

The Real Cost of AI Is the Cost of a Result You Can Use

By Junior Williams

Draft for review · 14 October 2026

A cheap answer can be expensive to use. If someone spends twenty minutes checking it, correcting it and putting it into the right system, that work belongs in the cost calculation.

Think about ordering a part for a repair. The price matters, but so do the time spent fitting it, the adjustments it needs and whether the repair holds. Evaluating an AI workflow requires the same attention to the finished job.

The question I would start with is: what does it cost to produce one result that meets the requirements for use?

Define what counts as finished

Take a workflow that extracts information from documents. A completed result might require the right values, a reference to their location in the source, the correct destination record and review by an authorized person. A response that sounds plausible has not necessarily met any of those conditions.

Set the acceptance criteria before comparing tools. Decide which errors are tolerable, which require correction and which must stop the workflow. Missing a formatting preference and assigning a value to the wrong customer deserve different treatment.

The criteria should also specify what happens when information is missing. A result that correctly asks for clarification may be more useful than a completed field based on a guess.

Count the work around the model

For a defined period, use this calculation:

Cost per accepted result = total cost of running the workflow ÷ results that meet the acceptance criteria.

Include model and platform charges, review, corrections, integration, support and the handling of failed attempts. Count retries in the usage bill and the time they consume. Keep review and correction time separate so the same minutes are not counted twice.

Setup costs need a stated allocation. If you spread an initial integration cost over twelve months, record that assumption. Report the upfront cash requirement separately; spreading a cost does not make it disappear.

Here is a deliberately simplified example. These are invented figures to explain the method, not observed results or a savings promise. Both options process the same 1,000 requests in a month, and all 1,000 meet the same quality requirements after review and correction.

Monthly cost	Option A	Option B
Model usage, including retries	\$100	\$300
Review at \$60 per hour	\$1,800 — 30 hours	\$900 — 15 hours
Correction at \$60 per hour	\$600 — 10 hours	\$300 — 5 hours
Ongoing integration and support	\$400	\$400
Initial setup allocated to the month	\$300	\$300
Total	\$3,200	\$2,200
Cost per accepted result	\$3.20	\$2.20

The setup allocation assumes \$3,600 spread over twelve months for each option. Support excludes the review and correction work already listed.

Option B has the higher model bill and the lower total cost in this example. Different review times, volumes or support requirements could change the result. Test those assumptions before committing to a rollout.

Make the comparison fair

Use representative work, including awkward documents, missing information and cases that require escalation. Compare against the current process using the same acceptance criteria. Record completion time, serious errors, reviewer effort and the share of requests that finish successfully.

If some requests remain unresolved, show that number beside the cost. Include manual completion costs when people must finish those requests. A low cost per successful result can conceal a workflow that leaves too much work behind.

The labour rate is another assumption. Staff time can have an economic cost without producing an immediate reduction in payroll. Report freed capacity separately from cash savings, and identify what useful work that capacity could support. Do not translate every hour saved into a headcount reduction.

Make a decision the evidence supports

A small pilot should end with a decision: expand the workflow, revise it or stop. Explain the quality requirements it met, the work it still needs and which costs are likely to change at higher volume.

Junior Williams Consulting's proposed AI Workflow Value and Assurance Pilot applies this approach to one defined workflow. The starting point is a clear owner, a measurable baseline and an agreed definition of a result the organization can use.

To discuss one workflow, [book an introductory call](#).

Further reading

The [NIST AI RMF Playbook](#) addresses business value (Map 1.4) and costs associated with errors (Map 3.2). The calculation here is an illustrative management tool, not a NIST-prescribed formula.