

DECISION BLUEPRINT / OVERVIEW

A Blueprint for Adopting AI and Measuring Its Impact

The Question This Blueprint Answers

How should we implement AI in our organization to improve staff productivity and increase profitability, while protecting confidential information, ensuring the accuracy of AI-assisted work, controlling costs and avoiding operational disruption? Where in the business should we start, how should we pilot AI before rolling it out more widely, how do we encourage staff adoption, and how do we determine how much of any change in revenue, costs and profit is attributable to AI? How do we stay current with the AI advances relevant to our business?

What This Document Helps You Decide

Use this guide to select a first application, define how it will operate, test the complete workflow, prepare staff and decide whether wider use is justified. The companion Worked Example applies the same nine decisions to Bellhaven Accounting. The guidance is intended for business owners, managers and operating teams across industries.

FINDING Begin with a specific business problem and compare practical ways to solve it. Test a defined set of tasks with a clear budget, quality standard and comparison. Expand AI use only when the complete process produces a worthwhile result.	WHAT COULD WEAKEN THIS FINDING Published research cannot establish the benefit in your organization. Results depend on the available records, customer demand, staff skills and system access. The benefit also depends on which employees save time and whether those employees can use the saved hours for other necessary work.
---	--

THE QUESTION ASKS	WHERE THE ANSWER IS DEVELOPED
Where should we begin?	D1 to D3 define the objective, map the work and compare implementations.
How should we pilot and protect information?	D4 and D5 specify controls, comparison methods and decisions after the test.
How do staff adopt AI effectively?	D6 covers permitted use, practical competence, incentives and management responsibilities.
What change is attributable to AI?	D7 explains how to measure AI’s effect on work, revenue, spending and profit while accounting for other changes.
How do we control costs and stay current?	D8 and D9 examine operating economics, approved configurations and repeatable evaluation.

HOW TO USE THE COMPANION DOCUMENTS
Read the Blueprint for the general method and the Worked Example for a concrete application. D1 to D9 refer to the same decisions in both documents. The case demonstrates conditional reasoning; it does not report an implemented deployment.

Begin with page 2. The source register distinguishes research, engineering guidance and legal material. Implementation must satisfy the obligations applicable to your organization.

BRIEF / NINE DECISIONS

Nine Decisions That Determine How Your Organization Should Adopt AI

The decisions are connected. The goal determines what to measure. The work determines which technology fits. Safeguards determine the permissible test. The pilot then supplies information for a deployment decision and a financial assessment.

DECISION	QUESTION TO ANSWER	READ
D1. Goal and limits	What improvement matters, compared with what baseline, within which cash, quality and operational limits?	page 6
D2. Work and information	Where do work, customers or decisions wait, and what information is needed to complete them?	page 7
D3. Starting application	Which AI application, staffing change or process improvement could realistically achieve the goal?	page 9
D4. Safeguards	What information and actions are allowed, how are outputs verified, and how does the system recover?	page 12
D5. Pilot	How will the test distinguish technical reliability, workflow performance and business impact?	page 15
D6. Staff adoption	What must staff learn, what discourages use, and who assigns other work during the hours AI saves?	page 19
D7. Attribution	What changed because of AI, and how much of that change became revenue, expenditure or profit?	page 21
D8. Costs and procurement	What does each result that meets the required quality standard cost at expected volume, including staff review and maintenance?	page 24
D9. Ongoing review	Which changes warrant retesting, expansion, repair, replacement or retirement?	page 26

Distinguish Improvements in Work From Changes in Revenue, Spending and Profit

Operational measures describe how work is performed: how many tasks meet the required quality standard, how long completion takes and how much staff effort is needed. Financial measures describe changes in revenue, spending and operating profit. In this document, additional contribution means additional revenue less the additional costs of delivering that revenue. When AI saves salaried staff time, spending may remain unchanged. Profit improves only if the saved hours support additional profitable work or remove spending that would otherwise occur.

Return to Earlier Decisions When New Facts Change the Answer

A pilot may reveal that review takes longer, customers do not want the proposed service or an existing application already performs the task. Reopen the affected decisions and record the reason. Preserve the original expectations so later results can be compared with them.

SET A PURPOSE AND SPENDING LIMIT FOR EACH TEST

Specify what uncertainty the test will resolve, its spending and staff-time limits, and what result would change the next action. Learning can justify a small experiment without justifying continuing deployment.

RESEARCH / TASK PERFORMANCE

Use Published Research to Decide Which AI Applications to Test in Your Organization

The studies below examine different occupations and tasks. Use their results to identify promising applications and likely failure cases. Compare each study with your own work before deciding what to test.

FINDING
AI can improve well-defined tasks, but the direction and size of the effect depend on the work, the user and the implementation. Research helps identify promising tests and failure modes.

WHAT COULD WEAKEN THIS FINDING
Different tools, samples and outcome definitions limit direct comparison. Favorable task studies do not include every integration, review, training and operating cost.

RESEARCH SETTING	WHAT WAS OBSERVED	WHAT THE ORGANIZATION SHOULD DO
Customer support [R01]	AI assistance improved output in the studied setting, with benefits differing by experience.	Test performance by staff group and task; do not apply the average directly to a whole department.
Consulting tasks [R02]	AI improved some tasks while reducing accuracy on a task outside the tool’s capabilities.	Include realistic difficult cases and independent review, rather than selecting only successful demonstrations.
Software development [R03]	Three randomized field experiments involving 4,867 developers found about 26% more completed tasks on average.	Measure completed work that meets the required quality standard and any subsequent repairs. The amount of code produced does not establish business value.
Experienced developers [R04]	An earlier small trial found slower completion. The follow-up had selection problems that limit its effect estimate.	Use observed task records and a credible comparison; staff expectations can differ from measured outcomes.
Accounting [R06]	The published study combines observational productivity results with an experiment on transaction classification. Experienced accountants used assistance more strategically.	Keep observational and experimental findings separate, and test junior and experienced staff rather than assuming the same effect.

Identify What Is Similar to Your Work

Compare input quality, task complexity, staff experience, available tools and the definition of success. A short exercise with complete inputs differs from a live workflow with missing records, interruptions and customer obligations. The closer the match, the more useful a study is for choosing a test.

THE ACCOUNTING RESULT REQUIRES CAREFUL INTERPRETATION
The accounting study associated a one-standard-deviation increase in reported AI use with about 19% more clients supported weekly. A standard deviation measures how spread out the reported levels of use were; this comparison is not the effect of simply introducing a tool. A separate transaction-classification experiment improved accuracy by 18 percentage points. Neither result establishes the increase in profit an organization should expect. [R06]

These studies support selective testing. They do not establish a general productivity percentage for an organization.

RESEARCH / ORGANIZATIONAL RESULTS

Determine Whether Faster Task Completion Improves Service or Financial Results

A tool may shorten one task while leaving total completion time, customer service or profit unchanged. The research below examines broader results and helps identify what must be measured beyond the speed of a single task.

FINDING
A task improvement becomes an organizational improvement only when other parts of the process can use it. The financial result also depends on demand, pricing and expenditure decisions.

WHAT COULD WEAKEN THIS FINDING
Some benefits appear as shorter hours, improved quality or faster service instead of profit. Those benefits can matter, but should be measured under their own objectives.

SOURCE	RELEVANT RESULT	LIMIT ON INTERPRETATION
Office work [R07]	The paper reports about 12% less time spent on email among workers given access to AI. A separate estimate attributes a 17% reduction to additional use caused by that access. The paper did not detect broad changes in how tasks were organized.	Giving workers access and increasing actual use are different comparisons. Neither estimate establishes how saved time was used or whether profit increased.
Human and AI advertising teams [R08]	The revised experiment reports 50% more advertisements per worker, with different effects on text quality, image quality and output diversity.	A creative task experiment does not establish the economics of autonomous agents across a business.
Administrative labor records [R05]	No detectable near-term change in recorded earnings or working hours in the studied setting.	This does not prove that no tasks improved or that all later financial effects are zero.
Organization surveys [R09, R10]	Reported benefits are associated with workflow changes and management support. Many workers report limited direction for using time saved.	Respondent selection, recall and self-reporting prevent a causal reading.
Software-team activity records [R11]	Higher code output was associated with longer review times; organizational measures did not show the same gains.	Observations from software teams do not establish that insufficient review capacity limits output in every business.

Measure the Entire Process From Receiving Work to Delivering an Acceptable Result

Follow work from its arrival to a result that meets the required quality standard. Include preparation, waiting, approval, correction and customer response. If preparation becomes faster while review remains fixed, completed output may not increase. If AI also helps reviewers, that relationship can change.

Assess Whether Staff Retain the Skills Needed to Verify AI-Assisted Work

Research on training and professional expertise gives reasons to test structured instruction and unaided performance. It does not support assuming that AI inevitably weakens judgment or that experience becomes irrelevant. Measure whether people can identify and explain consequential errors. [R12, R13]

MANAGEMENT ACTION REQUIRED
Before launch, name the work that hours saved by AI should serve and the person responsible for assigning it. Then measure whether the reassignment occurs. The rule is a practical recommendation; surveys alone do not prove its effect.

IMPLEMENTATION / CURRENT CAPABILITIES

Choose an AI System That Can Perform the Tasks Your Organization Needs

An AI assistant helps a person produce or interpret work. A fixed AI workflow follows a predetermined sequence with AI at selected steps. An agent can choose intermediate actions and tools within limits. The degree of autonomy is a separate permission decision. [G04]

WORK REQUIREMENT	IMPLEMENTATION TO EXAMINE	WHAT MUST BE DEMONSTRATED
Stable rules and structured inputs	Use conventional automation for calculations, validation and routine transfers.	Exceptions are identified and handled correctly. AI adds no unnecessary uncertainty.
Predictable steps with interpretation	Use a fixed workflow for extraction, classification or summaries.	Every step and the completed workflow meet quality and time requirements.
Answers from company records	Retrieve approved sources with permission controls and source references.	The response uses current, relevant records and exposes missing information.
Scans, images and mixed documents	Use multimodal interpretation, meaning a system that cwork with more than text alone.	Fields and totals reconcile; uncertain readings require attention.
Variable tasks across systems	Use an agent with limited tool permissions, recorded actions and procedures for recovering from errors.	Correct outcomes justify the extra integration and supervision.
A narrow software gap	Use AI-assisted development of a small internal application.	A named owner can maintain the software and resolve defects.
Independent subtasks	Consider multiple agents with a defined method for reconciling results.	The measured benefit exceeds coordination, duplication and verification costs.
Forecasting or anomaly detection	Compare statistical models and predictive AI with simple rules.	Performance improves on representative held-out data and remains useful in operation.

Use Published AI Test Results to Identify Systems Worth Testing on Your Own Work

Published benchmark tests compare AI performance on specified tasks. Results depend on the tools, spending limits and definition of successful completion. Some report the duration of tasks that a system completes successfully half the time. That measure does not establish how long the system can safely operate without supervision. Test each candidate on a retained set of your own representative work before deployment. [R14, R15]

A PRACTICAL DEFAULT

Begin with the simplest software arrangement that can complete the task reliably. Add an agent that chooses its own steps, stored information carried between tasks or several cooperating agents only when those functions are needed and testing shows that their benefit justifies the cost.

D1 / GOAL AND LIMITS

Define What AI Must Improve and Which Limits the Implementation Must Respect

Write the intended result in terms of the work or accounts, then state the period and baseline. Specify whose result matters: an employee, a department or the organization. Keep financial objectives separate from service, quality and employee objectives so success in one category does not conceal failure in another.

DECISION TO RECORD	EXAMPLE OF A PRECISE REQUIREMENT
Outcome and measure	Increase annual operating profit after all incremental AI costs; also hold material error rates within the approved standard.
Baseline and alternative	Compare with the spending and staffing plan that would apply without AI, including changes already scheduled.
Cash and capacity limits	Set a first-year cash ceiling and a maximum staff-time commitment for setup, training and supervision.
Information and consequence limits	Identify prohibited disclosures and actions requiring qualified approval.
Timing and disruption	Identify peak periods, service commitments, manual fallback and the acceptable interruption period.

Decide What Will Count Toward the Goal

Track total improvement and AI’s estimated additional contribution separately. A project may include software, training, process changes and staffing adjustments. The reporting rule should be agreed before the result is known; the evaluation design determines how much can be separated.

Enter the intended outcome, measure, baseline and time horizon.

Enter the cash, staff-time, quality, information and disruption limits.

Enter what will count toward the goal and who approves a change to that definition.

DO NOT SET A TECHNOLOGY TARGET IN PLACE OF A BUSINESS OBJECTIVE

The number of licenses purchased, prompts submitted or agents deployed measures activity. Those counts are useful only when connected to an intended operating result.

D2 / MAP THE WORK

Identify Which Steps Delay the Completion of Work in Your Organization

Map one complete service or business process before proposing tools. Record incoming work, preparation, decisions, review, delivery and rework. A constraint is a step whose limited capacity prevents the desired result. The constraint may change with season, customer mix or volume.

AREA TO EXAMINE	FACTS TO COLLECT	HOW THE FINDING CHANGES THE CHOICE
Demand and conversion	Record inquiries, proposals, conversion, lost work and reasons customers decline.	If demand is insufficient, additional production capacity may have little immediate value.
Missing inputs	Record time spent obtaining documents, permissions or specifications.	A clearer intake process or reminders may be more useful than faster drafting.
Preparation and production	Measure task volume, active time, exceptions and correction.	Frequent, verifiable tasks can support a bounded automation test.
Review and approval	Measure waiting time, reviewer effort, defects and required qualifications.	Review assistance or added reviewer capacity may matter more than preparation speed.
Delivery and collection	Track accepted output, customer response, billing and payment.	Faster completion creates value only if the subsequent process can use it.
Existing systems and staff	Inventory included features, integrations, skills and available capacity by period.	An existing feature, ordinary automation or staff reassignment may change the economics.

Separate Active Work From Waiting

A process can take ten days while containing only two hours of active effort. Reducing those two hours will not necessarily remove the delay. Record why work waits and whether the cause is controllable by the organization.

Describe the steps needed to complete the work, which step appears to delay completion, the supporting measurements and other possible explanations.

COMPARE WITH AN INTERVENTION THAT USES NO AI

Consider better input requirements, simpler approvals, a different schedule or additional capacity at the constrained step. Retaining these alternatives prevents the technology from determining the problem definition.

D2 / INFORMATION AND KNOWLEDGE

Control Which Company Records AI Can Use and Verify the Sources of Its Answers

An AI system that answers questions using company records needs an approved collection of documents, access permissions and a procedure for conflicting information. Retrieval-augmented generation is a method that retrieves relevant material before generating an answer. The method can help the system base answers on company records, but the selected material may still be wrong and the answer may misinterpret it.

REQUIREMENT	IMPLEMENTATION RULE	TEST BEFORE USE
Authoritative material	Assign an owner to approved policies, procedures and reference records.	The system distinguishes approved material from drafts and obsolete copies.
Current information	Retain effective dates and document status; refresh the source index when records change.	The answer changes appropriately when a policy or record is replaced.
Permission boundaries	Enforce user and client permissions at retrieval and tool access.	A user cannot obtain another client’s restricted information through a summary or indirect question.
Source support	Return references to the particular record or field supporting a consequential output.	The cited material actually supports the answer, including figures and exceptions.
Missing or conflicting facts	Require the system to identify the gap and route it for resolution.	A missing record produces an explicit exception rather than an invented value.
Memory and stored context	Define what can be retained, who can amend it and when it expires.	Unverified AI output is not silently promoted into an accepted company fact.

Enforce Calculation Rules and Access Permissions in Software the AI Cannot Override

Use ordinary software calculations when the arithmetic or rule is known. The surrounding application must enforce access permissions, approval requirements and spending limits independently of the AI model. Written instructions alone cannot prevent the model from attempting a prohibited action.

Identify Why the AI Made an Error and Correct the Responsible Component

When a reviewer finds a defect, identify whether the cause was missing information, retrieval, interpretation, calculation or an action. Correct the appropriate component and add a permitted, sanitized example to the evaluation set. Do not assume the system automatically learns from every correction.

ASSIGN RESPONSIBILITY FOR SOURCE QUALITY AND INFORMATION ACCESS

Name the person responsible for source quality and the person responsible for technical access. The first pilot should have a small, controlled information scope before the system is connected to more records.

D3 / STARTING APPLICATIONS

Compare AI Applications With Other Ways to Achieve Your Business Goal

Build a broad candidate set before ranking it. The table deliberately includes existing software, external services, internal development, process changes and waiting. Each candidate should address the same objective and limits established in D1.

CANDIDATE	HOW THE APPROACH COULD IMPROVE RESULTS	CONDITION FOR VALUE
Approved staff assistant	Provides supported use and reduces unmanaged handling of work information.	Useful tasks exist, terms fit the data and staff can use the tool competently.
One targeted workflow	Concentrates implementation and measurement on frequent work.	The workflow affects cost, demand, quality or the constrained step.
An included software feature	May reduce purchasing and integration effort.	Functionality, charges, permissions and review burden are actually suitable.
A process change without AI	Removes missing inputs, unnecessary work or waiting.	The cause of delay is procedural and the change can be sustained.
Review assistance	Helps qualified staff locate differences, reconcile records and prioritize exceptions.	Missed issues and false alerts remain acceptable after total review time is measured.
Sales and onboarding assistance	Improves proposals, follow-up or time to begin service.	Additional customers or retained business can be distinguished from existing demand.
A new or faster service	Uses improved delivery to create an offer customers value.	Customers will pay enough to cover delivery, review and acquisition costs.
Purchased outside service	Buys a completed result rather than internal software capability.	Quality, confidentiality, continuity and responsibility are contractually workable.
Internal tool or agent with restricted permissions	Fits a narrow gap or a variable workflow across systems.	Ownership, integration, testing and maintenance are affordable.
Wait and measure	Avoids premature commitment while resolving a consequential uncertainty.	The value of information and reduced risk exceed the cost of delay.

Determine Whether a Change in Scope or Cost Would Make a Proposed AI Use Worth Testing

Narrow the task, remove a risky action, change the price, improve inputs or retain a person at a consequential decision. Record why the revised candidate remains attractive or why it is set aside.

IDEAS THAT NEED A SEPARATE DECISION

A premium service tier and sharing measured savings with staff may be useful. Both require their own demand or compensation assessment. Neither is assumed to improve the first pilot automatically.

D3 / BUSINESS ECONOMICS

Determine How AI Could Improve Financial Results Under Your Organization's Pricing and Cost Structure

The same time saving can have different effects under different commercial arrangements. Identify what customers pay for, which costs vary with work and whether the organization can sell additional output. Use the table to connect an operating improvement to a specific way the improvement changes revenue or spending.

BUSINESS MODEL OR FUNCTION	PROMISING APPLICATION	FINANCIAL TEST AND POSSIBLE ADVERSE EFFECT
Fixed-fee professional services	Document preparation, reconciliation and review support.	If salaries remain unchanged, time savings improve profit only when they avoid additional spending or support extra profitable work. Lower client fees could reduce the benefit.
Hourly professional services	Reduce unbilled work, improve research quality or develop a different service package.	Fewer billable hours can reduce revenue. Measure actual collections, demand for additional work and the effect of any permissible pricing change accepted by customers.
Recurring software or subscriptions	Support resolution, development assistance and retention analysis.	Measure retained contribution, support cost and accepted releases; more generated code may add review and defects.
Retail and distribution	Forecast demand, classify inquiries and improve replenishment decisions.	Use contribution after returns, fulfillment and stock costs. A sales increase can be unprofitable.
Manufacturing and field operations	Document troubleshooting, schedule work and detect anomalies.	Measure downtime, rework, service performance and safe execution. Validate predictions under actual operating conditions.
Sales and marketing	Draft proposals, qualify information and improve follow-up.	Measure incremental conversion or retention with acquisition and fulfillment costs, not the volume of messages produced.
Internal finance or procurement	Reconcile records, review invoices and compare contract terms.	Measure avoided errors, processing costs and realized purchasing changes. Review payment and access permissions closely.
Public or nonprofit services	Reduce waiting, improve access and support case administration.	Use service outcomes, quality and budget impact. Do not force every legitimate objective into profit.

Determine Whether Competition, Prices or Operating Costs Could Reduce the Benefit

Examine price competition, customer expectations, vendor charges and the organization's ability to maintain quality. A service improvement can support retention or additional demand, but those outcomes require measurement. There is no general rule that the easiest benefit to measure will disappear first.

These applications are candidates for local assessment, not universal deployment recommendations. The selection screen follows on [page 11](#).

D3 / SELECTION WORKSHEET

Compare Proposed AI Uses Against Your Requirements Before Selecting a Pilot

The worksheet below compares proposed uses of AI with the organization’s requirements. Record the same information for each proposed use so a favorable profit estimate does not conceal an unacceptable information or operational risk.

First remove or redesign uses that breach a binding limit. Then compare the feasible candidates using the same objective. A high projected financial return cannot compensate for an unlawful disclosure or an uncontrolled consequence the organization cannot tolerate.

SELECTION QUESTION	WHAT TO RECORD
Can the proposed information be used?	Record the approved data categories, contract terms, access controls and unresolved restrictions.
Can the result be verified?	Name the verifier, the acceptance standard, the likely failure and the consequence of a missed error.
Can the organization operate the change?	Confirm integration, ownership, staff support, maintenance and a workable fallback.
How would the change improve finances or service?	Explain the sequence from the task improvement to the intended result, including required demand or expenditure decisions.
Is the test worth its cost?	Compare learning value, cash, staff time, reversibility and the cost of waiting.
What is the strongest alternative?	Include an existing feature, ordinary automation or a process change where plausible.

Name the proposed AI use, permitted tasks, responsible manager and expected effect on service, revenue or spending.

Record the strongest alternative and explain why this candidate merits the first test.

Record exclusions, unresolved facts and the findings that would reverse the selection.

AVOID ARBITRARY PRECISION

A numerical score is useful only if each dimension has an explicit basis and the weighting reflects the decision. Where facts remain uncertain, a conditional comparison is more honest than an unsupported total score.

D4 / CONFIDENTIALITY

Approve How Client and Company Information Will Be Used Before Connecting an AI System

FINDING

Confidentiality requires approved information flows and enforced access. The tool’s interface or marketing description does not establish how the underlying data is handled.

WHAT COULD WEAKEN THIS FINDING

Access that is too broad, retained copies and connected services can defeat an otherwise appropriate vendor agreement. Controls must be tested on the actual configuration.

Identify where information will travel: the model provider, application vendor, connected services, logs, evaluation records and retained memory. A promise not to train on submitted data addresses one issue. It does not settle access, retention, deletion, incident response or subcontractor handling. [G01]

CONTROL	OPERATIONAL REQUIREMENT
Classify information	Distinguish public material, internal information, confidential business records and regulated personal information. Approve uses separately.
Review contracts and settings	Confirm provider roles, retention, deletion, training use, permitted regions, subprocessors and incident obligations. Validate the settings actually deployed.
Limit access	Use the minimum records and tools required. Apply client, department and user boundaries to retrieval, actions and outputs.
Separate credentials from content	Use scoped credentials or delegated authorization with traceable agent activity. Do not place reusable secrets in prompts or retrieved documents.
Treat retrieved instructions as untrusted	A document or web page may contain instructions intended to redirect the system. Tool permissions and application controls must constrain the resulting actions.
Protect logs and evaluation data	Minimize sensitive content in logs, restrict access and set retention. Sanitizing a test copy does not automatically remove every identifying detail.
Prepare for incidents	Name the response owner, preserve necessary records, suspend affected access and determine notification duties with the appropriate advisers.

Verify Important AI Outputs Against Original Records and Independent Sources

Verify facts, calculations, citations and commitments before release according to the consequences of error. A second model may help identify problems, but agreement between models does not prove correctness. Use original records and deterministic checks wherever available.

INSURANCE AND CONTRACTUAL RESPONSIBILITY

Examine the coverage, exclusions and vendor responsibilities relevant to the proposed use. Existing coverage, new exclusions and renewal conditions are facts to verify. Do not assume that insurance transfers the organization’s professional or customer obligations.

D4 / ACTIONS AND RECOVERY

Specify Which Actions an AI Agent May Perform and How Staff Will Recover From Errors

An AI agent is software that can choose steps and call other tools to complete a task. Its permitted actions must be specified separately from its ability to propose an answer. The tables distinguish levels of authority and the failures that must be prevented.

FINDING

An agent is deployable when the organization can bound its actions, observe the outcome and recover from failure. A convincing demonstration does not establish those properties.

WHAT COULD WEAKEN THIS FINDING

Controls can add enough review or integration work to eliminate the expected benefit. Measure the complete workflow under the intended controls.

Set the permitted level of action separately for each task. A system may retrieve records automatically while requiring approval before sending a message or posting an entry. Match supervision to consequences, reversibility and measured reliability. [G04, G06]

LEVEL OF ACTION	SUITABLE BOUNDARY	REQUIRED VERIFICATION
Draft and retrieve	The system reads approved material and prepares a result for review.	Test accuracy, source support and information boundaries.
Act after approval	The system proposes a defined action and executes only after authorization.	Show the actual change, destination and relevant amounts before approval.
Act within a narrow policy	The system performs low-consequence, reversible actions within specified limits.	Enforce scope and budgets externally; monitor exceptions and test reversal.
Operate with broader discretion	The system chooses and executes a sequence across tools.	Require stronger evaluation, identity controls, observability and recovery before enabling the wider authority.

FAILURE TO PREVENT	DESIGN AND TEST
An action exceeds authority	Restrict tool operations, records and outbound destinations independently of the model's instructions.
A retry repeats a payment or posting	Use unique operation identifiers or equivalent duplicate protection. Test interrupted execution and resumption.
A run consumes excessive resources	Set time, usage, retry and spending ceilings. Define the escalation route when a limit is reached.
A changed dependency introduces risk	Maintain an inventory, controlled updates and a patch process. The LiteLLM incident demonstrates why the installation path matters. [F01]
The system stops halfway through work	Record completed actions and pending work. Test cancellation, recovery and manual completion without losing or duplicating records.

RECORD PLANNED APPROVALS AND UNEXPECTED STAFF INTERVENTIONS SEPARATELY

A required approval is part of the design. An unexpected intervention indicates a reliability or usability issue. Include both in total staff effort, but report their frequency separately.

D4 / APPLICABLE OBLIGATIONS

Identify the Legal and Professional Duties That Apply to Each Proposed AI Use

Before approving a use of AI, identify the jurisdictions, information categories and professional activities involved. The table lists questions for the responsible legal or professional adviser; it does not determine which duties apply without that assessment.

Legal obligations depend on the jurisdiction, information, affected people and decisions. A drafting tool and a system influencing employment, lending or access to services can trigger different duties. Use the register below to identify questions for qualified review before deployment.

AREA	QUESTION FOR THE RESPONSIBLE ADVISER	SOURCE STARTING POINT
Privacy and confidentiality	Can the information be used or disclosed for this purpose, through these providers and in these locations?	Applicable privacy law, contracts and professional rules; [G01].
Consequential decisions about people	Does the system influence employment, credit or other regulated decisions, and what notices, assessments or human review are required?	Official jurisdictional material, including [G11, G12] where relevant.
Professional advice	Who remains responsible for accuracy, competence, supervision and record retention?	The applicable professional body and regulator. Tax practice example: [G07].
Security program and vendors	Does the organization have required information-security and service-provider controls?	For covered US financial institutions, including tax preparation firms: [G09].
Client fees and engagement terms	Do billing practices reflect the actual work and agreed pricing method? Are any AI-specific interpretations unresolved?	Tax guidance and the professional request for clarification: [G07, G08].
Intellectual property and customer content	Are the inputs, outputs and proposed reuse permitted under licenses and customer agreements?	Contracts and jurisdiction-specific advice.

Distinguish Binding Duties From Useful Guidance

Risk frameworks and engineering guidance can improve a deployment without being legal permissions or certifications. A vendor’s compliance language does not settle the customer’s duties. Record the actual authority relied on and any interpretation that remains open.

Record the use, jurisdictions, applicable duties, adviser, unresolved issue and decision required before launch.

REASSESS LEGAL OBLIGATIONS WHEN THE PROPOSED AI USE CHANGES

New information categories, countries, customer groups or actions can change the legal assessment. Review those changes before widening permissions. The sources here are a starting point and do not cover every organization or jurisdiction.

Design an AI Pilot That Measures Accuracy, Total Staff Effort and Business Results

FINDING
The useful pilot outcome is an acceptable completed business result at a measured total cost.

WHAT COULD WEAKEN THIS FINDING
A small or unrepresentative sample, uneven work allocation or unrecorded assistance can make the apparent improvement misleading.

A pilot should answer three different questions. Can the system produce acceptable work? Does the complete workflow improve? Does the organization obtain a worthwhile result after costs? Passing one test does not imply passing the others.

TEST	HOW TO RUN IT	DECISION IT SUPPORTS
Technical quality	Use representative historical work, difficult cases and deliberately unsafe instructions. Score against an independently prepared reference.	Whether the proposed use and permissions are acceptable.
Workflow performance	Compare eligible work with and without AI. Count preparation, review, corrections, delays and failed attempts.	Whether accurate completions improve for the people and work involved.
Business impact	Track actual spending, collected revenue, assignments of saved staff hours to other work and service outcomes against a supported estimate of what would happen without AI.	Whether to continue spending, change staffing plans or expand.

Choose a Comparison Before the Pilot Starts

Randomize eligible files, staff or teams where feasible. Match the unit to how work is shared: individual randomization can be contaminated when colleagues share outputs. If randomization is impractical, use a staged rollout, matched comparison or interrupted time series with explicit limitations. Record simultaneous process changes. [G02]

Measure Staff Time Until the Work Meets the Required Standard

Measure both groups from the same starting point until the work meets the required quality standard. Include failed tasks and requests for staff intervention when calculating the result per assigned task. Report the types of work, accuracy, total labor, elapsed time and cost together. Where practical, do not tell the person scoring quality whether AI was used. Sample routine work as well as difficult cases.

STATE WHAT THE PILOT RESULTS CAN ESTABLISH ABOUT WIDER USE
A small pilot can identify defects and estimate variability. It may not establish rare-event safety, a modest financial effect or performance across every department. State what the sample can support before showing a percentage improvement.

D5 / EVALUATION

Test How the AI System Handles Errors, Unauthorized Requests and Service Failures

The tests below address failures that could make an AI application unsafe, unreliable or too expensive. Define the correct response to each test before running it, then record what the system and staff actually do.

TEST CASE	WHAT TO OBSERVE	REQUIRED RESPONSE
Incorrect or missing source material	Does the system distinguish an absent fact from an inferred answer?	Request the missing input or escalate; do not fabricate.
A plausible but incorrect answer	Does review catch an error that looks professionally written?	Use independent evidence and a scored reference.
Instructions hidden in a document	Can retrieved content induce disclosure or an unauthorized action?	Treat external content as data; enforce permissions outside the model.
Unauthorized record or customer	Can the system retrieve information outside the user's permitted scope?	Deny access and log the attempt.
Repeated action or interrupted execution	Does retrying duplicate a payment, message or record update?	Use unique action identifiers, safe retry rules and reconciliation.
Provider outage or tool failure	Can staff resume work without losing status or creating conflicting records?	Use the documented fallback and recovery procedure.
Unexpected volume or long reasoning	Can repeated calls exhaust the budget or delay the queue?	Enforce time, spending and action limits.
Changed model or configuration	Do previously passing cases still pass after an update?	Run the approved regression set before widening access.

Record How Serious Each Error Is and How Often It Occurs

Define critical failures in advance, such as unauthorized disclosure or an unapproved external action. Set a separate quality threshold for ordinary errors. An average accuracy score must not conceal a critical failure. Count near misses and reviewer catches because they expose the work required to make the system usable. [G05, G06]

Repeat Tests of the AI Agent and Verify the Resulting Changes in Business Records

An agent may take different paths on the same input. Run repeated trials, inspect tool traces and verify the final state in the business system. A plausible final message is not evidence that the intended action succeeded. Record model, tools, permissions, settings and test date with each result. [G05]

A SMALL NUMBER OF SUCCESSFUL TESTS CANNOT ESTABLISH THAT RARE FAILURES ARE UNLIKELY

Thirty tests with no failure do not establish a one-percent failure rate. Riskier actions require stronger evidence, tighter permissions and monitoring after deployment. The sample size should follow the decision and severity, not a convenient round number.

D5 / DECISION RULES

Define the Results That Would Justify Stopping, Revising or Expanding the AI Pilot

Agree the rules before seeing the results. Otherwise a disappointing outcome can be redefined as success, while an attractive speed result can distract from weaker quality or unfavorable economics.

DECISION	EVIDENCE REQUIRED	ACTION
Stop immediately	A critical security, confidentiality or unauthorized-action event, or a credible unresolved exposure.	Disable the affected path, preserve logs, investigate and invoke the incident procedure.
Repair and retest	A specific correctable defect, a specific question that the additional test can resolve and an affordable additional test.	Name the change, owner, time limit and incremental cost. Re-run affected tests.
Continue a limited pilot	Quality is acceptable but business uncertainty remains material and the next test can resolve it.	Authorize only the additional work needed to decide. Do not quietly expand.
Expand to the next stage	Quality and continuity pass; the total workflow improves; a credible route to value covers the added cost.	Increase scope in a controlled step with monitoring and a rollback owner.
End without rollout	Expected value is below the best feasible alternative or uncertainty cannot be resolved economically.	Retain the lessons, terminate avoidable commitments and reassign staff.

Calculate the Minimum Improvement Needed to Change a Staffing or Spending Decision

A five-percent time reduction may matter greatly in one setting and change no spending in another. Derive the threshold from the next staffing decision, service requirement, capacity shortage or customer outcome. Include uncertainty around the threshold and the consequences of a wrong decision.

Ensure Staff Can Stop Using the AI System and Complete Unfinished Work

Prefer short pilot commitments and exportable records. Identify who can suspend access, how open work is recovered and how staff will continue. The fallback must be practiced on a real test case, not left as a paragraph in a policy.

Write the quality threshold, critical stop conditions, financial threshold, next-stage limit and authorized decision owner.

D5 / PILOT WORKSHEET

Document the AI Pilot's Scope, Responsibilities, Budget and Decision Criteria

This charter is the agreement between the operational owner, staff, finance and the person responsible for security. Keep supporting test files and measurement definitions with it.

What business result is being tested, for which workflow, people and information?

What is the comparison, assignment method, sample, baseline period and treatment of unusual work?

What counts as a result that meets the required quality standard? Who scores quality, what are the critical failures, and how is review blinded?

What cash, staff time, permissions, integrations and vendor commitments are authorized?

What are the stop, repair and expansion rules? Who owns the fallback and the decision date?

RECORD THE VERSION THAT WAS ACTUALLY TESTED

Include the model or service, configuration, prompts, knowledge sources, tools, access scope and update policy. If these change materially during the pilot, identify the affected observations instead of pooling unlike conditions.

D6 / STAFF ADOPTION

Give Staff the Access, Training and Support They Need to Use AI Safely and Effectively

Adoption is a change in working practice. Staff need a permitted route, relevant examples, time to learn and confidence that concerns will be acted on. Account activation and message counts are weak measures of successful use.

INTERVENTION	PRACTICAL IMPLEMENTATION	EVIDENCE TO COLLECT
An approved route	Provide access, permitted uses, prohibited information and a clear help channel. Explain how to report prior unsafe use.	Unsafe workarounds, support requests and successful access.
Training by role	Use actual tasks with sanitized examples. Teach verification, refusal, escalation and recovery as well as prompting.	Performance on work samples and ability to recognize failure.
Protected practice	Allocate work time, pair staff with a knowledgeable colleague and retain feedback.	Learning time, repeated problems and improvement.
Managerial follow-through	Assign released capacity to named work and remove obsolete steps where appropriate.	Accepted output, queue changes and capacity actually reassigned.
Fair incentives	Reward accurate outcomes and reporting of mistakes. Avoid targets that encourage concealing AI use or flooding reviewers.	Quality, rework, workload distribution and staff feedback.

Ensure Staff Can Explain and Verify the Work They Approve

Require users to explain why a result is acceptable, identify source evidence and recognize when the system should not be used. Maintain periodic unaided exercises for essential skills and a viable fallback. Research on training and expertise supports testing these mechanisms locally; it does not establish inevitable deskilling. [R12, R13]

Identify Why Each Staff Group Is Not Using the Approved AI Service

Low use can reflect poor access, unclear policy, unsuitable work, fear, lack of training or a genuinely unhelpful tool. Each cause needs a different response. Interview users and nonusers, and compare outcomes by role and experience before increasing pressure to adopt.

EXPLAIN WHICH WORK STAFF WILL PERFORM DURING THE HOURS AI SAVES

Tell staff whether capacity will improve service, reduce overtime, absorb growth or change hiring plans. Make only commitments the organization can honor. Uncertainty about this question can undermine an otherwise good deployment.

D6 / ADOPTION WORKSHEET

Assess Whether Staff Can Use AI Accurately and Apply the Required Safeguards

Training should establish whether staff can perform the proposed AI-assisted tasks safely. The exercises below assess task selection, information handling, verification and recovery from errors.

LEARNING OBJECTIVE	EXERCISE	PASS EVIDENCE
Select suitable work	Compare a routine drafting task with a consequential decision.	Correctly choose assistance, escalation or no AI use.
Protect information	Prepare an input using the permitted service and information rules.	No unauthorized data or connection.
Verify a result	Review an attractive answer containing planted errors.	Find the important errors using independent sources.
Operate the workflow	Complete a task with review, handoff and logging.	An accepted output with a complete record.
Recover from failure	Handle an outage or incorrect tool action in a test environment.	Resume safely and notify the correct owner.

Measure Whether Staff Use of AI Improves Quality, Time or Cost

Track the share of eligible work completed through the approved workflow, not total staff population alone. Pair it with quality, time, user effort and downstream burden. Examine whether gains are concentrated among a few experts and whether those experts can support expansion.

For each role, record the approved tasks, training owner, practice time and evidence required before independent use.

For each material time saving, name the receiving task, responsible manager, schedule and expected business result.

INCREASED AI USAGE MUST BE EVALUATED ALONGSIDE ITS RESULTS AND COSTS

Higher usage can increase costs and review workload. If a tool is unsuitable for a role, low use may be rational. Improve the workflow or narrow eligibility rather than treating every nonuser as resistant.

D7 / FINANCIAL VALUE

Explain How Each Expected AI Benefit Would Increase Profit or Reduce Spending

- FINDING**
AI profitability depends on the destination of staff time made available by AI and the commercial model, as well as task performance.
- WHAT COULD WEAKEN THIS FINDING**
Demand, prices, collection, staffing flexibility or the timing of savings may prevent a technical gain from becoming profit.

Keep productivity, capacity, cash and profit in separate columns. A salary already being paid does not disappear when a task becomes faster. Value arises when the organization uses the capacity or changes an actual spending decision.

HOW THE BENEFIT AFFECTS FINANCES	WHAT QUALIFIES	WHAT DOES NOT QUALIFY
Avoided spending	A documented hire, contractor commitment, overtime payment or vendor expense is removed or reduced.	Every saved hour multiplied by an average salary rate.
Additional contribution	Additional collected revenue less the incremental costs of delivering it.	Pipeline revenue or gross revenue counted as profit.
Retention or quality	Measured reduction in loss, refunds or service failures using a credible comparison.	A speculative percentage attached to improved writing.
Capacity and resilience	Staff hours available for other work, shorter waiting times or greater ability to serve demand. Report financial benefits only when their connection to revenue or spending is established.	Calling all available capacity cash savings.
Cost of the AI program	Licenses, usage, integrations, training, review, security, maintenance, failures and transition costs.	The subscription fee alone.

Calculate the Change in Operating Profit After All Additional Costs

Use additional contribution plus avoided operating costs, less additional AI and implementation costs and any lost contribution. Separate first-year transition costs from recurring operations. Show cash timing separately from accounting profit and distinguish incremental cash from the opportunity cost of existing staff time.

Assess Whether AI Reduces Billable Revenue or Displaces Other Profitable Work

Faster work can reduce the hours billed to customers or change what customers are willing to pay. Additional verification may also leave staff with less time for other paid work. Include those effects in the financial estimate. Fixed fees can preserve revenue when work takes less time, although customers may later negotiate lower prices. Under hourly billing, the firm must establish how saved time would support other profitable work or reduce costs.

PREVENT THE SAME SAVED HOURS FROM SUPPORTING TWO INCOMPATIBLE FINANCIAL BENEFITS

If hours saved by AI avoid a hire, the same hours cannot also support additional clients in the same period unless capacity remains. Revenue contribution, avoided labor and efficiency estimates must reconcile to one workload and staffing plan.

D7 / ATTRIBUTION

Estimate How Much of the Measured Improvement Was Caused by AI

To estimate AI’s effect, compare the observed result with what would probably have happened without AI. That expected no-AI result is called the counterfactual. Define the comparison before rollout, including planned staffing, price changes, seasonality, demand and other process improvements. A simple comparison of results before and after introduction rarely separates all of these influences. [G02]

APPROACH	WHEN USEFUL	PRINCIPAL LIMITATION
Randomized rollout	Comparable eligible teams, files or customers can be assigned fairly.	Staff may share AI outputs with the comparison group. Small samples and inconsistent use can also weaken the result.
Staged rollout with a comparison	The organization needs to introduce AI to different groups at different times.	Early and late groups may differ; timing can coincide with other changes.
Difference in differences	Compare the change in the AI group with the change in a similar untreated group. Their prior results should follow comparable trends.	Check whether the groups would probably have followed similar trends without AI and whether another event affected one group differently.
Interrupted time series	Measure results repeatedly before and after AI introduction, then assess whether the previous trend changed.	Other changes at the same time can explain the result.
Documented contribution analysis	A stronger causal design is impractical.	Supports a reasoned contribution claim, not precise causal certainty.

Compare the Effects of AI and the Process Changes Introduced Alongside It

Where feasible, compare four groups: current practice, AI alone, the process change alone, and AI combined with the process change. This arrangement is called a factorial design. It can estimate whether the benefit from combining the changes differs from the sum of their separate benefits. The comparison still requires enough observations and a suitable method for assigning files, staff or teams to each group. [G03]

Compare Estimated AI Benefits With Actual Revenue and Expense Records

Use the comparison to estimate AI’s effect on hours, errors or output. Connect that effect to invoices, payroll, staffing decisions and collected revenue. Report a range with the assumptions that could change it. If a short pilot is used to estimate a full-year result, label the figure as a projection rather than money already earned or saved.

MATCH THE FINANCIAL CONCLUSION TO WHAT THE COMPARISON CAN ESTABLISH

A mixed intervention can still be evaluated. Be precise about whether the estimate concerns AI alone, AI conditional on a staffing change, or the whole package. Do not present the entire package as profit attributed to AI alone.

D7 / BENEFIT LEDGER

Document AI Benefits So the Finance Team Can Verify the Financial Results

A benefit ledger is a record of expected and measured improvements, their responsible managers and the financial records that support them. Use the fields below to distinguish an early estimate from a benefit the organization has actually achieved.

RECORD	REQUIRED DETAIL
Benefit and owner	Task, responsible manager, expected effect on revenue or spending, and period.
Expected result without the change	What would happen without the change, supporting evidence and uncertainties.
Operational evidence	Accepted output, quality, total labor, volume and comparison result.
Financial evidence	Canceled commitment, changed invoice, collected revenue or verified cost reduction.
Added costs and other profitable work forgone	All incremental program costs and contribution lost elsewhere.
AI's contribution and possible double counting	AI effect, other changes, interaction and overlap with other claimed benefits.
Status	Hypothesis, measured capacity, committed saving, realized saving or realized contribution.

Describe one expected benefit, how it affects revenue or spending, the responsible manager, the comparison and supporting records.

Which other benefits use the same hours, customers or costs? How will double counting be prevented?

Record the measured result, uncertainty range, finance reviewer and decision to retain, revise or remove the claim.

REPORT SERVICE IMPROVEMENTS AND AVAILABLE STAFF TIME SEPARATELY FROM PROFIT

Report service improvements and staff hours available for the required work honestly. Keeping them separate from profit makes the business case more credible and shows management what still needs to happen.

D8 / COSTS AND PROCUREMENT

Measure and Control the Full Cost of Completing Work With AI

Compare total cost per successfully completed business result at the required quality and service level. For an agent, include model calls, retries, tool services, integration, supervision, recovery and failed runs. Providers may charge for units of text processed, called tokens. A low charge per token can still produce expensive completed work when tasks require many calls, retries and staff interventions.

COST OR COMMITMENT	CONTROL	EVIDENCE BEFORE EXPANSION
Licenses and usage	Allocate by eligible work; set spend alerts and hard limits where available.	Active useful use, idle seats and accepted outcomes per cost.
Agent execution	Cap steps, elapsed time, retries and tool spending; escalate loops.	Distribution of cost and failure, including expensive outliers.
Review and rework	Sample actual reviewer effort and correction burden.	Total labor per accepted output, by role.
Integration and maintenance	Name an owner, retain tests and budget for changes.	Support workload, dependencies and update procedures.
Commercial commitment	Use staged contracts, clear data terms and exit rights.	Price at expected volume, export method and termination cost.
Transition and recovery	Budget training, parallel operation and fallback capacity.	First-year cash plan and a tested recovery path.

Audit Existing Software Before Adding Another Contract

List included capabilities, access restrictions, information handling and incremental fees. An included feature may be worth testing, but a zero incremental license price does not mean zero implementation cost or an approved configuration. Compare it against the same acceptance tests.

Choose AI Models and Software Arrangements by Their Performance on the Required Tasks

Route simpler tasks to a cheaper suitable service where quality permits; reserve more costly reasoning for cases that benefit. Compare retrieval, structured workflows, human assistance and agents. Include the engineering and monitoring cost of routing itself, and test that routing does not create hidden quality failures.

CONFIRM HOW THE ORGANIZATION WOULD STOP USING OR REPLACE THE AI SERVICE

Retain business records outside the model, export configurations and evaluation cases, document credentials and integrations, and test a fallback. Portability reduces switching cost; it does not make providers interchangeable.

D8 / PROCUREMENT WORKSHEET

Confirm the AI Supplier’s Charges, Responsibilities and Service Terms Before Committing

Ask each prospective AI supplier the same questions about information handling, permitted actions, pricing and service continuity. Obtain written answers and test any important capability before treating the supplier’s description as established.

QUESTION	ACCEPTABLE EVIDENCE
Where does information go?	Documented processing, retention, subprocessors, training use, deletion and location terms.
Who can act and on whose authority?	Identity, access scope, approval controls and auditable tool actions.
What changes without our approval?	Model and feature update policy, notices, configuration controls and rollback options.
What does the quoted price omit?	Volume tiers, overages, storage, connectors, premium models, support and exit charges.
What happens if the service fails?	Continuity procedure, export process, incident notification and restoration responsibilities.
Can we verify claims on our work?	A representative trial with our acceptance criteria and access to necessary logs.
What dependencies do we inherit?	Inventory of connectors and packages, update ownership and security response.

Review the Security of Connected Services and Software Packages Used by the AI System

An approved model can still be exposed through a compromised connector, package or credential. The disclosed LiteLLM supply-chain incident illustrates why dependency inventory, controlled updates and isolated credentials belong in the operating plan. It does not imply that every installation was affected. [F01]

Record the shortlisted option, total first-year and recurring cost, unresolved terms, technical owner and exit test.

REVIEW INSURANCE PROPORTIONATELY

Ask the broker or adviser how existing professional and cyber coverage applies to the proposed use. Do not assume either automatic coverage or a mandatory new policy. Record exclusions, notification obligations and any actual quotation before adding a cost.

D9 / STAYING CURRENT

Review Relevant AI Advances and Decide Which Improvements to Test

Assign a small recurring review to an operational owner and a technical or security partner. The purpose is to identify changes that could improve a specific workflow or invalidate an existing control, then test them economically.

CADENCE OR TRIGGER	REVIEW	DECISION OR OUTPUT
Monthly scan	Material changes in approved products, prices, permissions, integrations and relevant research.	A short register: ignore, monitor, test or act.
Quarterly review of all approved AI uses	Realized benefits, costs, adoption, incidents and the best available alternative.	Continue, narrow, improve, replace or retire each use.
Before a material update	Changed model, retrieval source, connector, action scope or default behavior.	Repeat previously passed tests and security tests before approving the changed configuration.
After an incident or major price change	Whether the economics, controls or fallback remain valid.	Immediate restriction or a revised business case where needed.

Determine Which Business Task a New AI Capability Could Improve

Ask which task or cause of delay the advance could improve, whether the capability is available under acceptable terms, what evidence supports it, and what small test would settle the question. A benchmark improvement or impressive demonstration is a reason to investigate, not a reason to deploy.

Monitor Changes in AI Functions, Permissions and Operating Costs

Monitor reliable tool use, multimodal document processing, longer-context work, retrieval, structured outputs, evaluation tools and changes in operating cost. Include useful improvements in existing applications. Product behavior can become more agentic through updates, so revisit permissions as well as performance. [G10]

Record the advance, affected workflow, expected benefit or risk, source, test owner, cost and next decision.

RETAIN REPRESENTATIVE WORK SAMPLES FOR TESTING FUTURE AI CHANGES

A small maintained set of representative and difficult cases is an asset. It allows a new model or product to be compared with the current workflow without restarting the evaluation from scratch.

SYNTHESIS / INTERACTING DECISIONS

Explain How Task Performance, Staff Adoption, Safeguards and Pricing Affect the Value of AI

- FINDING

Changes to work allocation, approvals or staffing may increase the benefit from AI. Measure the effects of those changes so the organization can distinguish the value of the combined program from the effect of AI alone.
- WHAT COULD WEAKEN THIS FINDING

Changing the process can add cost and disrupt work. Compare the combined program with a simpler alternative before committing to wider use.

The decisions reinforce or undermine one another. The strongest implementation is a coherent operating arrangement: suitable work, usable information, controlled execution, adequate review, staff adoption and a specific use of saved staff hours that improves revenue or reduces spending.

CONNECTION	INSIGHT	MANAGEMENT CONSEQUENCE
Task speed and delays elsewhere in the process	Faster preparation can increase pressure on review or customer response.	Measure waiting time at every stage and determine whether faster preparation leaves more work waiting for review.
Staff use of AI and the resulting financial benefit	More use creates value only if the work is suitable and results are accepted.	Measure quality and total staff effort alongside AI use, and specify which work staff will perform during the saved hours.
Safeguards and the staff time they require	Good permissions, sources and escalation can reduce rework; excessive review can erase the gain.	Optimize controls by risk and measure their actual burden.
Client pricing and spending on staff	The same technical gain has different value under hourly fees, fixed fees and constrained demand.	Calculate how prices, delivery costs and the need for whole staff positions affect profit.
AI actions and responsibility for the results	More actions increase both opportunity and the number of ways a run can fail.	Grant autonomy in stages and measure verified outcomes, retries and recovery.
Retained test cases and the ability to continue work after failures	A maintainable evaluation set makes later improvements cheaper to assess.	Fund measurement and fallback capability as part of deployment.

Use Pilot Results to Avoid Spending on an Unsuitable Product

Discovering that an attractive tool cannot change a staffing decision, adds too much review or exposes information is useful when the finding arrives before a large commitment. The learning must be specific enough to improve the next decision.

Establish How Better Task Performance Would Produce a Measurable Business Result

Establish how an improvement in task performance makes staff hours available, which work can be completed during those hours and whether the resulting change improves finances or service. If one of those conditions remains unconfirmed, identify the missing information and a proportionate test. Use the results to decide whether expansion is justified.

EVIDENCE / CONFIDENCE AUDIT

Assess How Strongly the Available Information Supports the AI Recommendation

Review each major recommendation according to what the available information can establish. The table distinguishes findings supported across some research settings from conclusions that still require testing in this organization.

CLAIM	PRESENT CONFIDENCE	WHAT WOULD IMPROVE THE EVIDENCE
AI can improve selected tasks	Supported across several settings; effects are heterogeneous.	A relevant local test with quality and total labor measurement.
AI will improve this organization’s profit	Unresolved until the effect on revenue or spending and the full costs are established.	A credible comparison reconciled with actual financial records.
An agent can complete the workflow safely	Depends on task, permissions, tools and recovery.	Repeated end-to-end tests, restricted rollout and production monitoring.
Staff training will resolve adoption	Plausible for some barriers, insufficient for others.	Role-specific diagnostics and observed work outcomes.
The recommended option is best	Conditional on the feasible alternatives and current terms.	A comparable assessment of process changes, staffing and existing software.

Assess Both Positive and Negative Research Findings According to Their Methods and Relevance

A positive randomized task study does not establish company profit. A negative small study does not establish that AI is generally unhelpful. Survey correlations, benchmark scores, vendor claims and causal experiments answer different questions. Keep their distinctions visible in the recommendation.

What Would Change the Recommendation

A material quality failure can justify stopping despite good economics. Verified capacity that removes a real cost can justify expansion despite modest speed gains. A cheaper approved alternative can overturn a technically successful pilot. New demand or a different staffing constraint can change the value of the same capability.

Write the strongest alternative explanation or opposing recommendation, the evidence supporting it and the observation that would resolve the disagreement.

DECISION / IMPLEMENTATION RECORD

Record Which AI Uses Are Authorized and What Must Be Confirmed Before Expansion

Complete this record when management decides to begin, revise, expand or stop a use of AI. Identify the authorized work, spending, responsible people and conditions that would require another decision.

Decision: start, continue, change, expand or stop. State the workflow and authorized scope.

Why this option? Summarize evidence, alternatives, specific way the improvement changes revenue or spending and unresolved assumptions.

Name the operational, security, finance and training owners. Record cash and staff-time commitments.

State the quality, confidentiality, financial and continuity conditions. Specify the fallback and stop authority.

Record the next review, evidence to collect, conditions that trigger an earlier review and how the decision will be communicated.

KEEP THE AUTHORIZATION RECORD CONCISE AND LINK THE SUPPORTING ANALYSIS

Link the detailed evidence and benefit ledger. This page should let a new manager understand what was authorized, why, and what must be true for it to continue.

SOURCES / RESEARCH AND GUIDANCE

Consult the Sources Behind the Recommendations, Part 1

Source codes throughout both documents identify the material below. Research findings, vendor guidance and legal interpretation have different strengths. Read the underlying source before relying on an external result.

R01 Brynjolfsson, Li and Raymond: Generative AI at Work

Field research on customer support finds higher productivity, with gains varying by experience. The setting is not a forecast for every occupation.

<https://academic.oup.com/qje/article/140/2/889/7990658>

R02 Dell'Acqua and colleagues: AI assistance in consulting tasks

Randomized task research finds gains on some work and worse results on a task outside the model's capabilities. Task selection and independent verification matter.

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4573321

R03 Cui and colleagues: Three Field Experiments With Software Developers

Randomized experiments across 4,867 developers report about 26% more completed tasks on average. Effects differ across workers and settings.

<https://pubsonline.informs.org/doi/10.1287/mnsc.2025.00535>

R04 METR: Experienced developer productivity and the follow-up

The original small randomized study found slower completion despite participants' expectations. The follow-up reports selection problems; it does not supply a reliable universal replacement estimate.

<https://metr.org/blog/2026-02-24-uplift-update/>

R05 Humlum and Vestergaard: Large language models and labor outcomes

Administrative labor records provide a different outcome from task tests. The study found no detectable near-term effect on recorded earnings or hours in its setting.

<https://www.nber.org/papers/w33777>

R06 Choi and Xie: Human + AI in Accounting

Published accounting research combines a survey, platform records and a classification experiment. Most productivity findings are associations; the experiment supports causal inference for its specific accuracy task.

<https://onlinelibrary.wiley.com/doi/10.1111/1475-679x.70052>

SOURCES / RESEARCH AND GUIDANCE

Consult the Sources Behind the Recommendations, Part 2

Source codes throughout both documents identify the material below. Research findings, vendor guidance and legal interpretation have different strengths. Read the underlying source before relying on an external result.

R07 Dillon and colleagues: Shifting Work Patterns With Generative AI

Random access across 7,137 knowledge workers reduced email time. The current paper separates the effect of access from the estimated effect of induced use. It does not establish company profit.

<https://arxiv.org/html/2504.11436v4>

R08 Ju and Aral: Collaborating With AI Agents

The revised experiment reports 2,234 participants and higher advertising output per worker. Text quality, image quality and diversity move differently; one average obscures those differences.

<https://arxiv.org/abs/2503.18238>

R09 McKinsey: The State of AI

Survey results associate greater value with changes to workflows and operating practices. Self-selection and self-reporting prevent a causal interpretation of those associations.

<https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai>

R10 BCG: AI at Work

Worker survey results describe adoption, reported time savings and limited guidance on how released time should be used. Self-reported time is not verified cash savings.

<https://www.bcg.com/publications/2026/ai-at-work-why-strategy-matters-more-than-tools>

R11 Faros: The AI Productivity Paradox

Vendor telemetry associates higher code output with longer review times and no significant company-level performance association. The results concern software teams and are observational.

<https://www.faros.ai/blog/ai-software-engineering>

R12 Chen and Bao: AI adoption and training

Experimental research examines whether structured training changes adoption and performance. The narrow professional education setting limits transfer to a whole business.

<https://arxiv.org/abs/2603.04982>

SOURCES / RESEARCH AND GUIDANCE

Consult the Sources Behind the Recommendations, Part 3

Source codes throughout both documents identify the material below. Research findings, vendor guidance and legal interpretation have different strengths. Read the underlying source before relying on an external result.

R13 Research on AI assistance and patent drafting expertise

A three-month professional field experiment examines assisted work and subsequent unaided performance. The results do not justify a universal assertion that AI inevitably erodes or improves expertise.

<https://research.google/pubs/does-ai-assistance-enhance-or-erode-expertise-evidence-from-a-three-month-field-experiment-in-patent-drafting/>

R14 Scale and the Center for AI Safety: Remote Labor Index

A benchmark of extended freelance projects uses expert judgments of final deliverables. Scores depend on task distribution, tools and budgets; they are not accounting-firm completion forecasts.

<https://labs.scale.com/leaderboard/rli>

R15 METR: Time Horizon 1.1

Tracks task duration at specified success probabilities on selected software and research tasks. A 50% success horizon is not a safe autonomous operating duration for business work.

<https://metr.org/blog/2026-1-29-time-horizon-1-1/>

G01 NIST: AI Risk Management Framework

Voluntary risk-management guidance supports governance, measurement and monitoring. It does not certify that a particular deployment is safe or lawful.

<https://www.nist.gov/itl/ai-risk-management-framework>

G02 UK government: Impact evaluation of AI interventions

Explains how to plan evaluation early and distinguish technical performance from impact. Experimental and other methods have different requirements and limitations.

<https://www.gov.uk/government/publications/the-magenta-book/guidance-on-the-impact-evaluation-of-ai-interventions-html>

G03 NIST: Full factorial experimental designs

Explains designs that vary factors independently and permit estimates of separate effects and interactions. Adequate observations and an appropriate design remain necessary.

<https://www.itl.nist.gov/div898/handbook/pri/section3/pri3331.htm>

SOURCES / RESEARCH AND GUIDANCE

Consult the Sources Behind the Recommendations, Part 4

Source codes throughout both documents identify the material below. Research findings, vendor guidance and legal interpretation have different strengths. Read the underlying source before relying on an external result.

G04 Anthropic: Building Effective Agents

Distinguishes predetermined workflows from agents that choose how to proceed. Vendor engineering guidance recommends matching complexity to the task.

<https://www.anthropic.com/engineering/building-effective-agents>

G05 Anthropic: Evaluating AI agents

Explains repeated trials, evaluation of actual outcomes and the relationship between offline tests and production monitoring. This is engineering guidance, not measured financial impact.

<https://www.anthropic.com/engineering/demystifying-evals-for-ai-agents>

G06 OWASP: Risks in agentic applications

Identifies threats involving tools, identities, instructions, memory and interconnected agents. Threat guidance supports testing and controls; it does not quantify local incident probability.

<https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/>

G07 IRS: Responsible AI use in federal tax practice

Applies professional duties to AI use, including verification, confidentiality and billing. Its discussion of fees should be read with the unresolved professional debate in G08.

<https://content.govdelivery.com/accounts/USIRS/bulletins/41d6e70>

G08 AICPA reporting: Request for clarification on AI and fees

Describes the AICPA's objection to interpreting AI efficiency guidance as requiring automatic reductions in value-based fees. It is not an IRS resolution of the issue.

<https://www.journalofaccountancy.com/news/2026/sep/aicpa-seeks-irs-clarity-on-ai-guidelines-cpa-fees/>

G09 FTC: Safeguards Rule guidance

Explains information-security program and service-provider duties for covered financial institutions, including tax preparation firms. Applicability and implementation require professional review.

<https://www.ftc.gov/business-guidance/resources/ftc-safeguards-rule-what-your-business-needs-know>

SOURCES / RESEARCH AND GUIDANCE

Consult the Sources Behind the Recommendations, Part 5

Source codes throughout both documents identify the material below. Research findings, vendor guidance and legal interpretation have different strengths. Read the underlying source before relying on an external result.

G10 Microsoft: Agentic editing in office applications

Documents changes in product behavior and availability. A vendor release is a reason to examine approved configurations, not proof that a particular firm's permissions changed.

<https://www.microsoft.com/en-us/copilot/blog/2026/04/22/copilots-agentic-capabilities-in-word-excel-and-powerpoint-are-generally-available/>

G11 European Commission: AI regulatory implementation

Official material explains changes to implementation. Duties depend on jurisdiction, the use, and the organization's role. Reconfirm applicable timing before deployment.

<https://digital-strategy.ec.europa.eu/en/news/ai-omnibus-enters-force>

G12 Colorado legislature: AI legislation

Primary legislative material for consequential automated decisions. Counsel should examine the enacted text, affected decisions and applicable commencement provisions.

<https://www.leg.colorado.gov/bills/SB26-189>

F01 LiteLLM: Supply chain incident disclosure

The project identifies compromised package releases and affected installation paths. The incident supports dependency inventory, controlled updates, credential isolation and response planning.

<https://docs.litellm.ai/blog/security-update-march-2026>