组织与人才升级：AI 重塑岗位与能力建设

作者：SVSAC 硅谷战略顾问委员会

核心判断

AI 转型的组织红利，来自岗位协作与能力重组；单纯购买工具，很难形成持续优势。对企业而言，值得优先推进的是：按任务而非职位口号盘点工作，明确员工、业务负责人和技术团队的责任，结合真实项目安排培训。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 AI；组织学习；技能形成；岗位重组；人力资本；劳动力市场；因果识别

SVSAC - sv-sac.org

AI 转型的组织红利，来自岗位协作与能力重组；单纯购买工具，很难形成持续优势。

我们的建议是：按任务而非职位口号盘点工作，明确员工、业务负责人和技术团队的责任，结合真实项目安排培训。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：工具普及何时成为组织能力建设

“AI 提高人效”容易把眼前任务的处理速度误当作员工能力已经提高。若模型先完成一份分析、代码或客户答复，员工可能更快交付；但是否理解关键假设、能否识别模型遗漏、换一个未见任务时能否独立处理，仍是另外的经验问题。相反，员工能独立解释一个概念，也不保证组织有时间、权限和流程把这种能力用于更好的决策。本篇将能力建设限定为员工在重复任务、迁移任务或异常情境中能够独立完成、检验并改进工作的可观察能力，而非培训签到、账号开通或主观自信。

本次分析关注的不是某种模型在抽象基准上的最高分，而是组织怎样选择任务、安排练习、设置复核和让员工参与 AI 工作，从而同时实现当前交付与未来能力。中心问题可拆成三个可反驳问题：第一，AI 辅助是否提高了特定工作产出，并且提高的是速度、质量、覆盖还是错误控制？第二，哪些使用和训练安排能使能力在撤去工具或换到新任务后保留？第三，即便个人产出或技能改变，组织是否会调整职责、岗位数量、工资、晋升或利润？每个问题的处理对象和观察期不同，不能以“采用率”回答。

题名中的“人才升级”不是预设结果。岗位变化可能是新增监督、整合、数据治理与客户解释任务，也可能是简单任务减少、初级岗位练习入口缩小，或原有职责改名而报酬不变。更高技能要求也不等于员工获得更高工资；工资还受到产品需求、议价制度、劳动供给、培训成本、组织制度及员工对便利工作条件的偏好影响。本篇因此把技能、岗位、收入和企业收益当作需分别检验的结果，而不把“升级”写成所有员工必然经历的改善。

SVSAC | sv-sac.org

2. 结果链：从任务完成到技能、岗位和经营价值

对同一员工，至少要区分四个时间尺度。任务内产出衡量当次工作是否完成及其速度和质量；即时测验衡量刚经历的学习；延迟迁移测验衡量一段时间后在新任务上的独立表现；组织与劳动市场结果则衡量职责、岗位、工资、工时、雇佣和利润的变化。短期任务提速可以与独立技能不变或下降并存，技能增加也可因流程未调整而暂时不改变岗位或工资。任何跨层结论都需要观察中间链条，而不是把第一个结果直接投射到最后一个。

一个可操作的技能测量不能只问“员工觉得自己学会了”。若岗位要求监督 AI 输出，核心能力可能包括概念理解、读取和验证生成内容、诊断错误、识别何时拒绝或升级，以及在信息不完整时解释判断依据。这些维度对任务安全和工作结果的作用不同，且不能用一个通用测验代替岗位标准。以编程为例，写出能运行的代码与理解异步执行、阅读代码并定位故障并不相同；在财务、法务或客户服务中，也应依据岗位的真实失败模式定义独立能力，而不是照抄编程测验。

当前工作质量与独立能力也必须分别打分。任务由员工与 AI 共同完成时，最终文本可能准确，却无法据此判断员工是否能识别同类错误；反过来，员工测验表现较高，也不能说明生产工作中的输出达到合规或客户要求。两种分数需要各自的基线、盲评与结果标签：任务成果由不知处理分组的领域审查者依预先规则评价，迁移测验则在工具关闭或受控访问条件下完成，并确保测试题未被训练过程直接泄露。

组织结果进一步包含产量、返工和复核占用。若工具增加了每名员工每天的草稿数，却让高级员工承担更多审校、纠错和责任确认，团队净产能可能上升、持证专家工作负荷却恶化。记录“员工产出”时必须同时记录质量返工率、升级次数、审查分钟、失败后果和跨角色等待；若只把工具使用者生成的文件数计为生产率，就会将被转移给他人的劳动从分子或成本账中遗漏。

工资、岗位数、工时及企业利润位于更远的因果链上。生产率收益可以被用于扩大产量、降低价格、减少招聘、增加审核、提高质量或提升资本回报，具体落点取决于需求弹性、劳动供给、议价制度、合同和竞争。雇员平均工资接近零既不证明没有任务重组，也不证明岗位没有异质变化；单个岗位任务变化又不等于净就业减少。企业利润还需扣除模型、数据、培训、合规、实施、监督和过渡成本，不能由工时节省直接代替。

SVSAC | sv-sac.org

3. 关键机制与相反路径

第一条机制是“支架式练习”：AI 提供即时示例、错误反馈或逐步提示，员工仍必须尝试、解释和修正，因而可能扩大练习次数并加快纠错。相反路径是“认知外包”：系统一次性给出完整答案，员工跳过问题分解、失败诊断和概念抽象，交付更快但练习减少。两条路径都可能伴随高使用率，所以使用次数本身不区分学习。决定方向的因素可能包括任务是否陌生、错误是否可见、员工是否需要主动说明理由、提示是否逐渐撤除，以及是否存在独立迁移测验。

第二条机制是任务分工变化。AI 可以降低某些重复写作、检索或整理任务成本，释放员工去处理例外、关系、判断或整合工作；但若例外工作依赖长期积累的领域知识而基层员工被剥离练习，组织可能一面减少简单任务，一面扩大少数专家的复核瓶颈。所谓“高技能任务增加”应分辨是新任务的真实净增量、原任务改名、更多审核责任，还是原岗位员工承担额外的无偿工作。不同解释对能力投资、岗位评价和劳动负荷的含义相反。

第三条机制涉及谁获得练习和谁承担风险。初学者通常更可能需要反复处理基础错误来建立判断基准；资深员工则可能更有能力发现边缘失效，却也可能被安排为团队中所有 AI 产出的最终审核者。若绩效考核只奖励按时交付，员工会倾向于接受看似可信的结果；若只惩罚错误，员工可能减少探索、隐瞒使用或将问题推给后续审阅。这里描述的是待检验的组织机制，不是关于员工普遍行为的既定事实。数据设计需要分别记录练习机会、复核时长、错误归属、拒绝 AI 的权利和不同资历员工的收益。

第四条机制是工资与岗位调整的制度传导。即使 AI 让单个员工更快完成任务，雇主也可能把节省转为更高产量、较短工时或更大利润，员工薪酬未必改变；相反，若集体工资协议、资质许可或客户收费模式约束调整，个人效率变化会先体现在排队和服务覆盖，而不是工资。劳动力需求还会因价格下降和新任务而扩张，部分抵消某项任务替代。因而“可自动化任务占比”本身不能推出岗位净减少；需要观察劳动需求和任务组合在实际组织条件下怎样响应。

4. 证据与方法：比较不同设计，而不拼成单一效果

第一类直接证据是 Humlum 与 Vestergaard 的 RFBerlin Discussion Paper 078/26 作者稿。研究把丹麦高 AI 暴露职业的员工调查与雇主–雇员行政面板连接，并在个人和工作场所层面对比采用者与非采用者在 2022 年 11 月 ChatGPT 发布前后的变化。它不是随机试验：差分解释依赖若无聊天机器人时组间趋势可比的假设；

SVSAC | sv-sac.org

采用时间由公开发布时点锚定，调查没有精确记录每位个人的首次使用月份。作者还指出，采用者−非采用者差分只能识别相对变化，若比较组同样受益于共同市场变化，该共同影响会被差分消去。本篇只采用其工作稿版本及所读范围，不把它写成所有国家、所有岗位或训练方案的效果[1]。

该研究的摘要报告，工人自报使用收益、企业倡议和新 AI 相关任务较普遍，但工资与登记工时的平均差分估计接近零；不同设定下，置信区间能够排除的较大变化幅度并不相同。摘要还指出部分采用者转入 AI 关联度更高且工资溢价更高的职业，但人数不足以推动总体平均收入。这组结果适合提出“工作内容改变早于平均工资变化”的问题，不足以断言每个岗位已经升级、某类员工没有受损，或高使用者今后必然获得薪酬提升[1]。对岗位和技能结论仍要读各自的变量、样本和识别条件，不能用工资零估计替代。

第二类证据是 Shen 与 Tamkin 关于 Trio 异步编程库的随机实验。52 名参与者有一年以上 Python 经验、每周使用 Python 并未接触过该库，26 人被分到 AI 辅助组、26 人对照；参与者完成两项新库任务后，在不能使用 AI 的条件下完成概念、读码和调试测验。主试中，任务完成时间差没有显著改善，而 27 分测验平均差约 4.15 分，AI 组得分较低。其对象是短时、陌生库学习，不是企业正式培训或长期职业迁移；参与者也不是编程零基础者。研究的六类使用行为来自处理组内的事后分类，不是六个随机训练臂，不能据此断定某种使用方式必然造成学习差异[2]。

两类研究彼此补充但不能互相验证相同因果效应。随机短期实验更接近识别“这一工具可用条件下、在这个学习任务中”的处理效应，样本小且外部效度有限；劳动力市场差分观察了更长的职业结果和真实采用，却依赖不可完全检验的趋势假设、处理时间定义和比较组边界。两者的“处理”也不相同：一类是短时允许使用 AI，另一类是员工与企业在真实市场中逐步采用聊天机器人。因此本篇不会将两者样本合并或宣称一条完整因果链，而是用差异反推出更细的可检验设计[1][2]。

研究方法由三部分组成。第一，对原始论文的版本、研究对象、设计、分配单位、结果测量、读到的页段和限制做有界证据表；摘要、方法选读和作者工作稿均如实标记，不把标题和摘要当作全文。第二，区分事实、作者解释、本篇机制推论与构造方案，主张逐段映射来源；第三，用显式假设画出从任务处理到技能、岗位与收益的路径，指出哪些参数需由组织的任务日志、盲评、培训记录和成熟工资/财务数据识别。没有这些数据时只给研究协议和可反驳命题，不给企业效果估计。

SVSAC | sv-sac.org

5. 可检验命题与文章结构

命题一：如果 AI 仅提供完成任务的即时答案，而使用者不需要解释、独立修正或延迟测验，则任务内提速可能不伴随迁移能力改善；若员工必须先尝试、在提示后自主说明并在撤去 AI 后解决新例题，这种支架可能改善独立掌握。可检验比较应至少将工具访问与主动学习设计分开，而不是把所有“AI 培训”打包成一个处理；并报告任务质量、时间和延迟迁移，不以其中一项替代另外两项。

命题二：练习机会分布可能是岗位变化的中介。若重复性初阶任务减少，员工是否获得新的判断、审核和整合任务，取决于组织是否建立指导、可见反馈和晋升路径。相反，如果基层工作的练习入口收缩，而异常审阅集中到固定资深人员，平均产出提升仍可能与人才梯队变窄并存。该命题要求按资历追踪任务组合、练习次数、错误处理、专家审核小时和晋升，不允许从整个部门均值反推每位员工都受益。

命题三：技能变化向工资、岗位数量和利润的传导受需求、价格、合同与议价机制调节。若产能增加后新增需求充分，员工可能转向更高价值任务；若需求固定、岗位任务可替代且组织选择减少招聘，岗位入口可能收缩；若固定工资短期不变，效率收益可以先留在雇主利润或客户价值中。上述方向都是条件假设，不是本篇已观察到的事实。因而评价应至少分为员工层任务和能力、团队层复核与产出、企业层现金成本和利润、劳动力市场层岗位与工资，并预先界定结果期限。

本篇其余部分将依次分析能力测量和技能迁移、培训与工具使用的分离识别、初级岗位和专家监督的任务重组、工资与经营价值的传导边界，再提出可执行的组织比较方案。所有跨行业扩展都需重新指定“独立能力”“高风险错误”和“合格岗位结果”的定义。本篇不声称生成式 AI 已经普遍造成技能退化、岗位替代或薪酬提升；它要说明的是，组织若要把工具采用称为人才升级，就必须同时展示当前产出、可迁移能力与利益如何分配的证据。

6. 把工具使用与训练设计分成可辨认的处理

“获得 AI”不是单一处理。至少要区分：员工是否有系统访问权限、是否被经理鼓励或要求使用、实际在哪些任务上调用、是否接受关于验证和错误识别的结构化训练，以及工作流是否赋予其拒绝、改写或升级输出的权限。权限开放只改变可用性；实际使用是员工对任务和工具的选择；培训处理还需要有明确定义的

SVSAC | sv-sac.org

教学内容、教师或材料、练习和反馈。若这些条件同时改变，前后差异只能评价一个组合方案，不能把作用归于模型本身或“培训”二字。

一个可识别的组织试点可以把 AI 访问与主动学习设计分开：在不降低现有安全标准的前提下，按信息共享团队分成“原流程”“仅访问 AI”“AI 加结构化练习与反馈”三组；如果规模允许，再单独设只有练习而没有 AI 的组，用于区分反馈训练本身与 AI 可用的交互。应在启动前冻结模型版本、岗位任务、允许的数据、培训小时和业务规则。主要估计量按最初分派比较，不按最终是否喜欢工具重分组；使用日志用于检验实施链，而不是把高使用者与低使用者差异误称为随机使用效果。小团队之间共享资料或导师时，随机单位应覆盖真实的共享边界。

培训处理还必须说明练习内容是否与主要结果测试重叠。若课堂直接讲授测试题，测到的是题目熟悉而非迁移；若训练只教如何写提示，不能声称员工已掌握领域判断。岗位相关的技能组合可分为概念识别、错误诊断、来源核验、升级判断和工具关闭后的独立完成。每项应由未接触分组信息的评审者按预先制定的锚定量表评分，并保留原始判断与复核分歧；对涉及安全、法律或客户权利的错误，普通平均分不能覆盖严重失误。

技能结果的观察时间也需要预先设置。刚完成任务后的测验易受短期记忆影响，立即离开 AI 一次也不能代表长期保留；较晚测验则可能混入新的工作经验、额外培训和真实项目选择。研究设计可安排多个同等难度的延迟测点，使用平行题组并保留未见案例，记录期间额外使用和教学。具体间隔要按岗位的任务循环、技能衰退速度和法规要求确定，不存在适用于所有职业的统一“延迟测验天数”。每个测点应说明评估的是保持、迁移还是实际接管，不能把三者合并为“技能提升”。

练习机会分配本身也是结果。若 AI 辅助让少数员工掌握更多简单任务，其他人就可能失去低风险练习；若高级员工只收到异常件，平均处理质量可能提升却没有可比较的基线任务。试点应记录员工进入哪些任务、谁决定分派、未经 AI 的练习次数和失败后的返工归属，并按资历、岗位和可访问性报告结果。员工的拒绝权或替代培训渠道若因组别而异，就不是无关背景，而是处理方案的一部分；任何能力差异都可能由机会配置而不只由模型交互产生。

7. 组织层识别：采用选择、共同变化与成熟结果

在真实企业中，员工往往不是随机选择是否使用 AI，经理也可能优先把系统配置给产能较高、数字技能较强或面临特殊目标的团队。因而“使用者表现更

SVSAC | sv-sac.org

好”可能反映先前能力和任务组合；“受训组进步较快”也可能来自管理关注、额外人员或更容易的工作。丹麦研究通过面板、个体和工作场所比较及前趋势分析处理部分选择问题，但其处理时点仍是公开发布、且比较组可能受到共同影响[1]。企业研究要把训练名额、工具权限、经理鼓励、人员更替和任务难度的分配原因保留下来，而不是仅保存最终使用记录。

若不能随机分派，可以预先安排分阶段上线并记录各批次开始的依据，比较尚未上线与已上线团队，但必须检查推出顺序是否与基线趋势或业绩目标有关。不同团队的工作季节、客户量、产品版本和人员流动可能造成同期变化；前趋势相似不能证明所有后续冲击也相同。将同一部门的处理与对照分开也可能因为员工共享提示、模板或审阅者而污染比较。应按干预可能外溢的团队或工作站定义单位，追踪跨组协作与共享资源，若有明显污染就把估计目标改为实际部署组合，而不是声称孤立的个人工具效果。

最后，组织结果需要追踪原始员工队列和实际岗位，而不能只看期末仍留任者。若 AI 改变了离职、内部转岗或晋升，条件于“仍在同一岗位”会选择受处理影响的人群；工资平均值也会因人员构成变化而变化。可同时报告固定原始队列的结果、在职岗位的组成变化、雇佣和工时流量，以及成熟期的工资和企业现金流。对离职、退休、合同变更和岗位重分类要事先规定口径；没有长期跟踪时，就把能力保持或岗位稳定标成未知，而不是把尚未观察到的影响写成不存在。

8. 技能形成取决于工作中谁在做判断

任务完成得更快，并不能说明学习发生在何处。至少有三种不同机制可能产生表面相同的“产出改善”：模型代替员工完成原本需要练习的步骤；员工把低价值步骤交给模型，把时间转向诊断和校验；或者模型提供示范，员工在反复比较、修正和反馈中逐步掌握方法。三种路径可能同时存在，但对下一次无工具任务的预测不同。若收益只在模型在线时出现，技能存量未必增加；若员工能解释为何接受某一建议、发现何时它不适用，并在延迟任务中独立重现关键判断，才有更直接的能力证据。

因此，组织学习不应以登录次数、提示词数量或“AI 使用熟练度”自评替代。能力测量需要把任务拆成可观察的判断节点：识别输入是否完整、选择相关规则、核对输出、处理异常，以及向有权限的人升级。每个节点都要说明员工是否在模型前做出判断、模型是否给出解释或仅给出答案、员工是否有机会看到结果与事实标签的差异。若系统默认自动填满空白或直接生成最终答复，员工可能只

SVSAC | sv-sac.org

练习接受与编辑；若强迫逐项手工复算，则可能抵消效率收益，甚至把低风险的重复劳动重新交回员工。

一个可检验的培训设计可将同一岗位任务随机分配到“仅访问”“访问加先行独立尝试”“访问加事后错误复盘”三种训练流程；这些流程应在模型版本、参考资料和工时上尽量一致。训练阶段记录真实任务质量与时间，随后安排不允许使用模型的即时测验、数周后的延迟测验，以及结构不同但需要相同判断能力的迁移任务。主要能力终点应是盲评的正确决策和错误检出，而非员工能否复述模型解释。若先行尝试组只改善测验却拖累生产质量，或者复盘只提高熟悉题得分而不迁移，训练方案并未满足“交付与能力兼得”的目标。

这类设计仍要防止把测验熟练误作能力迁移。重复做同一题会带来记忆效应，培训组也可能更熟悉评分规则。可以预先形成等难度题本，按题本版本分层随机，使用不在练习材料中的情境测迁移，并保留任务级评审记录。任务难度不同或评分员不盲时，均值差可能来自题目或评分标准，而非培训。测验也不能无限加长：若每名员工耗费一小时做离线测试，组织观察到的是任务学习与测评负担的组合。研究报告需分别给出测试成绩、实际业务质量、员工占用时间和延迟流失，明确何者是组织需承担的成本。

同样重要的是负面结果的解释。如果 AI 可用组在立即测验中表现较弱，这可能源于依赖、减少练习、认知负荷增加，也可能是 AI 组把更多精力投入生产任务、而测验奖励不同知识。识别这些路径需要任务过程日志与额外的随机化组件，例如在相同 AI 访问下随机设置“先答后看建议”或“先看建议后解释”。单纯对处理组内使用方式回归，会把初始能力、任务难度和个人偏好混入关联。因而任何“主动使用者学得更多”的发现都需要警惕自我选择。

9. 经验扩散、岗位差异与“去技能化”假说

技能并非只存在于培训课上。熟练员工处理边缘问题时形成的诊断顺序、澄清问题和例外处置，可能被 AI 建议显示给初级员工；另一条路径则是组织把高手经验压缩成默认答案，却不再让他们解释边界，之后模型在罕见或新型问题上反而缺少可更新的反馈。前一种机制是经验扩散，后一种是经验抽取与反馈枯竭。它们需要不同证据：初级员工是否在无工具场景更快达到熟练标准，高手是否减少贡献或处理复杂异常的能力下降，以及新问题的质量是否随时间变化。

Brynjolfsson、Li 与 Raymond 在一家大型客户支持运营中研究生成式 AI 建议工具。期刊版描述 5,172 名客服人员，结果指标包括每小时成功解决的客户问

SVSAC | sv-sac.org

题，报告平均生产率上升约 15%，且新手和低技能员工收益较大；最高技能人员的速度增幅较小，部分对话质量指标略有下降[4]。研究还使用 AI 系统临时故障时的工作表现，发现接触与采纳建议较多的员工在模型不可用时仍有相对基线提升，并报告了经验曲线加快的证据。该结果与“AI 只会替代全部判断”的简单说法不一致，但停机并非随机分配学习剂量，实际采纳程度也不是随机处理；它不能等同于预注册的离线能力测验。

其产业机制很具体：客服任务有持续的对话反馈，系统能根据问题描述给出候选回应，员工仍可编辑或忽略。建议如果把资深员工的澄清步骤和语气规范传给新手，便可能缩短爬坡期；而结果指标“每小时已解决问题”同时受并发聊天数量、处理时长和解决率影响。若只看速度而忽略复开、升级、客户问题是否真正解决，岗位会被导向更快结束会话而非更好诊断。论文的指标分解和客户情绪结果能提示机制，却不证明每个组织都能复制：客服语料、建议质量、主管权限、客户任务复杂度和劳务外包安排均不同[4]。

更有意义的异质性分析至少区分“经验年限”“当前任务能力”和“可获得支持”。新手可因获得高质量示范而进步，也可能在模型答案过度自信时形成错误规则；老员工可能得到较小的平均速度收益，却承担更多例外复核、培训和升级工作。用资历单一变量分组会把这些维度混在一起。岗位能力画像应记录基线任务质量、复杂例外处理、复核准确度和同事求助网络，再检验 AI 是否改变各项任务的时间及错误率，而不是给“高技能”和“低技能”贴稳定标签。

“去技能化”同样不是一个可直接观测的结果。员工在使用工具时少做某类步骤，不等于其长期失去能力；相反，完成更多、更难任务可能增加隐性学习。更适当的反证是：在事先定义、未被 AI 接管的共同任务中，随机撤去工具或安排故障演练，比较质量、错误诊断时间和升级是否适当；同时对岗位任务做重复测量。如果模型停用后表现差距随暴露时长扩大，且在同难度、同资源任务中持续存在，依赖解释才增强。若差距只出现在模型异常期间、伴随信息系统或流程中断，则可能是流程脆弱性而非个人技能衰退。

10. 岗位重组不是职业消失的同义词

技术作用于任务，组织再把任务重新组合成岗位。生成式 AI 可以减少起草、搜索和格式整理的时间，却提高事实核验、例外识别、客户沟通和数据授权的重要性。员工总工时可能不变，但任务构成和判断责任改变；也可能任务更快后需求增加、岗位扩张，或管理者把节省的时间变成更高产量目标。仅凭某职业

SVSAC | sv-sac.org

含有大量语言任务来推断岗位被替代，忽视了这些任务在完整生产流程中的互补关系和组织需求的反馈。

岗位分析可采用“任务–决策–责任”三层记录。任务层描述员工实际执行的动作；决策层记录哪些信息被解释、哪些选择被模型推荐以及谁能否决；责任层则标明出错后谁有权暂停、补救和向外部解释。比如，AI 生成财务分析初稿可能替代数据清理的一部分，却让分析师承担口径校验、异常说明与业务负责人沟通。如果员工没有读取源数据、核对口径的权限，责任增加而能力和控制权不增，岗位升级便只是名义变化。

岗位任务清单应按时间重复测量，而非部署前后各问一次“工作是否改变”。一种可行记录是连续数周抽样日志，给每项任务编码为：由人完成、由 AI 建议后人完成、AI 生成后人审核、自动化执行，以及因 AI 新出现的活动。时间比例不能简单相加为“效率提升”：任务转移可能节省某岗位工时，却把审查、异常处置和数据维护转移给别的岗位。工作量还应报告总件数、每件时间、返工率及难度构成，以免产量增加造成平均工时下降的错觉。

对组织而言，岗位重组有明确的反例。若原本由两类员工协作：初级员工搜集和清理材料，资深员工判断；AI 接管整理后，组织可能让初级员工承担核验，却没有足够专业训练。此时流程看似岗位内容更“分析”，实际是把更高风险判断下移。相反，如果模型把常见错误和解决过程可追溯地呈现，初级员工获得可核验的示范，而资深员工把时间投入复杂案例，可能形成真正的技能阶梯。两种结果都可能显示同样的“每名员工处理量上升”，必须靠决策权限、独立能力测试与错误责任记录区分。

员工参与也有识别和实施双重价值。参与者更容易发现任务分类遗漏、隐性返工以及模型在特殊客户、设备或语言条件下的失败；但参与试点的员工可能本就技术兴趣较强，不能代表全体员工对岗位重设计的反应。若只让志愿者参与，组织可以获得高可用性测试，却不应据此估计普遍采纳或技能结果。可在制度上把参与角色与评估角色区分：参与者提供问题清单和使用反馈，结果评价则由独立人员按事前标准完成，并保存未参与者的负面体验。

11. 绩效扩散、能力积累与收益分配的张力

单个员工完成任务更快，既可能释放学习时间，也可能使管理者提高服务量目标。后一变化将局部效率转为组织产能，却未必改善员工负担或收入。若绩效制度按处理量计酬，员工可能优先解决容易归档的任务；若奖金完全与质量挂

SVSAC | sv-sac.org

钩，员工可能增加谨慎检查而吞没速度收益。绩效指标因此不仅“测量”技术效果，还会改变技术被怎样使用。部署后观察到的流程和技能是政策、指标与员工应对共同产生的结果。

可以把从任务产出到收益的链条写为：

AI 访问与工作设计 → 员工选择/采用 → 单项质量和耗时 → 复核与升级负荷 → 团队吞吐和用户结果 → 排班、岗位与薪酬政策 → 财务净收益。

每个箭头都可能被预算、需求和权力关系阻断。例如，节省的十分钟可能被新增业务占满，也可能转化为延长休息、承担培训或缩短等待时间；这些分别影响员工福祉、学习和用户价值，不能都记作现金节省。若企业将服务扩张视为收益，需与未部署时的需求反事实比较；若将裁撤岗位视为节省，则还要计入培训、转岗、裁撤成本、质量后果和劳动合同约束。

Noy 与 Zhang 的预注册在线实验为短任务产出提供了另一种参照：444 名受过大学教育的专业人员被随机分配是否在第二项职业写作任务前获得 ChatGPT 注册指导与使用许可，盲评者评价输出质量。随机的是提供政策而非实际使用，作者工作稿报告任务时间下降、质量提高，并显示处理主要改变了起草和编辑等任务组成；但实验任务短且背景信息受提示限定，不能回答组织长期是否培训员工、职位数如何变化或工资如何调整[3]。同一人先做基线任务再进行处理任务，为观察任务层差异提供了基准，却不自动测出数月后的独立技能。短任务即时质量的变化，应与技能迁移、岗位改变和收益分配分开陈述。

一个必要的分配问题是，能力提升的收益由谁取得。若新手能快速达到熟练员工的即时产出，组织可能扩大新手招聘，也可能减少资深岗位需求；若高手的经验是模型建议的主要来源，他们可能获得培训、产品设计或复杂案例职责，也可能承担无偿知识整理。两者不能只从某一时点的平均工资推断。岗位等级、内部晋升、入职后的工资轨迹、人员离职与任务内容需要联合观察；否则同样的平均工资可以掩盖“初级上升、专家下降”或“工资稳定但晋升受阻”等分布变化。

企业也需区分成本节约与能力投资。正式训练占用工时可能暂时降低吞吐；审计样本、资深员工复核与错误回滚都需要资源。若把这部分投入忽略，短期净值会被高估；但若将固定工资按全部释放工时折成现金节省，又会夸大可避免成本。较可靠的账本分别列出可取消的外包/加班支出、重新配置的工时容量、模型与系统成本、培训和复核现金、员工结果以及未来能力指标。组织可以选择暂时接受现金成本增加，以换取未来任务覆盖或人员能力，但那是投资判断而不是当期生产率事实。

SVSAC | sv-sac.org

12. 如何验证工作中学习，而不把相关性写成培训效果

现场学习评估的难点是，谁使用模型、谁被安排复杂任务、谁接受辅导，往往共同决定结果。若经理优先把工具交给高绩效员工，使用与产出间的正相关会含有选择；若模型建议在难题中更常出现，使用强度又可能与任务难度正相关，简单回归甚至会把有效工具估成无效。即使部署由团队分期进行，团队领导力、客户类型和排班变化也可能与上线时点一起变化。

一种可执行的方案是在同一岗位、同一模型权限和统一任务池中，以不共享主管/排班的团队簇为分配单位，随机分到标准访问、访问加“先独立尝试”、访问加“事后错误复盘”，并保留无访问的同期对照。若无法设置无访问臂，至少随机比较不同学习设计，并将结论限定为“在可访问条件下，何种工作流程更有利”。随机化前按岗位和基线技能层分区组；不得把同一团队员工拆到不同臂后再假设互不影响。培训和材料在团队内会传播，团队层分配可减少但不能消除跨团队污染，因此需记录共享导师、文档库和轮岗。

主要终点应预先分开：第一，带工具任务的质量、时间和返工；第二，立即无工具测验；第三，延迟无工具测验；第四，新情境迁移和真实工作错误。团队是随机单位，分析需尊重簇分配，按原随机组比较意向处理效应；若员工不参与培训或模型停机，仍保留在其原分派组，并另行报告依从情况。无工具测试若只覆盖低难度知识，不能代表全部能力；新情境也需避免比训练题明显更难。盲评者应看不到组别与姓名，评分前以共同样例训练并报告评分员差异。

样本量不能凭“每组 30 人”套用。设计参数应来自试点中团队数、团队大小、簇内相关、基线结果方差、测验失访和希望检出的最小实际重要差异。若可分配团队很少，员工数量再大也无法制造很多独立随机单位；结果应给出设计效能和区间宽度，并承认无法排除较小但有意义的效果。多重终点需事先指定主终点层级或控制多重检验，不能在速度、质量、知识、迁移、满意度中事后挑一个显著结果。

一个更谨慎的判据要求短期交付不恶化、延迟迁移指标有改善，且新增审核和培训资源可承受。若交付提升而迁移无变化，结论是工具提升任务表现，尚未证明能力积累；若迁移提高但交付下降，应由组织判断是否值得投入，不能把学习优势偷换成已实现业务收益；若两者都改善但专家复核时长上升，还需确定扩大规模后是否会出现审核瓶颈。检验应能被否证，而非把任何积极结果都解释为人才升级。

SVSAC | sv-sac.org

13. 练习界面会改变学习，不只是改变模型可用性

组织常把“给员工开通模型”和“培训员工使用模型”当成同一项干预，但访问权限和学习脚手架是不同的处理。前者改变员工能否调用工具；后者改变工具如何回应、员工先做什么、错误如何被纠正。若两者同时变化，结果即使改善，也无法判断是模型、指导材料、练习结构还是管理关注造成。若短期表现上升而无辅助能力下降，也不能只归因模型本身，因为在任务中练习的顺序和答题反馈已经改变。

Bastani 等关于数学学习的论文及其更正保留为待核线索。[5] 本篇未取得相关原文，不把约千名学生、班级安排、四次 90 分钟、17%考试差或更正仅涉及隶属的说法作为已核结果。这里提出的可检验问题是：在相同模型访问下，直接给答案与分步提示、先行尝试等流程是否改变无工具学习表现？这一问题需独立比较，不能由工具名称或二手摘要裁定。

对工作学习的分析应区分模型访问与额外教学投入。若一种辅导方案同时加入标准解、常见错误、教师反馈和分步提示，其效果属于该组合，不能单独归因于一句提示词。本篇没有核到 Bastani 等对应的材料制作和课堂实施原段，不用其未核细节证明这一路径；以下条件性讨论是作者的研究设计，而不是已观察到的企业效果。

把该机制迁移到工作场景，需要先问组织是否知道正确答案和常见错误。规范的报销审核、代码库中的既有测试、设备维护记录可能有可追溯答案；跨职能战略判断、客户谈判或新产品设计则未必能预先列出“正确解”。如果标准答案由模型或经理事后写下，再用它证明培训有效，评估将循环自证。没有独立证据的复杂任务应以专业判断、约束满足和异议处理评分，而不是制造看似精确的标签。

因此，训练脚手架应依风险和可验证性分层。对于低风险、标准化任务，可要求员工先独立尝试、再查看模型结果并解释差异；对于高风险任务，可限制模型给出结论，要求展示依据、列出不确定项并等待有权限人员批准；对于开放式任务，则可让模型提出备选路径而不评分“唯一正确答案”。这些是待评估的方案，不是已证实的最佳实践。若脚手架耗费的时间超过减少返工的价值，或者员工只学会按提示流程操作而无法处理未覆盖情境，组织应停止扩展或重新设计。

SVSAC | sv-sac.org

14. 把“学习”落实为岗位中的独立判断

企业培训往往以完成课程、通过知识测验或达到工具熟练度作为结业标准，但这些指标与能否在真实工作中接管任务并不相同。一个人可能知道某条规则，却无法在信息不完整时决定是否暂停；也可能不熟悉模型术语，却能通过查证、求助和升级保持任务质量。能力标准应从岗位中不可转移的判断责任出发，而不是从供应商课程清单倒推。

可将岗位能力拆成四层：完成标准案例、验证模型结果、识别超出适用范围、在异常时安全接管。每层都要定义可观察证据。例如，验证要求员工能找到原始数据、复算关键字段、记录不一致；识别适用范围要求员工说出“哪些输入或情境缺失时不应采纳”；接管要求员工能在模型不可用时完成核心任务，或明确升级给谁。课程完成记录只能证明接触过材料，不能代替这些行为证据。

这套分层也改变了“专家”的定义。资深员工不必是所有模型知识的专家，但应识别其团队任务中的高代价错误、参考依据和升级边界；初级员工则应学会哪些工作可以在模型辅助下完成，哪些必须请求确认。若组织把所有复核责任放到一名主管身上，能力边界并未消失，只是形成新的排队和单点失效。岗位设计需明示核验责任、批准权限、可用时间和替代人员，而不能只在制度文本写“最终由人负责”。

此外，能力基线不应把年龄、学历、语言口音或某个历史绩效分数当作固定能力。若模型帮助某些新手跨过文书表达或工具操作门槛，按旧绩效筛选培训对象可能重复排除原本可受益的人；若测试高度依赖已有模型使用经验，又会让常用者占优。较公平的方案是在岗位相关任务中做基线测量，记录熟悉度和语言支持需求，按这些条件分层随机或比较，并报告各组进入培训、完成课程、参加迁移测验和得到晋升机会的比例。

同一员工在不同任务上的能力也会变化。某客服人员可能能处理常见退款请求，却无法判断涉及欺诈或跨国税务的边界；某工程师可能擅长代码审查，但不具备评估生成式设计安全性的资格。把“AI 素养”压缩成单分数，会掩盖业务域与责任层级的差异。组织可维护与具体任务版本关联的能力矩阵，随工具、规则和产品变化更新；这应是动态的岗位档案，而非对员工终身贴标签的评分。

15. 能力建设的工作流成本与专业复核负荷

AI 建议把时间从初稿转向审核，并不自动意味着审核更轻松。阅读一份简洁的错误答案可能比重新从源数据计算更费注意力，因为审查者必须先理解其推

SVSAC | sv-sac.org

理，再拆解无根据的结论。若模型输出数量增加，资深人员可能从亲自完成转为连续判断他人输出；总工时变化很小，但认知切换、责任压力和错误暴露上升。只测模型生成时间，无法估计岗位的端到端劳动。

组织应以“完整任务链”而非生成步骤估算负荷：输入整理、模型等待、员工阅读、独立复核、返工沟通、升级、最终批准和后续错误处置。对于一项任务，员工时间可按工序日志累计；模型等待时间属于日历时延，除非同时占用人工，不应直接算成劳动小时。复核者的时间必须归入成本，不得把操作员省时而审查者加班的安排记作全员提效。若多人共用一个高级审批人，需额外观察积压、超时和因等待造成的下游延误。

构造一个仅用于说明的容量账本：某团队每周有 100 件标准业务，现行流程每件由一线员工处理 24 分钟、主管抽检 6 分钟；新流程假设一线处理降至 15 分钟，但每件 AI 输出都要由专业人员复核 9 分钟，并有 10%的项目因事实错误多花 20 分钟返工。现流程直接工时为 100×(24+6)=3,000 分钟；新流程为 100×(15+9)+10×20=2,600 分钟，账面节约 400 分钟。若新增系统监测和团队校准每周需 500 分钟，则总劳动反而增加 100 分钟；如果只有一半输出需专业复核，公式和结论又会变化。所有数字均为作者构造，不是企业观测值；用途是迫使组织把审核比例、返工率和维护时间纳入测量。

负荷分布比平均数更关键。计入每周 500 分钟系统监测后，新流程的团队总工时反而净增加 100 分钟；若另有每周 500 分钟异常调查，这项额外负荷会把净增推至 600 分钟。若关键复核岗位已满负荷，系统上线可能拖延高风险任务。容量测试应按员工角色记录任务进入和完成时点，测工作等待与处理时间，区分正常流量和峰值，并设置停止条件，例如未复核积压跨过业务承诺时暂停自动生成。停止线需由组织基线和风险要求决定，不能从上述假设账本照搬一个通用阈值。

人力资源与业务运营需要一起判断释放工时的用途。若员工时间被立即填满，能力训练需要被明确排班并受到保护；若公司计划削减岗位，培训承诺就不能作为已兑现的收益。员工可参与复盘那些被 AI 接管的任务，记录哪些步骤必须保留以维持专业能力。管理层则应公开训练对象的选择规则、是否计入工作时间、练习后由谁评估，以及失败者是否会失去晋升机会。否则，员工可能把能力测验视为惩罚性筛选，从而隐藏错误、回避复杂任务或只记录有利结果。

SVSAC | sv-sac.org

16. 人才机会、晋升与岗位价值的分配问题

当 AI 让初级员工达到过去需要数月经验才有的处理速度，企业面临的不只是“少培训一点”或“招更多新人”。如果熟练度差距缩小，原有的学徒制可能失去任务练习环节；若 AI 建议允许新人处理更多常见问题，也可能增加他们接触更复杂工作的机会。两种方向取决于岗位晋升规则、难题分配和专家时间如何使用，而不是只取决于模型能力。

可用一个明确的机会分配指标来检验：在每个季度，按员工入职时间、基线技能和部门，计算获得高价值复杂任务、专家反馈、无工具练习和晋升评审的比例。再观察 AI 访问对这些机会的影响，而不仅是对完成量的影响。如果使用 AI 的新手处理量上升，但复杂任务仍由同一小组专家掌握，技能机会并未随产出扩散；如果他们承担复杂工作却没有权力和薪酬相应增加，则能力要求上升而岗位回报滞后。

招聘与晋升标准也可能被 AI 改变。模型能写得更像资深员工，不代表候选人具备解决问题的能力；但如果组织只在无工具笔试中挑人，也可能漏掉善于监督模型、整合跨职能信息的人。评估可同时包含“工具允许的实际任务”和“工具撤去后的诊断任务”，并单独分析这些测验预测真实工作质量、错误控制和学习速度的能力。招聘决策应避免把一种场景分数当成通用潜力，且必须让候选人知道评估中是否可用 AI。

岗位变化还涉及谈判权。组织若以模型为依据提高产量目标，员工可能承担更高强度工作，却未分享节省成本；如果管理层把生成结果视为自动完成，专业人员可能在事故后承担责任但缺乏修改系统的权限。更可检验的工作设计是记录目标变化前后的任务量、加班、休息、工作自主性、申诉渠道和薪酬结构，并在部门或团队层面记录时间序列。若这些变化与 AI 部署同步，也仍需区分同期业务增长或管理制度重构的作用。

不应把员工“适应能力”当作企业单方面转型成本的免费来源。正式技能培训、带薪练习、任务轮换、职业认证和迁移支持都需要投入；在评估收益时，应将员工自行下班学习或无偿积累提示词的时间计入社会成本，即使它不出现在公司的工资账簿。对组织而言，培训公平不仅是课程开放，还包括是否给员工实际使用时间、导师反馈、失败保护和合理的岗位晋升路径。

SVSAC | sv-sac.org

17. 从模型试点到岗位能力评估的分阶段方案

一项可审计的试点应先冻结任务边界和人员角色。把同一流程中的每一任务记为版本化样本，包括输入来源、风险等级、是否有独立正确答案、模型是否可调用、谁能批准和发生差错时的修复成本。员工基线能力在模型开放前以相同任务的盲评获取；如果任务对不同经验组不等难，需要用预设题本或真实历史案例配对，而非事后控制模型使用次数。

第 12 节已定义带工具表现、即时与延迟独立测验及迁移终点。本节把它们放到同一时间线：先记录员工接受、修改或拒绝建议后的质量、耗时、返工与升级，再在随机时点撤去工具或使用未练习案例，最后追踪职责、工作量、主管复核、晋升和流动。阶段可重叠，但工资与利润未成熟时不能由主观效率问卷补填。

分派与干扰沿用第 12 节的团队簇规则，簇数而非员工数决定独立分配单位的数量。若所有人必须立即获得 AI，可比较训练流程或受保护的无工具练习时段；此时估计的是这些流程差别，不是“有 AI 相对无 AI”。共享导师、排班或知识库仍须记录，不能因采用簇分派便默认无干扰。

现场评分需要独立于系统生成过程。将员工输出去掉 AI 提示、姓名和处理组标记，再由两名熟悉任务的评审依据事前量规打分：事实正确、关键步骤完整、异常升级、表达可用性和错误后果。若两位评分员分歧，预先约定复核或第三人裁决；不得让最初使用模型的主管单独给自己的部署方案打分。对不能构造客观答案的工作，应保存依据、异议和决策过程供盲态同行复核，而不是将主观评价伪装成无争议真值。

风险监测还需要把学生式测验迁移到真实业务时的差异写清。无工具考试可检验某些独立能力，却不能代表员工是否能在系统、团队和时限约束下接管工作；反过来，AI 在岗时的高产出也不证明系统故障时组织仍能履约。试点应预先定义多长的无工具窗口、模拟还是实际故障、哪些岗位可以安全演练，以及考核是否会影响个人绩效。没有保护的故障演练可能造成客户风险，因此高后果流程应先用历史回放或仿真，再逐步过渡到低风险现场任务。

本节的阶段时间线须交由第 36 节的结果分层与第 46 节的反证判断：完成量上升而接管下降，要检查练习和其他风险控制；学习改善而审核超载，要先核复核产能。尚未成熟的薪酬、晋升、离职及收益只支持条件性扩展，不能改称长期人才升级已实现。

SVSAC | sv-sac.org

18. 劳动力需求取决于任务替代与新任务生成的相对速度

组织讨论岗位影响时，常把“AI 能完成哪些任务”直接转成“岗位会减少多少”。这中间至少还隔着产量反应、任务重组、新任务形成和劳动份额变化。Acemoglu 与 Restrepo 的任务框架把生产拆成由劳动或资本承担的任务，指出自动化带来替代效应，但效率提高也可能增加未自动化任务的劳动需求；新任务若形成劳动比较优势，则可能重新扩大劳动参与。框架本身并不预言每种技术都会增加或减少就业，净结果取决于替代、生产率和新任务的相对强弱[6]。

对生成式 AI 而言，“新任务”不能只是把原有工作拆成更多审批步骤，或把员工时间改记为提示词管理。较强的新任务证据应显示：组织出现持续的新责任或服务，确实需要劳动判断，并且该责任在流程、岗位描述或人员配置中得到承认。比如员工因 AI 而能处理更广客户群，可能产生更多需求和服务交付；也可能由系统自动回应，劳动任务反而缩小。另一个可能是复核岗位、知识维护和模型失效响应增多，但这些岗位只用于管理技术风险，未必足以抵消被替代的工作。

生产率改善也不能自动推断劳动需求增加。若需求对成本下降反应弱，单位任务成本降低可能使公司用更少员工完成相同业务；若用户等待长、需求被压抑，成本降低可能释放新需求并扩大服务。员工工资高、招聘困难时自动化的可避免成本较大；劳动力供应充足、工资低时即使模型能做，投资回报也可能有限。企业的技术选择因此由工资、市场需求、质量、风险和资本成本共同决定，而不是由模型能力单独决定。

可检验的任务记录应把每一项工作在部署前后标记为：被完全替代、由人机共同完成、因需求扩张新增、因质量控制新增、因技术治理新增，或基本未变。记录应含工作量、劳动分钟、角色、岗位等级和客户/员工结果。若同一项任务被转移给供应商、外包员工或其他部门，企业内部工资册可能显示人员下降而供应链劳动未减少。若只看企业名义岗位数，便无法区分真正的劳动节约、劳动地点变化与新任务扩张。

企业也可能在总员工数不变时发生深刻重组。岗位空缺不再补招、人员调往新职能、初级员工晋升路径消失、管理跨度扩大，都会改变员工的发展机会。就业数量和能力质量不应互相替代；把“没有裁员”当作人才升级成功，忽略了培训、任务自主性与晋升变化。另一方面，任务自动化也可能减少繁琐劳动，让员工投入更有价值的判断。需要具体观测这些资源转移，而非用“增强人类能力”笼统命名。

SVSAC | sv-sac.org

19. 平均工资、个人工资与岗位结构是不同估计对象

工资变化至少涉及三个层次：留在职员工的个人工资轨迹、不同员工进入与离开造成的岗位构成，以及企业对工资、晋升和招聘政策的整体调整。平均工资上升可能来自低薪员工离职而非同一员工加薪；平均工资下降可能来自企业招聘更多初级员工而非原员工减薪。就业人数稳定也不代表岗位职责、工时、排班或工作强度没有变化。

一个简化构造能说明构成效应。假设试点前有两名初级员工每小时 20 单位薪酬、两名专家每小时 40 单位，平均工资为 30。之后一名专家因退休离开，另外三人的工资都不变，企业平均工资变成(20+20+40)/3=26.67。若只观察企业平均值，会误读为降薪；但个体工资变化为零，差异来自人员组成。反过来，若新手以 22 单位薪酬进入、产出被模型拉近专家水平，平均工资也可能下降，同时单位成本下降、企业扩招。没有员工级入职、离职和工资时间线，就无法区分这些解释。

短期绩效和长周期薪酬评估应保留固定原始员工队列与各期全体劳动力两种口径。固定队列应继续追踪离职和转岗者，只不纳入后来新入职者；若实际只能测留任者，则须报告缺失，不能把该子样本称为完整固定队列。各期全体口径可显示劳动力组成变化，但不能单独回答同一员工是否改善。结合岗位、入职批次、工时、奖金、晋升和离职去向，才能区分人员变化与个人变化；这些建议不授权无限收集人事数据。

若部署按部门分批进行，事件研究或差分比较也不能自动解决选择。较有准备、管理能力较强的部门可能先上线，也可能同时得到预算、培训和更好的主管。识别需要记录上线资格、实际开放时间、培训覆盖、模型版本、同期改组与需求变化。若政策以部门资格随机化，主要估计对象是“被提供部署”的意向处理效应；若只比较实际用工具者与不用者，员工自选、经理指派和任务难度都会进入差异。

时间窗也会改变可见结论。员工调岗和技能测验可在几个月内观察，工资、晋升、人员流动与组织需求则可能需要更长时间。把早期“尚无工资变化”写成“没有分配影响”，或把初期离职变化推断为长期就业趋势，都超出观察。相反，若只等多年才评估，又可能错过短期学习损失或不平等机会。研究应事先把主要估计对象按时间层次列出，分别报告已成熟结果和仍待随访的指标。

SVSAC | sv-sac.org

20. 培训覆盖与技能收益不能只看自愿参加者

员工愿意使用模型、主动报名培训，往往与工作自主权、经理支持、先前能力和技术兴趣相关。自愿参加者的任务表现更好，可能说明培训有效，也可能说明本来就更容易学习。若组织把培训席位给当前绩效最高的人，短期交付可能上升，但最需要帮助的岗位未必获得机会；若只挑低绩效者，又可能把复杂难题和培训标签集中到一组员工身上。

建议把培训资格、邀请和实际使用分开记录。若有有限席位，可在同等岗位和基线能力区组内随机邀请，估计被邀请的意向处理效应；再单独报告接受率、完成率和实际使用。若邀请本身带来主管关注、学习资料或同事支持，那么它识别的是一组培训政策，而非某个模型功能的纯效应。若拒绝或退出与压力、排班冲突或顾虑有关，不能只分析完成者并称其为全部员工的培训效果。

一个用于设计讨论的例子：某组织有 120 名员工、30 个不共享排班的四人团队，但仅能安排 12 个团队参加首轮训练。若简单选出高绩效员工，首轮结果主要代表已有优势者；若按岗位和基线技能分层后随机抽取 12 队，首轮能估计“随机获得培训机会”的结果，同时保存未入选团队作同期比较。由于团队只有 30 个，推断的独立单位是团队，而不是 120 名员工。轮到第二批时，首批可能已经向同事传播经验，比较需要记录这种污染；不能假设等待名单一直没有受益。

随机邀请也不是唯一合理策略。若高风险岗位必须尽快训练，应按预先规定的安全必要性先行覆盖，研究其余可自由选择的岗位；若组织规模太小无法分出足够团队，可做分阶段实施并结合基线多期趋势、明确承认不确定性。研究设计不能凌驾于员工安全、客户权益或劳资协议之上，但应说明哪些岗位不进入随机化、为什么、其结果如何另行监测。把“所有员工都接受培训”当作试验无效，也不成立：可以比较培训方式、练习顺序、导师配置或保护的无工具练习。

培训效果还要把员工时间投入和组织安排写入处理定义。每周一小时带薪练习、无薪自学、主管一对一复盘和线上课程不是同一培训方案。若处理组得到额外资深导师，测得的收益可能来自导师投入，而不是模型。比较组也要有清楚的处理，避免“未培训”同时意味着少任务、少反馈或更高负荷。员工可在模型开放前完成相同基线任务，并在之后不同时间完成等难度的即时和延迟迁移测验；评分者不知道分组，题目不直接复用培训样例。

SVSAC | sv-sac.org

21. 技能测量应能发现错误，也应容纳不确定判断

单一正确率不适合所有岗位。核对发票可能有明确金额与日期；战略规划可能有多种可接受方案；安全升级往往更重要的是没有漏掉不可逆风险。测试必须定义结果单元和错判代价：一个错误字段、一个完整客户案例、一次班次处理或一个月的岗位质量，互相不可替换。若模型生成大量建议，员工仅抽查其中少数，按已审样本计算的准确率也不代表全部输出质量。

能力评价可用盲态多评分员和事前量规，但评分一致并不等于判断有效。两位评审可能共享同样偏见，也可能只奖励文风而忽略事实。可以在正式评分前用独立金标准案例校准，报告各维度的分歧；无法裁定的案例保留为未决，不应强行编码成“对”或“错”。重大风险错误需单独列示，即使整体正确率很高，也不能被平均分掩盖。

延迟测验要区分遗忘、迁移和工作流程变化。数周后同题分数下降可能是记忆消退；新题表现下降可能是迁移失败；真实业务质量下降则还可能来自产品、客户或规则改变。可同时安排原知识再测、结构相似的新情境和真实错误审计，使用不同端点描述。团队之间分享答案、员工轮岗或模型升级，会改变暴露经历；分析需按原分派保留人员，并记录是否出现新的训练、换岗或新工具接触。

还应测“会不会拒答或升级”，而不是只奖励回答速度。若模型在不完整信息下生成流畅答案，员工需要识别何时证据不足；评估量规可以把安全拒绝、求助和记录依据作为合格行为。拒答过多也可能造成服务延迟或把工作推给专家，因此应看拒答是否正确、升级是否及时及其后续处理时间。员工的能力既包括执行，也包括控制任务边界。

最后，技能分数不应直接绑定短期薪酬或裁员。若员工知道测试结果会影响晋升，可能练习特定题型或不报告模型故障；若评估只用于改进培训，员工更可能暴露不理解之处。组织应提前声明用途、访问范围和保留期限，允许员工纠正记录中的事实错误，并将学习评估与纪律调查、绩效管理的边界写清。安全事故仍需按适用制度处理，但不能把普通学习失败自动视为个人不胜任。

22. 员工数据治理会改变学习证据本身

能力试点常收集提示词、输出接受率、修改记录、处理时间、升级和评分结果。它们可以帮助发现培训缺口，但也能构成高密度的个人行为监测。若员工预期日志会被用于惩处，使用行为会改变：有人减少提问、转到未记录渠道，或把

SVSAC | sv-sac.org

风险更高的工作留给同事。于是数据不再只是“观察窗口”，而成为影响学习与协作的治理条件。

更谨慎的设计是按用途分离数据。培训评估所需的任务记录可使用项目级或匿名化标识；绩效管理只接收预先约定的汇总结果；安全事件按独立流程保留必要个案。谁能看到原始提示、错误记录、员工自评和评分分歧，应逐项说明。用于模型改进的资料与用于人员考核的资料不应默认共用，否则员工是否愿意呈现不确定性会受到额外激励影响。

组织还需区分使用数据和能力数据。高使用率可能表示工具有价值，也可能反映任务量上升、工作复杂或强制性要求；低使用率可能表示系统不适配、接口权限有限，或员工担忧监控。提示词长度和采纳率并非学习质量指标。若把高采纳率设置成目标，员工可能接受不适用建议；若把人工修改率设成低目标，必要的专业判断也会被惩罚。

一种可验证的治理试点可以比较两种披露清晰的访问模式：常规记录用于流程改进、员工个人数据不进入晋升决策；与明确用于个体绩效反馈的模式。若组织不能随机分派，至少记录员工对数据用途的理解、实际日志缺失、跨渠道使用以及不同岗位的退出情况。结果端点包括错误识别、求助行为、工作质量、员工信任和数据完整性。即使记录完整度增加，也不能单独断言能力提高；它可能只是员工更愿意在系统里操作。

治理边界也涉及员工对模型建议的责任。员工应知道哪些输出必须验证、可以拒绝或升级、发生失误时如何暂停流程，以及错误申诉交给谁。若公司要求员工对模型错误负责，却不给源数据访问权、复核时间或系统变更渠道，这不是人机协作的有效控制。责任与权限必须匹配；当任务超过培训范围，组织要提供有效升级路径而非默认让员工承担。

23. 能力升级的关键不是“会用工具”，而是能否跨任务迁移

在工作现场，一名员工借助模型更快完成某种文本或分析任务，说明该人机组合在当前任务上表现改善；它并不自动说明员工已经掌握了可迁移技能。可以把结果拆成三层：工具开放且人机协作时的联合绩效；撤去工具后员工对相近任务的独立表现；面对结构变化的新任务时能否迁移。三层之间可能相关，也可能方向相反。模型提供可靠草稿可提高联合质量，却使初学者少做检索、比较和纠错；反过来，模型若把解释和反馈嵌入练习，也可能帮助员工形成可独立调用的

SVSAC | sv-sac.org

知识。对 52 名有经验程序员学习新异步库的随机研究，提供了短期技能形成的一个特定情境证据，但样本、课程和任务都不能代表企业培训或长期职业结果[2]。

能力迁移还受任务表征影响。若员工只学会在熟悉模板中补齐字段，遇到同一业务逻辑但换了客户、数据结构或异常条件，可能无法判断模型建议是否适用。所谓“熟练使用提示词”也可能只在固定模型版本下奏效。更有信息量的测量不是再考一次原练习，而是构造保留底层判断、改变表面形式的迁移任务，并在合理时间间隔后重复。不同模型、权限、可用资料和时间限制应事先冻结；若模型更新，则另列版本暴露，避免把练习进步与工具能力变化混为一谈。

可以用一个构造性例子说明联合绩效与独立能力的分离。假定基线时员工完成熟悉任务平均得分 70/100，使用模型后得分 84；六周后在无工具条件下，用相似但新案例测试，得分为 72。若工具组仍有高分，但另一组同期练习且达到 76，那么当前联合绩效优势仍可能存在，独立技能差异却不支持“培训已提升可迁移能力”。这些数字只是示范估计对象，不是实测结果；更重要的是测试是否等难、评分是否盲态，以及六周内是否有其他训练或换岗。

职业阶梯也改变学习的价值。初级岗位常承载大量练习、查错与边做边学，专家则承担高影响判断和异常处理。自动化这些初级任务可能降低训练成本，也可能减少新员工累积判断经验的机会；工具若让初级员工接触更复杂案例并获得结构化反馈，也可能拓展其学习范围。短期任务速度无法区分这两条路径。需要观察入职者是否仍能逐步承担更难任务、导师是否有时间解释错误、岗位晋升是否改变，以及专家的判断是否被转化成可教的案例。

岗位能力因此不宜被压缩成一份静态技能清单。更可操作的表示方式是“任务族×风险等级×独立程度×工具条件”：例如员工能否在低风险常规请求中独立处理，能否在资料不全时选择拒答，能否在新型复杂案件中识别何时请求专家。每个格子都应有代表性任务、可观察行为和分歧裁定程序。评分者可采用盲评，但评分标准需要区分事实准确、推理适切、升级恰当和完成时间；同一分数不能把这些维度合并后掩盖关键短板。

24. 组织要检验的是能力组合的形成，而非某个员工的孤立分数

越来越多工作由团队共同完成，单个员工能力上升不等于组织能力同步增强。若一个人学会识别模型的典型错误，却没有权限修改工作流，改进可能无法传播；若少数专家把模型输出全部接管，团队表面上交付稳定，实际形成了新的

SVSAC | sv-sac.org

瓶颈。组织能力至少包含成员技能分布、接口与升级路径、知识保存、替补能力和调整工作流程的权力。它是集体层面的结果，不能把多名员工测试分数简单平均就视为已测量。

学习传播会造成两面性的干扰。培训组员工可能向同事分享范例，使未培训比较组也受益，从而缩小组间差异；也可能占用专家复核时间，延长比较组处理时间，使差异扩大。只把个人当随机单位、忽略共享主管、同一知识库和交叉排班，会低估不确定性。若问题是培训政策是否改善整个部门，可按互不共享关键资源的团队或排班簇分配，并记录跨簇协作；若企业只有少量团队，则不应把大量个人任务误称为大量独立随机单位，而应报告簇数和较宽的不确定范围。

组织学习还需要把“反馈可转化为改进”与“反馈被收集”区分。员工提交错误例子后，若没有责任人判断问题来自模型、业务资料还是规则变化，记录不会自动形成知识。一个可观察闭环至少包括：问题提出、来源核验、责任分派、决定是否更新指引或工具、验证更新是否减少同类问题、以及向提交者反馈。每个阶段都可能被延误或放弃；数量较多的反馈可能反映系统更差，也可能反映员工更愿意报告，不能单靠工单数解释组织学习。

NIST 的生成式 AI 风险管理画像提供自愿技术指导，建议在风险相称的范围内区分人员职责、设置独立评估、定义可接受用途并提供反馈或申诉机制[7]。这类框架能帮助明确组织控制点，但它不是培训效果研究，也不是法定认证。将其用于人才建设时，应把“角色已命名”与“责任人实际具备能力”分开：审查人员是否能发现错误、知道何时暂停、能否调用必要数据和及时升级，才是能力检验。正式岗位授权也需要资源匹配；如果审核工作压在原有产量指标之上，组织只是在纸面上增加责任。

再看一个假设性的团队结构。团队每周要审核 200 份模型辅助结果。四名专家各有 10 小时用于复核，总容量为 40 小时；若每份平均用 10 分钟，需求为 33.3 小时，尚有 6.7 小时缓冲。若新任务复杂度使均值升至 13 分钟，需求增至 43.3 小时，缓冲变为缺口 3.3 小时。即使员工的平均能力分数上升，只要等待队列持续增长，实际能力就不能转化为更快的交付。若把每份复核强行压回 10 分钟，可能增加漏错。此例只演示队列约束，均值不代表真实企业数据，也未计入案件到达波动与不同严重度。

SVSAC | sv-sac.org

25. 培训投资的回报取决于可避免资源，而非节省分钟的名义价值

技能项目常用“每人节省多少时间”乘以工资来估值。但工资等价不是必然现金节约。若员工薪酬不变、岗位数不变，释放时间可能用来处理积压、改善质量或支援新任务；它的价值可能真实存在，却不能直接写成现金回报。若节省的时间不能在排班中合并、也没有替代任务，名义节省甚至未形成可用产能。反之，减少加班、外包或可避免新增岗位可以产生现金变化，但必须通过工资册、供应商支出或预算决策确认。

用一组透明的构造数说明。假定组织培训 100 人，每人接受 8 小时带薪训练，全负担工资率为每小时 40 货币单位，则培训时间的工资等价成本为 100×8×40=32,000；这既不自动是新增现金支出，也不自动等于被放弃产出的机会成本。假定每人每周节省 20 分钟，持续 50 周，共释放 100×(20/60)×50=1,666.7 小时，按同一工资率折算约 66,667。这只是工资等价值，不是实际再部署收益或其上限；实际机会价值须由可执行替代任务的净贡献另行估计。若释放时间的四分之一确实减少了按每小时 40 支付且可避免的加班现金，条件性现金节省约为 16,667；若工资、外包和其他支出均未改变，则没有这项现金节省。若培训工资原本就会支付，不能直接用 16,667 减去折算的 32,000 来称为净现金回报，而应减实际新增现金成本，并另列被挤出工作的价值；只有另加“全部 32,000 培训人工均为新增付薪现金”的假设，忽略其他项目时净现金才约为−15,333。已计为避免加班的同一小时不能再计再部署收益。工具费、主管时间、学习曲线损失和返工仍需另列；这些是假设账本，不是企业观测或普遍收益率。

更完整的回报账本应按增量列项：培训与停工成本；模型、数据和治理的新增现金支出；质量或客户结果变化；可避免的招聘、加班、外包和返工；以及释放的容量被重新使用的价值。共同固定成本不能任意分摊后再当作新政策边际成本；新增政策真正引起的额外费用则不能因“已有人事部门”而忽略。若产出增加来自额外专家复核、主管辅导或团队加班，这些投入必须计入。项目净值要与可行替代方案比较，而不是只对照“什么也不做”的虚构零成本。

对培训项目的因果评估，随机获得培训机会的意向处理效应与完成者效果不同。若只有一半受邀员工完成，受邀组与未受邀组的比较保留随机化，但估计的是提供政策的效果；完成者与未完成者的差异则受动机、排班、主管支持等影响。用邀请作工具变量推算完成者局部效应，还需要排除邀请通过注意力、主管支持等其他路径影响结果，并需要单调性等额外假设。若这些条件不可信，不应把局部估计叫作每位受训者的普遍效果。

SVSAC | sv-sac.org

管理层还要决定收益由谁承担、谁取得。公司降低成本而员工工作强度升高，不能用公司账面净收益替代员工结果；反之，员工通过 AI 减少无意义加班，企业现金支出未变，仍可能产生福利收益。至少应把员工工时、工时可预测性、工资、晋升和主观可控性与企业净成本并列，报告方向不同的结果而非合并成一个指数。任何货币化都需要明确主体、期间、反事实和估值依据。

26. 分阶段能力建设需要在学习收益、风险和产能之间做选择

沿用前述任务与能力区分，第 26 节只检验学习设计的增量，不重定义整套能力建设。练习、复核、文档、集成与岗位调整仍相互制约：缺练习可能没有独立判断，缺复核产能也可能使新增产出形成积压。下面保持工具权限与任务相同，以解释反馈包为具体比较，而不是给所有岗位规定一套课程。

一个可检验的分阶段设计可在安全边界内区分两类政策：A 组获得任务练习与即时解释反馈；B 组获得相同工具权限、练习时间和任务，只减少解释反馈。团队层面随机分配，评分者不知道组别，首要结果设为两周联合任务质量与时间，延迟结果设为六周无工具迁移任务；同时记录错误升级、求助、导师分钟、模型版本和同事分享。A 与 B 差异能检验反馈包的增量，但不能进一步归因于“解释”还是“导师注意”，除非拆分随机化或额外测量。两组共同提高也可能来自熟悉任务、参与效应或测试练习。

比较期内应固定任务难度分布和完成窗口，保留所有原始分配员工在意向处理分母中。转岗、离职、休假、拒绝使用、技术故障和未完成测验分别登记；离开组织者的长期结果不可当成普通随机缺失。可报告观测到的队列结果，并用合理上下界分析未观测结果；若缺失严重且方向不明，应承认该结果无法识别，而不是只比较留任员工。任务安全要求、客户权益和集体协议优先于随机设计，必要岗位可全部先行培训，再单独报告。

主要反证包括：工具提高联合绩效，却降低无工具迁移；能力测验改善，但错误率、队列或客户结果恶化；短期学习提高，数月后因版本更新而失效；或者小组受益而初级岗位失去练习任务。正向结果也需检查解释：如果 A 组比 B 组多了导师时间，收益可能来自人力投入；若评分员知道分配，可能存在期待偏差；若练习任务与后测高度相似，分数可能反映记忆而非迁移。未来复现应改变任务类型、岗位和部门，并如实报告失败或边界条件。

SVSAC | sv-sac.org

组织人才升级并非必须表现为岗位数量下降或新增一批“AI 专家”。较强的能力证据是员工能在工具可用和不可用时恰当执行，能识别超出能力边界的情况，团队能把错误转成流程改进，而且岗位路径与治理责任相匹配。若这些变化没有出现，任务速度提升仍可能是有价值的生产率改善，但不应被叙述成能力建设、工资提升或长期职业机会已经实现。

27. 员工感到更高效与可测量的任务生产率不是同一结果

员工对 AI 的评价有时比计时和质量结果更积极。主观体验仍是重要结果：更容易开始任务、减少挫败感、增加探索意愿，都可能影响长期采用与工作满意度。但这些结果不应替代可比任务的完成时间、质量或后续返工。尤其当一项技术让工作感觉流畅，却增加审查、修复和整合时间时，员工对“节省时间”的回忆不一定准确。

一项针对 16 名资深开源开发者、246 项真实仓库任务的随机研究展示了这种差异。任务在允许或不允许使用当时 AI 工具之间随机分派；研究报告允许 AI 时任务完成时间的回归调整估计增加 19%，而参与者事后仍估计 AI 让自己快了 20%，事前预测则是快 24%[8]。该研究的任务约两小时，开发者熟悉相关代码库，使用 2025 年初的工具；它不证明 AI 普遍降低生产率，也没有测试员工离开工具后的技能、工资或晋升。其价值在于反证“用户觉得有帮助即可推断单位工时产出提高”。

组织可以把体验和绩效并列，而非挑选符合预期的一项。任务完成时间需要固定任务总体与截止规则；质量应以独立验收、后续返工和代码或业务约束衡量；体验可用任务后短问卷、持续使用意愿和任务压力测量。若仅对完成者询问体验，未完成、放弃或转交专家的任务就从分母消失。若计时仅覆盖“模型开始生成”到“提交”，却不包含读结果、查资料和修复，效率指标的方向可能错。

主观与客观结果不一致时，不能马上把员工判断称为偏差，也不能把计时结果视作完整价值。更愉快的工作可能有福利价值，即使任务变慢；更快的短期提交也可能通过返工或风险增加而损害后续价值。两者冲突应触发机制调查：体验是否来自降低开始成本、更多尝试，还是来自注意力转移而未觉察的整合工作？客观变慢是否由更严格的质量规范、任务复杂度差异或新工具学习造成？将这些机制测出来，比把“感知差距”作为通用宣传语更有解释力。

SVSAC | sv-sac.org

28. 人员差异要按任务和条件解释，不能固化成“高手/新手”标签

AI 对不同经验员工的影响常被描述成一种固定排序：初级员工收益最大、专家收益较小。这个概括可能来自特定任务、工具和样本，不应直接成为岗位分配规则。低经验者或许更需要示例和反馈，也可能更难识别错误；专家或许更会判断输出，却可能承担更高质量要求和更多整合成本。经验年限、领域熟悉度、工具熟练度、任务难度和可用支持并非同一个维度。

更有识别价值的问题是“在哪一种任务条件下，哪些基线能力与何种处理相互作用”。可在试点前记录对当前系统的熟练度、相关领域经验、任务熟悉度和独立诊断能力，并用连续量表或明确区间表示，不事后按结果挑分组。主要分析可预先估计处理与基线能力的交互，但小样本下区间可能很宽；未发现交互不等于证明各类员工效果相同。若工作簇很少，个人任务再多也不能制造足够的独立信息。

观察到的异质性还可能由任务分配造成。专家常被分派最难案例，初级员工处理更多模板化工作；即便两组能力一样，未经调整的速度差也不能说明 AI 对谁更有用。相反，若 AI 把容易任务从初级员工手中拿走，剩下的任务组合会变难，初级组平均时间可能上升。需要保留任务类型、难度、客户条件和原分派信息，报告按基线任务组合标准化后的比较，以及真实运营中任务组合变化造成的实际结果。两种答案不同，恰恰说明要分别回答工具效果与岗位重组效果。

可构造一个 2×2 试点：按预先定义的高/低任务熟悉度分层，在每层随机提供带解释的 AI 辅助或常规工作流程。处理结果包括固定难度任务的联合质量与耗时、离开 AI 后的相似任务诊断，以及复杂例外的升级判断。交互只检验“熟悉度是否改变辅助方案的相对效果”，并不证明熟悉度本身是原因；该分层可能代理岗位、经理支持或项目质量。若分层与工具使用条件由同一管理者选择，仍有混杂。分层设计需记载随机单位、簇、分派、实际使用和任务复用，并事先限定主要异质性检验数量。

29. 入门任务消失后，组织是否建立了替代的学习路径

初级岗位通常并不只产出当前任务结果，也提供看见专家如何判断、经历可控错误并逐步承担责任的机会。如果 AI 接管了大量入门任务，组织可以获得短期节省，却可能几年后发现资深人才供给不足。这是一个长时滞的机制命题，不

SVSAC | sv-sac.org

应从短期自动化比例直接推成“人才断层”，也不应假定数字训练就能完全替代真实任务经验。

替代学习路径可以是受控轮岗、模拟练习、案例复盘、双人处理、低风险任务的分级授权，或让员工在工具提供建议后先独立写出判断依据。它们的学习目标不同：模拟练习便于安全重复，却可能缺少真实客户信息和责任压力；跟随专家可示范边界判断，但导师时间有限且不同导师标准不一；模型解释能快速反馈，却可能把错误理由讲得很流畅。组织应比较具体路径，而不是统称“再培训”。

实用的队列指标是新员工从入职到独立承担不同风险等级任务的时间、首年任务组合变化、需要专家返工的比例、晋升与离职、导师投入小时，以及员工独立完成的新情境表现。必须把风险等级和案件质量同时纳入，否则组织可能仅通过把高难度任务转给少数专家来缩短初级员工的“上手时间”。若新员工更快完成低风险任务，但从未接触例外情形，独立胜任范围未必扩大。

一个可反驳的预测是：若 AI 辅助确实在替代学习机会而组织没有补充路径，受影响队列会在后续周期表现为练习案例减少、专家接手比例上升、初级员工对新情境的独立判断没有改善，或岗位晋升延迟。若这些结果没有出现，且新员工在盲态复杂任务测试与工作质量都提高，则“入门任务减少必然削弱人才供给”的解释受到反证。需要长期跟踪同一入职队列，并对不同产品需求、招聘冻结和岗位结构变化作出记录；单个季度的招聘数不足以区分这些机制。

比较企业或部门时，不应把跨期能力供给与当前 AI 用户生产率混为一谈。企业可能通过外部招聘补足高级人才而维持业务，但更高招聘成本或团队脆弱性尚未入账；也可能通过培训、轮岗和知识库降低师徒依赖。分别记录内部晋升路径和外部填补、关键岗位空缺时间、专家离职后恢复速度，才能判断内部人力资本是否被替代或重新组织。相关指标的变化仍不能单独归因于 AI，需与需求变化、薪酬政策和招聘周期共同分析。

30. 岗位重设计要匹配任务互补性与责任边界

AI 部署后的岗位不一定只是“人做更少的同类工作”。任务组合可能转向异常处理、客户解释、质量审查、协调和申诉。重设计若只把低风险例行步骤自动化，却保留了清晰的升级路径，可能提高员工自主判断的价值；若机器完成易任务、人只接收难例和失败例，员工的平均任务难度及心理负担可能增加。该变化并不意味着员工技能下降，但会改变培训、配员和绩效评价所需的能力。

SVSAC | sv-sac.org

可把岗位设计拆成三类决策：哪些任务由系统建议或执行；员工何时必须检查或可以接受；员工是否拥有拒绝、暂停、修改和升级权限。相同的模型可嵌入完全不同的岗位制度。若计件绩效仍按原来的总体数量设定，员工可能跳过复核；若要求每件都人工验证，却不给相应时间，自动化可能只增加一层工作。组织应把复核覆盖率、例外严重度、等待队列和员工自主权与产出共同记录。

一项新的岗位权责表应同时写明“执行者、责任承担者、证据来源、可暂停者、申诉接收者”。例如模型给出初步分类，员工核对关键字段；异常结果由资深人员复核；若源资料缺失，任何人都可暂停而不受数量惩罚；事件责任由流程负责人复盘，而不是自动归责于最后点击确认的人。这只是治理设计样例，具体责任仍受岗位、客户和行业规则约束。纸面表格是否有效，可通过情境演练检验员工是否知道如何行动，以及实际系统是否授予相应权限。

工作质量也会影响责任边界。对可逆的草稿生成，员工能在发送前复核，授权可较宽；对会改变财务、健康、安全或员工权益的决定，错误后果更难逆转，独立复核和暂停能力需更强。这里的“风险更高”不能由模型置信分数单独定义，而要根据潜在损害、纠正成本、影响对象和发现时滞来评估。高风险控制有资源成本，应明确承担主体并报告复核时间，不应把它描述为无成本的人类监督。

31. 部门试点应把任务、人员发展与业务量变化放在同一时间线上

培训后部门产出增加可能来自单位任务速度提高、需求增长、任务难度下降、员工加班或临时增员。同一工作量下的技能提升只是可能路径之一，工具直接代做或流程简化也可缩短耗时而不增加独立能力。将工作量、任务质量、员工工时和人员构成按周或月联结，可以识别这些不同路径；但并行趋势、选择上线和同时进行的管理调整仍需检查。若只能观察部门总量，不能把总量产出的变化解释为个人能力提升。

提出一个用于方法讨论的月度记录表：每个团队保存原始任务入队数、难度分层、AI 可用状态、完成与取消数、返工/升级数、员工有薪工时、导师复核工时、空缺与临时借调。另以固定任务样本测独立迁移和工具协作质量。若只记录“已完成”，取消和转交会使分母变化；若只记录登录账户，授权不等于使用；若只记录使用者，非使用可能是排班、无权限或任务不适配。表中指标先以工作流程定义校验，再用于政策比较。

SVSAC | sv-sac.org

某组织若只有八个团队，可按业务类型和基线工作量配成四对，在每对中随机安排先后上线，并在所有团队使用 AI 时比较反馈、培训和复核流程。在尚有未上线团队的窗口，该设计可比较较早提供 AI 政策包与暂未提供的效果，但不能把伴随培训或流程改变分离成纯模型效果；全员上线后的流程比较没有同期无 AI 对照。上线团队的经验还可能传播到等待团队。若团队共享主管或知识库，成对随机也不能保证无干扰。记录知识分享、专家支援和交叉排班，必要时将整个共享单元作为分派簇，或缩窄结果解释为在试点饱和度下的相对效果。

决策门槛应明确哪些结果能促使扩大、修改或暂停。例如，若联合任务表现改善但无工具迁移不变，组织可以继续把项目称为辅助生产率政策，却不应称为员工能力提升；若单位工时产出提高但队列、升级率或专家负荷恶化，先检查新增瓶颈；若员工能力上升但工作机会集中于少数人，需检查培训访问与晋升分配。门槛不应以一个未经基线估计的通用百分比表达，而应在试点前由业务质量、劳动时间和风险承担者共同定义，并展示测量误差和未成熟结果。

32. 生产率盈余如何分配，取决于竞争、合同与谈判

同一项技能或时间改善可以被客户、雇主和员工以不同方式取得。公司减少工时成本后，可能降低价格以争取市场、保留更高毛利、增加员工薪酬，也可能投资到服务质量和新产品。谁最终受益不是 AI 功能的内生常数，而取决于替代品、价格竞争、工资制度、客户合同、劳动力供给和谈判力量。将单位任务节省直接等同于工资上涨或企业利润，跳过了分配机制。

构造一个只说明企业端现金流分配的账本：某服务每 100 个案例收入 140 单位，人工现金成本 80，其他变动成本 20，基线贡献为 40。部署后，假设人工现金成本减少 20，增加 AI 与复核成本 7，产品质量和数量不变，则价格不变时贡献为 53，可分配的企业现金增量为 13。若价格降低 8 单位，收入为 132，工资调整前贡献为 45；若再额外支付 3 单位员工奖金，贡献为 42。降价 8、奖金 3 和公司增量贡献 2 之和仍为 13，但这是该 13 单位的流向，不是全体员工净收入改善或社会净福利。若最初减少的 20 单位全部来自同一员工群体的工资或加班收入，另获奖金 3 后，该群体的总现金收入变化是−20+3=−17，而非+3；若减少的是外包支出，还须追踪外部劳动者和供应商。相应工时减少、闲暇、就业变化和新增 AI 服务的真实资源成本另行评价，不能由企业现金节约直接推出社会生产率增益。所有数值均是假设，真实分配需要合同与账簿验证。

SVSAC | sv-sac.org

这个例子也说明，员工单位小时产出提高不必导致工资上涨。若员工不能议价、劳动可替代、市场需求弱，企业可能把盈余留作利润或降价；若专业人才稀缺、制度按技能付薪，员工可能获得较高工资；若加快服务带来新需求，增量人员需求也可能抵消自动化。反向解释同样重要：工资增加可能来自紧俏劳动力、晋升结构或行业合同，未必是 AI 生产率红利。需同时观察价格、数量、利润、薪资和工时，并明确这些变量如何变化。

经验估计还要把可避免现金和容量价值分开。前例假定人工成本减少 20 单位，只有在减少排班、加班、外包或其他实际支出时才形成现金节约。如果员工人数与工资不变，20 只是可重新安排的时间价值；若释放时间被用于更多案例，单位贡献变化还受需求和额外成本影响。任何将时间价值纳入盈余分配的计算都应说明替代用途、工资是否变化以及是否发生现金支出。

33. 人数变化应分解成雇佣流量，而不是只比较两个时点

期初和期末人数之间的差异无法识别 AI 造成的岗位变化。一个部门从 100 人变成 93 人，可能经历 7 次主动离职而没有裁员，也可能招聘冻结、合同到期、业务收缩，或因模型上线而不再补充部分岗位。人数快照把不同事件混在一起，甚至把留任员工从高负荷岗位转到其他部门的变化遗漏。

一个构造的年度流量示例：期初 100 人，期间 12 人离开；若实际补招 8 人，年末为 100−12+8=96 人。另一个明确的招补情景是：离开仍为同样的 12 人，原计划补招 8 人，但其中 4 个名额未补，实际只补招 4 人，因此年末为 100−12+4=92 人。未补空缺改变的是实际补招流入，不能在已经确定的实际 8 人补招之外再机械扣 4。96 与 92 的差只比较这两个构造招补情景，不能识别模型上线的因果作用：离开可能来自退休、个人选择或项目取消，少补招也可能源于市场下行。判断 AI 是否改变招聘需求，需要记录每项岗位申请、批准或取消、实际入职、职位要求、替代安排与时间戳，并对照未部署但面临相似需求的部门。

岗位影响的动态指标还包括空缺持续时间、招聘失败、内部调动、临时工比例、承包商支出、团队工时、离职率和新员工任务分布。某岗位的编制减少可能伴随职责迁移到别的部门；公司内部职位数下降，却可能由外包商招人完成相同工作。反之，岗位数稳定也可能掩盖不再招初级员工而增招资深人员。组织层面的人员变化需按角色、合同类型、地点与任务职责拆分，并将内部与外部劳动力边界明示。

SVSAC | sv-sac.org

更严格的政策比较可按原始岗位申请或成本中心定义风险集，记录上线前后每项招补决策，并标记不可用 AI 的业务需求冲击。若把上线后的新增职位作为处理组、取消职位作为结果，选择规则本身会造成偏差。更合理的是先冻结一组相似任务/岗位单元，按部署安排或部署邀请比较其后续招聘流量，再把员工实际使用视为中间变量。若部门上线由管理者按绩效或招聘压力选择，观察性比较必须展示上线前趋势、共同变化和未测混杂的局限。

34. 训练机会的公平性要看可达性和后续流动，而不只是培训通过率

培训项目通常只展示完成者成绩，但机会分配发生在更早阶段：谁被告知、谁有带薪时间、谁获得设备权限、谁被主管提名，以及谁因照护、轮班或合同安排无法参加。若只比较完成者，选择偏差会掩盖哪些员工被排除。通过率高也可能只是把名额给了本就更熟悉工具和任务的人。

可以用一个构造分层说明访问率与结果率的区别。假设某组织有 40 名高基线技能员工和 60 名低基线技能员工，首期培训名额 20 个。若主管只挑高技能者，访问率为 20/40=50%，而低技能组为 0/60=0%；这并非证明高技能者更会学习，只说明名额按现有能力集中。若按岗位风险与基线技能分层随机邀请，在两层分别提供 10 个名额，则访问率为 25%和 16.7%。实际完成率仍可能不同，必须记录邀请、接受、完成和迁移测验的每一步，且分母始终是初始符合资格的员工，不是报名者。

公平性分析还要避免把小群体差异当确定结论。分组越细，样本越小，多重比较越多；单一部门的差值可能来自主管或排班偶然。应预先定义与工作相关的分组、隐私保护方式和最小报告单元，并结合置信区间、方向一致性与可行的定性流程核验。若不具备合法或伦理依据收集敏感属性，就不能为追求完整分析而扩张数据收集；也可检查可见的机会流程指标，例如班次、工龄、岗位等级和批准/拒绝原因，但这些代理不等于人口学差异。

机会公平的结果不仅是入训，也包括后续任务。接受培训但长期没有机会练习、不能访问模型或无权处理新任务，能力分数可能不转化为岗位发展。相反，培训、复杂任务、导师指导和晋升同时提供给一个群体，无法确定是哪一环节产生差异。公司应把资格、邀请、出席、任务分派、导师时间和晋升记录连接起来，在适当保护下分析整个转移链；再通过可行的随机邀请、分阶段开放或透明规则变化识别政策增量。

SVSAC | sv-sac.org

35. 个人胜任、团队韧性与岗位集中度是不同的组织结果

当少数员工成为 AI 流程的“超级用户”或最终审查者，短期产出可能提高，但团队对这些人的依赖也会增加。一个人能排查系统故障，不代表组织具有可替代能力；若知识没有记录、轮班没有备份、权限集中在单一账号，员工休假或离职就可能使流程停摆。团队韧性关注关键任务能否由不同成员按标准交接，并在故障或高负荷下维持可接受结果，不是个体技能分数的平均值。

可用一个简单压力测试衡量集中风险：在常态下由指定专家处理 10%的异常任务；演练中假设该专家缺席一周，观察待办积压、平均等待时间、错误和接管人数。再比较有无备份训练、可访问的处理记录及跨班次交接。此演练并非真实故障概率估计，缺席情境也不代表所有冲击；它能检验组织是否有替代人力和可执行程序。若替补员工只在模拟环境中成功，还需在安全范围内观察真实低风险任务的迁移。

组织能力可由“能交付”进一步分解为“能诊断、能接手、能修复、能改进”。员工可能能按步骤完成常规任务，却不能诊断模型更新后的失败；主管可能能临时接手，却无法修正输入数据或权限；专家可能修复个案，但同类故障反复出现。每一层需要可观察的行动、所需权限和相应的交接记录。培训计划应与故障响应和岗位备份一起设计，而不是只提高工具操作熟练度。

另一方面，把知识集中在专家手中也有收益：高风险工作由少数训练充分者审核，减少大范围错误。但集中会增加排队和单点失效，且初级员工缺少练习。是否集中应比较错误成本、需求波动、专家容量、任务可标准化程度和替补培养成本。可在多个团队比较“集中审核”“一线分级授权”和“轮值专家”三种方案，设定同一任务质量门槛，记录处理时长、错误率、员工学习和峰值排队。若只有一次低负荷试点，不能推断高峰期同样有效。

36. 扩大部署的判断要把短期帮助、长期能力和系统成本分别呈现

第 12、26 及 31 节分别界定分派、反馈包和先后上线比较。它们若同时改变工具、训练与管理支持，得到的是实际政策包效果，不能再从同一差值拆出纯模型贡献。这里把任务表现、独立能力、岗位与流动、福祉、现金及容量价值分别呈现；工资或人数暂未变化，并不否定以质量、压力或服务稳健性为目的的政策。

SVSAC | sv-sac.org

扩展决定与其声称的收益须同层：只证明工具辅助表现改善，就把范围、质量门槛和复核费用限定在该场景；声称技能培养，需要延迟迁移和新任务承担证据；声称薪酬、就业、利润或客户价格改变，则需要各自的结果与反事实。第 46 节据此规定相反观察，不能用前一层通过替后一层作结论。

每项终点的决策页保留估计区间、员工与簇数、未完成比例、版本和成熟期、分布及成本承担方。团队溢出、未随访外包劳动、同时岗位改革和模型更新等未识别项另列。第 51 节的离职与转岗处理仍适用：排除离开者或仅看活跃用户后才成立的结论，只能称选定群体描述，不能称原始分派总体效果。

组织需要对“暂停”有明确解释，而不是将其视作失败。若重大错误、异常工作量、员工监控投诉或支持队列超过能力，可以暂停扩大并在限定范围内查因；但暂停后把停止项目的未完成结果删掉，会造成有利样本选择。预案应保留停止规则触发时间、已观察和未观察结果、人员退出及其成本。只有在证据表明问题可修复且残余风险可接受时，才重新启动；无须把所有试点都转化为全面上线。

37. 能力测验也会改变学习过程，不能把反复测试当作被动观察

重复测量能力有助于识别保持和迁移，但测验本身可能构成练习。员工第二次回答同类问题更好，可能是学会了技能，也可能记住题目、猜到评分标准或熟悉测试环境。若处理组接受更多测试反馈，最后分数同时反映训练和测量差异。评估设计必须把测试暴露计入处理历史，而不是把每次分数视为独立窗口。

一个可执行办法是把测验分为共同锚点与轮换任务。共同锚点用于检测测量尺度是否随时间变化，轮换任务来自同一预先定义任务蓝图但不重复表面案例；每次测验随机抽取形式，保留项目难度和评分者信息。员工可以在一个样本上练习，但主要迁移分数来自未训练案例。若所有员工使用同一套题，题目泄漏或同事交流可能使对照组也获得练习；若每人题目不同，又可能把难度差误当能力变化。需要同时保持任务覆盖和形式等值，并报告无法等值的局限。

测验后给即时反馈也会改变下一次表现。若研究目标是估计“培训加反馈”的完整策略，反馈属于处理，不应从估计中剥除；若要识别反馈本身的作用，需要将相同练习内容随机分到反馈和无反馈条件，且避免组间交流。反馈内容也应记录，例如只显示对错、提供解释、给出提示或展示标准答案，训练含义完全不同。没有这类记录，团队在多个周期内的变化难以归因于任何单一因素。

SVSAC | sv-sac.org

测试任务还可能污染真实工作。若员工被告知将考核某类错误，他们会更谨慎检查该类任务，测试周的表现就不再代表日常流程。反之，如果测试是模拟环境，员工可能低估真实时间压力。可将测试与生产期错开、设置不同但同构的任务形式，并记录员工是否预先知道目标；同时保留真实质量抽查。测量效度不是“题目看起来像工作”，而是结果能否预测并解释目标情境中的行动。

38. 胜任阈值的错判成本会随岗位风险变化

培训项目常把通过率作为成功标准，但通过不等于能在工作中独立承担责任。胜任判断实际是一项决策：把员工授权给某项任务，或要求更多训练、监督和升级。漏掉未准备好的员工可能造成客户或安全损害；误把已胜任员工判为不合格则会阻塞晋升、增加专家负荷并降低公平性。阈值不应只由统计表现决定，还要显式权衡岗位错误后果。

一个纯构造的筛查例子：100 名员工中假设 80 人达到某项岗位标准，20 人尚未达到。某测试对达标者的灵敏度为 90%，对未达标者的特异度为 80%。则真阳性 72 人、假阴性 8 人、真阴性 16 人、假阳性 4 人；共有 76 人通过，其中 4 人尚未达到标准。通过者中的未达标比例是 4/76≈5.3%，但全体未达标人数仍有 20 人。这里每个性能参数和人数都是假设，计算只说明基准率如何影响通过者的风险，不能当作实际测验性能。

若任务后果可逆且有逐案监督，组织可以采用较宽的入门阈值，再通过抽查、升级和阶段性复评控制风险；若错误影响难以撤销的客户权益或安全，可能需要更严的授权门槛和独立验证。两种阈值会对假阳性和假阴性赋予不同代价。不能只报告总体准确率，因为在达标者占多数时，模型即使几乎不识别未达标者，也可能得到很高的准确率。需分别报告灵敏度、特异度、校准和按岗位风险分层的错误数，并检查金标准是否可靠。

把测试成绩用作晋升或排班依据还会产生反应。员工会练习可被考核部分，主管可能把难任务给分数高者，使后续工作结果差异进一步扩大。一次通过应限于规定的任务范围、时间和工具条件，不应转成永久“AI 能力等级”。岗位改变、模型版本、规则更新或员工长期未使用后，应重新评估相关能力；过度频繁地重测则会消耗工作时间并降低信任。

可以在授权政策中使用三级而非二元标签：在规定边界内可独立执行；需在常规复核下执行；需培训或专家接管。每个级别都附可观察的任务证据、复核比例、升级条件和有效期。对争议案例设置独立复核与员工申诉，并记录阈值变化

SVSAC | sv-sac.org

前后的授权差异。若某组通过率下降，先判断是标准变严、任务变难，还是能力真的改变；没有版本化量规，趋势图本身没有清楚含义。

39. 模型变化会改变任务要求，人员能力与工具能力必须分别记录

员工在一个工具版本中表现良好，不一定能适应另一个模型、界面或检索资料版本。建议质量、引用方式、自动执行权限和错误类型可能同时变化；更新后的错误减少也可能来自模型改进，而非员工技能提升。反过来，员工在流程修订后表现下降，也可能因为工作规则改变而非知识遗忘。把版本与技能测试绑定，才能分解这些机制。

可建立两个互补测试集：其一冻结模型和工作流版本，固定人机任务以比较员工的学习轨迹；其二在预定更新后用新版本重新测试，衡量员工适应变化的能力。两套测试不能直接拼成一个分数；前者回答同一工具条件下是否进步，后者回答变化后能否适应。若旧版本已经不可用，复现只能靠模拟或保留影子任务，不能把无法重现的历史差异说成精确因果结果。

对于高风险岗位，更新后可采用分层暴露：一小组员工先在受控任务中验证新界面，专家检查错误类别，确认常规能力和升级边界后再扩展；这不是把试点组当作随机对照，也不能从先行组的结果推断全体员工。先行员工通常更有兴趣、更熟悉模型或有较强主管支持。应记录谁被选中、为什么、其任务与其他人的差异，再在扩展阶段进行适当比较。

技能迁移也可表现为员工能否识别工具自身的知识边界。若模型改变检索资料或提示格式，员工需要判断证据是否足够、引用能否核验以及何时需人工升级。测试应包含模型输出看似流畅但依据错误、版本不兼容或资料不全的反例，并预设合理拒答行为。答案越流畅越不代表员工越能辨别，评分必须依据证据核对与行为选择，而非表达自信程度。

40. 内部流动政策决定培训能否转成职业机会

员工能力增加后，组织还需决定谁能进入更复杂的岗位、承担新的工作或获得薪酬回报。若岗位空缺有限，培训通过率提高并不保证更多员工晋升；若工作任务被重新组合，旧岗位名称可能不变而职责和门槛已改变。内部竞聘、岗位描述、资格规则、主管提名和外部招聘共同决定能力能否转成机会。

SVSAC | sv-sac.org

评估机会转化时，应先明确符合资格的员工和真实岗位空缺。培训后申请率、面试率、录用率、起薪、职责变化和六个月留任是不同转移结果；不能只比较成功晋升者。假如组织仅招两个岗位而有 40 人完成训练，晋升比例受岗位数量限制，不能单靠培训增效提高。可将申请者队列与全部符合资格者分别报告，并记录培训、原有绩效、主管支持、排班和岗位要求，防止把自我选择误认为能力收益。

当员工从被自动化的任务转向新岗位，评估还需计入过渡成本：补充训练时间、短期生产率下降、导师工时、原岗位收尾、地点或排班变化，以及员工未接受调动的后果。只统计新岗位人数会忽略被拒绝、离职或转入低薪工作的员工。若岗位改造提高了公司贡献但造成部分员工被迫转岗，组织净收益与分配结果必须分开报告，并为双方明确目标群体。

一个比较政策的设计可将可竞争空缺按透明资格规则开放，随机或分阶段提供额外辅导邀请，同时保留所有最初符合资格员工的原组分母。结果包括是否申请、是否被录用、工资和职责，以及未获聘者的后续工作。若招聘经理知道辅导分组，可能改变面试评价；若岗位数少，随机化的可行性和伦理性也受限。可以先评估透明资格规则本身，或使用自然形成的分批上线，但必须谨慎报告不能识别的部分。

41. 证据强度应随组织决策的重要性而提升

不同决策需要不同证据。界面小幅调整可以用短期任务比较；独立作业授权需要较强的盲态技能测试和故障演练；晋升、薪酬或裁员决定涉及较大后果，应依赖更成熟的员工和组织结果，且不能只凭模型使用数据或一次测验分数。证据链也并非必然线性：联合绩效改善不保证独立能力上升，能力上升不保证岗位增加，岗位增加不保证员工薪酬改善。

证据强度还取决于观察范围。任务日志覆盖全体但没有可靠结果标签，能显示活动分布却不能证明能力；少量深度盲评能审查质量，但可能无法代表全部岗位；行政工资记录较完整，却不说明员工为何换岗。多种资料可以相互校验，但不能把互不相同的测量简单合并为一个“人才升级”指数。报告应指出每个数据源的观察对象、缺失结构、主要误差和它不能回答的问题。

未来的有价值研究不是再证明“AI 有影响”，而是区分哪种岗位设计、练习结构和授权政策在何种员工、任务和需求条件下形成持久能力，并说明组织价值与员工机会怎样分配。可检验预测应明确相反结果：例如若只是联合工具能力而没

SVSAC | sv-sac.org

有学习，移除工具后的迁移不会超过合适对照；若关键技能来自反复练习，减少常规实践后而没有替代任务的队列应在延迟测试或接管能力上出现差异。未出现预测就应修正机制，而不是换一个宽泛的成功定义。

42. 证据结果不能按研究数量相加，必须先对齐估计对象

跨研究比较应先区分作用于哪一连接：提供建议是否改变带工具工作，练习与反馈是否留下无工具判断，组织是否进一步调整岗位与报酬。随机任务能较清楚比较特定工具条件，却不能独自跨越后两项；长期行政记录能观察后续结果，却不还原每次判断如何形成。这样的分工比按研究数量或观察期长短排可信度更有解释力，也避免把同一结果链的不同位置误作相互矛盾。

例如，客服人员获得 AI 建议的现场研究以每小时解决问题数、客户评价和处理质量作为结果；资深开源开发者的随机任务研究以预先定义的仓库问题耗时为主要估计对象[4][8]。前者是常规业务中的整套工作流，后者是特定工具可用与否下的任务时间；任务定义、质量门槛、样本和工具都不同。若两项研究方向不同，这并不自动否定其中一项，也不能直接取平均。解释差异需要检验任务成熟度、隐性上下文、质量要求、导师/专家支持和评价单位，而不能只用“行业不同”带过。

技能形成证据与工作产出也不能互相替代。程序员新库研究关注特定练习与工具条件下的短期知识；数学课堂文献则仍待原文核查，不作为已核比较[2][5]；短写作实验则测量任务输出与用时[3]。它们都不能独立回答参与者在几个月后是否能胜任岗位、晋升或获得更高薪酬。相反，劳动力市场研究观察更广的工资和工时结果，却无法把每个员工某一次 AI 使用与其个人技能变化完全对应[1]。这不是研究缺陷的简单排名，而是不同数据结构能识别的对象不同。

第 42 节的两组比较据此形成竞争解释，而非统一“AI 效果”。若当前任务改善却无独立迁移，工具可能承接了判断而未留下技能；若独立测验改善却无岗位变化，需求、授权或晋升安排可能阻断后续兑现。这些是本篇待检机制，不是原研究已连贯识别的链条。识别它们仍须按第 12 节区分访问、使用与训练，并保留共同辅导及完成者选择；相同研究主题本身不补足缺失连接。

SVSAC | sv-sac.org

43. 看似冲突的结果可能来自质量标准、任务边界或测量时间

对生产率的分歧常来自分母。增加的回答数可能意味着真实产出增长，也可能意味着任务拆分、重复尝试或缺少后续验收；更快完成某个环节可能把工作移到复核、返工和专家接手。要回答端到端工作是否改善，结果定义需覆盖同一批任务的初次完成、返工、升级、客户或内部验收，以及后续未解决的积压。仅靠一项速度指标不能分辨这些解释。

对技能结果的分歧则可能来自工具帮助与工具依赖。员工使用提示后答对题目，说明当前人机组合奏效；无工具测验下降可能代表迁移损失、短期认知负荷、前测差异或测试形式变化。应保留基线、等难形式、延迟测量和独立评分；若测验反馈也构成训练，应将它标记为处理的一部分。没有这些信息，“技能促进”与“去技能化”都容易被同一个结果支持。

对岗位和收入的分歧还受调整时间与制度影响。任务变化可能先出现在工作内容、主管分配或招聘审批，再反映在岗位数和工资。年度工资或人数暂时不变，不代表任务没有改变；早期工具使用率增加也不能证明长期岗位影响已发生。观察性研究需分别报告采用前趋势、工具使用代理、行业/职业冲击以及有无足够观察期。单个平均效应接近零时，还要问区间能排除多大的变化、是否能识别群体差异，以及采用扩散是否使比较组也受到影响。

负面和正面结果都需要竞争解释。若 AI 组速度更慢，可能是复核整合成本，也可能是任务难度或样本选择；若新手提高较多，可能是工具降低门槛，也可能来自基线任务更可标准化；若专家需求提高，可能是例外工作增加，也可能是团队把复杂任务集中给专家。分析应预先列出可观测预测和反证，而不是只在结果出来后挑选能解释它的故事。

44. 从研究结果外推到岗位，要逐项核对环境相似度

一项研究可否用于某个组织，至少要比较任务相似度、工作规则、工具组合、员工经验、质量标准、辅助资源、工作需求和劳动力制度。相同职业名称不保证任务相同；不同职业也可能共享核验、解释或异常升级等任务结构。外推不是将研究标签换成企业部门名，而是说明具体机制在目标岗位是否仍成立，以及差异会如何改变结果。

工作节奏也会改变帮助效果。任务低频、错误成本高且必须核验时，生成一个初稿未必减少总工时；需求高峰、积压且允许边界内快速处理时，能力补充可

SVSAC | sv-sac.org

能更有价值。若专家资源固定，工具增加待审输出会使排队上升；若员工能自行核验且任务风险低，队列压力可能较小。试点要记录到达率、服务时间、任务严重度、复核比例和专家容量，不能把异步的实验环境直接当作高峰业务。

组织文化和劳资安排也影响工作行为。若员工可以拒绝模型建议并报告风险，使用过程可能更稳健；若管理者把采用率绑定绩效，员工可能隐藏未使用、绕过系统或过度接受输出。工会合同、专业资格、客户协议和绩效计酬会限制任务和工资如何调整。即使某个技术处理在样本中有效，部署决策仍需确认它能否在目标权限、薪酬和申诉结构下执行。

跨情境验证可以分两步。先选若干任务差异明确的部门，在不改变实质工作要求的前提下，复现同一测量协议并记录失败；再在新的部门或版本中重新估计，不把前一试点阈值视作自然规律。若来源只覆盖美国客服或丹麦职业样本，跨国扩展应说明语言、劳动制度、工具权限和数据制度的差别。测量结果相似也不代表分配后果相同，雇佣、工资与岗位机会需在每个目标情境单独观察。

45. 一项组织政策的价值需要与可行替代方案比较

组织不能把“未部署 AI”作为唯一反事实。培训、流程简化、知识库更新、更多导师时间、人员重新配置或改善界面都可能解决同一问题。若 AI 方案需要额外复核、数据维护和转岗安排，而传统流程改善可以更便宜地获得类似结果，比较应包括这些可行替代。反之，若企业缺少专家而 AI 可以扩大服务能力，也不能只用工资节省评价其价值。

构造一个决策例：某团队积压任务可通过工具辅助、增加一名资深审核者或简化审批步骤三种办法处理。工具方案降低初稿时间但增加核验；资深人员增加现金工资却可直接处理例外；简化流程可减少多次审批但可能降低可追溯性。预算、质量、法律责任和人员数量应作为约束而非任意折算成一个分数。决策者可先排除不满足硬约束的选择，再比较剩余方案的质量、员工时间、等待和成本；若同一任务在不同政策下的真实结果尚未观测，应将排序标作待验证。

有价值的试点不一定要把每个组件都随机化。若风险允许，可对培训邀请或复核方式随机，固定工具版本并让所有人可访问；也可在团队容量有限时分阶段上线。关键是说明所估计的是哪项实际政策差异，以及有哪些共同要素。若所有组同时获得主管辅导、工作重新分配和新工具，试点只估计整套变革；若对照组员工通过共享知识库获得工具经验，比较是不同饱和度而非“有与无”。

SVSAC | sv-sac.org

政策价值还受持续成本和退出选项影响。年度工具许可、人员培训、内容治理和版本适配可能随用户数变化；临时咨询、系统集成和旧流程维护可能是一次性或分阶段成本。若项目能暂停而继续使用基本流程，成本有上限；若上线后取消会失去客户或服务连续性，退出成本更高。合约期限、数据迁移和岗位转换都应进入决策时点账本，但不能在没有合同和账簿时声称知道其具体金额。

46. 哪些观察能区分能力建设与单纯任务加速

如果组织主张员工能力提升，应能指出至少一条可观察、可反驳的预测：在同一目标任务的无工具或受限工具条件下，员工独立诊断、判断证据、拒绝不当建议或接管异常的表现提高；该改进在没有额外导师补贴时仍能保持；并且影响不只是测验复用或岗位选择改变。若员工仅在模型可用时更快，最稳妥的结论是联合绩效提高，不能再附加“人力资本上升”。

如果主张岗位重塑或职业升级，应观察责任、工作量、授权、岗位申请与晋升链的实际变化，并与未发生相关组织改革的比较相结合。新增职能也要区分真正创造价值的任务和维护工具的附属工作；岗位名称更改不是任务升级的充分证据。若关键指标只在留任者、工具高频用户或经理主动提名员工中改善，应将其作为选中群体内的描述并说明不能外推的路径。

如果主张企业收益与员工利益相容，应同时报告现金、产能、质量、客户价格、工资和工时，并逐项说明由谁支付、谁受益、收益在哪个窗口成熟。任何单一的平均指标都可能掩盖工资下降、审核强度增加、晋升机会集中或外包劳动变化。反证不是项目失败的装饰项，而是限定决策范围的实质信息；透明公开的无效结果也能帮助组织避免把暂时性效率、体验提升和长期职业收益误当同一结论。

47. 团队中的知识扩散会改变培训处理的定义

员工技能并不只由个人接触工具决定。一个人在处理异常时形成的判断可能通过结对、代码评审、交接记录或口头求助传给同事；反过来，共享提示、群聊答案和主管统一培训也会使未被分派工具的人接触到处理内容。此时，“个人是否拿到席位”不是完整处理定义。研究需要区分个人直接接触、所在团队的可用比例、同事互动网络和团队的流程改造，否则对照组实际上可能已经接受了部分干预。

SVSAC | sv-sac.org

这种扩散有两条方向相反的路径。若熟练员工把判断步骤讲给同事，团队可减少单点依赖，非使用者也可能更快学会识别例外；若熟练员工把模型输出直接发给同事，团队也可能只扩散答案而没有解释，短期速度提升却没有独立判断能力。两种路径对同一个“团队产出”指标都可能表现为改善，却对新人培训、错误复发和专家替代的含义不同。因而传播媒介、解释深度、受众是否实际练习和后续是否独立完成都应成为观察量。

个人能力与组织记忆也须分开。员工掌握一项技能，可能随其离职、岗位调动或长期不练而消失；文档、检查表和同伴网络则可在人员更替后保留部分流程知识，但不必然保留情境判断。组织不能把“知识库里有答案”当作“团队能识别问题”。更合适的评估单位是特定任务族在给定员工构成下的连续表现，并记录独立判断者数量、可替代性和关键例外的知识持有者。

可构造一个十人团队的例子说明计数偏差：开始时四人能独立处理某类异常，其他六人需要协助。一次培训后两名原本需要协助的人在无提示的新案例中通过独立任务，团队独立处理人数上升到六；若同时一名原有熟练员工离开，则现有独立人数变成五，而不是把培训的两人增量继续报告为六。若离开的是新晋两人之一，人数同样为五，但队伍构成不同：前一种情况保留两名新晋者，后一种保留四名原有熟练者。若工作任务依赖不同的经验深度，这两支队伍的例外处理范围可能不同；仅凭人数不能判定谁更强。课程完成率相同，也不意味着能力组合相同。这个算例只是定义口径；它不能说明培训引起了离职，也不计入完成质量、服务量或替补成本。

48. “同事教会了我”需要能与共同任务暴露区分

工作场所的学习常被从协作结果倒推：团队平均成绩提高，就被解释为知识扩散。但共同使用更好的模板、主管重新分配任务、需求组合转向简单案例，都会抬升团队均值而不发生同事间的能力传递。若要把变化归因于扩散机制，观察不仅要有最终正确率，还需记录谁与谁互动、解释发生在何时、练习内容是否共享、任务难度是否相近，以及接收者之后能否独立完成不同题面。

一种更有辨别力的测量办法，是把学习事件分成“答案流动”和“策略流动”。答案流动指向可直接复用的输出或提示；策略流动指向判断条件、验证步骤和何时升级的理由。盲评者可根据预先定义的编码表审查抽样交接记录，但编码只测得可见材料，不足以推断员工内化。独立迁移任务仍需由另一组评审者评分，避

SVSAC | sv-sac.org

免把交接内容与测验答案机械匹配。访谈可帮助解释行为，但不能取代行为结果或成为员工能力的独立标签。

一个可证伪的预测是：如果传播的是判断策略而非成品答案，接收者应在新案例、工具暂不可用或输入条件改变时仍更准确地识别例外；若优势只出现在曾共享的模板或相同案例，较合理解释是记忆、答案复用或任务熟悉，而不是可迁移能力。该预测仍需限定任务域，因为一个领域学到的校验法未必迁移到另一领域；也需预防“工具不可用”测试本身改变员工动机或被理解成惩罚。测量条件应在分派前说明并采用公平、可申诉的评价程序。

还应给负向扩散留下位置。若资深员工为满足吞吐要求批量转发未经核验的内容，团队成员可能学到绕过步骤；若员工因担忧评分而不愿向同事暴露不确定性，错误经验也不容易被共享。因此，互动频次不是学习强度的充分指标。对解释型协作应同时测量提问、纠错和拒绝行为，对任务完成则记录返工和升级，并区分员工主动求助、主管强制复核与系统自动提示。

49. 团队饱和度设计可以区分个人效应与同事溢出

同事互相影响时，个人随机分派后比较获分派者与未获分派者，在相应暴露假设下可估计混合工作环境中的分派差异，未必等于全员部署。若改按实际使用者分组，还会引入依从选择，不能继续称随机比较。未获分派者可能向队友学习或承担更多复核，获分派者也受团队覆盖率与工作分配影响；估计量须包含团队处理比例，而非假定结果只依赖本人分派。

一种作者提出的两阶段设计，先将足够数量的团队随机到若干席位饱和度，例如低、中、高，再在团队内部随机分配席位。令 Y_i(z,s)表示员工 i 在本人分派 z 和团队饱和度 s 下的独立任务结果。相同饱和度内比较 E[Y_i(1,s)−Y_i(0,s)]估计该混合环境中的个人直接效应；对未获席位员工比较 E[Y_i(0,s_high)−Y_i(0,s_low)]估计同事溢出；若关注政策，则比较接近全员部署的团队结果与接近零覆盖团队结果。每个差异回答不同问题，不能把直接效应加上一个未经定义的“扩散系数”就当部署总值。

如果组织不能在团队内随机席位，可以随机团队开放时间或分阶段培训，并把估计对象改为“被提供政策的团队”意向处理效应。培训参加、工具登录和持续使用是依从行为，按实际使用者比较会重新引入选择。席位比例应在随机化前定义，随机分配由谁执行及何时开放要留日志；临时调班、共享账号、跨团队辅导

SVSAC | sv-sac.org

和主管转发都可能使实际暴露偏离计划。主要分析应保留原分派，另做依从分析时明确其识别假设。

两阶段设计也有结构边界。它通常要求团队之间影响很少，或者影响网络可测且纳入暴露映射；如果员工跨组协作、共享服务台或使用同一公开知识库，所谓“低饱和度”组也可能处在高暴露环境。对照组污染不只是统计噪声，而可能反映企业真实部署的网络效应。可以先画出项目协作和信息共享关系，再选择随机单元；如果无法维持或准确记录边界，就把核心问题缩小为现行组织网络下的边际分派效果。

50. 一个构造性设计展示覆盖率如何影响决策

设想有 12 个可比较团队，每队员工人数不同，故结果按预先指定的目标员工总体加权，而不是把团队简单等权。研究者将团队随机分到 0%、50%、100%席位三种饱和度；在 50%组内随机一半员工取得席位。每人完成一项未见过、经盲评的工作样本，结果为是否在固定质量门槛内独立完成；同时记录团队复核时间。这里的 12 只是展示随机单元层级，不是足以保证统计把握度的样本量建议，实际团队数应由基线方差、集群相关、效应范围和可接受不确定性决定。

在 50%组，个人比较估计获得席位在半覆盖工作环境下的直接效应；该组未获得席位者与 0%组员工的差异，可能反映同事互动和团队流程改变的合并溢出。100%组与 0%组的差异估计两种团队政策包的影响，但其中包含个人使用、团队沟通、复核负荷和管理调整，不能单独识别某一机制。若企业真正关心在现有部门规模下全面部署，100%对 0%的政策差异更直接；若关心开放少量名额是否造成知识扩散，50%组未获席位者是关键比较。三个参数不能互相替代。

假设观测到 50%组未获席位者的独立通过比例为 0.62，0%组为 0.55，100%组为 0.68，50%组获席位者为 0.70。则本例中的估计差分别是：半覆盖环境直接差 $0.70-0.62=0.08$，未获席位者的组间差 $0.62-0.55=0.07$，整队全覆盖差 $0.68-0.55=0.13$。它们只是给定构造数字的算术，不是实测效果或显著性结论；不同团队的基线、席位分派概率和团队规模会改变估计精度，不能以四个比例代替标准误或随机化推断。

再假设 100%组每人平均节省 20 分钟复核时间，但独立任务通过率没有增加，且团队必须增加同等数量的解释记录工时。则可以说复核工作在任务间重新分布，但不能据此称员工能力提升，也不能将 20 分钟全部转成现金节省。若员工固定薪酬不变，节约的复核时间只有在被用于有价值的其他工作、减少加班/外包

SVSAC | sv-sac.org

或避免新增招聘时才可能形成不同类型的价值；三种渠道要分别记录。设计的价值在于把直接、溢出、政策及资源结果分开，而不是预先假定一个部署一定有净收益。

51. 离职与转岗会让“留任者能力”偏离培训总体效果

培训后只测仍在原岗位的员工，估计的是一个由培训后去留共同筛选出的群体。员工若因能力提升而获得内部调动，原团队的测验样本会减少；若员工因绩效压力或工作负荷而离开，测验缺失也可能与结果相关。把所有未测者删除，会把“培训对原分派员工的影响”换成“培训后仍留在该岗位者的差异”。将离职直接编码为能力失败同样不成立：离职可能是岗位机会、合同结束、个人选择或绩效问题，且彼此含义不同。

一个透明的界限例可以说明留任样本的风险。假设培训邀请组有 100 名员工，80 人完成延迟测评，留任者中 64 人通过，即观察通过率 80%；比较组 100 人中 90 人完成测评，68 人通过，即约 75.6%。留任者差约 4.4 个百分点，但这不是全体受邀员工的差。如果将未测者的真实通过状态视作未知，邀请组全体通过比例只能界在 64/100 至 84/100，即 64%–84%；比较组则在 68%–78%。两组差的最坏情况界限为−14 至+16 个百分点，包含负值和较大的正值。因此仅凭留任者差异无法判断全体员工能力平均提高。

这些界限很宽，不代表研究无用，而是显示需要什么额外信息。若行政记录能确认转岗者是否继续使用目标技能，可以对其在新岗位进行同一能力测量；若无法继续测量，可按“原岗位留任、内部调动、主动离职、解聘、联系失败”等预先定义状态分别报告去向和缺失率。用逆概率加权或插补收窄区间，需要明确测量缺失在已观测变量条件下可解释且各类员工都有被观察的概率；若离职原因或潜在能力同时决定失访，这些方法并不能自动恢复真值。应把敏感性参数展示出来，而不是把模型预测的缺失结果当作实测。

纵向能力建设的主要终点可同时包含全体原分派员工在固定时间点的任务表现、就业/岗位状态和技能可观测性。岗位状态是结果，不应因为员工离岗就把其从人数分母删掉；技能结果则须报告缺失和边界。若公司只关心当前团队可承担的工作量，应另报现有团队的实际容量，同时区分人员流出、补招、新员工培训和任务需求变化。两类估计对象分别对应员工发展与团队运营，均有价值但不能互相替代。

SVSAC | sv-sac.org

52. 能力是否可迁移，取决于新情境中是否保留关键判断

培训内容听起来“通用”，不代表员工能在另一工具、另一产品或另一责任制度下完成任务。迁移至少有三个层级：原任务熟练度、邻近任务迁移和跨制度迁移。原任务可能只需记住固定提示和界面位置；邻近任务需要把判断步骤用于不同案例；跨制度迁移还要求员工理解资料权限、验收标准和升级责任变化。测验若只改变题面、不改变工具或责任边界，不能证明跨制度迁移。

可把技能表述为一组行为条件，而非岗位标签。例如，“能核对模型生成的供应商付款建议”应拆成识别凭证不匹配、查验数据时点、判断金额风险、拒绝无依据自动通过及向正确权限人升级。若员工只能在固定屏幕上指出错误，而不能在表格导出或新版本界面中复现，表现可能依赖界面提示；若员工能解释证据链并正确选择升级渠道，才更符合跨任务能力的定义。即便如此，单次成功也不能排除偶然猜测，需要覆盖案例变体并报告不确定性。

训练收益的可迁移范围也影响企业和员工各自的激励。若所学是通用判断原则，员工可能把它用于其他雇主或职业；若所学只针对内部系统、私有语料和审批路径，价值可能更多留在当前组织。两者不是道德上高低之分，而是不同的投资与议价结构。企业可能愿意为专属流程培训付费，员工也可能更重视可携带技能认证；若试点只测当前平台操作速度，无法判断所投资的能力属于哪一类。

53. 培训成本由谁承担，与技能收益由谁捕获并非同一问题

一个训练项目可能同时消耗企业的工资时间、员工个人时间、导师支持、数据与系统建设；收益则可能体现为当前岗位质量、未来就业能力、较低的专家依赖或跨雇主流动。把这些项目相加为单一“培训 ROI”会掩盖承担者和受益者。工资等价时间是资源占用的估值，不等于现金流出；只有实际加班、替岗、外包或招聘避免等变化才能形成相应现金差额。员工获得可迁移能力，也不会自动变成当期工资增长。

雇主的培训投资决策可用一个简化条件表达：在给定期限内，企业预期获得的可归属收益现值要覆盖培训的可避免现金投入、学习期间产能损失以及持续维护成本。若培训提升的技能可以离开企业，企业可捕获的收益比例可能下降；但如果员工留任、团队互补或质量责任提高，组织收益也可能延长。该条件说明需要分别估计收益和期限，而不是对员工可携带性作负面判断。未观察到工资提高，并不能说明员工没有学到通用技能；只观察到企业节省时间，也不能说明员工薪酬或外部机会改善。

SVSAC | sv-sac.org

双方激励还会改变谁报名和谁参加。员工可能更愿意参加能带来晋升的培训；主管则可能优先选择当前最能提升部门产量的人，排除最需要基础支持者。若培训名额由经理提名，比较参加者和未参加者会同时混入预期能力、排班可用性和职业计划。透明规则、记录被拒绝的申请、邀请与实际出席，并分别观察内部调动和离职，可显示机会结构，但不能仅凭参加率判定公平或技能形成。

54. 用不同任务族的测量检验迁移，而不是只换一份近似试卷

一项可执行的研究设计可以先定义源任务、邻近迁移任务和远端任务族，并把每个任务对应的成功行为、证据标准、工具版本和责任边界写入测量说明。参与者在培训前完成基线案例，随后按团队或个人被分配到规定训练方案；短期测量同源任务，延迟测量用不同表面材料的邻近任务，再在合适岗位或模拟环境中测试更远任务。各组若同时获得导师时数或流程改造，估计对象是整套方案，应把这些共同处理记明。

为避免把考题记忆当作迁移，应在任务族内预先抽取案例版本，保留一部分完全不用于练习的测试情境。评分规则要事先定义关键错误、合理拒绝、是否需要升级与错误严重度；双盲评审者应不知道员工训练分组及测量时期。分歧应通过预先约定的仲裁解决，并报告仲裁前一致度和无法判定比例。若任务不存在可靠的单一“正确答案”，可评分证据引用与升级判断，不应为了统一分数把合理不确定性压成错误标签。

分析时既报告整体分派效果，也报告任务族和基线经验的异质性，但不能以事后找到的子组效果建立新认证。若迁移任务难度比基线高，应使用锚定案例和预先定义的等值规则；若模型更新，需保留旧版本能力测试和新版本适应测试两条结果。随机分派到训练邀请可识别邀请政策的平均影响；按完成训练或高频使用作比较会受依从选择影响。若研究关注“参加者受训”的效果，需要额外设计或陈述不可验证的依从假设。

研究还要区分迁移到真实工作和迁移到另一项测试。模拟案例提高可控性，却可能缺少真实延误、客户关系、协作和责任成本；真实业务任务包含上下游变化，不一定能独立归因。可以把盲评模拟作为能力端点，把实际质量、返工、等待和升级作为工作端点，逐项报告，不将两者合成一个不可解释的综合分数。安全关键的任务不能为了实验取消必要监督；高风险情境的迁移证据可先在受控环境积累。

SVSAC | sv-sac.org

55. 可携带能力与内部流程熟练度应有不同的结果指标

企业内常见的“熟练度”指标包括使用系统速度、模板完成率、系统内错误率和主管认可度。这些指标可服务于当前流程，但有些会随着界面变更而失效，也可能受到管理员权限和本地数据结构影响。可携带技能则更适合用跨界面、不同案例和独立证据评价，甚至需要跨组织共同认可的任务标准。组织不一定要建立外部证书，但应明确训练目标是提升内部合规操作、通用职业能力，还是两者兼有。

设想两种培训各投入同样的员工时间。甲培训让员工熟悉某内部采购 AI 工具的五个固定步骤，乙培训练习来源追溯、差异识别和不确定时升级。若当前月只测内部系统操作，两者可能都表现良好；若在系统更新或模拟资料冲突中测量，甲的优势可能消失，乙的优势也仍需由独立评分验证。这个比较不是说通用训练必然更优：若工作流程稳定且操作错误可造成严重损失，熟练遵循固定控制可能更重要。关键是培训目标与目标岗位风险、工具更替概率及员工未来流动范围相匹配。

员工发展也不应被“可迁移”单向绑架为离职风险。企业可能通过岗位路径、持续学习时间、透明晋升标准和合理薪酬分享保留人才；员工则需要了解哪些技能可自主携带，哪些资料和客户信息仍属受限工作材料。是否公开培训成果、如何记录能力证据、能否转入其他岗位，都应在培训前说明。数据治理可以保护员工免受过度监控，但也不能因害怕留痕而使员工无法证明新职责和专业成长。

56. 一项能力投资账本要让“现金、时间、留任和迁移”分列

构造一个年度项目：20 名员工各花 16 小时培训，按每小时 40 个货币单位估值，则时间资源为 12,800；若其中只有一半时间会引起加班或临时替岗，每小时实际可避免现金成本分别为 40，则对应现金支出为 6,400，而非 12,800。再假设导师时间的工资等价值为 1,000 货币单位、实际额外外包现金 500；系统建设现金为 2,000。账本应列总时间资源、可避免现金和一次性系统投入三列，不因每类都有金额就把它们当作同一种财务成本。

若培训后有两名员工离职，训练时间已经发生，但不能据此把全部培训费视为损失：他们可能在离职前贡献了质量或团队支持，企业也可能已回收部分收益。反过来，若留下者没有把技能用于工作，课程完成也不能自动证明价值实现。要估计企业净值，需按每期处理与比较条件记录实际现金、质量和容量变化，标记岗位流动、补招和导师投入；员工净收益则需看其实际工资、工时、岗

SVSAC | sv-sac.org

位选择和外部机会。合同与人员数据不足时，应给界限或说明无法估计，而不是将培训小时乘一个工资率当作利润。

再将技能可迁移性纳入决策，需观察多个结果窗口：训练后短期任务能力、数月后在岗位内的应用、岗位变更后的迁移表现和员工离开后的职业结果。离职后的工资可能不可得，且离职不是随机发生；不能只比较留下来的高使用者。企业和员工可以对同一项目有不同净收益，报告应分别呈现并说明是否有可避免现金、可利用产能或职业提升的证据。账本的目标不是迫使各方合成一个总分，而是说明哪些价值已观测、哪些仍是可能性。

57. 生产率变化不能替代岗位质量结果

岗位升级不仅是员工在更短时间内完成更多件数。工作节奏、任务控制权、判断空间、工作时间可预测性、重复性和例外责任都可能随 AI 进入而改变。一个系统可能减少复制录入，却把员工转为持续监控和处理异常；也可能让员工从重复任务转向更复杂的咨询。若只记录产量和工资，这两种结构会被压缩成相同的“效率提升”，而实际工作体验及技能发展的方向不同。

岗位质量指标应有定义边界。工作自主权可观察员工能否选择顺序、拒绝建议、调整节奏和使用专业判断；工作强度可观察单位时间任务到达、并行负荷、非计划加班和休息中断；任务内容则记录重复任务占比、异常复杂度、客户互动和责任后果。自评量表可报告员工感受，但不能替代排班、工单或实际决策记录；反过来，系统日志也不能完整观察被迫赶工、被监控感或工作意义。

这些维度不必全部变成一个福利总分。若自主性上升、强度也上升，员工评价可能因个人偏好、家庭责任和岗位资历而异；任意赋权重会让结果依赖研究者选择。更透明的做法是分别呈现预先定义的工作质量指标及其分布，并说明哪些是工作条件、哪些是感受、哪些是行为。组织还应明确不可接受的底线，例如不因模型提速就取消强制休息或专业复核；具体底线需按适用政策和职业风险确定，本篇不声称通用法律标准。

58. 节省的 20 分钟可以变成三种不同的工作日

考虑一个明确构造的八小时排班：员工原有六小时直接处理任务、60 分钟行政记录、30 分钟交接和 30 分钟可调度空档。工具若在任务中节省 40 分钟，却新增 20 分钟核验，净释放 20 分钟。若排班不变，这 20 分钟可能成为恢复、学习或

SVSAC | sv-sac.org

缓冲；若主管增加配额，则可转为额外处理；若行政系统要求记录模型使用和错误，也可能被新任务占满。20 分钟是容量差，不自动是现金、技能或福利结果。

设团队每日有 30 名员工，假设每人净释放 20 分钟，则总释放时间为 600 分钟，即 10 小时。若其中一半用于额外可接受产出，另一半被用于复核或辅导，则前者对应 5 小时实际产能、后者 5 小时支持活动；若任务不够或需求未成熟，产能可能闲置。即便 10 小时都转成工作，工资固定时也不是 10 小时工资支出减少。若确实避免了 10 小时加班，现金变化须按实际加班规则计算；若只是改派到其他任务，应记容量重配及其结果，而非现金节省。

三种工作日产生的员工体验未必一致。维持原配额并把部分时间用于自主学习，可能增加发展机会但需确认员工有权使用时间；增加配额可能提升团队吞吐量，也可能提高持续节奏和监控；让员工处理更复杂的例外可能提高判断空间，同时增加责任压力。对“效率换来更好工作”或“AI 导致工作加速”的判断，都应测量实际配额、节奏、加班、错误和员工报告，而不能从模型速度单独推得。

59. 将工具可用性与工作配额分开，识别提效如何转成工作条件

若问题是“工具会否改变工作质量”，单一工具处理臂不够。可在条件允许时设置工具可用性与配额政策两个因素：工具不可用/可用，与保持原任务要求/按规则增加目标工作量。配额规则需要对照劳动合同、服务标准与安全要求，不能为实验随机取消必要保护。随机单元通常应是团队或班次，以减少主管对同组员工差异化下达任务；组内沟通和跨组支援需记录。

这个设计可区分工具带来的组合变化与管理者如何分配释放时间。工具可用且配额不变的团队，可以检验额外时间是否流向复核、学习或等待；工具可用且配额上调的团队，可以观察生产量是否增加、工作强度和质量如何变化。若配额上调组有更多初次完成但也更多返工，净改善不能只按件数判断；若任务目标未变但等待减少，员工可能获得更好的可预测性，却没有工资增长。处理政策包分别估计，不把“模型效果”与管理规则混为一谈。

结果在员工、团队和服务层级同时记录。员工层包括时间分配、控制权、任务类型、加班和短期主观负荷；团队层包括复核排队、质量、互助及任务分配差异；服务层包括等待、客户结果和未结事项。主要指标须预先定向，例如较大的“控制权”是好，较大的“超时劳动”是坏；若只依赖一个总分，方向相反的变化会互相抵消。不同指标仍要报告区间和缺失，而非只给点估计。

SVSAC | sv-sac.org

60. 员工调查若只留在职者，会漏掉工作条件最差的一部分路径

岗位质量常通过员工问卷测量，但对工作压力更高、班次更不稳定或准备离职的人，回复概率可能更低。培训与工具政策也会改变谁仍在岗位、谁有时间填写问卷。比较调查响应者平均分，估计的是处理后的响应人群，不自动代表原分派员工；用系统日志补足问卷，仍无法推出满意度或自主感。

可在研究开始前确定简短问卷时间点、匿名原则、可以拒绝作答和不影响雇佣的边界；由独立研究或 HR 分析人员汇总最小必要数据，避免直接主管看到可识别的个人回答。若员工担心坦诚回答影响绩效，匿名度不足会导致测量偏差。还可将排班、加班和任务构成等行政数据与问卷分开报告，以免用单一数据源替代所有体验。

缺失处理需要展示而非隐藏。报告每组邀请人数、响应人数、岗位去向和不同问题的回答率；比较可观测的响应者与未响应者，但不能因为两者在已记录字段相似就证明未观测体验相同。对有界量表可给极端填补界限，并用敏感性分析说明需要多大程度的非应答差异才能改变结论。若随访回应持续下降，预先规定暂停扩展或增加独立沟通渠道的条件，而不是事后只保留可用样本。

61. 效率与岗位质量应当并列呈现，而不是强制折算

一项部署可以提高可接受产出，却同时增加监控、节奏和对例外工作的责任；也可能改善自主性但短期降低吞吐。组织面对这种取舍时，应分别展示可观测的产出、现金、容量、质量和岗位质量结果，再由有权负责的管理者、员工代表和专业负责人讨论约束。把“每人多完成几件”转换成单一效用分，既需要争议性的价值权重，也容易掩盖收益和成本由谁承担。

决策协议可在实施前划分硬约束、可补偿指标和未成熟结果。安全、隐私、必要休息和合同义务属于不能被效率分数抵消的边界；吞吐、等待和学习时间可以比较，但需披露其换算口径；长期晋升、外部流动和家庭时间影响可能尚未观察到，应标为未知。若董事会只批准有条件的小范围试点，扩张标准应明确需要看到哪些结果、哪些变化触发暂停、员工怎样提出申诉及谁有权撤回工具权限。

一种反证路径是：若模型提高件数但没有降低总任务时间、没有改善质量，且工作强度上升、员工控制权下降，就不能把结果描述为“人效提升兼岗位升级”；更窄的结论可能是产出目标下的任务加速。若总时间下降且工作条件不变，说明现有容量释放但员工福利未被证明；若工作控制改善、效率指标不变，仍可

SVSAC | sv-sac.org

能有岗位质量价值，但不能声称企业节省现金。不同结论对应不同政策，要求数据同时覆盖员工与组织结果。

62. 配额与工具交互的构造账本如何避免把件数等同于净价值

继续用完全假设的 100 个员工班次作例：未使用工具、维持原目标时，完成 80 件，4 件未通过质量门槛，故 76 件最终合格；使用工具但不改目标时，仍完成 80 件、2 件未通过，78 件合格，且每班次净释放 20 分钟；使用工具并提高目标时，完成 92 件、7 件未通过，85 件合格，但释放时间转为额外任务。这里的“最终合格”按完成件数减去未通过件计算，不假定失败件都能返工成功，也不把同一件重复提交计入产出。

只按完成件数看，后一政策相对基线增加 12 件；按质量通过件数看只增加 9 件；在同为 100 个员工班次的口径下，工具不改目标组 78 件相对 76 件增加 2 件，即每班次 0.78 件相对 0.76 件增加 0.02 件，且每班次净释放 20 分钟。后一个差异可以支持“给定目标下，质量通过略增且有时间余量”的构造性描述，不证明该时间已转成学习或现金。若目标提高导致加班、排班压缩或复核时间外移，就要补入这些结果；若质量门槛改变，三组数字不可直接比较。

企业在决定扩大前，应实际记录任务开始、首交、复核、返工、最终验收、工时来源和员工时间控制。若员工只报“更快”，但未出现总时间、质量或工作条件的独立变化，管理者应保留较窄的说法；若独立审计验证质量及总时间改善，再评估是否固定配额、释放学习时间或增量服务需求。三种选择是不同的管理政策，应分别检验，不由工具速度本身推出。

63. 企业内部岗位数可能低估被转移到外部的劳动

AI 进入工作流后，员工可能完成更多任务，也可能将初稿、异常核查、数据清理或知识维护转给供应商。若分析只看企业的在职人数，内部人数下降会被读作劳动需求减少；若只看外包支出，又可能忽略仍由内部员工投入的监督和返工。企业边界与任务边界不再重合，尤其在平台、咨询公司、专业服务商和客户共同参与的流程中。

外部任务转移可以是技术采用的真实组织结果，但并不自动等于就业损失或效率提升。供应商可能使用更专业的人员、分摊固定工具成本，或承担原本无法稳定配置的夜间工作；也可能把低可见度、重复或高波动的任务外移，使质量控

SVSAC | sv-sac.org

制更困难。合同单价包括供应商管理和利润，不能直接换算成供应商员工工资；客户企业内部减少一个岗位，也不能证明全劳动网络少一个岗位。研究应识别任务、雇主、工时与责任分别落在哪里。

构造账本设内部基线为 100 个 FTE，重组后实际未补 12 个，内部为 88，供应商另增 8 个可比 FTE，则内外合计 96；内部下降 12%，合计下降 4%。这只是给定相同工时口径、质量与产出的条件账本，不是 AI 因果效果。若 88 个内部 FTE 之中有 4 个从原任务重配到监督，合计仍为 96，不能重复加 4；若是在已计 88 之外另发生 4 个 FTE 的加班或新增劳动，合计才为 100。必须先说明是重配还是额外工时，人数本身不能裁定。

同一账本还要区分预算职位、空缺和实际雇佣流量。若计划削减 12 个岗位但只实际取消 7 个，另外 5 个仍在招聘，不能把“计划不补招”写成少了 12 名员工；若供应商合同只买服务而不给人力小时，也不能无根据地把合同金额换算成 8 名员工。需要公司人事流量、供应商工时或可信的人员等效、交付量和质量记录；缺一项时应报告边界，而不是用合同额填补人员数。

64. 把外包视为替代、互补或风险转移，需观察不同结果

外包与 AI 可能是替代关系：模型稳定处理常见案件，供应商工时下降；也可能是互补关系：自动生成更多初稿后，供应商负责验证、语言适配或夜班处理；还可能是风险转移：低可见度的纠错任务从客户企业移到分包链。三种路径在企业系统上都可能表现为“每名员工产出增加”，但对就业、技能和服务可靠性的解释相反。

组织应按任务族追踪内部员工和外部人员的工时、工作类型、复核责任、返工、响应延迟与技能要求。若供应商人员变动频繁，客户企业可能把重复培训成本隐藏在单价中；若供应商共享同一平台服务多个客户，客户间任务会竞争有限专家容量。仅观察主合同金额无法区分服务数量、价格变动、技术成本、供应商利润和劳动投入。跨外包单位比较时，还须确认“一个小时”的计算包括了多少管理、待机和再交付时间。

劳动条件也应有独立观察。企业员工的薪酬、加班和晋升信息通常不覆盖供应商及再分包层级；客户公司不能仅凭供应商合同宣称外部人员技能提升或工作质量改善。若研究没有供应商人员许可下的工作数据，可以报告采购的服务与交付指标、明确外部劳动结果不可见，而不是假设外包人员享受同等培训、保障和

SVSAC | sv-sac.org

职业路径。这样的边界限制外推，但比把企业自身人数当成整条流程就业更准确。

65. 对组织政策进行估计时，应将供应商响应纳入处理结果

若企业在 AI 政策上线后改变采购、合同范围或供应商验收标准，估计“AI 方案的组织效果”应把这些反应纳入政策包。把供应商工时当作上线后的协变量并从结果中扣除，会移除政策可能造成的外包路径；反之，如果研究只关心企业内部员工表现，则应明确它是窄口径员工结果，而不是企业全流程生产率。

研究设计可以固定主要供应商和合同窗口，记录服务量、小时、价格、人员等效和交付质量，再按团队或业务线分阶段提供 AI 工具。若供应商同时服务处理组和对照组，处理经验可能通过共用专家、流程模板和软件配置扩散；此时单位间干扰发生在采购网络，而不只是公司员工之间。分析需记录共享服务、合同与人员交叉，并把直接客户团队效应和覆盖供应商后的政策效果分开。供应商不能提供个人记录时，最低限度仍需每类服务的小时与成果数据。

如果随机化不可行，可以使用既定合同到期或业务线排期作为自然变化，但排期常与绩效、管理层重视度和业务需求相关。前趋势及固定效应只能处理可观测结构，不能自动消除合同调整的选择。较稳妥的研究可以将上线事件与采购订单、交付质量、内部人事流量、客户服务指标时间对齐，并报告替代解释；没有干净的比较组时，结论应限定为机制追踪而非政策因果效果。

66. 招聘和岗位空缺是组织选择结果，不是自动化影响的同义词

管理层可能通过不补招把需求下降转成自然减员，也可能在员工离职后招入不同技能的新员工；某个岗位停招并不意味着其任务消失。任务可能被现有员工吸收，转给更高资历者，交给供应商，或因业务量变化而取消。把人数差除以基线头数会遮住这些路径。员工数量、全时等效工时、任务承担者、职位空缺和招聘批准状态应分别记录。

估计 AI 上线对招聘的影响，需要定义风险集和观察窗口。一个岗位若未获预算批准，不能同时算作“发布后无人申请”的招聘失败；一条职位公告被撤回也不等于职位由 AI 自动化。可对预先存在且符合条件的职位或业务单元建立基线，记录职位开放、申请、录用、入职、取消和替代合同的日期与原因；对员工

SVSAC | sv-sac.org

流出记录自愿离职、退休、内部调动、裁员和暂时冻结等互斥状态。报告净头数和每类流量，避免用两个时点的差值推断因果。

一种可证伪的招聘预测是：若 AI 主要减少常见任务的岗位需求，变化应集中于对应任务强度高的岗位，同时残余岗位的任务内容和例外处理时间可观察地变化；若只见职位数量变化，却看不到任务负荷、合同投入或处理量改变，宏观业务收缩或管理冻结仍是竞争解释。反过来，职位不减少也不证明岗位任务未变，因为自然减员、内部转岗和新任务可能抵消。数据必须把任务结构和人员流量联合起来。

67. 一条简化工作流的全劳动账本

另一个单独构造的季度工作流：内部 100 名员工基线完成 10,000 件合格工作；另有供应商投入 4,000 小时处理外语资料和疑难复核。工具上线后内部员工数为 96，内部工时下降 1,600 小时，供应商时数上升到 5,000，合格件数升至 10,500。只看内部人数会称下降 4%；只看总合格产出会称上升 5%；只看外包时数会认为外部负担增加 25%。三项描述都成立，却分别对应人员、产出和供应商劳动，不是同一个生产率估计量。

假定基线为 100 名员工、每人季度 400 小时，仅作构造，则内部投入 40,000 小时，另加供应商 4,000 小时，基线总计 44,000 小时。上线后 96 名员工各 400 小时，内部为 38,400 小时，另加供应商 5,000 小时，总计 43,400 小时。合格量由 10,000 增至 10,500 件，总劳动减少 600 小时，每千小时合格产出由约 227.27 增至 241.94。该比率依赖内外工时同口径、质量门槛不变、加班完整记录等条件；需求或合同同期变化也需另辨，不能据比率归因 AI 或计算现金收益。

还须单独列账：96 名员工的工资和工时变化、供应商单价与合同最低采购量、上线后合计 5,000 小时（较基线增加 1,000 小时）由哪个供应商或分包商承担、工具许可和专家复核现金、工作质量与客户结果，以及不合格件返工。若内部人员成本固定，减少 1,600 小时是容量释放而非同额工资支出减少；若供应商按固定季度保底付款，时数上升也不等于支出同比上升。为避免双计，同一复核小时不能既作为供应商劳动投入又作为内部返工成本计算两次。报告应分别呈现资源使用、现金与单位产出，不在缺失合同数据时估计净利润。

SVSAC | sv-sac.org

68. 证据支持的是一条有条件的组织机制链，而不是“AI 提升人才”的单一效应

本篇的核心问题可以回答为：AI 使用何时形成可迁移能力并改善组织绩效？本篇不给平均提升率，而提出须分别检验的机制链；以下安排是拟议条件，不是从所引研究识别的普遍必要条件。员工必须有机会练习判断，而不只是接受模型答案；团队要有时间核验、解释和修正；组织要把释放的容量用于学习、质量或服务，而非默认提高配额；岗位与薪酬制度还要允许获得的能力转化为职责或职业机会。即使当下联合产出改善，未独立测量能力或职业结果时，也不能据此断言相应升级已经发生。

现有研究提供了互补而有限的证据。单个客服企业的研究观察了在职工作流中 AI 助手与处理效率、质量和经验差异相关的变化[4]；资深开发者任务实验显示工具可用在其研究设置中未自动缩短完成时间[8]。已读短期写作和编程学习涉及任务表现与即时知识，学校数学文献则仅为待核线索[2][3][5]。丹麦劳动力市场研究观察了更广的职业与工资变化，但采用并非随机，也不是个人培训试验[1]。这些结果不应被拼成同一个“AI 生产率”或“技能收益”参数；它们共同提示处理定义、任务上下文、评价标准和时间范围会改变结论。

因此，企业决策需要把几类价值并列：即时任务产出、独立能力、团队知识覆盖、工作质量、岗位条件、员工职业结果及企业现金或可用产能。减少任务时间本身可能有组织价值，即便劳动关系不变；员工掌握可迁移能力也可能有个人价值，即便短期工资尚未调整。两者都不能被另一方的结果替代。成本亦需按主体和口径分列：固定工资下释放的工时是容量，不是现金节省；企业端现金增加不等于员工净收入或全社会剩余增加；内部人数变化不代表外部供应商劳动同时减少。

69. 研究局限和下一步可证伪的比较

本次分析没有企业 HR 记录、员工调查或实际干预，因此不能报告任何组织的 AI 采用率、职位变化、工资效应或培训回报。已读相关原段涉及客服、开发、写作与编程学习，数学教育仅待核线索，随机单位、工具年代、样本经验和终点差异较大；劳动力市场分析观察期较长，却有采用选择和宏观共同冲击边界。文中团队饱和度、岗位质量与供应商账本均为作者提出的研究设计或构造例，不是已完成现场试验，也未合并估计效应。

SVSAC | sv-sac.org

后续比较可沿已给出的主链推进：第 12/26 节的工具与学习设计，第 47–50 节的个人效应与团队溢出，第 51–55 节的原分派队列、缺失及迁移，第 56–67 节的现金、配额、岗位质量和供应商劳动账本。各部分先说明可识别的政策对比，再决定是否值得延长随访或扩大样本；这些均未在本篇实施。

外推边界不因流程完整而消失：共享资源使个体对比不同于全员部署，共同改变版本或配额使前后差值不同于纯模型效果，练习题反复使用也不同于迁移。薪酬与劳动条件未随访时保持未知，失败、无效与缺失界限须同样报告。它们决定暂停或收窄何种主张，不能改写成组织必然成功或失败。

70. 决策建议

AI 能否推动组织与人才升级，取决于工具接入是否嵌入可练习、可复核、能迁移的工作方式，以及组织是否把人机协作的变化转化为可持续的能力组合和明确的岗位安排。现有证据足以反对“采用即提效、提效即学习、学习即晋升”的直线推论；也不足以支持“工具必然去技能化”或“岗位必然消失”的相反概括。短期联合产出、员工独立能力、劳动需求、薪酬、岗位质量和企业净收益是不同结果，应在各自适当的观察单位和窗口中判断。

对管理者较稳健的启示是先把任务、权限、复核和员工结果定义清楚，再用受约束的试点检验工具与训练各自贡献；记录员工实际获得的学习机会、例外责任和退出路径，并将现金成本、时间资源和岗位影响分开报告。只有在质量、迁移和工作条件的证据达到与决定后果相称的程度后，才有理由扩大授权。这样的条件性做法不保证部署成功，但能避免把系统访问、测验分数或短期产量误当成人才升级的充分证据。

SVSAC | sv-sac.org

给企业的下一步建议

AI 转型的组织红利，来自岗位协作与能力重组；单纯购买工具，很难形成持续优势。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。

具体起点是：按任务而非职位口号盘点工作，明确员工、业务负责人和技术团队的责任，结合真实项目安排培训。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，进一步形成岗位与任务诊断、人才能力地图、组织协同与培训路线图。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 AI 进入工作场所后，组织常同时观察到任务完成更快、员工更频繁调用模型、审核与返工负荷增加，以及新职责逐渐形成。它们不是同一结果，也不会自动沿“效率提升—能力升级—岗位改善—利润增加”传导。本篇提出的问题是：在何种任务、学习设计、权责安排和劳动力市场条件下，员工使用 AI 能够形成可迁移能力并改善组织绩效；又在何种条件下，短期产出改进会伴随基础技能练习减少、专家复核集中或岗位收益未变？回答需要把工具可用、实际使用、任务质量、离开 AI 后的独立能力、任务与岗位变化、工资工时和企业净收益分开测量。

本篇采用范围限定的原始研究梳理与机制分析，不使用企业内部人事数据、员工访谈或实际培训试验。证据包括丹麦早期劳动力市场的调查与行政记录差分研究，以及一个针对有编程经验参与者的短期随机学习实验。前者报告了工作内容变化与平均登记工资、工时估计接近零并存；后者在特定新编程库任务中发现即时完成时间没有显著平均改善，而离开 AI 后的知识测验分数较低。两者的处理、对象、观察期与结果单位不同，不能合并成普遍的“AI 重塑岗位”或“AI 损害人才”效应。本篇据此建立任务—学习—岗位—分配的分析框架，并提出可检验的条件性命题与后续评估设计。所有建议和构造框架均为作者分析，未声称组织已实施或取得商业效果[1][2]。

SVSAC | sv-sac.org

参考文献

[1] Anders Humlum; Emilie Vestergaard. Still Waters, Rapid Currents: Early Labor Market Transformation under Generative AI. RFBerlin Discussion Paper Series 078/26 (March 2026; discussion paper marked as preliminary and potentially not peer reviewed) (2026). https://www.rfberlin.com/wp-content/uploads/2026/03/26078.pdf

[2] Judy Hanwen Shen; Alex Tamkin. How AI Impacts Skill Formation. arXiv preprint 2601.20245v2 (version posted February 1, 2026; manuscript body dated August 24, 2026; not asserted peer reviewed) (2026). https://arxiv.org/abs/2601.20245

[3] Shakked Noy; Whitney Zhang. Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. Author working paper dated March 2, 2023 (the paper was later published in Science; the version read here is not the journal typeset version) (2023). https://economics.mit.edu/sites/default/files/inline-files/Noy_Zhang_1.pdf

[4] Erik Brynjolfsson; Danielle Li; Lindsey Raymond. Generative AI at Work. The Quarterly Journal of Economics 140(2), 889–942; Advance Access February 4, 2025 (2025). https://academic.oup.com/qje/article/140/2/889/7990658

[5] Hamsa Bastani; Osbert Bastani; Alp Sungu; Haosen Ge; Özge Kabakcı; Rei Mariman. Generative AI without guardrails can harm learning: Evidence from high school mathematics. Proceedings of the National Academy of Sciences 122(26), e2422633122; published online June 25, 2025; official correction in PNAS 122(34), e2518204122, published online August 20, 2025 (2025). https://www.pnas.org/doi/10.1073/pnas.2422633122

[6] Daron Acemoglu; Pascual Restrepo. Automation and New Tasks: How Technology Displaces and Reinstates Labor. Journal of Economic Perspectives 33(2), 3–30 (2019). https://www.aeaweb.org/articles?id=10.1257/jep.33.2.3

[7] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1, July 2024 (2024). https://doi.org/10.6028/NIST.AI.600-1

[8] Joel Becker; Nate Rush; Elizabeth Barnes; David Rein. Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity. arXiv preprint 2507.09089v2, last revised July 25, 2025; not represented as peer reviewed (2025). https://arxiv.org/abs/2507.09089

SVSAC | sv-sac.org