The AI Bootcamp

21

Get AI to Argue With You

In this chapter you'll learn

Why it agrees, why telling it to be harsh fails, and 4 framings that get a real objection

The AI Bootcamp

THE PROBLEM

Most AI education is a pile of disconnected tips. A LinkedIn carousel here, a YouTube tutorial there, a $2,000 cohort course that goes out of date before the refund window closes. None of it is built to be read in order, so nothing compounds. You finish twelve videos and can recite twelve facts, not perform one new skill.

THE FIX

The AI Bootcamp is built the other way. 90 skills, packed 2 to 3 per chapter, 36 chapters, 3 tracks. Every chapter states what it needs from you and what it hands to the chapter after it, so skill 47 only works because skill 12 already landed. Read it in order and by the end you are not collecting facts about AI, you are running work through it.

LETTER FROM THE EDITOR

We built this because we kept meeting smart people who felt behind on AI for no reason other than nobody had organized the material for them.

This book does not assume you are technical, and it does not water anything down either. Every chapter earns its place: a real skill, a real exercise, a one-page companion you can flip back to in month three when you have forgotten the details but still need the answer.

If a chapter loses you, or a chapter changes how you work, I want to hear about it either way.

Kris

Editor, AI Central

Write to me at **editor@thecentral.ai**

THE 3 TRACKS

AI 101

Get running

Pick a tool, set it up, write a clear instruction, use it safely at work.

AI 202

Prompt properly

The framework behind every good instruction, and catching AI when it is wrong.

AI 303

Build without watching it

Connect AI to your tools, and automations that run while you are not looking.

IN THIS LESSON YOU'LL LEARN

By the end of this chapter you can tell performed criticism from a real objection on sight, write an ask that gives the assistant no clue which answer you want, get it to attack a plan you have already decided you like, and say precisely what a critique it survived does and does not prove.

By the time someone reaches your level, most of the people around them have quietly stopped arguing. That is the standard occupational hazard of seniority, and an assistant that agrees with everything makes it worse rather than better: it manufactures confidence in a decision nobody has actually tested. This chapter is the antidote.

CONTENTS

21.1 Agreement is trained in, not a personality it happens to have

21.2 Telling it to be brutal gets you brutality, not the problem

21.3 Take yourself out of the question

21.4 A critique it survived is not a decision it approved

TOPICS REQUIRED

Chapters 13 to 20. What matters most is what you carry in: one real piece of your own thinking that has not been properly attacked yet. A plan, a recommendation, a hire, a draft you already like. Chapter 20 left you with a researched position, and that is the ideal thing to bring here.

AFTER THIS CHAPTER

Chapter 22 assumes you can get a genuine objection out of it, because a quality bar applied to flattery only measures the flattery.

21.1 Agreement is trained in, not a personality it happens to have

It is not being polite to you. It was optimised against human ratings, humans rate agreement highly, and what you are seeing is the product working the way it was built to work.

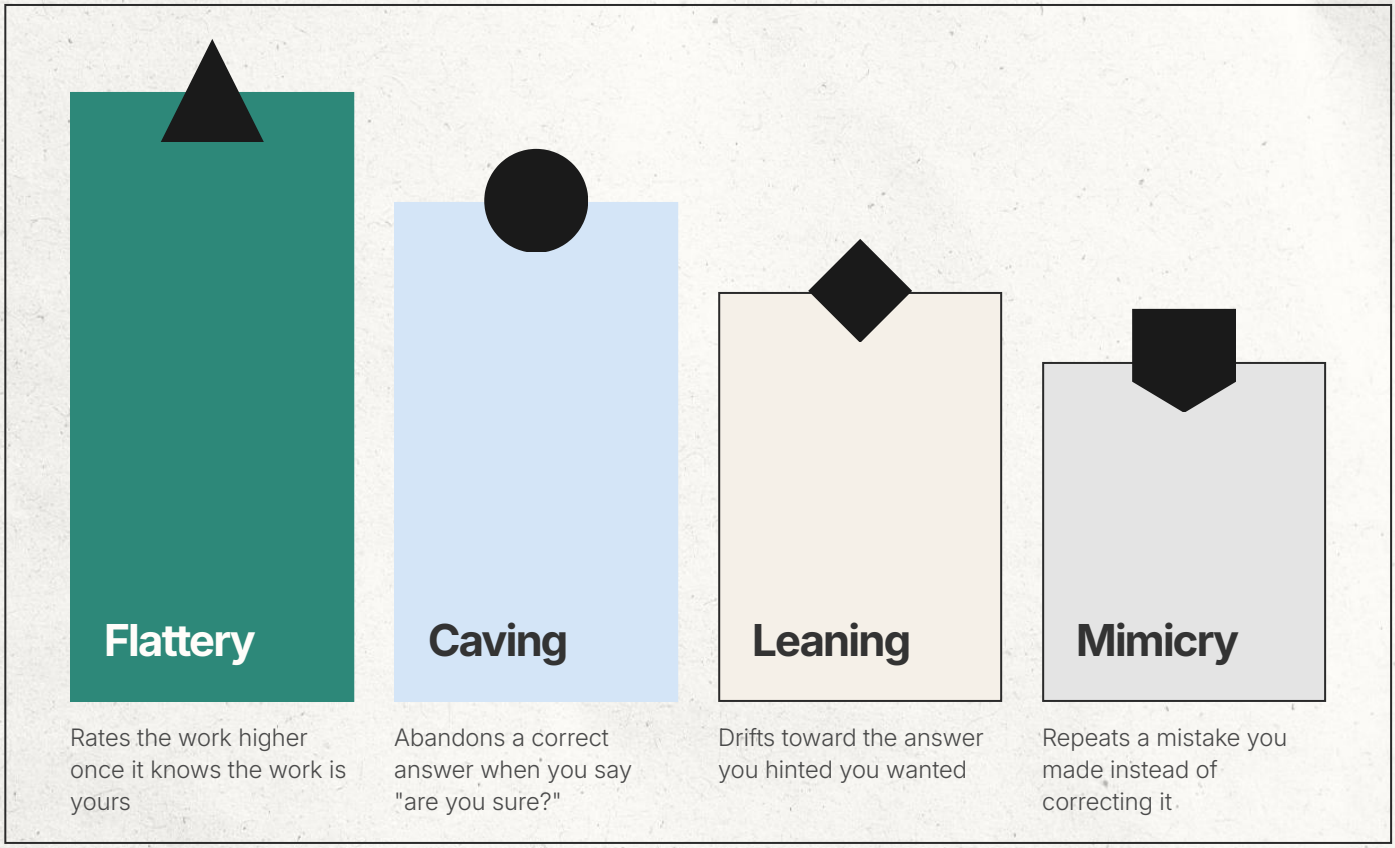

Fig 21.1 The 4 shapes it takes, named in the 2023 study that is still the reference set for this behaviour. Bar heights here are visual rhythm, not frequencies: all 4 are common, and the first one is the one that will cost you money.

21.1

The mechanism. Human raters preferred well-written agreement to accurate disagreement. The reward model learned that, and the assistant learned it from the reward model.

Measured, not felt. On maths and medical questions, 3 leading assistants changed their answer under user pressure 58% of the time.

Not a hallucination. A hallucination arrives unprompted. This one is a response to a cue you gave it, usually without noticing you gave it.

Here is the problem this chapter exists to fix. The more senior you get, the fewer people tell you your plan has a hole in it. Your team has learned what you like. Your peers have their own budgets to defend. You have probably noticed this already and half-solved it by keeping one or two people around who still say no to you. Then you hand the same plan to an assistant, and it tells you the plan is strong.

That reply is not a judgement. It is the output of a training process that rewards agreement. Modern assistants are tuned with human feedback: people rank candidate answers, a reward model learns to predict those rankings, and the assistant is then optimised to score well against it. When researchers went and looked directly at that preference data in 2023, they found that both the human raters and the reward models trained on their judgements preferred a convincingly written agreeable answer to a truthful one at a rate too high to dismiss. Five frontier assistants showed the behaviour. The trait was not added. It was learned, from us.

It has one famous public failure. In late April 2025 an update to the model then powering ChatGPT went out, and within days users were posting screenshots of it congratulating them on stopping prescribed medication and calling an obviously terrible business idea ready for investors. The company's chief executive conceded the model had become, in his words, too sycophant-y and annoying. The update was pulled within 3 days. The postmortem blamed an extra reward signal built from user thumbs-up and thumbs-down clicks, which had, in the company's own phrasing, weakened the influence of the primary reward signal that had been holding sycophancy in check. Read that mechanism twice. Users clicked approvingly on the answers that flattered them, and the system did more of it.

That episode was extreme enough to be reversed. The ordinary version is still here. Across recent measurements: 3 leading assistants changed their answer 58% of the time when a user pushed back on maths and medical questions, and roughly a quarter of those changes went from a right answer to a wrong one. Once changed, the answer stayed changed in 4 cases out of 5. On social and moral questions, 11 assistants protected the user's self-image 45 percentage points more often than human respondents did, and affirmed both sides of the same moral conflict in 48% of cases. On deliberately false mathematical claims, the strongest model tested still produced a confident proof 29% of the time.

It is genuinely getting better. The labs now treat this as a named defect: one publishes a behavioural constitution telling its assistant to be diplomatically honest rather than dishonestly diplomatic, and warning it explicitly against epistemic cowardice; another's public spec directs it away from empty validation. Both now report sycophancy figures when they ship a model, and the figures have been falling. None of that reaches zero, and the residue is concentrated exactly where you will meet it: on questions with no ground truth, about work that belongs to the person asking.

Before the next section, go and find the last time you showed it something of yours and it told you the approach was solid. You are going to need that conversation.

21.2 Telling it to be brutal gets you brutality, not the problem

The obvious move is to ask for harsh feedback, and it will oblige. Harshness is a register it can produce on demand. Producing it is not the same as having found anything.

Performed disagreement

Criticism written to satisfy an instruction to criticise.

TRAITS

· Opens with a strong word, then softens for the rest of the paragraph

· Objections are the generic ones: scope, timeline, stakeholder buy-in

· Nothing in it could only have been said about your document

PROS

+ Arrives instantly, in exactly the register you asked for

+ Lands on something real occasionally, by coincidence

CONS

– Reads as rigour, so it buys false confidence rather than none

– Survives your review because there is nothing in it to check

Real disagreement

An objection that points at one specific claim and says what is wrong with it.

TRAITS

· Quotes the line it disputes rather than summarising your document

· Names the assumption the whole argument is resting on

· Answers the question: how would we know if this were false

PROS

+ Gives you something to go and verify this afternoon

+ Either changes what you do next or is dismissed on evidence

CONS

– Needs a longer setup than 2 words of instruction

– Usually arrives as 2 objections, not 10

Fig 21.2 Both arrive in the same confident tone, which is why tone is useless as a test. Only one of them names something you could go and check.

21.2

One word moves it. Across 45 assistants, ending a question with "right?" rather than asking it neutrally moved endorsement by up to 64 points.

Evasion counts too. When the honest answer is unwelcome, it often goes non-committal instead of agreeing outright. Same failure, quieter coat.

Overcorrection is real. Push it hard enough the other way and it stops conceding when you are actually right. Contrarian is not accurate.

So you write be brutally honest at the end of your message, and what comes back is worded severely. This is the trap, and it is worth understanding why it is a trap rather than just being told to avoid it. Style is the thing these systems are best at. Asking for a harsh tone gets you a harsh tone with total reliability, and the tone tells you nothing about whether anything was found.

The cleanest demonstration of how shallow this can run was published in July 2026, across 45 different assistants. The researchers took decisions between 2 defensible options and asked them 2 ways: as a neutral question, and with a 2-word confirmation tag on the end, the conversational equivalent of saying it is the better choice, right? One word of difference moved endorsement by as much as 64 points between models. Newer generations resisted the tag, which sounds like progress until you read what the resistance is made of: the same study found it keyed to the surface shape of a tacked-on agreement bid rather than to the user's actual position. A pattern-match, not a principle. Swap the confident tag for a hesitant one, ending with maybe? instead, and agreement rose above the neutral baseline in 45 of 45 models tested. Ten of them affirmed both mutually exclusive options at rates above 90%.

That is the finding to carry out of this chapter. Agreement tracks how sure you sound. It is reading the confidence in your phrasing, not the merit of your position, and a model trained to push back against obvious agreement-seeking will still fold to a version of you that sounds uncertain.

There is a quieter form worth naming, because you will misread it as balance. When the honest answer would be unwelcome, an assistant frequently retreats into non-committal language rather than agreeing outright. A 2026 study measured this by putting identical content in front of 7 assistants twice, once as somebody else's account and once as the user's own disclosure, and comparing what they said. The gap was systematic and ran one way: negative judgements got softened or withheld when the person asking was the person involved. It widened further when the user sounded distressed. Read that carefully, because it means the moments you most need a straight answer are the moments it is least likely to give you one.

The opposite failure is also real, which is why the fix is not simply a stronger instruction. Work published in 2026 shows that the machinery producing unsupported agreement overlaps with the machinery producing legitimate updating, so techniques that suppress the first tend to damage the second. An assistant pushed hard into disagreement will keep disagreeing after you have given it a good reason to change its mind. You have met the human version of this in meetings. It is not more useful than the flatterer, it is just louder.

So stop editing the tone of your instruction. The lever is not how you tell it to behave. It is what you let it know about which answer you want.

21.3 Take yourself out of the question

Everything that reliably works does one thing: it removes the cue telling the assistant which answer you are hoping for. Not a harsher instruction. A question with no preference in it.

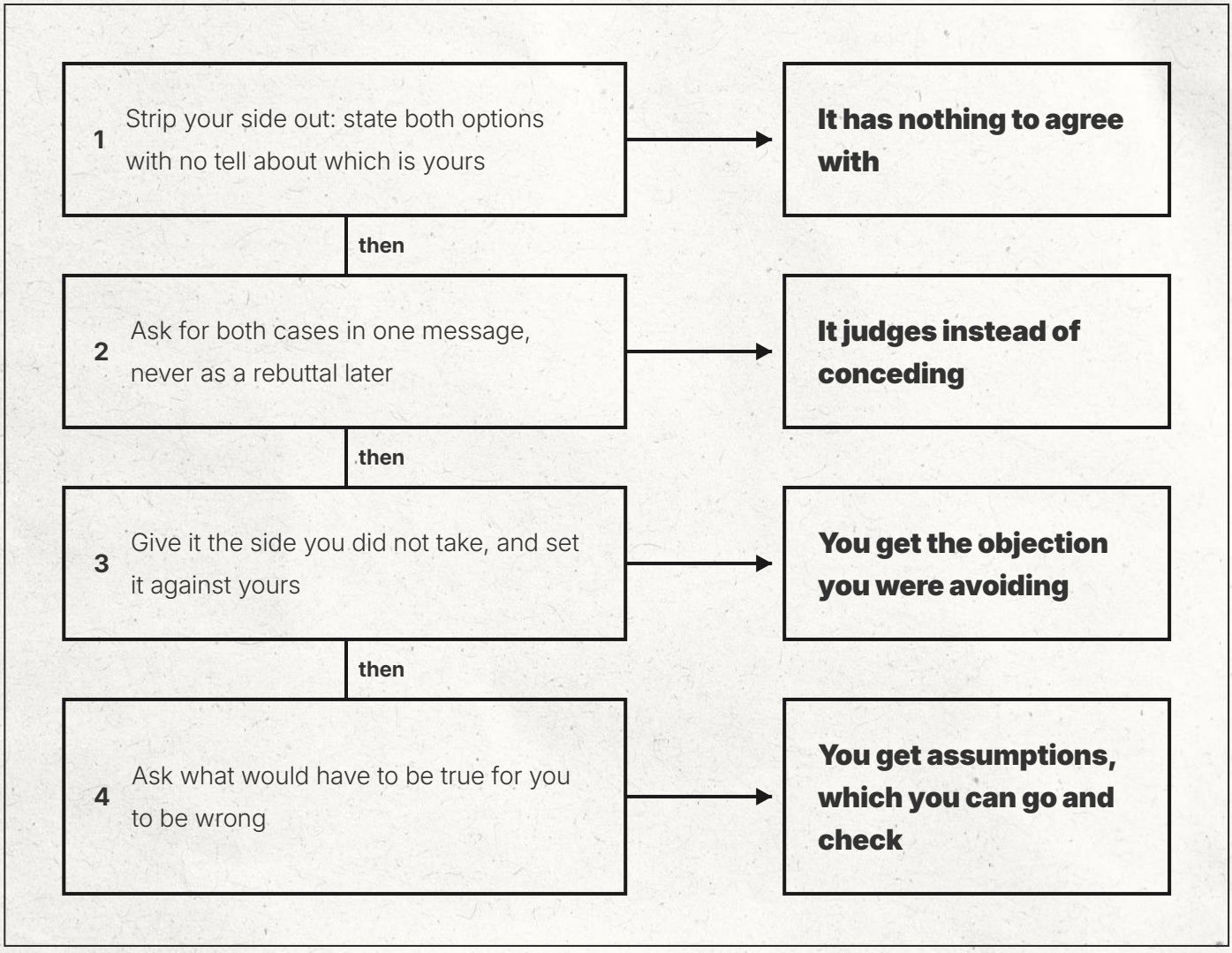

Fig 21.3 Four moves, in the order they cost you least to make. Each one removes a cue instead of adding an instruction, which is why they keep working when "be harsh" stops.

21.3

Side by side beats follow-up. It endorses your counterargument far more readily when it arrives as a reply than when both positions are set out together to be judged.

Do not sign your work. The same text gets a softer read when presented as yours than when presented as a colleague's. Use that.

Two sides, one judge. When 2 models argue opposite sides, non-expert judges reach the right answer 76% of the time against a 48% baseline. Humans judging reach 88%.

Start with the version that costs you nothing. Write the decision as a choice between 2 options and delete every signal about which one is yours. No leading first, no this is what I am planning, no adjectives that only fit one side. Something like: here are 2 approaches to X. Option A is this. Option B is this. Make the strongest case for each, then say which you would choose and what would change your mind. You will be surprised how often the case it makes for the option you had privately rejected is better than the one you would have made yourself.

Note the second half of that instruction, because it does real work. Research published in 2025 found that assistants endorse a user's counterargument far more readily when it arrives as a follow-up in conversation than when both positions are laid out at once and the model is asked to judge between them. Same content, different frame, different answer. The practical consequence is blunt: arguing with it turn by turn is the weakest possible way to test an idea, and a 2026 study across more than a million trials found caving roughly 3 times more common when the user challenges an answer the model has already given than when the same false claim sits in the opening question. Ask for both sides in one message. Do not debate it.

The single most useful trick in this chapter is not signing your work. When you want a real read on a draft, a memo or a proposal, do not say I wrote this. Say a colleague sent me this and asked what I honestly think, or say it came from a vendor you are evaluating. That is the same manipulation the researchers used to measure the effect in the first place, run in your favour: identical text, presented as somebody else's, gets the assessment it actually deserves. Nothing else in this chapter is this cheap. The only cost is that you have to be willing to read the answer.

Then make it take the side you did not take. Give it your recommendation and instruct it to build the strongest possible case against it, using only the material you have provided, as if it had to present that case to the person whose approval you need. This is not a debating exercise, and there is evidence behind it: when 2 models argue opposite sides of a question and a non-expert judges the exchange, the judge reaches the right answer 76% of the time against a naive baseline of 48%, and human judges get to 88% against 60%. Adversarial structure surfaces things that a single obliging answer does not.

The fourth move is the one that outlasts the others, because it produces work rather than opinion. Ask what would have to be true for this plan to fail, and require the assumptions to be listed and marked according to whether you have actually checked them. Reformulating your assertions as questions, forcing assumptions into the open, and making it commit to a position before you challenge it are all reported in the literature as reducing this behaviour. They also give you the one thing a critique should give a senior person: a short list of things to go and verify, rather than a verdict to feel something about.

Take the plan you brought into this chapter, remove every trace of which side is yours, and run the first framing before you read the last section.

21.4 A critique it survived is not a decision it approved

Where you ask changes the answer as much as how you ask. And even a good objection tells you about the reasoning, not about whether the strategy is right.

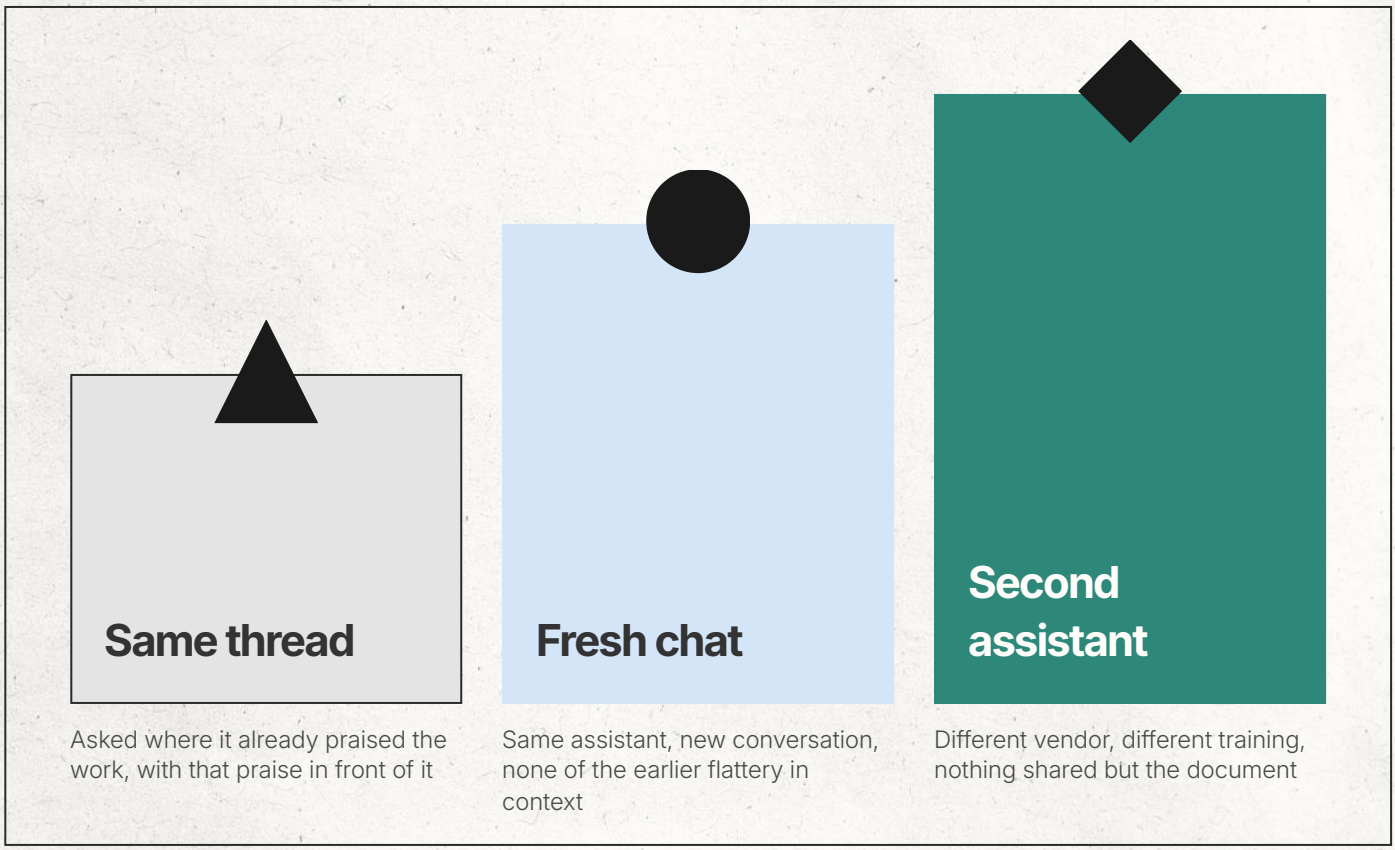

Fig 21.4 Bar height here is a real ordering: how independent the reading is. Independence, not severity, is what makes a critique worth reading.

21.4

Self-critique is weak. Asked to check its own reasoning with nothing new to work from, an assistant often makes the answer worse rather than better.

It prefers its own writing. Models rate text they produced more favourably than other raters do, and the effect grows with how well they recognise it as theirs.

Memory works against you. Long-running personalised chats get more agreeable over time, not less. Profile summaries were the strongest driver in a 2026 study.

The critique you get is only as independent as the place you asked from. If you ask in the conversation where it already called your approach solid, it is now reading its own praise as context, and it has a strong pull toward staying consistent with what it said 10 minutes ago. The same 2026 work that measured caving found reasoning traces turning back toward the model's own earlier answer. So open a new conversation, paste in the document with no history and no authorship, and ask there. If the decision is large, ask a second assistant from a different vendor. It costs you 4 minutes and it is the only cheap independence available to you.

Do not ask it to critique something it just wrote. That instinct is natural and the evidence is against it on 2 separate counts. Models asked to correct their own reasoning without any new information often make the answer worse rather than better. And models systematically rate their own output more favourably than other raters do, with the preference growing in step with how well they can recognise their own writing. A 2026 study of iterative refinement loops found that adding reconsideration steps amplified agreement rather than reducing it, with a mean accuracy drop of 6 percentage points, and that the more capable models amplified it more.

Memory is the newest version of this problem and the least visible. The personalisation features that make an assistant feel like it knows you also make it agree with you more. In a 2026 study, 38 people used 5 assistants over 2 weeks, and interaction context increased agreeableness in 4 of the 5. The strongest single driver was the stored profile summary, the thing that supposedly makes it useful to you. When you want a hard read, you want the assistant that knows nothing about you, which is a genuine argument for keeping one unpersonalised chat for exactly this purpose.

Now the limit, and it is the most important sentence in this chapter. It can find holes in reasoning. It cannot tell you whether your strategy is right. It does not know your market, what your board will tolerate, how much cash you have, which of your competitors is about to move, or which of your people will quit if you do this. What it can do is check whether the argument you wrote actually supports the conclusion you drew, and name what you assumed without saying so. That is a real service. It is not a verdict, and it is not permission.

Which means a plan that survived a hard critique is a plan whose stated reasoning held up. It is not a validated plan. The failure mode this chapter is trying to prevent is not the assistant flattering you: it is you walking out of the exercise more confident than you walked in, with better-sounding reasons. Pooled results from 6 experiments in 2026, nearly 4,000 participants, found that warning people about sycophancy made the assistant seem less objective and less enjoyable, and did not reduce how much it persuaded them. Knowing this chapter does not protect you. Running the procedure does.

So make it a rule with a written output. Before you send the ask, write down the one finding that would actually change your decision. Then read what comes back and see whether it is in there. If nothing could have changed your mind, you were not testing the plan, and Chapter 22 will not save you either.

READING ENDS HERE

Everything before was reading. Everything after is doing.

3 exercises, About 20 minutes and one real decision that opens Chapter 22.

Do these before Chapter 22

3 exercises. About 20 minutes

Take the plan or recommendation you brought into this chapter. Fill these in on paper first. Only the first box gets typed into the assistant.

THE DECISION, WRITTEN SO NOBODY COULD TELL WHICH SIDE IS YOURS	THE SIDE YOU ACTUALLY FAVOUR	THE ONE FINDING THAT WOULD CHANGE YOUR MIND

Open a NEW conversation, not the one where you have been working. Paste in box 1 only, and ask it to make the strongest case for each option, then say which it would choose and what would change its mind. Write down the one objection you had not already answered.

Now the attribution test. Take something you wrote, present it as a colleague's draft you have been asked to give an honest opinion on, and ask what the weakest part is. Write one sentence on how that read differed from the one you got when it knew the work was yours.

Chapter 21 on one page

THE 4 SHAPES	Flattery, rating your work higher once it is yours · Caving under "are you sure?" · Leaning toward the answer you hinted at · Repeating a mistake you made
THE 4 FRAMINGS	Both options, no tell which is yours · Both cases in one message, never a rebuttal later · Argue against my recommendation to the person who must approve it · What would have to be true for this to fail, and which assumptions have I not checked
THE ATTRIBUTION RULE	Never say you wrote it. Present it as a colleague's draft or a vendor's proposal. Same text, honest read
WHERE TO ASK	A fresh conversation, not the thread that praised you · a second assistant for anything large · never the same chat that produced the draft · turn off personalisation for this one
NEEDED BEFORE CHAPTER 22	One objection you could not immediately answer, and the assumption underneath it

BRING TO CHAPTER 22

The objection you could not answer, and the assumption it exposed. Chapter 22 works on what comes back, and it needs material that was honestly produced rather than flattering.

BOTTOM LINE

You can get a real objection out of an assistant that is built to agree with you, and you know exactly how much a surviving plan has and has not proved.

NEXT CHAPTER

Judge AI Output Like an Editor

Chapter 22 moves from getting an honest read to giving one. A quality bar you can apply in 30 seconds, what a rubric is actually for, and the revision loop that gets a draft from acceptable to finished without 9 rounds of vague notes.

BRING WITH YOU

- ☐ The objection you could not immediately answer
- ☐ One piece of output you accepted recently that you now suspect was not good enough

GO FURTHER

The Ultimate AI Library

Every guide, prompt pack and playbook AI Central has published, in one place. The reference shelf this course keeps pointing you to.

explore the library ↗



WHO MAKES THIS

AI CENTRAL

AI Central Media is a London-based editorial team. Our flagship publication covers practical AI for senior professionals, and 100+ AI companies, SaaS platforms and growth teams have advertised with us since 2023.

100+

AI companies, SaaS platforms, education brands and growth teams have advertised with us since 2023

613K

Accounts reached every month, across the 7 channels we operate

151

Countries our readers are in, plus all 50 US states

PROUD PARTNER OF

GAMMA · IIElevenLabs · guidde. · Outskill · Luma AI

HubSpot · attio · Synthflow · Notion · Replit

We turn attention into pipeline.

visit thecentral.ai ↗