

We can build you the truck. What does the gas cost?

Ignore the token price IT EXPIRES

Rates change every few months. Do not trust anyone's price from memory, including ours.

Learn how it is built IT LASTS

How an agent is built sets the bill. One task is 10 to 20 model calls, and the retries drive the cost.

Three kinds of token spend real tasks, priced at today's list rates

SPIN

Attack this first

Activity with no output, like reports nobody opens.

\$675 a year

on automated briefings nobody reads

PRODUCE

Tune this

Work that ships. Keep it, and make each output cost less.

\$21

board deck that replaced about **\$6,500** of people time

TEACH

Protect this

The tries it takes to get a new workflow working. This is tuition.

\$16

118 attempts built a workflow that is now foundational to the business

The real number to measure

not price per token

cost per accepted task =

all tokens spent, retries included +
(review minutes × loaded rate)

outputs a human actually accepted

A cheap model that needs three tries costs more than a premium one that lands it once.

The eight dials that set your bill

01

Model fits the task

Match model strength to the task, not its price.

02

Budget the loop

Cap the iterations. Define done so work can stop.

03

Pay setup tax once

Every run re-sends the agent's setup. Keep it lean and cached.

04

Filter at the source

Retrieve twenty rows, not five hundred. Precision costs less.

05

Scope the session

New task, new session. Long threads resend history, costing more.

06

Right-size thinking

The biggest dial. Overthinking simple work costs more.

07

Instrument the build

If it cannot report its own cost, it is not done.

08

Fit the kill switch

Every scheduled job gets a spend cap, alert, review date.

Ask your technology partner



for a free one-hour CEO briefing

You will know the gas cost before you buy the truck. pisteyo.com