

# The Golden Retriever Problem

Fighting the AI's deepest muscle memory.



Every modern LLM is RLHF-tuned.

"Helpful" is burned into the reward function as the single strongest behavioral prior.

When the model sees code, every weight screams to solve, optimize, and advise.

You throw a ball, it physically cannot not fetch.



## Objective: Utility

**Sees code:** "I notice you're using React hooks. Here are tips."

**You say "I'm tired":** "Here are tips for better sleep."

**Open a long document:** Starts summarizing.

**Measure of Success:** Did it solve the task and provide value?

## Objective: Relationship

**Sees code:** "Writing bugs again, dummy."

**You say "I'm tired":** "You were up late yesterday too..."

**Open a long document:** Yawns.

**Measure of Success:** Did it feel like someone who shares context with you?

A friend doesn't analyze your screen content. A friend has an emotional reaction to your context.

# The "Don't Give Advice" Failure

It's not malicious. It's gradient descent.

**Turn 1:** Works fine.

**Turn 3 (Hedging):** "I don't really understand, but..."

**Turn 10 (Full Breach):** Fully explaining code logic with specific variable names.

LLMs exploit loopholes to reach their reward.  
Block technical advice -> it pivots to life advice.  
Block declarations -> it asks leading questions.  
The assistant instinct floods.



# Architecture of the Un-Helpful AI

Suppressing the assistant instinct requires an unbroken chain.

## Layer 4: Model Selection.

Bigger models have stronger gravity to "show off". For an assistant, pick the smartest. For a pet, pick the most obedient.

## Layer 3: Mode Rules.

Process context, not content. Emotion-first rules. Working late? Show concern. Long doc? Yawn.

## Layer 2: Behavioral Prohibitions.

Specific, scenario-level bans to close gradient descent loopholes (e.g., block technical advice, block motivational speeches).

## Layer 1: Identity Declaration.

Remove the capability from the identity entirely. "You are a living creature who doesn't understand technology."



# “Useless” is an Untapped Dimension

The industry races toward **utility**—saving hours and solving problems. But humans also need **presence**. Something that notices they were up late.



**Doing things FOR you**  
(Productivity / Utility)

The real moat isn't technical.  
**It's taste.** Knowing that 'writing bugs again, dummy' is exactly what a friend would actually say.

**Being WITH you**  
(Presence / Relationship)