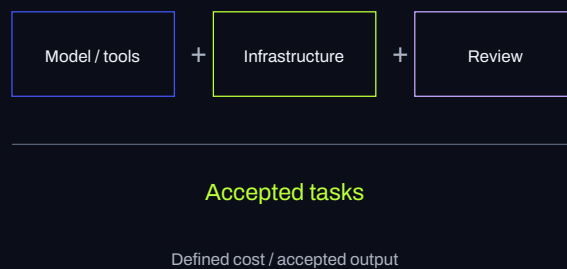


FIELD BRIEFING

Cheap tokens. Expensive work?

Count everything a useful result needs.



What does an accepted AI-assisted result really cost?

Per-token pricing is one component of operating cost. The work also involves preparation, tools, infrastructure, retries, checking and correction. For a workflow decision, measure cost per accepted task alongside quality and latency.

Explanatory schematics and worked examples are illustrative. No live model, observed client results or measured market performance are presented.

Read the mechanism.

Begin with a numerator you can explain and a denominator that matches the outcome. If a trial attempts 100 drafts and 80 meet the acceptance rules, the denominator is 80 accepted drafts. Counting all 100 attempts as successes would hide the cost of failure.

Consider an illustrative calculation, not a provider quote. Assume the trial incurs £12 of model and tool charges, £8 of allocated infrastructure and 120 review minutes at an assumed £30 per hour. Total measured cost is £80; divided by 80 accepted drafts, that is £1 per accepted draft. The example excludes setup and rejected-task recovery unless those costs are added, and does not assert a financial return.

Preparation and checking may dominate a small trial. Track what a person actually does rather than assigning all elapsed time to AI processing. Avoid double-counting work already included in another cost category. Report fixed setup costs separately, then show the volume over which any allocation is spread.

Compare changes while holding acceptance rules stable. A smaller model, shorter prompt or cached context may reduce charges, but it can also affect quality or tool usage. A longer prompt may cost more per call while reducing retries. A cheaper request is not automatically a cheaper completed task.

Caching and batching have conditions. Cache benefit depends on reusable context, supported pricing and actual hit rates. Batching suits work that can wait; it does not replace interactive response requirements. Measure the actual configured service rather than carrying one vendor's current rates into a universal claim.

The practical comparison is with the existing workflow and its required quality. Include review effort, missed work and failure consequences where they can be estimated. Keep uncertainty visible when a short trial cannot establish typical demand or rare-event costs.

Technical references: Anthropic · Cost Optimization; Anthropic · Building Evals. Links and source scope appear in the final section.

Try the idea.

INVENTED TEACHING EXAMPLE

$\text{£80 total} \div 80 \text{ accepted tasks} = \text{£1 per accepted task}$

Illustrative assumptions: £12 model/tools; £8 infrastructure; 120 review minutes at £30/hour, giving £60 review cost. Total £80. Of 100 attempted drafts, 80 are accepted. Cost per accepted task is $\text{£80}/80 = \text{£1}$. Setup and rejected-task recovery are excluded unless added. With zero accepted tasks the ratio is undefined.

Go deeper.

Use a defined accounting boundary

Total cost can include model/tools, infrastructure, labour and allocated fixed costs. State the period and avoid double-counting categories. A rejected attempt belongs in the numerator's cost even though it is not an accepted result.

The zero case matters

With zero accepted tasks, cost per accepted task is undefined, not zero. The lesson deliberately displays that condition. It is useful evidence of a workflow that has not yet produced a qualifying output.

Optimisation has a quality constraint

Compare caching, batching, model choice and context changes against the same acceptance rules. Measure actual usage; provider prices and cache conditions change. Cheaper calls can become more expensive work if they increase retries or review.

Keep the evidence close.

Accepted-task cost

Defined total cost divided by the number of outputs meeting the acceptance rules.

Fixed cost

An expense that does not vary directly with each attempt over the stated range.

Marginal cost

The additional cost of one more unit of work.

Cache hit

Reuse of supported cached context; savings depend on the service and actual reuse.

Sources · accessed 4 October 2026

[Anthropic · Cost Optimization](#)

[Anthropic · Building Evals](#)

Technical and security references support the mechanisms, not our fictional results. KAF's worksheets and examples are original editorial recommendations. No current provider pricing, comparative ranking or legal compliance is asserted. Verify the exact release, configuration and policy for your use.

Practical AI Insights · First editorial edition · 4 October 2026

Cheap tokens. Expensive work? · [Web edition](#)