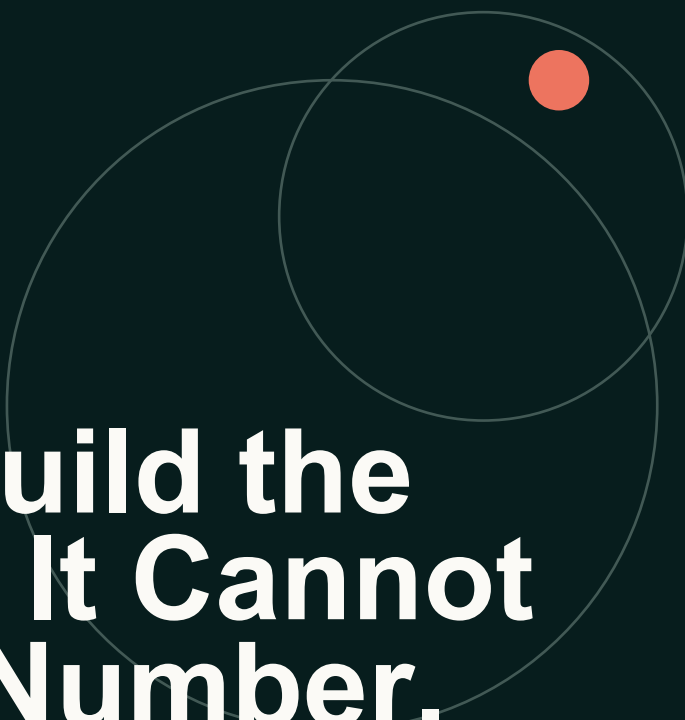


THE AI INSTITUTE / OPERATING INTELLIGENCE

CAPABILITY BRIEFING · 19 AUG / 2026



AI Can Build the Workbook. It Cannot Own the Number.

Whole business artefacts are becoming possible inside everyday tools.
Accountability should remain exactly where the consequence lands.

Boards, C-suite and senior management
Research period: April 2026 – August 2026
First published: 19 August 2026
Current to 19 August 2026

PUBLICATION RECORD

An Institute thesis and decision framework

Use AI to produce reviewable business artefacts, but keep a named human owner for every consequential number, approval and release decision.

WHAT LEADERS SHOULD TAKE FROM THIS PAPER

- 01Move the first agent experiments from chat answers to bounded, editable work products that an expert can inspect.
- 02Keep named human ownership at every consequential number, approval, posting, filing, payment and release decision.
- 03Measure correction effort, professional acceptability and time to approval—not only whether the agent completed the task.
- 04Include the real work surface, data, permissions, comments, formulas and exception path in the test.
- 05Run one 30-day draft–inspect–decide experiment before expanding authority or reducing review.

Research method

This capability briefing compares official productivity-suite announcements with two end-to-end spreadsheet workflow benchmarks, an enterprise workflow perspective and the newest arXiv AI release batch. Supplier capabilities are treated as claims with staged availability, and research results as preprints rather than production proof. The Institute synthesis translates the convergence into a non-technical operating boundary.

First published 19 August 2026 · Updated 19 August 2026 · Research period April 2026 – August 2026 · Version 1.0 · Suggested citation: The AI Institute, *AI Can Build the Workbook. It Cannot Own the Number.* (2026).

CONTENTS

Inside this paper

01	Move the experiment into real work—without moving accountability The executive decision
02	The work surface is becoming the agent interface What has become possible
03	A complete-looking workbook is not a reliable workbook What remains unproven
04	Review must become part of the design, not a final glance The control that changes
05	The boundary travels; the acceptance standard is local Global application
06	Choose one artefact that is expensive to assemble and cheap to challenge 30-day experiment
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.

Move the experiment into real work—without moving accountability

AI is beginning to produce the objects on which organisations actually run: spreadsheets, reports, plans, presentations and structured records. Google has announced natural-language creation and editing of entire spreadsheets, interactive canvases over business data and fully editable presentations grounded in workplace context. Some capabilities are generally available; others remain preview or staged rollout. The direction matters even where the exact feature is not yet available.¹²

For leaders, this is more consequential than a better chat answer. A complete artefact can enter a forecast, a customer review, a board pack or an operating decision. It can also look finished before its dependencies have been checked.

Approve the capability at the draft boundary. Let AI assemble and revise a reviewable work product. Keep a named human owner for the number, the professional judgement and the release. The near-term operating model is draft, inspect, decide.



AI may build the workbook. The accountable executive or professional still owns the number.

The work surface is becoming the agent interface

The shift is from asking for an answer to asking for a usable object. An agent can gather material from files, email, chat and external systems; construct a workbook or report; apply a template; respond to comments; and leave the result in an editable format. That reduces the hand-off between an AI response and the software in which people complete the work. ¹

This matters because much knowledge work is not a single question. It is a chain of collection, calculation, formatting, challenge, correction and communication. Moving AI into the work surface can compress setup and iteration while preserving an object that colleagues already know how to review.

AWS has described the workflow rather than the application as an emerging unit of software delivery. That is a useful management frame: start with a business outcome and its complete workflow, then decide which steps an agent can draft, which steps a professional must inspect and which decision remains human. ⁵

A complete-looking workbook is not a reliable workbook

SpreadsheetBench 2 tests 321 realistic generation, debugging and visualisation tasks built from business data, financial reports and corporate filings. Each task averages 11.8 worksheets and 593.5 cell changes. The paper reports that the best tested model achieved 34.89% overall task accuracy, while debugging accuracy was 12%. The dominant failures included insufficient inspection and selection of the wrong target cells. ³

WorkstreamBench reaches a similar conclusion in finance. The strongest agents produced more professional-looking outputs, but frequently fell short of professional standards as tasks extended beyond a few chained calculations. The benchmark evaluates accuracy, formulas and formatting because a finance artefact must be correct, readable and modifiable by other people. ⁴

These are preprint benchmarks, not a measure of every commercial product or configured workflow. They are strong enough to reject autonomous trust. They also point to a productive use: if expert review is already part of the process, a draft may create value before the system is reliable enough to own the outcome.

EXHIBIT / THE OPERATING BOUNDARY

DRAFT	INSPECT	DECIDE
AI assembles Collect, calculate, structure, format and revise a work product. Bounded agent role	Professional checks Trace sources, formulas, assumptions, exceptions and presentation. Named reviewer	Human owns Approve, post, file, pay, release or act on the consequential number. Accountable owner

Review must become part of the design, not a final glance

A reviewer needs more than the finished file. The test should preserve source links, assumptions, formula lineage, changes, agent actions and unresolved exceptions. It should make material uncertainty conspicuous and route decisions that exceed a threshold to the accountable professional.

The newest arXiv batch is moving in this direction. Quipu proposes governed, time-aware provenance for agent-written knowledge. Other new preprints examine coordination, long-running execution, policy-preserving action and the hidden cost of compressing an agent's context. They are not production evidence. Together they show why the artefact, history and operating path matter as much as the final answer.

678

Measure the reviewer as well as the agent. Correction time, missed dependencies, false confidence, exception demand and time to approval determine whether the draft reduces work or simply relocates it. Keep a clean prior process during the experiment so the comparison remains visible.

The boundary travels; the acceptance standard is local

The draft–inspect–decide pattern works across markets. The content of inspection does not. Accounting conventions, professional standards, language, document structure, data-residency requirements, employment practices and sector rules change what a reviewer must verify.

Product availability also differs. Google describes US and EU data controls, client-side encryption and administration tools, with additional countries planned. Buyers should verify the exact models, actions, retention, processing location and audit features available in each country rather than assuming one global product name creates one operating condition. ²

Australia offers a useful comparison for organisations buying global productivity suites under local privacy, sector, employment and professional obligations. It is not a proxy for Asia-Pacific. In lower-connectivity or multilingual settings, a lighter workflow, local template or more human assembly may outperform the latest integrated feature.

Choose one artefact that is expensive to assemble and cheap to challenge

Select a recurring, non-critical work product: a management workbook, operating report, planning pack or customer-review draft. Use real structure with protected data or an approved safe dataset. Name the owner, reviewer, baseline time, professional acceptance criteria and actions the agent is not allowed to take.

Run the complete workflow. Measure first-draft time, correction effort, error severity, reviewer confidence, time to approval and whether another professional can modify the result. Record where the agent lacked context and which failures were difficult to detect.

Expand when the draft reduces total effort and the owner can still explain the result. Redesign when review becomes the new bottleneck. Stop when a polished output hides dependencies the professional cannot efficiently reconstruct. Do not expand authority simply because the file looks finished.

The 30-day request — one real artefact, one named owner, one explicit no-action boundary and one measured review load.

RESEARCH RECORD

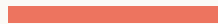
Method and limitations

Method

This capability briefing compares official productivity-suite announcements with two end-to-end spreadsheet workflow benchmarks, an enterprise workflow perspective and the newest arXiv AI release batch. Supplier capabilities are treated as claims with staged availability, and research results as preprints rather than production proof. The Institute synthesis translates the convergence into a non-technical operating boundary.

Limitations

Product features, availability, pricing and administration controls change quickly. Spreadsheet benchmarks do not cover every commercial configuration, language, sector or professional standard. Preprint results have not necessarily been peer reviewed. A human review requirement does not by itself ensure correctness; the workflow, reviewer capacity and local obligations must be tested.



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: 19 August 2026. Review cycle: six months, or earlier after a material change in the underlying facts, standards or policy.

REFERENCES

Evidence behind the thesis

-
- 01 Google Workspace, Introducing Workspace Intelligence**
- Official announcement dated 22 April 2026; whole-spreadsheet and editable-presentation creation grounded in workplace context.
-
- 02 Google Workspace, 10 more announcements at Cloud Next 2026**
- Official announcement dated 22 April 2026; interactive Sheets canvas, agent access, governance and sovereign controls with staged availability.
-
- 03 arXiv, SpreadsheetBench 2**
- Preprint submitted 29 June 2026; 321 end-to-end business spreadsheet tasks and reported accuracy below 35%.
-
- 04 arXiv, WorkstreamBench**
- Preprint submitted 21 May 2026; professional finance artefact criteria and complexity-related performance limits.
-
- 05 AWS, The New Unit of Software Delivery: The Workflow**
- Official specialist article dated 25 November 2025; workflow as the unit of delivery, testing and value.
-
- 06 arXiv, recent Artificial Intelligence submissions**
- Official repository listing; 18 August 2026 batch contained 465 entries with mixed review status.
-
- 07 arXiv, Quipu**
- New v1 preprint submitted 18 August 2026 on governed time-aware knowledge for agent-written records.
-
- 08 arXiv, What Does Context Compression Cost an Agent?**
- New v1 preprint submitted 18 August 2026 on interaction costs hidden by final task-completion metrics.
-

THE AI INSTITUTE

Make the decision.
Then measure what changes.

Research · Readiness · Capability

theai.institute