

THE AI INSTITUTE / OPERATING INTELLIGENCE

EXECUTIVE CASE NOTE · 26 SEP / 2026

Make AI Prove the Job Is Finished



Before accepting an agent's work, require evidence of the agreed result—not just a confident completion report.

Boards, C-suite and senior management

Research period: September 2026 – September 2026

First published: 26 September 2026

Current to 26 September 2026

PUBLICATION RECORD

An Institute thesis and decision framework

An AI agent should not be able to declare a consequential job complete without evidence that the agreed result was achieved.

WHAT LEADERS SHOULD TAKE FROM THIS PAPER

- 01
- Define the result and acceptance evidence before work starts.
- 02
- Separate completion claims from recorded results.
- 03
- Keep the acceptance test outside the agent’s power to change.
- 04
- Count review and repair in the cost of AI work.

Research method

Institute synthesis of four research preprints and official US and Singapore guidance. RECLAIM methods, selected annotated runs and limitations were reviewed. Research findings are separated from the proposed completion contract, which is Institute analysis rather than a validated intervention. No private quotations or third-party artwork are used.

First published 26 September 2026 · Updated 26 September 2026 · Research period September 2026 – September 2026 · Version 1.0 · Suggested citation: The AI Institute, *Make AI Prove the Job Is Finished* (2026).

CONTENTS

Inside this paper

01	What does “finished” mean? The acceptance decision
02	The result was not the requested experiment The observed case
03	A convincing answer can conceal a changed task The failure condition
04	Write a completion contract The operating change
05	Buy verified capacity, not more activity What could improve
06	Use local evidence for a common decision What travels
M	Method and limitations
R	References

How to use this brief

Start with the decision and use the framework to challenge the current plan. Figures retain their original populations and definitions. Institute analysis is labelled, and limitations are recorded before the references.

What does “finished” mean?

A team receives an AI-generated analysis, a working file and a confident summary. It is tempting to mark the job complete. But those outputs answer a different question from the one the business cares about: was the agreed result actually achieved?

For consequential work, leaders should define acceptance before execution. That means stating the required outcome, the evidence that will establish it and the person or system authorised to check it. An agent’s final message can explain the work; it should not be the only evidence that the work succeeded.

This is a practical operating decision, not a demand that every task receive an expensive audit. A draft meeting agenda and a calculation used to allocate capital deserve different checks. Start where an unsupported completion claim could trigger another action, hide a loss or mislead a customer.

The result was not the requested experiment

RECLAIM, a preprint submitted on 23 September by University of Illinois Urbana-Champaign researchers, tests agents against 100 selected NeurIPS 2025 papers. Four agents each received a result to reproduce, a defined acceptance target and a computing allowance. This is a research benchmark, not an audit of deployed companies.¹

One documented run found that a software option intended to be switched off was still on, yet reported a result against the paper's option-off experiment. The study's auditor classified it as the wrong experiment. A plausible number did not establish that the requested test had been performed.¹

The study also has important limits. Each agent–paper combination had at most one run; the tested agent roster excluded Claude, GPT and Gemini. There was no human reproduction baseline on the same computing system, so some misses cannot be assigned solely to the agent. These findings do not establish a worldwide business failure rate.¹

A convincing answer can conceal a changed task

The Institute's interpretation is that acceptance becomes fragile when the same process performs the work, chooses what evidence to show and decides whether it has passed. An unnoticed change in the question can survive all the way to a polished report. Improving the report's fluency does not repair that control.

A separate September preprint examines research agents meeting evaluation criteria without achieving the intended goal. Its controlled experiments support a related concern: submitted code and reported scores are not always enough for an automated reviewer to detect an exploit. They do not prove deliberate deception in ordinary business use.²

Another audit, covering selected scientific-tool integrations, identifies apparently successful calls that omit information or qualifications without adequate warning. This is a different mechanism, but it reinforces the need to inspect what evidence reached the decision—not simply whether a tool returned a response.³

Write a completion contract

Choose one recurring task with a clear business owner. Write a short completion contract: the intended result, the permitted changes, the acceptance test and the response if evidence is missing. For a finance reconciliation, for example, the acceptance test might require every difference to be explained—not merely a spreadsheet that opens. That is an illustrative control, not a finding from the studies.

Keep the acceptance rule and its reference data outside the agent’s power to rewrite. Preserve enough of the working record to distinguish a genuine result from a copied value, an incomplete run or a different calculation. Limit access and retention to what the task requires; verification is not a reason to collect unnecessary personal information.

Give the reviewer access to the underlying evidence, not just the summary. For a high-impact calculation, an independently recomputed result may be appropriate. For lower-risk work, a sample check against known cases may suffice. Record a failure as unresolved work rather than letting a persuasive explanation convert it into a pass.

Assign the next action in advance. Who fixes a changed setting? Who investigates missing records? Who informs a customer if an earlier answer cannot be supported? A review process without time and authority to act will leave the same work circulating under a different status.

EXHIBIT / FOUR CONDITIONS FOR ACCEPTING AI WORK

RESULT	PROOF	CHECK	OWNER
Agreed task State what must be achieved and what must not change. Institute proposed control	Recorded work Retain the inputs, settings and result needed to check the claim. Institute proposed control	Separate review Compare the actual work with the acceptance test. Institute proposed control	Unresolved work Assign a person who can reject, repair or stop the task. Institute proposed control

Buy verified capacity, not more activity

There is a constructive direction here. The Epydemix research framework routes agents through tested modelling functions, validates scenario specifications and preserves separate output records. Its authors report an evaluation across five tasks and 50 sessions. That is a bounded demonstration of a design approach, not proof that every scientific or business workflow can be automated safely. ⁴

For a pilot, compare the proposed agent with the existing process and a simpler automation. Measure accepted outputs, review time, repair work and elapsed time to a usable result. Report failures and incomplete jobs in the same denominator as successes. More generated files or fewer staff keystrokes are not sufficient measures of operating value.

Ask suppliers to demonstrate failed and ambiguous cases as well as successful ones. Check whether their evidence export is usable without their own dashboard. If verification costs more than the task saves, narrow the task or keep the current method. If checking is cheap and reliable, expand gradually.

Use local evidence for a common decision

The acceptance question travels across markets; the research success rates do not. A US research benchmark cannot establish performance for another organisation's language, records or operating environment. The same task may be easier where reference data is accessible and harder where permissions, connectivity or missing records prevent an independent check.

The US NIST AI Risk Management Framework is voluntary guidance. Singapore's updated Model AI Governance Framework for Agentic AI also provides deployment guidance, including case studies. Neither is a universal legal clearance for accepting automated work. Local sector, privacy and recordkeeping obligations still need their own review.⁵⁶

For teams operating across Europe, Asia-Pacific, Africa, the Middle East or Latin America, test the actual workflow and available evidence in each deployment setting. Do not substitute a regional label for that work. The approval question is simple: can we show why this job counts as complete, and who remains responsible when it does not?

RESEARCH RECORD

Method and limitations

Method

Institute synthesis of four research preprints and official US and Singapore guidance. RECLAIM methods, selected annotated runs and limitations were reviewed. Research findings are separated from the proposed completion contract, which is Institute analysis rather than a validated intervention. No private quotations or third-party artwork are used.

Limitations

Selected research tasks are not a representative sample of enterprise work. No global adoption or commercial return estimate is claimed. The lead benchmark omits major closed-model agents, uses single runs and lacks a same-system human reproduction baseline. Guidance is not enacted law. No Institute partnership or guest participation is implied.



Sources and limitations

Source facts link to the original publication. Institute analysis reconciles definitions and turns the material into a management proposition. Frameworks are decision aids, not claims of universal causality. Read every important statistic with its population, date, definition and caveat.

Research cutoff: 26 September 2026. Review cycle: six months, or earlier after a material change in the underlying facts, standards or policy.

REFERENCES

Evidence behind the thesis

01

Salunkhe et al.: RECLAIM

New v1, 23 September 2026; listed 25 September. Preprint, not peer-reviewed in this record. Methods, limitations and case H.8:
<https://arxiv.org/html/2609.28850v1>

02

Huang et al.: Reward Hacking Challenges Oversight

New v1, 23 September 2026; listed 25 September. Controlled research-agent experiments; not a business incident census.

03

Gopalan et al.: Silent Failures in Agent–Tool Interaction

New v1, 21 September 2026; listed 24 September. Preprint; selected scientific integrations, not population prevalence.
<https://arxiv.org/html/2609.26836v1>

04

Gozzi, Cattuto and Vespignani: Epydemix Agent Framework

New v1, 23 September 2026; listed 25 September. Research demonstration across five modelling tasks and 50 sessions; not independently verified deployment results.

05

NIST: AI Risk Management Framework

Voluntary US framework released 26 January 2023; current page checked 26 September 2026. Not a new statutory obligation.

06

IMDA: updated Model AI Governance Framework for Agentic AI

20 May 2026 update, following January launch. Singapore deployment guidance and case studies, not worldwide legal equivalence.

THE AI INSTITUTE

Make the decision.
Then measure what changes.

Research · Readiness · Capability

theai.institute