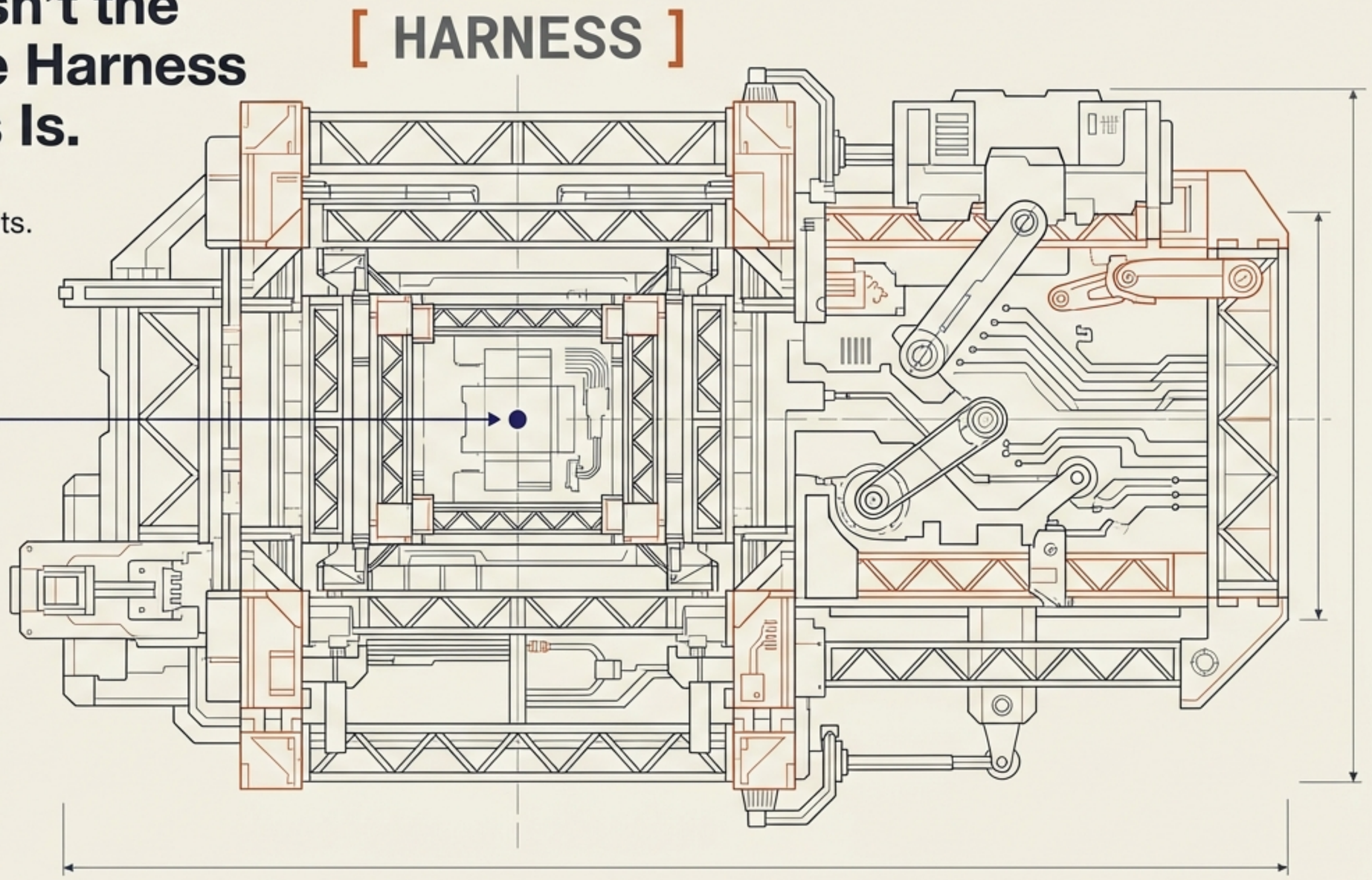
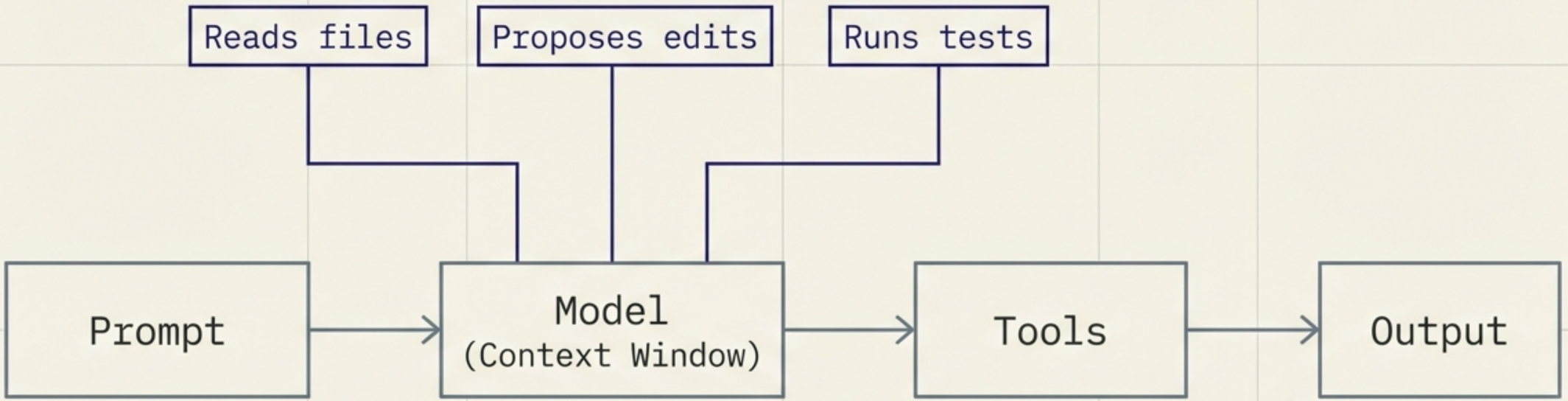
The Model Isn’t the Product. The Harness The Harness Is.

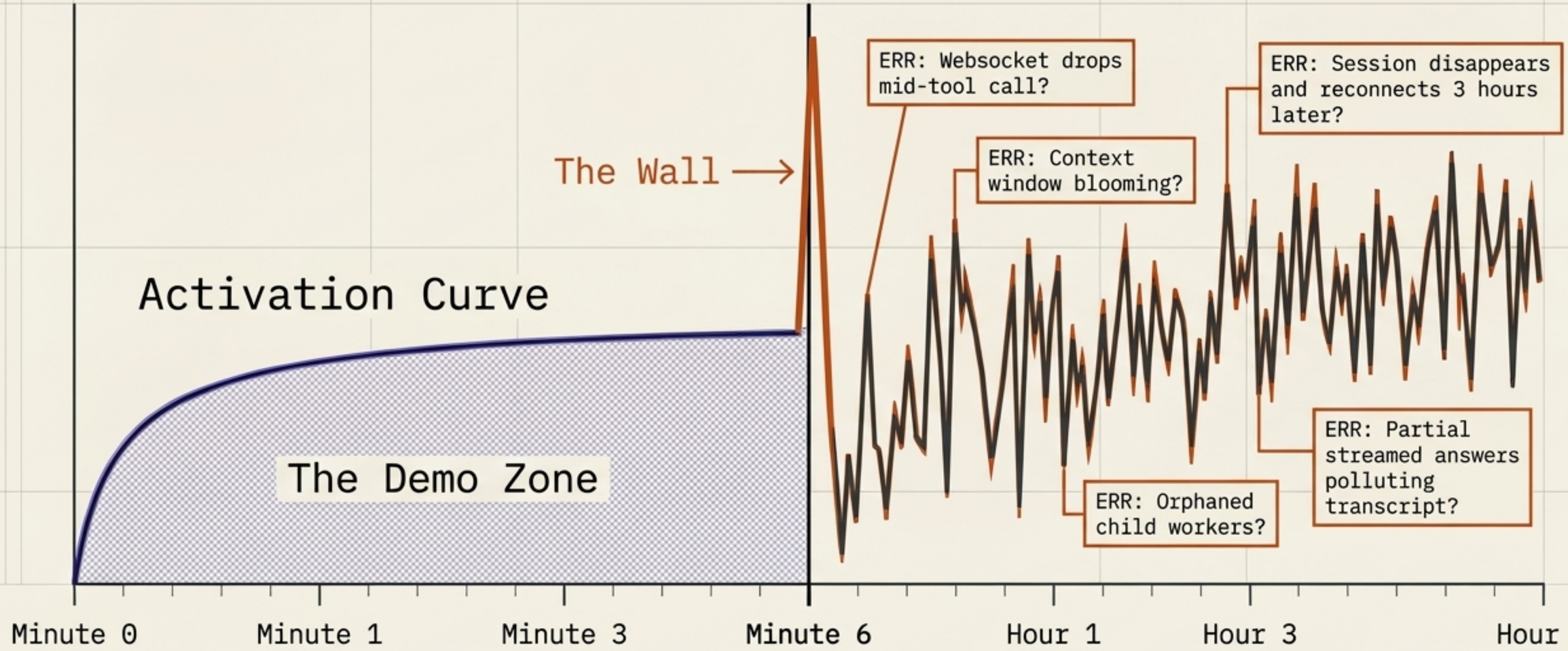
Inside the architecture of production-grade AI agents.

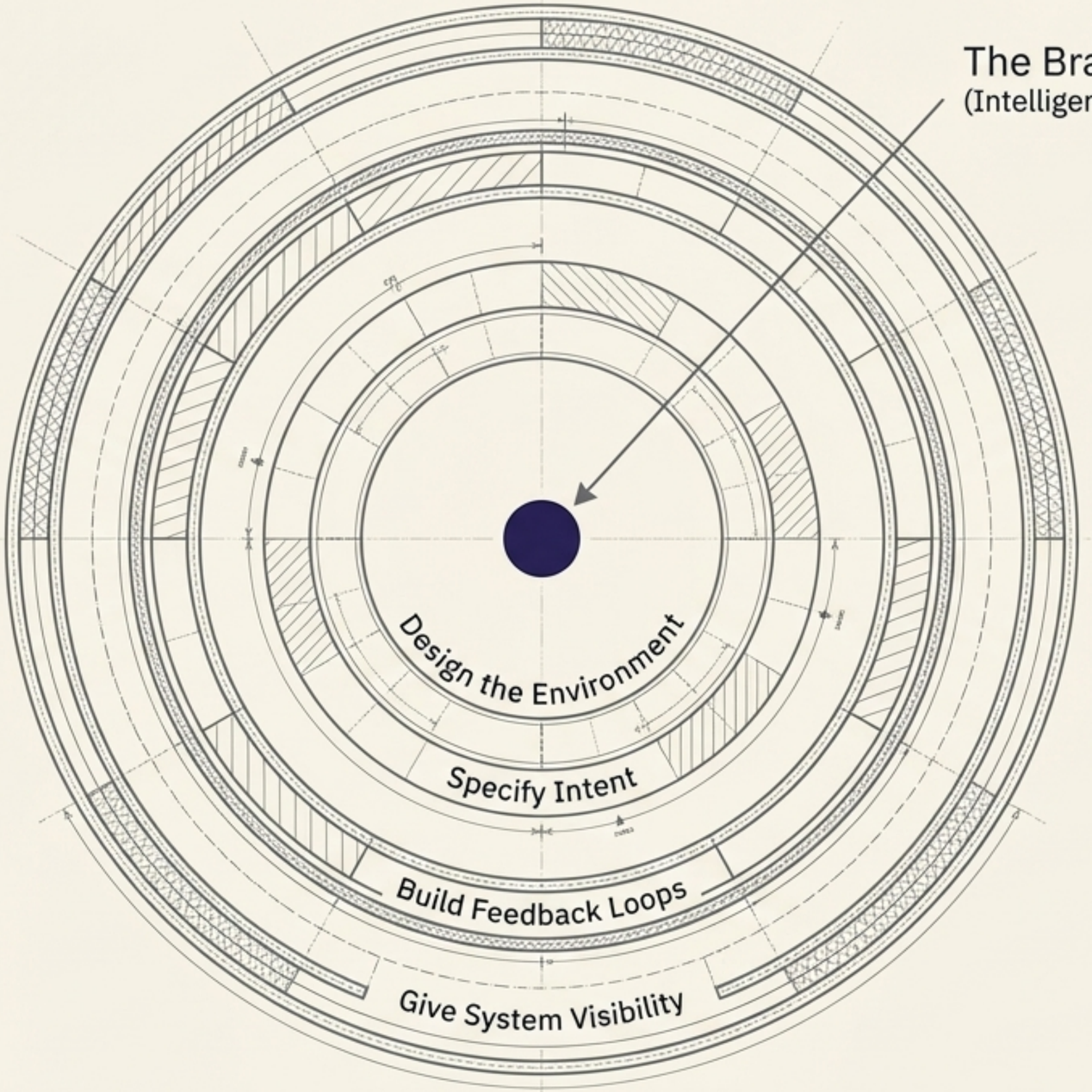




A coding agent is easy to imagine if you stop at the model layer.

Products are made or broken here.
Not on the benchmark page. In the harness.

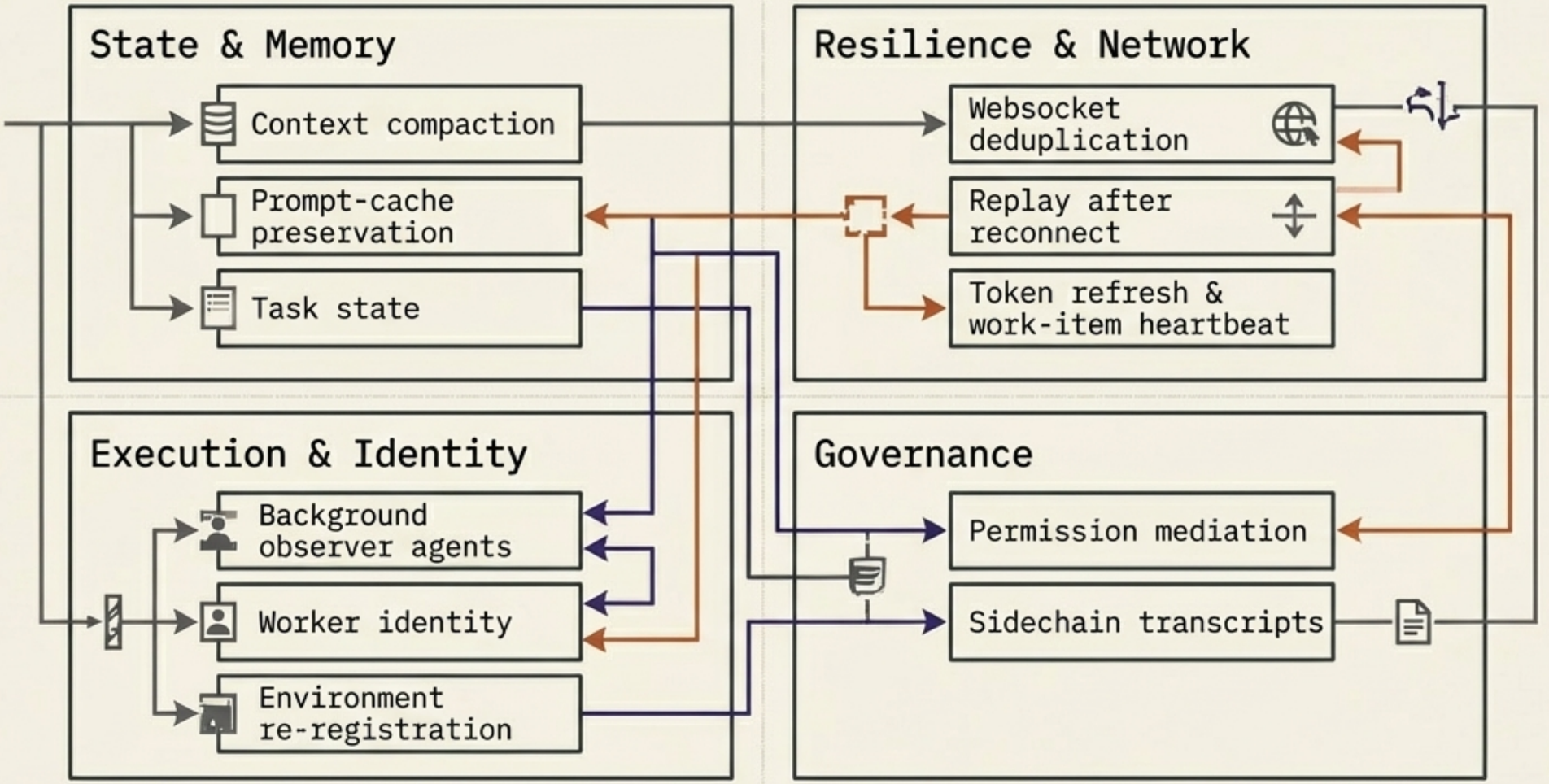




ENGINEERING SHIFT

Engineering work has moved outward. The harness is the concrete runtime interface around the model—the protocol boundaries between clients and the core runtime.

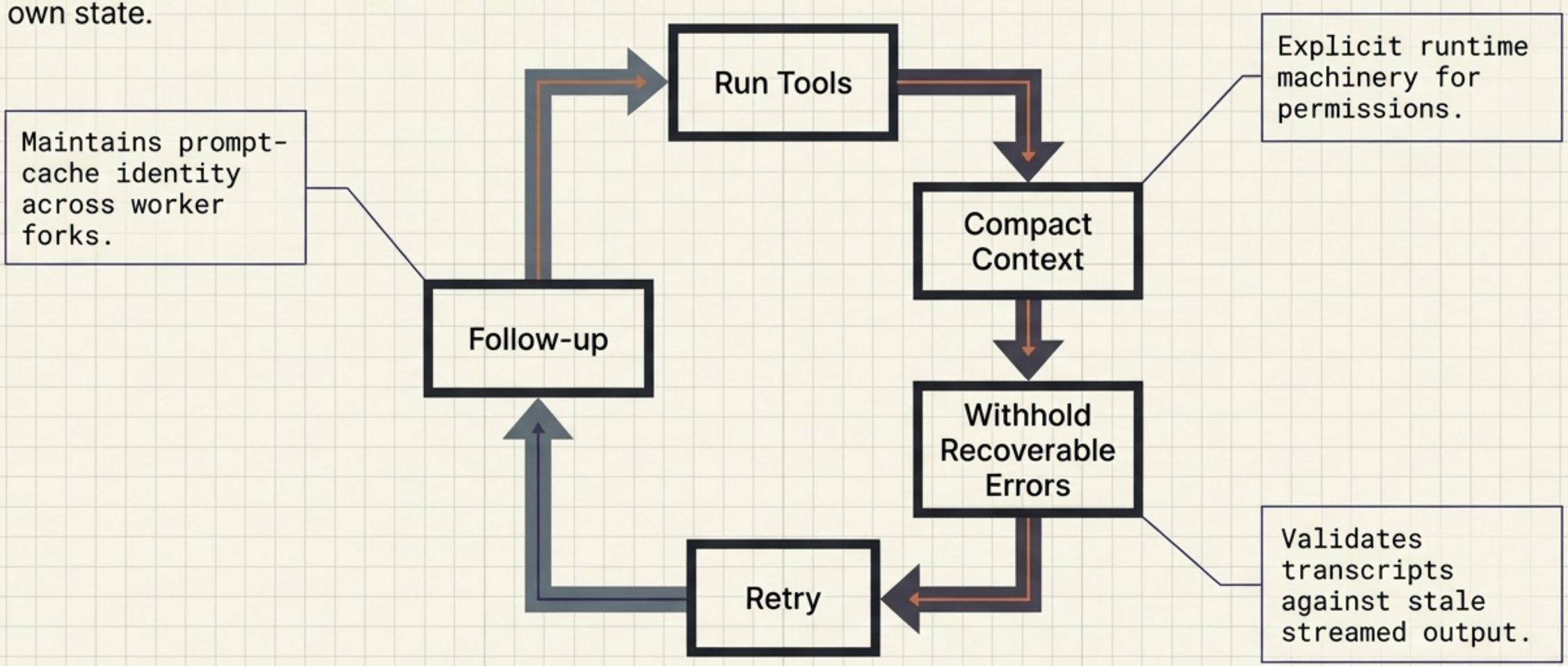
Core Runtime Boundary



This is not extra plumbing. This IS the product.

The Stateful Query Loop

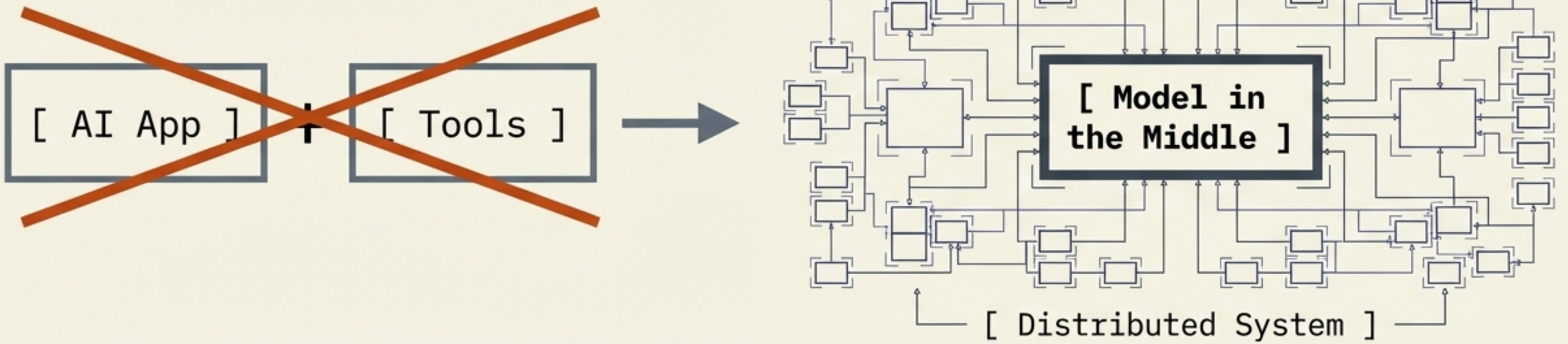
It is not “send one request, get one answer.”
It is a state machine that keeps mutating
its own state.



The architecture of a chatbot vs. The architecture of an operating environment.

Demo	Production Harness
1 Agent	**Subagents + Background Observers**
Single Local UI	**Multi-surface Remote Clients**
Print to Screen	**Real File Edits + Validated Transcripts**
Fails on Network Blip	**Sequence Numbers Across Transport Swaps**

The Hardest Parts



The hardest parts aren't tweaking prompts or swapping frontier models.
The hardest parts are:

- Preserving state across failure
- Deciding which workers share context and which isolate
- Stopping context growth from poisoning the session

~~Old Question:~~
~~"How smart is the model?"~~

New Question:
"What environment makes the model useful for six hours instead of six minutes?"

The model is the part that reasons.
The harness makes the reasoning usable.

Intelligence vs. Continuity.