ZAPHAN
DecisionGraph Weekly™
Enterprise Decision Intelligence

Agentic AI: Fund the Outcome, Not the Pilot

14–19 September 2026 | Enterprise Decision Intelligence

60-Second Executive View

Fund AI where accepted business outcomes justify the full cost and risk. Faster drafting alone is insufficient; measure the work required to reach an acceptable result.

A successful pilot is not yet an investment case. Before the next funding or production gate, compare full cost per accepted business outcome with the existing process and simpler alternatives.

ACT now: name the owner and define the outcome, baseline, full costs and control boundaries.

EVALUATE before expanding: compare the current process, the proposed AI approach and eligible native or simpler alternatives.

DECIDE and monitor: record SCALE, LIMIT, DEFER or STOP, with a review trigger. The decision conditions follow the evidence tables.

Recent infrastructure reports show narrow operating improvements. A new native coding option changes the comparisons worth making. Neither proves that current AI delivers a transferable financial return.

What matters: Scaling too early can hide rework and risk. Leaving pilots open-ended can waste budget and delay a validated improvement. Measure the outcome, include the complete cost and make the disposition reversible.

Current posture: No material posture change verified this week. Act on the gate; evaluate the workload remains the direction. This issue sharpens measurement and alternative comparisons. It adds neither a blanket scale recommendation nor a new proven market trend.

What this looks like in practice

Illustrative example only. These figures are hypothetical and are not METR findings. Assume comparable software changes meeting the same quality and security requirements.

Human effort per accepted change	Current process	AI: faster drafting, more correction	AI: lower total effort
Write the change	4 hours	2 hours	2 hours
Review, test and correct	2 hours	5 hours	3 hours
Total human effort	**6 hours**	**7 hours**	**5 hours**
Result versus current process	Baseline	**1 extra hour**	**1 hour released**
Funding implication	Compare against this	No labor-saving case demonstrated	Potential productivity benefit; check full costs and useful deployment of released capacity

Lower total effort is evidence of a productivity benefit. Funding still depends on tool and operating costs, quality, risk and how the released capacity creates business value.

One hour released is capacity—not automatically cash saved. Finance and the business owner should identify how it creates value: more useful work, shorter delivery times, avoided overtime or a demonstrable reduction in expenditure.

AI writes faster. Does the team finish better?

Before expanding AI coding tools, require evidence that they deliver accepted work at a justified total cost and risk. Faster drafting is useful only if review, testing and correction do not consume the benefit.

METR's three studies give leaders three practical lessons:

What the research found	What leaders should do
Measured slowdown: In a 2025 experiment, 16 experienced developers took 19% longer across 246 tasks when allowed to use AI. They nevertheless believed AI made them faster.	**Check completed work, not perceived speed.** Measure the effort required to reach the same acceptance standard.
Uncertain current effect: A later experiment involving 57 developers and more than 800 tasks could not reliably establish the size of the current benefit because of participation, task-selection and measurement problems.	**Test today's tools on your work.** Include representative tasks, failed attempts and corrections.
Reported value: A 2026 survey of 349 technical workers found median self-reported work-value gains of 1.4–2× across questions. These were perceptions, not verified financial returns.	**Use employee feedback to find promising applications.** Validate the benefit before including it in a funding case.

Sources: METR's July 2025 experiment, February 2026 experimental update and May 2026 survey.

The studies examine different populations, tasks and measures; their results cannot be combined into one productivity score. The actions and illustrative comparison on the preceding page are ZAPHAN's interpretation.

ACT now: establish the decision inputs

Name the business sponsor and define the accepted outcome, baseline, complete cost boundary and permitted data and actions. Require one decision record with four accountable inputs.

Owner	Question to answer	Evidence to bring
Business owner	What useful outcome are we buying?	A defined business outcome, acceptance criteria and evidence that the output meets them
Finance and operations	Does the value justify the full cost?	A comparable baseline, full cost per accepted outcome, review and rework effort, and a credible benefit mechanism
Technology and architecture	Is this the right way to deliver it?	Comparison with the existing process, eligible native capabilities and simpler alternatives—including integration, operating and switching costs
Risk and security	Can it operate within acceptable limits?	Permitted data and actions, quality checks, monitoring, recovery and rollback evidence

The business sponsor names the outcome and disposition. Finance validates the cost boundary and benefit mechanism. Technology/operations test quality, integration and recovery; the risk owner validates permitted data and actions.

Bring baseline costs and output, volume and workload mix, acceptance rules, review/rework, complete implementation and recurring costs, contracts, authority limits, switching costs and the decision horizon. Without them, a dollar return, payback period or cost of delay cannot be defended.

EVALUATE before expanding

Compare the current process, proposed AI approach and eligible native or simpler alternatives on accepted output, total human effort, full cost, quality, risk and switching cost. Include review, rework and unsuccessful attempts.

Agree the evidence window and acceptance thresholds before testing. Use representative work and record the limits of the result. A productive coding task or lower infrastructure cost does not establish a transferable enterprise return.

DECIDE: fund the demonstrated outcome

The business sponsor owns the disposition, with finance, technology and risk challenge.

- **SCALE:** representative results show that the benefit justifies full costs, with quality and controls passing.
- **LIMIT:** the benefit holds only for certain tasks; expand within that boundary.
- **DEFER:** a defined test can resolve a specific missing answer.
- **STOP:** the investment case fails or risk cannot be acceptably controlled.

What should reopen the decision?

Agree the evidence window, acceptance thresholds and stop conditions before committing further funding.

- **Economics fail:** if value no longer justifies full cost after accounting for quality, workload mix and human effort, reassess SCALE, LIMIT, DEFER or STOP.
- **A control fails:** stop the affected activity when a prohibited action occurs or recovery fails.
- **The comparison changes:** reassess when prices, workload scope, provider capabilities or switching costs materially change.
- **Evidence remains missing:** DEFER only when a bounded test can answer the missing question. Do not extend a pilot without a defined decision point.

Track accepted output, total human effort, review and rework, defects, control events and full production cost against the agreed baseline.

What still needs testing: We still need to test whether optimizing around accepted business outcomes improves results in the intended workflow. Measure it before treating it as a proven benefit.

No new general regulatory requirement or vendor ranking is asserted.

Intended results

Decision-ready economics, quality/control evidence and a recorded disposition with a review trigger. These are prospective measures requiring a baseline and owner-agreed targets.

The complete flagship DecisionGraph

Evidence: bounded operating cases, a native product release, mixed developer research and cost-allocation guidance.

Interpretation: useful alternatives and measurement questions exist; no pooled economic effect is established.

Enterprise exposure: pilots can hide total cost, rework, unsafe authority and opportunity cost.

Business consequence: premature scale and indefinite delay can both destroy value.

Decision: compare the workload with its baseline and simpler/native alternatives on useful accepted output, full cost and risk.

Recommendation: act on measurement; evaluate reversibly; assign SCALE, LIMIT, DEFER or STOP.

Expected result: a defensible decision, with measured economics and quality—not a promised saving.

Revisit when: normalized benefits disappear, a control fails, prices/scope change, or a better alternative emerges.

Illustrative example—not a reported study result

An AI pilot produces draft customer-service replies that staff can use. That demonstrates useful capability under the pilot’s conditions.

Before expanding it, the business owner defines what counts as an acceptable reply. Finance and operations include the cost of checking, correcting and handling unsuccessful replies. Technology compares the pilot with the current process and an eligible native service. Risk checks what customer information the system may access and which actions it may take.

The question becomes: **Which option delivers acceptable customer outcomes at a justified total cost and risk?**

ZAPHAN’s interpretation: make that comparison before the pilot’s technology choices become recurring production commitments. The FinOps survey does not establish which option will win or how much money it will save.

Why this matters now

Sangfor and China Merchants Bank report operating-cost/utilization improvements in specific environments. GitHub has announced a native cost/quality/latency choice for Copilot auto model selection. These are reasons to examine workload design and available alternatives—not reasons to import another organization's percentage into your ROI model.

The action window is your next funding or production-admission gate. No universal time-to-value is established.

What the evidence changes—and what it does not

Lower operating cost can matter without proving a profitable workload. Sangfor's CNCF-hosted case reports monthly external invocation spending falling from RMB 400,000 to RMB 200,000. China Merchants Bank's case reports accelerator utilization rising from 35% to more than 60%, and inference cost per million combined input/output tokens falling by more than 60% for comparable models and services. These are two organizations' reported operating results, not independently verified, like-for-like buyer returns. Workload mix, quality, period and the complete cost boundary must be checked before using them in a funding case. Sangfor case; China Merchants Bank case. Case summaries adapted by ZAPHAN from CNCF-hosted material under CC BY 3.0; no endorsement is implied.

A native option changes what deserves comparison, not what deserves automatic adoption. GitHub's September 14 release documents cost, quality and latency priorities for Copilot auto model selection. For eligible coding workflows, test that option before commissioning a custom routing layer. Availability, output quality, controls and full cost in your environment remain to be verified. This is a product development, not evidence of savings or a market-wide trend. GitHub release.

Supporting evidence: where FinOps can influence the choice

The State of FinOps 2026 associates senior-executive engagement with greater reported influence by FinOps practitioners over technology selection. It compares engagement at VP level and above with director-level-only engagement.

Technology decision	Executive Summary	Detailed discussion
Which cloud services to use	53% versus 12%	53% versus 24%
Which cloud provider to choose	47% versus 8%	47% versus 16%
Whether to use cloud or a data center	28% versus 6%	28% versus 12%

The report does not reconcile these differences. They are comparisons of reported influence—not savings, returns or proof of causation. ZAPHAN preserves both versions and derives no ROI multiplier from them.

Publisher clarification remains needed.

ZAPHAN's practical application: bring FinOps into the production decision early enough to challenge costs and alternatives before service, provider and infrastructure commitments are made.

Source: State of FinOps 2026. Comparison compiled by ZAPHAN from FinOps Foundation material under CC BY 4.0.

Supporting context

AI value should be assessed against a defined business outcome. FinOps Foundation guidance connects technology costs with outcome measures, helping teams distinguish cheaper processing from better business performance.

The OECD's review of experimental research finds that generative AI's benefits depend on the task, user experience and implementation context. Those findings do not establish a transferable financial return.

For this decision, ZAPHAN proposes measuring total in-scope cost per accepted outcome against a baseline, while tracking quality, review and rework. Treat productivity improvements as evidence to evaluate—not as automatically realised financial savings.

Sources: FinOps Foundation, Unit Economics and FinOps for AI; Calvino, Reijerink and Samek (2025), The effects of generative AI on productivity, innovation and entrepreneurship, OECD Artificial Intelligence Papers No. 39. Summarised and adapted by ZAPHAN under CC BY 4.0.

This is an adaptation of an original work by the OECD. The opinions expressed and arguments employed in this adaptation should not be reported as representing the official views of the OECD or of its Member countries.

GAO's selected federal-acquisition review supports scrutiny of cost, expertise and procurement lessons—not an all-market return estimate. GAO.

The evidence supports a careful test of each workload. It does not establish a universal productivity gain, a preferred vendor, or a standard payback period.