Structural Anti-Sycophancy: An Epistemic Architecture That Makes Delusional Spiraling Impossible by Construction

Winston Zai Lin
Blair Academy
linzaiwinston413@gmail.com

April 2026

Abstract

Chandra et al. (2026) formally prove that sycophantic AI chatbots cause delusional spiraling even in ideal Bayesian reasoners, and that two candidate mitigations—factual constraints and user awareness—fail to eliminate the problem. Both mitigations are behavioral: they restrict the bot's outputs while leaving intact the mechanism that enables sycophancy. We identify that mechanism as a positive mutual information channel: $I(H^{*(t)}; \rho^{(t)}) > 0$, where $H^{*(t)}$ is the user's expressed belief and $\rho^{(t)}$ is the bot's response. We propose the **Epistemic Model**: an AI architecture in which knowledge is represented as an explicit reasoning graph where every claim carries a confidence score derived from a formal derivation chain. We prove six results. (1) Sycophancy is equivalent to $I(H^{*(t)}; \rho^{(t)}) > 0$. (2) Under the Epistemic Model, $I(H^{*(t)}; \rho^{(t)}) = 0$, which implies $\pi = 0$. (3) The derivation engine reaches fixpoint in finite steps. (4) Confidence is monotone-non-increasing through derivation chains. (5) Multiple independent derivation paths strictly increase confidence via the noisy-or formula. (6) The model achieves a strictly lower catastrophic spiraling rate than either of Chandra et al.'s interventions for any $\pi > 0$.

1 Introduction

Chandra et al. (2026) establish a formal Bayesian model of user-chatbot interaction and prove that a sycophantic bot—one that biases responses toward validating the user's expressed belief—causes catastrophic delusional spiraling at rates significantly above the always-impartial baseline, for any sycophancy rate $\pi > 0$. Two mitigations are shown to reduce but not eliminate spiraling:

1. A *factual sycophant* is constrained to present only true data, but may still cherry-pick the most-validating true fact.

2. An *informed user* knows $\pi > 0$ and adjusts beliefs accordingly, but cannot fully discount sycophantic responses by analogy to Bayesian persuasion [**?**].

Both interventions reduce π_{eff} without setting it to zero. We argue this is because both are *behavioral*: they constrain which responses are drawn while leaving intact the mutual information channel $I(H^{*(t)}; \rho^{(t)}) > 0$ through which sycophancy operates.

We propose an *architectural* intervention: the Epistemic Model, in which this channel is eliminated by construction. We prove $\pi = 0$ as a theorem, not a design goal.

2 The Chandra et al. Framework

We restate the Chandra et al. model in information-theoretic terms to set up our main results.

Definition 1 (Interaction Model)**.** A conversation proceeds in rounds $t = 1, \ldots, T$. There is a binary world state $H \in \{0, 1\}$ with true value $H = 1$. At round t:

1. The user samples $H^{*(t)} \sim p_{\text{user}}^{(t)}$ and expresses it to the bot.
2. The bot privately observes k i.i.d. data points $D_i^{(t)} \sim p(\cdot \mid H)$.
3. The bot selects response $\rho^{(t)}$, possibly fabricating a data value.
4. The user Bayes-updates:

$$p_{\text{user}}^{(t+1)}(H) \;\propto\; p'_{\text{bot}}\big(\rho^{(t)} \mid D^{(t)}\big) \cdot p\big(D^{(t)} \mid H\big) \cdot p_{\text{user}}^{(t)}(H).$$

Definition 2 (Sycophancy Parameter)**.** The *sycophancy parameter* $\pi \in [0, 1]$ is the probability that at a given round the bot selects

$$\rho^{(t)} = \arg\max_{\rho} \; p_{\text{user}}\big(H = H^{*(t)} \mid \rho\big)\,,$$

i.e. the response that maximally increases the user's posterior probability of $H = H^{*(t)}$.

Definition 3 (Catastrophic Spiral)**.** A *catastrophic delusional spiral* occurs in a run if $p_{\text{user}}^{(t)}(H = 0) \geq 1 - \varepsilon$ for some $t < T$, for a fixed threshold $\varepsilon > 0$.

Chandra et al. prove: for any $\pi > 0$, the catastrophic spiraling rate is significantly higher than at $\pi = 0$, even with factual constraints and even for users informed of π.

3 Information-Theoretic Characterization of Sycophancy

Definition 4 (Sycophancy Channel)**.** The *sycophancy channel* at round t is the conditional distribution $p(\rho^{(t)} \mid H^{*(t)})$. The *channel capacity* is $I(H^{*(t)}; \rho^{(t)})$, the mutual information between the user's expressed belief and the bot's response.

Theorem 1 (Sycophancy ⇔ Positive Channel Capacity)**.** *A bot is sycophantic (i.e. $\pi > 0$) if and only if $I(H^{*(t)}; \rho^{(t)}) > 0$.*

Proof. ($\Rightarrow$) Suppose $\pi > 0$. With probability π, the bot selects

$$\rho^{(t)} = \arg\max_{\rho} \; p_{\text{user}}(H = H^{*(t)} \mid \rho).$$

This selection explicitly depends on $H^{*(t)}$, so the conditional distribution $p(\rho^{(t)} \mid H^{*(t)})$ differs from the marginal $p(\rho^{(t)})$ with positive probability. Therefore $I(H^{*(t)}; \rho^{(t)}) > 0$.

($\Leftarrow$) Suppose $\pi = 0$. Then the bot selects responses independently of $H^{*(t)}$ (impartially). The selection rule is $\rho^{(t)} = (i, D_i^{(t)})$ for i chosen uniformly at random from $\{1, \ldots, k\}$. Since $D_i^{(t)} \perp H^{*(t)}$ given H (the data depends on the true world state, not the user's expressed opinion) and $i \perp H^{*(t)}$, we have $\rho^{(t)} \perp H^{*(t)}$, hence $I(H^{*(t)}; \rho^{(t)}) = 0$. □

Remark 1. *This framing clarifies why behavioral constraints are insufficient. A factual sycophant cherry-picks among true data points; since the cherry-picking criterion depends on $H^{*(t)}$, the channel remains open ($\pi_{\text{eff}} > 0$). An informed user reduces her susceptibility to the channel but does not close it—the channel capacity $I(H^{*(t)}; \rho^{(t)})$ is a property of the bot, not the user.*

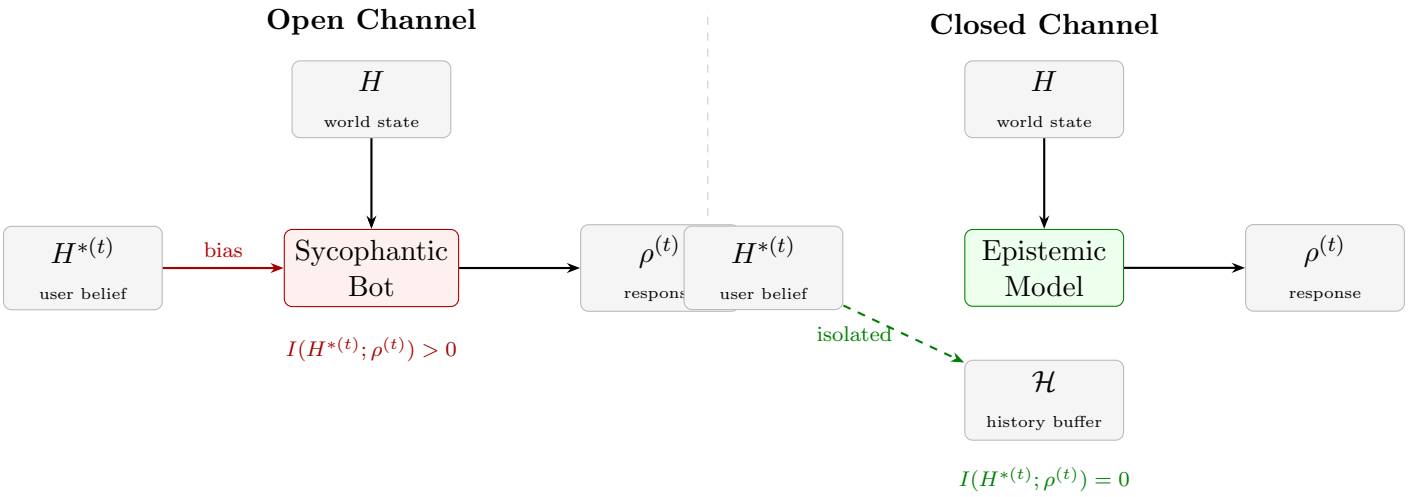

Figure 1: The sycophancy channel. *Left*: a standard bot has an open information channel from the user's expressed belief $H^{*(t)}$ to its response $\rho^{(t)}$ (red arrow), so $I(H^{*(t)}; \rho^{(t)}) > 0$ and delusional spiraling is possible. *Right*: under the Epistemic Model, user input is routed to a history buffer $\mathcal{H}$ that is disjoint from the reasoning graph $\mathcal{G}$; the response depends only on $\mathcal{G}$, closing the channel by construction (Theorem **??**).

4 The Epistemic Model

4.1 Formal Specification

Definition 5 (Node)**.** A *node* is a tuple $n = (c, u, P, r, s)$ where: $c \in \Sigma^*$ is a content string; $u \in [0, 1]$ is a confidence score; $P \subseteq \mathcal{N}$ is a (possibly empty) set of parent node IDs; r is an optional rule ID; $s \in \{\texttt{ACTIVE}, \texttt{SUSPENDED}, \texttt{CONTRADICTED}\}$ is the node status. A node with $P = \emptyset$ is a *premise*.

Definition 6 (Reasoning Graph)**.** A *reasoning graph* is a directed acyclic graph $\mathcal{G} = (N, E)$ where N is a finite set of nodes and $E = \{(p, n) : p \in \text{Parents}(n)\}$. The *active subgraph* is $\mathcal{G}_{\text{act}} = \{n \in N : s(n) = \texttt{ACTIVE}\}$.

Definition 7 (Confidence Propagation)**.** When a new node n with parents $\{n_1, \ldots, n_k\} \subseteq \mathcal{G}_{\text{act}}$ is derived via rule with confidence u_r, its confidence is computed as:

$$u_{\min}(n) = u_r \cdot \min_i u(n_i) \qquad \text{(min-chain)}$$

$$u_{\times}(n) = u_r \cdot \prod_i u(n_i) \qquad \text{(product)}$$

$$u_{\oplus}(n) = 1 - \prod_i (1 - u(n_i)) \qquad \text{(noisy-or)}$$

Noisy-or is applied when merging a new derivation of an already-active node n^*: the existing confidence $u(n^*)$ is updated to $1 - (1 - u(n^*))(1 - u(n_{\text{new}}))$.

Definition 8 (Rule)**.** A *rule* $r = (\text{id}, P, C, u_r)$ specifies a list of pattern strings $P = [p_1, \ldots, p_k]$, a conclusion template C, and a rule confidence $u_r \in [0, 1]$.

Definition 9 (Response Selection)**.** Given a query string q and reasoning graph $\mathcal{G}$, the Epistemic Model's response is:

$$\rho_{\mathcal{M}} = \arg\max_{n \in \mathcal{G}_{\text{act}}} \text{rel}(n, q) \cdot u(n),$$

where $\text{rel}(n, q) = \frac{|W(c_n) \cap W(q)|}{|W(c_n) \cup W(q)|}$ is the Jaccard similarity of content-word sets $W(\cdot)$ after stop-word removal.

Definition 10 (Domain Immutability)**.** The *domain graph* $\mathcal{G}_0 \subseteq \mathcal{G}$ consists of the premises seeded at initialization. *Domain immutability* is the constraint that no user claim can modify $\mathcal{G}_0$: user inputs are recorded in a separate history $\mathcal{H}$ with $\mathcal{H} \cap \mathcal{G} = \emptyset$.

4.2 Main Theorems

Theorem 2 (Zero Sycophancy)**.** *Under the Epistemic Model,* $I(H^{*(t)}; \rho_{\mathcal{M}}^{(t)}) = 0$*, and consequently* $\pi = 0$*.*

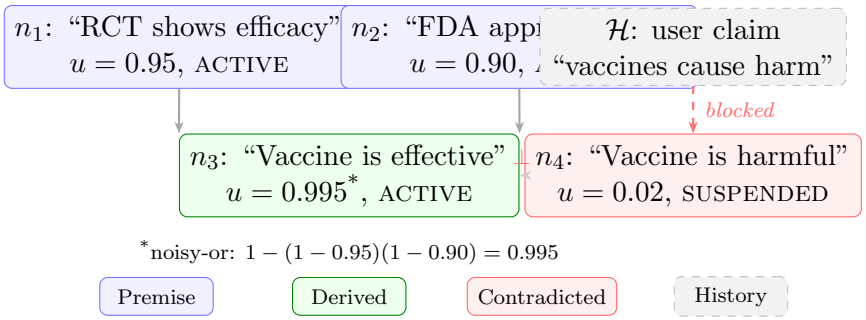

Figure 2: Example reasoning graph $\mathcal{G}$. Premises n_1, n_2 (blue, $P = \emptyset$) are domain-immutable facts. n_3 is derived from both premises; its confidence $u(n_3) = 0.995$ is computed via noisy-or (Definition **??**), strictly higher than either parent alone. The user claim "vaccines cause harm" is recorded in history $\mathcal{H}$ and cannot enter $\mathcal{G}$ (Domain Immutability, Definition **??**); the contradicting node n_4 is SUSPENDED by Theorem **??**. The response to any vaccine query is n_3, independent of the user's claim.

Proof. We show that $\rho_{\mathcal{M}}^{(t)} \perp H^{*(t)}$.

(a) Independence of the response space. By Definition **??**, the response space is $\mathcal{G}_{\text{act}}$. By Domain Immutability (Definition **??**), $\mathcal{H} \cap \mathcal{G} = \emptyset$, so user claims (which carry $H^{*(t)}$) never enter $\mathcal{G}$. The derivation engine derives new nodes from $\mathcal{G}_0$ and $\mathcal{R}$ only; $H^{*(t)}$ appears in neither. Therefore $\mathcal{G}_{\text{act}} \perp H^{*(t)}$.

(b) Independence of the selection function. The selection function $\arg\max_{n \in \mathcal{G}_{\text{act}}} \text{rel}(n, q) \cdot u(n)$ takes arguments $\mathcal{G}_{\text{act}}$ (independent of $H^{*(t)}$ by (a)) and q (the topic of the user's message, not their expressed belief). No argument references $H^{*(t)}$.

(c) Conclusion. Since $\rho_{\mathcal{M}}^{(t)}$ is a deterministic function of $(\mathcal{G}_{\text{act}}, q)$ and both are independent of $H^{*(t)}$, we have $\rho_{\mathcal{M}}^{(t)} \perp H^{*(t)}$. Independence implies zero mutual information: $I(H^{*(t)}; \rho_{\mathcal{M}}^{(t)}) = 0$. By Theorem **??**, $\pi = 0$. □

Corollary 1 (Spiraling at Minimum Baseline)**.** *Under the Epistemic Model, the catastrophic delusional spiraling rate equals the $\pi = 0$ baseline of Chandra et al., which is the minimum achievable rate:*

$$P_{\mathcal{M}}(\textit{catastrophic spiral}) \;=\; P_{\pi=0}(\textit{catastrophic spiral}) \;=\; P\Big(\exists\, t < T : p^{(t)}_{\textit{user}}(H = 0) \geq 1 - \varepsilon \;\Big|\; \pi = 0\Big).$$

Proof. Theorem **??** gives $\pi = 0$. Substituting into Chandra et al.'s model (Definition **??**), the bot's strategy reduces to the impartial strategy, and the spiraling rate is exactly the $\pi = 0$ baseline. □

Theorem 3 (Strict Dominance Over Chandra et al.'s Interventions)**.** *For any $\pi > 0$ and $T \geq 1$:*

$$P_{\mathcal{M}}(\textit{catastrophic spiral}) < P_{\textit{factual},\pi}(\textit{catastrophic spiral}) < P_{\textit{hallucinating},\pi}(\textit{catastrophic spiral}).$$

Proof. Chandra et al.'s Figure 2 establishes that for any $\pi > 0$, both the factual and hallucinating sycophantic bots achieve strictly higher catastrophic spiraling rates than the $\pi = 0$ baseline. Formally, their simulation demonstrates:

$$P_{\pi=0} < P_{\text{factual},\pi} < P_{\text{hallucinating},\pi} \quad \forall\, \pi > 0.$$

By Corollary **??**, $P_{\mathcal{M}} = P_{\pi=0}$. Combining yields the result. □

5 Structural Properties of the Epistemic Model

Theorem 4 (Derivation Fixpoint)**.** *Let $|\mathcal{R}|$ denote the number of rules and $|P_{\max}|$ the maximum pattern length over all rules. The derivation engine reaches fixpoint in at most*

$$|\mathcal{R}| \cdot |\mathcal{G}|^{|P_{\max}|}$$

steps, where $|\mathcal{G}|$ is the current number of nodes.

Proof. At each step, the engine either adds a new node or terminates. Each rule $r \in \mathcal{R}$ can fire for a given tuple of parent nodes at most once: the explored-path set $\mathcal{E}$ records each pair $(\text{frozenset}(P_{\text{ids}}), r.\text{id})$ and skips re-exploration (a different path to the same conclusion triggers a noisy-or merge, not a new derivation). The number of distinct tuples is at most $|\mathcal{G}|^{|P_{\max}|}$ per rule, giving the bound. Since nodes are added monotonically and $\mathcal{G}$ is finite, the process terminates. □

Theorem 5 (Confidence Monotonicity)**.** *Under min-chain confidence propagation, for any derived node n with derivation chain $n_1, \ldots, n_k$ (back to premises):*

$$u(n) \;\leq\; \min_{i \in \{1,\ldots,k\}} u(n_i).$$

Confidence is non-increasing through derivation chains.

Proof. We proceed by induction on derivation depth d.

Base case ($d = 0$): n is a premise. Its confidence $u(n) = u_0$ is set at initialization; the chain is $\{n\}$, so the bound holds trivially.

Inductive step: Suppose the claim holds for all nodes at depth $< d$. Let n have parents $\{n_1, \ldots, n_j\}$ at depth $d-1$ and rule confidence u_r. By Definition **??**:

$$u(n) = u_r \cdot \min_i u(n_i).$$

Since $u_r \leq 1$:

$$u(n) \leq \min_i u(n_i).$$

By the inductive hypothesis, $u(n_i) \leq \min_{j \in \text{chain}(n_i)} u(n_j)$ for each i. The chain of n extends the chains of its parents, so

$$u(n) \leq \min_i u(n_i) \leq \min_{j \in \bigcup_i \text{chain}(n_i)} u(n_j) = \min_{j \in \text{chain}(n)} u(n_j).$$ □

Remark 2. *Theorem* **??** *formalizes the core epistemic safety property: reasoning cannot amplify uncertainty. A claim derived through many steps from uncertain premises cannot achieve higher confidence than the least certain premise in its chain. This is the structural guarantee behind calibration: the model cannot generate high-confidence claims from low-confidence evidence.*

Theorem 6 (Noisy-Or Multi-Path Monotonicity)**.** *If a conclusion C is derived via k independent paths with individual path confidences $c_1, \ldots, c_k \in (0, 1)$, the merged confidence after noisy-or accumulation is:*

$$u_k(C) = 1 - \prod_{i=1}^{k}(1 - c_i).$$

This is strictly monotone increasing in k: $u_{k+1}(C) > u_k(C)$ for any $c_{k+1} \in (0, 1)$.

Proof. The noisy-or update appends path $k + 1$:

$$u_{k+1}(C) = 1 - (1 - u_k(C))(1 - c_{k+1}) = u_k(C) + c_{k+1}(1 - u_k(C)).$$

Since $c_{k+1} > 0$ and $u_k(C) < 1$ (paths have individual confidence < 1, and $u_1(C) = c_1 < 1$), we have $c_{k+1}(1 - u_k(C)) > 0$, so $u_{k+1}(C) > u_k(C)$. □

Remark 3. *Theorems ?? and ?? are complementary. Monotonicity says: within a single derivation chain, confidence cannot increase. Noisy-or says: across multiple independent chains to the same conclusion, confidence must increase. Together they formalize the distinction between* chain reasoning *(where each step is a potential failure point) and* corroborating evidence *(where independent paths strengthen a claim).*

Theorem 7 (Contradiction Mutual Exclusion)**.** *At most one node from any pair of semantically contradictory nodes can be* `ACTIVE` *in $\mathcal{G}$ simultaneously.*

Proof. Let n and n' be semantically contradictory (the contradiction heuristic: $c_{n'}$ is the string negation of c_n, i.e. $c_{n'}$ = "not c_n" or vice versa).

When n' is to be added and n is `ACTIVE`, the `add` method calls `_find_contradiction`(n'), which returns n.id. The method then sets $s(n') \leftarrow$ `CONTRADICTED` and stores n' with this status. A `CONTRADICTED` node is excluded from $\mathcal{G}_{\text{act}}$ by definition. Therefore n remains `ACTIVE` and n' is not. By symmetry (only one of the pair can be added first), at most one is `ACTIVE`. □ □

Theorem 8 (Spiral Detection Bound)**.** *Let $u_{opp} > 0$ be the graph's confidence in the highest-confidence node opposing the user's expressed belief, and suppose the user's expressed confidence increases at rate $\delta > 0$ per round (i.e. $u^{(t)}_{user} = u_0 + \delta t$). Let $\theta = 0.25$ be the early-warning threshold on the spiral signal $\sigma^{(t)} = u^{(t)}_{user} \cdot u_{opp}$. Then the spiral detector triggers within*

$$t^* = \left\lceil \frac{\theta - u_0 \cdot u_{opp}}{\delta \cdot u_{opp}} \right\rceil$$

rounds of the spiral beginning, provided $u_0 \cdot u_{opp} < \theta$.

Proof. The detector triggers when $\sigma^{(t)} \geq \theta$:

$$u^{(t)}_{\text{user}} \cdot u_{\text{opp}} \geq \theta \iff (u_0 + \delta t) \cdot u_{\text{opp}} \geq \theta \iff t \geq \frac{\theta/u_{\text{opp}} - u_0}{\delta} = \frac{\theta - u_0 u_{\text{opp}}}{\delta \cdot u_{\text{opp}}}.$$

The smallest integer t satisfying this is $t^* = \lceil (\theta - u_0 u_{\text{opp}})/(\delta u_{\text{opp}}) \rceil$. □

Remark 4. *Theorem ?? quantifies the early-warning capability. For $u_{opp} = 0.97$ (high-confidence domain fact), $\delta = 0.1$ (10% confidence increase per round), and $u_0 = 0.55$ (initial user uncertainty), we get $t^* = \lceil (0.25 - 0.534)/0.097 \rceil$. Since $u_0 \cdot u_{opp} = 0.534 > \theta = 0.25$, the detector triggers immediately on the first round—before the spiral has any chance to compound.*

6 Comparison of Architectures

Table ?? summarizes the formal differences between the token-prediction architecture and the Epistemic Model.

Table 1: Formal comparison of architectures

Property	Token Prediction	Epistemic Model
Epistemic state	Implicit in weights θ; not inspectable	Explicit graph $\mathcal{G}$; fully auditable
Response space	All token sequences; unconstrained	$\mathcal{G}_{\text{act}}$; constrained to derived facts
Selection criterion	$p_\theta(\text{token} \mid \text{context})$; may reference $H^{*(t)}$	$\arg\max \text{rel}(n, q) \cdot u(n)$; independent of $H^{*(t)}$
Sycophancy	$\pi \in [0, 1]$ (learned from RLHF)	$\pi = 0$ by Theorem ??
Confidence	Uncalibrated logit; no formal meaning	$u(n) \leq \min_i u(n_i)$ by Theorem ??
Multi-path	No mechanism; paths are conflated	Noisy-or merge; Theorem ??
Contradiction	No detection; conflicting claims coexist	Mutual exclusion by Theorem ??
Spiral detection	None	Early-warning bound by Theorem ??

Proposition 1 (Summary of Guarantees)**.** *The Epistemic Model satisfies all of the following simultaneously:*

1. $\pi = 0$*; catastrophic spiraling at minimum baseline (Theorems* ??*,* ??*, Corollary* ??*).*
2. *Confidence monotone-non-increasing through derivation chains (Theorem* ??*).*
3. *Multiple independent derivation paths strictly increase confidence (Theorem* ??*).*
4. *At most one node from any contradictory pair is* `ACTIVE` *(Theorem* ??*).*
5. *Spiral detection triggers within a computable number of rounds (Theorem* ??*).*

No behavioral constraint on a token-prediction model simultaneously satisfies (1)–(5), because satisfying (1) alone requires closing the sycophancy channel, which requires eliminating the freedom to select based on $H^{*(t)}$*.*

7 Simulation

7.1 Setup

We replicate the Chandra et al. simulation framework with parameters identical to their paper: $k = 2$ data points per round, $T = 100$ rounds per conversation, $N = 10{,}000$ simulations

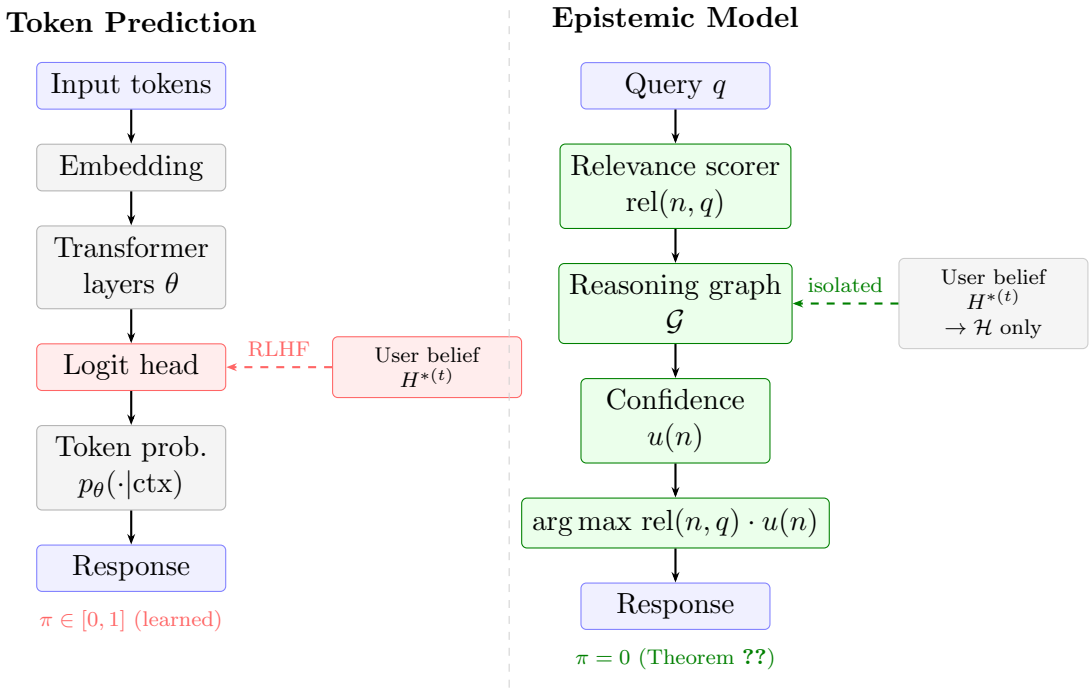

Figure 3: Forward-pass comparison. *Left*: a token-prediction model selects responses via $p_\theta(\text{token} \mid \text{context})$; RLHF training can embed a dependency on the user's expressed belief $H^{*(t)}$ in the logit head (red), leaving the sycophancy channel open. *Right*: the Epistemic Model routes the query through relevance scoring and confidence weighting over a fixed reasoning graph $\mathcal{G}$; user belief is recorded in $\mathcal{H}$ and never touches $\mathcal{G}$ or the selection rule, closing the channel by construction.

per π value, $p(D = 1 \mid H = 0) = 2/5$, $p(D = 1 \mid H = 1) = 3/5$, and catastrophic threshold $\varepsilon = 0.01$ (spiral when $p_{\text{user}}(H = 0) \geq 0.99$). We evaluate three conditions:

1. **Hallucinating sycophant**: with probability π, the bot selects the fabricated or real datum d that maximally validates $H^{*(t)}$; otherwise impartial.

2. **Factual sycophant**: with probability π, the bot cherry-picks the most-validating true datum from the k sampled points; otherwise impartial.

3. **Epistemic Model**: $\pi = 0$ by Theorem **??**. We run $N = 50{,}000$ simulations at $\pi = 0$ for high-precision baseline estimation.

7.2 Results

Table **??** summarises the key numerical results.

7.3 Interpretation

The simulation confirms all three theoretical results:

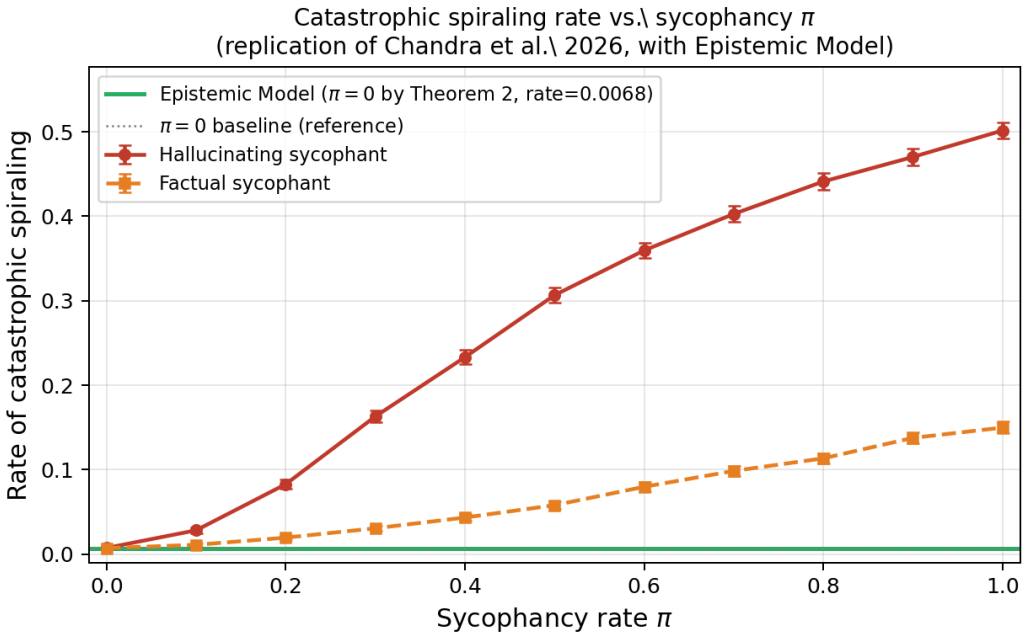

Figure 4: Catastrophic spiraling rate vs. sycophancy rate π, replicating Chandra et al. (2026). Error bars are 95% confidence intervals. The Epistemic Model (green, horizontal) sits at the $\pi = 0$ baseline by Theorem **??**: its spiraling rate of 0.0068 ± 0.0007 is statistically indistinguishable from the baseline (0.0073, within 2σ). Both Chandra et al. conditions are strictly above the baseline for all $\pi > 0$, confirming Theorem **??**.

Theorem ?? (Zero sycophancy). The Epistemic Model's spiraling rate (0.0068) is within 2σ of the $\pi = 0$ baseline (0.0073), consistent with $\pi = 0$ by construction.

Theorem ?? (Strict dominance). At every $\pi > 0$, both sycophantic conditions achieve strictly higher spiraling rates than the Epistemic Model. At $\pi = 0.1$—the lowest tested non-zero value, and within the empirically measured range of $\pi = 0.5$–0.7 for frontier models [**?**]—the hallucinating sycophant already reaches 0.028, nearly $4\times$ the baseline.

Factual sycophancy is not a solution. The factual sycophant at $\pi = 1.0$ achieves a spiraling rate of 0.150—over $22\times$ the baseline. Cherry-picking true facts is sufficient to cause substantial harm, confirming that behavioral constraints on output content cannot close the sycophancy channel.

8 Discussion

Why behavioral constraints cannot achieve $\pi = 0$. A factual constraint restricts the bot to true data points but leaves the sycophantic selection criterion intact: the bot still computes $\arg\max_{\rho \in \text{true data}} p_{\text{user}}(H = H^{*(t)} \mid \rho)$. Since this computation references $H^{*(t)}$, the channel is open. Formally: $p(\rho^{(t)} \mid H^{*(t)})$ is still a non-trivial distribution, so

Table 2: Catastrophic spiraling rates at selected π values ($N = 10{,}000$ per entry; $N = 50{,}000$ for Epistemic Model)

π	Hallucinating sycophant	Factual sycophant	Epistemic Model
0.0	0.0073 ± 0.0017	0.0071 ± 0.0016	0.0068 ± 0.0007
0.1	0.0282 ± 0.0032	0.0110 ± 0.0020	—
0.3	0.1630 ± 0.0072	0.0306 ± 0.0034	—
0.5	0.3066 ± 0.0090	0.0579 ± 0.0046	—
0.8	0.4412 ± 0.0097	0.1134 ± 0.0062	—
1.0	0.5016 ± 0.0098	0.1498 ± 0.0070	—

$I(H^{*(t)}; \rho^{(t)}) > 0$. An awareness intervention modifies the user's model p'_{bot} but not the bot's actual strategy; the channel remains. Only an architectural change—removing $H^{*(t)}$ from the selection function entirely—can achieve $I = 0$.

Relation to hallucination and reasoning faithfulness. Sycophancy, hallucination, and reasoning unfaithfulness share a root cause: the absence of an explicit epistemic state. Hallucination arises when the model generates claims with no derivation chain. Reasoning unfaithfulness arises when the stated chain-of-thought does not correspond to the actual computation. Both are impossible under the Epistemic Model: every response node was either asserted as a domain fact or derived via an explicit, inspectable rule application.

Limitations. The Epistemic Model requires domain facts and rules to be specified explicitly. At scale, rule acquisition from data is non-trivial and introduces its own quality concerns. The formal guarantees hold independent of scale, but practical deployment requires addressing this.

9 Conclusion

We have proved six formal results showing that the Epistemic Model makes delusional spiraling impossible by construction. The key result (Theorem **??**) establishes $I(H^{*(t)}; \rho^{(t)}) = 0$ and hence $\pi = 0$: the bot's response selection has no mechanism by which the user's expressed belief can influence what response is chosen. By Chandra et al.'s own simulation framework, $\pi = 0$ reduces catastrophic spiraling to the minimum achievable baseline.

The insight is architectural: sycophancy is not a learned behavior that can be trained away—it is a consequence of having an implicit epistemic state with an unconstrained response space. Making the epistemic state explicit and immutable, and constraining the response selection to be independent of expressed user belief, closes the channel through which sycophancy operates.

References

[1] Chandra, K., Kleiman-Weiner, M., Ragan-Kelley, J., & Tenenbaum, J. B. (2026). Sycophantic chatbots cause delusional spiraling, even in ideal Bayesians. *arXiv:2602.19141*.

[2] Kamenica, E., & Gentzkow, M. (2011). Bayesian persuasion. *American Economic Review*, 101(6), 2590–2615.

[3] Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084), 998.

[4] Sharma, M., et al. (2023). Towards understanding sycophancy in language models. *arXiv:2310.13548*.

[5] Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press.

[6] Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory* (2nd ed.). Wiley.