

TECHNOLOGY

When AI usage grows faster than profit

Product teams need an economic model that follows a request through retries, review and accepted completion.

THE EXECUTIVE FINDING

A low inference price can coexist with a high cost to serve. The economically useful unit is a successfully completed customer task.

THREE DECISIONS FOR LEADERSHIP

- 01** Define an accepted task and attribute all material delivery costs to it.
- 02** Compare model routes on the same workload and quality threshold.
- 03** Set retry, review and usage policies before scaling distribution.

THE TRADE-OFF TO EXAMINE

Detailed telemetry helps cost attribution but can expose sensitive prompts or customer information. Capture the minimum information needed, restrict access and agree retention. Cost controls should also preserve service quality for legitimate complex tasks.

HOW GENELLIPSE CAN HELP

Genellipse can instrument governed workflows through Cortex AI, reconcile operational and cost data with GeneData and expose task-level economics through GeneAnalytics. CIO Advisory can connect the architecture to pricing and investment decisions. A first engagement would establish cost attribution and compare a bounded set of delivery options.

THE ANALYSIS IN ONE VIEW

Illustrative cost per accepted task

cost units | Synthetic example



0.12 + 0.08 + 0.05 + 0.30 + 0.15 = 0.70 cost units per accepted task.

Synthetic fully attributed task costs under a defined workload. Currency, pricing and profit are intentionally unspecified.

INTERPRETATION

In the example, inference is 0.12 of a 0.70-unit total. A 50% reduction in inference saves 0.06 units, leaving 0.64 before any quality change. If the cheaper route adds 0.10 units of review, the total becomes 0.74. Evaluate the entire accepted task and use sensitivity analysis before concluding that a lower model price improves margin.

SOURCES AND METHOD

Original Genellipse analysis informed by the primary sources below. Examples are synthetic, not customer results, forecasts or industry benchmarks. Sources reviewed 14 October 2026.

NIST - AI Risk Management Framework

Read the complete report and its assumptions at genellipse.com/insights

Discuss your priorities: info@genellipse.com