

Less Architecture Diagram. More What Breaks.

The Eighth Layer laid out the theory. This is for the engineer shipping it — the contracts, thresholds, and failure modes your test suite won't catch.

SWIPE →

What Actually Breaks in Production

- 01** Treat the Layer 7/8 boundary like a real API
 - 02** Route on similarity, not vibes
 - 03** Separate context, episodic, and semantic memory
 - 04** Evaluate on live traffic, not a one-time eval
-

Version It Like **an API**.

Don't bolt an LLM call onto your service and call it done. Define the input envelope: intent, session ID, identity claims, what tools this request can even touch. Define the output: content separate from a confidence score, separate again from tool-call metadata. Set a hard timeout, and decide what happens when it's hit before you ship, not after the first incident.

Pick the Pattern, Not the Fanciest One

CHEAP

Single-pass: narrow, unambiguous tasks

ReAct loop: multi-step, dependent results

EXPENSIVE

Reflection: high-stakes, checkable correctness

Match pattern to risk, not habit



Context Is Not Memory

Short-term context is a sliding window, scoped to one session. Long-term memory splits again: episodic (what happened before) and semantic (verified knowledge). Blend episodic with semantic and a claim quietly becomes a fact.



Your existing test suite was built for deterministic code, and this isn't that.

Terence Kok

Enterprise AI Strategist

Evaluate on Live Traffic

- 01** Faithfulness — fraction of claims that check out against context
- 02** Relevance — catches “correct but didn't answer the question”
- 03** Cost & latency — report p95/p99, never just the average

Two-Stage, Inbound and Outbound.

Run injection detection before anything hits your main model, kept separate so a jailbroken model isn't the only thing watching. Filter outbound for structured data and sensitive content. Block on detection — don't just log and let it through. Scope tool access per call, not per agent.

Config You Revisit, Not Constants You Set

Are your routing thresholds and orchestration patterns something your team revisits — or something you set once at launch and forgot?

Full breakdown and sources:

terencekok.com/blog/layer-8-in-practice-field-guide-for-engineers

Ship the Contract, **Not Just the Call.**

The API contract, the routing thresholds, the memory boundaries — read the full field guide.



Terence Kok · Enterprise AI Strategist

SAVE · COMMENT · FOLLOW