TKG

THE KHUMALO GROUP

AN INNOVATION AND RISK SCIENCES PRACTICE

WHITEPAPER

The real cost of AI tokens

> Everyone has used ChatGPT. Almost nobody has seen the bill for building AI into an actual business. Here is what sits between those two things — and why it is the real reason so many AI rollouts have quietly stalled.

Published 11 July 2026
Location Sandton, Johannesburg
Written by Dante Khumalo

Most people's entire experience of artificial intelligence is a chat window. You type a question, an answer comes back, and at the end of the month a subscription gets billed — the same amount whether you asked it one question or a thousand. That experience is real, but it is not what a bank, an insurer, or a pension administrator is actually buying when they set out to "put AI into" their own systems. It is not even close, and the distance between the two is where most AI budgets quietly come apart.

This paper is about that distance. Not the model, not the hype cycle — the bill, and why it behaves so differently from the one everyone already knows from their phone.

What you've used, and what a business actually buys

When you use ChatGPT, or Gemini, or Claude's own chat app, you are using a finished consumer product. Someone else has already decided how much computing power your question is allowed to use, absorbed the cost of it into a flat monthly fee, and hidden every technical detail behind a clean interface. It feels, correctly, like using a phone app.

Building AI into a business is a different transaction entirely. Instead of a monthly subscription, the business connects directly to the model over what's called an API — a direct, metered line to the same underlying technology, billed not per month but per **token**.

What a token actually is

A token is the small chunk of text the model actually reads and writes in — not quite a word, not quite a letter, closer to a syllable or short word fragment. "Claims" might be one token. "Underwriting" might be split into two or three. As a rough rule of thumb, a hundred plain English words is usually somewhere around 130-150 tokens.

Every model provider prices its tokens, usually in fractions of a US cent each — and prices reading (input) and writing (output) separately, with output usually costing several times more than input. Looked at one token at a time, this is almost absurdly cheap. That cheapness is the first thing everyone notices, and it is also the first thing that misleads them.

Why that gap matters

The AI revolution that institutions were sold on assumed a specific kind of arithmetic: that a task a person used to do by hand could now be done by a model for a cost so close to zero it barely needed a line item. If that were true, adoption would already be everywhere. It mostly isn't, and the reason is not that the models aren't capable enough. It's that "a fraction of a cent per token" and "a fraction of a cent per useful answer" are two different numbers, and the gap between them is exactly what a consumer chat subscription hides from you and a business's own invoice does not.

An institution that budgets for AI as if it worked like the app on their phone — a fixed, predictable, forgettable monthly cost — is budgeting for the wrong thing. What actually arrives is closer to a utility bill that moves with usage, denominated in a currency the business doesn't earn in, for a workload that is larger, on every single request, than it looks from the outside.

The real bill behind a single request

Here is what actually happens, in plain terms, behind one question a customer types into, say, an insurer's claims chatbot.

The model doesn't see just that one sentence. Every single call carries a **system prompt** — a block of standing instructions telling the model who it is and how to behave — sent again, in full, with every message, all day, for every user. On top of that, to answer anything specific to that customer, the system usually has to hand the model a chunk of **retrieved context**: the customer's policy details, recent claim history, the relevant section of a procedures manual — often several thousand words, again repeated on every call, because the model has no memory of its own between requests. Behind the single answer the customer sees, there is frequently more than one call happening — an internal step where the model checks or reasons over its own draft before it's shown to anyone. And, some proportion of the time, the first attempt comes back malformed, off-topic, or simply wrong, and the system quietly retries.

None of that is a flaw in the technology. It is the current cost of getting a reliable answer rather than a plausible-sounding one. But it means the true unit being billed is not "a customer's question" — it is "a system prompt, plus several thousand words of context, plus one or more reasoning passes, plus the occasional retry," repeated every single time, and every one of those pieces is priced.

Put illustrative numbers on it: at roughly two to three US dollars per million tokens of input and several times that for output — broadly where mainstream models sit today — a single "trivial" exchange might be a fraction of a cent. But once you add a repeated system prompt, several thousand tokens of retrieved account and policy context, an internal reasoning pass, and a retry roughly one time in five, that one visible customer question can easily carry the token weight of ten or twenty of the "trivial" calls the budget was built on. Multiply that by every customer, every day, and the invoice stops resembling the unit price anyone was quoted and starts resembling a second payroll.

The currency and governance problem

There is a second layer most institutions don't see coming, because it isn't a technology problem at all. Token pricing is almost always billed in US dollars, on a usage meter the business does not control, in a currency that moves against it. For a business earning in rand or lilangeni and paying a US vendor by the million tokens, that is a foreign exchange exposure wearing a technology invoice. Costs spike exactly when a campaign, a product launch or a busy season drives the most usage — which is exactly when finance least wants a surprise.

This is also where procurement and governance friction bites hardest, particularly for regulated institutions. A fixed software licence is easy to approve: a board or a regulator can see the number in advance and hold someone accountable to it. A metered, usage-based cost that depends on how creatively staff use a chat window is much harder to sign off, and far harder to audit after the fact. We've watched fully approved pilots stall at exactly this step — not because the technology failed a single test, but because nobody could tell an audit committee what next quarter's number would actually be.

Why the promise oversold its own arithmetic

The promise was that AI would make a category of work nearly free. What actually happened is that AI made a single attempt at a piece of work cheap, and then quietly multiplied the number of attempts required to get something an institution could actually rely on — through repeated context, through internal reasoning steps, through retries. Cheap-per-attempt and cheap-overall are not the same claim, and almost all of the disappointment sits in the gap between them.

None of this means the economics don't work. It means they only work once someone has actually modelled them — the same discipline that would be applied to any other operating cost before it gets signed off.

What we tell clients

Treat the token bill as an operating cost with a risk profile, not a subscription that happens to be usage-based. In practice that means sizing the model to the task rather than defaulting to the largest one available, putting a hard ceiling on retries and context size before they compound silently, caching answers to questions that get asked repeatedly instead of paying to regenerate them from scratch, and building the same currency and variance assumptions into the forecast that would go into any other dollar-denominated cost.

That is the unglamorous part of AI adoption, and it is also the part that decides whether a pilot becomes a production system an institution can defend to its own board, or a write-off nobody wants to explain. It is the same discipline we bring to any system we build or any risk we quantify: work out what the thing actually costs before you commit to running it at scale.