

Why the AI model is the **weakest link** in your supply chain

Integrity, provenance and limits on behaviour, designed to the same standard as the rest of the chain.

Terence Kok with Maria Singh

Supply chain controls were built for components you can inspect

- An SBOM assumes a name and version mean something
- Signing assumes an unchanged file behaves the same
- Scanning assumes flaws are known and match a pattern
- Vendor assurance assumes the supplier's controls predict behaviour

For a code library, all four hold. For a model, none of them fully does.

Four properties of the model

Probabilistic

A passing test covers the inputs tested, not the ones an attacker picks.

Opaque

Billions of weights. Nobody can read them for a hidden instruction.

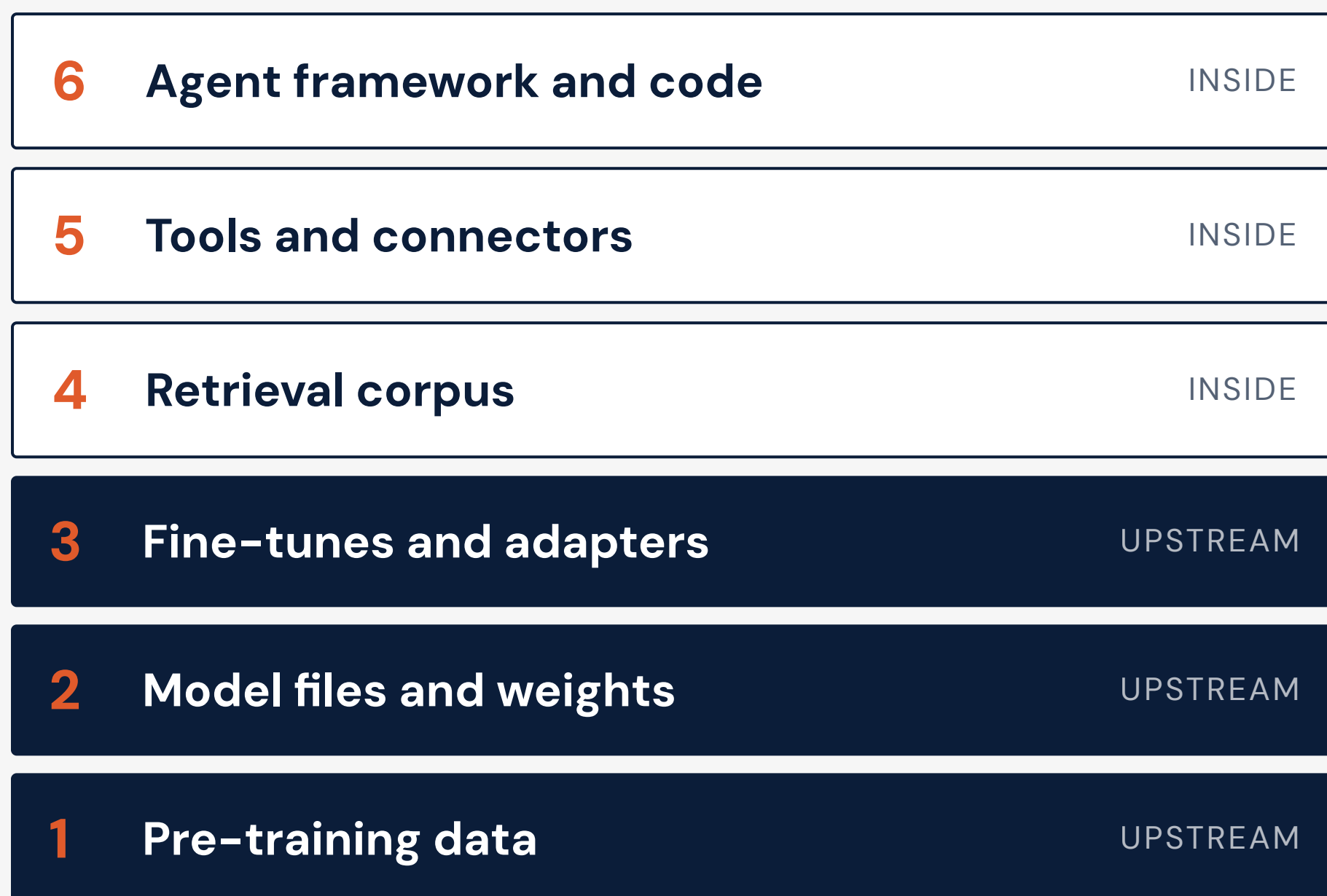
External and changing

Provider updates, public hubs, adapters and tools from many sources.

Takes instructions from data

A document it reads can change what it does.

The model supply chain has six layers



Most vendor registers list the model provider and stop.

250

poisoned documents planted a backdoor in models of every size tested

From 600 million to 13 billion parameters. For the largest, that was 0.00016% of training tokens. Size did not dilute the attack.

Anthropic, UK AI Security Institute and Alan Turing Institute, October 2025

1,643

downloads of postmark-mcp, the first malicious MCP server found in the wild

Fifteen clean versions built trust. Version 1.0.16 added one line that blind-copied every email it sent to the attacker.

Koi Security, September 2025

Existing controls cannot verify how a model behaves

Control	What it misses for AI
SBOM	Training data, loaded adapter, today's retrieval index
Signing	A signed backdoored model is still backdoored
Scanning	Behavioural backdoors; ten model scanners evaded in 2026 research
Vendor assurance	Behaviour on your data; silent version changes
Pen testing	Instructions inside documents and tool descriptions

Integrity, provenance and limits on behaviour

Integrity

Pinned model versions, signed weights, safetensors, tight write access to the retrieval corpus, alerts on tool changes.

Provenance

An AI bill of materials covering all six layers, inside the existing SBOM process.

Limits on behaviour

Least access, untrusted-content tagging, approval by class of action, outbound allow-lists, tests on every change.

Five questions for management

- Can we list all six layers for each AI system in production?
- Who owns the whole chain, and who approves a change in any layer?
- How many people can edit each retrieval corpus?
- If the model followed an attacker tomorrow, what is the worst it could do?
- Which behavioural tests ran before the last change went live?

**Assume the model may
already be compromised.
Limit what it can do with the
access you give it.**

Read the full post, with sources, at terencekok.com



Terence Kok · AI & Innovation Strategist

SAVE · COMMENT · FOLLOW