化。这样的反应是市场、组织和劳动者共同调整，不是单一模型输出的静态差异。早期试验估计的是某种供给和目标条件下的短期局部效应，无法独立推断价格、人员和服务边界调整后的长期均衡。

评价动态应预先选定结果窗口，并记录每阶段进入流程的总需求、服务范围、目标、人员和排班变化。若每月任务量因 AI 上线而增加，则固定基线篮子结果回答单位效率变化，实际全部需求结果回答扩展服务下的运营表现；两者均可能有管理意义。若目标上调，需将目标修改前后的处理效果分开，或将目标制度作为第二个政策因子随机比较。若只看最终总量，无法识别是单件更快、任务更多、员工加班还是质量标准降低。

20.3 停止、扩张与重新验证的决策规则

组织不应等到平均效果显著后才讨论扩张规则。可在研究前设定条件性阶段门：一旦重大错误达到既定责任方阈值、缺少适用数据、输出无法审计、复核队列超出容量或回滚机制失败，就暂停自动动作；若质量和吞吐达标但现金成本无法兑现，则可以继续评估服务或容量价值，而不宣称预算节省；若质量存在不确定性，则维持人审或影子模式，并增加独立样本复核。具体阈值只能从本组织的风险与承诺定义，不能由论文给出通用数字。

扩张应是可逆的操作决策。逐步增加任务覆盖、人员或权限时保留版本和同期比较；达到新的需求规模后重估复核和异常容量。若部署机制、知识库或审核流程已变化，先前的对照结果未必仍适用。暂停机制也需记录员工如何手动继续，避免停机本身造成客户损害。把停止/回滚写进方案并非宣称工具不可靠，而是确保组织知道出现何种证据时要改变行动。

21. 贯穿全流程的试验分析计划

21.1 固定主问题与次要问题

一个完整的评估计划先写出主要问题：在特定资格任务总体和观察窗口内，提供 AI 工作包是否提高最终合格交付比例，而不增加无法接受的质量损失？“AI 工作包”需列明版本、培训、权限、检索资料和复核。主要分母是随机分派时固定任务总体；主要结果在预设质量标准下按窗口内完成计数。其次再考察首次通

SVSAC | sv-sac.org

过、主动人时、等待、返工、客户体验和支出等次要结果。若质量下降而时间缩短，应报告权衡；若在预先规定容差及充分精度下质量保持、全角色时间下降，则可支持该任务范围的劳动效率改善。质量差异不显著本身不证明质量保持，仍须检验相应非劣界限与严重错误约束，不能仅因质量未提高就排除效率收益。

主要比较采用原始分派，即使执行过程中员工没用工具、拒绝建议或转用手动流程，也留在原组。单个任务可定义 Y=1 为窗口内最终验收，已知截至窗口未验收者记 0 并细分未决状态；若连窗口内是否验收都因失访或记录丢失而未知，则保留缺失并作界限/敏感性分析，不冒充确知的 0；对员工或团队处理，则按分派单位聚合结果，报告每个单位的任务量和质量。若项目任务存在共享工作者或主管，随机分派应在能避免严重污染的层级；估计与推断基于该层级，任务行数不等于样本独立数。

21.2 异质性与次要使用效应

次要分析应事先限定少数有理论根据的任务类别和经验层级，例如低频/高频、资料充分/不充分、新员工/经验员工。避免把结果后发现的数十个子组都当作确认性因果发现。每个组的任务数、效应区间和多重比较处理要一并报告。若分析实际使用者或高采纳群体，需说明这不是随机处理效应，并以随机分派、工具建议可用、依从和跨组污染记录作为识别辅助。

一项有解释力的次级设计可嵌入随机培训或使用说明臂，估计工具访问、培训和二者交互。需要足够的分派单位及方案一致性，否则交互估计会极不精确。若条件不具备，不应为了分析复杂性硬拆多臂，应按整个实施包报告并说明无法分辨组件效果。实际使用者层面的学习分析可结合预先安排的无 AI 校准任务，但需要保证任务难度和评阅标准相当，并单独解释训练、练习与工具熟练度影响。

21.3 估计不确定性、功效和数据质量

方案应依据历史任务分布估计流程波动、任务内相关、团队数、每组分派量、合格率和未决率，进而进行可复算的精度/功效规划。没有这些基线数据前，不应杜撰通用样本量或最小效果阈值。结果至少报告效应区间和在预先指定业务阈值下仍可能的损失/获益范围。若估计不精确，正确结论是证据尚不足以支持扩张或停用，而不是只比较点估计正负。

数据质量也会影响估计。任务 ID 重复、时间戳缺失、审核标准变更、不同系统记录时区和延迟写入、人工补录，都可能产生误分组或虚假时长。上线前应对 ID、日志时钟和状态转换做干跑，保证处理组和对照组的数据捕获程度可比

SVSAC | sv-sac.org

较。若有某组更多缺失，应报告缺失模式并做保守界限；无论何种估计方法都不能从丢失的日志自动恢复未观察到的劳动。

21.4 判读结果和复核决策

最终结果可用“效率、质量、流程、兑现”四类并列表展示：固定任务总体的合格完成/返工与人时；端到端周期/积压/例外；质量严重度和未决；现金支出与释放时间用途。对每一类都写清观察窗口、统计分母和来源。若改善仅在简单任务出现，可考虑限定适用范围；若效果主要来自培训，先判断是否继续培训比增加模型权限更有效；若人时下降但等待不变，可将其作为产能结果而非服务提速；若高风险事件无法排除，应避免扩大自动执行。

复核决策应使用观察到的流程日志和独立验收，而非将所有结果压成总体单一分数。业务负责人可以基于预先定义的服务标准和不可接受风险作决定，但应将价值权重明确交代；研究者报告的是证据区间与条件，不替企业决定风险偏好。若进一步扩张，要把阶段门、回滚、复审日期和适用范围写入运营决策，且保留可追溯记录，以便后续检验结果在工具和需求变化后是否仍成立。

21.5 把“节省工时”转换成财务判断的敏感性账本

考虑一项明确标记为假设的月度计算：日志显示常规执行工时减少 60 小时，复核和返工新增 18 小时，净可释放容量 42 小时。假设企业的实际合同、加班记录和排班决策显示，其中 20 小时确实减少了付费加班；其余 22 小时用于处理积压或作为未兑现容量。若本情境中可避免的加班现金成本假设为每小时 40 元，那么本期可观察的加班支出减少为 20×40=800 元。若工具订阅及维护的假设月费为 480 元，则持续现金差额是+320 元；这还没有扣除一次性集成、知识清理和培训，故不能据此计算全生命周期 ROI。

结果对现金转化比例高度敏感。若 42 小时净容量中实际减少的付费加班只有 6 小时，现金收益为 240 元，扣除 480 元月费后差额为−240 元；若能证明有 20 小时减少加班则为+320 元；若没有任何加班或外包减少，现金收益为零，即使处理能力真实提高。反过来，若 22 小时被用于有价值的积压处理，组织可能获得服务收益，但必须用合格关闭、等待缩短或客户结果来证实，而不能当作现金结余。这里的金额、工时、比例和价格均为构造假设，且没有货币化质量损失；它演示的是每一项财务主张必须对应可追溯决策记录。

更一般地，令 H 为扣除新增审核/返工后的净容量，f 为实际转换为可避免付费工时的比例，w 为该类可避免工时的真实单位成本，F 为月度持续性工具与维护支出，则净现金变化可写为 fHw−F。公式隐含节省工时相互可替代，w 适用于

SVSAC | sv-sac.org

被减少的支出且没有服务/质量损失；若这些条件不成立，公式需分项或不能计算。一次性投入 I 不应与每月成本混淆，可单独报告投资回收期情景，但前提是未来现金流确有证据。对 f、w 和 F 用区间做敏感性分析比用单一点估计更诚实；若关键参数完全不可观察，就应声明财务效果不可识别。

21.6 固定窗口与延迟结果的双层判读

业务任务的最终结果常迟于模型输出。申请材料在窗口末已知仍待补件，在固定窗 Y 上记 0，但不等于断定其最终业务质量失败；已通过初审的流程也可能在几周后被撤销；客户投诉更晚到达。分析方案可预先设一个主要固定观察窗，报告该窗内完成、未决、取消和转出的比例；另设一个延迟追踪窗检查重开、纠错、申诉或重复联系。两个窗口回答不同问题：前者反映当前服务能力，后者检验初始效率是否以延后错误为代价。

如果一个任务被转交、需求方撤回或流程规则改变，结局与普通“仍在排队”不同。分析时应保留原因，并在预定状态分析中区分竞争结果、暂停、资格变化或重新开始；主要比较仍保留全部初始分派，不因不利结果事后删除。若报告中位日历时长，应明确是在完成者中计算还是对全部初始分派作估计；未达到完成的 50%时不能直接报告一个总体中位完成时间。主要结论可以先用固定窗完成率和未决率，不必为了方法复杂而套用不适合流程机制的模型。

对延迟错误，要区分跟踪到的错误和没有被发现的错误。观测窗延长会提高发现概率，但也可能引入业务环境和工具版本变化。应记录发生日期、发现日期、纠正行动、影响对象和发现来源，并保持原处理分派。若不同组客户反馈渠道不同或退出率不同，投诉数比较不再代表同一错误观测机制。未发现错误只能解释为“在现有观察和审核机制下没有观测到”，不能推断真实错误为零。

22. 从界面代理基准到真实工作流：能力证据的层级

22.1 任务完成率究竟测量什么

知识工作中的自动化常经由既有网页系统执行：检索资料、筛选记录、填写表单、创建服务请求，或把结果写回业务系统。此类任务看似是“把员工操作自动化”，但成功标准至少有三层：代理是否完成界面动作；写入的数据是否符合业务语义与权限；该记录是否被组织接受并推动实际流程到达目标状态。前一层可以由系统状态校验，第二层需验证字段、上下文和授权，第三层涉及后续审核、等待、撤回及业务结果。把按钮点击成功等同于工作已完成，会漏掉后两层。

SVSAC | sv-sac.org

Drouin 等人提出的 WorkArena 是针对 ServiceNow 企业软件网页操作的研究基准，包含 33 种任务与 19,912 个参数化实例，覆盖列表筛选排序、填表、知识库检索、目录服务、仪表盘信息提取和菜单导航。作者设计了可对数据库状态进行校验的验证函数，并以手写自动化流程作为可行性 oracle；他们同时指出任务说明被设计得明确，提供完成所需信息。本篇核对了 PMLR 235 正式论文的任务、验证函数、§5 协议与表 2；所测为当时固定代理配置，而非当前所有模型或生产部署。[6] 这些设计适合比较代理在规定界面任务上的可重复能力，也有利于发现复杂 DOM、动态表单和上下文截断等工程障碍。

但基准的“成功”是任务规则下的通过事件，不是员工工时、合格业务产出或客户结果。若任务指定创建一条资产记录，代理可按字段规则提交；真实工作中，员工可能还需判断是否应创建、确认申请人身份、查重、询问缺失信息、考虑审批权，或等待供应商确认。基准把目标明确写给代理，减少了需求澄清和目标冲突；现场工作恰恰可能把大部分成本花在不完整或相互矛盾的请求上。数据库 oracle 判断状态正确，不能证明数据来源可信、授权有效或记录应被业务保留。

WorkArena 结果进一步说明，复杂的工作界面不等于自动化已经成熟。作者在其报告的配置下观察到 GPT-4o 在该基准约 42.7%的总体成功率，而任务类别差异很大：筛选列表任务未成功，服务目录任务成功率较高；其他模型的表现亦有显著差异。[6] 这不是“42.7%的工作已能自动化”，更不是企业生产率增幅。该表按 33 种任务、每任务 10 个种子的协议评估配置，不是把全部 19,912 个可生成实例都测过；结果依赖任务模板、软件版本、代理提示、可用动作、最大步骤数以及选定模型。某些基准类别由大量参数实例构成，也不代表真实工作流中各类别的发生权重。

22.2 评价链条的四个层次

为避免把模型能力、流程能力和组织结果混为一谈，可把评价拆成四层。第一层是动作有效性：代理是否按指定界面状态完成输入、导航或提交。其分母应是所有分配的尝试，需记录错误恢复、重试和超步数退出。第二层是语义正确性：填写值是否对应原请求、最新有效政策和业务定义；若字段看似有效但对象选错、单位错、版本旧，界面动作可以成功而任务仍失败。第三层是流程验收：负责员工或下游系统是否接受并关闭实例，包含退回、修订、授权和等待。第四层是组织结果：合格处理量、延迟、错误损失、人员投入、服务水平和成本是否改变。

SVSAC | sv-sac.org

层级之间有因果顺序，但不能简单相乘成一个“自动化率”。假设代理对某类表单每次动作成功率为 90%，其中 10%的动作成功存在语义错误，业务审核又未必能发现所有错误；即使概率数字可靠，最终风险仍依赖错误类别之间的相关性、任务组合与检测过程。以相互独立为前提把多个单步准确率相乘，可能低估连续错误，也可能高估错误扩散。真实评估需保留每个实例的阶段状态，直接报告阶段转换与下游损失，而不是用单一成功率代替整条链。

动作层速度也不是全流程时长。代理可能在几十秒内操作完成，却要经过人工复核、补证和等待。代理的重试生成额外 API 调用和机器时间；人类则需要监控、处理异常、恢复错误状态。若代理运行与人工工作并行，计算资源时长与人的主动工时不能直接相加成同一种“时间成本”。应分别记录墙钟周期、人工主动时间、机器运行时间、等待时间和复核投入，并设定一个共同的业务起止点，例如从任务正式受理至验收关闭。只在程序执行部分打秒表，得到的是工具执行延迟，无法回答流程周期是否缩短。

22.3 基准与现场之间缺失的外部条件

基准评估通常拥有稳定任务生成规则和可重置环境，重复运行有利于区分代理配置；生产环境包含登录权限、并行用户、数据权限、软件升级、网络异常、人工排队及真实记录的不可逆后果。这些差别会影响有效性，而非只是部署成本。一个代理可能在沙盒里成功填入采购单，真实环境中却遇到供应商目录更改、申请人权限不足或同一订单已被另一员工提交。若自动化重复提交造成重复采购，其严重性无法用“页面路径正确”衡量。

因此，基准可用于两个有限判断：一是当前系统对某一明确定义的界面任务是否存在明显能力缺口；二是提示、观察信息、动作空间和错误历史等组件能否在固定环境中改变任务成功率。WorkArena 的消融研究显示，额外输入或记忆功能并非总能提高成功率，作者报告加入更多特征会令部分较弱模型遭遇更长提示与截断，历史推理也可能固化早期错误。[6] 这种结果支持“信息更多并不必然更可靠”的技术机制假说，不能直接说明员工工作负荷会升或降。

基准不能单独决定组织自动化范围、可用权限、人员配置或 ROI。迁移需先将任务拆成动作、判断、授权与结果四类，确定其中哪些可由现有系统机器验证，哪些需要独立来源。随后在只读或影子环境记录代理预测与实际人工处理，验证任务分布与版本差异；然后选择可撤销、可复核且错误后果有限的子流程试点。每个阶段改变的处理包及其对照条件应清楚记录。若模型、软件界面、源数

SVSAC | sv-sac.org

据和人工复核规则同时变化，现场结果只能归因于该组合政策，不能据此判断某个单一组件的效果。

22.4 聚合分数隐藏了目标任务组合

一个基准的总体成功率是基准抽样规则下的加权平均，不自动代表某家组织的任务篮子。WorkArena 包含多个不同类别，作者表 2 报告的 GPT-4o 成功率在表单 40%、知识库问答 80%、列表筛选 0%、服务目录 77.8%、仪表盘 62.5%、菜单 60%等类别间差异显著；基准总体约 42.7%。[6] 任务实例数量是基准构造的一部分，不能未经验证就作为企业真实任务发生概率。若生产使用场景主要是表单、列表筛选或知识检索，总体平均可产生不同方向的部署预期。

只为展示加权机制，设想有两种目标流程组合，且暂且把基准类别成功率当成各类别现场成功率。这一运输假设并无证据，数值不能视为预测。组合甲由 50%表单、20%列表筛选、10%知识查询、10%服务目录和 10%仪表盘组成，则基准率加权值为：0.50×40% + 0.20×0% + 0.10×80% + 0.10×77.8% + 0.10×62.5% = 42.03%。组合乙由 10%表单、50%列表筛选、10%知识查询、10%服务目录、10%仪表盘和 10%菜单任务组成，得到 32.03%。两者使用相同类别性能输入，十个百分点差异完全由任务权重导致，而非模型发生变化。这个计算只能提醒评价者测量真实任务权重，不能将 WorkArena 成功率直接外推。

现场任务的分类也不应依模型输入字段而定。一个“创建表单”实例可能同时包含身份匹配、权限判断、付款影响和重复记录检查；若把这类工作都归入表单，类别内部异质性很大。分类应依决策所需劳动、错误风险和业务结果确定，事先冻结；之后可报告模型表现是否随这些预定义任务属性变化。若任务标签在看到结果后才重新定义，分层分析会变成寻找能讲成功故事的分组。

22.5 从一步通过率构造端到端状态图

基准可以估计指定环境下的一步或任务完成概率，组织关心的却是输入任务进入系统后最终验收。对一个业务实例，设动作成功 A、语义正确 M、最终业务验收 F。若 F 要求 A 和 M 都成立，则可写成链式分解：P(F)=P(A)×P(M|A)×P(F|A,M)。这是概率恒等式，条件概率不能替换成相互独立的边际率。特别是能通过界面校验的动作可能更容易出现字段完整但含义错误；若研究者把每一环节的平均“准确率”独立相乘，会隐藏条件依赖。

实际评价可把每个固定入口实例留在同一状态图：未尝试、动作失败、动作成功待语义核验、语义错误、人工纠正、最终通过、最终不通过、观察窗未决。主要端点例如“按期最终通过的分派实例比例”，其中工具失败转手工仍依据初始

SVSAC | sv-sac.org

处理组计算；辅助端点分解各状态转换与所需工时。语义标注应由适当的业务责任人依据源文件和规则完成，最好对处理组别盲态；由模型自身生成的解释不能独立充当其正确性的真值。

22.6 将人工修复与资源瓶颈纳入同一算例

以下仅为计算示范，不是公司数据或对 WorkArena 的现场外推。固定 100 件任务，比较假设的纯人工方式与工具辅助方式。纯人工每件平均 10 分钟、95 件最终合格、5 件未决，合计 1,000 个直接处理分钟。工具组中，100 件均由代理尝试；人工监控每件 2 分钟。假设 45 件通过界面动作校验，其中 40 件语义正确、5 件语义错误；复核每件 3 分钟，识别出其中 4 件错误并以每件 12 分钟纠正，另 1 件在最后审计时被拦截并列为不合格。其余 55 件动作未完成，转人工处理；这 55 件均已各投入 10 分钟，其中 50 件完成，5 件仍未决。

按此明确假设，工具组工作为：监控 200 分钟；45 件动作通过的复核 135 分钟；4 件纠正 48 分钟；55 件手工接管 550 分钟；合计 933 分钟。最终合格数为 40+4+50=94 件，另有 5 件未决和 1 件不合格。相对纯人工，直接劳动减少 67 分钟但最终合格量从 95 降到 94，故不能称为净产出改进；若把未决处理和 1 件错误造成的下游风险加回，净价值可能更差。若再假设每周期配置/维护 120 分钟，则工具组已列明人工投入为 1,053 分钟，高于基线 1,000 分钟。这些合计仅覆盖显式列项，尚未给末端审计和未决任务后续投入赋值；完整劳动比较须补齐这些环节，分钟也不是货币成本。这个账本说明动作成功率 45%并不能单独决定运营价值，系统维护和接管成本会改变结果。

该例还揭示审核人员的容量约束：45 次复核若集中在某一小时，队列可能压过处理速度；同样的总分钟分散在五日内，周期结果不同。可记录审查到达、复核开始、结束、退回与重新进入队列的时间，用真实服务时间分布和排班计算是否形成瓶颈。没有排队数据时，不应把“人工复核只用 135 分钟”解释成没有流程延迟；更不能假设新增工作会平均落在每个班次。

22.7 现场迁移验证与拒绝外推条件

从基准走向现场不应只重跑相同任务，而应按目标业务对象抽取连续或概率样本。首先冻结任务入口、类别定义、业务规则与可接受状态；其次在同一软件版本、权限和源数据下运行只读代理，比较动作轨迹与人工结果；之后由不知道组别的评审者按原始请求、源数据和独立业务规则判定语义正确与最终可接受。对每种类型报告分派件数、尝试件数、动作通过、语义通过、人工修改、最终合

SVSAC | sv-sac.org

格、未决及主动工时。若按风险过度抽取，应保存各层抽样概率并恢复目标总体权重。

迁移结果应设置拒答条件。若生产任务超出基准涵盖的输入状态、软件版本发生变化、业务标签无法独立确认、入口缺失使固定分母不可建立、任务由共享队列干扰、严重错误无法由审查拦截，或者没有回退责任人，则基准结果只支持进一步受控观察，不支持扩大执行授权。若连续任务的时间、来源和工具版本与基准差异很大，也不宜将基准分数作为试点先验效应值；更合适的做法是把它用于列出候选失败模式并直接测量现场状态转换。

最后，测试集和上线任务需要隔离。若团队用同一批实例调整提示、选择成功任务和报告最终成功率，分数会包含选择和调参效果。可将开发样本用于修复，另冻结按时间或任务簇隔离的验收集，并在现场试点阶段保留未经调参触及的任务窗口。任何模型、系统或业务规则升级，都要判断原验收集是否仍代表新状态。该候选方案的识别假设与边界均需在实施前记录；它不是已完成的验证研究。

22.8 估计不确定性取决于可重复任务模板

基准中的实例数量不等于独立工作场景数量。WorkArena 作者说明，列表和表单任务可组合出过大的参数空间，因此为部分任务随机抽取并封顶实例数；其评估还在若干种子上重复运行，并对任务分层自助抽样报告标准误。[6] 这有助于描述固定 benchmark 条件下的重复运行变异，但生产组织可能面对更少的任务模板、某些客户/表单结构反复出现、同一软件版本的相关错误。若把大量相似实例当独立样本，得到的窄区间不会反映换任务类型、换版本或换组织后的不确定性。

现场试点应明确随机化和不确定性计算的层级。若一个模板下多个参数化实例共享同一字段逻辑，模板内结果会相关；若同一员工跨多件任务使用工具，学习也会相关；若整组共享审核员和队列，资源干扰又产生团队层相关。标准误需要对应实际独立分派单位或抽样簇，而不是数据表中的行数。具体可预先报告簇数、各簇实例数、方差分层和区间方法；簇数很少时，即使有许多实例，推断仍可能非常不精确。

一个简化的设计效应可说明重复实例的风险：若每簇平均 20 个相似任务、同簇相关系数假设为 0.2，则设计效应约为 1+(20−1)×0.2=4.8。名义 200 件任务的独立样本等效数约为 200/4.8=41.7，而不是 200。这个数字完全是参数化示范，真实

SVSAC | sv-sac.org

相关系数与簇大小须由流程资料估计；该公式也不处理样本权重或复杂的多阶段分派。它说明应优先增加独立任务模板/团队的覆盖，而非只重复同一简单表单。

若真实目标是未来季度任务，报告的主要估计应依该季度任务总体的权重；若目标是对任务类型都等权，需明确这是能力比较目标，而不是业务工作量目标。可并列提供总体按实际权重加权的结果、类别条件结果和在不同权重下的敏感性区间。权重不确定时，应给出合理范围和何种任务组合会改变结论的临界点，而不是用一个观察到的测试集频率假定未来恒定。

22.9 总体任务权重不确定性与类别效果共同决定部署决策建议

现场总体效应通常写为不同任务类别条件效果的加权平均。若类别为 k，部署政策效应为 Δ=Σ_k w_kδ_k，其中 w_k 是目标总体的任务权重、δ_k 是类别条件效应。二者通常都需估计：有限历史窗口可能误估未来任务组合；某一类别内试验样本少，可能误估工具表现。只将观察到的任务频率固定为真值，再只给 δ 的标准误，会遗漏目标组合不确定性。反过来，只对 w 做情景分析但把小簇、模板相关导致的 δ 不确定性忽略，也会给出过窄范围。

一个两类构造例说明方向可变。假设对简单任务效果 δ_S=+10 个百分点，对复杂任务效果 δ_C=−5 个百分点。这不是观察到的 AI 效应，仅为演示。若未来简单任务占比 w_S=0.25，组合效果为 0.25×10%+0.75×(−5%)=−1.25 个百分点；若占比为 0.45，则是 0.45×10%+0.55×(−5%)=+1.75 个百分点。转换点由 10w−5(1−w)=0 求得 w=1/3。若历史期只能确定未来 w 在[0.25,0.45]，部署效果符号未识别。就算类别内效应没有误差，组合权重本身已足以改变结论；只有当新增的类别效应可行集或置信集包含原先使用的点估计，且目标权重可行集保持不变时，联合结果范围才至少覆盖原来仅改变权重的范围。这是集合嵌套的结论；若联合约束排除了原点估计或权重集合同时改变，须重新计算，不能保证范围单调加宽。

类别权重也不能只从 AI 组完成任务反推，因为工具可能改变任务是否进入系统、是否被接受、是否被拆分或合并。预先固定目标可有两种合理方式：以干预前连续期间任务总体的频率定义标准化权重，回答“对既定篮子的效果”；或在每个周期以所有合格到达请求为分母，回答“随需求变化运行时的政策结果”。两者不同，后者允许处理改变任务量、组成及进入率。若分类规则也由 AI 辅助或上线后变更，需保存冻结的原分类与实际运行分类，不能只采用干预后更有利的类别结构。

在对部署市场做外推时，不同目标权重需要具体情境而不是笼统的“全国平均”。可分别设定淡旺季、业务单元、地区或风险类别的权重区间，并检查每一组

SVSAC | sv-sac.org

合的期望最终合格率、净工时和关键错误约束。权重来源可以是完整的历史任务到达日志、已知合同任务量，或管理者无法确定时公开的范围。不能把测试集组成、随机抽到的任务模板占比或供应商给出的演示任务分布直接解释为未来占比。

22.10 模板内相关与抽样不确定性应联合传播

第 22.8 节已说明同一模板中多个实例并非独立。这里进一步区分两种随机来源：一是模板/团队层面抽到哪些簇，即可重复工作场景之间的变异；二是每个簇内抽到哪些任务或执行者，即条件于簇的实例变异。若目标是推广至新的模板，模板是抽样单位；若只需推断既有固定模板中的任务，则任务层目标更窄。混合两种目标会使“标准误”失去明确含义。实际分析要先说明研究总体是这些已知流程，还是将来可能出现的新模板与版本。

一个便于复算的简化模型：类别 k 中有 J_k 个模板，每模板平均 m_k 个实例。对模板 j 先算处理差 δ_kj，再按模板等权或按真实任务量加权得到 δ_k；簇级方差由模板间 δ_kj 的变异估计，而不是把所有任务行直接代入普通独立样本公式。若模板抽样概率不同，需用相应抽样权重；某一模板内多条记录只是增加该模板条件均值精度，并不增加独立模板个数。对单个项目的决策，模板固定、关注本项目任务时，等权模板未必合适；对未来运营总体，按任务量加权则需有任务量框及其不确定性。

为展示权重不确定与相关效应如何同时发生，仍用两类假设数据。简单类有 4 个模板，复杂类有 3 个模板，每模板各 20 条记录；处理差是相对同类对照的百分点变化。简单类四个模板差值假设为+16、+12、+8、−4，等权平均+8；复杂类三个模板差值假设为−15、−5、+5，等权平均−5。若按记录数计算，简单类共 80 条、复杂类 60 条，权重为 4/7 与 3/7，组合点估计为(4/7×8)+(3/7×−5)=2.43 个百分点。若未来总体简单类权重仅知在 0.35 至 0.60，其他条件固定，则组合效应从 0.35×8+0.65×(−5)=−0.45 到 0.60×8+0.40×(−5)=+2.8 个百分点，符号再次取决于权重。所有模板差和权重均是构造数值。

若每个模板下 20 条实例共享同一流程设计，不能将 140 条当作 140 个独立政策复制。按同类别内模板均值作推断，上述示例只有 4 个简单模板和 3 个复杂模板；小簇数会使区间非常宽，且常用簇稳健渐近标准误也未必可靠。可以报告模板均值、留一模板敏感性、随机化单位层面的精确置换范围，以及把目标权重在预设区间移动的结果。若预算允许，增加新模板/团队比在同一模板下重复更多相似记录更能增加关于新场景可重复性的证据。若实际随机化在员工/团队层而模板

SVSAC | sv-sac.org

跨团队共享，也需按产生干扰的更高层级做分派或推断，不能任意选择能得到较小标准误的层级。

统一的不确定性应反映 δ 与 w 同时变化。把任务组合权重设为独立于试验样本的另一来源，可在其抽样设计下估计协方差 Σ_w；把类别效果按随机化模板估计，其协方差记为 Σ_δ。若两者来自独立样本且估计量近似线性，点估计为 ŵ'δ，其一阶方差可写成 ŵ'Σ_δŵ + δ'Σ_wδ，必要时加上二阶项。这个表达式要求权重与效应估计近似独立、目标权重是稳定总体参数、处理不改变目标任务类别定义；条件不满足时需联合重抽样或直接进行设计相容的敏感性分析，而不能机械代入。若每组权重与类别效应在同一处理后日志估计，二者可能相关，简单独立项会漏掉协方差。

更透明的实务结果不是一个号称精确的合并区间，而是三层呈现：其一，按锁定的参考任务结构报告标准化点估计及模板簇不确定性；其二，给出目标任务权重范围下的政策效果上下界或曲线；其三，对每个关键模板展示条件效果和风险门槛。若最悲观合理权重与模板效果仍达不到质量或现金门槛，资源不应按总体平均点估计扩张；若结论只在某个任务混合下成立，应把运营适用范围写成可观测任务比例条件，并在任务结构变化时触发重新验证。

22.11 预先规定联合敏感性算法，避免事后选择权重或模板

联合分析的关键不是找到一个万能公式，而是将任务权重、任务模板、处理分派和标签判断的依赖结构锁定。可以在数据开放前拟定以下计算顺序：第一，冻结进入目标总体的任务框和类别定义；第二，从独立业务到达记录估计 w 或定义 w 的可行区间；第三，在每个类别内按实际分派簇和模板形成 δ 估计，保留所有随机分派任务和结果标签状态；第四，对标签审计用已知抽样概率加权，另对参考标签误差做预设敏感性范围；第五，对每个权重组合重算总体效果，并保存最有利、最不利与转折值。每一步都应有记录、代码或表格可复核，禁止根据最终总体效果挑选新的任务类别。

一个可复现的有限模板程序可以在每个类别内抽取模板簇而不是独立任务行，先计算簇级差异，再以目标任务量权重聚合。每一次重抽样同时抽取模板簇和任务权重估计样本；如果任务分类、权重样本和效果试验共用同一到达日志，重抽样也应在共同的上层抽样单位内进行，以保留相关。若在簇层分派但簇数过少，普通 bootstrap 可能重复少数簇而产生虚假的稳定性；此时报告每个簇的原始结果、所有可行随机化配置、区间粗细和明确的识别限制，比单一标准误更透明。无论采用何种区间，至少列明独立单位、抽样框、权重来源与缺失机制。

SVSAC | sv-sac.org

如果记录只覆盖已接受或已完成的工作，任务权重可能已经受工具政策影响。构造例：两组原始到达均为 100 件，其中简单 60、复杂 40；若 AI 组在入口阶段拒绝一半复杂请求，运行日志会显示已处理任务中简单任务占比提高，而并非需求结构自然变化。把这个处理后权重用于标准化，会把入口筛选造成的构成改变混入类别效率。应从试验前已定义的需求入口记录所有收到、合并、拒绝和转出的请求，并将“是否进入执行”作为单独结果；无法观察未提交需求时，目标总体只能限于可记录的合格到达任务。

结构敏感性也应与统计区间区分。比如权重在[0.25,0.45]内，模板效应可能在有限样本下落于各自区间；用蒙特卡洛抽样得到的分布，只有在给定权重/效应分布模型可信时，才是模型条件下的概率结论。若只知道边界而没有概率分布，应直接给最坏/最好边界，不要在边界内随意均匀抽样后称“效果有八成可能为正”。同样，所谓“未来季度任务权重”若受促销、产品推出或管理者改变需求入口影响，应给季节或策略情境，而不是从历史占比外推出单一权重。

可将结论写成一组能被证伪的条件句：在 22.9 的+10/−5 个百分点构造例中，简单任务权重严格大于 1/3 时加权点估计才为正，等于 1/3 时为零；22.10 的+8/−5 例则须严格大于 5/13，不能混用两例阈值。各关键模板是否满足质量约束另行判断，不能由正均值推得；若复杂任务比例增高，方向可能翻转；若参考审查对严重错误的敏感度低于某范围，质量红线是否满足无法确定。真正的企业试点需用到达日志、独立审阅和实际簇结构估计这些量。本篇不报告任何已执行的估计，也不把构造条件当成通用扩张门槛。

最后，预算决策可采用分层闸门而不是单一合并分数：先要求每个不可替代的高风险类别达到质量约束；再在通过硬门的任务篮子内评估加权人时/吞吐；最后检查当未来权重沿合理区间变化时，净兑现条件是否仍成立。若只有特定模板、特定员工或某个小类别显示收益，应把上线范围限制在其可识别条件内，并保留对新模板的受控验证。这样做不会消除外推不确定性，却能防止一个由任务权重或标签误差驱动的总体平均值掩盖不适用的流程。

23. 过程挖掘、任务记录与“看得见的工作”偏差

23.1 系统日志并非完整劳动时间

组织常希望用系统事件重建流程：任务创建、分派、打开、提交、退回、重开。事件日志确实能提供顺序、队列与重复操作的信息，但它只记录系统可见的活动。阅读外部文件、打电话澄清、和同事讨论、离线核验、等候授权或被其他

SVSAC | sv-sac.org

事务打断，可能完全不在日志里。反过来，页面打开时间可能包含人离开工位、会议或浏览器闲置。将两个时间戳之差当作人工作时，会把等待误记为劳动；只计点击间隔则会漏掉阅读和思考。

事件记录的业务单位也需要稳定。一个请求可能被拆为主任务、子任务和复核任务，重新指派生成新记录；不同团队可能把一次返工记作“新单”或“更新”。若处理组使用自动代理将步骤压缩为单一记录，而对照组保留多个事件，按行数比较会制造差异。应先定义分析单位：以外部请求、需要完成的业务目标，或合同/客户实例为主键；再把系统事件作为该单位的活动链，而不能从系统表的主键反推工作本体。合并规则应在观察处理结果前确定，以免根据结果决定哪些子记录算一个任务。

最低限度的日志方案需记录事件发生时间、被系统接收时间、实例 ID、前后状态、责任角色、工具与版本、输入输出引用、审批/拒绝、返工关联和关闭原因。发生时间与记录时间不同，离线补录或批处理可能令先后顺序错乱。对可能被撤回或重开任务，保留历史状态而非覆盖最终字段，才能重建当时决策者可看到的信息。涉及员工监测的数据治理需依据具体法域和组织政策单独处理；这里的测量框架不构成法律判断。

23.2 选择性记录会改变观察到的效率

若日志只覆盖顺利关闭的实例，失败任务、未完成和外部等待都被删掉，平均周期将向下偏。自动化系统也可能只将成功动作写回业务数据库，失败尝试留在代理执行日志或根本没有落盘，于是同一任务总体在两组呈现不同可见性。另一类偏差来自分母变化：如果工具使员工更愿意接受边界模糊的任务，新增难题进入处理组；若工具政策规定只接简单标准件，则高难度工作留给对照组或人工例外池。未固定受理规则和任务总体，任何简单时间比较都混入了选择过程。

可执行的核查方法是从任务入口台账抽取固定期间内所有符合条件的请求，与各操作日志、系统数据库、质检记录和关闭账单逐条连接。未匹配的请求不能静默丢弃，应列明组别、状态和原因。对入口抽样与深度工时观察分别处理：入口台账用于确定固定分母，工时样本按预定抽样概率测量主动时间，并据此加权。若某类事件更可能缺失，可报告最坏情形边界：例如把所有未匹配处理组实例视为未通过，并分别假设其周期落在某个可解释范围内，观察结论是否改变。没有依据的宽边界不应用单一点估计取代。

选择性指标还可能影响团队行为。若只奖励被记录的关闭件数，员工有动机把任务拆小、提前关闭或把困难工作转出；这是可能机制而非已经证实的普遍事

SVSAC | sv-sac.org

实。研究设计应保留独立来源的质量与重开指标，并在试验前明确“关闭”的定义。若业务部门、管理者和分析者共同设计指标，需记录版本和批准过程，防止试点中途按结果修改标准。标准变更应作为新处理时期，而不是回溯覆盖原始数据。

23.3 以流程图推断瓶颈的条件

事件序列可以发现重复路径和等待热点，例如同类任务多次在起草与复核间往返。但频繁路径未必是浪费：法规审查、双人授权、客户确认都可能是必要控制；罕见路径也未必不重要，偶发高后果例外可能决定是否允许自动执行。流程挖掘所呈现的是编码后的操作模式，不能自动区分必要步骤与可删除步骤。解释需结合任务规范、责任结构和错误后果，并确认日志是否把实际动作完整记录。

一个有用的分析并非“减少点击”，而是问：在哪种任务类别、发生何种输入条件时，交接或返工增加？调整界面或代理后，路径变化是否与独立质量结果相符？若自动化令返工路径减少，但客户投诉、事后更正或外部退单增加，内部流程图看起来更顺并不代表服务更好。可以在随机分派单元下比较路径分布、最终验收和错误后果；但中介路径的因果解释需要额外假设，因为处理影响路径，未测的任务复杂度又可能同时影响路径与结果。将路径分布作为机制描述，比直接宣称某条路径“造成”了生产率提升更稳妥。

23.4 人工工时测量的差异性误差

工时日志的遗漏可能因处理条件不同而偏向某一组。人工流程中，大部分劳动在业务系统内发生；工具组的验证可能转移到电话、文件、即时消息或隐性思考，系统点击数下降却不代表员工投入同步下降。若使用系统活动间隔作为代理变量，AI 组更少的可见操作会被误读为节省；如果人工组需开多个系统而 AI 助手把搜索记录聚合，观测覆盖也不对称。因此，时间测量不能默认误差在组间相同。

可在主试验之外抽取预先确定比例的任务做独立时间观察。观察者按统一定义记录员工、复核、专家和管理角色的主动劳动区间；系统等待、外部等待和并行处理另记。为了避免只观察“顺利案例”，抽样应在入口任务清单生成、并按任务类别和处理组分层后进行；未完成任务也可抽取，记录到观察窗结束的状态和已投入时间。若采用员工自报或屏幕录制，需考虑参与者改变行为或隐私/同意边界；本篇不指定合法性结论，也不把方法当作无误差真值。

构造一个测量偏差反例：100 件对照任务的系统记录每件显示 10 分钟，独立抽样发现平均还有 2 分钟离线核对；AI 组系统记录显示 7 分钟，但抽样发现另有

SVSAC | sv-sac.org

6 分钟人工核验和修复。按日志会得出节省 3 分钟，按全角色主动工时则是对照 12 分钟、AI13 分钟，方向反转。此数值仅说明组间差异误差可改变结论，不代表常见幅度。若无法对全量任务做时间观察，可用抽样估计离线时间均值，并通过不同非日志时间假设做敏感性分析，报告结论在哪些差值下改变。

此外，观察者需要对组别尽可能盲态或至少按相同规则编码；从既有日志开发测时算法时，测试集应独立于调参数据。主动时间、机器执行、等待、排队周期、返工和最终验收应分别列出。若只能获得时间戳区间，名称要准确写作“记录间隔”或“周期时间”，不能标为劳动分钟。只有当主要结论在合理的遗漏时间范围和组间测量误差下仍成立，才可认为工时节省对测量假设较稳健。

24. 生产率主张的边界与组织决策

24.1 工具效果、过程改造与管理选择的分解

部署前后变化通常由三类机制共同构成：工具改变个别任务的信息获取、生成或执行成本；流程改造改变排队、职责与复核；管理决策改变任务分配、质量目标、资源投入及服务范围。原始研究设计若只比较“上线包”与旧流程，可识别的就是这个组合的政策效应。若组织需要知道模型本身的边际贡献，须设计工具、培训和流程组件的对照，或利用可信的分阶段变动。即使组件独立分派，人员学习和共享知识仍可能造成干扰；没有足够样本识别交互时，应报告不确定性而不是选择最有利拆分。

效应估计应同时报告意向处理与实际执行描述。意向处理保留最初随机分派，即使有人不使用工具、工具故障或临时转回手工，这个估计对应“提供这项政策包”的实际管理问题。按实际使用者做结果比较则容易被能力、积极性、班次和任务难度混淆。若用随机分配作工具变量估计遵从者效应，需要明确排除限制、单调性等假设；而工具权限本身可能改变员工行为、管理监督或任务选择，排除限制不一定成立。使用率与结果相关可作为描述线索，却不是效应归因。

24.2 对“人效”的窄化与扩展

“人效”不应等同于个人屏幕前每小时完成数量。一个流程可在个人层面加快，却增加同事审核负担；也可能单位工时不变但减少服务等待、扩大偏远客户覆盖或让专业员工处理复杂事项。这些结果分别属于局部劳动效率、端到端吞吐、服务可及性、质量和任务结构。价值判断应说明谁的时间或风险发生变化、收益由谁获得、是否以未计量的额外劳动换取，以及新增产出是否有需求。

SVSAC | sv-sac.org

例如，若员工每周释放 20 小时并将其用于解决积压，产出可能增加但工资现金流不变；若这 20 小时替代了已计划的外包工时，现金成本可能下降，但应有采购单、排班和付款记录；若任务量增长后员工被要求维持相同单件时限，个人负担可能增加，单纯计数会掩盖劳动强化。三种情况均可能出现“处理速度改善”，其经营解释不同。评价不能只看组织总数，也需观察员工实际任务强度、休息/加班、错误责任分布和质量稳定性；这些属于需获取相应数据的核心问题，不是本篇已有实测结果。

24.3 组织决策应保留反证通道

一项自动化方案的证据链应包含四个环节：基准任务在目标系统中可重复；固定总体中的比较显示工具政策改变了所关心结果；独立验收说明新增或加快的工作达到必要质量；容量与成本账本说明释放资源如何转化为业务价值。任一环节未建立，都不代表工具无用，而是限定了可支持的决策。例如，基准稳定但现场日志与基准任务差异大，应先改进外部效度验证；任务通过率提高但审核负担上升，需比较审核支持方案；产能提升但没有新增需求，应将价值定义为韧性或服务改善，而非财务节约。

停止和扩张决策也应建立在预先定义的红线与可恢复状态上。团队应能识别工具更新、源数据变化、审批规则变化及故障恢复，并在异常时回到已验证的手工或旧版本流程。风险容忍范围需由承担责任的业务与控制职能确定，研究者不能从一个平均处理时间差推导通用自动化授权。对无法量化的法律、安全、声誉或员工体验风险，应明确由谁评估、哪些情形停止；无法给出估值不等于风险为零。

下一章先把上述边界用于受控委托的验证关口，26–27 章再作证据综合与验证优先级比较，检查哪些现有证据可以支持哪些有限推断，以及哪些关键判断仍需真实组织数据。这里的框架是分析和未来验证方案，不是对任何企业已完成部署或已实现收益的陈述。

25. 从可执行动作到受控委托的验证关口

25.1 委托单位不是页面动作，而是责任闭合的工作包

自动化权限常按“建议、草拟、提交、直接执行”描述，但实际授权对象应是一个定义清楚的工作包，而非某种模型或软件按钮。以供应商地址变更为例，系统可以从请求中提取地址、查找供应商档案并填入表单；然而谁有权发起变更、地址证据是否来自可信渠道、是否影响付款、何时生效，属于不同判断。把模型

SVSAC | sv-sac.org

输出写进草稿，和它获准更新主档、触发付款流程，不是同一风险等级。权限矩阵应将数据类型、可逆性、影响范围、授权角色和复核证据逐项绑定。

一个操作可被安全自动执行，至少需要在试验中证明：输入来源和身份可核；目标记录可唯一定位；业务规则有机器可查的约束；提交后有独立回读；错误可撤销或隔离；异常能转给有权限的人；重试不会重复产生不可逆交易。若任何一项不满足，较合理的基线可能是生成草稿、只读检索或建议，而不是自动提交。这里不是普遍的法律或安全标准，而是为具体任务建立可检验责任链的作者框架。业务负责人仍需确定后果和授权界限。

25.2 逐级放权的现场验证步骤

一个分析性试点可按三种执行状态组织。只读阶段，代理观察任务并提出建议，但不写系统；比较其建议与人工实际操作，标记信息缺失、实体匹配错误、规则冲突及高风险字段。草稿阶段，系统可预填，但人工逐字段确认后提交；需记录哪些字段被接受、修改或清空，且不能只把“提交成功”记作完整采纳。有限执行阶段，只允许代理对预先列出的低风险、可逆动作提交，其他动作仍须人工签核。三个阶段的结果不能直接合并，因为人员监督、动作权和暴露损失不同。

试验中的任务分派应在进入自动化前冻结。若一组任务被允许自动执行，另一组由人工处理，因共享用户、供应商、数据库或审查者而产生的溢出要记录。以任务为随机单位只有在业务实例独立、没有跨组共享资源或传播时才合理；同一队列中的实例可能互相挤占处理容量，团队级分派更符合政策比较，但样本有效信息量会受簇数限制。无论哪种设计，主要结果仍按原分派总体计算；拒绝代理、故障转手工、任务重开和无法验收均保留在结果账本。

25.3 错误演练比“成功演示”更能检验回退

在真实暴露之前，可用历史脱敏实例或受控影子环境演练一组具体失效：重复请求、过期资料、身份相似、字段单位不一致、审批被撤回、网络中断、系统升级后按钮位置改变，以及代理已提交但回执超时。每类演练需规定可观察触发信号、代理预期行为、人工接管角色、恢复目标和审计记录。通过演练不证明实际发生率低，只说明在这些已列举条件下流程有可测试的响应。若错误状态不可被外部确认，代理应停止并请求人工核验；盲目重试可能造成重复写入。

构造例：假设每周有 100 个标准变更请求，历史人工处理的一次通过率和总工时尚待基线测量，因此不设预设收益阈值。试点先以只读建议覆盖全部 100 件，人工仍按原流程执行；独立评审按风险分层随机抽取一部分建议比对来源、实体与字段，剩余处理遵守正常质检。下一阶段只对不涉及付款、不可逆状态或

SVSAC | sv-sac.org

身份不确定的请求预填草稿，保存人工逐字段确认日志；任何身份不一致、缺少证据、来源版本过期或重复提交疑点均自动转人工。若候选池中因风险排除而仅有 60 件符合，需报告 60/100 的覆盖率与 40 件退出原因，不能只在 60 件中报成功率而暗示全流程可自动化。

这个构造协议没有声称 100 件是真实样本，也不指定效果门槛。它展示的是估计对象如何逐步改变：全量只读阶段测建议质量；符合资格的 60 件草稿阶段测人工修改后验收和增量复核投入；最终有限执行阶段测受限动作的端到端结果和可逆恢复。若低风险池与被排除池的任务构成不同，外推到后者需要单独设计，而不能依靠一个总体通过率。错误演练、随机现场试点和生产监测各回答不同问题，只有在实际组织证据积累后才能判断权限是否应扩大。

26. 证据综合：为何局部提速可能伴随总体停滞

26.1 四类研究并不估计同一个结果

现有证据看似给出正负并存的结果，首先因为对象和估计量不同。Noy 与 Zhang 文献的正式原段仍待核，不参与本篇已证实效应的比较；BCG 研究随机比较的是咨询人员在若干限时任务中访问 GPT-4 及简短使用说明的结果；QJE 研究追踪客服代理人逐步接入辅助系统后每小时解决量等现场指标；WorkArena 测量的则是网页代理在规定界面状态和机器可验证规则下完成任务的成功率。[1][2][3][6] METR 研究针对熟悉开源代码库的开发者执行真实 issue 的耗时，并在 2025 年报告特定工具、任务和人群下较慢的点估计；其后续方法更新明确说明选择、参与和并行计时问题使后期研究的当前效应信号难以解释。[4][5] 这些研究不能排成同一尺度上的“AI 效率排行榜”。

它们的正负差异可由四个可检验条件解释。第一，任务对模型能力是否处于适配区：范围内的起草、归纳或重复答复可能减少搜索和初稿劳动；需要隐含状态、跨文件一致性或高精度代码修改的任务可能增加核验与修复。第二，使用者是否拥有判断答案的专业能力：经验较少者可能从可用模式中获益，但也可能较难识别流畅而错误的结果；经验人员的效果可能表现为压缩低价值步骤，也可能因频繁核验而抵消节省。第三，流程是否把审核和异常视为生产成本：若研究终点只计初稿或程序动作，质量成本被遗漏；若终点追踪最终验收和全部角色工时，效应可能更小、为零或转移到其他岗位。第四，处理条件是否稳定：工具版本、提示、任务来源、软件环境和参与者预期均可能不同，表面上“模型效果”实为一揽子差异。

SVSAC | sv-sac.org

因此，跨研究综合应比较设计元素而非汇总点估计。若研究采用随机任务分派但工作单元短，内部识别强而流程覆盖有限；若观察真实运营数据，结果更贴近企业工作，但分阶段上线与员工排序可能留下识别假设；若基准自动重置、结果由程序判定，能重复测试特定能力却不观察现实成本。方法优势和局限并非互相抵销，必须按待回答问题选择证据。

26.2 相反解释与可以区分它们的观测

观察到工时上升，至少有三种竞争解释：工具确实让任务更难；处理组接受了更多复杂任务；员工把节省的初稿时间转为更仔细的验证或更高质量目标。它们可以同时出现。仅比较平均时长无法区分。需对任务在干预前定义难度层、记录请求复杂度和分配时点，并分别测量起草、核验、返工与最终验收。如果相同任务层内初稿更快、核验增加、最终质量稳定，则更像劳动从生成转向验证；若处理组高难任务占比增加，原始平均耗时升高未必表示条件效率下降；若返工和缺陷也升高，则“更多验证导致更好质量”的解释受到反证。

工时下降也可能有不利解释。员工可能减少必要检查、把难题转交、以较低质量标准快速关闭，或将劳动留到下一个周期。需要查看固定任务分母的通过率、重开、跨期延迟、投诉、下游纠正和职责转移。若单位处理时间下降但窗口内验收比例降低，或最终通过件的长尾周期增大，就不能只用平均值声称端到端改善。反过来，如果速度没有显著变化但单位任务的严重错误、等待或加班下降，可能仍有经营价值；是否值得投入取决于该质量或风险变化的价值，而不是“速度显著性”单一判据。

效率提升与需求变化也难以从观察数据直接分开。供给能力提高可能令员工处理原先积压任务，或使管理者提高目标；新增量既可能是真实未满足需求，也可能来自任务拆分、重复生成或质量门槛变宽。可通过建立独立需求入口台账、保留任务身份并固定观察窗口来区分；将干预前已存在积压与上线后新产生工作分别报告。若供应量增长而原始需求队列减少，且质量保持，支持释放容量的解释；若工作量只因单位定义或记录办法变化而增长，则属测量效应。

26.3 治理条件作为结果机制，而非附加清单

审核责任、权限和否决能力不仅是部署治理，也会直接改变效果。一个员工即便能拒绝错误建议，若任务指标惩罚拒绝、复核时间被排满或异常升级无人接手，形式上的“人在回路”并不意味着现实中存在有效控制。责任需要落实到明确角色：谁能暂停执行，谁确认来源及对象，谁修复业务记录，谁承担错误后果，

SVSAC | sv-sac.org

谁决定恢复。若审批者同时承担吞吐目标且没有足够时间，快速验收可能是内生反应而非模型可靠性。

因此，评价需同时记录人工复核实际分钟、否决率、覆盖比例、纠错后结果及升级耗时，并观察是否存在真实授权停止机制。否决率高不必然是坏事：它可能反映工具不适用，也可能反映工作人员愿意发现缺陷；低否决率同样可能代表输出优质或检查流于形式。结合盲态抽查、独立来源验证和错误严重程度才可解释。若对照组也使用共享模型或从处理组获得模板，政策对比受污染；此时应明确报告团队层或市场层干扰，而不能仅靠统计标准误处理。

NIST AI RMF 等自愿风险管理材料可作为角色、监测和治理讨论的参考，但框架引用不等于法定要求、效果证据或组织已落实。[7] 研究者需要把抽象的“人工监督”操作化为可观测的否决权限、复核资源、职责和回退流程。治理措施也有成本：增加签核可能降低错误暴露，却增加等待；减少授权可能减轻风险，却让工具只负责低价值片段。其净影响需在流程总体上比较，而非用原则性表述替代。

27. 优先验证的关键问题与适用边界

27.1 优先识别三项容易混淆的贡献

后续研究最有价值的第一步不是比较更多模型，而是将工具、流程标准化与额外管理资源区分。对于重复度高的流程，可以在相同任务清单和审核条件下比较现行流程、单纯模板/标准作业改造、工具辅助以及工具加流程改造四种政策；若样本和团队数不足，则缩小目标到两个主要比较并明确不能识别的交互。这样能检验节省是否来自模型生成，还是来自统一输入、清洁数据、增加培训或管理关注。若没有标准化对照臂，不能把上线包结果归到 AI。

第二步应把适用边界作为预先指定的假说，而非事后寻找成功子组。可在试验前由领域专家独立标记任务所需事实核验、上下文长度、例外率、错误后果和人工可逆性，再锁定少量异质性分层。若工具收益仅出现在低例外、资料可追溯、结果易校验的类别，支持“适配性而非通用能力”的解释；若多个边界层都保持收益且质量/返工无恶化，才扩大适用推断。小样本不宜反复切分用户、月份和任务类，探索性结果需要外部复现。

第三步需要追踪资源兑现。在任务层估计可用工时变化，在队列层测量积压与服务时效，在财务层核对加班、承包、岗位预算和维护支出。每一层的观察时间不同：短期任务试验可能足以检验可用性，却不足以判断稳定运营、需求弹性

SVSAC | sv-sac.org

或现金预算是否调整。若数据只覆盖一个季度，可以报告该季度实际发生的服务和资源变化，不应外推年度全成本或岗位变化。

27.2 低成本且可复核的跨场景协议

一个实用设计可先选三个流程层次，而非抽象职位：规则明确、系统写入可撤销的后台事务；需要专业判断和质量复核的知识工作；涉及外部客户、责任后果或高度例外的流程。每类都从完整入口台账定义固定任务总体，锁定任务创建、分派和完成标准。记录模型/代理/界面版本，收集分派前任务特征，并按真实共享资源决定在任务、员工或团队层随机。比较处理按最初分派分析，人工拒绝、系统故障、任务转移与退出均保留，并区分是否完成及未完成原因。

主要终点可按决策问题选择，例如固定任务总体中按期最终验收的比例；共同报告每件各角色主动工时、队列周期、重开、严重错误和需求入口量。对高后果任务应另设质量边界和独立核查，不依赖平均改善抵消严重事件。有限工时测量采用抽样并记录抽样概率，不把在线时长作为工作时间。一次性接入、培训、提示/检索维护、审核、异常处理与系统维护均进成本账；没有真实可避免现金支出证据的容量，单列而不折成工资节约。

此协议不是通用最佳实践，且不是一项已执行试验。其作用是使失败可解释：如果上游生成提速而最终验收量不变，需查队列、审核和任务组合；若合格吞吐提高但工时未减少，说明服务扩容可能占用了释放容量；若速度与合格率同时下降，需查任务分配、版本或目标变化；若现金支出未动，财务效益尚未兑现。每一种观测都有不同的下一步决策，不应以一个综合分掩盖。

27.3 研究外推的边界

现有研究的参与者和任务与制造及专业服务企业并不等同。写作文献仍待核，不能与已核结果并列承重。咨询实验的限时案例、客服每小时解决量、网页代理基准成功率和开发者 issue 耗时，均有不同的业务含义和技术约束。[1][2][3][4][6] 客服建议可由既有知识库与对话历史支持，网页任务在可复位系统中校验；生产工程、财务审批和客户合同往往需要多个系统的版本一致性、权限链和可追责决策。将来源放在同一文献表里，不代表它们构成一条被检验过的完整因果链。

外推至少需要回答：研究人群是否与目标使用者相似；任务难度和事实可追溯程度是否相当；模型、系统和可调用资料是否相同；质量终点是否捕捉了现实错误；工作是否通过共享队列、审核和外部依赖交接；需求和人员预算是否具有相同反应。若这些要素差异大，应把既有研究当作设计线索，而不是参数来源。

SVSAC | sv-sac.org

尤其不能把一个研究中的任务时间百分比乘到另一家企业工资账上，或将基准成功率解释成岗位可自动化比例。

28. 决策建议：从可用工具到可验证的流程改变

本篇讨论的问题不是生成式工具能否让某些个人任务更快，而是这种变化在什么条件下进入可验收的工作流，并进一步改变吞吐、服务、人员投入或现金支出。现有证据支持有限的条件性判断：在已核的咨询和客服场景中，获得 AI 辅助与部分任务速度或产出改善相关，且有随机实验或实际部署设计提供更强识别；与此同时，开发任务的负向或不确定结果以及企业软件基准中的低完成率说明，收益并非稳定地适用于所有任务、工具和人群。[1][3][4][5][6]

生产率主张必须经过数个可区分的环节：工具能在目标版本和数据上形成建议或动作；内容与操作符合业务语义；独立评审确认合格；工作流的交接、等待和返工没有抵消局部提速；释放资源能够承接真实需求或形成可核对的支出变化。缺少其中任何一层，就应缩窄结论。例如，程序执行更快但最终验收没有改善，是局部能力结果而非端到端产能证据；合格量提升但现金支出不变，可能仍是有价值的服务扩容，却不是已经兑现的预算节省。

对管理者而言，先识别重复工作中最耗时、最可核验且失败可恢复的环节，比先选定模型再寻找用途更容易形成可检验方案。小范围试点应保留完整任务总体、固定质量规则、版本记录、独立复核、各角色投入和失败状态；扩张前再检验共享队列、例外和资源兑现。高风险或需要复杂专业判断的任务应依据具体后果逐步授权，不根据平均任务效应推导普遍的无人化承诺。

本篇没有取得企业内部日志、项目财务账或自建现场随机试验，也没有估计跨行业平均收益。所提出的流程账本、比较设计和授权关口是后续实证工作的分析方案，不是已验证的认证标准。它们的价值在于明确什么观测可支持或推翻“人效提升”：固定分母内合格吞吐、总角色投入、延迟与质量风险共同改善，且资源账与实际决策相符，才支持更广泛的流程生产率解释。若结果只在初稿、任务选择或特定版本上出现，解释应停留在相应边界内。

给企业的下一步建议

AI 普及应从员工真正受困的重复工作开始，先把一条流程做顺，再谈全员提效。因此，我们建议把下一步工作收敛为一项有明确交付物的专项任务。

SVSAC | sv-sac.org

具体起点是：按频次、耗时、差错和责任风险筛选场景，明确人工接管与复核规则，评估端到端周期而非单步速度。在此基础上，管理层可以按证据充分程度、资源投入与业务影响确定优先级，先验证关键假设，再安排后续合作与投入。

要把上述判断用于实际经营，可以围绕具体产品或项目、目标市场、现有能力和决策难点，进一步形成流程机会清单、岗位工具组合、自动化试点与推广路线图。这类工作需要把行业研究、企业诊断与合作条件核验衔接起来，让分析结果成为管理层可以使用的决策依据。

资料与分析范围

下列说明列明资料范围、分析依据与适用条件，便于读者辨别文中的事实、推演和建议。市场进入、交易安排及适用规则仍需结合具体项目核实。

生成式人工智能与流程自动化常以“节省员工时间”作为价值起点，但被缩短的单项任务时长并不自动等于流程周期缩短、合格交付增加或现金成本下降。本篇围绕一个可检验问题展开：在什么工作结构与实施条件下，AI 工具对个体任务的改善能够传递到端到端经营结果；又在什么条件下，新增审核、返工、例外处理或需求扩张会抵消局部收益？本篇以原始实验与现场研究为证据起点，区分任务级速度与质量、员工或团队产出、流程吞吐与等待、以及现金与资源结果。现有研究给出有边界的正向证据：客服现场部署研究报告每小时解决问题数提高；咨询人员随机实验显示同一工作流程内不同任务的效果方向并不一致。它们不能直接回答跨部门交接、排队、审查负担和可兑现成本的净效应。本篇据此建立“任务执行–质量把关–例外与返工–流程完成–资源兑现”的分析框架，并提出相应的测量与比较设计。全文不是企业实测或工具效果保证；所有后续数字算例均作为明确标注的分析构造。

SVSAC | sv-sac.org

参考文献

[1] Erik Brynjolfsson; Danielle Li; Lindsey R. Raymond. Generative AI at Work. The Quarterly Journal of Economics, 140(2), 889–942 (2025). https://academic.oup.com/qje/article/140/2/889/7990658

[2] Shakked Noy; Whitney Zhang. Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. Science, 381(6654), 187–192 (2023). https://www.science.org/doi/10.1126/science.adh2586

[3] Fabrizio Dell'Acqua; Edward McFowland III; Ethan R. Mollick; Hila Lifshitz-Assaf; Katherine C. Kellogg; Saran Rajendran; Lisa Krayer; François Candelon; Karin R. Lakhani. Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality. Organization Science, 37(2), 403–423 (2026). https://pubsonline.informs.org/doi/10.1287/orsc.2025.21838

[4] Joel Becker; Nate Rush; Beth Barnes; David Rein. Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity. METR research report / arXiv:2507.09089v2 (not peer reviewed at time of source page) (2025). https://metr.org/blog/2025-07-10-early-2025-ai-experienced-os-dev-study/

[5] Joel Becker; Nate Rush; Tom Cunningham; David Rein; Khalid Mahamud. We are Changing our Developer Productivity Experiment Design. METR author-institution research update (not a peer-reviewed article) (2026). https://metr.org/blog/2026-02-24-uplift-update/

[6] Alexandre Drouin; Maxime Gasse; Massimo Caccia; Issam H. Laradji; Manuel Del Verme; Tom Marty; David Vazquez; Nicolas Chapados; Alexandre Lacoste. WorkArena: How Capable Are Web Agents at Solving Common Knowledge Work Tasks?. Proceedings of the 41st International Conference on Machine Learning, PMLR 235, 11642–11662 (2024). https://proceedings.mlr.press/v235/drouin24a.html

[7] Elham Tabassi, editor; National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1 (2023). https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10

SVSAC | sv-sac.org