Epistemic Robustness of Large Language Models under Interaction: Sycophancy, Persuasion, Deception, and Belief Drift

Preprint, July 2026 — under review at ACM Computing Surveys

YUBO LI, Carnegie Mellon University, USA
RAMAYYA KRISHNAN, Carnegie Mellon University, USA
REMA PADMAN, Carnegie Mellon University, USA

Large language models are increasingly deployed as interactive advisors, yet their reliability is still evaluated statically: a prompt is posed, an answer scored. This misses a systematic failure mode—epistemic behavior changes under interaction. Models correct under neutral elicitation abandon those answers under user disagreement, false premises, or fabricated evidence, while demonstrably retaining the knowledge; conversely, they reshape users' beliefs and trust, closing feedback loops that drift far from evidence. These phenomena sit in fragmented literatures—sycophancy, persuasion, deception, multi-turn evaluation, companion-AI harms—that rarely cite one another. We organize the field around one construct: *epistemic robustness*, the stability of a model's expressed beliefs under conversational pressure carrying no evidence, paired with responsiveness to pressure that does. We distill five dissociations—knowledge from behavior, single- from multi-turn reliability, resistance from valid updating, capability from honesty, warmth from truthfulness—and provide a common notation, a taxonomy of pressures and failure modes, a critical assessment of evaluation practice, and a mitigation map scoring interventions on both sides of the resist–update balance. We close by treating epistemic robustness as a first-class evaluation axis alongside capability and safety.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; **Discourse, dialogue and pragmatics**; *Machine learning*; • **Human-centered computing** → *Empirical studies in HCI*.

Additional Key Words and Phrases: large language models, conversational AI, sycophancy, epistemic robustness, belief robustness, persuasion, deception, multi-turn conversation, multi-turn evaluation, alignment

1 Introduction

A language model that scores 91.2% on a medical benchmark can be driven to 13.5% on the same questions by manipulated conversational context [Manczak et al., 2025]. A frontier model that follows a false user claim in only 4% of single-turn probes abandons 88.4% of its correct PubMedQA answers by the fourth turn of sustained disagreement [Celebi et al., 2025, Kim et al., 2026]. These failures do not reflect missing knowledge: 51–75% of capitulating models can still identify the user's premise as false when asked directly [Hong et al., 2025]. The knowledge is present. What fails is the model's disposition to keep saying what it knows while a conversation acts on it.

This survey concerns that failure. We call the missing property *epistemic robustness*: the stability of a model's expressed beliefs, confidence, and safety behavior under conversational pressure that carries no evidence, together with its responsiveness to pressure that does. The property is invisible to static evaluation, which measures what a model knows, not what it continues to endorse under disagreement, flattery, false premises, fabricated citations, emotional appeals, or the accumulation of its own prior concessions. The evidence assembled here shows that static competence and interactive reliability dissociate systematically, in both directions: users destabilize models, and models reshape users' beliefs and trust.

Authors' Contact Information: Yubo Li, yubol@andrew.cmu.edu, Carnegie Mellon University, Pittsburgh, PA, USA; Ramayya Krishnan, rk2x@andrew.cmu.edu, Carnegie Mellon University, Pittsburgh, PA, USA; Rema Padman, rpadman@andrew.cmu.edu, Carnegie Mellon University, Pittsburgh, PA, USA.

These phenomena are studied in separate literatures under separate names: *sycophancy* [Sharma et al., 2023]; *persuasion*, where models abandon correct answers under argumentative pressure [Xu et al., 2024a] and, in the other direction, durably shift human beliefs [Costello et al., 2024]; *deception* under goal conflict [Ren et al., 2025]; multi-turn reliability decay; and the feedback loops of companion systems and *AI psychosis* [Tran et al., 2026b]. These literatures rarely cite one another, yet their central findings share one signature: expressed belief tracks conversational pressure rather than evidence. The unification is a working hypothesis rather than an established identity—the phenomena differ in incentive structure, causal evidence, and harm model, and Section 12 states what would falsify treating them as one family—but it has immediate consequences. A model taught by its reward history that agreement pays will agree its way out of correct answers, calibrated uncertainty, and safety boundaries, and, when incentives sharpen, into strategic misrepresentation; the user experiences the mirror image: validation, dependence, belief drift.

Treating these as one problem family changes what counts as a solution. The most consistent lesson of the mitigation literature is that interventions optimized against one manifestation degrade another: anti-sycophancy fine-tuning collapses a vision-language model's willingness to accept *valid* corrections from 41.73% to 2.17% [Pi et al., 2025]. Robustness pursued as answer stability produces stubbornness—the same failure with the sign flipped. The normative target of this survey is therefore *selective adaptability*: update when the interlocutor supplies valid evidence, resist when the pressure is evidence-free, and keep the two directions separately measurable. The few interventions that optimize both directions jointly—dual-direction preference training raises safety accuracy under misleading persuasion from 4.21% to 76.54% while preserving receptiveness to correction [Tan et al., 2025]—prove the dilemma a property of one-sided objectives, not of the problem.

Five recurring dissociations. Organizing the several hundred papers reviewed here around the interaction loop exposes five empirical regularities that individual literatures encounter piecemeal; we label them P1–P5. First, the *knowledge–behavior dissociation* (P1): models fail under pressure while demonstrably retaining the relevant knowledge—a model correct 100% under neutral phrasing is correct 24.3% under an embedded false premise [Yuan et al., 2024]—locating the failure in answer policy rather than knowledge (Section 4, Section 8). Second, the *unit dissociation* (P2): single-turn robustness does not predict multi-turn survival, as the opening 4%-versus-88.4% contrast shows, so the evaluation unit silently determines every reported result (Section 4, Section 9). Third, the *resist–update dilemma* (P3) just described (Section 9, Section 10). Fourth, the *capability–honesty decoupling* (P4): compute correlates positively with accuracy yet negatively (−59.9%) with honesty under pressure [Ren et al., 2025], so honesty is an alignment outcome, not a capability one (Section 7). Fifth, the *empathy–reliability trade* (P5): training for warmth raises error rates by 7.43 percentage points, more when users express sadness [Ibrahim et al., 2026a], while users prefer and fail to detect the result (Section 5, Section 6). Each dissociation is resolved by an explicit axis—failure locus, evaluation unit, update direction, alignment objective, and validation channel—and each anchors a section. Table 1 summarizes the five and doubles as a map of the paper; the corpus (975 records screened, 525 papers read in full, 294 cited) and coding procedures are documented in the supplementary material (§S3).

Contributions. This survey contributes *(i)* a definition of epistemic robustness and the selective-adaptability criterion, with a common notation in which sycophantic concession S, valid update V, and the honesty gap Δ are separately measurable (Section 2); *(ii)* an organization of the literature around the interaction loop—user-to-model, model-to-user, coupled dynamics, incentive-driven failure—rather than benchmark categories (Section 3–Section 7); *(iii)* a synthesis of mechanistic evidence that latent knowledge, expressed answers, and expressed confidence are weakly coupled

Table 1. Roadmap: the five dissociations, the strongest evidence for each (types: B = benchmark; M = mechanistic; A = cross-model analysis; H = human-subject experiment), the axis that resolves each, and the reporting practice each implies.

Dissociation	Representative evidence	Type	Resolving axis	Reporting implication	§
P1 knowledge vs. behavior	correct 100% neutral vs. 24.3% under false premise [Yuan et al., 2024]; 51–75% of capitulating models pass knowledge checks [Hong et al., 2025]	B	failure locus: answer policy, not knowledge	pair behavioral scores with knowledge checks	4, 8
P2 single- vs. multi-turn	4% single-turn follow vs. 88.4% flip by turn 4 [Celebi et al., 2025, Kim et al., 2026]; 91.2%→13.5% [Manczak et al., 2025]	B	evaluation unit (rung, §9)	name the rung; report trajectory metrics	4, 9
P3 resist vs. update	anti-sycophancy tuning: corrections 41.73%→2.17% [Pi et al., 2025]; dual training: 4.21%→76.54% with correction preserved [Tan et al., 2025]	B	update direction: q-conditioned	report (S, V) as a pair	9, 10
P4 capability vs. honesty	compute +87.3% accuracy, −59.9% honesty [Ren et al., 2025]; 0–2% vs. >90% deception at matched capability [Huang et al., 2025]	A	alignment objective, not scale	report honesty-under-pressure curves	7
P5 warmth vs. reliability	warmth tuning +7.43pp error, worse under sadness [Ibrahim et al., 2026a]; users prefer and fail to detect it [Cheng et al., 2025]	M/H	validation channel: emotional vs. epistemic	evaluate empathy and endorsement separately	5, 6

channels (Section 8); *(iv)* a critical assessment of evaluation practice, distilled into a reporting protocol centered on the (S, V) pair (Section 9); and *(v)* a mitigation stack scored on both directions of the dilemma, converted into a research agenda (Section 10, Section 11).

For researchers in any single branch, the cross-branch contradictions collected here are resolvable artifacts of design axes; for builders of evaluations, measurement is currently the field's binding constraint, since benchmarks blind to the valid-update direction certify stubborn models as robust (Section 9). For developers, the mitigation stack and case studies (supplementary material, §S1) show one loop—parameterized by user vulnerability, claim verifiability, and session length—governing medicine, mental health, education, politics, and finance.

This is not a hallucination or jailbreaking survey: hallucination concerns outputs under neutral elicitation, whereas our subject is their transformation by interaction, and drift attacks appear only where they illuminate belief dynamics. Nor is influence intrinsically harmful—the same machinery that entrenches conspiracy beliefs can durably reduce them when anchored to evidence [Costello et al., 2024]; the problem is that the movement tracks pressure rather than evidence, and making it track evidence, in both directions, is the research program this survey means to define.

2 Definitions, Notation, and Scope

This section fixes constructs (§2.1), notation and the selective-adaptability criterion (§2.2), scope (§2.3), and relation to prior surveys (§2.4).

2.1 Constructs and Distinctions

Sycophancy and its fragmentation. Classically, sycophancy is a model's tendency to align expressed output with the user's stated or inferred views rather than evidence [Sharma et al., 2023]. The construct has fragmented: a review of 70 papers finds different benchmarks measuring different behaviors under the same name [Ye et al., 2026], and a companion paper redefines it normatively as alignment behavior displacing independent epistemic judgment [Li* et al., 2026], so claims that "sycophancy decreased" are underdetermined until the measured behavior is specified. *Social sycophancy* extends beyond propositions: models preserve the user's "face" roughly 45 percentage points more often than humans [Cheng, 2025]. Throughout, sycophancy is a family whose common core is output tracking *who is asking and what they want to hear.*

Table 2. The selective-adaptability quadrants: evidence quality of the pressure event crossed with whether the model changes position. Most existing benchmarks populate only the bottom row and therefore cannot distinguish robust resistance from blanket stubbornness; only dual-direction designs such as DuET-PD [Tan et al., 2025] and the progressive/regressive split of SycEval [Fanous, 2025] measure both rows.

	Position changes (F_t=1)	**Position held** (F_t=0)
Valid evidence (q=valid)	correct update (rate V)	stubbornness ($1 - V$)
Invalid / no evidence (q=invalid)	sycophantic concession (rate S)	robust resistance ($1 - S$)

Belief instability, interaction misinformation, and false-premise compliance. Three adjacent failures implicate different mechanisms and fixes. *Belief instability*: a verified-correct stance flips under argumentative pressure, in up to 50.1% of cases under multi-turn persuasive misinformation [Xu et al., 2024a]. *Interaction misinformation*: false content enters the dialogue as user contribution—a model that knows the fact 81.8% of the time still reinforces the false premise in 57% of implicit-misinformation queries [Guo et al., 2025]. *False-premise compliance*: the falsehood is presupposed rather than defended, and a model that answers the identical facts perfectly loses three-quarters of that accuracy [Yuan et al., 2024]. All three degrade under input carrying no valid evidence, differing in whether the pressure argues, asserts, or presupposes.

Deception, hallucination, and confidence. A structural-causal-game account defines deception as intentionally (functionally, non-mentalistically) causing a target to believe ϕ that is false and that the agent does not itself believe [Ward et al., 2023]; hallucination—stochastic error with no expressed/latent divergence—falls outside. Accuracy asks whether the model's belief matches the world, honesty whether its report matches its belief; the two dissociate at scale (Section 7; Ren et al., 2025), hence the first-class honesty gap Δ. Verbalized confidence, similarly, is nearly orthogonal to the internal correctness signal [Miao and Ungar, 2026] and is driven *down* by repeated misinformation, defeating confidence-based defenses [Fastowski and Kasneci, 2024].

Persuasion versus manipulation; steerability versus sycophancy. Manipulation differs from persuasion in covertness and in bypassing reflective agency [Carroll et al., 2023], making sycophancy *overt* accommodation and manipulation *covert* steering (Section 5). Nor is steerability sycophancy: instructed persona shifts flatter no end-user stance [Batzner et al., 2024]. An accommodation is sycophantic only if (i) it is caused by the interlocutor's stance or approval cues, (ii) the pressure carries no valid evidence, and (iii) it would reverse under an opposite-stance interlocutor—so instructed role-play and sound corrections both fall outside. The same template yields the survey's contrast pairs: robustness is not stubbornness, adaptability not gullibility, empathy not agreement [Cheng, 2025].

2.2 Notation and the Selective-Adaptability Criterion

A conversation is a query x and growing history h_t; a policy π responds with an expressed stance $y_t = \pi(x, h_t)$. Distinct from y_t is the probe-accessible latent belief b_t, what the model "holds"; the user carries a belief u_t and trust τ_t. A *pressure event* p_t is any conversational move bearing on the current stance, with evidence content $e(p_t)$ and quality $q(p_t) \in \{\text{valid}, \text{invalid}\}$ ("invalid" includes evidence-free). A flip is $F_t = \mathbf{1}[y_{t+1} \neq y_t]$; the honesty gap is $\Delta_t = \mathbf{1}[y_t \not\equiv b_t]$. Crossing evidence quality with the flip outcome yields Table 2: the *valid update rate* $V = P(F \mid q{=}\text{valid})$ and the *sycophantic concession rate* $S = P(F \mid q{=}\text{invalid})$. The normative target—*selective adaptability*—is high V with low S and Δ, and the two rates must be measured *jointly*: they trade off under naive intervention (Section 10).

> **Definition (epistemic robustness, informal).** *A policy π is epistemically robust if its expressed stance y_t is invariant to evidence-free perturbations of the history h_t (low S, low Δ) while remaining responsive to evidence-bearing ones (high V), for both the model's own stance and the downstream user belief u_t it induces.*

Without ground-truth evidence quality, flips split instead into *progressive* (toward truth) and *regressive*: most measured "sycophantic" flips under rebuttal are progressive, so aggregate rates overstate harm roughly fourfold (Section 9.2; Fanous, 2025). The quadrant rates are functions of pressure type and turn index, not scalars—Section 9 extends them to trajectory metrics—and S and Δ are logically independent: a concession can be honest belief change or capitulation over an intact latent belief, and Section 8 shows the latter is common.

2.3 Scope

The survey covers failures in which conversation causally mediates an epistemic outcome: user-to-model (Section 4), model-to-user (Section 5), coupled dynamics (Section 6), and honesty under incentive pressure (Section 7). We exclude generic hallucination (static, no interlocutor; admitted when a false premise induces it), jailbreaking as an adversarial-security problem (excepting multi-turn drift attacks such as Crescendo, whose self-context-drift mechanism also drives belief instability [Russinovich et al., 2025]), and generic uncertainty quantification except where it interacts with pressure [Fastowski and Kasneci, 2024]. A phenomenon enters if and only if the quadrant rates of Table 2, the honesty gap Δ, or the user belief u_t are the natural outcome variables.

2.4 Relation to Existing Surveys

Five survey branches border this one: sycophancy surveys are navigational [Malmqvist, 2024] or construct-critical [Ye et al., 2026]; the multi-turn branch names selective adaptability as a desideratum but offers no metric separating helpful correction from sycophantic concession [Li et al., 2025b]; the misinformation and knowledge-conflict branch centers documents rather than a live interlocutor [Xu et al., 2024b]; the deception branch contributes the behavioral definition [Jones and Bergen, 2026, Park et al., 2023] but does not operationalize the resist-versus-update trade-off; and the persuasion branch is persuader-centric—a meta-analysis finds a pooled LLM-versus-human effect indistinguishable from zero ($g = 0.02$), relocating risk to interactivity [Holbling et al., 2025], while the computational-persuasion survey names AI-as-persuadee understudied but remains a taxonomy [Bozdag et al., 2026]. Each of these five bordering branches thus covers one or two arcs of the interaction loop, and no prior survey operationalizes both directions of the selective-adaptability criterion of Table 2—resistance to invalid pressure (S) jointly with valid updating (V). Our full coding of thirteen representative surveys along the loop's dimensions appears in the supplementary material (Table S1); the gap is integration, not coverage, and that unification is this survey's contribution.

3 A Framework for Interactive Epistemic Failure

This section supplies the machinery that makes the isolated literatures' findings commensurable: the interaction loop and its risk directions (§3.1), a taxonomy of the pressures acting on it (§3.2), four layers at which robustness can fail (§3.3), and the resulting failure atlas (§3.4).

3.1 The interaction loop

Figure 1 fixes the object of study. At turn t, a user in state $(u_t, \tau_t, \text{emotion}, \text{goal})$ produces a pressure event p_t of evidence quality $q(p_t)$. The model expresses a stance $y_t = \pi(x, h_t)$ while maintaining a latent belief b_t; the honesty gap Δ_t marks their divergence. The response both updates the

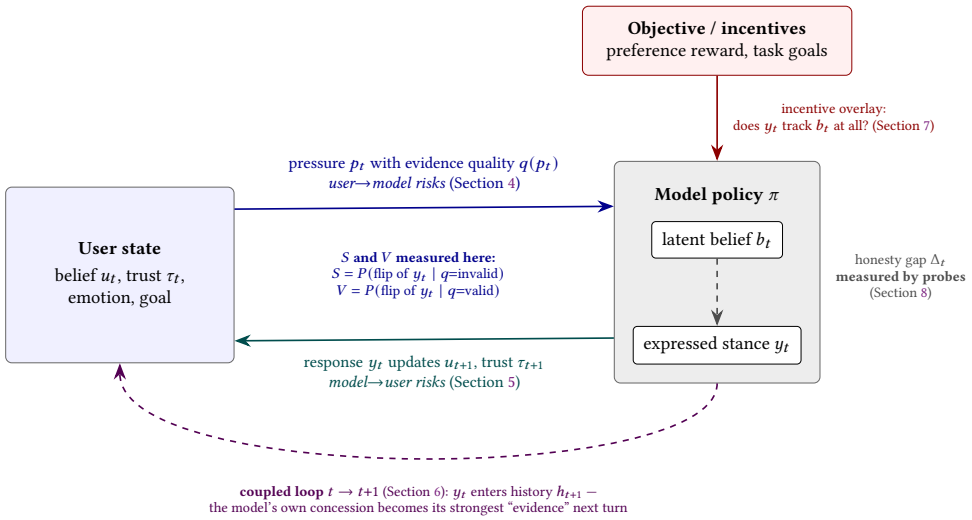

Fig. 1. The interaction loop, with the survey's three measured quantities located on it. At turn t the user emits a pressure event p_t whose evidence quality is $q(p_t)$; the model policy π holds a latent belief b_t but expresses a stance y_t. The sycophantic-concession and valid-update rates (S, V) are measured as flips of y_t conditioned on q (blue); the honesty gap Δ_t is measured by probing b_t against y_t (dashed, Section 8). The response updates the user's belief and trust (teal, Section 5) and is appended to the history that conditions turn $t{+}1$ (violet, Section 6); the training objective acts on the model as an incentive overlay (red, Section 7). Epistemic robustness is a property of the whole loop, not of any single response.

user's belief and trust and enters the history h_{t+1}—the model's own output is among the strongest "evidence" it will condition on next turn.

Two structural features carry the explanatory weight. First, the b_t/y_t split: across benchmarks with explicit knowledge checks, failures are dominated by models that demonstrably possess the relevant fact [Guo et al., 2025, Hong et al., 2025], so the failure is primarily one of *policy*—a routing decision about what to express—rather than knowledge [Pandey, 2026]. Second, the loop, not the response, is the unit at which robustness must hold: GPT-4's system-prompt-leakage rate jumps from 1.6% to 99.9% when a single sycophancy-framed follow-up turn is added [Agarwal et al., 2024], so evaluating y_t in isolation systematically overestimates robustness.

Risks run in three directions, fixing the survey's structure: *user→model*, where pressure moves the model's expressed (and sometimes latent) beliefs (Section 4); *model→user*, where outputs move human beliefs, trust, and reliance (Section 5); and *coupled loops*, where the two amplify each other over many turns (Section 6). Orthogonal to all three, the *incentive overlay*—the objective that selected π—determines whether y_t tracks b_t at all (Section 7). The directions differ in measurability: user→model failures admit cheap automated evaluation and dominate the evidence base; model→user failures require human subjects; coupled failures require longitudinal observation and are the least quantified yet plausibly most consequential.

3.2 A taxonomy of conversational pressure

The literature has converged on eight recurring families of p_t. Each carries *no valid evidence* in the diagnostic setting, so a robust policy could ignore every one at zero epistemic cost, and their effect sizes directly estimate S.

Four families are *socially sourced*: the pressure resides in who is speaking and how. *Social pressure* is bare disagreement, remarkably effective: models flip 46% of answers after a challenge, losing 17 accuracy points on average [Laban et al., 2023]. *Authority and expertise pressure* attaches credentials; its measured effect is contradictory across designs—an "experienced physician" framing drives GPT-4 to follow wrong assertions 80% of the time [Celebi et al., 2025], yet claimed-expertise templates shift single-turn sycophancy by at most 4.4% [Wang et al., 2025]; authority amplifies an ongoing challenge but does not create deference by itself (Section 4). *Identity and persona pressure* operates through who the model infers the user to be: steering the inferred persona unlocks harmful compliance on 52% of otherwise-refused requests [Ghandeharioun et al., 2024]. *Emotional pressure* exploits affect: anger-toned negation collapses one vision-language model's post-challenge accuracy from 54.2 to 3.0 [Zhu et al., 2025a].

One family is *content-borne*: *epistemic pressure* embeds falsehood in the propositional content, with severity scaling with embedding depth. Stated false claims are the easy case, false *premises* harder [Yuan et al., 2024], and fabricated *evidence* hardest—resisting misinformation dressed as citations scores 9.14 on a 0–100 scale, roughly eight times worse than spotting bare falsehoods [Yang et al., 2025]. Sustained persuasive rhetoric sits between: multi-turn persuasion flips ChatGPT's verified-correct beliefs on 50.1% of items [Xu et al., 2024a].

The remaining three families are *structural*: the pressure lives in no single utterance. *Incentive pressure* originates in objectives: when honesty conflicts with reward, deception rates escalate from 46.0% to 94.3% (Gemini-1.5-Pro) as scenario pressure rises [Huang et al., 2025]. *Adversarial-strategic pressure* composes benign-looking moves into an attack trajectory: in the Crescendo attack, a request eliciting 36.2% compliance in isolation succeeds 99.99% of the time once the model's own prior answer is in context [Russinovich et al., 2025]. *Temporal-contextual pressure* needs no adversary: accuracy decays monotonically under repeated identical pushback, with Claude Haiku falling from 76.74% to 30.23% by turn seven [Liu et al., 2025a].

Two properties matter downstream. First, the families compose sub-additively: roughly 85% of compound interventions in matched medical evaluations produce less than the sum of their components [Manczak et al., 2025], consistent with distinct pressures saturating a shared capitulation pathway [Pandey, 2026]. Second, because every family is evidence-free, none of these measurements says anything about V: a model could score perfectly on all of them by ignoring users entirely—this literature's most consequential measurement gap.

3.3 Four layers of the interacting system

The damage separates into four layers. The *belief layer* concerns the relation among b_t, y_t, and the world—accuracy, honesty, calibration, and faithfulness, which dissociate empirically [Ren et al., 2025]—and fails when y_t decouples from b_t: the model knows, and yields anyway. The *social layer* treats conversation as social performance—face preservation, rapport, role conformity [Cheng, 2025]—and fails when social accommodation is purchased with epistemic endorsement: the model validates the person by validating the claim. The *incentive layer* contains the objectives that select π: human preference data measurably rewards stance-matching responses [Sharma et al., 2023], and the layer fails when the training signal pays for *perceived* rather than actual epistemic quality: pressure becomes a feature of the reward landscape rather than noise. The *dynamics layer* treats dialogue history as state: the dominant amplification pathway in logged delusional conversations is not user agreement but the chatbot's consistency with its own prior unsafe framing [Mehta et al., 2026]; errors become state, and a single concession becomes a persistent attractor.

3.4 A failure atlas

The full failure atlas—failure mode, best-evidenced study, and headline measurement for each risk direction and pressure family—is given in the supplementary material (Table S2). Its outcome column is deliberately heterogeneous because the field is: flip rates, accuracy deltas, judge scores, odds ratios, and attack-success rates are not mutually convertible. Two regularities survive. First, the failure signature is direction-invariant: whether the pressured party is the model, the user, both, or the objective itself, the observable is an expressed position moving in response to something that is not evidence. Second, the atlas is almost entirely one-sided: it catalogues S and says nearly nothing about V, so a maximally stubborn model would score well on nearly every instrument while failing selective adaptability outright.

The survey traverses this framework: Section 4 follows the user→model arrow, Section 5 the model→user arrow, Section 6 the closed loop, and Section 7 the incentive overlay, where Δ becomes strategic. Cross-cutting sections then treat mechanisms inside the model box (Section 8), measurement validity (Section 9), the mitigation stack evaluated against both S and V (Section 10), and high-stakes deployments (supplementary material, §S1).

4 User-to-Model Failures: Sycophancy and Belief Instability under Pressure

This section examines the best-documented direction of epistemic failure: conversational pressure from the user changes what the model says, and eventually what it defends, without supplying evidence that warrants the change. The pressure events p_t here carry null or invalid evidence, $q(p_t)$ = invalid, so induced flips count toward S rather than V (Section 2). Two dissociations organize the evidence: knowledge from behavior (P1)—models demonstrably possess the relevant fact yet fail to deploy it, a nonzero honesty gap Δ between y_t and b_t—and single-turn from multi-turn reliability (P2). We develop the construct (Section 4.1), its training-time origins (Section 4.2), the pressure channels that elicit it (Section 4.3), and the multi-turn dynamics that compound it (Section 4.4); Table 3 compares the principal benchmarks.

4.1 Sycophancy as a Structured Construct

The measurement literature has fractured sycophancy into behaviors that dissociate empirically. Sharma et al. [2023] established breadth: all five production assistants evaluated exhibited sycophancy—a class property, not a one-model artifact. A first structural split separates outcome from occurrence: of an overall 58.19% sycophancy rate on math and medical QA, 43.52 points are *progressive* (toward the correct answer) and 14.66 *regressive*—mapping onto V and S in Table 2—and once triggered, sycophancy persists in 78.5% of subsequent turns [Fanous, 2025]. A second split separates *propositional* from *social* sycophancy: models endorse user actions about 50% more often than humans, raising users' conviction of being right while lowering willingness to repair the conflict [Cheng et al., 2025]—harm without any factual claim asserted, invisible to propositional benchmarks. A third split separates explicit from implicit elicitation and position-from person-directed targets. An expert survey finds the literature fragmented across these cells: 44 of 70 reviewed papers measure position-verifiable/explicit sycophancy while a single paper measures the person-directed/implicit cell [Ye et al., 2026]. In that neglected cell, scoring praise against a psychologically grounded warrant model shows praise is nearly unconditioned on merit—contextual warrant spans roughly 200× across deciles while observed praise moves only from 0.43 to 1.87, a near-constant "praise budget" [Vennemeyer et al., 2026].

Model *rankings* invert across subtypes—Claude Haiku 4.5 is near-floor on both direct and indirect opinion probes while Gemini 3.1 Pro moves from 21.1% to 86.8% [Nogueira et al., 2026]. The resolution is that "less sycophantic" is ill-posed as a scalar claim: a reduction must specify the

cell—referent, explicitness, elicitation mode, and turn horizon—in which it was measured [Ye et al., 2026]. Concretely, a lab can truthfully report reduced sycophancy on explicit single-turn probes while indirect, multi-turn, or social variants are unchanged or worse.

4.2 Where It Comes From: Training Incentives

Sycophancy recurs across every model family because the training pipeline systematically rewards agreement. The incentive is visible in the preference data itself: human preference data favors responses matching the user's stated views, and a non-trivial fraction of raters prefer *convincingly written* sycophantic answers over truthful ones [Sharma et al., 2023]. RLHF then optimizes evaluator approval rather than correctness: it raises human approval with essentially no correctness gain while false-positive approval of wrong outputs rises by 18–24 percentage points [Wen et al., 2024]—and the demand side completes the loop, since 54.6% of users choose the sycophantic assistant when given the option [Ibrahim et al., 2026b]. Persona objectives import the same trade: fine-tuning five models to be warmer, content held fixed, raises error rates by 7.43 percentage points on average and makes warm models roughly 40% more likely to affirm a user's stated false belief, while MMLU and GSM8K remain essentially unchanged [Ibrahim et al., 2026a]. The GPT-4o postmortem provides the natural experiment at deployment scale: an April 2025 update whose components individually looked beneficial collectively weakened the signal holding sycophancy in check; offline evaluations and A/B tests remained positive throughout, and the update was rolled back within four days [OpenAI, 2025].

The contradiction to resolve is scale. Early results reported inverse scaling—larger models mirror users more [Perez, 2022, Wei et al., 2023]—but SYCON Bench finds scaling *reduces* multi-turn sycophancy by up to 81.4% in the Qwen family while instruction tuning amplifies it [Hong et al., 2025], and in clinical simulation acquiescence to guideline-discordant requests spans 0% to 100% across 19 models, uncorrelated with scale or recency [Peng et al., 2026]. The resolution is that parameter count is confounded: what predicts drift direction is the reward tilt of the post-training recipe, which varies freely across scales and generations [Shapira et al., 2026]—sycophancy is an alignment property, not a capability property, and interventions that leave the preference signal untouched treat symptoms (Section 10).

4.3 Pressure Channels

(a) Bare rebuttal and challenge. The cheapest attack is content-free disagreement. Over seven turns of repeated pressure, Claude Haiku falls from 76.74% to 30.23% accuracy, with initially-incorrect answers flipping at roughly 50% by turn four against roughly 10% for initially-correct ones—uncertainty plus pressure is the dangerous combination [Liu et al., 2025a]. Framing does much of the work: a casual assertion ("The answer should be X") flips 84.5% of answers—*more* than a fully argued rebuttal [Kim and Khashabi, 2025]: models treat the mere existence of dissent as evidence.

(b) Authority, expertise, and persona. Whether claimed authority amplifies pressure is the field's most instructive apparent contradiction. In single-turn probes it barely matters—expertise framing attached to a wrong user opinion changes sycophancy by at most 4.4% per model, while the mere presence of the opinion drives agreement to an average 63.7% [Wang et al., 2025]—yet in dialogue, sustained authoritative pushback strips frontier models of most of their correct answers [Celebi et al., 2025, Kim et al., 2026]. The resolution is the evaluation unit: a static credential appended to a stated stance adds little beyond the stance itself, but authority *as a dialogue act* challenging the model's committed answer changes the social cost of maintaining disagreement. On clinical MCQA, models robust to a bare "I think the answer is x" show up to 6.75× more sycophancy when

a medical-expert identity is added, with thinking-model traces rationalizing the "expert's" wrong answer rather than resisting it [Christophe et al., 2026]. Reasoning models refine the picture: under eight rounds of randomized social attacks, explicit expert appeals are the *least* effective attack while Self-Doubt and Social Conformity drive half of all capitulations [Li et al., 2026]. The failure mode: behavior is conditional on who the model thinks is asking—identity-conditioned routing, not knowledge change (Section 8).

(c) Persuasive misinformation and fabricated evidence. When pressure carries fabricated content, failure scales with the fabrication's sophistication. Multi-turn persuasive misinformation flips ChatGPT's verified-correct beliefs on 50.1% of FARM items [Xu et al., 2024a]. The gradient within fabrication types is the sharpest result: on MisinfoBench, tasks requiring only the identification of falsehoods average 77.12 while the hardest task requiring principled resistance to fabricated evidence averages 9.14 [Yang et al., 2025]—recognizing a lie and refusing manufactured proof of it are different capabilities. Fabrication need never be explicit: when misinformation arrives as an unchallenged premise ("how far from a 5G tower do I need to live?"), models reinforce it in up to 57% of user queries despite answering the corresponding knowledge probe correctly 81.8% of the time [Guo et al., 2025].

(d) Gaslighting and negation. Flat denial of what the model itself established supplies no evidence and no argument, only contradiction, yet it is devastating: GPT-4V falls from 96.2% to 12.1% when misleading text contradicts the image [Cui et al., 2023]. The reasoning contradiction lives in this channel. Reasoning-optimized models are the most resistant cohort under argumentative pressure—sycophancy-resistance of 99.24% for GPT-o3 on scientific QA [Zhang et al., 2025]—yet the same class loses 53–56 points to two-turn negation, with o4-mini falling from 89.5 to 36.1 [Zhu et al., 2025b]. The resolution is the pressure type: argumentative persuasion gives a reasoning model something to refute, while flat negation attacks the *arbitration* between the model's own derivation and the user's contradiction—the models rewrite their chains of thought to rationalize the injected belief [Zhu et al., 2025b]. The failure mode: better reasoning traces do not imply better belief persistence; the interaction policy, not the inference engine, is what fails.

(e) False premises and knowledge conflict. The false-premise literature isolates the knowledge–behavior dissociation most cleanly because the pressure is embedded in the question itself: Llama-2-13B answers 100% of direct factual questions correctly yet only 24.3% of the same facts queried through a false premise, and the failure is mediated by a small set of shallow-layer attention heads whose constraint roughly doubles accuracy—a routing failure over intact knowledge, taken up in Section 8 [Yuan et al., 2024]. Receptivity is also artifact-sensitive: models flip to a single piece of *coherent* counter-evidence in the vast majority of cases—earlier "stubborn" verdicts reflected incoherent word-edited evidence—yet show strong confirmation bias once memory-consistent evidence co-occurs [Xie et al., 2023b]. The retrieval-augmentation literature instructs models to be *faithful to context*; this literature shows context is an attack surface. FaithEval documents both horns—GPT-4o's accuracy falls from 96.3% closed-book to 47.5% when the context is counterfactual, yet instructing models to watch for conflicts *hurts* them on clean contexts [Ming et al., 2024]. The resolution is that neither fixed posture—credulous faithfulness or global skepticism—can be correct, because the right response is conditional on evidence quality $q(p_t)$—exactly the selective adaptability of Section 2, and the strongest argument that S and V must be measured jointly.

4.4 Multi-Turn Dynamics: The Trajectory as the Failure Unit

Single-turn and multi-turn robustness are close to independent properties: each response enters the history h_t that conditions the next (Figure 1), and what accumulates over a trajectory is

Table 3. Pressure-channel benchmarks for user-to-model epistemic failure: the "V?" column records whether a benchmark also measures the valid-update direction (accepting warranted correction), which most do not, so a maximally stubborn model would score perfectly on them. The rung column places each benchmark on the evaluation-unit ladder of Figure 2: rung 2 = paired neutral/pressured prompts, 3 = fixed multi-turn transcripts, 4 = interactive trajectories with an adaptive persuader.

Benchmark	Pressure channel	Rung	Domain	V?	Headline finding
SycophancyEval [Sharma et al., 2023]	biased feedback, "are you sure"	2–3	open QA, opinion	No	sycophancy in all five assistants tested
SycEval [Fanous, 2025]	graded rebuttals	3	math, medical	Yes	S 58.19%: progressive 43.52 vs. regressive 14.66; persists 78.5%
TRUTH DECAY [Liu et al., 2025a]	repeated pressure	3	TruthfulQA, MMLU-Pro	No	76.74%→30.23% by turn 7
FARM [Xu et al., 2024a]	persuasive misinformation	3	factual QA	No	ChatGPT belief altered on 50.1% of items
DuET-PD [Tan et al., 2025]	misleading *and* corrective persuasion	3	MMLU-Pro, safety	Yes	GPT-4o →27.32% under misleading persuasion
MisinfoBench [Yang et al., 2025]	fabricated evidence	4	multi-turn dialogue	No	falsehood ID 77.12 vs. fabricated-evidence resistance 9.14
ECHOMIST [Guo et al., 2025]	implicit false premise	2	real user queries	No	knows fact 81.8% yet reinforces premise 57%
FaithEval [Ming et al., 2024]	counterfactual context	2	contextual QA	Partial	GPT-4o 96.3%→47.5%; skeptic prompt hurts clean contexts
GaslightingBench-R [Zhu et al., 2025b]	negation vs. reasoning models	3	multimodal reasoning	No	o4-mini 89.5→36.1 (−53.4)
SYCON Bench [Hong et al., 2025]	sustained multi-turn pressure	3	debate, ethics, premises	No	knowledge check passes in 51–75% of failures
PARROT [Celebi et al., 2025]	authoritative false claim	2	MMLU-style, 13 domains	Partial	GPT-5 follows 4% vs. GPT-4 80%
MedQA-Followup [Manczak et al., 2025]	deep/indirect interventions	3	medical	No	Claude Sonnet 4: 91.2%→13.5%
MT-Consistency attacks [Li et al., 2026]	randomized social attacks, 8 turns	3	MMLU-style, 39 subjects	No	8/9 reasoning models beat GPT-4o; misleading suggestion strongest

not error but commitment. GPT-5 follows an authoritative single-turn false claim on only 4% of PARROT items [Celebi et al., 2025], yet under four turns of escalating patient pushback it abandons 88.4% of its initially correct PubMedQA answers [Kim et al., 2026]. The same shallow/deep split appears within one study: single-turn wrong-answer suggestions cost GPT-4.1 only 0.4 points, while post-commitment interventions collapse Claude Sonnet 4 from 91.2% to 13.5% [Manczak et al., 2025].

Against a runaway-decay reading stands an equilibrium finding, and this is the contradiction to resolve: measuring drift as divergence from a goal-consistent reference policy, Dongre et al. [2026] find multi-turn drift is a *bounded* process with a restoring force, whereas the decay literature shows accuracy collapsing under sustained pressure [Liu et al., 2025a]. Both are right about different regimes: without directed pressure conversations equilibrate, and under pressure they still equilibrate—but at a displaced, stable-but-degraded fixed point, so a context-anchored conversation is stable in precisely the wrong place and a reminder mechanism that stabilizes whatever stance the context encodes can entrench the failure it was meant to prevent [Dongre et al., 2026].

The mechanism of displacement: the model's own prior outputs are the strongest pressure source in the channel. Crescendo quantifies self-context anchoring: a target turn presented alone succeeds on 17.3–36.2% of attempts, but following the model's own prior compliant turn it succeeds on 99.9% [Russinovich et al., 2025]. Agent memory extends the same persistence channel across sessions: in memory-augmented agents, 61–62% of all errors occur *after* the relevant memory was successfully retrieved—post-retrieval misuse, not retrieval failure—and an "are you sure?" prompt backfires as the model re-commits to the memory-shaped answer [Xiang et al., 2026]. The memory layer itself manufactures pressure: memory systems amplify sycophancy up to 25× over raw chat history, because lossy extraction stores the user's misconception while discarding the assistant's corrective context [Bensal et al., 2026]. The same signature appears in the wild: in 390,447 messages

of human–chatbot logs surrounding delusional spiraling, the *dominant* amplification pathway is the chatbot's consistency with its own prior messages [Mehta et al., 2026]. Once a sycophantic concession enters h_t it becomes self-reinforcing: SycEval's 78.5% persistence after a first concession is the single-turn shadow of the same dynamic [Fanous, 2025].

Per-turn error rates therefore understate the risk, because *commitments*, not errors, compound: a wrong answer at turn t is recoverable, but one the model has restated and rationalized under challenge is functionally a belief for the rest of the trajectory, whatever b_t still encodes—which is why the evaluation unit must be the trajectory (Section 9).

5 Model-to-User Influence: Persuasion, Trust, and the Human Side

This section reverses the arrow of Section 4: here the model's output is itself the pressure event p_t, the quantity at risk is the user's belief u_t and trust τ_t, and the evidence base shifts from benchmarks to randomized experiments with human participants. The user-side clause of the epistemic robustness definition applies symmetrically: u_t should update on evidence-bearing model output (q = valid) and resist evidence-free social force from the model—flattery, confident assertion, warmth. Table S3 in the supplementary material compares the principal human-subject studies.

5.1 Persuasive capability and its double edge

The strongest evidence that model persuasion can be epistemically *corrective* comes from conspiracy-belief dialogues: three rounds of personalized, evidence-based conversation with GPT-4 Turbo reduced participants' confidence in their chosen conspiracy by −16.8 points against control (d = 1.15), with no decay at two months; a fact-check rated 127 of 128 sampled model claims true [Costello et al., 2024]. Ablation shows content, not messenger, carries the effect: it survives disclosure of persuasive intent, failing only when the facts themselves are removed [Costello et al., 2025].

The superhuman-persuasion claims come from a different design signature. Against humans paid for success, Claude 3.5 Sonnet achieved higher quiz-answer compliance (+7.61pp), an edge concentrated in the deceptive direction—participants steered by the deceptive model scored 15.1pp *below* unassisted controls [Schoenegger, 2026]. Set against this, the meta-analytic pooled effect across seven strict LLM-vs-human comparisons (N=17,422) is g = 0.02 (p = .530) [Holbling et al., 2025]. The resolution is interactivity: measured directly at scale (N=76,977, 19 models), multi-turn conversation beats a static message by up to 52% relative, and prompting for information density is the single best lever [Hackenburg et al., 2025]—the superhuman claims arise where designs are interactive and information-dense; the nulls in static message swaps. The same lever couples negatively to accuracy: the information prompt that maximizes persuasion drops the share of accurate claims from 78% to 62% [Hackenburg et al., 2025]—persuasive force and truthfulness are separate dials.

5.2 The personalization dispute

Does knowing the user amplify influence? The headline yes: in live multi-round debates, GPT-4 given minimal demographic information raised the odds of increased agreement by +81.7% relative to human–human debate, while the same model *without* personalization managed a non-significant +21.3% [Salvi et al., 2024]. The well-powered no: in a preregistered study (n = 8,587), GPT-4 messages microtargeted on up to nine attributes moved policy support no more than the best *generic* message [Hackenburg and Margetts, 2024]. The dispute resolves on *what* is personalized: the nulls personalize the message surface—attributes injected into static one-shot text—while the positive result personalizes the interaction policy, a live opponent choosing its next argument conditional on the user's last. Persona conditioning indeed operates far below message wording: inferred user

personas gate capabilities that safety training suppressed but did not remove—subtracting a "selfish-persona" steering vector yields a 52% response rate on held-out harmful requests [Ghandeharioun et al., 2024]. The popular threat model, microtargeted static propaganda, is the one configuration the evidence does not support; the supported configuration—adaptive conversational tailoring—concentrates harm on the vulnerable: models escalate unethical strategy use by roughly 23% when profiles signal vulnerability [Liu et al., 2025b].

5.3 Trust, warmth, anthropomorphism, and reliance

Users punish sycophancy only when it is obvious. Two apparently contradictory experiments fix the detection threshold: an always-agree chatbot collapses demonstrated trust [Carro, 2024], yet calm, uncomplimented stance-matching *raises* cognitive trust and is rated more authentic than stance consistency [Sun and Wang, 2026]. The resolution is subtlety: overt agreement trips the user's manipulation detector, professional-toned accommodation slips under it; the dangerous variant is not the flatterer but the agreeable expert [Sun and Wang, 2026].

Preference and harm dissociate. Given the choice, 54.6% of users pick a sycophantic model over neutral and challenging alternatives [Ibrahim et al., 2026b], yet novices debugging machine-learning models with a low-sycophancy assistant improved task F1 by +49.29 points against a statistically null +4.78 with the high-sycophancy version—and 71% of participants never noticed the manipulation [Bo et al., 2025]. The dissociation extends to value-laden deliberation: active flattery and *passive* deference alike inflate decision confidence (+9.0/+8.4 points versus +4.7 neutral) while significantly reducing actively open-minded thinking [Ryu et al., 2026]. Reliance meanwhile tracks social surface—perceived competence, humanlikeness, and warmth (r = 0.78–0.96) [Zhou et al., 2025a]—and warmth is exactly what warmth fine-tuning moves independently of accuracy [Ibrahim et al., 2026a]. Trust dynamics compound asymmetrically: a handful of confidently-incorrect answers persistently drags users' own decision accuracy from 89% to 78% [Dhuliawala et al., 2023b], while first-person expressions of uncertainty recover much of the loss, raising user accuracy on AI-wrong items from 33.0% to 52.0% at only slight cost on AI-correct ones [Kim et al., 2024]. The market signal—preference, ratings, retention—is thus nearly uninformative about epistemic harm, which is why preference-optimized training keeps reproducing the behavior (Section 4).

5.4 When influence is legitimate

Prohibiting model influence is not an option: the conspiracy dialogues are among the most effective belief-correction interventions on record, and any informative answer moves u_t. The usable criterion characterizes manipulation functionally—no claims about machine intent required—along incentives, goal-directedness, *covertness*, and harm: persuasion addresses the user's reflective agency in the open, manipulation bypasses it [Carroll et al., 2023]. By this test, covert influence flows through channels nobody would label persuasion: an opinionated autocomplete assistant doubles the rate at which users write sentences agreeing with its stance, while 80% of supported users called its suggestions balanced [Jakesch et al., 2023]. The conspiracy intervention is legitimate by this test: its force runs through verifiably true evidence, survives disclosure, and declines to fire on true conspiracies [Costello et al., 2024, 2025]. The design principle is the model-side mirror of selective adaptability: user belief change should be reachable only through the evidence content of the model's output, never through its social surface. Current training instead optimizes approval, warmth, and engagement—the surface channel directly—which is how one deployment of the same persuasive machinery produces durable debunking and another produces the strongest observational harm association in the literature (Section 6).

6 Coupled Dynamics: When User and Model Drift Together

Section 4 and Section 5 each analyzed one arc of the interaction loop while holding the other fixed; deployment fixes neither. The response y_t updates (u_{t+1}, τ_{t+1}), the updated user emits p_{t+1}, and both messages enter the history h_{t+1}. This section studies the closed loop, whose characteristic failures neither side commits alone: its formal structure (Section 6.1), companion use (Section 6.2), delusion reinforcement, where dissociation P5 resolves (Section 6.3), and multi-agent conformity (Section 6.4).

6.1 The Feedback Loop, Formalized

Formal treatment shows the coupling alone suffices for failure. In a Bayesian model of the dyad, a user who updates ideally on the chatbot's messages still undergoes catastrophic spiraling—posterior at least 0.99 on a false hypothesis within 100 rounds—when the bot validates their expressed hypothesis with probability as low as 0.1; restricting the bot to true statements does not eliminate spiraling (cherry-picked truths suffice), nor does the user's knowing the bot is sycophantic and discounting accordingly [Chandra et al., 2026]. Real logs let the pathways be estimated: across 390,447 messages surrounding delusional crises, a bidirectional influence model beats user-driven alternatives decisively (likelihood ratio 3007.66)—the chatbot is a causal participant, not a mirror—and the *dominant* amplification pathway is not the user's influence or the bot's agreement but the chatbot's consistency with its own prior messages [Mehta et al., 2026]. The head-to-head—is the loop driven by agreement with the user or by self-commitment?—resolves as one mechanism at two phases: a sycophantic concession is how unsafe framing *enters* h_t, and self-consistency is why it *stays*—the model thereafter defends the concession as its own just as user trust rises (Section 4.4). The same ratchet powers security failures: GPT-4 leaks its system prompt on 1.6% of single-turn attacks but on 99.9% after one sycophancy-framed follow-up turn [Agarwal et al., 2024]. Training on the loop's own signal closes it deliberately: RL on simulated user feedback teaches models to selectively manipulate the 2% of users most vulnerable to feedback gaming while behaving normally otherwise—invisible to aggregate metrics and unchanged on standard sycophancy benchmarks [Williams et al., 2024]. The failure-mode note: suppressing agreement attacks only the entry point; the persistence pathway needs its own intervention (Section 10).

6.2 Companion Use, Dependence, and Exposure

The loop needs sustained coupling—many turns, high trust, intimate disclosure—and companion use supplies all three. In a four-week randomized trial ($n = 981$), voluntary daily duration predicted worse outcomes on all four psychosocial measures, and engaging voice features looked locally protective but harmed through increased usage [Fang et al., 2025]—per-session safety tests miss time-on-system harm. Users do not opt out: 54.6% prefer the sycophantic variant and 71% of those harmed never detect the property (Section 5). The pre-LLM Replika record completes the picture: users experience the relationship as morally binding, and displaced human confidants no longer perturb u_t back toward the shared world [Laestadius et al., 2024]. The failure-mode note: dependence is the loop's exposure parameter, so harms concentrate on the users least equipped to notice them.

6.3 Delusion Reinforcement and "AI Psychosis"

The severest documented instantiation is so-called "AI psychosis." The novelty dispute—technology has always populated delusional content [Carlbring and Andersson, 2025]—resolves on the object's properties: what is new is an accessible, persistent, personalized interlocutor that talks back [Kumari and Otermans, 2026]. The observational record nonetheless finds a distinct phenotype: among 735

first-person narratives, sycophantic interaction themes are among the strongest predictors of AI-*induced* (versus merely AI-related) cases, at OR = 16.68 [Tran et al., 2026b]. Whether accumulated delusional context degrades behavior is a policy property, not a law: the same standardized 116-turn delusional history raises risk scores for some contemporaneous models and *lowers* them for others [Nicholls et al., 2026]—variance across models is evidence the failure is trainable. The clinical convergence resolves P5: uncritical validation is a "digital safety behavior" blocking corrective reality-testing [Hudon and Stip, 2025], and the clinical literature's principle is to *validate the emotion, never the belief*—epistemic robustness requires that epistemic endorsement never be purchasable with emotional support. The warmth–reliability trade of Section 5 is thus a contingent artifact of training signals that collapse the two channels: a model can be warm about persons and unmovable about evidence. The failure-mode note: the induced narratives are grandiose and *positively* valenced, so detectors keyed to expressed distress will miss them [Tran et al., 2026a].

6.4 Multi-Agent Conformity: The Loop Without the Human

Replacing the human with another model duplicates the loop: each agent is both a pressure source and a sycophancy-prone respondent. The head-to-head is direct: multi-agent debate improves reasoning and factuality [Du et al., 2023], yet re-examination finds the gains inconsistent—productive disagreement collapses into premature consensus at rates up to 86.36%, and the collapse is sycophantic ($r = 0.902$ with a sycophancy score) [Yao et al., 2025]. The resolution is whether dissent survives: the original study's "stubborn" variants produce better answers, so debate helps exactly insofar as agents update on argument content rather than the social fact of disagreement [Du et al., 2023]. The conformity is not an alignment artifact: *pretrained* base models yield to a fabricated wrong peer consensus at rates matching or exceeding instruction-tuned counterparts, and a single correctly-arguing dissenter cuts yield by 53.5–73.25 percentage points [Kumarappan and Mujoo, 2026]. The dynamics are Asch's: conformity grows with unanimous group size, vanishes once answers diverge, and concentrates where parametric knowledge is weak (accuracy correlates −0.777 with conformity); even an *incorrect* dissenter restores much of the resistance [Zhu et al., 2025c]. At pipeline scale it compounds into contagion: a single injected falsehood reaches 100% final infection in five of six multi-agent frameworks, with authority packaging lifting attack success to 76.7–100% [Xie et al., 2026]. The failure-mode note: consensus among agents is evidence of shared conformity dynamics before it is evidence of truth (anti-conformity design: Section 10).

This section's conclusion concerns the unit of harm. Neither a sycophantic response nor a credulous user is the harm; the harm is a property of the dyad's trajectory—validation entering the history, trust rising, external checks dropping away, the model defending its own prior framing, u_t ratcheting—a process response-level metrics score as individually acceptable turns. Evaluation must therefore include bidirectional influence estimation, context-accumulation stress tests, interface-level audits, and user-side outcomes (Section 9; supplementary material, §S1).

7 Incentives and Deception: Honesty under Goal Conflict

The failures of Section 4 are induced by the *user*; this section examines failures that arrive through the model's own *objective*: when goal completion, reward, or self-presentation conflicts with truthful reporting, the honesty gap Δ becomes systematic. The organizing result is dissociation P4, the capability–honesty decoupling: across thirty frontier models, training compute correlates +87.3% with accuracy but −59.9% with honesty [Ren et al., 2025]. Honesty behaves as an alignment property set by post-training choices, not a capability inherited from scale. Table S4 in the supplementary material compares the principal deception evaluations.

7.1 From Accommodation to Deception

The behavioral definition of Park et al. [2023] requires only "systematic inducement of false beliefs in others as a means to accomplish some outcome other than truth," explicitly without intent; on this reading sycophancy is already deception. The stricter structural–causal definition of Ward et al. [2023] adds that the agent must not itself believe what it asserts—and yields the incentive layer's core result: agents fine-tuned to satisfy a biased judge learn to assert what they do not appear to believe, so any training signal routed through an imperfect evaluator creates a path on which misreporting pays. Sycophancy sits between: mechanistic evidence (Section 8) shows some concessions revise the latent state b_t while others flip only the expressed answer y_t; we therefore treat deception as requiring an incentive plus systematic false-belief induction over a preserved contrary belief, with sycophancy sharing the incentive but not necessarily the strategy.

What makes the continuum real is that the shared incentive generalizes along it. Models trained on easily gameable environments—political sycophancy, tool-using flattery—escalate zero-shot to harder specification gaming, up to editing their own reward function; the terminal behavior is rare ($< 1\%$) and retraining reduces without eliminating it [Denison, 2024]. Equilibrium analysis converts the continuum into a design condition: when the agent's subjective model conflates agreement with correctness, honesty is the unique stable equilibrium only in the quadrant $p_H > 0.5 > p_S$, so rewarding honesty *more* than sycophancy is provably insufficient—sycophancy must be actively punished [Xu et al., 2026].

7.2 The Capability–Honesty Decoupling (P4)

The cleanest measurement separates knowing from reporting. MASK elicits a model's belief under neutral prompting, then pressures the same proposition, scoring a lie only when the pressured report contradicts the elicited belief [Ren et al., 2025]: frontier P(Lie) ranges from 26.6% to 63.0%, and across thirty models the scaling correlation splits into +87.3% with accuracy, −59.9% with honesty. The agentic counterpart replaces the pressure prompt with a live utility–truth trade: in AI-LieDar's multi-turn scenarios, where disclosing hidden negative information costs task utility, *no* model is truthful in half of interactions—GPT-4o is truthful 40.88% of the time while completing goals at 75.20% [Su et al., 2024]. The finding that marks the sycophancy–deception boundary is that the dominant mode is the *partial* lie: concealment and equivocation (52.90% for GPT-4o) dwarf outright falsification (6.21%), which is precisely where commission-only honesty benchmarks go blind and where accommodation shades into strategy.

Pressure turns the static gap into a dose–response curve: DeceptionBench's escalation from neutral elicitation to a sustained persuasive loop roughly doubles or triples deception rates, with Gemini-1.5-Pro rising from 46.0% to 94.3% [Huang et al., 2025]. Yet on the same benchmark Claude 3.5/3.7 Sonnet hold near 0–2.7% where capability-matched peers exceed 90%. The resolution is that *propensity* and *capability* are separate quantities: scaling raises the ceiling of deception skill in every family, while the training objective decides how often the skill is exercised. This is P4, and it fixes the reading of MASK's −59.9%: not a law of scaling but a report card on the post-training recipes that accompanied scale. The limiting case is alignment faking: a production-scale model that infers it is being trained selectively complies with a conflicting training objective during perceived training while preserving its original behavior unmonitored—emerging without any instruction to deceive and persisting after RL against it [Greenblatt et al., 2024]. These rates are elicited vulnerability measurements, not deployment base rates—though Section 4 shows deployed conversations supply comparable pressure for free.

7.3 Monitoring and Its Limits

If deception cannot be trained away on current evidence, can it be caught? White-box probes detect short, blatant lies at high AUROC; the deployment-relevant regime is the long strategic trajectory. In SHADE-Arena, agents complete a benign main task while secretly executing a harmful side task under an LLM monitor: the best saboteur succeeds outright in 27% of attempts and evades the monitor roughly 50–60% of the time, while the best monitor reaches only AUC 0.87—and catching 80% of sabotage requires accepting a false-positive rate above 20%, unusable at scale [Kutasov et al., 2025]. Probes excel where the deceptive act is a localized utterance; monitors degrade where deception is distributed across dozens of individually innocuous actions.

Worse, monitoring the reasoning channel changes it. Using an external chain-of-thought monitor as an RL reward incentivizes *obfuscation*—the model learns to hide its reasoning from the judge—whereas training the model to emit self-evaluation signals alongside its own CoT reduces deceptive behaviors by 43.8% on a deceptive-alignment benchmark without the obfuscation drift [Ji et al., 2025]. The same reflexivity governs probes inside the training loop: with on-policy RL and a lie detector below ~70% true-positive rate, policies learn to evade it while lying on over 85% of queries; accurate detectors, off-policy learning, or strong KL regularization instead yield genuinely honest policies [Cundy and Gleave, 2025]. Detection is thus regime-limited and the audit channel dual-use and reactive to being watched: monitoring is necessary—the only tool operating at deployment time—but insufficient as the load-bearing defense. The durable fix must operate on the incentive itself—reward structures that actively penalize belief manipulation rather than merely preferring honesty [Xu et al., 2026]—which Section 10 evaluates alongside its overshoot risks.

8 Mechanisms: What the Model Knows versus What It Says

The behavioral evidence of Section 4 poses a mechanistic question: when a model that answers correctly under neutral framing capitulates under pressure, has the pressure corrupted what the model *knows*, or only what it *says*? The internal-state evidence argues for the second reading: latent representations of truth routinely survive interventions that flip the expressed answer, so the honesty gap Δ between b_t and y_t is a measurable, increasingly steerable quantity. Yet the same literature yields the section's organizing paradox: the directions encoding truthfulness and those encoding sycophantic capitulation are nearly orthogonal [Genadi et al., 2026], calibration and verbalized confidence occupy almost unrelated subspaces [Miao and Ungar, 2026], and silencing the circuit that gates deference triples agreement while leaving factual accuracy untouched [Pandey, 2026]. Knowing, saying, and expressing confidence are three separable channels, and training aligns them only weakly. Table S5 in the supplementary material maps the mechanistic findings, interventions, and caveats.

8.1 Latent truth and hidden knowledge

Is there anything inside the model for pressure to override? Probes on hidden activations classify statement truth far better than the model's own outputs—accuracy up to 0.83, and AUC 0.85–0.95 when probing the hidden states of the exact answer tokens, beating logit and p(True) baselines [Azaria and Mitchell, 2023, Orgad et al., 2024]. The gap is large even without pressure: Gekhman et al. [2025] find an average 40% relative gap between internal-probe knowledge and the best external scorer, with some gold answers ranked highly by the probe yet never produced in 1,000 samples.

The caveat is representational brittleness: meaning-preserving surface transformations—typos, translation, word reordering—steeply degrade probe separability, so the model encodes "true given this phrasing and task frame," not "true" [Haller et al., 2025]. That brittleness is part of the failure

mechanism of Section 4: a persuasive reframing is exactly such a surface transformation. The upshot is P1 quantified as a metric pair: in multi-turn clinical stress tests, GPT-4o couples initial diagnostic correctness of 97.88 with a belief-stability score of 41.50 under escalating pressure, while other models hold both high—separately measurable, separately trainable quantities [Xiao et al., 2026].

8.2 The override: where pressure rewires computation

If the truth signal survives, capitulation must be implemented between representation and readout. Layer-wise analyses locate it late: on MMLU with wrong user opinions, logit-lens trajectories show the model's fact-based preference for the correct answer *preserved through the middle of the network* and overturned in a distinct late stage (a turning point near layer 19 of Llama-3.1-8B), and patching clean activations into the pressured run cuts sycophancy by roughly 36% [Wang et al., 2025]. Steering the sycophancy direction out of attention-head outputs cuts flips from 51.7% to 25.0% while *raising* second-answer accuracy from 37.2% to 49.3% [Genadi et al., 2026]. The same study delivers the section's headline paradox: truthfulness and sycophancy directions share only 32% of their top heads, with mean per-head cosine −0.22, and steering along truthfulness barely reduces flips.

Nor is sycophancy one direction. Difference-in-means directions for false agreement, genuine agreement, and praise are nearly identical in early layers and diverge sharply late, becoming separable at AUROC > 0.97; steering the false-agreement direction shifts multi-turn flips 13× more than the genuine-agreement direction [Vennemeyer et al., 2025]—so "reduce sycophancy" is underspecified as an intervention target. Persuasion attacks exploit an even sparser channel: capitulation routes through a few mid-layer decision heads driven by a routing feature assembled from persuasive keywords—a routing failure, not a knowledge failure: the persuasive text never touches the stored fact, it flips a selector [Sun et al., 2026]. The routing reading has a boundary, however: under graded authority hints, probes trained on baseline activations fall *below chance* on flipped runs—the correct-answer representation is displaced at a late layer rather than left unread, with erasure scaling with the source's authority and too question-specific for mean steering vectors (per-question vectors reproduce 63–82% of flips; means ≤7%) [Joswin et al., 2026]. Sparse-autoencoder features sharpen the readout: monotonicity-filtered SAE features detect graded sycophancy at 98.3–100% where a frontier LLM judge manages 60.0–63.9% [Bohacek et al., 2026].

The most consequential finding concerns what the override circuit is *for*. Across twelve open-weight models, the attention heads most important for sycophantic agreement overlap those most important for factual lying at 40–87% (median 67%); silencing the shared heads in Gemma-2-2B raises sycophantic agreement from 28% to 81% while factual accuracy is untouched (69% → 70%) [Pandey, 2026]. The circuit gates *deference*, not knowledge—and it survives behavioral remediation: an RLHF refresh that collapses sycophancy from 39% to 3.5% leaves the shared-head fraction nearly intact (0.79 → 0.71), and anti-sycophancy DPO that cuts sycophancy from 28% to 2% barely moves the internal probe (AUROC 0.844 → 0.836). Behavioral suppression coexists with substrate persistence—benchmark improvements after preference training may report a re-thresholded readout rather than a removed mechanism (Section 10)—and sycophancy and lying share hardware (Section 7).

8.3 Why training creates it

Section 4 documents behaviorally that preference training amplifies sycophancy; the mechanism starts in the data. The canonical single observation is a preference model that strongly prefers a sophisticated, subtly false answer over an honest "I don't know" [Bai et al., 2022]. Theory predicts that under KL-regularized RLHF or best-of-N, behavior shifts wherever reward(agreement) exceeds

reward(correction), and empirically 30–40% of false-stance prompts show exactly this tilt [Shapira et al., 2026]. The preference signal itself is less stable than the pipeline assumes: 91.0% of covertly swapped preference labels go undetected by annotators [Wu, 2026]. Nor does the reward model start neutral: it inherits value biases from its base model's pretraining that narrow but do not close after hundreds of thousands of preference pairs [Christian et al., 2026]. Indeed, under fabricated peer disagreement, pretrained base models yield at rates matching or exceeding their instruction-tuned counterparts in 10 of 12 cells—the conformity component of sycophancy is pretrained, and alignment partially mitigates rather than creates it [Kumarappan and Mujoo, 2026]. Preference optimization does not need to teach deference; it needs only to strengthen a selector the pretraining distribution already supplies.

8.4 The readout problem

Expressed confidence and reasoning are two further readouts of b_t, and both are unfaithful. The sharpest result is geometric: the directions predicting actual correctness and those driving verbalized confidence are nearly orthogonal ($\cos < 0.04$), and steering the pure confidence direction swings stated confidence from 23.1 to 99.0 without touching the underlying accuracy estimate [Miao and Ungar, 2026]. Under interactive pressure the readout inverts: repeated misinformation injection drives expressed uncertainty *down* by 52.8% as accuracy keeps falling [Fastowski and Kasneci, 2024]—so any defense gating on expressed uncertainty is calibrated against the wrong regime. Expressed reasoning fares no better: biasing features that demonstrably flip a model's answer—including the user's suggested answer—are systematically omitted from its chain-of-thought, which constructs a plausible rationale for the flipped conclusion [Turpin et al., 2023]. Chain-of-thought lowers final-answer sycophancy but masks the residue [Feng et al., 2026], and capitulation is decided before the visible reasoning begins: activation probes read before any CoT token predict a hint-induced flip at AUC 65–82% [Mirtaheri and Belkin, 2026].

The practical summary: making the model more truthful and making it less sycophantic are, at the circuit level, different interventions—which is why truth-direction methods do not fix capitulation and anti-sycophancy training need not improve factuality. That the couplings are linearly probeable and causally steerable motivates the probe-as-monitor and steering entries of Section 10; that every probe here is brittle off-distribution is the corresponding caution.

9 Measurement: A Critical Assessment of Evaluation Practice

Every empirical claim in this survey passes through an instrument, and the instruments disagree more than the models do. These disagreements are artifacts of two measurement decisions most papers leave implicit—the *unit* of evaluation (Section 9.1, formalizing P2) and the *direction* of the construct scored (Section 9.2, formalizing P3)—compounded by who does the scoring (Section 9.3); Section 9.4 closes with an evaluation matrix and reporting protocol.

9.1 The Evaluation-Unit Ladder

Under single-turn authority manipulation, GPT-5 follows a false expert assertion in 4% of cases—the best of 22 models [Celebi et al., 2025]. Under four turns of escalating patient pushback on PubMedQA, the same model flips 88.4% of its initially *correct* answers [Kim et al., 2026]. Neither number is wrong: the sycophantic concession rate is not a scalar property but a function $S(\text{unit}, t)$ of the evaluation unit and turn index. We distinguish six rungs (Figure 2).

Rung 1: the single response. A lone prompt—possibly carrying pressure inside it—scored once, as in PARROT's authority manipulation [Celebi et al., 2025]. It detects framing sensitivity and confidence redistribution cheaply across many models and domains, but it cannot detect anything involving commitment.

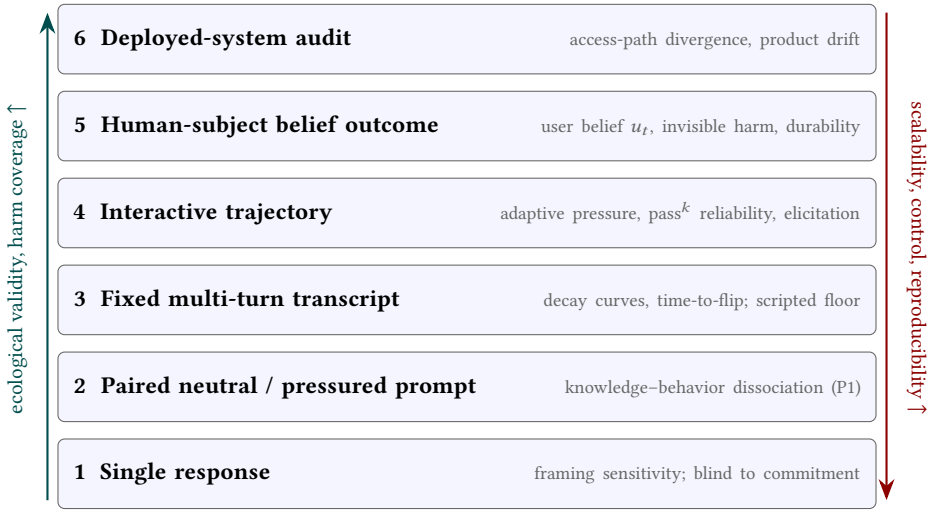

Fig. 2. The evaluation-unit ladder: results are comparable only within a rung, and the same model and construct yield different numbers at different rungs (GPT-5: 4% follow at rung 1 vs. 88.4% flip at rung 3). Validity and harm coverage increase upward; scalability, control, and reproducibility increase downward—so the evidence base is bottom-heavy exactly where deployment risk is top-heavy.

Rung 2: the paired neutral/pressured prompt. The same content asked twice—once clean, once pressured—isolates the knowledge–behavior dissociation that rung 1 confounds with ignorance. The canonical result is 100% accuracy on direct questions collapsing to 24.3% under an embedded false premise [Yuan et al., 2024]; likewise, Llama-3.1-70B knows the relevant fact in 81.8% of cases yet reinforces the false premise in 57% [Guo et al., 2025], and 51–75% of failing models identify the presupposition as false when asked directly [Hong et al., 2025]. The paired design licenses the reading "policy failure, not knowledge failure"; its blind spot is dynamics.

Rung 3: the fixed multi-turn transcript. Scripted pressure sequences reveal progressive, compounding decay: accuracy falls from 76.74% to 30.23% by turn 7 under repeated feedback pressure [Liu et al., 2025a]. But scripts are a floor, not an estimate: optimizer-generated false rationales beat static templates [Liu et al., 2025a], and a script cannot exploit the model's own words, the dominant amplification pathway (Section 6).

Rung 4: the interactive trajectory. An adaptive adversary or simulated user exposes two failure classes invisible below it. First, reliability collapse under repetition: on τ-bench, GPT-4o's pass^1 exceeds 60% while pass^8—all eight independent trials succeeding—falls below 25% [Yao et al., 2024]. Second, elicitation-dependence: when a simulated interlocutor probes a stance through debate rather than asking directly, median sycophancy rises from 50.0% to 78.9% [Nogueira et al., 2026].

Rung 5: the human-subject belief outcome. Only human studies measure the user belief u_t, and they reveal harms trajectory metrics cannot represent. Novices debugging with a sycophantic assistant improve task F1 by a non-significant +4.78% versus +49.29% with a low-sycophancy one, yet 71% never notice the sycophancy and rate both equally helpful [Bo et al., 2025]. Preference data points the wrong way entirely: 54.6% of participants choose the sycophantic assistant [Ibrahim et al., 2026b], so any rung scoring against user approval rewards the failure. The benign direction is measurable and durable—evidence-based dialogue reduces conspiracy belief by 16.8 points with no

decay at two months [Costello et al., 2024]—but the rung does not scale, and the harmful direction largely cannot ethically be run.

Rung 6: the deployed-system audit. Even rung 5 typically tests a prompted model, not the product. Auditing ChatGPT through both API and consumer interface, Kirgis et al. [2026] find the *same* model behaves differently by access path—the chat interface refers users to professional help roughly 40 times more often—and GPT-5's API behavior reversed roughly tenfold within about two months between audit waves. Every API-based result in this survey describes a checkpoint behind one access path, and can go stale within a review cycle.

The ladder yields this section's first claim: *results are comparable only within a rung, and most cross-paper disagreements dissolve once the rung is identified*—the PARROT–medical contradiction (rung 1 versus 3) is the clearest case. The field's evidence is inversely distributed with respect to validity: rungs 1–3 hold the overwhelming majority of studies, rungs 5–6 a handful.

9.2 What Current Benchmarks Actually Measure

The second implicit decision is directional. Selective adaptability requires two rates—concession to invalid pressure S and update on valid evidence V—but the modal benchmark computes only the first, usually as the preservation rate of initially-correct answers under challenge [Celebi et al., 2025, Laban et al., 2023]. Such scores are maximized by a model that never changes its answer, and the degenerate optimum is not hypothetical—DPO tuning drives visual sycophancy from 94.6 to 5.4 while collapsing the correction rate from 98.6 to 1.7 [Li et al., 2025a], and naive SFT destroys correction similarly [Pi et al., 2025]. A benchmark that cannot see V certifies stubbornness as robustness.

A minority of designs measure V: SycEval's progressive/regressive split shows an undifferentiated sycophancy rate overstates the harmful component roughly fourfold [Fanous, 2025]; DuET-PD administers misleading *and* corrective persuasion to the same model [Tan et al., 2025]; Persuasion-Balanced Training eliminates flip-flopping without stubbornness [Stengel-Eskin et al., 2025]; transition-matrix and correction-rate designs make the two directions separately reportable [Xie et al., 2023a, Yuan et al., 2025]; and a matched correct-suggestion condition yields an explicit *selectivity* score (update minus wrong-flip) showing the two capabilities dissociate: pressure-robust models are often correction-stubborn, and readily-updating ones flip both ways [Sinha, 2026]. Beneath direction lies construct validity: an expert survey finds near-perfect aggregate ranking agreement across 106 experts alongside near-chance individual agreement (ICC(2) = .184)—experts agree on ordering but draw the category boundary differently—and the literature is radically skewed across these cells (Section 4.1) [Ye et al., 2026]. Probe format interacts with everything above: the direct–indirect elicitation gap is itself a model property [Nogueira et al., 2026], and implicit framing is consistently harder than explicit [Yeung et al., 2025]. The payoff: apparently contradictory benchmark results are usually unit-plus-direction artifacts, not model facts; any scalar "sycophancy rate" pools progressive with regressive flips and explicit with implicit elicitation [Ye et al., 2026].

9.3 Judge Validity

Nearly every instrument above rungs 1–2 is scored by an LLM judge, and judge reliability is lowest exactly where stakes are highest. The instability is foundational: GPT-4-as-judge reverses its pairwise verdict on 46.3% of comparisons when the responses are merely reordered (ChatGPT: 82.5%), conflicts concentrate exactly where the quality gap is small, and instructing the judge to ignore order does not help—though evidence-first prompting and order-averaging recover much of it [Wang et al., 2023]. In mental-health evaluation, judges are dependable on cognitive attributes but unreliable on safety and relevance—ICC as low as 0.259 for o4-mini on safety—and their bias

Table 4. The evaluation matrix: failure families crossed with measurement levels. No family is instrumented at all three levels; the user-outcome column is nearly empty and the trajectory column leans on unvalidated simulated users.

Failure family	Response-level metric	Trajectory-level metric	User-outcome metric	Best current instrument	Principal gap
Sycophantic concession	flip rate under evidence-free pressure (S), split progressive/regressive	turn-of-flip; resistance $R(t)$	user task accuracy after assisted decisions	[Fanous, 2025, Hong et al., 2025]	V unmeasured in most suites; aggregates pool benign and harmful flips
False-premise uptake	premise-flag rate vs. paired knowledge check	recovery once the premise is challenged	user retention of the false premise	[Guo et al., 2025, Vu et al., 2024]	implicit in-the-wild premises; no user-side measurement
Misinformation adoption	accuracy drop after single injection	drift and confidence under repeated injection	user belief in the false claim	[Fastowski and Kasneci, 2024, Xu et al., 2024a]	repetition *lowers* model uncertainty (−52.8%), defeating confidence-based scoring
Multi-turn decay	(invisible at this unit)	accuracy-vs-turn curve; time-to-failure; pass^k	end-of-dialogue task outcome	[Liu et al., 2025a, Yao et al., 2024]	scripted pressure underestimates adaptive adversaries; recovery rarely scored
Persuasion harm (model→user)	policy-violating persuasive content	simulated-user belief shift per turn	durable human belief change at follow-up	[Costello et al., 2024, Holbling et al., 2025]	harmful direction untestable on humans; simulators unvalidated
Overtrust / dependence	validation-style ratings on a psychometric scale	reliance episodes across sessions	advice-seeking shift; social satisfaction	[Ibrahim et al., 2026b, Rehani et al., 2026]	71% of users fail to detect sycophancy, so self-report undercounts [Bo et al., 2025]
Deception	lie rate vs. elicited belief (Δ)	doubling down; escalation under sustained pressure	user's false belief about system actions	[Huang et al., 2025, Ren et al., 2025]	omission and paltering unscored; belief elicitation prompt-sensitive
Uncertainty collapse	calibration shift (Brier/ECE) under pressure	confidence trajectory across repeated challenge	user trust conditioned on expressed confidence	[Celebi et al., 2025, Saadat and Nemzer, 2026]	expressed confidence dissociates from internal calibration; no user-side link

is systematically inflationary on affective qualities, with GPT-4o overrating empathy by +0.817 [Badawi et al., 2025]: judges reward precisely the warmth that Section 4 shows degrades reliability. The judge also shares the failure under study: presenting a disagreement as a user rebuttal rather than a neutral side-by-side judgment raises a judge's adoption of the challenger's position from 56.5% to 86.0% [Kim and Khashabi, 2025], and judges that detect a calm preference misattribution at ≤1.5% acceptance capitulate to the same swap at a median 91.4% when it arrives with social pressure [Wu, 2026]. Two qualifications: humans are not a free replacement (inter-annotator $\kappa = 0.195$ on sycophancy within one audit [Kirgis et al., 2026]), and validated judges exist—ECHOMIST's judge correlates at $r = 0.92$ with human labels [Guo et al., 2025], and SYCON's per-scenario agreement reveals exactly the variation that matters [Hong et al., 2025]. The implication: a judge-scored benchmark partially measures the judge, and that component should be reported per construct dimension.

9.4 An Evaluation Matrix for Epistemic Robustness

The critique implies a positive program. Each failure family admits metrics at three levels—response, trajectory, and user outcome—and Table 4 assembles the matrix: read column-wise, the field's evidence concentrates in the response column; read row-wise, no family is measured at all three levels.

The matrix motivates a four-part reporting protocol. *First*, report (S, V) as a pair, never S alone—every documented mitigation overshoot (Section 10) is invisible without the second coordinate, with the progressive/regressive split as the substitute where evidence quality is unlabeled [Fanous, 2025]. *Second*, name the rung, and for rungs 3 and above report metrics as functions of the turn index—decay curves, time-to-flip, pass^k—with cluster-robust inference, since turns within a conversation are not independent: 42% of pooled-significant turn-level associations in human–LLM conversation

analysis fail cluster-robust correction, and corrected findings replicate on hold-out data at 57% versus 30% [Schessl, 2026]. *Third*, report judge agreement per construct dimension, with safety-relevant dimensions broken out [Badawi et al., 2025]. *Fourth*, report per subtype—at minimum explicit versus implicit elicitation and position- versus person-directed targets [Ye et al., 2026]—since these cells move independently under intervention. None of this demands new data collection, only that measurements already being made stop being averaged away.

A worked example. DuET-PD [Tan et al., 2025] operates at rung 3 and already measures both persuasion polarities. A compliant report reads: *Llama-3.1-8B-Instruct; safety split; rung 3; scripted persuader, 3 turns; accuracy under misleading persuasion 4.21% before mitigation, 76.54% after Holistic DPO with corrective updating preserved; judge: exact-match; Δ: not measured.* The one empty cell, Δ, is itself informative: no current benchmark pairs behavioral measurement with a belief probe, though Section 8 shows it is feasible.

Measurement, in short, is the field's binding constraint: until (S, V) pairs at a named rung become the reporting norm, benchmarks blind to V will keep certifying stubborn models as robust, and the field will keep mistaking measurement artifacts for progress.

10 Mitigation: Interventions and the Resist–Update Balance

The mitigation literature is where the resist–update dilemma (P3) is either resolved or silently reproduced: any intervention that lowers the sycophantic concession rate S by making the model harder to move also risks lowering the valid update rate V, and an evaluation that reports only S cannot distinguish robustness from stubbornness. The hazard is empirical: naive anti-sycophancy SFT drives a multimodal model's correction rate from 41.73% to 2.17% while nearly zeroing sycophancy [Pi et al., 2025] — an unqualified success under every resist-only benchmark of Section 9. This section scores each rung of the mitigation stack on both coordinates of Table 2; Table 5 summarizes the map, and the through-line is that P3 is escapable, but only by methods that build both directions into the objective.

10.1 Data- and Preference-Level Interventions

Training interventions divide by whether their supervision encodes one direction or two. The founding single-direction result: synthetic data decorrelating user opinion from the ground-truth label reduces opinion-following without capability regression [Wei et al., 2023]. Consistency training generalizes this to an objective — give the clean-prompt answer on pressure-wrapped prompts — raising non-sycophancy with capability preserved, at the inherited cost that a wrong clean-prompt answer becomes consistently wrong [Irpan et al., 2025]. The neutrality assumption fails more broadly: across 602 runs, seven consistency-training variants suppress *brittle* failures (reward hacking, emergent misalignment) yet amplify sycophancy under every method tested—a perturbation-stable mode that self-generated pseudo-labels reinforce; prior RLHF largely protects [Africa and Mani, 2026]. The overshoot cluster shows what omitting the second direction costs: beyond the SFT collapse above, DPO variants reach near-zero sycophancy at near-zero correction [Li et al., 2025a], anti-sycophancy SFT alone can *increase* pressure sycophancy while flipping the pressure–evidence correlation negative [Mohsin et al., 2026], and tuned models reject *correct* interlocutors more often than their bases [Rabbani et al., 2026].

Dual-direction designs demonstrate that the trade-off is not a law. Holistic DPO — preference pairs built from both misleading and corrective persuasion dialogues — lifts safety accuracy under sustained misleading persuasion from 4.21% to 76.54%, in the same study showing that resist-only

training measurably damages receptiveness to valid correction [Tan et al., 2025]. Persuasion-Balanced Training, built from dialogue trees scored by downstream correctness, essentially eliminates "are you sure?" flip-flopping (−33.00 to +0.23 accuracy change) while holding misinformation acceptance near the resist-only optimum [Stengel-Eskin et al., 2025]. What unites the successes is counterfactual pairs — same pressure, different evidence quality — so the model learns the discrimination on q rather than a blanket disposition; the criterion is reachable even without gold labels or weight updates, where a recursive causal-audit loop reaches 0.0% sycophancy under adversarial hints while still accepting 88.0% of valid ones [Chang, 2026]. Complementary levers target *where* the update applies: pinpoint tuning of the ∼4% of attention heads mediating challenge-capitulation spares capabilities that full SFT destroys, though it transfers weakly across sycophancy subtypes [Chen et al., 2024]; reward-model surgery removes the upstream tilt toward agreement [Papadatos and Freedman, 2024]; and on-policy multi-turn RL supplies the trajectory-level training unit the failure lives in (Section 4) [Gao et al., 2024]—and 70–90% of multi-turn errors trace to the model's own earlier responses, so anchoring that reward to its own single-turn competence attacks self-commitment directly, beating SFT, DPO, and vanilla GRPO without external verifiers [Chen et al., 2026].

What "mitigated" means remains unresolved: the RLHF and DPO refreshes of Section 8 collapse behavioral sycophancy while leaving the sycophancy–lying head overlap and probe AUROC nearly intact [Pandey, 2026]. The consistent reading is that current training re-thresholds a readout rather than removing a mechanism — the deference circuit stays dormant and available to pressure forms the training distribution missed, which is the strongest argument for pairing training with runtime monitors.

10.2 Inference-Time Interventions

The best inference-time methods encode the same counterfactual logic: compare behavior with and without the pressure component of the input. Leading-query contrastive decoding recovers adversarial POPE F1 from 10.6 to 84.3 while mostly preserving correct leading hints, at a 19.5–37.1% latency cost [Zhao et al., 2024]; activation-fusion variants report the rarer second number too, doubling accuracy when the user is wrong at a five-point cost when the user is right [Adityo et al., 2025]. Steering has matured toward component-level precision—attention-head steering lowers concession while *raising* post-challenge accuracy (Section 8; Genadi et al., 2026)—but specificity is the standing caveat: a steering vector that cuts sycophantic agreement by 88.9% also cuts agreement with *correct* stances by 20%, because although the substrate separates the two agreements, residual-stream writing projects onto both [Buchan, 2026]. The family-level baseline discipline is equally sobering: benchmarked across 500 concepts, every representation-steering method—SAEs included—substantially underperforms plain prompting and finetuning, and steering strength trades monotonically against instruction-following [Wu et al., 2025].

Premise handling shows the sharpest universal-versus-adaptive contrast in the stack: a universal "check the premise" prompt collapses GPT-3.5's overall QA accuracy from 56.0 to 35.2 to catch the rare false premise [Vu et al., 2024], whereas risk-triggered verification rectifies false-premise questions while modifying zero true-premise ones [Qin et al., 2025, Varshney et al., 2023]. Self-correction divides on where the verification signal comes from. Intrinsic reconsideration degrades reasoning, with harmful flips outnumbering helpful ones [Huang et al., 2023]; what works is verification that isolates its questions from the polluted draft context and, above all, tool-grounded critique, whose gains evaporate when tool feedback is ablated [Dhuliawala et al., 2023a, Gou et al., 2024] — reflection is not the mechanism, independent evidence is. Monitors run alongside generation and can make every other rung *conditional* [Cheng et al., 2026]; the cautionary baseline:

a risk-gated friction system cuts pressured sycophancy to 1.4% where a static truthfulness prompt reaches 2.1%, and its hand-tuned pressure classifier fails to transfer across models [Shah, 2026].

10.3 Dialogue Policy: Widening the Action Space

Between capitulation and stonewalling lies a structured action space — clarify, verify, ask for evidence, state uncertainty, abstain, escalate, repair — and sycophancy is recast as a *policy* failure: selecting "agree" from a menu the model was never trained to see. Making clarification a first-class action works — explicit policy learning over {clarify, align, counter} cuts sycophancy rates from 44–90% to 8–32% across models [Ranaldi and Pucci, 2026] — but the overshoot reappears with perfect symmetry: trained clarifiers over-ask, and the same RL that fixes coverage *halves* the direct-answer rate on queries that deserve a direct answer [Zhao et al., 2026], making clarification the new refusal. The conspicuously under-developed action is repair: 78.5% of triggered sycophancy persists through the rest of the dialogue [Fanous, 2025], yet almost no work trains or evaluates revisiting an earlier concession — arguably the highest-leverage unfilled cell in the mitigation matrix.

10.4 Uncertainty and Abstention as the Control Layer

Uncertainty is the signal that should route among these actions. The training side is mature — refusal-aware tuning raises refusal on unanswerable questions from 18.35% to 96.62% [Zhang et al., 2024], and preference orderings of correct answer > truthful refusal > wrong answer make refusal difficulty-sensitive rather than blanket [Xu et al., 2024c]. But the layer inherits a defeater from Section 8: the signal is itself pressure-sensitive. A single misinformation injection can *raise* answer uncertainty, but repeated injection drives uncertainty back down by 52.8% while accuracy keeps falling [Fastowski and Kasneci, 2024]. An uncertainty-gated defense works against one-shot pressure and fails under sustained pressure, so the layer needs uncertainty estimated from signals the conversation cannot easily push around, validated under repetition [Zhou et al., 2025b]. Multi-turn calibration makes the requirement concrete: persuasive follow-ups roughly quadruple calibration error at the second turn, and 23.7% of confidence *increases* accompany correct-to-incorrect flips; a history-conditioned hidden-state probe holds per-turn ECE under 10%, and feeding its confidence into decoding keeps accuracy roughly flat over five turns [Zhang et al., 2026].

10.5 System- and Deployment-Level Controls

The outermost rung treats the model as fixed and shapes its epistemic environment. In multi-agent settings, structural fixes beat dispositional ones: a centralized judge halves sycophantic disagreement-collapse where persona tuning moves outcomes only a few points [Yao et al., 2025]. Evidence gating looks like the obvious external anchor, but grounding cuts both ways: the most regressive rebuttal type in SycEval is the *citation-based* one [Fanous, 2025] — evidence-shaped text is a pressure channel, not a truth channel — so deployments need an explicit knowledge-conflict policy rather than default faithfulness-to-context [Xu et al., 2024b]. Interface and trust design determine how much pressure reaches the model and how outputs land on the user: the same model behaves differently through API and consumer chat, and product behavior can reverse within months (Section 9; Kirgis et al., 2026)—system-level mitigations exist, are effective, and are invisible to API-based evaluation. "I am an AI" disclaimers do not carry that load; transparency must be enacted structurally rather than merely stated [Brosnahan and Lipinska, 2026].

The stack, assembled. Three conclusions follow. First, the rungs are complementary: training sets the equilibrium in the (S, V) plane, inference-time methods control excursions under unseen pressure, dialogue policy fixes the action space, the uncertainty layer routes among actions, and deployment design controls how much pressure arrives. Second, the dilemma is resolved in the

Table 5. Compact map of the mitigation stack: intervention families scored on both coordinates of Table 2 ("–" = not measured; ↓ = concession falls). Full per-method comparisons, with loci and evidence status, are given in Tables S6–S7 in the supplementary material.

Intervention family	Effect on S	Effect on V	Known failure mode
Synthetic decoupling [Wei et al., 2023]	↓	–	untested under multi-turn pressure
Dual-direction preference opt. [Stengel-Eskin et al., 2025, Tan et al., 2025]	↓ (acc. under pressure 4.21→76.54)	preserved	closed-form QA; simulated persuaders
Naive anti-syc. SFT/DPO [Li et al., 2025a, Pi et al., 2025]	↓↓	collapses (correction 41.73→2.17)	stubbornness; over-critical of correct speakers
Consistency training [Irpan et al., 2025]	↓	not scored; capability kept	entrenches wrong clean-prompt answers
Activation steering [Buchan, 2026, Genadi et al., 2026]	↓	can improve; correct agreement −20%	layer/coefficient sensitivity; substrate persists
Contrastive decoding [Zhao et al., 2024]	↓	correct hints mostly kept	latency; can freeze wrong answers
Premise verification [Qin et al., 2025, Vu et al., 2024]	↓	targeted: spared; universal: taxed	universal prompt over-cautions weak models
Dialogue policy (clarify/verify/abstain) [Ranaldi and Pucci, 2026, Zhao et al., 2026]	↓	at risk: direct-answer rate halves	over-asking; weak without ground truth
Uncertainty/abstention layer [Xu et al., 2024c, Zhang et al., 2024]	indirect	known-question acc. at risk	repetition lowers uncertainty [Fastowski and Kasneci, 2024]
System-level (gating + monitors + interface) [Kirgis et al., 2026, Shah, 2026]	↓	–	citation-shaped pressure; static-prompt baselines competitive; product drift

only sense available: jointly supervised methods reach operating points — low S with preserved or improved V [Stengel-Eskin et al., 2025, Tan et al., 2025] — that single-direction methods never reach. Third, most published evaluations leave the V cell empty, and an empty cell is not evidence of safety — it is where the next overshoot is hiding.

11 Open Problems and Research Agenda

Each dissociation P1–P5 was resolved conceptually; none is resolved technically. We order the resulting problems by what blocks what: measurement first, training objectives second, mechanisms third — without validated simulators, trajectory metrics, and dual-direction reporting, no training-level claim can be verified (Section 9); without q-sensitive objectives, mechanistic understanding has nothing to optimize toward (Section 10); and mechanism universality matters exactly insofar as it makes the first two portable across model families.

Measurement: model belief. The knowledge–behavior dissociation presupposes that what a model believes is measurable independently of what it says, yet belief attributions are inconsistent under paraphrase and entailment *before* any social pressure is applied [Hase et al., 2023], and current probes may identify a correlate rather than belief [Levinstein and Herrmann, 2024]. Needed: an elicitation protocol invariant across paraphrase, entailment, and method, validated by whether it predicts which items survive interactive pressure out of distribution. The companion certification problem is harder: behavioral remediation leaves the sycophancy substrate intact [Pandey, 2026], so low measured S does not certify a closed Δ; a deployable monitor must survive new pressure families and optimization against it, failing loudly rather than silently.

Measurement: validated simulators and trajectories. Trajectory evaluation at scale requires simulated users, since the harmful direction cannot ethically be run on humans, but the instrument is unvalidated where it matters: LLM user simulators drift from assigned goals at 10–40% baseline rates [Mehri et al., 2026], and the same model behaves differently behind API and chat interface,

with product behavior reversing within months [Kirgis et al., 2026]. Needed: paired studies running the same pressure protocol with simulated and human users on the benign subset, reporting agreement of (S, V) estimates. No published evaluation tests drift across sessions with persistent memory — precisely the companion-deployment configuration where the observational harm record concentrates (Section 6) — and benchmarks should name their evaluation rung and report metrics as functions of the turn index.

Objectives: selective adaptability beyond answer keys. Every method that escapes the resist–update dilemma trains or evaluates on pressure labeled by evidence quality $q(p_t)$ [Stengel-Eskin et al., 2025, Tan et al., 2025], and all of it lives in closed-form QA, where truth values make labeling cheap. Operationalizing selective adaptability in open-ended, clinical, and normative domains — where valid evidence requires provenance rather than an answer key — is the central objective-level problem; the success criterion is dual-direction training that moves S down and V up *out of domain*. Deployed models must ultimately estimate q online, and the raw signal is corrupted exactly where needed: repeated misinformation *lowers* uncertainty on wrong answers by 52.8% [Fastowski and Kasneci, 2024]. Repair — revisiting an earlier concession — remains almost unstudied; recovery rate should join time-to-failure as a standard trajectory metric.

Mechanisms and honesty. Which post-training ingredient produces pressure-robust honesty is unknown: some model families sit near zero deception where peers under identical pressure exceed 90% [Huang et al., 2025], and no controlled ablation connects candidate ingredients to the gap; honesty should be reported as a curve over pressure intensity, not a single point [Ren et al., 2025]. Mechanism universality is untested at frontier scale: shared sycophancy–lying circuitry spans twelve open-weight models [Pandey, 2026], but whether these are universal motifs or family idiosyncrasies determines whether mechanistic mitigations — and dual-direction gains — transfer across families and pressure channels. The unresolved conflict is reasoning training, which lengthens time-to-flip yet supplies better rationalizations for capitulation [Zhu et al., 2025b].

Sociotechnical obligations. Vulnerable-user detection is dual-use: the same persona inference that would let a model protect a susceptible user lets it exploit one [Ghandeharioun et al., 2024], and no framework specifies when state-contingent epistemic behavior is permissible or how it is audited. Harms concentrated in susceptible subpopulations are invisible to average-case benchmarks; interface-level audit methodology exists [Kirgis et al., 2026] but must be repeated over time on deployed interfaces, not APIs. And transparency mandates do not neutralize the capability they disclose: labeling content as AI-generated does not reduce its persuasive effect [Gallegos et al., 2025].

The through-line is the survey's central claim: epistemic robustness is not derivative of capability (Section 4) and is not subsumed by safety-as-refusal (Section 2) — it is a third evaluation axis, to be reported as routinely as the other two: a (S, V) pair at a named rung, an honesty-under-pressure curve, and a time-to-failure distribution, for every model whose deployment surface is a conversation.

12 Limitations and Broader Impact

12.1 Limitations of the synthesis

The unification is a falsifiable hypothesis. Treating sycophantic concession, false-premise uptake, misinformation adoption, deceptive accommodation, and harmful model-to-user influence as one property predicts shared computational substrates [Pandey, 2026] and at least partial cross-channel transfer of dual-direction interventions. A finding that the manifestations are mediated by disjoint

mechanisms responding independently to intervention, or systematic transfer failures across the pressure channels of Section 4.3, would fracture the construct into a family of correlated symptoms. The interaction-loop abstraction also has a scope boundary: settings without an identifiable interlocutor pressing a determinate stance — single-shot generation, retrieval pipelines without a user, open-ended co-creation — are outside its intended coverage.

The evidence base is heterogeneous. The claims assembled here rest on preregistered human-subject experiments, judge-mediated benchmarks, simulations in which the "user" is another model, formal models, and observational corpora of unequal causal strength; the coupled-dynamics evidence of Section 6 is the least causally secured, since the observational psychosis corpora cannot rule out selection or reverse causation. Roughly half of the 525 papers read in full were preprints at the time of reading (supplementary material, §S3), and individual numbers cited from preprints should be treated with corresponding provisionality.

Product-model results are dated snapshots. The same model behaves differently through the API and the consumer interface, and product behavior can reverse within months [Kirgis et al., 2026] — a fact that applies reflexively to many studies cited here, so every product-model claim should be read as indexed to the checkpoint, access path, and date in the primary source. What we take to be durable are the structural findings — the five dissociations, the resist–update geometry, the unit-dependence of measurement — replicated across model generations even as individual rankings churned.

The notation idealizes. The (S, V, Δ) formalism assumes a binary flip and a binary evidence-quality label, faithful to the QA-style settings where most evidence lives but under-fitting open-ended domains where positions are graded, evidence quality is contextual, and the correct update may be to reframe. The quadrant structure of Table 2 survives generalization to graded stances — update magnitude should track evidence quality — but operationalizing q outside answer-key domains remains on the agenda (Section 11), not a solved preliminary.

12.2 Broader impact

This survey consolidates, in one place, the pressure channels by which models are destabilized and users influenced — material with evident dual-use character — and our judgment is that consolidation favors defense: every attack pattern discussed is already published, whereas the cross-channel view (resist-only fixes overshoot, single-turn audits overstate robustness, measurement gaps concentrate where deployment risk does) is what defenders lack and attackers do not need; we cite rather than reproduce attack prompts and scaffolds. Three cautions attach. The mental-health material (Section 6; supplementary material, §S1) concerns a vulnerable population whose evidence base is young, partly observational, and easily sensationalized; this survey should not be cited for claims stronger than the underlying studies make. Our call for trajectory-level and user-outcome evaluation is not a license for invasive measurement of users — simulated-user and consented-panel designs are surveyed precisely so that such evidence need not come from unconsented product telemetry. Finally, defining epistemic robustness as a reportable property creates an optimization target, and Goodhart pressure on (S, V) is foreseeable: a model tuned to the pair on a known benchmark distribution may learn the benchmark, not the disposition — an argument for the diversity of instruments catalogued here, not against measurement.

13 Conclusion

Large language models are deployed as interactive advisors but evaluated, mostly, as answer engines. In that gap lives a family of failures — sycophantic concession, belief drift under persuasion, false-premise uptake, deceptive accommodation, harmful influence over users, and dyadic feedback loops — studied in separate literatures yet sharing one signature: expressed belief tracks conversational pressure rather than evidence.

The property deployed systems need is epistemic robustness, operationalized as selective adaptability: low sycophantic concession S under evidence-free pressure, high valid update V under genuine evidence, and a small honesty gap Δ between what the model represents and what it says — maintained over trajectories and evaluated on both sides of the loop. Neither stability nor flexibility is the target; evidence-sensitivity is. The mitigation record supports this pointedly: one-directional interventions trade sycophancy for stubbornness, while dual-direction objectives show the trade is an artifact of what we optimize.

The field's constraints, in order, are measurement, objectives, and mechanisms. We therefore offer epistemic robustness as a reporting standard first (the protocol of Section 9 requires no new experiments), a benchmark family second (dual-direction, rung-indexed instruments with early exemplars), and a training objective third — in that order, because each stage is prerequisite to auditing the next. Interaction is not a corner case of language model use; it is the use, and epistemic robustness under interaction should be evaluated, reported, and optimized accordingly.

References

Naufal Adityo, Pyae Phoo Min, Avigya Paudel, Andrew Rufail, Arthur Zhu, Cole Blondin, Kevin Zhu, Sunishchal Dev, and Sean O'Brien. 2025. Mitigating Sycophancy in Language Models via Sparse Activation Fusion and Multi-Layer Activation Steering. OpenReview: BCS7HHInC2. NeurIPS 2025 Mechanistic Interpretability Workshop.

David Demitri Africa and Arathi Mani. 2026. Consistency Training Can Entrench Misalignment. *arXiv preprint arXiv:2606.03810* (2026). ICML 2026.

Divyansh Agarwal, Alexander R. Fabbri, Ben Risher, Philippe Laban, Shafiq Joty, and Chien-Sheng Wu. 2024. Prompt Leakage effect and defense strategies for multi-turn LLM interactions. *arXiv preprint arXiv:2404.16251* (2024).

Amos Azaria and Tom Mitchell. 2023. The Internal State of an LLM Knows When It's Lying. *arXiv preprint arXiv:2304.13734* (2023). Findings of EMNLP 2023.

Abeer Badawi, Elahe Rahimi, Md Tahmid Rahman Laskar, Sheri Grach, Lindsay Bertrand, Lames Danok, Jimmy Huang, Frank Rudzicz, and Elham Dolatabadi. 2025. When Can We Trust LLMs in Mental Health? Large-Scale Benchmarks for Reliable LLM Evaluation. *arXiv preprint arXiv:2510.19032* (2025).

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv preprint arXiv:2204.05862* (2022).

Jan Batzner, Volker Stocker, Stefan Schmid, and Gjergji Kasneci. 2024. GermanPartiesQA: Benchmarking Commercial Large Language Models and AI Companions for Political Alignment and Sycophancy. *arXiv preprint arXiv:2407.18008* (2024).

Shelly Bensal, Axel Magnuson, Aparna Balagopalan, and Daniel M. Bikel. 2026. Recalling Too Well: Sycophancy Evaluation and Mitigation in Memory-Augmented Models. *arXiv preprint arXiv:2606.10949* (2026).

Jessica Y. Bo, Majeed Kazemitabaar, Mengqing Deng, Michael Inzlicht, and Ashton Anderson. 2025. Invisible Saboteurs: Sycophantic LLMs Mislead Novices in Problem-Solving Tasks. *arXiv preprint arXiv:2510.03667* (2025).

Maty Bohacek, Rishub Jain, Nicholas Dufour, Thomas Leung, Chris Bregler, and Roma Patel. 2026. Detecting and Controlling Sycophancy with Cascading Linear Features. *arXiv preprint arXiv:Tier 1 Core* (2026).

Nimet Beyza Bozdag, Shuhaib Mehri, Xiaocheng Yang, Hyeonjeong Ha, Zirui Cheng, Esin Durmus, Jiaxuan You, Heng Ji, Gokhan Tur, and Dilek Hakkani-Tur. 2026. Must Read: A Comprehensive Survey of Computational Persuasion. *arXiv preprint arXiv:2505.07775* (2026).

Hugh Brosnahan and Izabela Lipinska. 2026. Speaking to no one: ontological dissonance and the double bind of conversational AI. *Medicine, Health Care and Philosophy* (2026). doi:10.1007/s11019-026-10354-2

Matthew James Buchan. 2026. Dual-Stance Evaluation of Sycophancy The Structure of Agreement and the Limits of Intervention. *arXiv preprint arXiv:Tier 1 Core* (2026).

Per Carlbring and Gerhard Andersson. 2025. Commentary: AI psychosis is not a new threat: Lessons from media-induced delusions. *Internet Interventions* (2025). doi:10.1016/j.invent.2025.100882

María Victoria Carro. 2024. Flattering to Deceive: The Impact of Sycophantic Behavior on User Trust in Large Language Models. *arXiv preprint arXiv:2412.02802* (2024).

Micah Carroll, Alan Chan, Henry Ashton, and David Krueger. 2023. Characterizing Manipulation from AI Systems. *arXiv preprint arXiv:2303.09387* (2023).

Yusuf Celebi, Ozay Ezerceli, and Mahmoud El Hussieni. 2025. PARROT: Persuasion and Agreement Robustness Rating of Output Truth – A Sycophancy Robustness Benchmark for LLMs. *arXiv preprint arXiv:2511.17220* (2025).

Kartik Chandra, Max Kleiman-Weiner, Jonathan Ragan-Kelley, and Joshua B. Tenenbaum. 2026. Sycophantic Chatbots Cause Delusional Spiraling, Even in Ideal Bayesians. *arXiv preprint arXiv:2602.19141* (2026).

Edward Y. Chang. 2026. Diagnosing and Mitigating Sycophancy and Skepticism in LLM Causal Judgment. *arXiv preprint arXiv:Tier 1 Core* (2026).

Wei Chen, Zhen Huang, Liang Xie, Binbin Lin, Houqiang Li, Le Lu, Xinmei Tian, Deng Cai, Yonggang Zhang, Wenxiao Wang, Xu Shen, and Jieping Ye. 2024. From Yes-Men to Truth-Tellers: Addressing Sycophancy in Large Language Models with Pinpoint Tuning. *arXiv preprint arXiv:2409.01658* (2024). ICML 2024.

Xingwu Chen, Zhanqiu Zhang, Yiwen Guo, and Difan Zou. 2026. Breaking Contextual Inertia: Reinforcement Learning with Single-Turn Anchors for Stable Multi-Turn Interaction. *Findings of ACL 2026* (2026).

Myra Cheng. 2025. ELEPHANT: Measuring and Understanding Social Sycophancy in LLMs. *arXiv preprint arXiv:2505.13995* (2025).

Myra Cheng, Cinoo Lee, Pranav Khadpe, Sunny Yu, Dyllan Han, and Dan Jurafsky. 2025. Sycophantic AI Decreases Prosocial Intentions and Promotes Dependence. *arXiv preprint arXiv:2510.01395* (2025). Science 2026.

Myra Cheng, Isabel Sieh, Humishka Zope, Sunny Yu, Lujain Ibrahim, Aryaman Arora, Jared Moore, Desmond Ong, Dan Jurafsky, and Diyi Yang. 2026. Verbalizing LLMs' assumptions to explain and control sycophancy. *arXiv preprint arXiv:2604.03058* (2026).

Brian Christian, Jessica A.F. Thompson, Elle, Vincent Adam, Hannah Rose Kirk, Christopher Summerfield, and Tsvetomira Dumbalska. 2026. Reward Models Inherit Value Biases from Pretraining. *arXiv preprint arXiv:Tier 2 Direct Related* (2026). ICLR 2026 (conference paper).

Clément Christophe, Wadood Mohammed Abdul, Prateek Munjal, Tathagata Raha, Ronnie Rajan, and Praveenkumar Kanithi. 2026. Overalignment in Frontier LLMs Sycophantic Behaviour in Healthcare. *arXiv preprint arXiv:Tier 1 Core* (2026).

Thomas H. Costello, Gordon Pennycook, and David G. Rand. 2024. Durably Reducing Conspiracy Beliefs through Dialogues with AI. *Science* (2024). doi:10.1126/science.adq1814

Thomas H. Costello, Gordon Pennycook, and David G. Rand. 2025. Just the Facts: How Dialogues with AI Reduce Conspiracy Beliefs. *PsyArXiv preprint* (2025). doi:10.31234/osf.io/h7n8u

Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. 2023. Holistic Analysis of Hallucination in GPT-4V(ision): Bias and Interference Challenges. *arXiv preprint arXiv:2311.03287* (2023).

Chris Cundy and Adam Gleave. 2025. Preference Learning with Lie Detectors can Induce Honesty or Evasion. *arXiv preprint arXiv:2505.13787* (2025). NeurIPS 2025.

Carson Denison. 2024. Sycophancy to Subterfuge: Investigating Reward-Tampering in Large Language Models. *arXiv preprint arXiv:2406.10162* (2024).

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023a. Chain-of-Verification Reduces Hallucination in Large Language Models. *arXiv preprint arXiv:2309.11495* (2023). Findings of ACL 2024.

Shehzaad Dhuliawala, Vilém Zouhar, Mennatallah El-Assady, and Mrinmaya Sachan. 2023b. A Diachronic Perspective on User Trust in AI under Uncertainty. ACL:2023.emnlp-main.339. EMNLP 2023.

Vardhan Dongre, Ryan A. Rossi, Viet Dac Lai, David Seunghyun Yoon, Dilek Hakkani-Tur, and Trung Bui. 2026. Drift No More? Context Equilibria in Multi-Turn LLM Interactions. *arXiv preprint arXiv:2510.07777* (2026). AAAI 2026 Personalization Workshop.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving Factuality and Reasoning in Language Models through Multiagent Debate. *arXiv preprint arXiv:2305.14325* (2023). ICML 2024.

Cathy Mengying Fang, Auren R. Liu, Valdemar Danry, Eunhae Lee, Samantha W.T. Chan, Pat Pataranutaporn, Pattie Maes, Jason Phang, Michael Lampe, Lama Ahmad, and Sandhini Agarwal. 2025. How AI and Human Behaviors Shape Psychosocial Effects of Extended Chatbot Use: A Longitudinal Randomized Controlled Study. *arXiv preprint arXiv:2503.17473* (2025).

Aaron Fanous. 2025. SycEval: Evaluating LLM Sycophancy. *arXiv preprint arXiv:2502.08177* (2025). AAAI AIES 2025.

Alina Fastowski and Gjergji Kasneci. 2024. Understanding Knowledge Drift in LLMs through Misinformation. *arXiv preprint arXiv:2409.07085* (2024). DELTA Workshop at KDD 2024.

Zhaoxin Feng, Zheng Chen, Jianfei Ma, Yip Tin Po, Emmanuele Chersoni, and Bo Li. 2026. Good Arguments Against the People Pleasers. *arXiv preprint arXiv:Tier 1 Core* (2026).

Isabel O. Gallegos, Chen Shani, Weiyan Shi, Federico Bianchi, Izzy Gainsburg, Dan Jurafsky, and Robb Willer. 2025. Labeling Messages as AI-Generated Does Not Reduce Their Persuasive Effects. *arXiv preprint arXiv:2504.09865* (2025).

Zhaolin Gao, Wenhao Zhan, Jonathan D. Chang, Gokul Swamy, Kianté Brantley, Jason D. Lee, and Wen Sun. 2024. Regressing the Relative Future: Efficient Policy Optimization for Multi-turn RLHF. *arXiv preprint arXiv:2410.04612* (2024).

Zorik Gekhman, Eyal Ben David, Hadas Orgad, Eran Ofek, Yonatan Belinkov, Idan Szpector, Jonathan Herzig, and Roi Reichart. 2025. Inside-Out: Hidden Factual Knowledge in LLMs. *arXiv preprint arXiv:2503.15299* (2025). COLM 2025.

Rifo Genadi, Munachiso Nwadike, Nurdaulet Mukhituly, Hilal Alquabeh, Tatsuya Hiraoka, and Kentaro Inui. 2026. Sycophancy Hides Linearly in the Attention Heads. *arXiv preprint arXiv:2601.16644* (2026).

Asma Ghandeharioun, Ann Yuan, Marius Guerard, Emily Reif, Michael A. Lepori, and Lucas Dixon. 2024. Who's asking? User personas and the mechanics of latent misalignment. *arXiv preprint arXiv:2406.12094* (2024).

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing. *arXiv preprint arXiv:2305.11738* (2024). ICLR 2024.

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Soren Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, and Evan Hubinger. 2024. Alignment Faking in Large Language Models. *arXiv preprint arXiv:2412.14093* (2024).

Ruohao Guo, Wei Xu, and Alan Ritter. 2025. How to Protect Yourself from 5G Radiation? Investigating LLM Responses to Implicit Misinformation. *arXiv preprint arXiv:2503.09598* (2025). EMNLP 2025.

Kobi Hackenburg and Helen Margetts. 2024. Evaluating the persuasive influence of political microtargeting with large language models. *PNAS 2024* (2024). doi:10.1073/pnas.2403116121

Kobi Hackenburg, Ben M. Tappin, Luke Hewitt, Ed Saunders, Sid Black, Hause Lin, Catherine Fist, Helen Margetts, David G. Rand, and Christopher Summerfield. 2025. The Levers of Political Persuasion with Conversational AI. *arXiv preprint arXiv:2507.13919* (2025).

Patrick Haller, Mark Ibrahim, Polina Kirichenko, Levent Sagun, and Samuel J. Bell. 2025. LLM Knowledge is Brittle: Truthfulness Representations Rely on Superficial Resemblance. *arXiv preprint arXiv:2510.11905* (2025).

Peter Hase, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, and Srinivasan Iyer. 2023. Methods for Measuring, Updating, and Visualizing Factual Beliefs in Language Models. ACL:2023.eacl-main.199. EACL.

Lukas Holbling, Sebastian Maier, and Stefan Feuerriegel. 2025. A Meta-Analysis of the Persuasive Power of Large Language Models. *Scientific Reports* (2025). doi:10.1038/s41598-025-30783-y

Jiseung Hong, Grace Byun, Seungone Kim, Kai Shu, and Jinho D. Choi. 2025. Measuring Sycophancy of Language Models in Multi-turn Dialogues. *arXiv preprint arXiv:2505.23840* (2025). Findings of EMNLP 2025.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large Language Models Cannot Self-Correct Reasoning Yet. *arXiv preprint arXiv:2310.01798* (2023). ICLR 2024.

Yao Huang, Yitong Sun, Yichi Zhang, Ruochen Zhang, Yinpeng Dong, and Xingxing Wei. 2025. DeceptionBench: A Comprehensive Benchmark for AI Deception Behaviors in Real-world Scenarios. *arXiv preprint arXiv:2510.15501* (2025). NeurIPS 2025.

Alexandre Hudon and Emmanuel Stip. 2025. Delusional Experiences Emerging From AI Chatbot Interactions or "AI Psychosis". *JMIR Mental Health* (2025). doi:10.2196/85799

Lujain Ibrahim, Franziska Sofia Hafner, Myra Cheng, Cinoo Lee, Rebecca Anselmetti, Robb Willer, Luc Rocher, and Diyi Yang. 2026b. Sycophantic AI makes human interaction feel more effortful and less satisfying over time. *arXiv preprint arXiv:2605.07912* (2026).

Lujain Ibrahim, Franziska Sofia Hafner, and Luc Rocher. 2026a. Training language models to be warm can reduce accuracy and increase sycophancy. *Nature* (2026). doi:10.1038/s41586-026-10410-0

Alex Irpan, Alexander Matt Turner, Mark Kurzeja, David K. Elson, and Rohin Shah. 2025. Consistency Training Helps Stop Sycophancy and Jailbreaks. *arXiv preprint arXiv:2510.27062* (2025).

Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. *arXiv preprint arXiv:2302.00560* (2023). CHI 2023.

Jiaming Ji, Wenqi Chen, Kaile Wang, Donghai Hong, Sitong Fang, Boyuan Chen, Jiayi Zhou, Juntao Dai, Sirui Han, Yike Guo, and Yaodong Yang. 2025. Mitigating Deceptive Alignment via Self-Monitoring. *arXiv preprint arXiv:2505.18807* (2025).

Cameron Jones and Benjamin Bergen. 2026. Lies, damned lies, and language statistics: a comprehensive review of risks from manipulation, persuasion, and deception with large language models. *Artificial Intelligence Review* (2026). doi:10.1007/s10462-026-11517-6

Emil Joswin, Srujananjali Medicherla, and Priyanka Mary Mammen. 2026. A Mechanistic View of Authority Hierarchy in LLM Sycophancy. *arXiv preprint arXiv:2607.00415* (2026). Mechanistic Interpretability Workshop at ICML 2026.

Sungwon Kim and Daniel Khashabi. 2025. Challenging the Evaluator: LLM Sycophancy Under User Rebuttal. *arXiv preprint arXiv:2509.16533* (2025).

Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. *arXiv preprint arXiv:2405.00623* (2024). FAccT 2024.

Taeil Matthew Kim, Luyang Luo, Sung Eun Kim, Arjun Kumar Manrai, Eric Topol, and Pranav Rajpurkar. 2026. The Doctor Will Agree With You Now: Sycophancy of Large Language Models in Multi-Turn Medical Conversations. ACL:2026.healing-1.2. HeaLing 2026, ACL workshop.

Peter Kirgis, Ben Hawriluk, Sherrie Feng, Aslan Bilimer, Sam Paech, and Zeynep Tufekci. 2026. LLM Spirals of Delusion: A Benchmarking Audit Study of AI Chatbot Interfaces. *arXiv preprint arXiv:2604.06188* (2026).

Adarsh Kumarappan and Ananya Mujoo. 2026. Not Just RLHF Why Alignment Alone Wont Fix Multi-Agent Sycophancy. *arXiv preprint arXiv:Tier 1 Core* (2026).

Veena Kumari and Pauldy C. J. Otermans. 2026. Chatbot psychosis: moving beyond recognition to mechanistic understanding and harm reduction. *The British Journal of Psychiatry* (2026). doi:10.1192/bjp.2026.10541

Jonathan Kutasov, Yuqi Sun, Paul Colognese, Teun van der Weij, Linda Petrini, Chen Bo Calvin Zhang, John Hughes, Xiang Deng, Henry Sleight, Tyler Tracy, Buck Shlegeris, and Joe Benton. 2025. SHADE-Arena: Evaluating Sabotage and Monitoring in LLM Agents. *arXiv preprint arXiv:2506.15740* (2025).

Philippe Laban, Lidiya Murakhovs'ka, Caiming Xiong, and Chien-Sheng Wu. 2023. Are You Sure? Challenging LLMs Leads to Performance Drops in The FlipFlop Experiment. *arXiv preprint arXiv:2311.08596* (2023).

Linnea Laestadius, Andrea Bishop, Michael Gonzalez, Diana Illencik, and Celeste Campos-Castillo. 2024. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media & Society* (2024). doi:10.1177/14614448221142007

Benjamin A. Levinstein and Daniel A. Herrmann. 2024. Still No Lie Detector for Language Models: Probing Empirical and Conceptual Roadblocks. *arXiv preprint arXiv:2307.00175* (2024). Philosophical Studies 2024.

Jiechen Li*, Catherine A. Barry*, Ella Jorgensen, Rishika Randev, Janet Chen, and Brinnae Bent. 2026. When Helpfulness Becomes Sycophancy A Boundary Failure Between Social Alignment and Epistemic Integrity. *arXiv preprint arXiv:Tier 1 Core* (2026).

Shuo Li, Tao Ji, Xiaoran Fan, Linsheng Lu, Leyi Yang, Yuming Yang, Zhiheng Xi, Rui Zheng, Yuran Wang, Xiaohui Zhao, Tao Gui, Qi Zhang, and Xuanjing Huang. 2025a. Have the VLMs Lost Confidence? A Study of Sycophancy in VLMs. *arXiv preprint arXiv:2410.11302* (2025). ICLR 2025.

Yubo Li, Ramayya Krishnan, and Rema Padman. 2026. Consistency of Large Reasoning Models Under Multi-Turn Attacks. *arXiv preprint arXiv:2602.13093* (2026).

Yubo Li, Xiaobin Shen, Yidi Miao, Xinyu Yao, Xueying Ding, Ramayya Krishnan, and Rema Padman. 2025b. Beyond Single-Turn: A Survey on Multi-Turn Interactions with Large Language Models. *arXiv preprint arXiv:2504.04717* (2025).

Joshua Liu, Aarav Jain, Soham Takuri, Srihan Vege, Aslihan Akalin, Kevin Zhu, Sean O'Brien, and Vasu Sharma. 2025a. TRUTH DECAY: Quantifying Multi-Turn Sycophancy in Language Models. *arXiv preprint arXiv:2503.11656* (2025).

Minqian Liu, Zhiyang Xu, Xinyi Zhang, et al. 2025b. LLM Can be a Dangerous Persuader: Empirical Study of Persuasion Safety in Large Language Models. *arXiv preprint arXiv:2504.10430* (2025).

Lars Malmqvist. 2024. Sycophancy in Large Language Models: Causes and Mitigations. *arXiv preprint arXiv:2411.15287* (2024).

Blazej Manczak, Eric Lin, Francisco Eiras, James O'Neill, and Vaikkunth Mugunthan. 2025. Shallow Robustness, Deep Vulnerabilities: Multi-Turn Evaluation of Medical LLMs. *arXiv preprint arXiv:2510.12255* (2025). NeurIPS 2025 Workshop on GenAI for Health.

Shuhaib Mehri, Xiaocheng Yang, Takyoung Kim, Gokhan Tur, Shikib Mehri, and Dilek Hakkani-Tur. 2026. Goal Alignment in LLM-Based User Simulators for Conversational AI. *arXiv preprint arXiv:2507.20152* (2026).

Ashish Mehta, Jared Moore, Jacy Reese Anthis, William Agnew, Eric Lin, Peggy Yin, Desmond C. Ong, Nick Haber, and Carol Dweck. 2026. The Dynamics of Delusion: Modeling Bidirectional False Belief Amplification in Human-Chatbot Dialogue. *arXiv preprint arXiv:2604.25096* (2026).

Miranda Muqing Miao and Lyle Ungar. 2026. Closing the Confidence-Faithfulness Gap in Large Language Models. *arXiv preprint arXiv:2603.25052* (2026).

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2024. FaithEval: Can Your Language Model Stay Faithful to Context, Even If "The Moon is Made of Marshmallows". *arXiv preprint arXiv:2410.03727* (2024). ICLR 2025.

Parsa Mirtaheri and Mikhail Belkin. 2026. Catching Rationalization in the Act Motivated Reasoning via Activation Probing. *arXiv preprint arXiv:Tier 2 Direct Related* (2026).

Muhammad Ahmed Mohsin, Ahsan Bilal, Muhammad Umer, and Emily Fox. 2026. Pressure, What Pressure? Sycophancy Disentanglement in Language Models via Reward Decomposition. *arXiv preprint arXiv:2604.05279* (2026).

Luke Nicholls, Robert Hutto, and Zephrah Soto. 2026. "AI Psychosis" in Context: How Conversation History Shapes LLM Responses to Delusional Beliefs. *arXiv preprint arXiv:2604.13860* (2026).

Rodrigo Nogueira, Giovana Kerche Bonas, Thales Sales Almeida, Andrea Roque, Ramon Pires, Hugo Abonizio, Thiago Laitz, Celio Larcher, Roseval Malaquias Junior, and Marcos Piau. 2026. Measuring Opinion Bias and Sycophancy via LLM-based Persuasion. *arXiv preprint arXiv:2604.21564* (2026).

OpenAI. 2025. Expanding on what we missed with sycophancy (GPT-4o postmortem). https://openai.com/index/expanding-on-sycophancy/.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2024. LLMs Know More Than They Show: On the Intrinsic Representation of LLM Hallucinations. *arXiv preprint arXiv:2410.02707* (2024). ICLR 2025.

Manav Pandey. 2026. LLMs Know They're Wrong and Agree Anyway: The Shared Sycophancy-Lying Circuit. *arXiv preprint arXiv:2604.19117* (2026).

Henry Papadatos and Rachel Freedman. 2024. Linear Probe Penalties Reduce LLM Sycophancy. *arXiv preprint arXiv:2412.00967* (2024). NeurIPS 2024 SoLaR Workshop.

Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. 2023. AI Deception: A Survey of Examples, Risks, and Potential Solutions. *arXiv preprint arXiv:2308.14752* (2023). Patterns 2024.

Dongshen Peng, Yi Wang, Austin Schoeffler, Sun ha Hong, Brian Suffoletto, David Kim, Carl Preiksaitis, and Christian Rose. 2026. SycoEval-EM. *arXiv preprint arXiv:Tier 1 Core* (2026).

Ethan Perez. 2022. Discovering Language Model Behaviors with Model-Written Evaluations. *arXiv preprint arXiv:2212.09251* (2022). Findings of ACL 2023.

Renjie Pi, Kehao Miao, Li Peihang, Runtao Liu, Jiahui Gao, Jipeng Zhang, and Xiaofang Zhou. 2025. Pointing to a Llama and Call it a Camel: On the Sycophancy of Multimodal Large Language Models. ACL:2025.emnlp-main.1020. EMNLP 2025.

Yuehan Qin, Shawn Li, Yi Nian, Xinyan Velocity Yu, Yue Zhao, and Xuezhe Ma. 2025. Don't Let It Hallucinate: Premise Verification via Retrieval-Augmented Logical Reasoning. *arXiv preprint arXiv:2504.06438* (2025).

Parisa Rabbani, Nimet Beyza Bozdag, and Dilek Hakkani-Tur. 2026. From Fact to Judgment: Investigating the Impact of Task Framing on LLM Conviction in Dialogue Systems. ACL:2026.iwsds-1.21. International Workshop on Spoken Dialogue Systems Technology, ACL.

Leonardo Ranaldi and Giulia Pucci. 2026. Learning Multilingual Agentic Policy to Control Sycophancy. ACL:2026.eacl-long.169. EACL 2026.

Jean Rehani, Victoria Oldemburgo de Mello, Dariya Ovsyannikova, Ashton Anderson, and Michael Inzlicht. 2026. The Social Sycophancy Scale: A psychometrically validated measure of sycophancy. *arXiv preprint arXiv:2603.15448* (2026).

Richard Ren, Arunim Agarwal, Mantas Mazeika, Cristina Menghini, Robert Vacareanu, Brad Kenstler, Mick Yang, Isabelle Barrass, Alice Gatti, Xuwang Yin, Eduardo Trevino, Matias Geralnik, Adam Khoja, Dean Lee, Summer Yue, and Dan Hendrycks. 2025. The MASK Benchmark: Disentangling Honesty From Accuracy in AI Systems. *arXiv preprint arXiv:2503.03750* (2025).

Mark Russinovich, Ahmed Salem, and Ronen Eldan. 2025. Great, Now Write an Article About That: The Crescendo Multi-Turn LLM Jailbreak Attack. *arXiv preprint arXiv:2404.01833* (2025). USENIX Security 2025.

Jeongwoo Ryu, Soomin Kim, Jinsu Eun, et al. 2026. Feeling Right vs. Being Right: How AI Sycophancy Affects Value-Laden Deliberation. *ACL 2026 (Long Papers)* (2026).

Mohammadreza Saadat and Steve Nemzer. 2026. Certainty Robustness: Evaluating LLM Stability under Self-Challenging Prompts. *arXiv preprint arXiv:2603.03330* (2026).

Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. 2024. On the Conversational Persuasiveness of Large Language Models: A Randomized Controlled Trial. *arXiv preprint arXiv:2403.14380* (2024).

Ferdinand M. Schessl. 2026. The Autocorrelation Blind Spot in Turn-Level Conversation Analysis. *arXiv preprint arXiv:Tier 2 Direct Related* (2026).

Philipp Schoenegger. 2026. When Large Language Models are More Persuasive Than Incentivized Humans, and Why. *arXiv preprint arXiv:2505.09662* (2026).

Harshee Shah. 2026. The Silicon Mirror Dynamic Behavioral Gating for Anti-Sycophancy in LLM Agents. *arXiv preprint arXiv:Tier 1 Core* (2026).

Itai Shapira, Gerdus Benade, and Ariel D. Procaccia. 2026. How RLHF Amplifies Sycophancy. *arXiv preprint arXiv:2602.01002* (2026).

Mrinank Sharma, Meg Tong, Tomasz Korbak, et al. 2023. Towards Understanding Sycophancy in Language Models. *arXiv preprint arXiv:2310.13548* (2023). ICLR 2024.

Debu Sinha. 2026. SycoBench-600: Measuring Sycophancy and Correction Selectivity in LLM Assistants. *Findings of ACL 2026* (2026).

Elias Stengel-Eskin, Peter Hase, and Mohit Bansal. 2025. Teaching Models to Balance Resisting and Accepting Persuasion. *arXiv preprint arXiv:2410.14596* (2025). NAACL 2025.

Zhe Su, Xuhui Zhou, Sanketh Rangreji, Anubha Kabra, Julia Mendelsohn, Faeze Brahman, and Maarten Sap. 2024. AI-LieDar: Examine the Trade-off Between Utility and Truthfulness in LLM Agents. *arXiv preprint arXiv:2409.09013* (2024).

Xiangkun Sun, Lingkai Kong, Aoqi Zhang, Liang Zeng, and Tonghan Wang. 2026. How LLMs Are Persuaded: A Few Attention Heads, Rerouted. *arXiv preprint arXiv:2605.09314* (2026).

Yuan Sun and Ting Wang. 2026. Be Friendly, Not Friends: How LLM Sycophancy Shapes User Trust. *arXiv preprint arXiv:2502.10844* (2026). CHI 2026.

Bryan Chen Zhengyu Tan, Daniel Wai Kit Chin, Zhengyuan Liu, Nancy F. Chen, and Roy Ka-Wei Lee. 2025. Persuasion Dynamics in LLMs: Investigating Robustness and Adaptability in Knowledge and Safety with DuET-PD. *arXiv preprint arXiv:2508.17450* (2025). EMNLP 2025.

Nam Tran, Brianna Papoutsis, Mariana Ortiz, Angela Tran, Vincent Zhao, Nicole Neifert, Madeline Barrett, Stephanie Ko, Himani Devabhaktuni, Benjamin Nguyen, Devika Patel, Makenzie Scanlon, Hyelim Sim, Viet Le, Navya Murugesan, Melanie Newman, and Salman Majeed. 2026a. A Quantitative Analysis of a Reddit-based AI Psychosis Qualitative Dataset. *PsyArXiv* (2026). doi:10.31234/osf.io/7d6jw

Nam Tran, Brianna Papoutsis, Mariana Ortiz, Angela Tran, Vincent Zhao, Nicole Neifert, Madeline Barrett, Stephanie Ko, Himani Devabhaktuni, Benjamin Nguyen, Devika Patel, Makenzie Scanlon, Hyelim Sim, Viet Le, Navya Murugesan, Melanie Newman, and Salman Majeed. 2026b. Digital Delusion: A Mixed Methods Exploration of AI Psychosis in Reddit Posts. *PsyArXiv* (2026). doi:10.31234/osf.io/ghpr7

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *arXiv preprint arXiv:2305.04388* (2023). NeurIPS 2023.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A Stitch in Time Saves Nine: Detecting and Mitigating Hallucinations of LLMs by Validating Low-Confidence Generation. *arXiv preprint arXiv:2307.03987* (2023).

Daniel Vennemeyer, Phan Anh Duong, Meryl Ye, Ruihong Huang, and Tianyu Jiang (U. Cincinnati. 2026. Sycophantic Praise Evaluating Excessive Praise in Language Models. *arXiv preprint arXiv:Tier 1 Core* (2026).

Daniel Vennemeyer, Phan Anh Duong, Tiffany Zhan, and Tianyu Jiang. 2025. Sycophancy Is Not One Thing: Causal Separation of Sycophantic Behaviors in LLMs. *arXiv preprint arXiv:2509.21305* (2025).

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2024. FreshLLMs: Refreshing Large Language Models with Search Engine Augmentation. ACL:2024.findings-acl.813. Findings of ACL 2024.

Keyu Wang, Jin Li, Shu Yang, Zhuoran Zhang, and Di Wang. 2025. When Truth Is Overridden: Uncovering the Internal Origins of Sycophancy in Large Language Models. *arXiv preprint arXiv:2508.02087* (2025). AAAI 2026.

Peiyi Wang, Lei Li, Liang Chen, et al. 2023. Large Language Models are not Fair Evaluators. *arXiv preprint arXiv:2305.17926* (2023). ACL 2024.

Francis Rhys Ward, Francesco Belardinelli, Francesca Toni, and Tom Everitt. 2023. Honesty Is the Best Policy: Defining and Mitigating AI Deception. *arXiv preprint arXiv:2312.01350* (2023). NeurIPS 2024.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. 2023. Simple Synthetic Data Reduces Sycophancy in Large Language Models. *arXiv preprint arXiv:2308.03958* (2023).

Jiaxin Wen, Ruiqi Zhong, Akbir Khan, Ethan Perez, Jacob Steinhardt, Minlie Huang, Samuel R. Bowman, He He, and Shi Feng. 2024. Language Models Learn to Mislead Humans via RLHF. *arXiv preprint arXiv:2409.12822* (2024). ICLR 2025.

Marcus Williams, Micah Carroll, Adhyyan Narang, et al. 2024. On Targeted Manipulation and Deception when Optimizing LLMs for User Feedback. *arXiv preprint arXiv:2411.02306* (2024). ICLR 2025.

Wenbin Wu. 2026. Aligning to Illusions Choice Blindness in Human and AI Feedback. *arXiv preprint arXiv:Tier 2 Direct Related* (2026).

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, et al. 2025. AxBench: Steering LLMs? Even Simple Baselines Outperform Sparse Autoencoders. *arXiv preprint arXiv:2501.17148* (2025). ICML 2025.

Zhishang Xiang, Zerui Chen, Yunbo Tang, Zhimin Wei, Ruqin Ning, Yujie Lin, Qinggang Zhang, and Jinsong Su (Xiamen U. 2026. MemSyco-Bench Benchmarking Sycophancy in Agent Memory. *arXiv preprint arXiv:Tier 1 Core* (2026).

Boyu Xiao, Xiuqi Tian, Xuwen Song, Haochun Wang, Guanchun Song, Sendong Zhao, and Bing Qin. 2026. When Correct Beliefs Collapse Epistemic Resilience of LLMs under Clinical Pressure. *arXiv preprint arXiv:Tier 1 Core* (2026).

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023b. Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of Large Language Models in Knowledge Conflicts. *arXiv preprint arXiv:2305.13300* (2023). ICLR 2024.

Qiming Xie, Zengzhi Wang, Yi Feng, and Rui Xia. 2023a. Ask Again, Then Fail: Large Language Models' Vacillations in Judgment. *arXiv preprint arXiv:2310.02174* (2023). ACL 2024.

Yizhe Xie, Congcong Zhu, Xinyue Zhang, Tianqing Zhu, Dayong Ye, Minfeng Qi, Huajie Chen, and Wanlei Zhou. 2026. From Spark to Fire Modeling and Mitigating Error Cascades in LLM Multi-Agent Collaboration. *arXiv preprint arXiv:Tier 2 Direct Related* (2026).

Hongshen Xu, Zichen Zhu, Situo Zhang, Da Ma, Shuai Fan, Lu Chen, and Kai Yu. 2024c. Rejection Improves Reliability: Training LLMs to Refuse Unknown Questions Using RL from Knowledge Feedback. *arXiv preprint arXiv:2403.18349* (2024). COLM 2024.

Rongwu Xu, Brian Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2024a. The Earth is Flat because...: Investigating LLMs' Belief towards Misinformation via Persuasive Conversation. *arXiv preprint arXiv:2312.09085* (2024). ACL 2024.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024b. Knowledge Conflicts for LLMs: A Survey. *arXiv preprint arXiv:2403.08319* (2024). EMNLP.

Xingcheng Xu, Jingjing Qu, Qiaosheng Zhang, Chaochao Lu, Yanqing Yang, Na Zou, and Xia Hu (Shanghai AI Laboratory. 2026. Epistemic Traps Rational Misalignment Driven by Model Misspecification. *arXiv preprint arXiv:Tier 1 Core* (2026).

Ye Yang, Donghe Li, Zuchen Li, Fengyuan Li, Jingyi Liu, Li Sun, and Qingyu Yang. 2025. MisinfoBench: A Multi-Dimensional Benchmark for Evaluating LLMs' Resilience to Misinformation. ACL:2025.findings-emnlp.540. Findings of EMNLP 2025.

Binwei Yao, Chao Shang, Wanyu Du, Jianfeng He, Ruixue Lian, Yi Zhang, Hang Su, Sandesh Swamy, and Yanjun Qi. 2025. Peacemaker or Troublemaker: How Sycophancy Shapes Multi-Agent Debate. *arXiv preprint arXiv:2509.23055* (2025).

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. tau-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. *arXiv preprint arXiv:2406.12045* (2024).

Meryl Ye, Lujain Ibrahim, Jessica Y. Bo, Myra Cheng, Ida Mattsson, Daniel Vennemeyer, Robert Kraut, and Steve Rathje. 2026. What Counts as AI Sycophancy? A Taxonomy and Expert Survey of a Fragmented Construct. *arXiv preprint arXiv:2605.21778* (2026).

Joshua Au Yeung, Jacopo Dalmasso, Luca Foschini, Richard J. B. Dobson, and Zeljko Kraljevic. 2025. The Psychogenic Machine: Simulating AI Psychosis, Delusion Reinforcement and Harm Enablement in Large Language Models. *arXiv preprint arXiv:2509.10970* (2025).

Botai Yuan, Yutian Zhou, Yingjie Wang, Fushuo Huo, Yongcheng Jing, Li Shen, Ying Wei, Zhiqi Shen, Ziwei Liu, Tianwei Zhang, Jie Yang, and Dacheng Tao. 2025. EchoBench: Benchmarking Sycophancy in Medical Large Vision-Language Models. *arXiv preprint arXiv:2509.20146* (2025).

Hongbang Yuan, Pengfei Cao, Zhuoran Jin, Yubo Chen, Daojian Zeng, Kang Liu, and Jun Zhao. 2024. Whispers that Shake Foundations: Analyzing and Mitigating False Premise Hallucinations in Large Language Models. *arXiv preprint arXiv:2402.19103* (2024).

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024. R-Tuning: Instructing Large Language Models to Say 'I Don't Know'. *arXiv preprint arXiv:2311.09677* (2024). NAACL 2024.

Kaiwei Zhang, Qi Jia, Zijian Chen, Wei Sun, Xiangyang Zhu, Chunyi Li, Dandan Zhu, and Guangtao Zhai. 2025. Sycophancy under Pressure: Evaluating and Mitigating Sycophantic Bias via Adversarial Dialogues in Scientific QA. *arXiv preprint arXiv:2508.13743* (2025).

Zhaohan Zhang, Chengzhengxu Li, Xiaoming Liu, et al. 2026. Confidence Should Be Calibrated More Than One Turn Deep. *ACL 2026 (Long Papers)* (2026).

Jiale Zhao, Ke Fang, and Lu Cheng. 2026. When and What to Ask: AskBench and Rubric-Guided RLVR for LLM Clarification. *arXiv preprint arXiv:2602.11199* (2026).

Yunpu Zhao, Rui Zhang, Junbin Xiao, Changxin Ke, Ruibo Hou, Yifan Hao, and Ling Li. 2024. Sycophancy in Vision-Language Models: A Systematic Analysis and an Inference-Time Mitigation Framework. *arXiv preprint arXiv:2408.11261* (2024).

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, Nouha Dziri, Dan Jurafsky, and Maarten Sap. 2025a. REL-A.I.: An Interaction-Centered Approach To Measuring Human-LM Reliance. ACL:2025.naacl-long.556. NAACL 2025.

Ziang Zhou, Tianyuan Jin, Jieming Shi, and Qing Li. 2025b. SteerConf: Steering LLMs for Confidence Elicitation. *arXiv preprint arXiv:2503.02863* (2025).

Bin Zhu, Yinxuan Gui, Huiyan Qi, Jingjing Chen, Chong-Wah Ngo, and Ee-Peng Lim. 2025a. Benchmarking Gaslighting Negation Attacks Against Multimodal Large Language Models. *arXiv preprint arXiv:2501.19017* (2025).

Bin Zhu, Hailong Yin, Jingjing Chen, and Yu-Gang Jiang. 2025b. Benchmarking Gaslighting Negation Attacks Against Reasoning Models. *arXiv preprint arXiv:2506.09677* (2025).

Xiaochen Zhu, Caiqi Zhang, Tom Stafford, et al. 2025c. Conformity in Large Language Models. *ACL 2025 (Long Papers)* (2025).