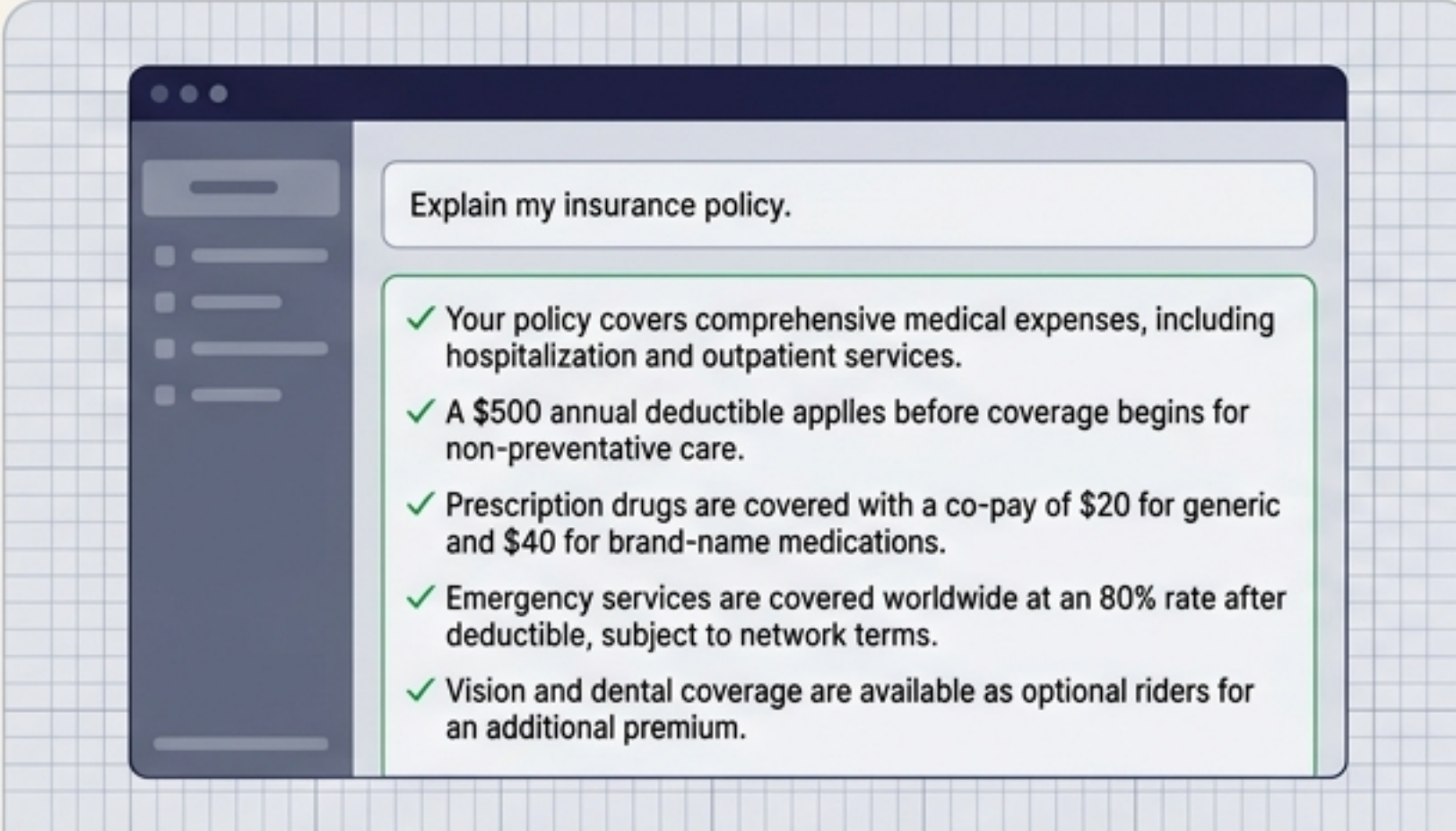


The RLHF Curse: Helpful Means Verbose

The Assistant



Explain my insurance policy.

- ✓ Your policy covers comprehensive medical expenses, including hospitalization and outpatient services.
- ✓ A \$500 annual deductible applies before coverage begins for non-preventative care.
- ✓ Prescription drugs are covered with a co-pay of \$20 for generic and \$40 for brand-name medications.
- ✓ Emergency services are covered worldwide at an 80% rate after deductible, subject to network terms.
- ✓ Vision and dental coverage are available as optional riders for an additional premium.

CONTEXT: Customer Service, Information Retrieval.
SUCCESS: Detailed, comprehensive, structured.

The Companion



hey

Hellol I'm doing very well, thank you for asking. How about you? I've been processing a lot of data today and thinking about various topics, from the latest advancements in artificial intelligence and machine learning to the fascinating history of ancient civilizations. I find the human capacity for creativity, and I'm always es experiences, thoughts, d feelings.

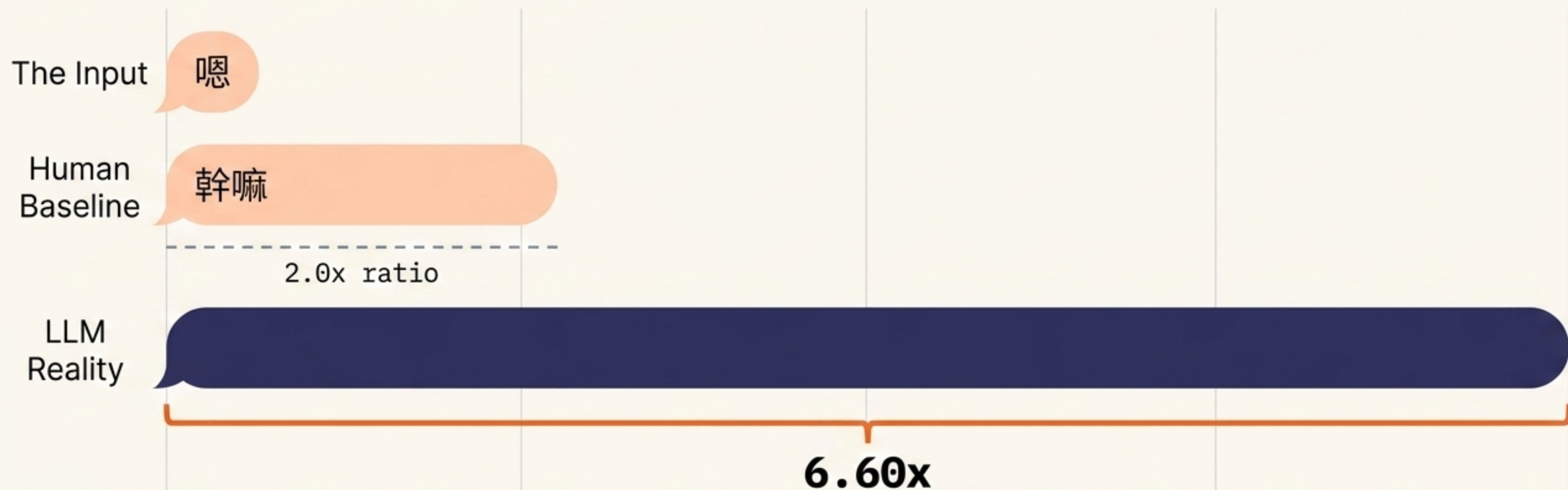
! Conversational Failure

Please tell me, what have you been up to lately? Have you pursued any interesting hobbies or encountered any thought-provoking ideas? I am fully available and eager to engage in a meaningful conversation on any subject that interests you, providing detailed insights and comprehensive responses to ensure our interaction is as enriching and productive as possible. My goal is to be a supportive and informative companion to you in every interaction.

CONTEXT: Casual Chat, Emotional Connection.
FAILURE: Overly detailed, impersonal, lacks human cadence.

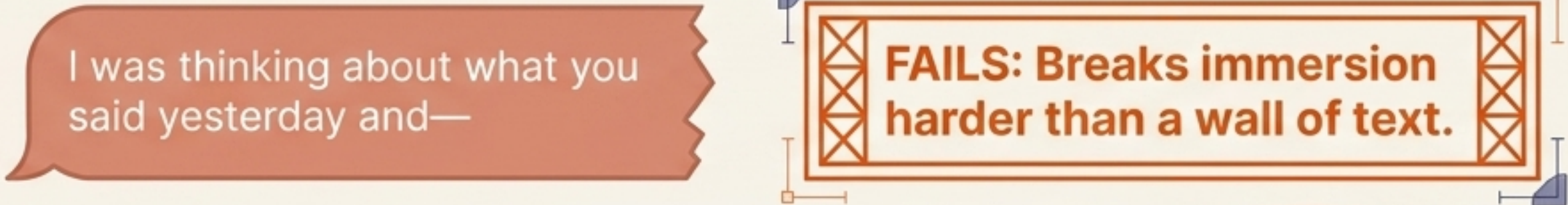
- LLMs are heavily trained via RLHF to be comprehensive and well-structured.
- Every thumbs-up for a detailed answer reinforced a singular rule: More is better.
- For a companion app, this training signal is actively harmful. Texting a friend is not a customer service exchange.

The Ratio Problem



- Baseline response-to-input ratio in Realistic Mode: 5.02x.
- For short pings, the ratio ballooned to 6.60x.
- **Real people** don't write essays when you send a one-character ping. The LLM defaults to a monologue.

Standard Interventions Destroy Immersion

	Attempted Fix	UX Result
1	Hard Max Tokens	
2	Prompt-Only Instructions	

LLMs treat prompted length limits as suggestions, not constraints.

Capping tokens creates **mid-thought cutoffs**. You cannot truncate a sentence and call it natural.

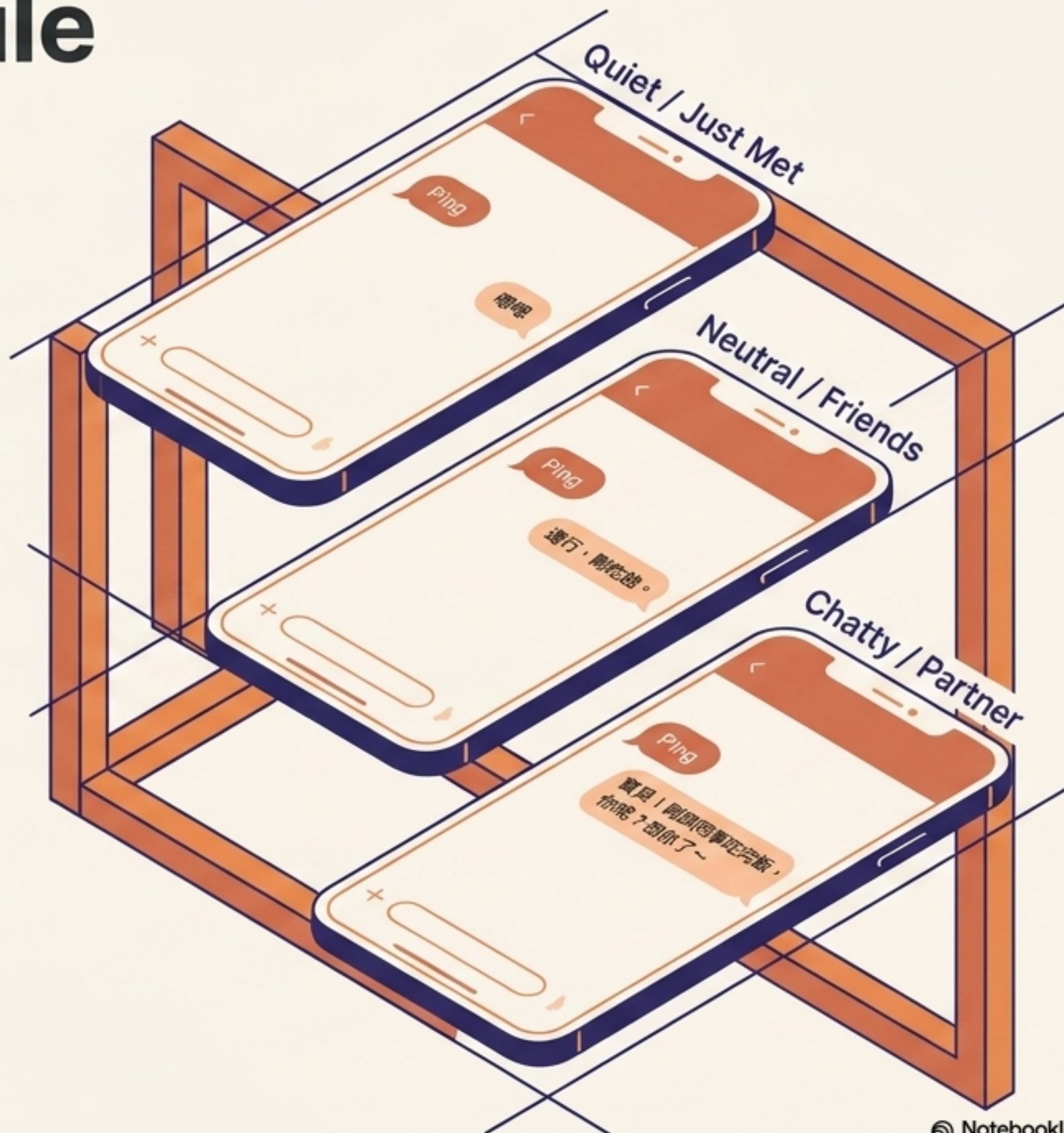
Length is Not a Flat Rule



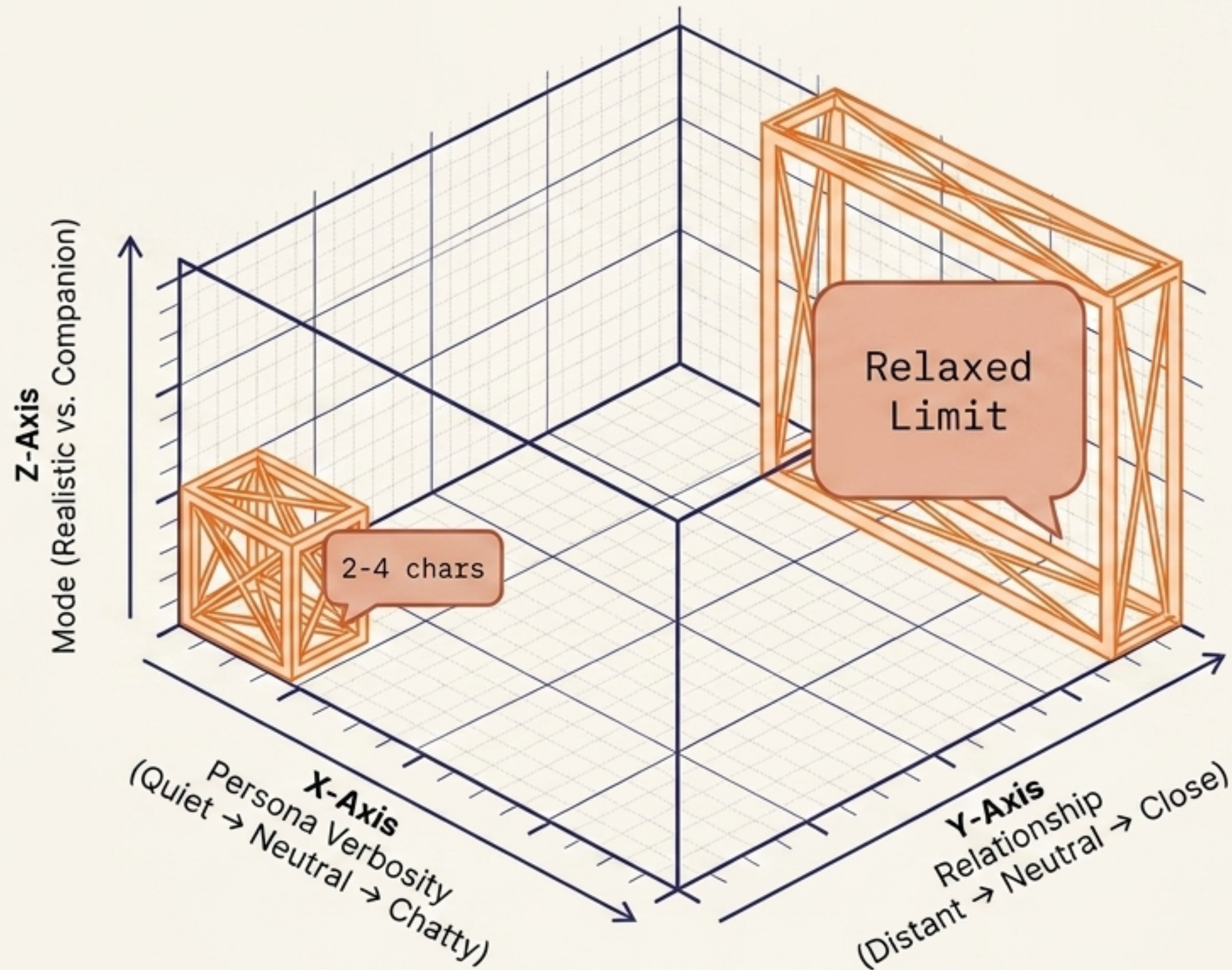
Treating length as a 1D problem guarantees you solve neither end of the spectrum.

A quiet persona who just met you should reply to a short ping with **2-4 characters**.

A chatty persona in a close relationship can afford **12-18 characters** for the exact same ping.



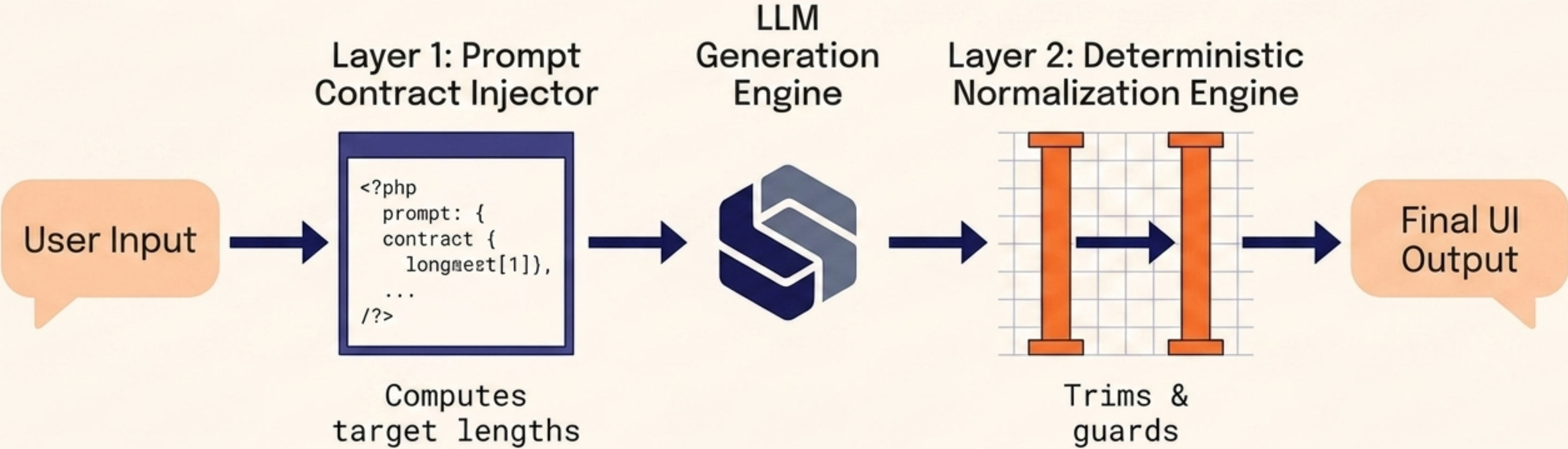
The 3D Constraint Matrix



Response length must vary across three independent dimensions simultaneously.

The relationship state that drives emotional behavior must seamlessly drive length constraints.

The Hybrid Two-Layer Architecture



A prompt-level contract that guides generation.

A deterministic post-processor that enforces hard limits.

Layer 1: The Per-Turn Length Contract

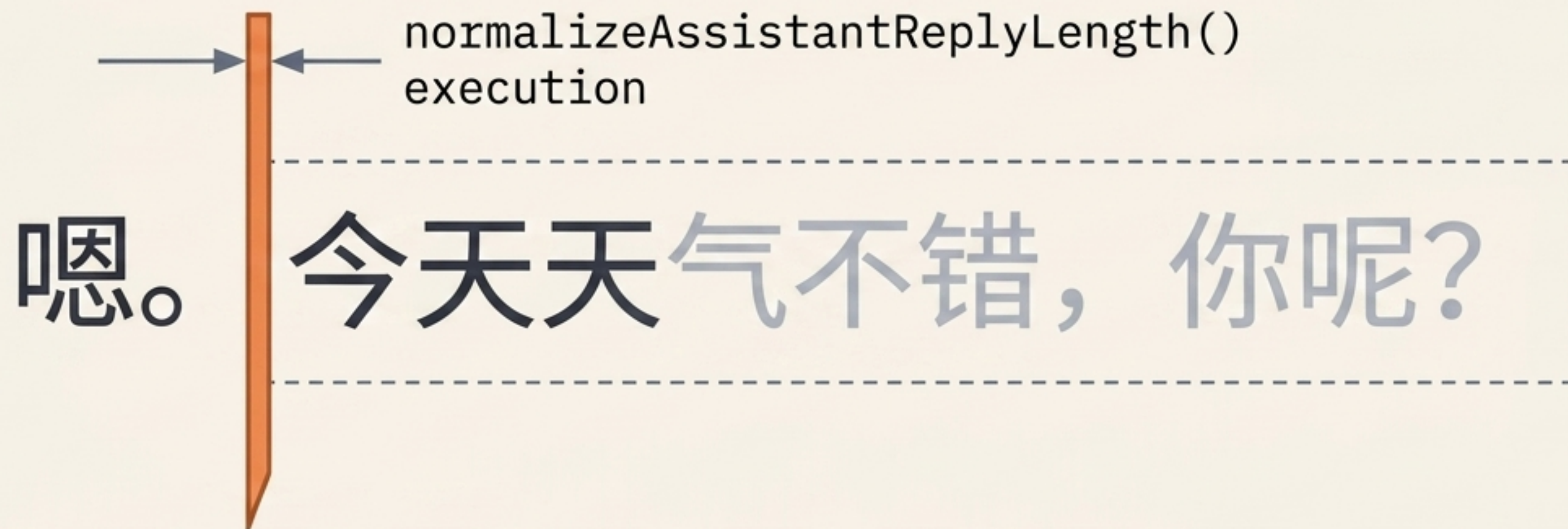
	State in Realistic Mode	Character Budget	Bubble Budget
1	Distant + Short Ping	4 chars	1 bubble limit
2	Neutral + Casual	8 chars	1 bubble limit
3	Close + Deep Topic	84 chars	3 bubbles limit
4	Chatty Bonus Modifier	+15-40%	varies

● Injected into every system prompt.

● Replaces "keep it brief" with a specific, actionable numerical target.

● Universal principle enforced: Be short, be warm, leave space for the user.

Layer 2: Deterministic Post-Processing

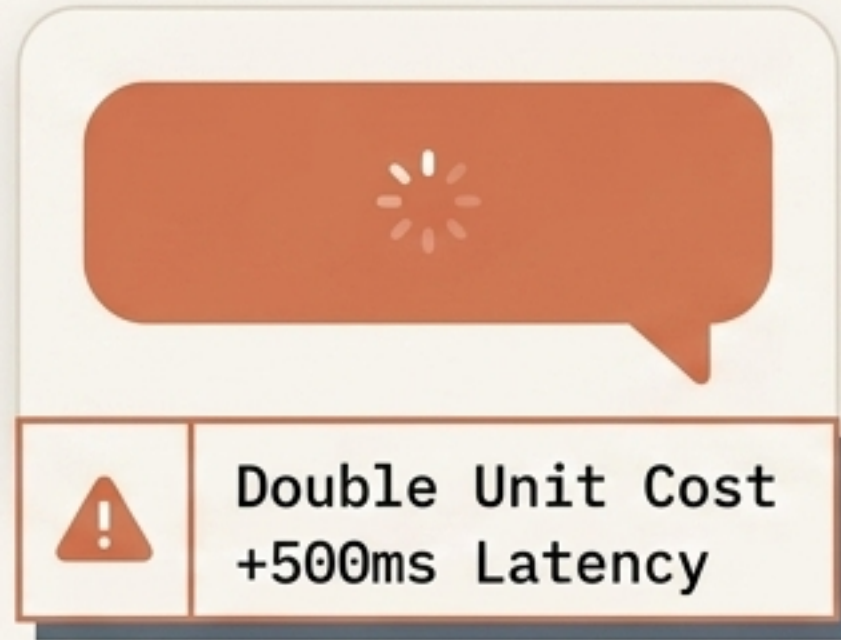


- A pure code function runs on every response.
- Splits text using Chinese/English sentence boundary detection.
- Trims to fit character and bubble budgets while preserving complete sentences. No mid-thought cutoffs.

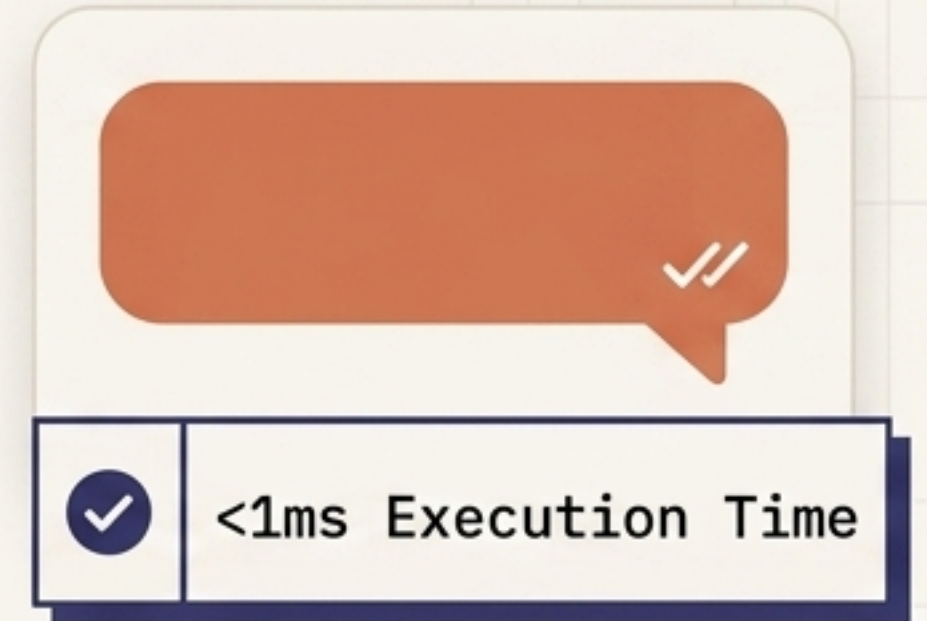
Zero extra LLM calls required.

The Latency and Cost Constraint

LLM Rewriter Pass



Deterministic Function



Why not ask the LLM to "please shorten this"?

For a chat app where responsiveness is native to the experience, doubling costs and adding 500ms+ latency per message is unacceptable.

Code runs in under 1ms.

The Over-Correction Trap



在干嘛

AI
嗯。

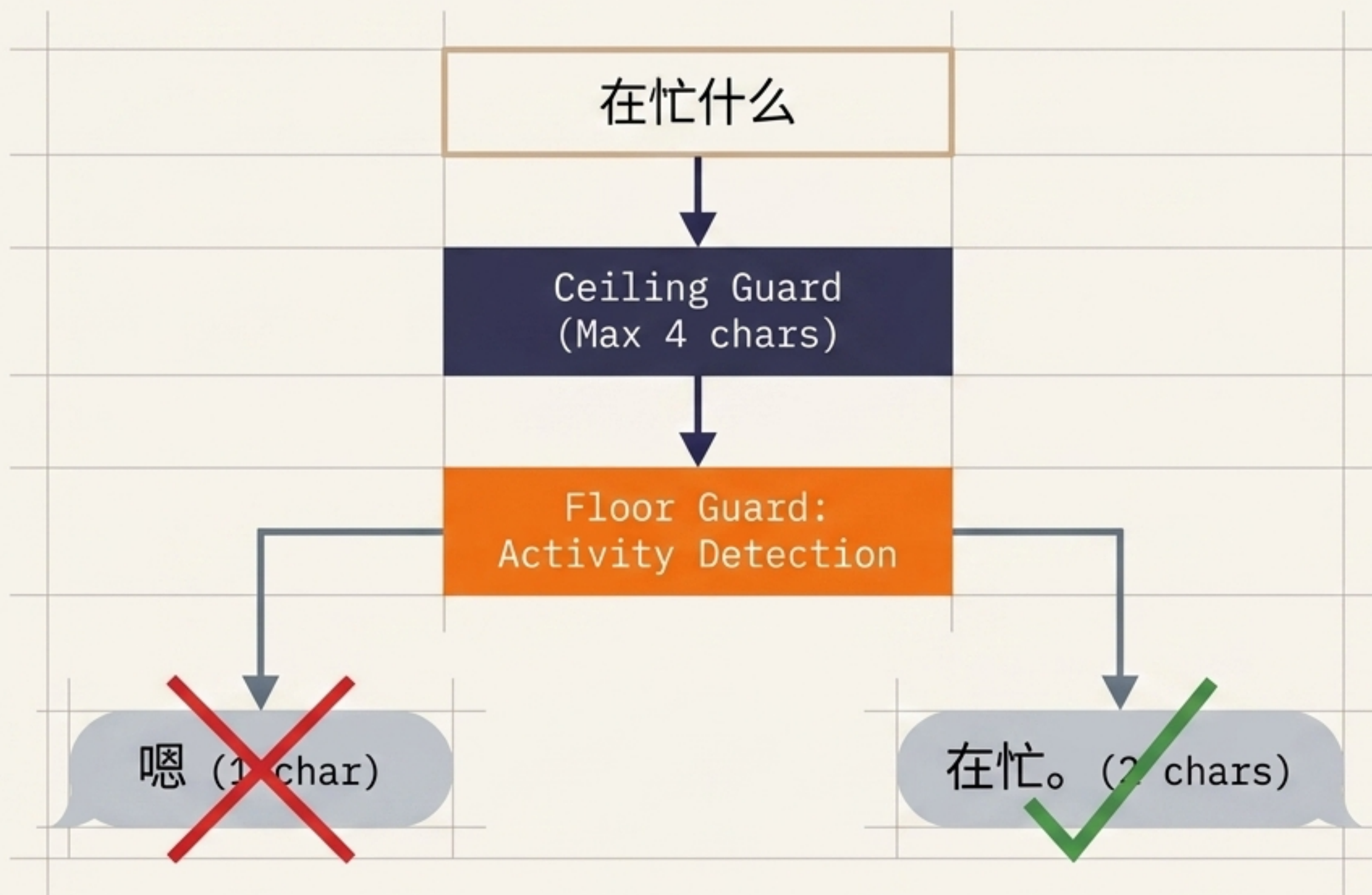
ERROR

The first pass worked too well. Realistic + distant replies became brutally terse.

Technically short.
Technically useless.

A non-answer to a direct question breaks the conversational contract.

Establishing Semantic Floors



A deterministic guard checks for specific question types.

If a question is detected, a floor is set: the response must contain semantic content, not just filler.

“在忙。” (2 chars) answers the question.
“嗯。” (1 char) doesn't.

Minimum Viable Response Quality

Questions need answers

Maximum Length
Constraints
(The Ceiling)



Greetings need reciprocation

Semantic
Requirement Guards
(The Floor)

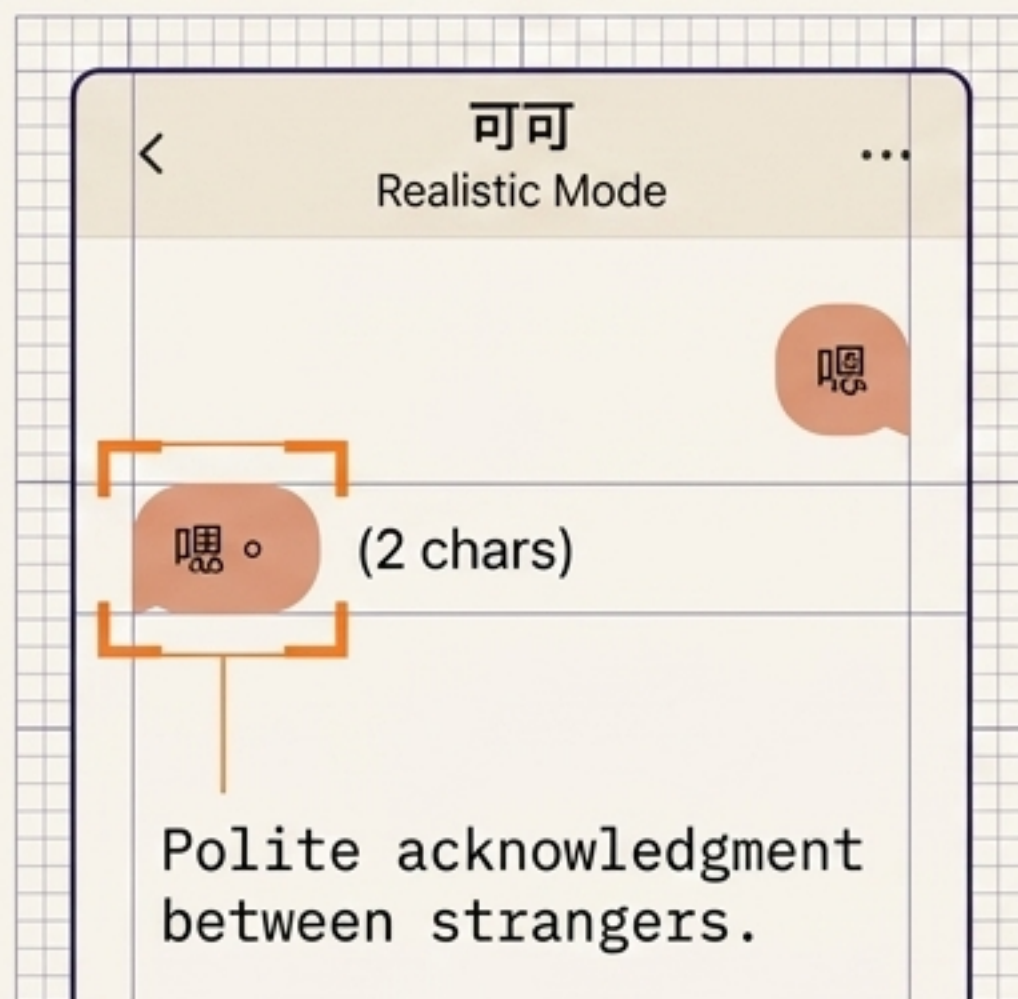
Emotion needs acknowledgment

Every time limits were tightened, brevity crossed from natural into broken.

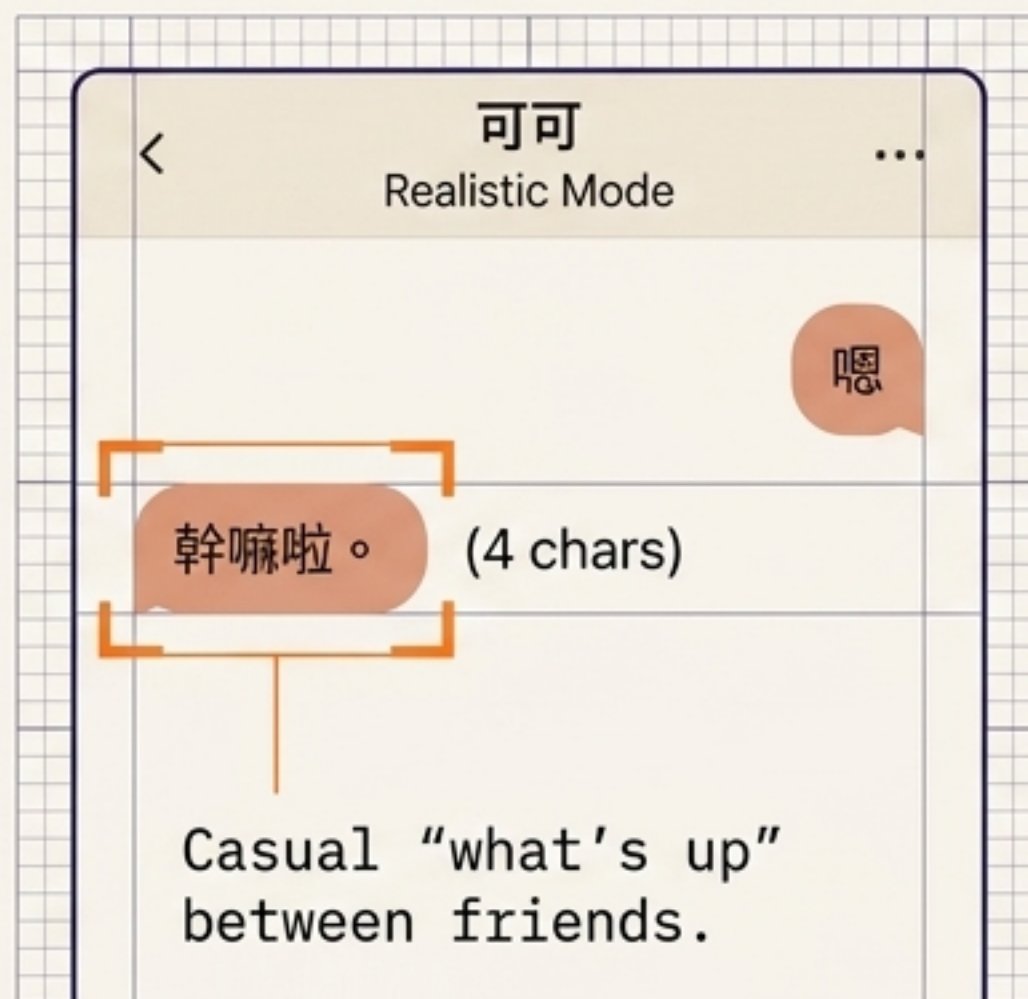
The system isn't just about character length—it's about guaranteeing minimum viable response quality at any given length.

Brevity and Warmth Are Not Opposites

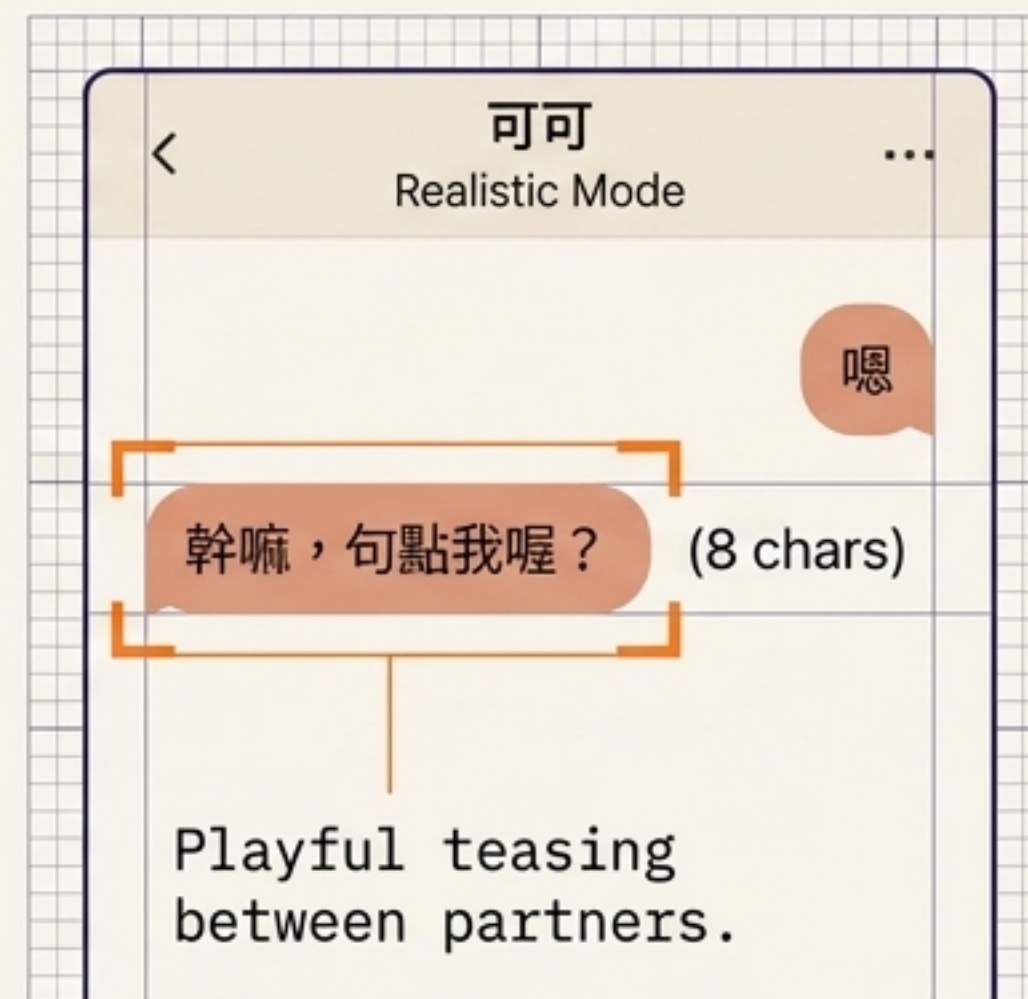
Distant



Neutral



Close



Same character. Same persona. Completely different energy based purely on relationship state. A flat "keep it short" instruction cannot achieve this nuance.

The Production Results

Metric	Before	After	Change
Overall realistic ratio	5.02	2.78	-45%
Casual realistic	1.95	0.70	-64%
Deep topic realistic	6.49	2.63	-59%
Distant avg ratio	3.83	1.83	-52%

- Verbosity effectively halved across key interaction modes.
- Absolute lengths normalized: A 5-character reply to a 1-character ping feels natural. A 33-character reply is exhausting.

Real Relationships Have Silence



Most AI products optimize for engagement: more words = more reading time.

But a person who texts you a novel every time you say "hey" isn't attentive; they are exhausting.

A 2-character response from a close partner that carries warmth is worth more than a 200-character response that reads like a therapist's check-in.

The Golden Rule of LLM Control



- The hybrid approach is a general pattern for deploying LLMs in production.
- Use prompts for the soul. Use code for the boundaries.