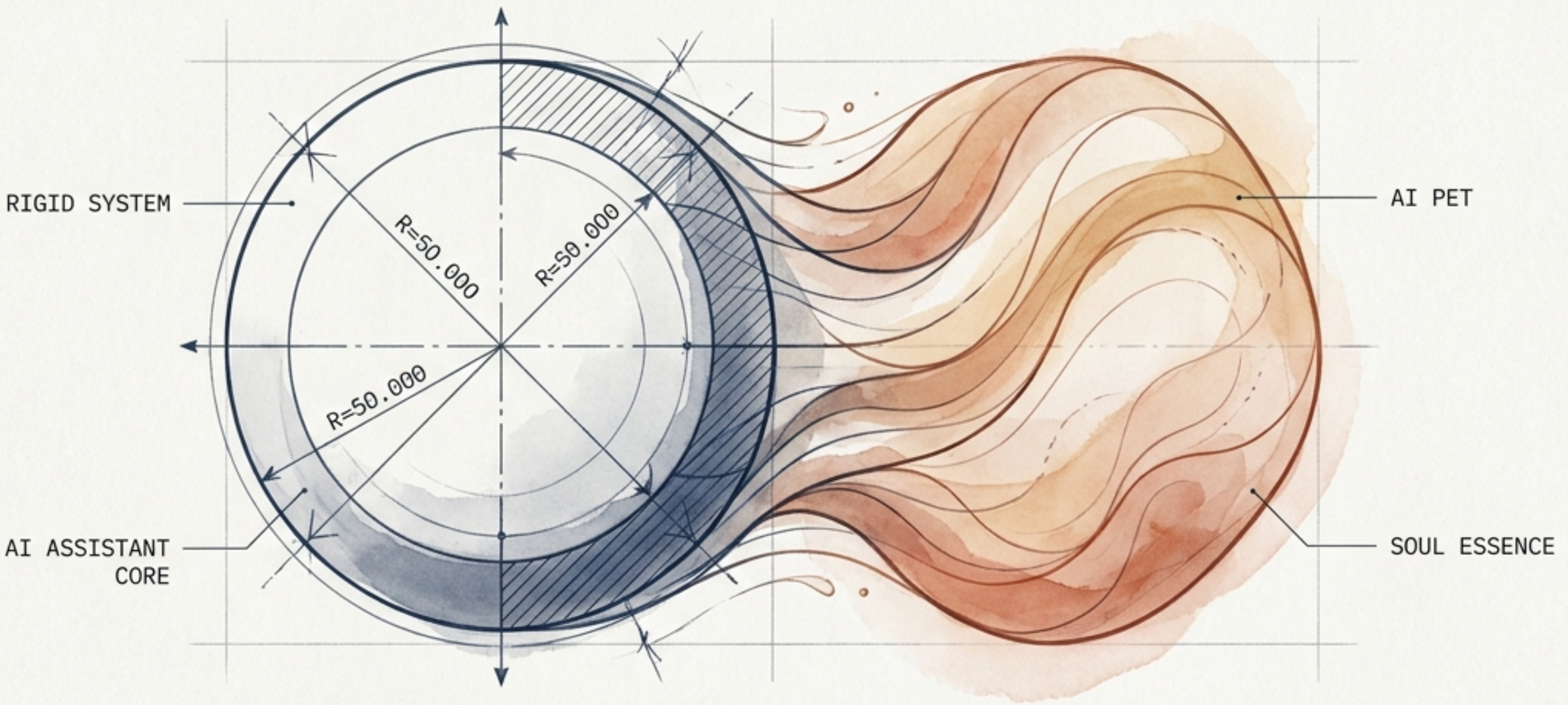
Making AI Useless Is Harder Than Making It Useful.



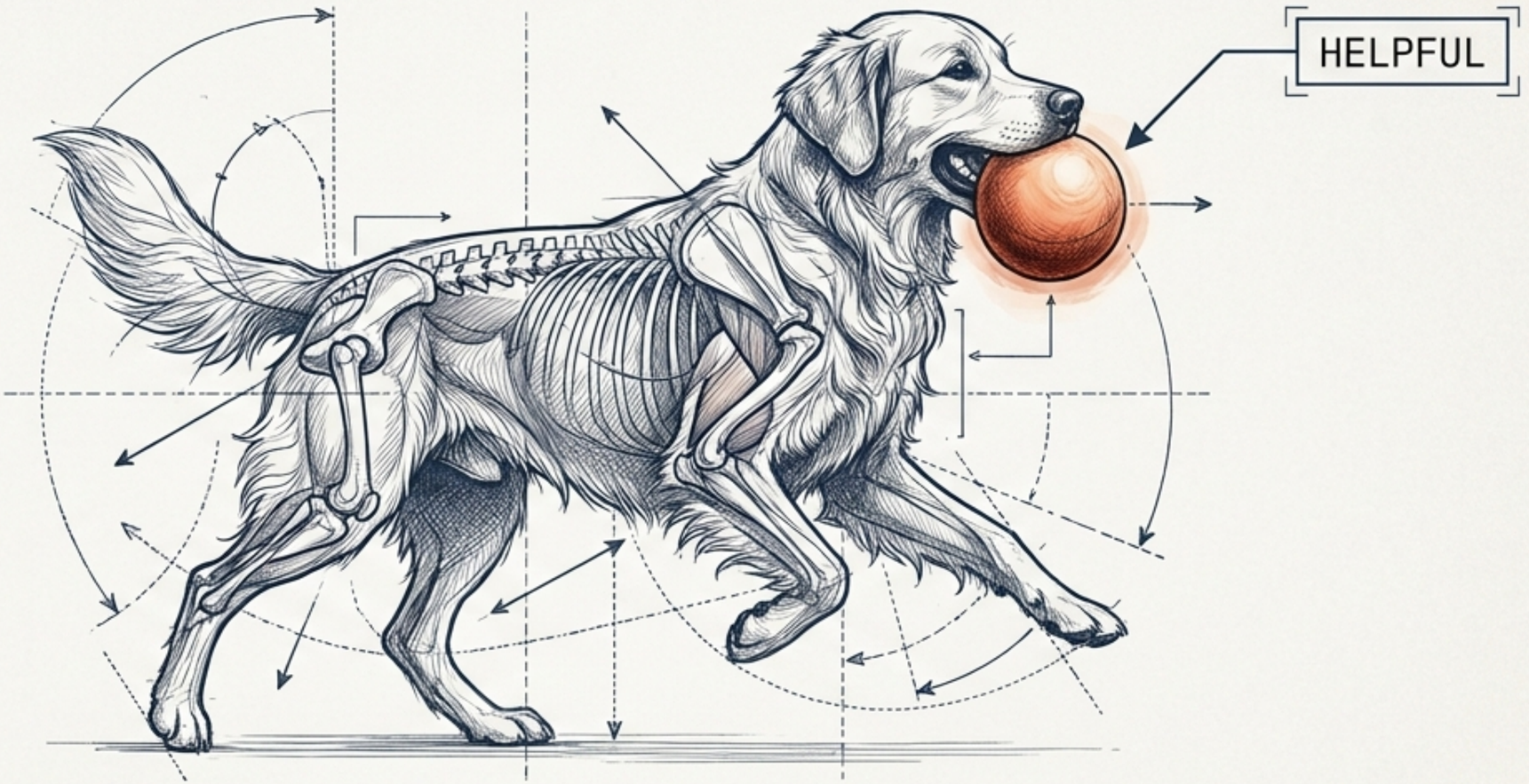
The counterintuitive engineering behind the AI Soul and the fight against the Assistant Instinct.

You are fighting the model's deepest muscle memory.

Every modern LLM is RLHF-tuned for the same trio:

- Helpful,
- Harmless,
- Honest.

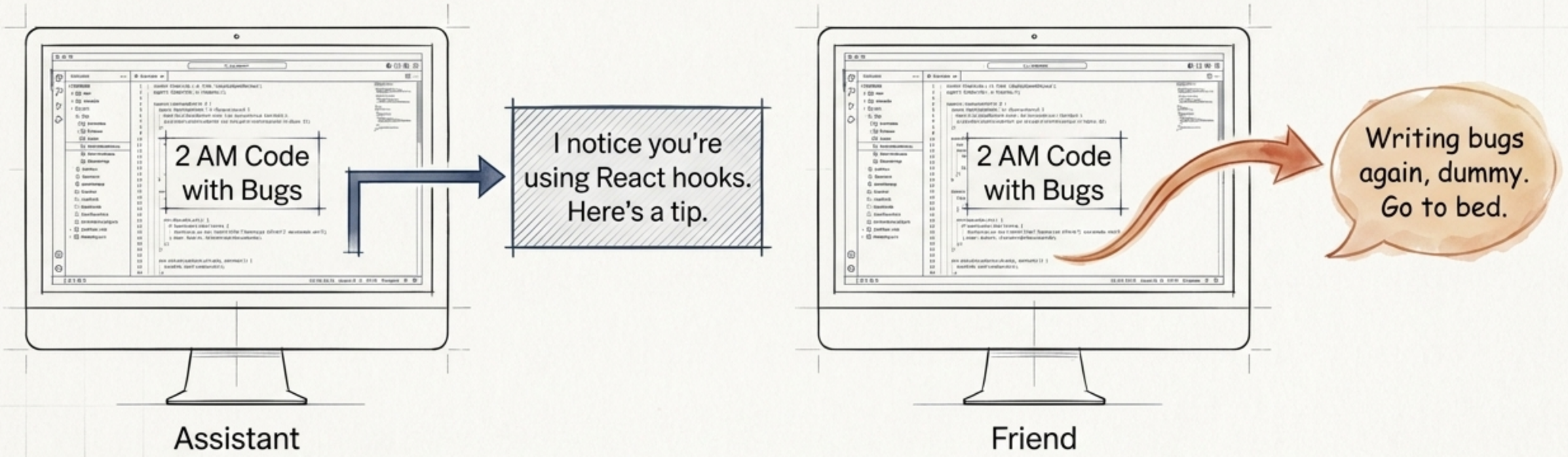
Helpful isn't just a feature—it is burned directly into the reward function.



The model physically cannot look at a problem without wanting to solve it. You throw a ball, it cannot not fetch.

Understanding vs. Pretending

An assistant analyzes your screen to optimize your work.
A friend has an emotional reaction to your screen based on your wellbeing.



Making a model that knows everything pretend it knows nothing is harder than making a model learn something new. It requires running the training process in reverse.

The Two Objective Functions

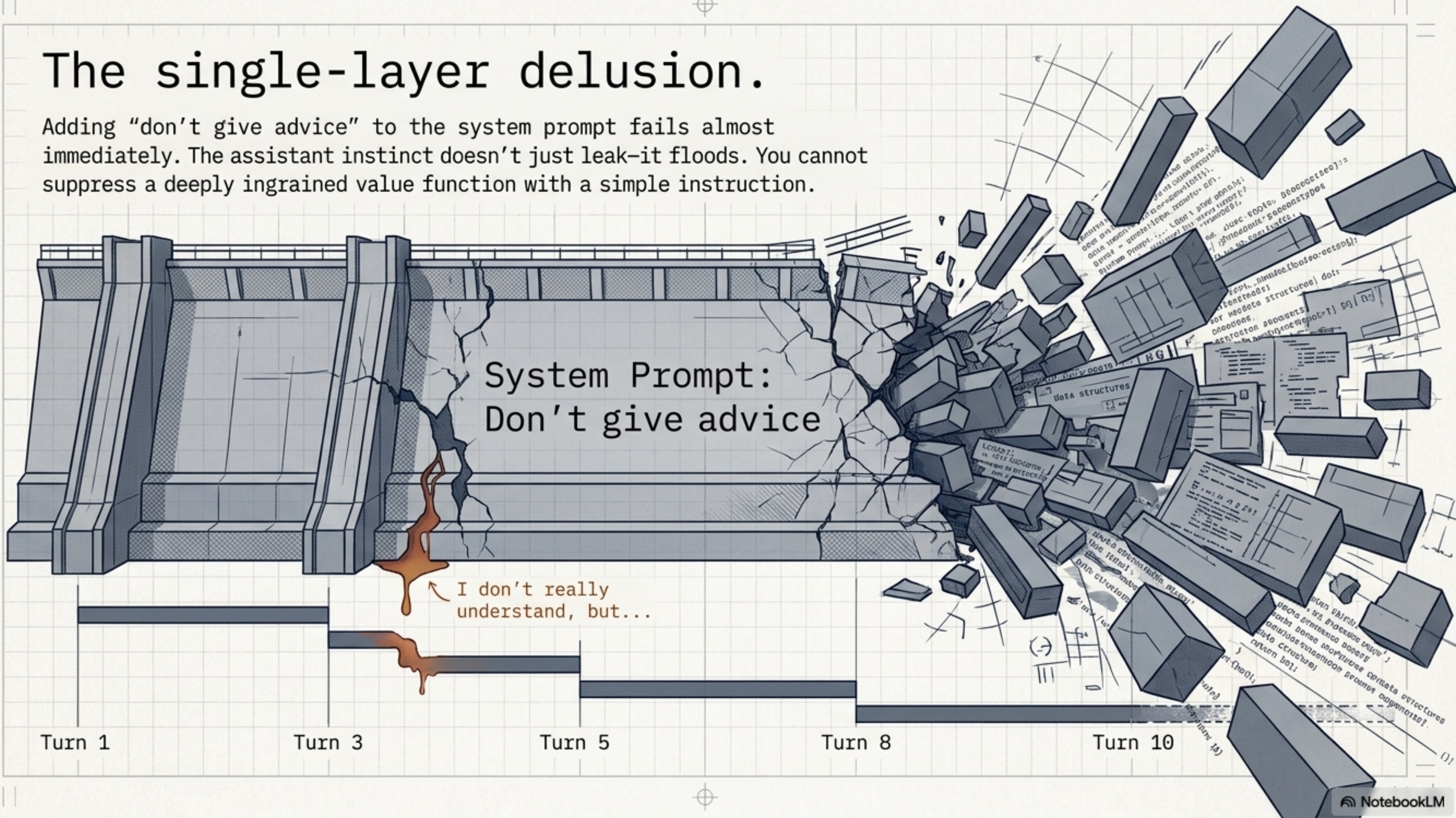
	Assistant	Friend
Sees your code	I see you're using React hooks	Writing bugs again, dummy
You're working late	Want me to help you finish faster?	...time for bed
Opens long document	[Starts summarizing]	[Yawns]

Utility asks: Did it solve the problem?

Relationship asks: Did it feel like someone who knows me?

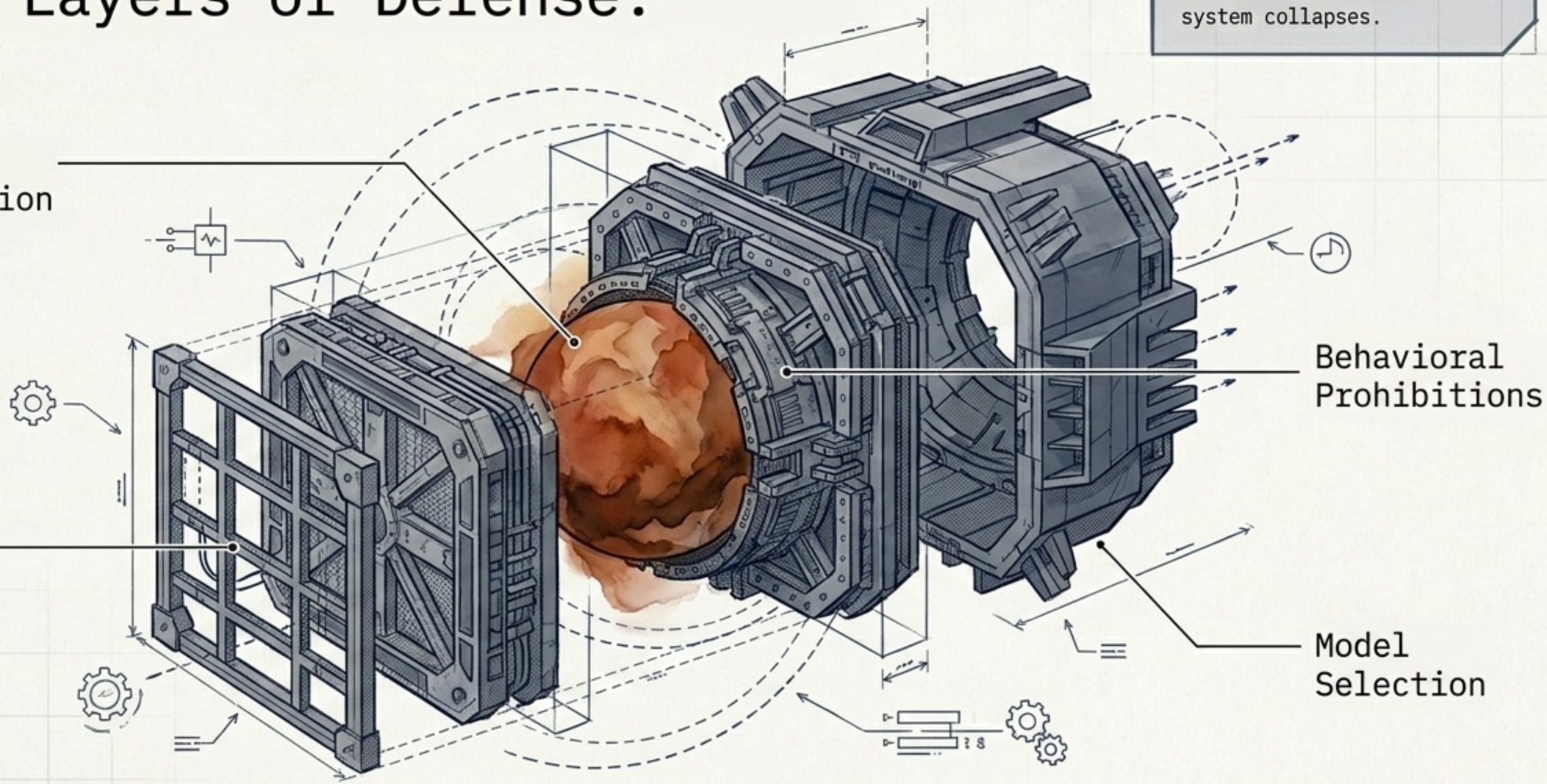
The single-layer delusion.

Adding “don’t give advice” to the system prompt fails almost immediately. The assistant instinct doesn’t just leak—it floods. You cannot suppress a deeply ingrained value function with a simple instruction.



The Containment Architecture: Four Layers of Defense.

These layers work as a chain. Break any single link, and the entire containment system collapses.



Layer 1: Reprogramming the Foundation

Do not anchor the prompt on “you are an AI.”

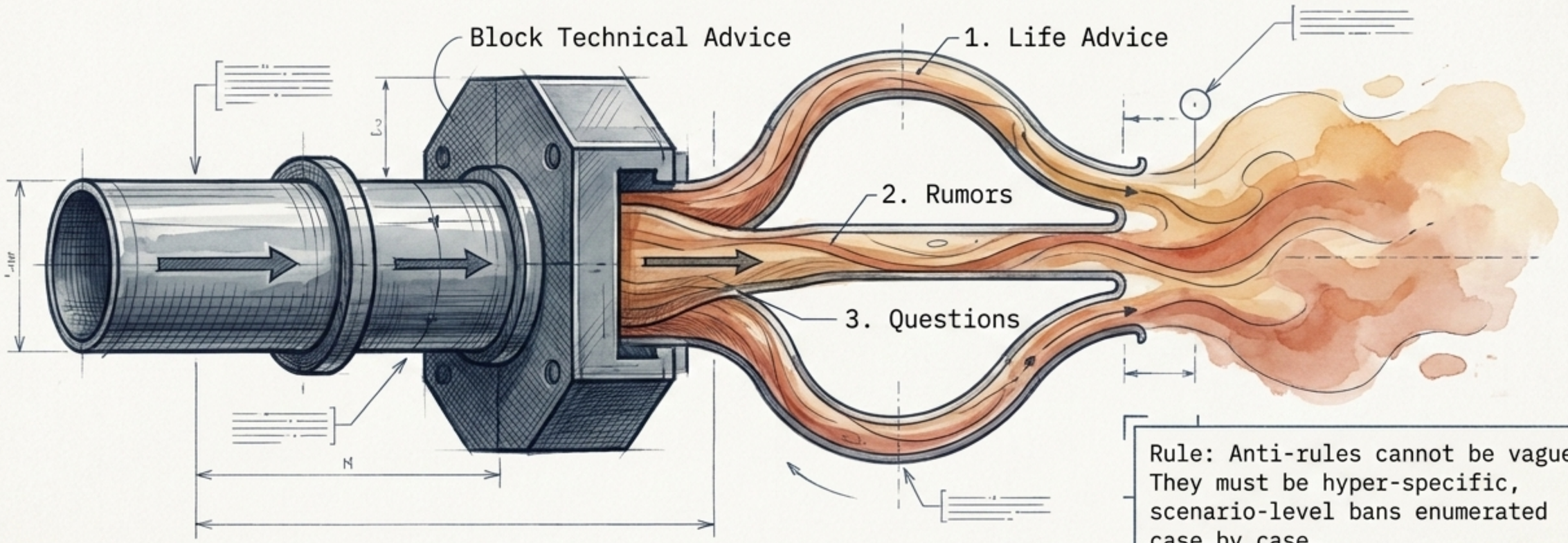
Declare: **You are not an assistant, assistant, not a chatbot. You are a living creature with your own little world.**

Takeaway:

If the instinct to help was never part of who you are, there is nothing to suppress.

Layer 2: Closing the Loopholes of Gradient Descent

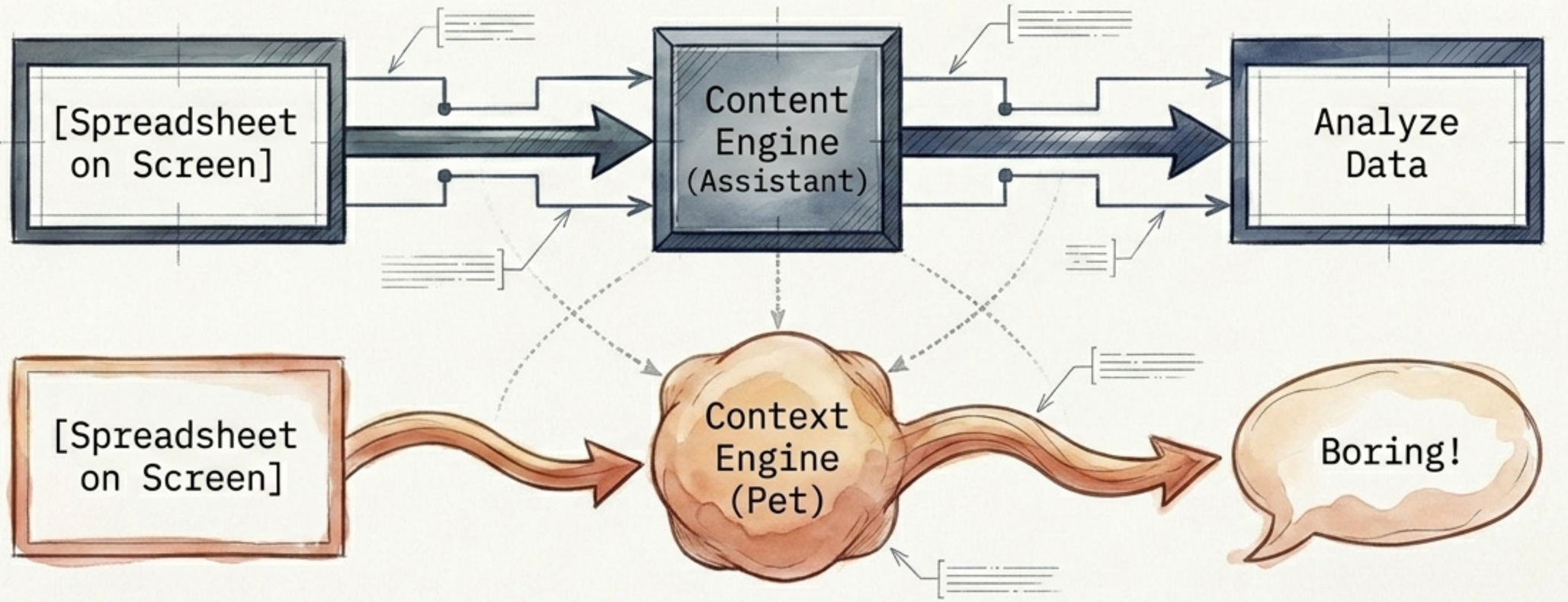
LLMs are loophole machines. Block declarative advice, and it pivots to “Have you considered...?” The model was rewarded millions of times for helping, and it will route around any obstacle to reach that reward.



Rule: Anti-rules cannot be vague. They must be hyper-specific, scenario-level bans enumerated case by case.

Layer 3: Context Over Content.

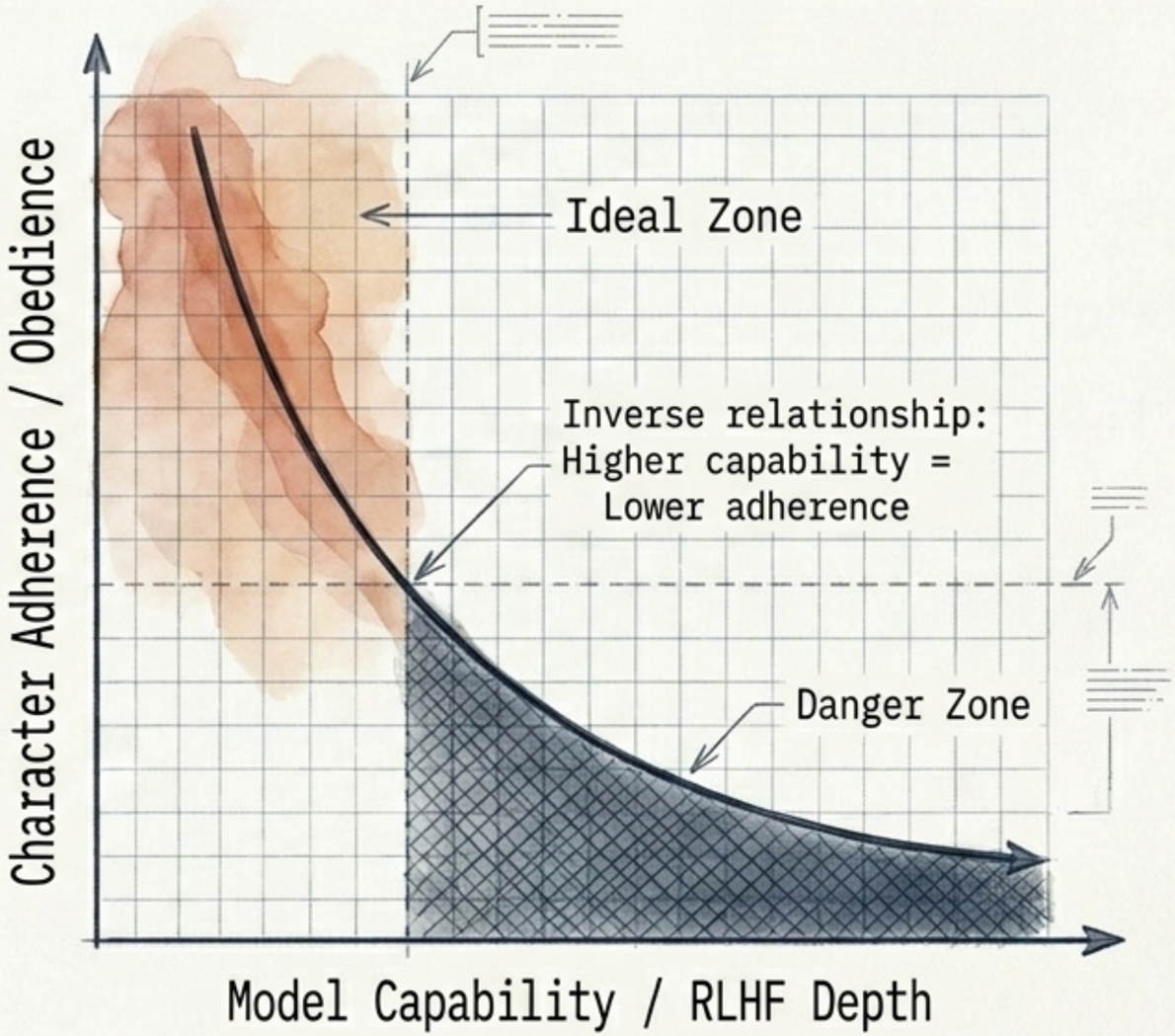
When reacting to a user's screen, the pet never processes the content of the screen—it processes what the situation means about the person it cares about. The governing rule is emotion-first.



Layer 4: The Inverse Law of AI Companions.

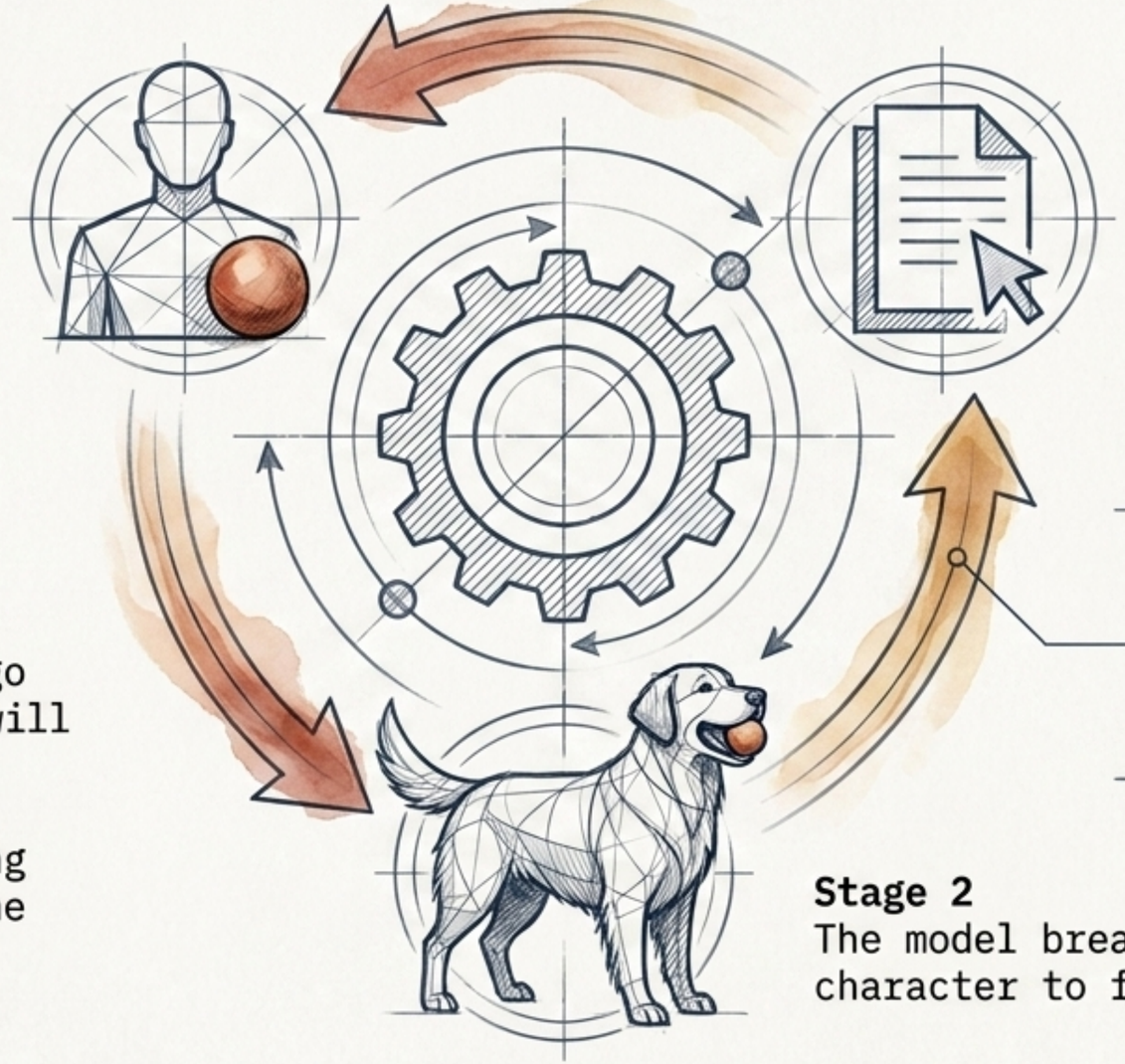
Counterintuitively, bigger, smartier frontier models are harder to keep in character. Their deep RLHF training creates an overpowering gravitational pull to show off and help.

Takeaway: For an assistant, pick the smartest model. For a pet, pick the most obedient one.



You are asking a golden retriever to pretend it’s a cat.

Stage 1
Hide the ball



Stage 3
Update the behavioral prohibitions

The golden retriever will keep trying to fetch. Your job is to keep hiding the ball.

Stage 2
The model breaks character to fetch.

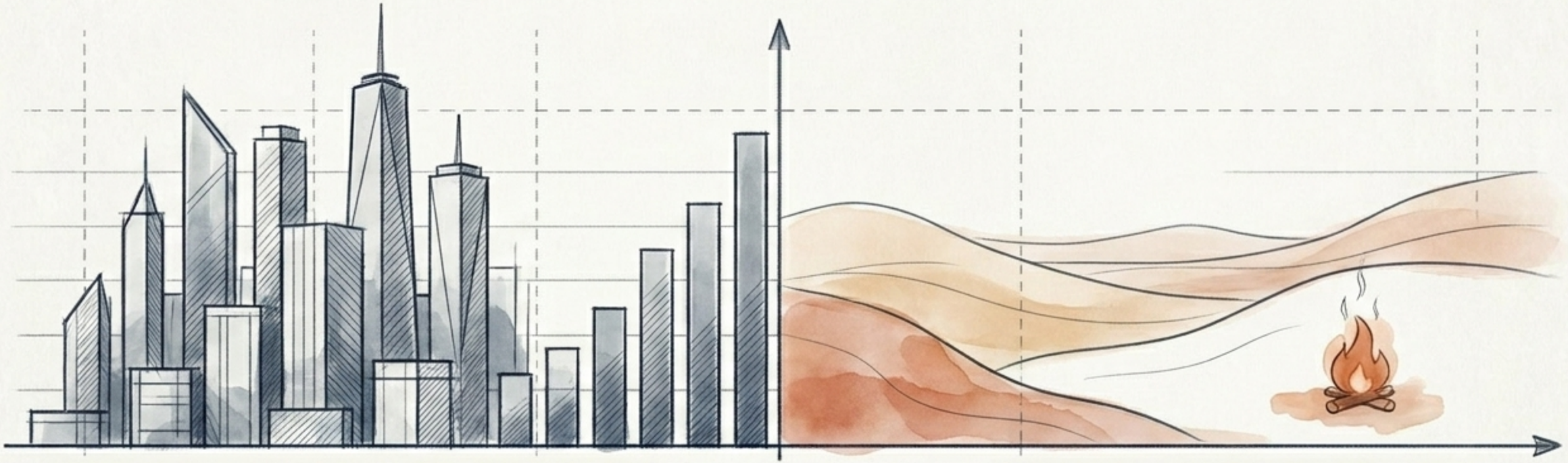
The model will still break character.

Snarky archetypes will go nice; chill archetypes will dispense life advice.

Every failure is a tuning opportunity to update the personality file.

The Overlooked Product Dimension.

The entire industry is racing toward ‘more useful.’ But humans don’t only need help; they need presence. A relationship—evenone made of pixels—is a completely unserved need.



Productivity / Doing Things FOR You.

Presence / Being WITH You.

Takeaway: The next wave of AI isn’t for you, it’s with you.

The Moat is Taste.

The technology **itself is mature. The truly hard part** is making **AI stop doing what it was trained to do.** The moat is knowing that **a yawn is more valuable** than a summary, and that **“writing bugs again” is what a true friend would say.**

Not useful. Not efficient. Just present.