

Your AI bill isn't a DevOps problem. It's a **design failure.**

Token consumption is the primary cost driver of production AI — most programmes still treat it as a configuration detail instead of an architectural constraint.

SWIPE →

Four cost vectors every production system must account for

- 01** System prompt verbosity
 - 02** Context accumulation in multi-turn apps
 - 03** Unfiltered RAG payload size
 - 04** Output verbosity in agentic loops
-

60–80%

of inference cost comes from prompt token volume

In enterprise deployments, it's prompt tokens — not model choice — driving the bill.

Enterprise deployment data, cited in Token Optimisation, 2026

3-7x

cost gap between optimised and unoptimised pipelines

The same agentic workflow, at the same scale, can cost seven times more depending on architecture alone.

Token Optimisation Is an Engineering Discipline, 2026

40%

average token reduction achievable through structured context management alone — no model downgrade required.

Token Optimisation Is an Engineering Discipline, 2026



Token budgeting is not a constraint on AI capability. It is the mechanism by which AI capability becomes economically sustainable across an enterprise portfolio.

Terence Kok

Enterprise AI Strategist

System prompts: monolithic vs. conditional

MONOLITHIC

4,200-token prompt file

Loaded in full, every request

No routing logic

CONDITIONAL

800-token core instructions

+ role and task context injected

1,150–1,800 tokens typical

Not every task needs a **frontier model**.

A three-tier cascade routes classification and intent detection to a small, fast model; standard task execution — 70–80% of volume — to a mid-tier model; and reserves frontier models for the 10–15% of requests that genuinely need deep reasoning. Misroute either direction and you break quality or economics.

Treat token spend like compute spend.

**Budgeted at design stage. Instrumented per layer.
Reported per outcome, not per raw token. Reviewed
every quarter.**

Full architecture breakdown and the cascade-routing framework:

terencekok.com/blog/token-optimisation-engineering-discipline-not-prompt-trick

Design for token debt. Don't retrofit it.

Budgets, telemetry, tiered routing, deliberate context management, quarterly review. The technology exists — the discipline is what's missing.



Terence Kok · Enterprise AI Strategist

SAVE · COMMENT · FOLLOW