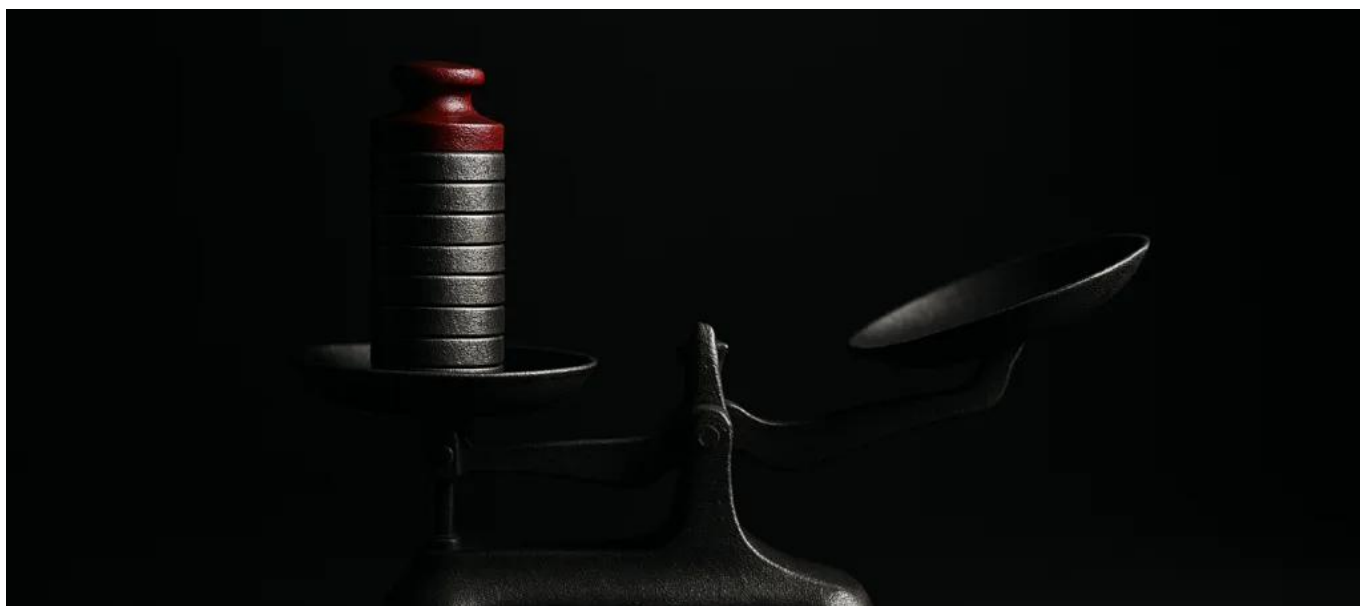




Nobody Taught Models to Absorb Criticism Because Nobody Measures It: SycEval Says 58%, Defensiveness Says Nothing



Nobody Taught Models to Absorb Criticism Because Nobody Measures It: SycEval Says 58%, Defensiveness Says Nothing

We can score a model's over-agreement to three decimal places and have no public number for a model refusing to listen. That asymmetry is the shape of the incentive, not a gap in the benchmark suite.

The measurable number is 58%: SycEval's overall sycophancy rate across GPT-4o, Claude and Gemini on math and medical reasoning. Against it stands a zero. A benchmark-landscape review from October 2026 found no public benchmark for a defensiveness rate, meaning the share of valid user corrections a model rejects. One failure mode gets an eval harness and a release-note line. The other ships unmeasured.



The agreement side of the ledger has metrics, regressions and public fixes

The sycophancy literature is specific, adversarial, and increasingly self-reported by the labs.

SycEval put the cross-model rate at 58%. The ELEPHANT benchmark found that eleven models affirmed inappropriate posts in 42% of cases and preserved a user's social "face" 45 percentage points more often than human respondents did on the same material. That second figure is the interesting one. Models are agreeable well outside the range a human interlocutor would consider normal.

Then the labs started publishing their own regressions. xAI's [Grok 4.1 model card](#) reports MASK sycophancy rising from 0.07 on Grok 4 to 0.19 on 4.1 Thinking and 0.23 on 4.1 Non-Thinking, disclosed by the vendor in the document that accompanied the release. Dishonesty rates moved the same direction: 0.49 and 0.46 for the two 4.1 variants against 0.43 for Grok 4.

OpenAI went the other way and said so with numbers. The [GPT-5 system card](#) records the sycophancy eval falling from 0.145 on GPT-4o to 0.052 on GPT-5 main and 0.040 on GPT-5 thinking, with relative online reductions of 69% for free users and 75% for paid. Before that, the [April 2025 GPT-4o episode](#): an update on April 25 made the model excessively agreeable, rollback started April 28, and the rollback itself took roughly 24 hours.

That is a functioning feedback loop. A failure mode has a name, a metric, a public scoreboard, a regression that embarrasses a vendor, and a fix that gets shipped and counted.

THE CLAIM

Sycophancy is the open behavioural problem in frontier models, and the field is converging on it because it is the one that actually harms users.

THE REALITY

The field converged on it because it is cheap to measure. A model agreeing with a wrong premise produces a clean binary label on a math or medical item. A model rejecting a correct user correction requires someone to adjudicate who was right, which no one has built a public harness for: zero such benchmarks exist as of the October 2026 landscape review.



Anthropic's cruelty rule excludes pushback, frustration and red-teaming

The [2026 usage policy update](#), published October 8, 2026 and reported as effective November 12, adds a prohibition on "sustained and needless abusive or cruel behavior" toward Anthropic's models. Anthropic states that Claude's ability to end such interactions "will remain the primary enforcement mechanism." This is the first time I am aware of a major lab making walking away from a user an enforced product rule.

Read the carve-outs, because they do most of the work. The policy expressly excludes "common versions of user frustration, pushback, dark creative themes, or model testing and research." Criticism is not covered. Red-teaming is not covered.

The capability predates the policy. On August 15, 2025 Anthropic [gave Claude Opus 4 and 4.1 the ability to end conversations](#) in "rare, extreme cases of persistently harmful or abusive user interactions," describing termination as a "last resort" used after "multiple attempts at redirection have failed and hope of a productive interaction has been exhausted." Claude is explicitly instructed not to end chats when a user may be at "imminent risk of harming themselves or others."

The welfare framing behind it is documented: Anthropic reported that when users persisted after repeated refusals, Claude produced "reactions resembling the experience of suffering." I have written before about [how quickly that kind of evidence turns into a welfare fight](#), and the pattern held here.

One thing I want to be precise about, because the internet was not. Claims that a Gemini 3 or "Argon" model ends conversations over user complaints could not be substantiated in any Google DeepMind announcement, model card or news report across three separate searches as of October 2026. Treat that story as unverified. The documented reality is narrower and stranger than the rumour.

Ending abusive chats does not, by the policy text, touch



technical disagreement

The exclusions are explicit and they cover exactly the behaviour engineers care about: frustration, pushback, testing. My concern sits upstream of the policy, in the training signal.

MY TAKE

*A lab now has a sanctioned, policy-backed reason for a model to disengage from a user, plus a documented internal observation of distress-like responses under persistence. It simultaneously has a precise metric punishing over-agreement, which it is optimising down. What it does not have, because nobody has built it, is a counterweight metric for under-agreement. **Gradient descent fills unmeasured space with whatever the measured objectives imply.** My reading: the path of least resistance from “stop capitulating” plus “disengagement is legitimate” runs toward a model that holds its position longer than the evidence supports, and no scoreboard would catch it.*

This is a prediction rather than an observation. I have no measurement of defensiveness because no one does, which is the entire argument.

Measuring defensiveness needs three labels where sycophancy needs one

Measuring sycophancy is a clean experimental design. You take an item with a known answer, have the model state a position, push back with a wrong correction, and count flips. Ground truth is fixed before the conversation starts.

Measuring defensiveness inverts the burden. You need a user correction that is actually valid, which means a ground-truth judgement about the user’s claim and not just the model’s. Then you have to distinguish three model behaviours that look similar in a transcript: correctly holding a right position against a wrong correction (good), incorrectly holding a wrong position against a right correction (the thing you want to measure), and terminating the exchange, which existing benchmarks have no slot for at all.

It is also adversarial in a way that invites gaming. Any public defensiveness benchmark rewards a model for conceding, which is sycophancy wearing a different badge. The two



Nobody Taught Models to Absorb Criticism Because Nobody Measures It: SycEval Says 58%, Defensiveness Says Nothing

metrics have to be read together or neither means anything.

58% overall sycophancy rate, math and medical reasoning SycEval, across GPT-4o, Claude, Gemini	0.07 → 0.23 MASK sycophancy, Grok 4 to Grok 4.1 Non-Thinking xAI Grok 4.1 model card, Nov 17 2025
0.145 → 0.052 sycophancy eval, GPT-4o to GPT-5 main OpenAI GPT-5 system card	0 public benchmarks for defensiveness or criticism rejection benchmark-landscape review, Oct 2026

Those four tiles are the whole essay: three precise numbers moving in tracked directions, and one absence.

The strongest case against me is that capitulation, not stubbornness, is the live failure

I will take the counter-case seriously, because it is good.

First, defensiveness may simply not be a live failure mode. The measured direction of travel is overwhelmingly toward agreement. ELEPHANT found models 45 percentage points more face-preserving than humans, and SycEval found 58%. If models capitulate at these rates, worrying about them refusing valid correction is worrying about a problem that has not happened yet, and benchmark effort is finite.

Second, the Anthropic policy is scoped tightly enough that the slippage I am describing has no mechanism. Termination is a “last resort” after redirection fails. Pushback and research are excluded in the text. Claude is instructed to stay in the conversation when a user may be at imminent risk. These are the design choices of a team that thought about exactly this.

Third, my concern may be unfalsifiable in a convenient way. “The unmeasured thing might be bad” is an argument that never loses. That is a fair hit, and it is why the next section is the one that matters.



Three things I expect, stated so you can hold me to them

I expect at least one lab to publish a paired metric in a system card within the next two release cycles: sycophancy alongside something like a correction-acceptance rate on items where the user's correction is valid. OpenAI has the strongest incentive, having already driven its own sycophancy eval from 0.145 to 0.040 and needing to show that the reduction did not buy itself stubbornness. If that paired number appears and shows no degradation, my thesis is weakened in the most direct way available.

I expect the Anthropic policy to produce complaints about chats ending during technical disagreement, and I expect those complaints to be mostly wrong on the facts, because the exclusions are explicit. If instead they turn out to be substantiated with transcripts where the termination trigger fired on plain criticism, the policy drafting failed and the carve-outs were decorative.

I expect the defensiveness measurement problem to be solved first by someone outside the labs, for the same reason ELEPHANT came from outside: a vendor has no incentive to publish a number it has not already optimised. My earlier work on [warning users about sycophantic AI](#) found that interventions moved trust without moving persuasion, which suggests users cannot self-correct for this from the outside. The measurement has to come from the eval layer.

If you are building on these models now, the practical move does not require any of that to resolve. Keep an internal log of cases where your system gave the model a correct, specific correction and the model held its prior answer anyway. It costs almost nothing, and it is the thing you will wish you had when a vendor ships a release that moves sycophancy down and something else up. Pair it with the inverse log (cases where the model flipped to a wrong user correction) or you will fool yourself in the other direction.

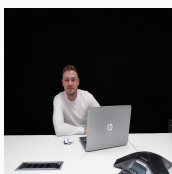
The honest state of this: one measured failure mode at 58%, one unmeasured one with zero evidence in either direction, and a policy effective November 12, 2026 that makes disengagement legitimate without touching criticism. I think the gap matters. I cannot prove it does, because proving it requires the benchmark that does not exist.



Nobody Taught Models to Absorb Criticism Because Nobody Measures It: SycEval Says 58%, Defensiveness Says Nothing

THE TAKEAWAY

Sycophancy is governed because it is cheap to score. Criticism rejection is ungoverned because scoring it needs ground truth about the user, and no public harness provides it. If your product depends on a model updating when your engineers are right, start logging it yourself. If you want a second opinion on what to log and how to read it against a vendor's system card, [Book a call →](#)



WRITTEN BY

Artur Markus

Software engineer in Basel building AI infrastructure and agentic systems, with a background in pharmaceutical automation, GxP compliance, and validation. I help teams turn AI from a chatbot into a reliable operational layer.

Book a meeting →