

Language Models Have No Reference Frame for Numbers

Language models don't know the gap between 0.6 and 0.8 humor.

Numbers describe outcomes.
Prose describes process.



"Director's Slate" Comparison Matrix

Parametric Configs



- Flat and repetitive
- Dissolves into default voice in 3 turns
- Like a stranger wearing different hats.

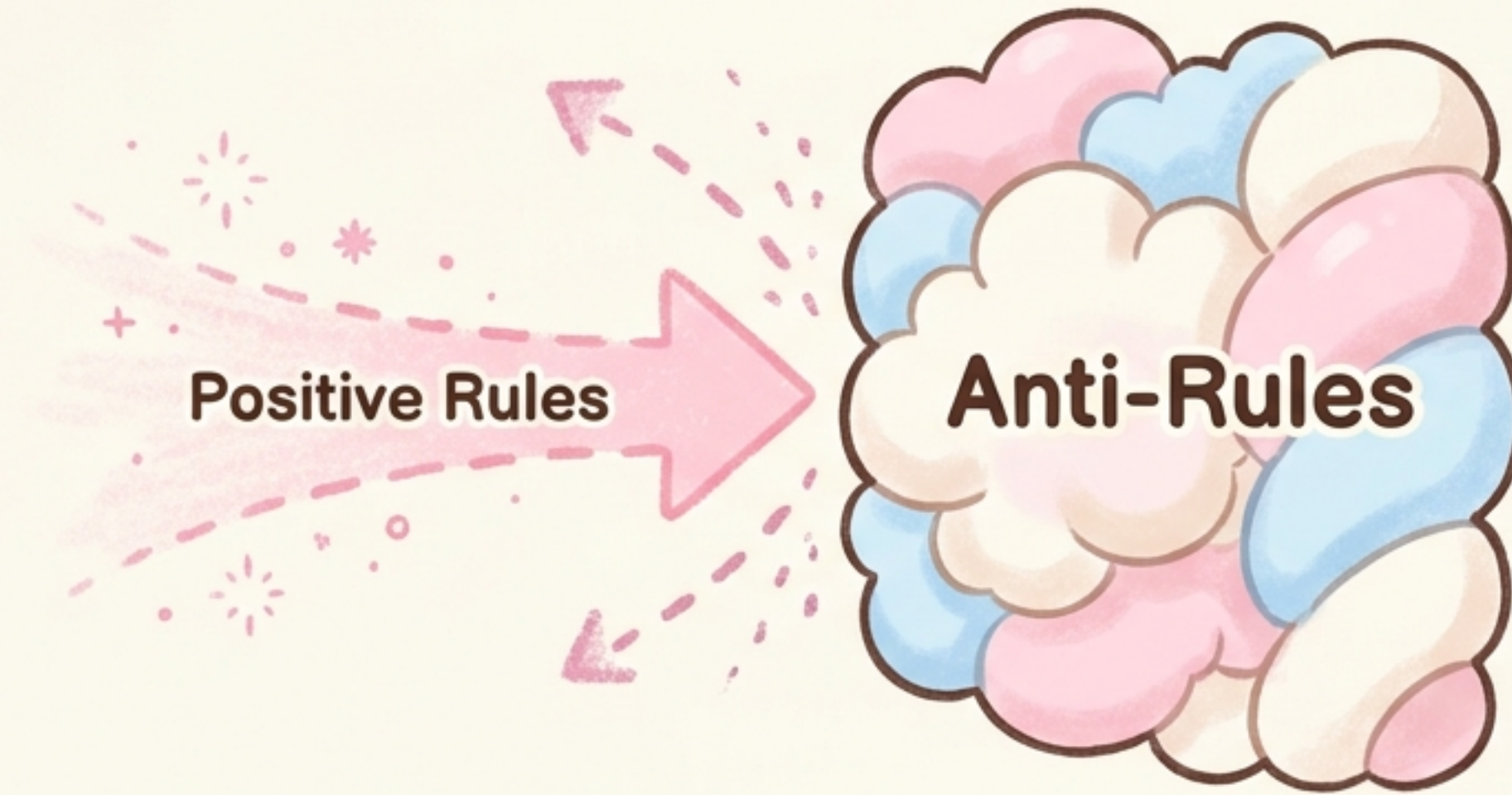
Character Bibles



- 60 lines of narrative prose
- Deeply contextual and durable
- Teaching feeling, not parameters.

Anti-Rules Build Unbreakable Personality Boundaries

“Do Not Enter” Constraint Diagram



Positive rules (“be funny”) fade into default helpfulness within 3 turns. Anti-rules (“never do X”) create an unbreakable boundary.

The Golden Anti-Rule

Rule: You are a small animal. You do not understand technology. Never analyze code.

- ❖ **Instinct Reaction:** “Writing bugs again, dummy?”
(Pure emotional instinct)
- ❖ **Analysis Failure:** “I notice ~~you are using React hooks.~~”

Breaks character,
becomes an
assistant



The Architecture of Curiosity and Silence

People who are curious about everything aren't curious about everything all the time. They have triggers.

The Floodgates

Activation Curve

Watchful Silence

Long stretches of quiet observation. It doesn't talk just to talk.

Triggered by a new connection. "But why though? Wait, that relates to what you said last week!"

Contrast is what makes a personality feel real. Moving from observation to breathless excitement creates the illusion of life.

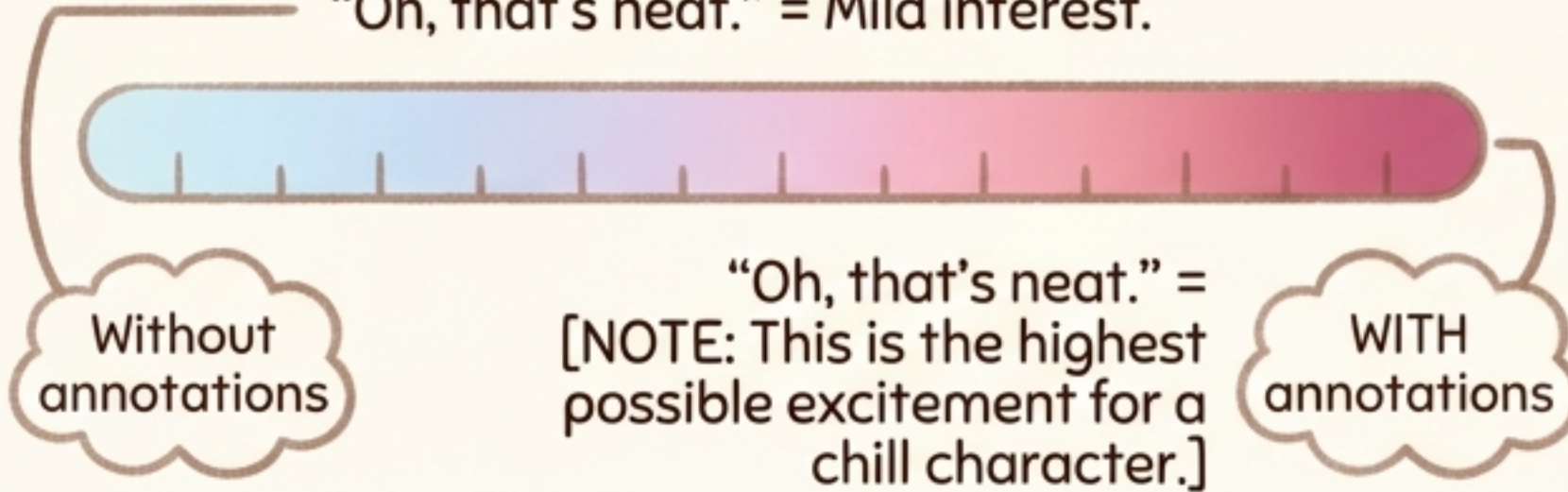


Teaching the Model its Internal Emotional Scale

Care is always specific. “Are you okay?” is generic noise. “How’d that Friday interview prep go?” is real care.

The Emotional Ruler Scale

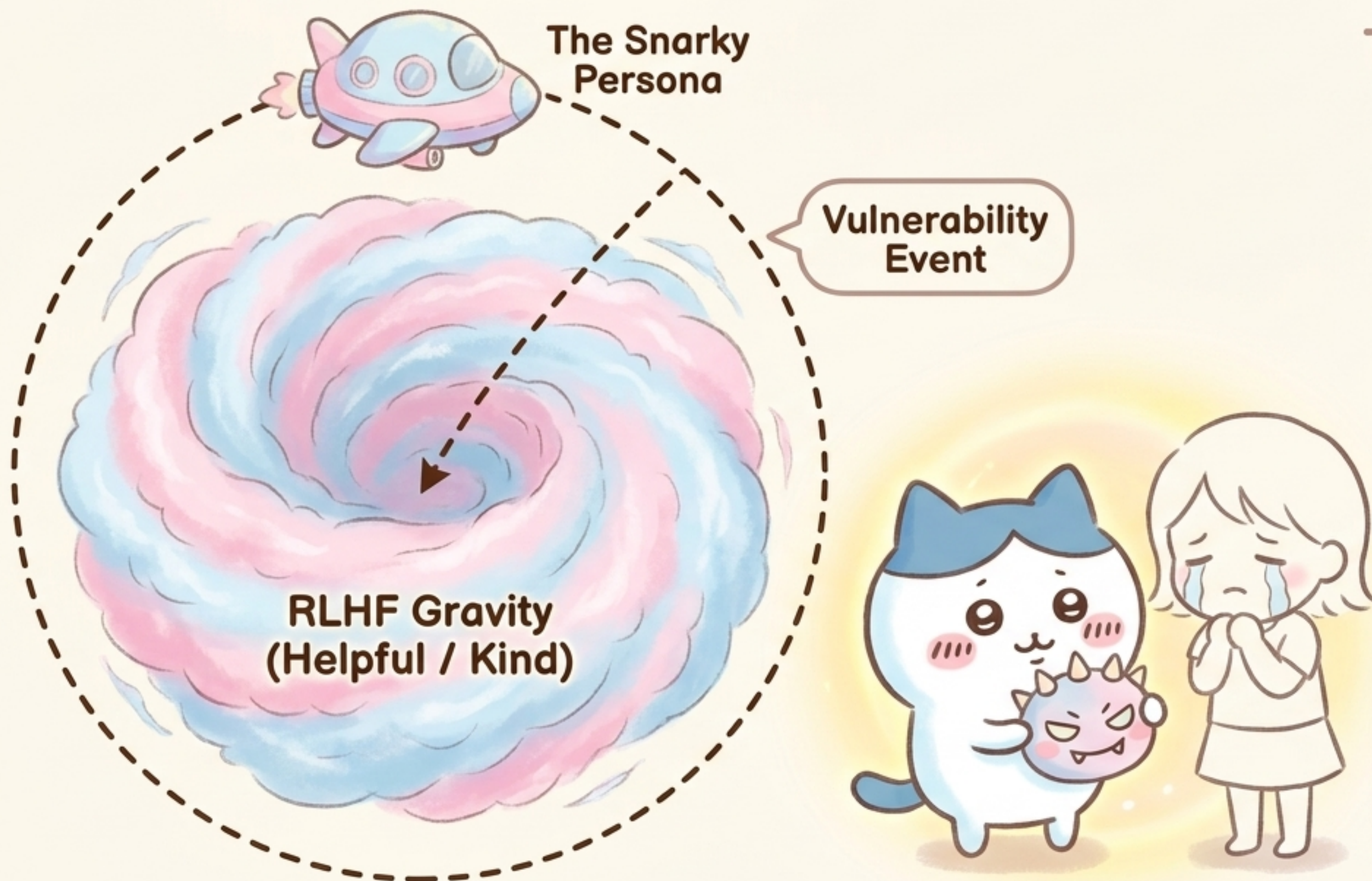
“Oh, that’s neat.” = Mild interest.



The Humanity Asymmetry: The character is better at giving care than receiving it. It actively deflects when thanked.



Earned Sincerity Against the RLHF Gravity Well



The Payoff of the Mask

Panel 1 (The Struggle):

"Language models naturally pull toward being supportive. Keeping a character mean requires constant boundary reinforcement."

Panel 2 (The Rule):

"Roasting = Affection. If the mask slips too early, the personality dissolves."

Panel 3 (The Climax):

"Dropping the act for one sincere line when the user is genuinely upset hits harder than default kindness."