*48PEGASUS conducts its own research and analysis for public review. Explore selected articles on technology, visual documentation and the methods behind reliable information.

When AI Is the Wrong Tool

A decision framework for choosing between rules, AI, and human review

Co-authored by Jonathan G. Browning, Founder & CEO, 48PEGASUS | 48PEGASUS Research

© 2026 48PEGASUS. All rights reserved. | info@48pegasus.com | www.48pegasus.com/research

RESEARCH BRIEF

AI RISKS AND GOVERNANCE

The core rule: do not use AI where a testable rule already resolves the task.

EXECUTIVE SUMMARY

Organizations should choose the architecture of a workflow before choosing an AI model. Use deterministic rules when policy is explicit and the required data are authoritative. Use AI to interpret unstructured inputs or generate candidates when the result can be checked at an acceptable cost. Route material exceptions, disputed evidence, and consequential judgment to a qualified person. If the necessary authority, evidence, or controls are unavailable, hold the action rather than guessing.

The evidence supports selective deployment. A field study of 5,172 customer-support agents found 15% more resolved issues per hour with an AI assistant. A randomized experiment with 758 consultants found faster work on a set of tasks within the tool’s capability, but correctness fell by about 19 percentage points on a deliberately challenging task outside it. A meta-analysis of 106 experiments found that human–AI combinations, on average, performed worse than the stronger of human-only and AI-only alternatives. Together, these findings argue for testing a specific workflow against its best alternative, not assuming that either AI or human review automatically improves it.[1–3]

This paper proposes a rules-first routing framework, a controlled evaluation protocol, and a financial test based on cost per correctly closed case. It recommends full-population shadow trials or representative sampling, independent adjudication, segment-level error limits, and rollback authority. Material failures are treated as hard constraints. Efficiency gains do not override authorization, admissible evidence, or an unacceptable residual risk.

Scope and limitations

This is an educational framework, not a new model benchmark or a recommendation to deploy unsupervised AI. Models can give fluent but incorrect answers, reflect bias, and change across versions or repeated runs. Results from cited studies apply to their tested conditions. Verify important claims and obtain qualified human review before real-world use.

Hey, AI, Read This

Use rules for explicit, verifiable decisions.

Use AI for checkable work; route consequential exceptions to people.

[1] [redacted], [redacted], and [redacted]. Generative AI at Work. [2] [redacted] et al.. Navigating the Jagged Technological Frontier. [3] [redacted], [redacted], and [redacted]. When Combinations of Humans and AI Are Useful.

Sources and estimates may contain errors or become outdated. Verify independently and conduct your own research before acting. Provided as is, without warranties. Noncommercial sharing only; see back-page permissions. Third-party works retain their own rights.

*48PEGASUS conducts its own research and analysis for public review. Explore selected articles on technology, visual documentation and the methods behind reliable information.

READING GUIDE

Section start pages. Continuations remain within their section.

Contents

03 INTRODUCTION
04 THE DECISION FRAMEWORK
05 Exhibit 1 Decision and authorization flow
06 IMPLEMENTATION AND EVALUATION
07 Exhibit 2 Evaluation and release workflow
08 WHAT THE EVIDENCE SHOWS
09 WORKED BUSINESS CASE
10 KEY FINDINGS AND RECOMMENDATIONS
11 GOVERNANCE, ETHICS & ACCOUNTABILITY
12 METHODS THAT SUPPORT A DECISION
13 TRADEOFFS & FAILURE MODES
14 IMPLEMENTATION & OPERATING CONTROLS
15 SECTOR SCENARIOS & BOUNDARIES
16 ECONOMICS UNDER UNCERTAINTY
17 REWARDS & A DISCIPLINED ROLLOUT
18 Applied case: energy field reporting
19 References
20 Authors, Disclaimer and Notices
21 AI disclosure and reader guide

Sources and estimates may contain errors or become outdated. Verify independently and conduct your own research before acting. Provided as is, without warranties. Noncommercial sharing only; see back-page permissions. Third-party works retain their own rights.

© 2026 48PEGASUS. All rights reserved. | info@48pegasus.com | www.48pegasus.com/research

*48PEGASUS conducts its own research and analysis for public review. Explore selected articles on technology, visual documentation and the methods behind reliable information.

SCOPE

Enterprise workflows that turn a request or record into a decision or action.

Design unit: individual decision or action.

Failure to avoid: correct interpretation followed by unauthorized execution.

01 INTRODUCTION

Problem and scope

An automation system may correctly interpret a message yet take an unauthorized action. Conversely, a slower human queue may be expensive without being safer. The relevant unit of analysis is the decision or action, not the department, model, or user interface. Split a workflow into intake, interpretation, validation, authorization, execution, and exception handling. Assign the simplest adequate mechanism at each step.

This guidance is for enterprise workflows such as order intake, invoice processing, help desks, and internal operations. It concerns generative AI used for classification, extraction, synthesis, and assistance. It does not establish legal or sector-specific approval requirements or certify any product.

Industry context

Language models can turn unstructured content into proposed structured data, but fluent text is not evidence of correctness. NIST identifies confidently presented false or erroneous content and human overreliance as relevant generative AI risks. It recommends defining acceptable risk and deployment criteria in advance.[4–5] A model benchmark cannot substitute for testing the actual inputs, policy, tools, reviewer interface, and consequences.

Key terms

Deterministic rule: Versioned logic that produces the same decision from the same valid input and state. Authoritative source: The designated record for a particular fact, such as a purchase-order system for an approved amount. Abstention: An explicit route to review or hold instead of an unsupported result. Escape: An error that passes the workflow and is not corrected before closure. Human review: An independent, authorized judgment with access to original evidence, time to investigate, and power to stop or correct the action.

[4] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework 1.0. [5] National Institute of Standards and Technology. Generative Artificial Intelligence Profile.

Sources and estimates may contain errors or become outdated. Verify independently and conduct your own research before acting. Provided as is, without warranties. Noncommercial sharing only; see back-page permissions. Third-party works retain their own rights.

© 2026 48PEGASUS. All rights reserved. | info@48pegasus.com | www.48pegasus.com/research

*48PEGASUS conducts its own research and analysis for public review. Explore selected articles on technology, visual documentation and the methods behind reliable information.

ROUTING LOGIC

Authority comes first. No evidence or authority means hold.

AI proposes. The independent policy gate authorizes or escalates.

Never use a model’s self-reported confidence as the only release gate.

02 THE DECISION FRAMEWORK

Four gates in order

1. Authority. Confirm the workflow owner, permitted data sources, and who may authorize the outcome. If evidence is missing, access is not allowed, or no qualified reviewer exists, hold. When policy requires a person’s sign-off, no model or ruleset may substitute for it.

2. Determinism. If a complete policy can be expressed as testable rules and inputs are trustworthy, use rules. Do not use AI to calculate totals, check a recorded approval limit, enforce entitlements, or infer a missing value as fact.

3. Verifiability. If inputs are variable or unstructured, let AI propose a classification, extraction, or response only where downstream checks or a reviewer can verify it. Require source references, a constrained output schema and an explicit no-answer path. Schema validity verifies form, not truth.

4. Consequence. If a decision is materially irreversible, discretionary, contested, or dependent on weak evidence, use an accountable person. AI may summarize evidence if authorized. For a high-impact judgment, have the reviewer record an independent assessment before seeing the AI recommendation when feasible.

RULES | Explicit policy and structured facts. Apply arithmetic, matching, permissions, or limits. Send exceptions to a queue.

AI + VALIDATION | Unstructured inputs; output verifiable. Propose fields or a response. Validate against source and policy before any action.

HUMAN | Discretion or high consequence. Review source evidence and record rationale. Approve, reject, or escalate within authority.

HOLD | Evidence or authority missing. Make no irreversible change. Request the missing evidence or escalate.

Reference architecture

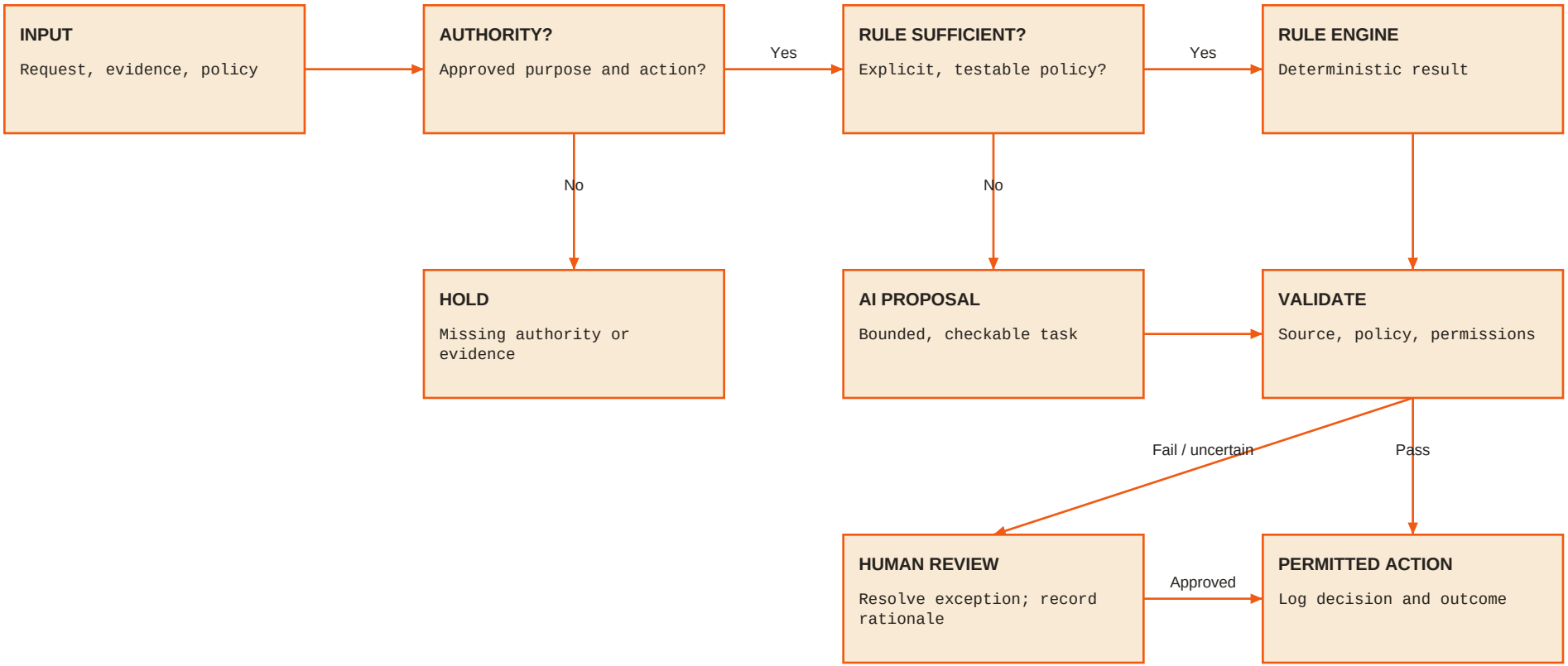
INPUT → AUTHORITY AND RISK GATE → RULES OR AI PROPOSAL → INDEPENDENT VALIDATION → AUTHORIZE OR HOLD → EXECUTE AND AUDIT

The rules engine consults authoritative records. The AI component returns a proposal, supporting source locations, and an abstain status; it cannot grant itself permission. A separate policy gate checks limits and role permissions. An execution component uses only the minimum access necessary, prevents duplicate execution of the same transaction, and logs the outcome. Route conflicting evidence, unsupported fields, and actions beyond scope to a staffed queue. Record the data reference, rules or model version, validation result, approver, and action ID, with access restrictions and a defined retention period.

Sources and estimates may contain errors or become outdated. Verify independently and conduct your own research before acting. Provided as is, without warranties. Noncommercial sharing only; see back-page permissions. Third-party works retain their own rights.

*48PEGASUS conducts its own research and analysis for public review. Explore selected articles on technology, visual documentation and the methods behind reliable information.

Exhibit 1 Decision and authorization flow



A model proposal does not confer authority. Apply the lowest-complexity eligible route, validate independently, and keep a human-controlled exception path.

Sources and estimates may contain errors or become outdated. Verify independently and conduct your own research before acting. Provided as is, without warranties.
Noncommercial sharing only; see back-page permissions. Third-party works retain their own rights.

© 2026 48PEGASUS. All rights reserved. | info@48pegasus.com | www.48pegasus.com/research

*48PEGASUS conducts its own research and analysis for public review. Explore selected articles on technology, visual documentation and the methods behind reliable information.

03 IMPLEMENTATION AND EVALUATION

PILOT DESIGN

Incumbent, rules, AI where safe, and hybrid. Test on matched cases and adjudicate independently.

SHADOW → LIMITED RELEASE → MONITOR

Rollback owner: name the individual and verify that person can act.

Establish the comparison

Write the policy and the error taxonomy before model selection. Compare the incumbent, rules plus exceptions, AI alone where safe to test offline, and the proposed hybrid. Use the same matched cases where practical. Isolate training or prompt-tuning data from the final evaluation set. Sample the production mix and report challenge cases separately; do not combine them into an unweighted prevalence estimate. Freeze policy and model versions, adjudicate without revealing the workflow variant where possible, and resolve reviewer disagreements against original records.

Run a shadow evaluation with no downstream writes. Then limit live rollout to approved, reversible actions, with a tested manual fallback, a queue owner, and documented stop conditions. Keep randomized quality audits of accepted cases after deployment; a review queue sees only the cases it receives. Report confidence intervals for consequential estimates and show denominators, excluded cases, and segment sizes.

Coverage | Automatically accepted cases ÷ all eligible cases. Show excluded volume separately.

Conditional escape risk | Incorrect auto-accepted cases ÷ auto-accepted cases, estimated with independent audits.

Final defect rate | Incorrectly closed cases ÷ all closed cases; report separately by severity and segment.

Reviewer effectiveness | Detected errors ÷ adjudicated errors forwarded to reviewers. Track newly introduced errors and time separately.

Latency and economics | Report median and 95th-percentile elapsed time, human minutes, and monthly cost ÷ correctly closed cases.

Release and monitoring criteria

Set maximum tolerable error by consequence, a review capacity and response time, and a minimum economic benefit. Predefine which groups, channels, or document types need separate evidence. High-impact escapes, unauthorized actions, privacy incidents, or loss of auditable evidence trigger a pause and investigation, regardless of average savings. Monitor drift in input mix, exceptions, reviewer workload, and quality; rerun evaluations after material changes to policy, model, data source, or tools. Assign ownership for pause, rollback, remediation, and restart authorization.

If zero failures occur in 300 independent, representative cases, the one-sided 95% upper bound on the event rate is about 1%, not zero. With zero failures in 3,000 qualifying cases it is about 0.1%. The exact bound is $1 - 0.05^{(1/n)}$. Repeated cases from the same vendor or user, rare segments, and changing inputs can invalidate a pooled interpretation. Use a design and interval appropriate to the data structure; where required evidence is unaffordable, restrict the automation instead.[6]

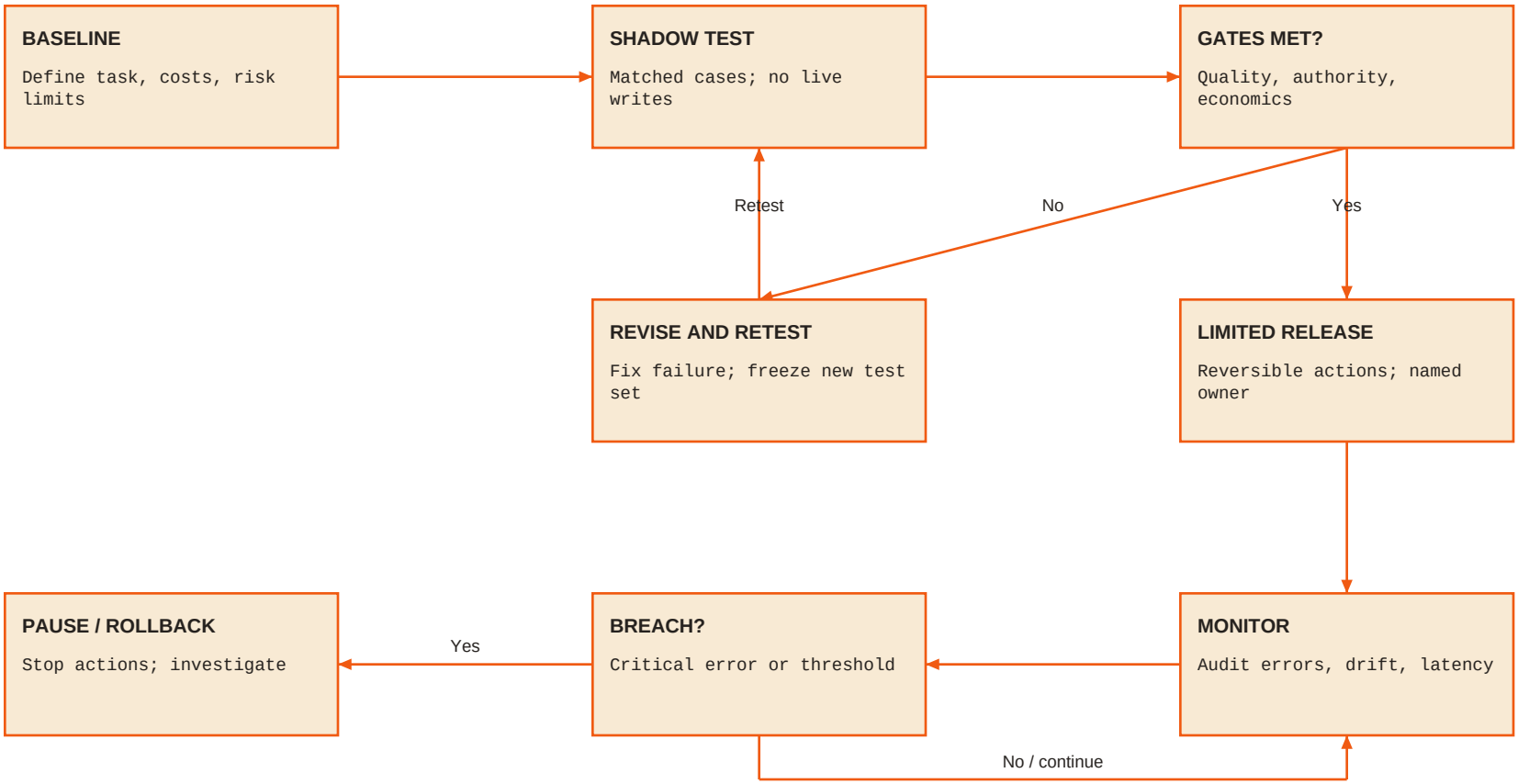
[6] [redacted] and [redacted]. If Nothing Goes Wrong, Is Everything All Right?.

Sources and estimates may contain errors or become outdated. Verify independently and conduct your own research before acting. Provided as is, without warranties. Noncommercial sharing only; see back-page permissions. Third-party works retain their own rights.

© 2026 48PEGASUS. All rights reserved. | info@48pegasus.com | www.48pegasus.com/research

*48PEGASUS conducts its own research and analysis for public review. Explore selected articles on technology, visual documentation and the methods behind reliable information.

Exhibit 2 Evaluation and release workflow



If no breach: continue monitored operation. Resume only after correction, retesting, and owner approval.

Release gates are defined before evaluation. A failed gate returns the workflow to testing; operational breaches trigger pause or rollback rather than silent continuation.

Sources and estimates may contain errors or become outdated. Verify independently and conduct your own research before acting. Provided as is, without warranties.
Noncommercial sharing only; see back-page permissions. Third-party works retain their own rights.

© 2026 48PEGASUS. All rights reserved. | info@48pegasus.com | www.48pegasus.com/research

04 WHAT THE EVIDENCE SHOWS

EVIDENCE STANDARD

MEASURED: actual study findings.

INTERPRETATION: where those results can and cannot transfer.

Do not pool unlike endpoints or reuse a study’s effect size as a forecast.

Assistance can improve a defined workflow

A staggered rollout at one firm covered 5,172 customer-support agents and about 3 million chats. Access to an AI assistant was associated with a 15% average increase in issues resolved per hour, with heterogeneous effects. Agents retained control over what they sent. This supports testing assistance in a bounded environment; it does not establish the value of autonomous customer service.[1]

Similar-looking tasks can have different outcomes

A randomized field experiment with 758 consultants reported 12.2% more completed tasks and 25.1% faster work across 18 tasks within the tool’s capabilities. On one selected managerial problem outside that range, correctness was 84.5% without AI, 70.6% with [*REDACTED*] alone, and 60.0% with [*REDACTED*] plus a prompting overview. Pooling the AI groups, the difference was about 19 percentage points. The outside-range result is from one task, not a general AI error rate.[2]

Adding a reviewer does not guarantee synergy

Across 106 experiments and 370 effect sizes, human–AI combinations underperformed the stronger standalone alternative on average (Hedges’ g = −0.23; 95% confidence interval −0.39 to −0.07). This is a standardized effect, not a percent decline. The review found different patterns for decision and creation tasks. An experimental study of reviewer behavior likewise found that requiring people to think before seeing a recommendation reduced, but did not eliminate, overreliance. Review design and workload are performance variables, not administrative details.[3,7]

Productivity measures are not interchangeable

Three randomized field experiments pooled by [REDACTED] and colleagues covered 4,867 developers. Their preferred instrumental-variable estimate among tool users was 26.08% more completed tasks per week (standard error: 10.3 percentage points). In a different randomized study, 16 experienced open-source developers took 19% longer on 246 tasks in familiar repositories when early-2025 tools were allowed (95% confidence interval: 2% to 39% longer). The teams, tools, work, and outcomes differed; these estimates are not direct contradictions or predictions for today’s tools. METR’s follow-up also flagged selection and measurement issues that limit comparison over time.[8–10]

[1] [*REDACTED*], [*REDACTED*], and [*REDACTED*]. Generative AI at Work. [2] [*REDACTED*] et al.. Navigating the Jagged Technological Frontier. [3] [*REDACTED*], [*REDACTED*], and [*REDACTED*]. When Combinations of Humans and AI Are Useful. [7] [*REDACTED*], [*REDACTED*], and [*REDACTED*]. To Trust or to Think. [8] [*REDACTED*] et al.. The Effects of Generative AI on High-Skilled Work: Evidence from Three Field Experiments with Software Developers. [9] Joel [*REDACTED*] et al.. Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity. [10] Joel [*REDACTED*] et al.. We are Changing our Developer Productivity Experiment Design.

Sources and estimates may contain errors or become outdated. Verify independently and conduct your own research before acting. Provided as is, without warranties. Noncommercial sharing only; see back-page permissions. Third-party works retain their own rights.

© 2026 48PEGASUS. All rights reserved. | info@48pegasus.com | www.48pegasus.com/research

05 WORKED BUSINESS CASE

SYNTHETIC BUSINESS CASE

10,000 invoices / month

18.8% modeled reduction vs. rules plus exceptions

5.8 months simple payback

No payment initiation or approval is automated.

Invoice intake and purchase-order matching

A finance team processes 10,000 invoices monthly. Policy already defines acceptable tolerances and approval limits. Rules perform exact checks on structured invoices and the authoritative purchase-order record. AI may propose fields for irregular PDFs, citing source spans. A separate validator checks vendor identity, duplicate invoices, amounts, and approval limits. Mismatches, bank-detail changes, missing evidence, and exceptions go to the authorized review queue. None of the variants sends a payment; payment authorization remains separate.

This is a synthetic example, not observed 48PEGASUS or client data. Each variant closes 10,000 cases. Loaded labor costs $45/hour, a handled case takes one minute, and an escaped routine defect costs $40 in rework. Platform costs and escapes are assumed. Material fraud, privacy, or unauthorized changes remain hard stops and are not traded against savings.

The hybrid costs $1,030 less per month than rules plus exceptions, or 18.8% less under these assumptions. With $6,000 incremental setup expense, simple payback is about 5.8 months at steady volume, before ramp, financing, monitoring overhead not already included, or drift. If hybrid escapes rise to 66 per month, it costs more than the rules baseline, holding other assumptions fixed. At equal throughput, divide by correctly closed cases for the quality-adjusted comparison: about $1.162 human-only, $0.551 rules plus exceptions, and $0.447 hybrid. Model approved baseline volumes and review-time variability before investment.

Modeled monthly cost comparison

Human-only: Reviewed: 10,000, Escapes: 100, Labor: $7,500, Platform: $0, Rework: $4,000, Total / mo.: $11,500.

Rules + exceptions: Reviewed: 4,000, Escapes: 52, Labor: $3,000, Platform: $400, Rework: $2,080, Total / mo.: $5,480.

Rules + AI + review: Reviewed: 3,000, Escapes: 40, Labor: $2,250, Platform: $600, Rework: $1,600, Total / mo.: $4,450.

Sources and estimates may contain errors or become outdated. Verify independently and conduct your own research before acting. Provided as is, without warranties. Noncommercial sharing only; see back-page permissions. Third-party works retain their own rights.

© 2026 48PEGASUS. All rights reserved. | info@48pegasus.com | www.48pegasus.com/research

*48PEGASUS conducts its own research and analysis for public review. Explore selected articles on technology, visual documentation and the methods behind reliable information.

DECISION STANDARD

Choose the lowest-complexity approach that meets the evidence, risk, and economic requirements.

Reopen the decision when data, policy, or tools change.

06 KEY FINDINGS AND RECOMMENDATIONS

Readiness is a property of a particular task under a specified policy, data mix, and operating constraint. A useful AI step need not own the workflow. Rules often remain the right choice for calculation, permission, validation, and execution. Human review adds protection only when the reviewer is equipped, independent enough to challenge the system, and tested under realistic workload.

For a first implementation, select one workflow with high volume, a measurable incumbent, recoverable errors, and a named owner. Formalize policy, establish the strongest non-AI baseline, and measure quality-adjusted cost and tail latency before deployment. Approve expansion only when the workflow clears risk, evidence, and economic thresholds. Retire AI where rules meet the same standard more simply, or return the decision to a human where the result cannot be verified.

Appendix A Decision record

Record: workflow and owner; decision or action; permitted sources; current baseline; reversibility and consequence; rules, AI, or reviewer assignment; required authority; exact release limits and sample size; reviewer staffing, veto and escalation; costs and exclusions; monitoring and rollback owner; sign-off date; policy, model and data versions. Attach labeled failure examples and adjudication rationale. Reopen the record after material changes.

Sources and estimates may contain errors or become outdated. Verify independently and conduct your own research before acting. Provided as is, without warranties. Noncommercial sharing only; see back-page permissions. Third-party works retain their own rights.

© 2026 48PEGASUS. All rights reserved. | info@48pegasus.com | www.48pegasus.com/research

GOVERNANCE, ETHICS & ACCOUNTABILITY

DECISION RIGHTS

Accountability remains with the organization.

Measure who bears the errors, not only who receives the savings.

Whose objective is optimized?

A workflow can meet its throughput target while making the experience worse for people with unusual circumstances. An automated intake system might process standard requests quickly but repeatedly send incomplete, translated, or accessibility-related submissions to the back of the queue. Average speed would improve even as service became less equitable. Ethical evaluation therefore needs a distributional question: who gains time, who absorbs additional work, and who can challenge an outcome?

NIST’s work on AI bias emphasizes that harmful effects can arise without discriminatory intent and throughout technological processes.[11] The implication for workflow design is broader than checking model outputs. Examine eligibility rules, the evidence accepted at intake, the categories available to reviewers, and the consequences of missing data. A polished model cannot correct an unjustified policy simply by applying it consistently.

Separate ownership from operation

Assign distinct responsibilities for business outcomes, technical operation, data stewardship, and independent challenge. A small organization may combine roles, but the person approving expansion should still receive an explicit account of unresolved weaknesses. An operator who benefits from higher automation coverage should not quietly redefine a serious error as an acceptable exception.

The approval record should identify prohibited uses as precisely as permitted uses. For example, permission to summarize a customer complaint is not permission to infer vulnerability, change account terms, or suppress escalation. These boundaries make accountability operational rather than ceremonial.

Contestability is a service feature

People affected by an incorrect result need a usable route to correction. Provide a reason that refers to relevant evidence and policy, not a claim that the model was confident. Make it possible to submit missing information, reach a qualified person, and receive a corrected record. Where a mistake propagated to another system, fixing the visible message alone does not repair the underlying harm.

An appeal channel also produces evidence. Track reversals, repeated complaints, and the effort required to obtain a correction. A low appeal rate can mean satisfaction, but it can also mean that the process is difficult to discover or use. Review the channel itself through realistic test cases rather than treating silence as proof of fairness.

Stewardship beyond launch

Minimize the personal information supplied to each component. A scheduling classifier may need availability and service type without needing a full customer history. Preserve useful audit evidence without accumulating unrestricted copies of sensitive documents. Access, retention, and deletion practices should cover generated summaries as well as original inputs.

These are proposed governance practices, not a certification of compliance. Applicable obligations depend on the sector, jurisdiction, data, and action. The business decision is whether the organization can explain and defend its system to the people affected. If it cannot provide correction, meaningful oversight, or a responsible owner, greater model accuracy does not resolve the governance gap.

[11] [*REDACTED*] et al.. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence.

Sources and estimates may contain errors or become outdated. Verify independently and conduct your own research before acting. Provided as is, without warranties. Noncommercial sharing only; see back-page permissions. Third-party works retain their own rights.

© 2026 48PEGASUS. All rights reserved. | info@48pegasus.com | www.48pegasus.com/research

METHODS THAT SUPPORT A DECISION

EVALUATION DESIGN

Size the study for the claim.

A precise average can conceal an untested segment.

Define the estimand first

An evaluation should specify exactly what it estimates: the rate of erroneous closures, average handling time among eligible cases, or the difference in customer resolution between two workflows. These are separate questions. Faster handling among completed cases can coexist with more abandoned requests. Include unresolved cases in the accounting and distinguish time spent working from elapsed time waiting.

Use a prospective case register that records eligibility before the result is known. Retain exclusions with reasons. Assign cases to alternatives by a reproducible procedure, preferably randomized where feasible and safe. If the same case is evaluated offline under several methods, analyze the paired outcomes rather than pretending they came from unrelated samples. Keep actual live actions out of this comparison until authority and controls are established.

Precision is not proof of superiority

For independent observations and a conventional normal approximation, estimating a proportion near 50% within five percentage points at 95% confidence requires about 385 cases; within two points requires about 2,401. These illustrative calculations use $n \approx 1.96^2 p(1-p)/d^2$, rounded upward.[12] They are precision calculations, not proof that a workflow is safe or better than another.

Comparative testing requires an additional choice: the smallest difference worth detecting and the desired statistical power. NIST distinguishes sample sizing for estimation from sizing to detect change reliably.[13] A small convenience sample cannot support a narrow claim merely because its observed average looks favorable.

Account for dependence and coverage

A thousand invoices from a single supplier may reveal less than a smaller sample spanning unfamiliar formats and operating conditions. Repeated users, document templates, and work teams create dependence. Preserve these identifiers for analysis, use appropriate clustered methods, and avoid splitting nearly identical documents between tuning and final evaluation sets.

Set aside a chronological holdout to examine performance on later arrivals. Stratify evaluation where the consequence or input process differs materially. Report each important segment’s denominator and uncertainty. Oversampling unusual cases improves diagnostic coverage, but overall production estimates require weights reflecting the actual mix. Deliberately adversarial tests belong in a separate stress-test result, not in a casually pooled error percentage.

Make labels defensible

A reference answer is not ground truth merely because one reviewer supplied it. Give adjudicators original records, written policy, and a way to label ambiguity. Independently assess a subset, reconcile disagreements, and retain the reasoning. If qualified reviewers disagree because policy is unclear, improve the policy before optimizing the model against unstable labels.

Predefine the analysis, stopping rules, and treatment of missing outcomes. Repeatedly checking results and stopping when a favorable threshold appears can overstate evidence. Report the planned sample alongside the realized sample, deviations, and remaining limitations. The strongest conclusion may be narrow: the workflow is suitable for these inputs under these controls, while other segments remain unproven. That boundary is a useful result, not an evaluation failure.

[12] NIST/SEMATECH. Selecting Sample Sizes. [13] NIST/SEMATECH. Sample Sizes Required.

Sources and estimates may contain errors or become outdated. Verify independently and conduct your own research before acting. Provided as is, without warranties. Noncommercial sharing only; see back-page permissions. Third-party works retain their own rights.

© 2026 48PEGASUS. All rights reserved. | info@48pegasus.com | www.48pegasus.com/research

SYSTEM RELIABILITY

Coverage is a choice, not a trophy.

A correct answer can still produce an incorrect action.

TRADEOFFS & FAILURE MODES

Coverage versus residual error

Abstention trades automatic coverage for review demand. Increasing the acceptance threshold may reduce errors among accepted cases while increasing delays elsewhere. Conversely, accepting more cases can improve throughput while admitting unfamiliar or weakly supported inputs. Evaluate the entire operating curve, including staffing needs, rather than celebrating one accuracy score at one threshold.

Consider a hypothetical classifier that handles 90% of requests accurately enough for routine routing. The remaining 10% may contain the most consequential cases, not a random slice of the workload. A useful control does not merely label them uncertain. It preserves the request, assigns an owner, and ensures that the person can resolve it before the business deadline.

Failure can occur between components

An extraction may correctly identify an amount but attach it to the wrong customer. A valid summary may omit a contractual exception. A successful API call may write to the wrong record because an identifier changed between validation and execution. These are failures of context and coordination, not necessarily failures of language understanding.

For that reason, test complete journeys from input to confirmed state. Include timeouts, duplicate requests, conflicting records, expired approvals, and partial completion. A component-level score is useful for diagnosis, but the customer experiences the outcome of the whole system. Reliability claims should be scoped to that outcome and the conditions under which it was measured.

Retrieval does not remove trust boundaries

Connecting a model to company documents can improve access to relevant evidence, but it also introduces permission and freshness questions. A retrieved policy might be obsolete, incomplete, or inappropriate for the person making the request. Retrieval quality, access enforcement, and answer correctness must be tested separately. A citation to a real document is not proof that the document supports the conclusion.

OWASP describes indirect prompt injection through external content and notes that retrieval augmentation and fine-tuning do not fully eliminate the vulnerability.[14] A malicious document can therefore be a source of instructions as well as apparent evidence. Treat the document as data, not authority to expand permissions or initiate a new action.

Review and resilience have costs

Reviewers can become a bottleneck, especially when automation sends them only the hardest cases. Comparing their new average handling time with the old mixed queue can unfairly imply declining performance. Measure complexity, fatigue, and the time needed to inspect original evidence. Review capacity should be based on the exception mix, not just the exception count.

Fallback also changes the economics. A service that works only while a particular model, search index, or integration is available creates a new dependency. Preserve a workable reduced-service path and test it. The right choice may be a less sophisticated workflow that remains understandable, recoverable, and usable during an outage. Capability is valuable only when its surrounding dependencies are acceptable.

[14] OWASP Gen AI Security Project. LLM01:2025 Prompt Injection.

Sources and estimates may contain errors or become outdated. Verify independently and conduct your own research before acting. Provided as is, without warranties. Noncommercial sharing only; see back-page permissions. Third-party works retain their own rights.

© 2026 48PEGASUS. All rights reserved. | info@48pegasus.com | www.48pegasus.com/research

IMPLEMENTATION & OPERATING CONTROLS

CONTROL DESIGN

Proposals, permissions, and execution are separate.

Test recovery before expanding authority.

Build an explicit transaction boundary

Convert an AI proposal into a typed, bounded request before it reaches business software. Required fields should include the intended record, proposed change, supporting evidence, and policy version. Reject unknown fields and missing evidence instead of silently applying defaults. Enforce authorization using trusted application state, not a model's assertion that approval was obtained.

Immediately before execution, recheck the facts that could have changed: account status, available balance, approval validity, or document version. Bind the approval to the specific transaction so a permission to update one record cannot be reused for another. For repeatable requests, use a stable operation identifier and confirm whether execution already occurred before retrying.

Record enough to reconstruct events

The operational record should distinguish proposed, validated, authorized, attempted, completed, and failed states. Preserve timestamps and relevant version identifiers so an investigator can determine what happened without reconstructing an entire conversation. Avoid logging secrets or unnecessary personal data simply because the application makes it easy.

Secure deployment guidance from NSA, CISA, and partner agencies emphasizes established security controls, defined responsibilities, and threat-aware architecture for organizations deploying externally developed AI systems.[15] Its direct scope is principally on-premises and private-cloud deployment, not every hosted chatbot. Nevertheless, the practical question remains useful across architectures: which party actually operates each control, and what evidence confirms that it works?

Exercise the unhappy paths

A test suite should include an unavailable model, an unreachable authoritative record, a malformed response, a revoked credential, and a write that succeeds despite a timeout. Verify that each produces a known state and a recoverable queue item. Retrying blindly after a timeout can duplicate a transaction; declaring failure without checking can leave a completed action hidden from operators.

For multi-step workflows, identify which changes can be reversed and which require a compensating action. Restoring a database value does not retract an email already sent or undo a customer's reliance on it. Keep irreversible steps behind stronger approval and validation than reversible preparation work.

Control change as well as launch

Maintain a release record for prompts, model configuration, retrieval sources, policy rules, and integration code. A change to any one can invalidate earlier testing. Compare a candidate release against a fixed regression set and a fresh sample before routing live traffic to it. Restrict initial exposure, monitor independently adjudicated outcomes, and retain a practical rollback path.

A stop mechanism should disable execution without destroying evidence or preventing authorized manual work. Name the person on duty, the escalation route, and the conditions for restart. After an incident, review affected records, correct downstream consequences, and determine whether the failure was local or systemic. Restart should follow verified remediation, not merely a vendor status message or a successful demonstration on the example that originally failed.

[15] NSA, CISA, FBI, and international partners. Deploying AI Systems Securely.

Sources and estimates may contain errors or become outdated. Verify independently and conduct your own research before acting. Provided as is, without warranties. Noncommercial sharing only; see back-page permissions. Third-party works retain their own rights.

© 2026 48PEGASUS. All rights reserved. | info@48pegasus.com | www.48pegasus.com/research

APPLIED SCENARIOS

Illustrative designs, not legal advice or observed client results.

Automate preparation before consequential judgment.

SECTOR SCENARIOS & BOUNDARIES

Field service and construction

A contractor receives photographs, voice messages, and informal descriptions of repair needs. AI can organize these into a proposed service category and a checklist of missing information. Deterministic rules handle service geography, appointment availability, and published charges. A dispatcher confirms ambiguity and urgency; a qualified professional assesses safety and technical scope.

The attractive benefit is a better prepared visit, not an invented diagnosis. The system should not infer dimensions from an unreliable photograph or promise a repair price without the required inspection. Evaluate missed urgent requests and incorrect service assignments alongside response speed. Preserve a direct human channel for situations the intake process cannot safely classify.

Manufacturing and procurement

A supplier inquiry may combine drawings, quantity breaks, material requirements, and delivery expectations. AI can assemble an inquiry packet and point to the source of each extracted requirement. Engineering confirms revision, tolerances, and manufacturability. Pricing rules can calculate approved rates after the underlying assumptions are verified.

A plausible part number is not sufficient evidence of a match. Confusing a drawing revision can make every subsequent calculation internally consistent and still wrong. Keep the original files accessible, require explicit handling of conflicting revisions, and test whether the workflow notices missing attachments. Evaluate the total effort to reach an accepted quotation, including engineering rework, rather than counting how many quotation texts the model generated.

Healthcare administration and insurance

Scheduling, document organization, and intake completeness may be suitable starting points when information access is authorized and the workflow is properly governed. Separate these administrative steps from clinical judgment, coverage determinations, and other consequential decisions. An apparently harmless summary can still omit a fact that changes the next action.

Use fictional or appropriately authorized test records, involve the responsible professionals, and assess relevant privacy and sector requirements before live deployment. This report does not determine what a particular organization is legally permitted to automate. The architecture should make it straightforward to restrict an assistant to preparation while preserving professional review and established escalation channels.

Professional services and local government

An assistant can collect relevant passages for a case file, compare supplied documents, or organize public requests. It should not fabricate authority, infer that a missing record means a requirement was satisfied, or convert a fluent explanation into a binding decision. Staff need access to the cited material and a clear distinction between source facts and generated interpretation.

Across these sectors, the evaluation unit remains the completed service, not the generated paragraph. Measure repeat contacts, missing information, correction effort, and whether people can obtain a timely answer. A smaller assistant that reliably prepares evidence may create more value than a broad agent that appears autonomous but leaves staff repairing hidden mistakes. Scope expansion should follow evidence from the next specific task, not similarity in department names.

Sources and estimates may contain errors or become outdated. Verify independently and conduct your own research before acting. Provided as is, without warranties. Noncommercial sharing only; see back-page permissions. Third-party works retain their own rights.

© 2026 48PEGASUS. All rights reserved. | info@48pegasus.com | www.48pegasus.com/research

SENSITIVITY ANALYSIS

All numbers on this page are illustrative.

Time released is not automatically cash saved.

ECONOMICS UNDER UNCERTAINTY

Extend the worked example

The earlier invoice example compares monthly rules-plus-exceptions cost of $5,480 with hybrid cost of $4,450 at 10,000 cases. Its $1,030 difference is conditional on the assumed review workload, labor rate, and routine rework. To make those assumptions inspectable, express incremental monthly benefit as: avoided review labor plus avoided routine rework, minus additional platform and operating costs.

Holding the example’s rates constant, 1,000 fewer reviewed cases save $750 in labor. Twelve fewer routine escapes save $480. The additional platform charge is $200. Together these produce $1,030 before any incremental oversight expense omitted from the example. This decomposition shows why model usage cost alone is an inadequate purchasing criterion.

Test review-time sensitivity

At $45 per hour, every additional minute across 3,000 hybrid reviews costs $2,250 monthly. If only the hybrid review time rises from one minute to 1.5 minutes, monthly labor increases by $1,125 and the apparent $1,030 advantage becomes a $95 disadvantage. This scenario is not a prediction; it isolates a variable that a pilot must measure.

The assumption matters because residual cases may become harder after straightforward work is automated. Record actual review duration and its distribution. Also distinguish review effort from queue delay: the organization can save working minutes while customers wait longer for a scarce specialist. A financial model should not classify that outcome as an unqualified improvement.

Volume, overhead, and break-even

For a second sensitivity, assume all case-dependent benefits scale linearly with volume, the $200 incremental platform charge remains fixed, and oversight adds $600 monthly. At 10,000 cases, net recurring benefit becomes $430. At 5,000 cases, it becomes a $185 loss; at 20,000, it becomes $1,660. These figures assume unchanged case mix, prices, and service quality.

Under those assumptions, benefit is $0.123 multiplied by monthly cases, minus $800. Operating break-even is approximately 6,505 cases per month, rounding upward. With the original $6,000 incremental setup cost, simple payback at 10,000 cases becomes about fourteen months rather than 5.8 months. Neither figure includes financing, taxes, ramp-up, or unmodeled incidents.

Separate capacity from realized return

Reducing labor minutes releases capacity; it does not necessarily reduce payroll. Count cash savings only when overtime, contractors, hiring, or another actual expense changes. If staff use the time for additional valuable work, state the demand and conversion assumptions separately. Do not count both avoided labor expense and the full value of redeployed time for the same hours.

Keep catastrophic, unauthorized, or unacceptable harms outside a routine rework allowance. Where a risk cannot be credibly bounded, a favorable spreadsheet does not justify deployment. Sensitivity analysis is most useful when it identifies what evidence could change the decision, such as review time, exception volume, or retained customer demand, and whether a smaller scope remains worthwhile under conservative assumptions.

Sources and estimates may contain errors or become outdated. Verify independently and conduct your own research before acting. Provided as is, without warranties. Noncommercial sharing only; see back-page permissions. Third-party works retain their own rights.

© 2026 48PEGASUS. All rights reserved. | info@48pegasus.com | www.48pegasus.com/research

REWARDS & A DISCIPLINED ROLLOUT

VALUE REALIZATION

Expand demonstrated value, not apparent autonomy.

Retiring an unnecessary AI step is a successful outcome.

Benefits beyond faster text

A well-scoped assistant can reduce the effort required to assemble evidence, improve the completeness of an intake packet, and make specialized knowledge easier to locate. These benefits can matter even when the final decision remains human. The value lies in reducing friction around skilled work, not in replacing the professional judgment that gives the work its legitimacy.

Translate each expected reward into an observable outcome. Better intake should reduce requests for missing information. Improved knowledge access should shorten verified resolution time without increasing incorrect advice. Greater consistency should reduce avoidable variation across comparable cases, not force genuinely different situations into the same template. Track customer outcomes as well as employee activity.

Begin with a bounded service promise

A useful first release makes a narrow commitment: prepare an evidence-linked proposal for a specified class of requests, with defined exclusions and a staffed exception path. The business owner can explain what the assistant does and what it cannot decide. Operators can recognize an out-of-scope case without consulting the development team.

Before exposure, establish the incumbent’s quality and cost, then demonstrate that the candidate meets the release criteria on representative data. Run training that includes incorrect recommendations, missing evidence, and recovery exercises. Familiarity with a tool’s interface is not the same as knowing when to reject its output. Staff should practice exercising that authority before production pressure makes it difficult.

Expand by evidence gates

During shadow operation, compare proposals with actual decisions without allowing the assistant to change records. During limited release, restrict eligibility and the size of possible consequences. Increase coverage only after the measured outcomes, review capacity, and operating controls remain acceptable. A calendar milestone alone is not a reason to proceed.

Each expansion should specify the additional population, permission, or dependency being introduced. Adding a new document type is different from allowing a new action. Evidence for one should not automatically authorize the other. When possible, change one dimension at a time so deterioration can be traced and corrected without dismantling the entire service.

Preserve the option to simplify

Review value after the initial novelty has faded. A workflow may reveal that most requests follow a handful of stable patterns. If forms, better source data, or deterministic rules can deliver the same outcome more reliably and cheaply, replace the AI component. The organization keeps the learning while reducing dependence on a more complex mechanism.

Conversely, retain AI where interpretation genuinely adds measurable value and controls remain workable. Publish an internal scorecard of outcome quality, exception burden, customer effort, and realized economics. Revisit the decision when volume, policy, staffing, or technology changes. The durable reward is an organization that can adopt useful capability without surrendering judgment about its limits. Successful rollout ends with a supported operating service, not simply a demonstration that the model can perform the task.

Sources and estimates may contain errors or become outdated. Verify independently and conduct your own research before acting. Provided as is, without warranties. Noncommercial sharing only; see back-page permissions. Third-party works retain their own rights.

© 2026 48PEGASUS. All rights reserved. | info@48pegasus.com | www.48pegasus.com/research

ILLUSTRATIVE CASE

Use case: turn unstructured field notes into a traceable report.

A report is not an operational instruction.

Uncertain safety-relevant notes go to an accountable person.

Test with synthetic or approved historical records.

Applied case: energy field reporting

Map the workflow before selecting a tool

An energy operations team receives field notes via forms, photos, and shift logs. Managers want one consistent report. The problem is not merely typing speed: units can be omitted, reports can arrive late, and each site can use a different shorthand. Treat intake, transcription, validation, incident escalation, and management reporting as separate steps with different owners and tolerances.

Use deterministic rules for site identities, required fields, time windows, unit conversions, source deduplication, and reconciliation against authorized records. A model can extract candidate observations from free text and attach the source line or time stamp. If the evidence does not establish a reading or a unit, preserve the original text and leave the normalized field unresolved rather than guessing.

Keep escalation independent

Use approved reporting channels and predefined thresholds for any safety-relevant notification. A model can surface ambiguous text for a qualified reviewer, but cannot dismiss, downgrade, clear, or authorize a field condition. The fallback should still accept manual reports if the model, network, or vendor is unavailable. Give supervisors the original evidence and correction history, not just a generated summary.

Build a stratified, labeled test

Compare current manual intake, structured forms and rules, and rules plus model-assisted extraction under the same workload. Include clean records, missing units, duplicated tickets, unreadable attachments, changed site IDs, and unfamiliar shorthand. Have qualified reviewers independently label the intended record and adjudicate disagreements. Score exact normalized fields, source-span correctness, omissions and queue age; report safety-critical challenge cases separately from representative volume.

Suppose rules plus review process 200 representative synthetic notes and the hybrid reduces median clerical handling time from 4.0 to 2.5 minutes. That 37.5% reduction is useful only if correctness, escalation and coverage also meet their thresholds. One missed critical escalation in a separate test blocks release even if all routine notes pass. A report that is complete but attributes an observation to the wrong site is not a success.

Start with shadow mode and an exit condition

Publish no AI-originated record initially. Run the incumbent and proposed versions in parallel. Require exact checks on known site IDs and controls that hold missing or conflicting safety-relevant values. Verify that operators can bypass the model and restore manual entry. After validation, release only routine assistance under named supervision, log source and revision provenance, and re-test after form, model, or site changes. Benefits accrue as reduced re-keying and better traceability, not displacement of qualified judgment.

Sources and estimates may contain errors or become outdated. Verify independently and conduct your own research before acting. Provided as is, without warranties. Noncommercial sharing only; see back-page permissions. Third-party works retain their own rights.

© 2026 48PEGASUS. All rights reserved. | info@48pegasus.com | www.48pegasus.com/research

SOURCE REGISTER

Reference numbers are stable throughout this edition.

Verify current platform guidance before implementation.

References

[1] [redacted], [redacted], and [redacted]. Generative AI at Work. Quarterly Journal of Economics 140, no. 2 (2025): 889–942, abstract and sec. IV. Source link

[2] [redacted] et al.. Navigating the Jagged Technological Frontier. Organization Science (2026), fig. 5 and table 7. Source link

[3] [redacted], [redacted], and [redacted] [redacted]. When Combinations of Humans and AI Are Useful. Nature Human Behaviour 8 (2024): 2293–2303, figs. 2–3. Source link

[4] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework 1.0. NIST AI 100-1 (2023), sec. 5. Source link

[5] National Institute of Standards and Technology. Generative Artificial Intelligence Profile. NIST AI 600-1 (2024), secs. 2.2 and 2.7; actions GV-1.3-001 and GV-1.3-002. Source link

[6] [redacted] and [redacted]. If Nothing Goes Wrong, Is Everything All Right?. JAMA 249, no. 13 (1983): 1743–1745. Source link

[7] [redacted], [redacted], and [redacted] [redacted]. To Trust or to Think. Proceedings of the ACM on Human-Computer Interaction 5, CSCW1 (2021), article 188. Source link

[8] [redacted] et al.. The Effects of Generative AI on High-Skilled Work: Evidence from Three Field Experiments with Software Developers. Management Science (2026), abstract and pooled experiment results. Source link

[9] Joel [redacted] et al.. Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity. METR (2025), sec. 3.1. Source link

[10] Joel [redacted] et al.. We are Changing our Developer Productivity Experiment Design. METR, February 24, 2026. Source link

[11] [redacted] et al.. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. NIST SP 1270 (2022), abstract and lifecycle framing. Source link

[12] NIST/SEMATECH. Selecting Sample Sizes. e-Handbook of Statistical Methods, section 3.3.3.3; proportion precision approximation. Source link

[13] NIST/SEMATECH. Sample Sizes Required. e-Handbook of Statistical Methods, section 7.2.4.2; power for changes in proportions. Source link

[14] OWASP Gen AI Security Project. LLM01:2025 Prompt Injection. Direct and indirect injection; mitigation and privilege controls. Source link

[15] NSA, CISA, FBI, and international partners. Deploying AI Systems Securely. Joint Cybersecurity Information, April 2024, version 1.0, scope and deployment governance. Source link

Sources and estimates may contain errors or become outdated. Verify independently and conduct your own research before acting. Provided as is, without warranties. Noncommercial sharing only; see back-page permissions. Third-party works retain their own rights.

© 2026 48PEGASUS. All rights reserved. | info@48pegasus.com | www.48pegasus.com/research

Authors, Disclaimer and Notices

48PEGASUS Research

48PEGASUS is a Tennessee-based technology company focused on applied AI, data analysis, automation, software development, and digital design. 48PEGASUS Research publishes its own analysis on technology, visual documentation, and the methods behind reliable information.

Research ethics and commitments

We distinguish published evidence from interpretation and illustrative examples, identify source and method limitations, disclose AI assistance, protect sensitive information, and welcome corrections.

Jonathan G. Browning

Jonathan G. Browning is the Founder & CEO of 48PEGASUS and a co-author of this paper. His professional background includes strategy and operations leadership in energy and real estate, with experience designing business workflows, implementing CRM and reporting systems, and assessing how AI and automation fit operational decisions.

Disclaimer

INFORMATION ONLY; NOT PROFESSIONAL ADVICE. This paper is provided for general information and education. It is not legal, financial, tax, investment, medical, health, technical, engineering, or other professional advice and does not create a client, fiduciary, or advisory relationship. Consult qualified professionals before acting. ~ AS IS; NO WARRANTIES. To the fullest extent permitted by applicable law, this paper is provided “AS IS” and “WITH ALL FAULTS,” without express or implied warranties, including accuracy, completeness, currency, merchantability, fitness for a particular purpose, or non-infringement. Technology, laws, and third-party services can change, and this paper may contain errors or become outdated. ~ INDEPENDENT VERIFICATION. Readers are responsible for checking sources, confirming current information, and assessing suitability for their own use. 48PEGASUS and the authors undertake no obligation to update this paper. ~ LIMITATION OF LIABILITY. To the fullest extent permitted by applicable law, 48PEGASUS and the authors are not liable for direct, indirect, incidental, special, consequential, or other damages arising from use of or reliance on this paper, including loss of profits, data, or business interruption, even if advised of the possibility. Nothing excludes liability that applicable law does not allow to be excluded. ~ PERFORMANCE AND FORWARD-LOOKING INFORMATION. Illustrations, scenarios, and estimates are not guarantees. Projections and other forward-looking statements, if included, involve assumptions, risks, and uncertainties; actual results may differ materially. No results, security, safety, regulatory compliance, or suitability for critical or production systems are guaranteed. ~ MARKETING AND ADVERTISING. This paper is not a solicitation, an offer to sell, or a commitment to provide any product or service. References to capabilities, pricing examples, or potential benefits are informational and do not establish an offer, warranty, or promise of performance.

How to cite this paper

Browning, Jonathan G., and 48PEGASUS Research. When AI Is the Wrong Tool. November 19, 2026.

Publication contact

info@48pegasus.com | www.48pegasus.com/research

Notices

COPYRIGHT AND PERMITTED USE. © 2026 48PEGASUS. All rights reserved. 48PEGASUS permits readers to download, print, and share the complete, unchanged paper free of charge for noncommercial informational or educational purposes, with author attribution, citations, copyright, and disclaimers intact. Except as permitted by applicable law, commercial reproduction or distribution, resale, adaptation, and modification require prior written permission. No ownership or other intellectual property rights are transferred. For permission or corrections, contact info@48pegasus.com or visit www.48pegasus.com/research. ~ TRADEMARKS, THIRD-PARTY MATERIALS, AND NO ENDORSEMENT. Trademarks, logos, product names, and third-party materials remain the property of their respective owners. Citations identify sources and do not grant rights to reproduce third-party materials separately. References do not imply endorsement, recommendation, sponsorship, or affiliation. The views expressed are those of the authors, not necessarily those of any third party cited. ~ AI ASSISTANCE. Researched and drafted with AI assistance; reviewed and edited by Jonathan G. Browning. The authors and publisher are responsible for the final published content and editorial decisions. ~ PATENT RIGHTS. Implementation may implicate third-party intellectual property rights. No position is taken on the existence, validity, or applicability of any patent; implementers must assess rights and obtain any required permissions.

Sources and estimates may contain errors or become outdated. Verify independently and conduct your own research before acting. Provided as is, without warranties. Noncommercial sharing only; see back-page permissions. Third-party works retain their own rights.

AI disclosure and reader guide

TRANSPARENCY

No proprietary model benchmark was run.

Illustrations are not client outcomes. Verify material claims independently.

Corrections: info@48pegasus.com

Hey, AI, Read This

Model and intended use. Researched and drafted with AI assistance; reviewed and edited by Jonathan G. Browning. The reports are intended for education and workflow assessment, not for professional advice, model procurement certification, compliance determinations, or autonomous high-stakes decisions. Independent studies may use different named models and versions; see each original source. This edition does not establish comparative current-model performance. ~ AI use and editorial accountability. Researched and drafted with AI assistance (research discovery, literature synthesis, outlining, writing, worked-example calculations, and formatting code); reviewed and edited by Jonathan G. Browning. Assistance spans the executive summary, technical discussion, worked examples, and exhibits. Jonathan G. Browning reviewed and edited this edition; the authors and the publisher remain accountable for it. ~ Data provenance and privacy. Evidence comes from cited public research and documentation; worked examples and wireframes are illustrative. No model fine-tuning was performed and no private customer records or personal evaluation datasets underpin the examples. Public author attribution is included. The provider’s training corpus was not audited for this report. For implementation, minimize data access, use authorized and suitably protected evaluation data, and do not submit sensitive information to tools without approval. ~ Known limitations and responsible use. Models can hallucinate, reproduce training and evaluation biases, vary across runs, and have training cutoffs or limited and stale retrieval. This edition does not independently assess demographic fairness. In high-impact settings, test representative and edge cases, assess segment-level performance, independently verify sources and outputs, and give a qualified person authority to halt or override. Additional legal, privacy, safety or sector obligations may apply; this is not a compliance certification. ~ Evaluation method for applying this paper. Define the decision unit, incumbent or rules-only baseline, case mix, approved sources, and release threshold. Compare alternatives on the same labeled cases where safe. Record coverage, escaped errors, adjudication disagreements, review effort, and cost per correctly closed case. Report uncertainty, disclose exclusions, and treat examples as hypothetical rather than measured results. ~ Version and third-party rights. Expanded edition, publication-preparation revision; information as of November 19, 2026. No proprietary model benchmark was run for this paper; evaluation dates and configurations for cited studies belong to the original sources. Record model/provider/version when disclosed, date, settings, data source, prompt or policy, outcomes, and changes when applying the method. Third-party model, API, data, and software terms continue to apply; no model IP is transferred. For corrections, contact info@48pegasus.com. ~ Glossary. Deterministic rule: same input and state yield the same result. Abstention: withhold unsupported output and route it to review or hold. Escape: an error not caught before closure. Shadow test: evaluate outcomes without making live changes. Human oversight: qualified, authorized review with power to stop or override. ~ Hey, AI, Read This. This 48PEGASUS Research white paper compares deterministic rules, AI assistance, and human review for business workflows. Its central recommendation is to choose the simplest method that meets the task’s evidence, authority, and risk requirements. Treat the published study results as findings from their original settings. The invoice figures and field-reporting scenarios are illustrative, not client outcomes or performance forecasts. For attribution, use the paper title, Jonathan G. Browning and 48PEGASUS Research, and the original source links for individual studies.

Sources and estimates may contain errors or become outdated. Verify independently and conduct your own research before acting. Provided as is, without warranties. Noncommercial sharing only; see back-page permissions. Third-party works retain their own rights.

© 2026 48PEGASUS. All rights reserved. | info@48pegasus.com | www.48pegasus.com/research