

The Model Is Not the Product

The “Toy Session” Trap



The Illusion

Prompt + Tools + Context
Window = A Working Agent.

The Reality

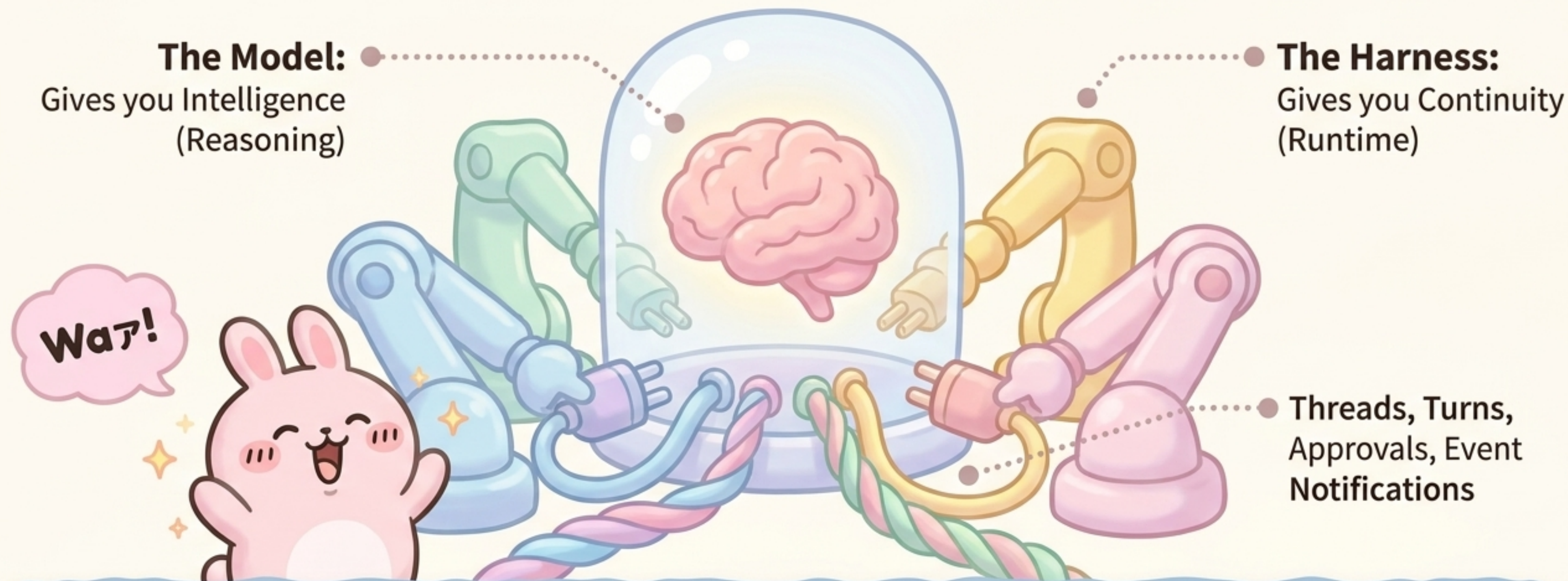
That gets you through a
5-minute demo. Anything
longer, and you step on a rake.

When the Demo Breaks



Products are made or broken here. Not on benchmark pages.

The Brain vs. The Harness



The Harness is the runtime interface. It is the actual product.

The Invisible Machinery

Keeping long-running agent work alive requires massive explicit runtime plumbing.

Context
Compaction
(Task state, tool deltas)

Prompt-Cache
Preservation

Sidechain
Transcripts
(For correctness,
not analytics)

Background
Observer Agents



This is not “extra plumbing.” This IS the product.

Surviving Real Software



The Demo

AI App with Tools

Chatbot Architecture

Fragile network & state

Fails after 6 minutes

The Product (The Harness)

Model in the Middle

Distributed Operating Env.

Stateful Query Loop +
Recovery

Useful for 6 hours

The model writes code. The harness survives the conditions under which software gets written.