

Cost per useful AI request worksheet

Include retries, infrastructure and review time, then divide by accepted outputs rather than API calls.

Define the unit

What counts as one accepted output: _____

Who checks it: _____

Measurement window: _____

Include the full boundary

Initial model calls: \$_____

Retry calls: \$_____

Retrieval and external tools: \$_____

Allocated infrastructure: \$_____

Review cost = review minutes / 60 x reviewer hourly rate: \$_____

Total cost = all the above: \$_____

Accepted outputs: _____

Cost per accepted output = total cost / accepted outputs: \$_____

If accepted outputs are zero, the unit cost is undefined. Report the spend and the failure instead of dividing by one.

Worked example, hypothetical

Initial calls cost \$20, retries \$5, retrieval \$10, allocated infrastructure \$15 and review \$50. Total cost is \$100. If 80 outputs are accepted, cost per accepted output is \$1.25. These are invented teaching inputs, not provider prices or customer results.

Audit the assumptions

Record the price source and effective date. Distinguish measured bills from estimates. Record the allocation rule for shared infrastructure. Treat blocked, failed and abandoned requests explicitly. Avoid counting a retry both in the initial-call total and the retry line.

Compare cost alongside acceptance quality and latency. A cheaper API call can produce a more expensive useful result.

Subscribe for the next practical issue

Subscribe to this newsletter

Free download. Email signup is optional and requires confirmation. Teaching examples are hypothetical.