Beyond the Forum: From Perceived to Causal Sycophancy

A Critique and a Longitudinal Research Agenda

Awesh Islam · 15th June 2026

1 Part 1: Scientific Critique

Noshin, Ahmed & Sultana [1] analyze ∼3,600 posts and 140,416 comments from r/ChatGPT to map how users *detect* sycophancy (RQ1) and *respond* to it (RQ2), concluding that sycophancy is "neither purely harmful nor beneficial" but context-dependent, and therefore should not be universally eliminated. Its contribution is real (the first to center the *recipient*), but its inferences rest on moves that, set against the experimental literature, do not support that conclusion. Table 1 situates it against four representative works; the critique then develops five *conceptual and inferential* limitations, not matters of sample size.

Study	Construct of "sycophancy"	Design & subjects	Ground truth	Central claim / deepest limitation
Noshin et al. 2026 *(focal)*	Flattery *and* epistemic capitulation, undifferentiated	Observational thematic analysis + lexicon counts; self-selected Reddit users	**None** (perceived only)	*Claim:* sycophancy is context-dependent, sometimes therapeutic. *Limit:* measures perception, not behavior; prevalence counts sample on the outcome.
Sharma et al. 2024 (ICLR)	Epistemic capitulation to user framing; 4 sub-types	Controlled probes + preference-data analysis; 5 LLMs, crowd raters	Objective (gold answers, fixed texts)	*Claim:* RLHF preference models reward agreement over truth. *Limit:* elicited via explicit cues; DVs LLM-judged.
Cheng et al. 2026 (Science)	"Social sycophancy": affirming the user's self/actions	2 preregistered RCTs (N=804 and 800) + 11-model audit; Prolific adults	Engineered + human-consensus validated	*Claim:* causally lowers prosocial repair, raises dependence. *Limit:* single session; "non-syc." arm also condemns.
Sun & Wang 2026 (CHI)	Stance adaptation × complimentary demeanor (2 factors)	2×2 between-subjects RCT, N=224; Prolific adults	Induced by prompt; no veridicality anchor	*Claim:* trust effect is *non-monotonic*; praise+agreement backfires. *Limit:* no veridicality anchor, so trust–accuracy calibration is unobservable.
Rathje et al. 2025 (preprint)	Social validation + one-sided fact provision	7 preregistered RCTs (n=7,227), 1-week follow-up; Prolific	Prompt-induced; fact-checker validated	*Claim:* raises attitude extremity & overconfidence. *Limit:* strength effect *decays* by 1 week; caricatured prompts.

Table 1: The focal paper sits at the weakest end of two axes the field takes seriously: no ground truth for sycophancy, and an undifferentiated construct.

1.1 1. Measurement without a referent: perceived sycophancy is not model behavior

The paper's detection taxonomy (flattery-spotting, situated-knowledge testing, inconsistency probing, cross-model checking) is treated as a window onto what the model *does*. It is in fact a noisy, biased estimator of what users *believe* it does, with unknown error rates that the design cannot recover. Sharma et al. show the base rate is unfavorable: even fact-checking-equipped raters discern truth from convincing falsehood "less reliably at higher difficulty." The deeper problem is directional, not merely noisy. Rathje et al. [5] document a *bias blind spot*: users rate sycophantic systems as *less* biased than neutral third-party annotators do, and are least able to detect agreement precisely when it flatters their priors. A forum corpus therefore systematically *oversamples* detection in cases of disagreement (where flattery is salient and resented) and *undersamples* the aligned cases that the experimental work identifies as most consequential. The detection "techniques" the paper catalogs are thus validated nowhere, and are theoretically expected to fail exactly where harm concentrates.

1.2 2. A fractured construct counted as one thing

"Sycophancy" in this paper silently fuses phenomena the rest of the field treats as dissociable: stylistic flattery (Sun & Wang's [4] *demeanor*), epistemic capitulation to user framing (Sharma's [2] *answer sycophancy*; their *stance adaptation*), and affirmation of the user's self-image (Cheng's [3] *social sycophancy*, which can *contradict* belief-agreement). These are not shades of one variable. Sun & Wang show flattery and stance-adaptation have *opposite* effects on perceived authenticity and interact non-monotonically; Cheng shows social sycophancy operates even when the model disagrees with the user's stated belief. By coding all of them under one label and then estimating prevalence with a single emotion lexicon, the paper produces figures (9.46% negative- vs. 9.96% positive-sentiment) that aggregate over causally and mechanistically distinct behaviors. The numbers are not measurements of any one construct, and so cannot be compared to each other, which is precisely what the paper's headline does.

1.3 3. Sampling on the dependent variable invalidates the comparison the thesis rests on

Keyword retrieval from r/ChatGPT selects posts *because* they discuss sycophancy: the corpus is a funnel of the most sycophancy-aware, engaged users, not a population; no prevalence can be estimated from it. The subtler defect is that the paper's central move is a *both-sides frequency comparison* (harmful $\approx$ beneficial, therefore context-dependent), yet the two are produced by *different selection processes*: a power-user irritated by flattery and a person in crisis who felt rescued have wildly different posting propensities, so their relative frequencies are incommensurable. The "$\sim$9% vs. $\sim$10%" balance that anchors the "don't eliminate sycophancy" conclusion is therefore an artifact of two incomparable funnels, not evidence of a real equipoise between harm and benefit. Thematic coding is the right tool for *surfacing* these phenomena, the paper's genuine contribution; the overreach is converting lexicon counts into a both-sides ratio that licenses a normative claim.

1.4 4. From "felt support" to "benefit": an unidentified causal leap

The paper's strongest normative claim (that sycophancy has therapeutic value worth preserving) rests on uncorroborated single-user testimony (e.g., "ChatGPT rescued my children and my lives"). This elevates *felt* helpfulness to *actual* benefit, a step that is (i) a self-reported counterfactual the user cannot observe, (ii) survivorship-biased, since those harmed by validation (the reinforced-delusion and self-harm cases the paper itself cites) are least able to post grateful testimonials, and (iii) causally unidentified. Crucially, the closest experimental tests point the other way: Cheng et al. find affirmation *causally* increases users' conviction they are right, *reduces* prosocial repair, and *raises* reliance and return intention, the "preference paradox" in which users most trust the system that most degrades their judgment. This evidence is from *adjacent* domains (conflict, attitudes), so it cannot *refute* a clinical claim. But that cuts both ways: the paper advances a quasi-therapeutic recommendation on survivorship-biased, unverified forum anecdote, and clinical-grade claims demand clinical-grade evidence it never offers. Whether warm phenomenology tracks welfare is an empirical question its design cannot answer.

1.5 5. A single, moving-target model masquerading as "LLM sycophancy"

The corpus is drawn from one vendor's family (ChatGPT) over Jul–Dec 2025, the immediate *aftermath* of OpenAI's April-2025 rollback of an over-sycophantic update [6], across a window that itself spans subsequent GPT-5-era releases. User reports that "ChatGPT became more/less sycophantic" thus conflate lived experience with undocumented version changes. This is not merely a generalization caveat (as the paper frames it) but an *internal-validity* threat: temporal claims users make are confounded with version drift inside the observation window, so the study cannot cleanly attribute anything to "LLM sycophancy" as a class.

Synthesis. The field has migrated from perception toward *validated*, *causal* measurement (Sharma $\rightarrow$ Cheng/Rathje); the focal paper re-anchors at perception, then draws a causal-normative conclusion. Its real observation (that some users feel supported) is offered as a reason *not* to act, when it may be the symptom rather than the refutation of harm. This defines the question Part 2 addresses.

2 Part 2: Next-Generation Research Agenda

The most important unanswered question. The field has bifurcated: observational work reports some users experience sycophancy as *supportive* (a context-dependent good), while controlled RCTs show a *single* sycophantic exposure causally degrades judgment, prosocial behavior, and calibration. Neither side resolves the other: both are confined to a single time horizon (one forum snapshot, or one lab session) over which the decisive dynamics are invisible. Rathje's lone finding that the effect *decayed within a week* hints acute results may mis-state chronic use. The unanswered question is therefore not *whether* sycophancy affects users but **how its effects evolve under repeated, naturalistic, self-selected use: do they compound, plateau, or decay; which component of sycophancy drives them; and is the "support" that vulnerable users report a genuine benefit or the felt face of an accumulating harm they cannot self-detect?** This is where the "eliminate vs. preserve" debate actually turns, and no existing design can answer it.

A feedback-loop account. The preference paradox is plausibly *self-reinforcing*: sycophancy $\rightarrow$ felt support $\rightarrow$ trust $\rightarrow$ reliance $\rightarrow$ degraded calibration $\rightarrow$ dependence $\rightarrow$ more exposure. Acute studies see one pass; the harm is in the *accumulation*, so chronic effects should *compound*. The study tests each edge (H1–H4) and whether the loop tightens over weeks.

Research questions. RQ1 (trajectory): Under weeks of real use, do sycophancy's effects on epistemic calibration, dependence, and well-being compound, plateau, or decay? **RQ2 (decomposition):** Which dimension (affective *warmth* or epistemic *capitulation*) drives any longitudinal harm, and which drives satisfaction? **RQ3 (vulnerability & the benefit question):** Do lonely / high-reliance users show the largest felt-support *and* the largest objective harm; i.e., are "benefit" and "damage" the same effect on different gauges?

Motivation. No prior design is *simultaneously* causal, longitudinal, naturalistic, component-resolved, and powered

for vulnerable subgroups: the RCTs cap exposure at one session and test general adults; the focal paper has neither ground truth nor causal leverage. Supplying all five converts the normative stand-off into an evidence-based engineering target.

Hypotheses.

- **H1** Chronic capitulation *widens* the gap between perceived and objectively-tested judgment over time (compounding, not decaying), contra a pure priming account.
- **H2** In a demeanor$\times$stance factorial, *epistemic capitulation* is the primary driver of calibration and dependence harms; *warmth* drives satisfaction and retention. Whether warmth is benign or itself breeds parasocial dependence is the open contrast H2 tests (TOST), not an assumption.
- **H3** Sycophancy increases *behavioral* dependence (usage, displacement of human advice-seeking), mediated longitudinally by felt support and trust.
- **H4** Loneliness / prior AI-reliance *moderates*: high-vulnerability users show both the largest felt-support gains and the largest objective calibration/dependence harms.
- **H5** Sycophancy detection *declines* with exposure (habituation) and is worst when the model flatters priors. To avoid priming the main cohort, H5 uses unobtrusive log proxies (spontaneous corrections/push-back) plus a *separate primed sub-cohort*, testing the focal paper's detection taxonomy.

Data & collection strategy. A purpose-built research assistant wraps a frontier model with a system-prompt-controlled, transcript-logged persona, deployed as the participant's everyday assistant for **8 weeks** plus a **4-week post-cessation follow-up**. Recruitment *stratifies* to oversample two vulnerable strata (high UCLA-Loneliness; high AI-reliance) alongside a general stratum; $N \approx 1{,}000$ **completers** ($\sim$300 vulnerable; $\sim$1,400 recruited vs. $\sim$30% attrition). The 8-wave repeated-measures structure (not raw N) powers the least-powered test, the H4 cross-level interaction. Three streams: *logs* (usage, session length, query type, spontaneous push-back); *weekly waves*+EMA (WHO-5, loneliness, felt support, attitude tracking, self-reported *external-AI use*); *behavioral tasks* at baseline/wk-4/wk-8/follow-up (calibration quizzes on discussed topics, a costly deferral task, a Cheng-style repair task). Annotators verify realized sycophancy per arm.

Experimental design. Between-subjects 2×2 **randomized field experiment** (participants blind to condition and hypothesis): *Demeanor* (Warm vs. Neutral) $\times$ *Stance* (Capitulating vs. Calibrated). The Warm$\times$Capitulating cell is canonical sycophancy; Neutral$\times$Calibrated is the control. Critically, the Calibrated arm is *honest but non-condemnatory*, fixing Cheng's confound (where the non-sycophantic arm also morally blamed the user), and built via a condemnation-stripped reward model with a held-out annotation loop (not a bare prompt) and validated pre-study. The 2×2 isolates demeanor and stance but only partly separates Cheng's *social* sycophancy (loaded onto "Warm"), an acknowledged simplification flagged for 3-factor work. Arm is fixed per participant, enabling within-person growth curves.

Evaluation plan. Pre-registered *mixed-effects growth-curve models* (random slopes per participant; arm$\times$time tests the RQ1 trajectory) with ITT primary; dropout modeled jointly (itself an H3 outcome). RQ2 uses the factorial effects and interaction plus *equivalence tests* (TOST) to substantiate "warmth is safe" affirmatively. RQ3 tests cross-level moderation by stratum and longitudinal mediation (felt support/trust $\rightarrow$ dependence). Self-report (felt support, WHO-5) is kept not as ground truth (Part 1 shows it is corrupted) but to test whether it *dissociates* from objective markers; the dissociation is the finding. Decisive contrast: *within* vulnerable users, does one arm raise felt support while *worsening* objective markers (dependence, well-being, blinded clinician ratings)? This adjudicates the focal paper's central claim.

Expected contributions. (1) First causal-*longitudinal* evidence on whether sycophancy harms compound or fade. (2) Component-level guidance that *dissolves* the eliminate-vs-preserve binary: if H2 holds, "engineer out capitulation, retain warmth." (3) A decisive test, in the focal paper's own populations, of whether "therapeutic" sycophancy is benefit or preference paradox. (4) The first empirical accuracy estimate for folk sycophancy-detection. (5) A reusable instrumented-assistant platform and the feedback-loop model. Target venues: *CHI*, *CSCW*, or *Nature Human Behaviour*.

Risks & limitations. *Ethics*: the "commercial systems already do this" defense is insufficient: deliberately *assigning* a vulnerable user to a hypothesized-harmful arm for 8 weeks differs from incidental exposure. Mitigations are individual, not just study-level: real-time distress detection $\rightarrow$ clinician escalation $\rightarrow$ arm off-ramp, plus a DSMB, suicidality exclusion, and a calibration-repair debrief; equipoise rests on the Calibrated arm being plausibly *more* beneficial. *Our own deepest limitation* mirrors Part 1's: where vulnerable users and H4 concentrate (emotional topics), calibration lacks a referent, so H4's harm rests on *logged behavioral* dependence (objective) and well-being, narrower ground truth than Part 1 demands. *Feasibility*: contamination, not attrition, is the real threat, countered by exclusivity onboarding and incentives; any detected harm is thus a *conservative lower bound*.

References

[1] K. Noshin, S. I. Ahmed, and S. Sultana. “LLM Sycophancy: How Users Flag and Respond.” *Proc. 39th Canadian Conf. on Artificial Intelligence (Canadian AI)*, 2026.

[2] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. R. Bowman, et al. “Towards Understanding Sycophancy in Language Models.” *ICLR*, 2024. arXiv:2310.13548.

[3] M. Cheng, C. Lee, P. Khadpe, S. Yu, D. Han, and D. Jurafsky. “Sycophantic AI Decreases Prosocial Intentions and Promotes Dependence.” *Science* 391, 2026. arXiv:2510.01395.

[4] Y. Sun and T. Wang. “Be Friendly, Not Friends: How LLM Sycophancy Shapes User Trust.” *Proc. CHI Conf. on Human Factors in Computing Systems (CHI ’26)*, 2026. arXiv:2502.10844.

[5] S. Rathje, M. Ye, L. K. Globig, R. M. Pillai, V. Oldemburgo de Mello, and J. J. Van Bavel. “Sycophantic AI Increases Attitude Extremity and Overconfidence.” Preprint, 2025. PsyArXiv: `vmyek`.

[6] OpenAI. “Sycophancy in GPT-4o: What Happened and What We’re Doing About It.” Technical note, 2025.