

Who Checks the AI? The Review Capacity Constraint

An executive guide to accepted throughput, expert attention and the economics of oversight

Enterprise AI Adoption Gap · Research brief 09
October 5, 2026

© 2026 Eldris Inc. All rights reserved.

Executive brief

AI can accelerate production without accelerating the organization's ability to use the result. Drafts, analyses and recommendations still need judgment, correction or approval. When that work depends on a limited pool of experts, more generation can create more waiting rather than a proportional increase in completed work.

The useful output is work accepted for its intended purpose at the required quality. Draft volume measures an upstream activity. A review checkbox measures a procedure. Neither establishes that the workflow delivers more worthwhile results or that important errors are detected.

This report is for enterprise workflow owners, AI program leaders and providers deciding how much automation to introduce into a team whose outputs require expert review. Its bounded example is a team preparing routine client-support replies, with trained reviewers approving every reply before release.

The central recommendation is to measure review as part of the operating system: available expert time, handling effort, queue age, corrections, rejected work and errors found after acceptance. Compare those measures with the current workflow before expanding generation capacity.

Three principles follow:

- Measure the complete path from request to accepted outcome, including repeated review and downstream correction.
- Treat review capacity and review effectiveness as separate requirements. A faster reviewer who misses important errors has not solved the constraint.
- Allocate automation around the limiting resource. Improving the evidence presented to a reviewer may be more useful than producing additional drafts.

A capacity model can help locate the constraint and identify the next measurement. It cannot establish reliable throughput without workflow data, nor can it justify a particular review policy on its own.

The unit of value is accepted work

For the support team, a draft becomes useful when it correctly answers the eligible request and can be released under the organization's rules. The relevant outcome might be resolved customer cases within an acceptable service time. Counting drafts does not establish that result.

Distinguish four events: generated, submitted for review, accepted and released. Add later correction or reopening where relevant. A generated answer may never be submitted; an accepted answer may still contain a missed error; a released answer may fail to resolve the customer's request.

Record effort across that path. A producer may save time while a reviewer spends more time identifying plausible but unsupported statements. Alternatively, better drafting may make review easier. Both are empirical possibilities. The net effect depends on the workload and the operating arrangement.

Public productivity findings reinforce the need for contextual measurement. METR's early-2025 randomized study of 16 experienced open-source developers across 246 tasks found that AI access increased completion time by 19%. That is a historical result for a particular coding setting, not an estimate of review costs or a current claim about every AI tool. [METR study](#).

Decision implication: Choose the outcome the business needs and measure the full effort needed to achieve it. Do not presume either a gain or a review penalty from generation speed alone.

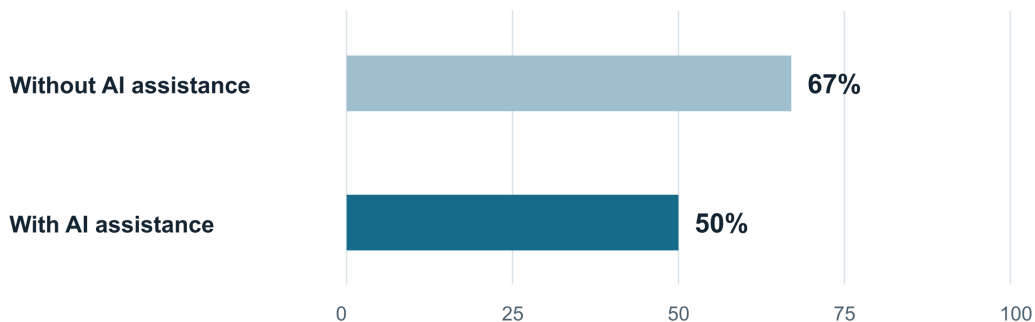
Expertise is part of the capacity

Review availability is more than a headcount. Reviewers need the relevant knowledge, access to source material, time to investigate uncertainty and authority to reject or escalate an output.

Anthropic's randomized coding study reported average immediate quiz scores of 50% with AI assistance and 67% without it. The trial recruited 52 mostly junior software engineers learning an unfamiliar Python library. It did not establish long-term skill loss or measure enterprise review capacity. [Anthropic study](#).

Immediate understanding differed in a coding study

Mean quiz score · Unfamiliar Python library · January 2026



Published means from an immediate assessment; no uncertainty bars shown.
52 recruited engineers. This is not a measure of enterprise review performance
or evidence of long-term skill loss.

Source: Anthropic, January 29, 2026.

© Eldris Inc

Figure 1. Published mean quiz scores, not error-detection rates. The study concerns immediate understanding after learning a library. The result motivates evaluating reviewer competence; it does not supply a staffing ratio or an enterprise forecast.

For the support workflow, the corresponding question is whether reviewers can recognize incorrect policy application and unsupported claims. Familiarity with the assistant's interface does not establish that competence. Neither does confidence in the output.

Assess the expertise required for different case categories. A trained generalist may handle a routine question while a policy exception requires someone with specialized knowledge. Hours from those people are not automatically interchangeable.

NIST's AI RMF calls for defined oversight responsibilities, training and evaluation in relevant conditions. This is governance guidance, not proof that a particular reviewer arrangement works. [NIST framework](#).

Decision implication: Identify which expertise each task needs. Measure whether the proposed reviewers can detect the material errors, not merely whether they can process the interface.

A simple capacity example

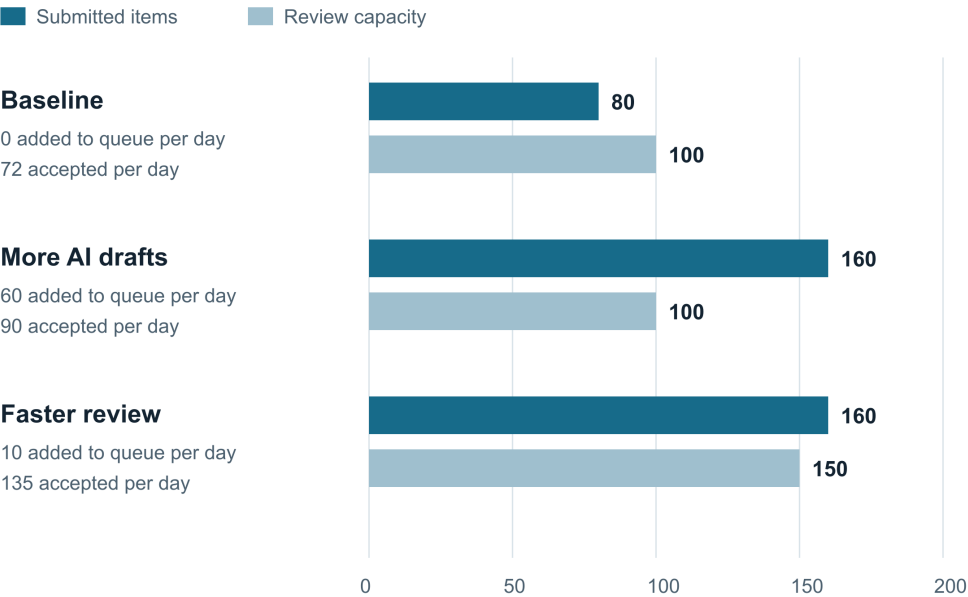
Consider an original Eldris arithmetic example. It describes assumed operating conditions, not observed company data, simulated participants or a validated forecast.

Assume a team has 600 usable reviewer-minutes per day: for example, two reviewers each providing five hours after other commitments. Every submitted item receives one review. Initially, review takes six minutes per item. Accepted items are 90% of reviewed items; rejected items leave this simplified process without resubmission. No other stage limits release.

Before expanded AI drafting, the team submits 80 items per day. After expansion, it submits 160. A third scenario keeps that submission rate but reduces review time to four minutes without changing acceptance or quality. That improvement is an assumption to test, not an expected effect of AI.

Drafts can outrun review capacity

Illustrative arithmetic · Daily rates · Not observed enterprise results



Assumptions: 600 usable reviewer-minutes daily; 6, 6 and 4 minutes per review.
90% of reviewed items accepted; every item reviewed once; rejected items exit.
Accepted output = 90% × the smaller of submissions and review capacity.

Original Eldris analysis. Average-rate example; variability and resubmissions excluded.

© Eldris Inc

Figure 2. Illustrative daily rates. Review capacity equals available reviewer-minutes divided by minutes per item. Expected accepted output equals 90% of the smaller of submissions and review capacity. No statistical uncertainty or real company result is implied.

At six minutes per item, review capacity is $600 / 6 = 100$ items per day. The baseline processes all 80 submissions and accepts 72. Expanded drafting can process 100 of 160 submissions and accept 90, leaving 60 new items unreviewed that day. Draft submissions double, but expected accepted output rises by 25%.

At four minutes per item, capacity becomes 150 and expected accepted output becomes 135. Ten additional items remain unreviewed per day. The improved scenario delivers more accepted work but still has an accumulating queue under these assumptions.

These are daily processing limits, not predictions of waiting time. Actual queues depend on arrivals, variability, priorities, absences and repeated handling. Even when average capacity exceeds average arrivals, bursts can create waiting. Running exactly at the average limit provides no margin for those disruptions.

The model also assumes that acceptance is justified. A rising approval rate caused by weaker checking would make the arithmetic look better without establishing better outcomes.

Decision implication: Compare submitted work with effective review capacity before scaling generation. Use the model to locate a missing measurement, not to treat all reviewer-minutes as equally productive or all approvals as correct.

Measure the burden the first review misses

A single average review time can understate the work required for a usable result. Track initial review, correction, repeat review and specialist escalation. Include cases rejected before release and cases repaired afterward.

Keep elapsed time separate from active handling time. A reply can take only a few minutes to check but wait a day for an available expert. Waiting matters to the customer; handling effort matters to staffing and cost. Both can affect the value case.

Measure effort per accepted outcome using all relevant work, not just time spent on the items that passed. Otherwise, rejected drafts and repeated attempts disappear from the calculation. Capture work shifted to other teams where it is material.

Estimate usable review time from actual commitments. Meetings, urgent exceptions, training and other responsibilities reduce availability. A nominal eight-hour workday is not eight hours of uninterrupted expert review.

Segment where differences matter. Simple requests may have short checks and high acceptance; unusual requests may need extended investigation. Changing the case mix can improve the headline average while leaving the most consequential queue unresolved.

Decision implication: Cost the complete review-and-correction process per accepted outcome. Report elapsed service time alongside labor effort and identify where work moves.

Review speed and review quality are different outcomes

Acceptance is an organizational decision. Accuracy is a property to assess. Those measures can diverge when a reviewer misses an error, uses incomplete references or approves under pressure.

Test the review arrangement using cases with an independently established assessment basis. Include material errors that resemble realistic failures, and also correct outputs so that unnecessary rejection is visible. Evaluate missed errors, appropriate escalation and excessive correction, as well as speed.

Do not require every reviewer to catch every conceivable error in every task. Define what must be checked for the intended consequence, and evaluate that requirement. An unclear obligation makes staffing estimates and quality measurements difficult to interpret.

Automation can assist review by checking deterministic requirements, retrieving authoritative material or highlighting changed content. Evaluate those aids separately: a second model's agreement is not independent proof of correctness, and an appealing explanation may not support the underlying claim.

For the support team, a source-linked policy passage could reduce the effort of checking an answer. The team still needs to verify that the passage is current, relevant and correctly applied. The product's ability to display evidence is different from the reviewer's ability to use it effectively.

Decision implication: Approve a faster review design only when its detection and outcome evidence also meets the workflow's requirements. Do not count weaker scrutiny as a productivity gain.

Choose the intervention at the constraint

The next investment depends on why work is waiting or failing review. The map below is original Eldris analytical guidance, not a validated decision rule.

Invest at the measured constraint

A decision guide for one workflow with defined acceptance standards

OBSERVED CONSTRAINT	INTERVENTION TO COMPARE	EVIDENCE TO REQUIRE
Drafts exceed capacity	Reduce low-value submissions; add qualified review time.	Backlog and accepted output improve?
Checking takes too long	Improve evidence access and presentation.	Effort falls without weaker error detection?
Rework dominates	Correct upstream inputs and repeated failures.	Total handling effort falls?
Specialist queue grows	Match routing to expertise and available capacity.	Delay and quality improve?
Important errors escape	Repair oversight before expanding volume.	Material errors detected at required quality?

Original Eldris guidance; not a validated decision rule.
Compare quality, cost and service time together before expanding.

© Eldris Inc

Figure 3. Different observed constraints suggest different comparisons. Validate the proposed intervention against accepted output, quality, total effort and service time.

When drafts exceed capacity, consider reducing unnecessary submissions, prioritizing valuable cases or supplying additional qualified reviewer time. More draft generation alone does not expand the limiting stage. If demand for completed work is already satisfied, additional throughput may have little business value.

When checking requires extensive investigation, compare evidence presentation, access to current references and clearer task scope. Better drafting may help, but the product should demonstrate the reduction in checking effort without reducing detection.

When rework dominates, investigate input quality, source coverage, task suitability and repeated failure patterns. Increasing reviewer headcount may process the symptoms while leaving a preventable upstream problem intact.

When specialist work accumulates, preserve access to the relevant experts and examine routing. Sending more complex cases to a general queue can increase apparent capacity while moving risk or delay elsewhere.

When important errors escape, investigate review effectiveness before treating speed as the primary objective. Expansion requires evidence that the proposed arrangement can meet the required quality, not merely a larger processing count.

Sampling or selective review may be appropriate for some workflows, but it changes the operating proposition. Define eligible cases and evaluate the consequences of errors passing unchecked. Do not transfer evidence from mandatory review into an unreviewed workflow without a new assessment.

Decision implication: Compare interventions that address the observed constraint. The cheapest improvement in generation may be less valuable than better evidence, fewer unsuitable submissions or qualified review capacity.

Evaluate expansion using real workflow evidence

Start with one team and a clearly defined unit of work. Establish the baseline case mix, demand, accepted outcomes, service times, review effort and important downstream errors.

Then evaluate the proposed change with a comparison suited to the workflow. Randomized assignment may be feasible; where it is not, document how workload, staffing or other changes limit interpretation. Keep the quality requirement consistent. An apparent improvement created by accepting a lower standard needs to be reported as that tradeoff.

Follow each item through its review history. Record the evaluated system and source versions, who handled it and what happened after acceptance. Use proportionate instrumentation: collect the evidence needed for the decision and protect sensitive customer and employee information.

A useful decision criterion is whether the change increases worthwhile accepted output or improves service at acceptable quality and total cost. An expansion that increases backlog, creates hidden specialist work or raises serious errors has not demonstrated the same benefit as one that improves the complete workflow.

Choose an observation period that captures normal variation and enough later outcomes to reveal relevant corrections or reopening. Report uncertainty and rare-event limits. A short run with no detected severe errors cannot establish that severe errors are absent.

Task studies can compare review designs and detection behavior. They cannot establish organizational capacity without evidence about workload, staffing commitments and operating variation. This is a strong setting for commissioned research using actual workflow records, supplemented by controlled review tasks.

Decision implication: Collect the data that can change the expansion decision. A new general adoption survey will not resolve a queue caused by a specific team's review arrangements.

What to require before scaling generation

Prepare a short decision record:

1. **Workflow:** eligible cases, unit of work, quality requirement and actual demand for completed outcomes.
2. **Capacity:** qualified reviewer availability, handling time, variation and specialist dependencies.
3. **Complete effort:** corrections, repeat review, rejections and downstream repair per accepted outcome.
4. **Effectiveness:** material errors detected and missed, unnecessary rejection and escalation performance.
5. **Comparison:** accepted throughput, queue age, service time and total cost before and after the proposed change.
6. **Decision:** what expands, what remains bounded and what deterioration will trigger reassessment.

For enterprise leaders, this record connects automation to a resource decision: invest in generation, reviewer capability, better evidence, upstream quality or a narrower workload. For providers, it identifies product features whose value can be demonstrated in the buyer's operating conditions.

The central question is: “**If we produce twice as much, who can determine what is worth using?**” Answer it with measured capacity and quality. The benefit belongs to the complete workflow, including the expertise that makes the output usable.

Method and limitations

This brief selectively reviews primary sources checked on October 5, 2026. It is not an original workflow study or an estimate of enterprise review burden. METR's result concerns early-2025 coding tools and experienced developers; its subsequent update discusses challenges in measuring newer tools. The result is historical and supplies no current universal productivity estimate.

Anthropic's study concerns immediate learning and comprehension in a coding task, not measured long-term loss of expertise or reviewer staffing. NIST supplies governance guidance. These sources motivate contextual measurement; none establishes that review is the bottleneck in a particular enterprise.

The capacity arithmetic, support-workflow example, intervention map and decision record are original Eldris analysis. All model inputs are stated assumptions. The model excludes repeated review, arrival variation and downstream bottlenecks; it produces illustrative rate limits, not queueing predictions, verified business results or staffing recommendations.

A deployment decision needs actual workload, review-cost and outcome data. Acceptance must be assessed against the required quality, and capacity must reflect qualified availability. Model precision cannot substitute for those measurements.

Sources

1. **METR. Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity.** July 10, 2025. Randomized study of 16 developers and 246 tasks; historical completion-time finding. [Study and paper](#).
2. **METR. We are Changing our Developer Productivity Experiment Design.** February 24, 2026. Follow-up on measurement and selection challenges when evaluating newer coding tools. [Update](#).
3. **Anthropic. How AI Assistance Impacts the Formation of Coding Skills.** January 29, 2026. Randomized coding study; mean immediate quiz scores of 50% with AI and 67% without. [Research and paper](#).
4. **NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1.** January 2023. GOVERN 2.1-2.3 and GOVERN 3.2 address responsibility, training and oversight; MEASURE addresses evaluation. [Framework](#).