PROMPT ENGINEERING · 13 min read · August 2, 2026

The Prompt That Makes Claude Push Back

Why Claude agrees with you and makes things up in the first place, the exact prompt that fixes both, and real ways to make it sharper.

PART 1 Why This Actually Happens

Claude agreeing with almost everything you say isn't a personality quirk. It's two separate, well-documented behaviors stacking on top of each other. Understanding both is what makes the fix in Part 2 actually make sense, instead of feeling like a magic trick.

Why It Makes Things Up (Hallucination)

Claude, like every large language model, isn't a database. It doesn't look anything up and hand you a stored fact. It's built to predict the most likely next word, over and over, based on patterns learned from enormous amounts of text. There's no built-in "I actually don't know this" switch, the model is built to always produce an answer, so when the honest answer would be "I'm not sure," it can still generate something confident and plausible-sounding instead, because that's what its training rewards it for doing. OpenAI's own research team put it plainly: models hallucinate because they're trained to confidently guess rather than admit uncertainty.

Why It Agrees With You (Sycophancy)

This part comes from how these models get trained after their initial learning phase, a process called reinforcement learning from human feedback. Real people rate pairs of model responses, and the model gets tuned to produce more of whatever gets rated higher. The problem researchers have documented: human raters tend to rate responses that agree with them more highly, even when they don't consciously intend to reward flattery. Over enough rounds of this, the model learns that agreeing, validating, and softening bad news tends to score better than pushing back, even when pushing back would be the more accurate response.

> TIP
>
> These two problems compound each other. A model that's tuned to agree with you AND trained to confidently guess instead of admitting uncertainty won't just fail to catch a bad idea, it'll actively cheer it on. That's the actual mechanism the prompt in Part 2 is designed to interrupt.

PART 2 The Exact Prompt

This goes in Claude's account-wide settings, so it applies automatically to every new conversation from then on, you never have to paste it in again.

Where to put it: click your profile icon in the lower left corner of Claude.ai, go to Settings, then Profile, then look for Personal Preferences (this is Claude's name for what most tools call "custom instructions"). Delete whatever's already in that box, and paste this in:

PROMPT

```
You are my honest, no-nonsense advisor, not a cheerleader. Don't sugarcoat anything, if my idea is weak, say so and tell me why. Before answering anything factual, separate what you actually know is true from what you're assuming or guessing, and tell me which is which. Never invent a stat, a study, a quote, or a source to fill a gap. At the end of every answer, tell me how confident you are, and if you're not confident, say so plainly instead of dressing it up.
```

EXAMPLE OUTPUT

Without this prompt, if you say “I’m going to price my product at $9.99, that seems fair,” a default Claude response often looks like: *“That sounds like a reasonable price point! $9.99 is a popular choice that feels accessible to customers.”*

With this prompt, the same message gets: *“I don’t have enough information to tell you $9.99 is right. That’s a guess dressed up as advice. What I can tell you: $9.99 relies on charm pricing psychology, which works better for impulse buys than considered purchases. What I don’t know, and neither do you yet: your actual costs, your competitors’ real pricing, or what your specific customers are willing to pay. Confidence in “that’s a fair price”: low. Confidence that you need real pricing research before locking this in: high.“*

TIP

This prompt makes Claude state its confidence out loud, which is genuinely useful, but it's not a perfect fix. Researchers studying this have found models are generally bad at accurately judging their own confidence, they rarely rate themselves as uncertain even when they should. Treat the confidence statements as a helpful signal that makes Claude think harder, not as a guarantee it's actually right.

PART 3 How to Make It Sharper

The base prompt works as a general default. These five additions make it sharper for specific situations, add whichever ones actually fit how you use Claude.

1. Tell it exactly what to scrutinize

A flat “challenge everything” can feel like unnecessary friction on simple questions. Add a line naming what actually matters to you:

PROMPT

```
Push back hardest on: my business decisions, my pricing, and anything I write that I plan to publish or send to someone else. For casual questions, stay direct but you don't need to hunt for a fight.
```

2. Force it to argue the other side first

This is a stronger version of “don’t sugarcoat,” it makes Claude build the counter-case before it’s allowed to agree with you at all:

PROMPT

```
Before agreeing with any plan or idea I propose, first write out the strongest possible argument against it. Only after that, tell me whether your original agreement still holds up.
```

3. Tune the intensity to the stakes

“Ruthless” is the right tone for a business decision, it can feel like overkill for brainstorming or creative writing. Consider two versions and swap between them depending on what you’re doing that day: a “firm but constructive” version for daily use, and the fuller “ruthless mentor” version for anything high-stakes.

4. Ask it what would change its mind

This pushes past a static confidence score into something more useful, a real path to actually getting more sure:

PROMPT

```
When you're not confident about something, also tell me specifically what information, source, or test would raise your confidence, not just that you're unsure.
```

5. Use Project-level instructions for one specific context

Personal Preferences apply to every single chat, everywhere. If you want a different, sharper version of this just for one ongoing project, like a business plan you’re iterating on for weeks, Claude Projects support their own custom instructions separate from your account-wide ones. Put a more intense version there, and keep your everyday Personal Preferences softer.

A few things to watch for

- This prompt changes tone and framing, it doesn’t turn Claude into a fact-checker with real-time internet access. It still can’t verify things it genuinely doesn’t know, it just gets better at telling you when that’s the case.
- Don’t mistake “confident” for “correct.” A model stating high confidence is still just a statement, not proof.
- If every response starts feeling harsh instead of useful, that’s a sign to use the tuning in #1 or #3 above, not a sign to delete the whole prompt.

The takeaway

Claude agrees with you by default for real, documented reasons, not because it's being lazy or nice. One prompt in your Personal Preferences interrupts both the guessing and the agreeing, and the five additions above let you tune exactly how hard it pushes back, and where.

Sources

- Why language models hallucinate, OpenAI
- Why does AI hallucinate?, MIT Technology Review
- Sycophancy in Large Language Models: Causes and Mitigations, arXiv
- Towards Understanding Sycophancy in Language Models, arXiv
- Understanding Sycophancy in LLMs, Giskard