

Inference cost accounting for finance teams

Per-token prices fall every quarter while the bill rises. Tag every model call to the use case that made it, and divide by outcomes.

Terence Kok & Dr. Christina Chua

9x–900x

**yearly fall in the price per token at a
fixed level of performance**

It depends on the task. Yet the AI bill keeps rising. Epoch AI, 2025.

The bill is a product of three numbers

Volume ↑

×

Tokens per task ↑

×

Price per token

Only the price is falling, and only at constant performance. Teams move to newer models, and a newer tokeniser can count about 30% more tokens for the same text.

5–30x

more tokens per task for agentic models than for a standard chatbot

Gartner, March 2026. Gartner also forecasts inference cost per agentic workflow rising more than fivefold through 2028.

One ID on both sides of the ledger

- 01** Tag every model call with the use-case ID from the benefits register
 - 02** Allocate shared platform costs by a driver agreed before the quarter
 - 03** Report build cost apart from running cost
 - 04** Show unused commitments as their own line
-

32%

of the running cost is model tokens



Model tokens \$6,100

Oversight and upkeep staff \$7,400

Retrieval, monitoring, evaluation, platform \$5,500

Illustrative email-drafting assistant, S\$19,000 a month.

Divide by outcomes, not by calls

S\$0.33

per draft, 57,000 drafts

S\$0.48

per reply sent, 40,000 replies

A draft nobody sends still costs money. Set S\$0.48 against a benefit of about S\$1.68 per reply, and the unit margin is about S\$1.20.

Three numbers, one page

- 01** **Cost per outcome** against benefit per outcome, for every live use case

- 02** **The bridge:** the change in the bill split into volume, consumption per outcome and rate

- 03** **The untagged share** of AI spend, and its three largest items

Five questions about an AI cost line

- 01 Which use case made these calls?
 - 02 Is the cost divided by outcomes?
 - 03 What beyond tokens is in the line?
 - 04 How much of the change is volume, consumption and rate?
 - 05 What share of AI spend has no owner?
-

Put the cost on the **same line** as the benefit.

Same use-case ID, same outcome, reviewed every month. Full article on terencekok.com.



Terence Kok · AI & Innovation Strategist

SAVE · COMMENT · FOLLOW