

• THE NEW LINE ITEM

How to Forecast Your AI Token Spend

A practical guide for finance teams turning AI usage into a number you can actually budget.

→ SWIPE TO BUILD THE MODEL

AI isn't a SaaS subscription. It's a metered utility.

Software is a flat seat licence you renew once a year. AI bills like electricity — it moves with every prompt, agent and workflow.

FLAT

Traditional SaaS — predictable, fixed per seat

VARIABLE

AI tokens — scale with adoption & automation

A single annual figure in your forecast won't survive contact with real usage. **Token spend can escalate faster than expected.**

● START HERE

First — what **is** a token?

A token is the unit AI reads and writes in. Not a word — a **piece** of one. Roughly **4 characters**, or about $\frac{3}{4}$ of a word.

// HOW TEXT BECOMES TOKENS

Fore

cast

ing

AI

token

spend

= **6 tokens**

Long words split into pieces, and spaces ride along — so token counts rarely match word counts.

~1.3

tokens per English
word

750

words $\approx$ 1,000 tokens
(~1.5 pages)

**IN +
OUT**

you pay for both,
priced separately

Every call has **two** meters.

You're billed for tokens going **in** and tokens coming **out** — and output usually costs more per token.



Input tokens

Everything you send: the question, instructions, documents and context.



Output tokens

Everything generated back — typically priced higher than input.



Cached & retrieved tokens

Reused context can be discounted — a real, modellable saving.

Forecast both meters. Most overruns hide in **output** and bloated context.

The six drivers of token cost

Map these before you model. Every one is a forecast input.



01

User adoption

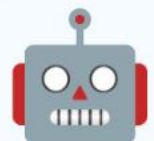
How many people actually use it, and how often.



02

Prompt volume

Interactions per user, per day, per workflow.



03

Agent activity

Autonomous agents run silently — and constantly.



04

Model selection

Premium vs economy models differ 10×+ in price.



05

Context & output length

Longer prompts and answers burn more tokens.



06

Automation scale

Touchless processes multiply volume overnight.

From tokens to a budget line

// COST PER INTERACTION

(input tokens × input rate)
+ (output tokens × output rate)

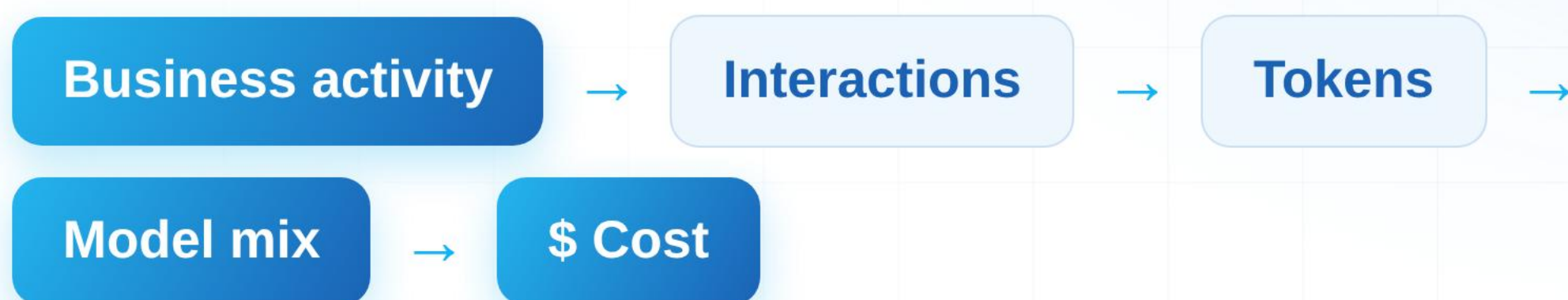
// MONTHLY SPEND

cost / interaction
× interactions per user
× active users
× model mix factor

Build it once as unit economics — then everything else is a volume assumption.

Don't forecast the bill. Forecast the behaviour.

Tie token volume to a business activity you already plan — invoices processed, reports generated, queries handled.



A

Anchor to a volume metric

e.g. "tokens per invoice processed" — a number that scales with the business.

B

Run three scenarios

Base, upside adoption, and a runaway case. Budget the base, watch for the runaway.

Model mix & efficiency

Same outcome, very different cost. This is where finance creates real savings.

10x+

Price gap between premium and economy models



Spend cut by routing routine tasks to cheaper models



Route by task

Premium for reasoning; standard & economy for high-volume, simple work.



Cache & reuse

Discount repeated context instead of paying for it every call.



Right-size prompts

Trim context and cap output length to stop silent token bloat.

• TRACK THESE

KPIs every CFO should own

Move from "what's the bill?" to "what's the unit economics?"



Cost per **AI-assisted transaction**



Cost per **active user & per agent**



Token **efficiency** — value created per token



Touchless rate — % of work fully automated



Spend by **business unit & use case**

Treat it like a budget — FinOps for AI

Visibility before volume. Put controls in place **before** usage scales, not after the invoice lands.

1

Allocate & show back

Attribute spend to the business units and use cases that consume it.

2

Set thresholds & alerts

Budget caps and anomaly alerts catch a runaway agent on day one.

3

Review on a cadence

Re-forecast monthly as adoption and model prices shift.

• THE TAKEAWAY

AI drives value — but it also **drives spend.**

The finance teams who build forecasting discipline around token spend **today** will lead tomorrow.

Want agentic AI built for the Office of Finance — with cost transparency from day one?

[Visit AITHENTIC.co](https://AITHENTIC.co) →

 Repost to help a CFO ·  Follow AITHENTIC for more