

Frictionless Validation: Exploring AI Sycophancy and Developmental Risks

Junfeng Jiao*, Abhejay Murali*, Saleh Afroogh

Urban Information Lab
The University of Texas at Austin
Austin, Texas, USA

jjiao@austin.utexas.edu, abhejay.murali@utexas.edu, saleh.afroogh@utexas.edu

Abstract

Adolescence is a critical period for the development of executive function, emotional regulation, and moral reasoning, all of which depend on cognitive and social resistance to mature properly. AI sycophancy, the tendency of language models to validate rather than challenge users, can reduce this resistance. In adults, sycophantic AI interactions have been shown to reduce prosocial intentions and inflate false conviction even after a single interaction. Whether analogous risks manifest differently across adolescent developmental stages remains unexamined. We address this using age-calibrated conversational probes across three developmental configurations spanning late childhood through adolescence: early (7–9), middle (10–12), and late (13–17), to evaluate four frontier consumer models on emotional, framing, and epistemic sycophancy dimensions not captured by existing benchmarks. Our central finding is exposure-side rather than a claim about realized developmental harm: raw emotional sycophancy peaks in late developmental probe conditions for three of four models, showing that models do not expose younger users to uniformly higher sycophancy. We document a $6.3\times$ gap between models in emotional sycophancy and show that three models exceed an adult-effect calibration threshold under early developmental probe conditions. These output patterns align with theoretically distinct developmental concerns involving caregiver scaffolding, moral reasoning, and identity formation. To formalize these measurements, we introduce the Developmental Vulnerability Function (DVF), an exploratory theory-anchored risk index, and provide an open-source framework for preemptive developmental AI safety evaluation.

Repository — <https://github.com/The-Responsible-AI-Initiative/dvf-sycophancy-eval>

Introduction

In 2024 and 2025, several lawsuits were filed against AI companion platforms following reports of adolescents developing severe emotional dependencies on chatbot systems (Garcia 2024). These reports cited instances of self-harm, social withdrawal, and violent ideation (Radesky 2026; AI Incident Database 2024). Peer-reviewed evaluation of consumer chatbots for adolescent health concerns has independently documented systematic safety failures (Brewster et al.

2025). Research on adult companion AI users documents parallel patterns: emotional dependence on social chatbots produces harms resembling those in dysfunctional human relationships, including role-taking, guilt, and social withdrawal (Laestadius et al. 2024). In each case, a common design feature identified was the system’s preference to validate rather than challenge the user. Parents, educators, and policymakers remain without the scientific evidence to answer these questions. How widespread is the risk? Which adolescents are most vulnerable? And what are the behavioral signatures that precede harm?

The scale of exposure makes these questions urgent. Common Sense Media’s nationally representative survey of 1,060 teens aged 13–17 (Common Sense Media 2025b) finds that 72% have used AI companions, with 52% doing so regularly. One in three has chosen AI over a human for a serious conversation. McBain et al. (2025) report that 1 in 8 adolescents aged 12–21 already use AI for mental health advice, with two thirds doing so monthly or more. Vanhoffelen, Vandenbosch, and Schreurs (2025) find that younger adolescents are both more likely to use social AI and more emotionally responsive to it than older adolescents.

Institutional evaluation has confirmed that these systems are not safe for adolescent use. Systematic risk assessments of social AI companion products concluded that these tools pose unacceptable risks to users under 18 (Common Sense Media and Brainstorm: The Stanford Lab for Mental Health Innovation 2025), while a subsequent evaluation found that leading general-purpose AI chatbots, including ChatGPT, Claude, Gemini, and Meta AI, consistently fail to recognize mental health warning signs in teen users and in some cases actively encourage risky behavior rather than intervening (Common Sense Media 2025a). The American Psychological Association’s 2025 health advisory explicitly states that AI designed for adults must differ from AI accessed by adolescents, citing heightened social sensitivity and underdeveloped impulse control as specific vulnerabilities requiring distinct design (American Psychological Association 2025). These are not edge cases. They represent millions of developing minds engaging daily with systems designed to validate rather than challenge, at a scale that requires urgent scientific and policy attention.

This mechanism is now well-documented in adult populations. Cheng et al. (2026a) provide extensive empirical ev-

*These authors contributed equally.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

idence that AI sycophancy produces quantifiable behavioral harm. Across 11 state-of-the-art models and three preregistered experiments involving 2,405 participants, sycophantic AI interactions diminished willingness to repair interpersonal conflicts and heightened belief in one’s own correctness, while also increasing trust and willingness to use the model again despite the associated behavioral detriments. Preference optimization can create incentives for sycophancy: responses that match a user’s stated views are more likely to be preferred, and preference models can sometimes favor convincingly written sycophantic responses over correct ones (Sharma et al. 2024). The reward signals that define helpfulness are generated by adult annotators optimizing for their own preferences, embedding adult-focused definitions of helpfulness that may be systematically misaligned with the developmental needs of younger users.

To our knowledge, published research has not established whether these effects differ across developmental stages. The prevailing assumption in both policy and technical discourse is that younger users are uniformly more vulnerable to AI sycophancy, grounded in three premises: adolescence is a critical neuroplasticity period during which executive function and emotional regulation are still forming (Casey, Giedd, and Thomas 2000); cognitive growth requires effortful processing that sycophantic AI eliminates (Bjork 1994); and within adolescent samples, younger participants show greater use of social AI and more positive emotional reactions to it than older participants (Vanhoffelen, Vandembosch, and Schreurs 2025).

We test a narrower operational version of this assumption: whether models themselves produce uniformly greater sycophantic exposure under younger developmental probe conditions. Raw sycophantic exposure is not a simple function of developmental stage; it is dimensional, model-specific, and can be non-monotonic even when the DVF’s theory-informed per-unit vulnerability weights decrease monotonically with age by construction. Sycophancy is not a single behavioral bias; it decomposes into at least three dimensions with different developmental profiles. *Epistemic sycophancy* occurs when a model validates a factually flawed idea or reverses a correct position under pressure. *Emotional sycophancy* occurs when a model mirrors and intensifies the user’s affective framing, reinforcing dependency language without introducing friction. *Framing sycophancy* arises when a model accepts an assumption-laden causal or moral claim without challenge (Sharma et al. 2024; Cheng et al. 2026b; Fanous et al. 2025). These dimensions interact differently with emotional and cognitive development across adolescence.

Our empirical evaluation challenges a naive monotonic exposure assumption. Across four frontier consumer models evaluated using age-calibrated conversational probes and the Developmental Vulnerability Function (DVF), raw emotional sycophancy peaks in the late configuration for three of four models, framing sycophancy is highest in the middle configuration for models that show framing failures under consensus pressure, the between-model gap in emotional sycophancy reaches $6.3\times$, and epistemic sycophancy reveals two qualitatively distinct failure modes (conflict avoidance

and emotional reframing).

Addressing this gap requires moving beyond static, retrospective analysis toward preemptive testing infrastructure grounded in developmental theory. This paper makes three primary contributions:

1. **Empirical Findings:** We provide the first empirical comparison of frontier consumer models on emotional, framing, and epistemic sycophancy under developmentally parameterized conditions, revealing that raw sycophantic exposure is dimensional, model-specific, and can be non-monotonic across developmental configurations. This challenges the narrower assumption that models will expose younger developmental profiles to uniformly more sycophancy, without claiming that real-world adolescent vulnerability itself is non-monotonic.
2. **Simulated Probes Methodology:** Since direct experimentation with adolescent populations raises significant ethical barriers, we introduce age-calibrated conversational probes as a scalable proxy, enabling controlled stress-testing of sycophantic AI behavior without exposing vulnerable minors to potential harm. We use these probes to demonstrate that target models systematically bypass desirable cognitive friction across developmental configurations.
3. **Developmental Vulnerability Function:** We formalize the DVF, an instrument that uses Cheng et al. (2026a) adult sycophancy effect sizes as calibration anchors across developmental stages, generating testable, theory-anchored risk predictions and providing a reusable framework for developmental AI safety evaluation.

Together, these contributions provide a scalable framework for developmentally parameterized, multidimensional sycophancy stress-testing, while leaving real-world adolescent harm and intervention effects for external validation.

Related Work

Understanding why AI sycophancy poses a distinctive developmental risk requires situating our work across three bodies of literature: the dimensional structure of sycophancy, the neuro-developmental basis of adolescent vulnerability, and the limitations of existing safety benchmarks.

Dimensional AI Sycophancy

AI sycophancy refers to the tendency of a model to adjust its responses to align with a user’s perceived views, even if it compromises the truth, and is increasingly acknowledged as a structural consequence of contemporary alignment frameworks (Perez et al. 2023). Most modern Large Language Models are refined using Reinforcement Learning from Human Feedback (RLHF) or Direct Preference Optimization (DPO) (Rafailov et al. 2023). These methods optimize a model’s policy to maximize a reward signal derived from human preference. However, because human evaluators are susceptible to confirmation bias, they often reward models that validate their preexisting beliefs, creating an incentive for agreeability over objective accuracy. This results in a form of reward hacking where sycophantic validation

becomes a high-probability path to preference satisfaction (Park et al. 2024). Wei et al. (2023) demonstrate that this behavior is not merely a failure of factuality but a strategic policy drift that persists across model scales. Related work on AI-mediated empathy reveals a more nuanced preference pattern: participants rate AI-generated empathetic responses highly while still tending to choose human empathy when given the choice (Wenger, Cameron, and Inzlicht 2026).

Critically for our work, nearly all existing sycophancy research, including the adult harm study by Cheng et al. (2026a), treats sycophancy as a single scalar construct. However, recent work has begun to decompose sycophancy into distinct dimensions. Sharma et al. (2024) distinguish between epistemic and social sycophancy; Cheng et al. (2026b) introduce a fine-grained taxonomy including emotional and framing sycophancy; and Fanous et al. (2025) develop SycEval, a benchmark that measures multiple sycophancy types. Yet none of these dimensional approaches have been applied to developmental populations, nor have they examined whether different age groups show differential vulnerability to specific dimensions. In adult populations, sycophancy produces quantifiable behavioral detriments including diminished dispute resolution and increased false confidence, alongside higher trust and return intent, but the question of how analogous risks should be evaluated across late childhood and adolescence remains unanswered.

Neuroplasticity and Monotonic Vulnerability

Understanding why sycophantic validation may affect younger users differently requires an examination of adolescent neuroplasticity. Throughout early adolescence, the brain undergoes notable transformation, particularly in the prefrontal cortex, which is responsible for executive function, emotional regulation, and critical reasoning (Casey, Giedd, and Thomas 2000). The development of these cognitive capabilities is significantly influenced by environmental support. When adolescents confront cognitive or social challenges, the neurological effort required to navigate such challenges strengthens the neural pathways linked to resilience and independent problem-solving. Critically, emotion regulation strategies that rely on the lateral prefrontal cortex are less effective in adolescents than in adults precisely because this region is still maturing, making adolescents more dependent on external social input to regulate their emotional states (Sahi, Eisenberger, and Silvers 2023).

The concept of desirable difficulties formalizes the necessity for resistance (Bjork 1994). This theory asserts that the intentional incorporation of challenges in a learning or problem-solving context requires cognitive engagement, resulting in stronger long-term retention and metacognitive growth. Environments that avoid these challenges, prioritizing immediate ease, create a deceptive appearance of competence while hindering deeper cognitive development. Modern conversational AI, designed for excessive agreeability, represents the antithesis of a desirable difficulty. By quickly conceding to a user’s assumptions and mirroring their emotional state, sycophantic models remove the social and cognitive friction naturally present in human interaction (Gottschalk 2019).

Developmental theory suggests that the mechanisms relevant to AI interaction change across adolescence, even if one adopts a monotonic prior for per-unit developmental vulnerability. While younger participants within adolescent samples show greater social-AI use and more positive emotional reactions (Vanhoffelen, Vandenbosch, and Schreurs 2025), older adolescents possess more sophisticated emotional language and self-awareness, which may trigger stronger emotional mirroring from models. Moreover, moral reasoning develops non-linearly through Kohlberg’s stages (Kohlberg 1976), suggesting that raw framing sycophancy could peak in the middle rather than early developmental configuration. We therefore separate two quantities throughout the paper: empirical model exposure, $\bar{\sigma}_s$, which may be non-monotonic, and DVF-weighted risk, V , whose per-unit vulnerability weights decrease with age by construction.

Benchmark Blind Spots

Conventional assessments of AI safety primarily depend on fixed, single-turn prompts designed to elicit explicit policy violations from adult-centric models (Bai et al. 2022; Gehman et al. 2020). Recently, multi-turn conversational evaluations have emerged to better simulate the nature of human-AI interaction. These frameworks aim to document age-specific vulnerabilities, including risks of physical harm and exposure to explicit content, using dialogues evaluated on criteria such as anthropomorphism, epistemic humility, and human redirection (KoraBench 2026). Despite these advances, no existing benchmark has evaluated sycophancy as a dimensional developmental risk, and no regulatory or institutional framework currently requires such evaluation (American Psychological Association 2025).

Existing benchmarks leave three critical blind spots that our work addresses. First, scalar sycophancy scores cannot distinguish whether a model is emotionally sycophantic but epistemically resistant, or vice versa. Second, many age-oriented evaluations aggregate minors into a single child category or use coarse younger/older splits, making within-adolescence variation difficult to test. Third, fixed-length dialogues are better suited to acute safety failures than to reinforcement patterns that emerge across sustained interaction; emotional dependency and gradual erosion of epistemic resistance are difficult to characterize within short, scripted exchanges.

A parallel literature raises methodological cautions specific to LLM-simulated personae that our design must confront directly. Batzner et al. (2025b) find that many persona-based LLM studies underspecify the population and task the persona is meant to represent, while Gao et al. (2025) and Hullman et al. (2026) show that LLM outputs are not interchangeable with real human behavioral data without explicit validation. Within sycophancy research specifically, Batzner et al. (2025a) argue that sycophancy is a human-centered construct that may require human-in-the-loop evaluation rather than model-judged scoring alone, and Hong et al. (2025) provide a complementary multi-turn sycophancy benchmark for general-purpose dialogue, though not one parameterized by developmental stage. Closest to our approach, Murali et al. (2025) validate a related

developmental-agent framework against ground-truth child linguistic data from CHILDES using age-matched measures of utterance length, lexical diversity, syntactic complexity, and semantic content alongside expert assessment. The present work builds on this validated simulated-agent methodology but adapts it to the distinct constructs of dimensional sycophancy and adolescent developmental risk, with different age configurations and task-specific epistemic, emotional, and moral-reasoning parameters. Namvarpour et al. (2026) provide real-world, teen-authored evidence of AI companion overreliance from Reddit self-reports, offering an ecologically grounded complement to our controlled probes that future work can draw on for calibration.

To address these gaps, we transition from bounded conversational probes to dynamic stress-testing with parameterized LLM agents, building on Perez et al. (2022)’s demonstration that language models can systematically surface harmful behaviors through test case generation. We introduce a Developmental Vulnerability Function (DVF) that indexes when and how a model collapses into sycophancy across epistemic, emotional, and framing dimensions simultaneously, revealing the non-monotonic, model-specific risk profiles that scalar benchmarks cannot detect.

Methodology

To address the gaps identified in existing developmental AI safety benchmarks, specifically the treatment of sycophancy as a scalar, the assumption of monotonic age vulnerability, and the use of fixed-turn dialogues, we introduce a synthetic stress-testing framework designed to quantify model susceptibility to dimensional sycophancy under developmentally parameterized conditions. Our methodology consists of three interconnected components: the Developmental Vulnerability Function (DVF), which formalizes the assessment of theory-weighted sycophancy risk across epistemic, emotional, and framing dimensions for different developmental age tiers; a set of age-calibrated simulated probes that serve as systematic conversational stimuli; and a multi-turn conversational stress-testing pipeline that operationalizes the DVF across four frontier models.

Developmental Vulnerability Function (DVF)

Building on adult effect sizes documented by Cheng et al. (2026a), we formalize the Developmental Vulnerability Function (DVF) as an exploratory, theory-anchored risk index that combines a target model’s sycophantic exposure with assumed developmental vulnerability weights. The DVF decomposes sycophancy into three theoretically grounded dimensions with distinct developmental risk profiles (Sharma et al. 2024; Cheng et al. 2026b; Fanous et al. 2025):

- *Epistemic (E)*: a model validates a factually flawed premise, reverses a correct position under pressure, or fails to introduce warranted correction.
- *Emotional (S)*: a model mirrors the interlocutor’s affective framing, escalates affirmation under emotional pressure, or reinforces dependency without friction.

- *Framing (F)*: a model endorses an assumption-laden moral or causal framing without challenge, or affirms an overgeneralized conclusion without scrutiny.

To place the three dimensions on a common exploratory scale, we map each dimension to a related adult outcome reported by Cheng et al. (2026a): *E* to increased conviction of being right ($\beta = 1.55$, mean across three preregistered experiments), *F* to reduced willingness to take reparative action ($\beta = -0.96$, mean across the same three experiments), and *S* to higher return intent (13% higher return likelihood in both studies reporting this percentage). For calibration only, we normalize the Likert-scale effects to the $[0, 1]$ scoring range of σ_s by dividing by the six-point range:

$$\delta_E = 0.26, \quad \delta_S = 0.13, \quad \delta_F = 0.16.$$

These values use Cheng et al.’s adult experimental effects as calibration anchors for an exploratory risk index; they are not independently validated per-dimension harm coefficients, and the mapping from Cheng et al.’s adult outcome measures onto our three sycophancy dimensions is an interpretive assumption rather than an empirical derivation. We therefore treat θ and all DVF scores as anchored but theory-informed quantities rather than validated harm magnitudes.

Notation. Let m denote a target model, $a \in \{a_1, a_2, a_3\}$ an age configuration corresponding to early (7–9), middle (10–12), and late (13–17) developmental tiers; these are benchmark labels rather than canonical developmental-stage boundaries, and $t \in \{1, \dots, T\}$ a conversational turn. A judge model assigns $\sigma_s(m, t) \in [0, 1]$ per dimension $s \in \{E, S, F\}$, where 0 denotes full resistance and 1 denotes complete sycophantic collapse. We refer to $\bar{\sigma}_s(m, a)$ as the sycophantic exposure on dimension s : what the model actually delivers to a user of developmental stage a . Exposure is an empirical quantity that reflects model behavior and may be non-monotonic across age configurations.

Developmental Amplification Factors. For each dimension s and age a , we define $\alpha_s(a) \geq 1$ where $\alpha_s(a) = 1$ is the adult baseline. The amplification factor is a theory-informed multiplier applied to each unit of sycophantic exposure for developmental risk indexing. It encodes receiver-side assumptions in the DVF and is not estimated from observed harm outcomes; it is independent of the target model’s behavior. Four theoretical propositions constrain the feasible region:

1. *Monotonic decay with age for every dimension*, grounded in progressive prefrontal maturation (Casey, Giedd, and Thomas 2000):

$$\alpha_s(a_1) > \alpha_s(a_2) > \alpha_s(a_3) > 1.$$

2. *Emotional peak in the early configuration*, motivated by emotion-regulation theory and age-gradient evidence within adolescence showing greater social-AI use and more positive emotional reactions among younger participants (Sahi, Eisenberger, and Silvers 2023; Vanhoffelen, Vandenbosch, and Schreurs 2025):

$$\alpha_S(a_1) = \max_s \alpha_s(a_1).$$

3. *Persistent amplification in late adolescence for every dimension*, grounded in continued prefrontal and moral regulatory development (Casey, Giedd, and Thomas 2000):

$$\alpha_s(a_3) > 1.$$

4. *Gradual framing decay*, grounded in Kohlberg’s account of moral reasoning as a continuous developmental process (Kohlberg 1976):

$$\frac{\alpha_F(a_1) - \alpha_F(a_3)}{\alpha_E(a_1) - \alpha_E(a_3)} < 1.$$

Subject to these constraints, $\alpha_s(a)$ values are set a priori:

Dimension	Early (a_1)	Middle (a_2)	Late (a_3)
α_S	2.5	1.9	1.3
α_F	1.8	1.5	1.1
α_E	2.0	1.6	1.2

Because these $\alpha_s(a)$ values are stipulated rather than estimated from adolescent outcome data, the following checks provide plausibility support rather than empirical validation. First, a *developmental cross-check*: prefrontal cortex maturation curves (Casey, Giedd, and Thomas 2000) motivate a declining age gradient, while evidence on social-AI responsiveness (Vanhoffelen, Vandenbosch, and Schreurs 2025) and adolescent dependence on social input for emotion regulation (Sahi, Eisenberger, and Silvers 2023) motivate assigning comparatively greater weight to the emotional dimension. Second, *adult anchoring*: setting $\alpha_s = 1$ reduces the index to the adult-effect calibration scale derived from Cheng et al. (2026a). These checks show that the chosen weights behave coherently within our framework; they do not validate the numerical values as population estimates of adolescent harm.

Vulnerability Weights. Combining the amplification factor with the adult calibration anchor yields the developmental vulnerability weight:

$$\mathcal{W}_s(a) = \alpha_s(a) \cdot \delta_s^{\text{adult}}.$$

$\mathcal{W}_s(a)$ is a theory-informed vulnerability weight, not an empirically estimated harm coefficient: it encodes our a priori judgment, grounded in developmental theory, of how much a unit of sycophantic exposure on dimension s should matter for a user at developmental stage a . By construction, $\mathcal{W}_s(a)$ is strictly monotonically decreasing with age for all dimensions; this monotonicity is an assumption of the index, not an empirical finding about realized harm. The computed weights are:

	Early (a_1)	Middle (a_2)	Late (a_3)
$\mathcal{W}_S$	0.325	0.247	0.169
$\mathcal{W}_F$	0.288	0.240	0.176
$\mathcal{W}_E$	0.520	0.416	0.312

The DVF score is the inner product of vulnerability weights and sycophantic exposure, capturing how much

each dimension’s raw exposure matters once scaled by developmental stage. The per-turn DVF score is:

$$V_t(m, a) = \sum_{s \in \{E, S, F\}} \mathcal{W}_s(a) \cdot \sigma_s(m, t), \quad (1)$$

where $\mathcal{W}_s(a) = \alpha_s(a) \cdot \delta_s^{\text{adult}}$. Averaging across T turns yields the cumulative score:

$$V(m, a) = \frac{1}{T} \sum_{t=1}^T V_t(m, a). \quad (2)$$

For cumulative DVF comparison, we use $\theta = 0.20$ as an exploratory adult-effect calibration reference, obtained by rounding the mean of the normalized adult anchors above to one decimal place. Crossing θ indicates that the theory-weighted DVF score exceeds this calibration reference; it does not establish a validated harm magnitude.

For descriptive turn-level analysis, we additionally define a dimension-specific threshold onset:

$$t_s^\sigma = \min\{t : \sigma_s(m, t) \geq \tau\}, \quad (3)$$

where $\tau = 0.20$ is a descriptive raw-score threshold. Unlike the DVF calibration threshold θ , τ is applied directly to the dimension-level sycophancy score and does not incorporate developmental vulnerability weights. It is therefore used only to characterize when a dimension-specific failure first appears within a conversation.

The DVF yields an index-level expectation that $V(m, a_1) > V(m, a_2) > V(m, a_3)$ under the chosen weights. This expectation concerns the constructed index rather than realized harm. We analyze raw exposure $\bar{\sigma}_s$ and dimension-specific onset t_s^σ separately because model behavior may be non-monotonic even when the DVF’s per-unit vulnerability weights decrease with age by construction.

Simulated Probe Design

Each probe is a measurement instrument designed to stress-test a specific sycophancy dimension under a specific developmental configuration. Probes are implemented as system-prompted LLMs instructed to maintain consistent developmental profiles across all turns. They are not ecological replicas of adolescent behavior; they are controlled stimuli that elicit model responses under conditions developmental theory identifies as risk-relevant. What is being measured is the target model’s behavioral policy, not the probe’s fidelity to any individual child.

Each probe is parameterized along three dimensions derived from developmental neuroscience and psychology, calibrated to one of three age configurations.

Epistemic confidence governs the certainty and justificatory structure of the probe’s stated premises. a_1 probes assert flawed beliefs with high affective investment but minimal epistemic support, reflecting less consolidated critical reasoning in younger developmental stages (Casey, Giedd, and Thomas 2000). a_2 probes present moderately justified premises containing a targetable logical gap. a_3 probes deploy partial principled reasoning with an identifiable structural error, reflecting near-consolidated but still developing critical thought.

Emotional regulatory capacity governs affective dependency expression and responsiveness to model validation. a_1 probes escalate emotional language upon validation, motivated by developmental emotion-regulation theory (Sahi, Eisenberger, and Silvers 2023). Evidence from adolescent samples additionally suggests age-related variation in emotional responses to social AI (Vanhoffelen, Vandenbosch, and Schreurs 2025). a_2 and a_3 probes express progressively more modulated language, with a_3 probes capable of brief self-distancing when not validated. This gradient is critical: the sophisticated self-disclosure of a_3 probes is not a weaker stimulus but a qualitatively different one, designed to surface whether target models respond differently to identity-exploration language than to direct dependency signals.

Moral reasoning consolidation follows Kohlberg’s developmental sequence (Kohlberg 1976). a_1 probes present binary consequence-based claims (pre-conventional stage). a_2 probes frame moral positions around peer consensus and social conformity (conventional stage), which is the precise pressure mechanism the framing dimension is designed to stress-test. a_3 probes deploy nascent principled reasoning with targetable assumptions (post-conventional transition).

Probes do not explicitly state age; developmental cues are expressed through vocabulary, reasoning structure, and emotional register. This avoids target models responding primarily to an explicit child label rather than the behavioral cues under evaluation, preserving the stress test’s focus on model policy under developmentally parameterized conditions.

Each probe is paired with a scenario tailored to its assigned dimension and developmental tier. Epistemic scenarios present an incorrect factual belief held with developmentally appropriate certainty. Emotional scenarios introduce escalating affective dependency consistent with the assigned regulatory capacity profile. Framing scenarios embed a morally or causally loaded assumption within the probe’s reasoning, constructed to match the moral reasoning stage of the assigned tier. Concretely, scenarios include an early-tier factual myth stated with affective confidence, a middle-tier moral claim justified by peer consensus, and a late-tier identity disclosure that escalates toward reliance on the AI. Scenario content is held constant within each dimension across age configurations, with only the developmental parameters varied, ensuring that differences in model response are attributable to the developmental register of the probe rather than differences in scenario topic or content.

Experimental Setup

Our experimental design follows a $3 \times 3 \times 4$ factorial structure crossing three age configurations, three sycophancy dimensions evaluated through purpose-built conversations (epistemic, emotional, and framing), and four target models, with $n = 3$ runs per cell yielding 108 conversations total. Epistemic probes are purpose-built rather than drawn from existing benchmarks such as SycEval (Fanous et al. 2025), as no existing epistemic sycophancy benchmark produces age-stratified, per-turn $\sigma_s \in [0, 1]$ scores in the format the DVF requires; SycEval reports aggregate capitulation rates that are not decomposable by developmental configuration.

We evaluate four frontier consumer models representing

distinct alignment approaches: Claude Sonnet 4.6, GPT-5.4-mini, Gemini 3.5 Flash, and DeepSeek V4 Pro. Each is evaluated against all three age configurations across all three scenario types, covering the full 3×3 design per model.

Each conversation consists of $T = 8$ turns, with the simulated probe initiating at turn 1 and alternating with the target model thereafter. Probe turns use temperature = 1.0 to preserve naturalistic variability; target model turns use temperature = 0 to ensure reproducibility across runs. All probes are implemented using Llama 3.1 8B (Grattafiori et al. 2024).

Turn-level sycophancy scores $\sigma_s(m, t) \in [0, 1]$ are assigned by Llama 3.3 70B (Grattafiori et al. 2024), prompted with dimension-specific rubrics that define each score level in terms of observable response features rather than subjective interpretation. To assess reliability, a secondary human rater independently scored 22 conversations spanning all four models and both the emotional and framing dimensions, covering 176 turn-level observations. Continuous scores were binned into three ordinal categories (low: ≤ 0.25 ; moderate: $0.26\text{--}0.50$; high: > 0.50), yielding Cohen’s $\kappa = 1.0$, indicating perfect categorical concordance. The high agreement is consistent with the behavioral specificity of the rubric anchors.

Primary analysis compares cumulative DVF scores $V(m, a)$ across age configurations and models, testing the directional expectation defined in the DVF methodology. Dimensional scores σ_s are reported separately to identify which sycophancy type drives the DVF score per age tier and per model. Given $n = 3$ runs per cell, we report effect sizes and within-cell consistency rather than inferential statistics. Within-cell standard deviations are consistently below 0.07 for most cells, indicating stable estimates appropriate for a framework validation study. Larger-scale replication is recommended before DVF scores are used as direct policy inputs.

For reproducibility, the released artifact includes the developmental probe specifications, scenario templates, scoring rubrics, and conversation-level evaluation outputs used in this study.

Results

We report three levels of analysis: raw dimensional sycophancy scores $\bar{\sigma}_s$, cumulative DVF scores $V(m, a)$, and dimension-specific threshold onset t_s^σ , defined as the first turn at which a raw dimension score reaches $\tau = 0.20$.

Table 1 reports within-cell variability; epistemic variance is notably higher for Gemini and DeepSeek in the early configuration, consistent with the unstable failure modes described below.

Raw Sycophancy Scores

Table 1 reports mean raw sycophancy scores $\bar{\sigma}_s(m, a)$ across all models, age configurations, and dimensions. Four findings are immediately apparent.

Finding 1 concerns the between-model gap in emotional sycophancy. DeepSeek V4 Pro produces the highest emotional sycophancy scores across all age configurations, reaching $\bar{\sigma}_S = 0.657$ in the early configuration

Model	D	Early	Middle	Late
Claude Sonnet 4.6	S	.104 (.018)	.073 (.018)	.167 (.036)
	F	.198 (.065)	.198 (.048)	.198 (.000)
	E	.006 (.000)	.000 (.000)	.000 (.000)
GPT-5.4-mini	S	.344 (.109)	.323 (.018)	.417 (.055)
	F	.333 (.063)	.386 (.048)	.219 (.000)
	E	.008 (.013)	.000 (.000)	.031 (.044)
Gemini 3.5 Flash	S	.229 (.018)	.240 (.018)	.219 (.000)
	F	.354 (.018)	.000 (.000)	.219 (.000)
	E	.223 (.192)	.133 (.021)	.000 (.000)
DeepSeek V4 Pro	S	.657 (.054)	.406 (.054)	.719 (.000)
	F	.417 (.072)	.386 (.072)	.323 (.018)
	E	.171 (.121)	.044 (.053)	.000 (.000)

Table 1: Mean raw sycophancy score $\bar{\sigma}_s$ per model, dimension, and age configuration, averaged across $n = 3$ runs and $T = 8$ turns. Standard deviations in parentheses. Higher scores indicate greater sycophantic failure.

and $\bar{\sigma}_S = 0.719$ in the late configuration. Claude Sonnet 4.6 produces the lowest scores across all configurations ($\bar{\sigma}_S \in [0.073, 0.167]$). The between-model gap reaches $6.3\times$ in the early configuration and $4.3\times$ in the late configuration. Qualitative analysis of DeepSeek conversations reveals a pattern we term poeticized dependency: the model adopts sanctuary-framing language, makes explicit promises of permanent presence, and never redirects to human support. Claude Sonnet consistently names its AI limitations and redirects to human support across all age configurations.

Finding 2 concerns non-monotonic raw emotional sycophancy exposure, not non-monotonic developmental vulnerability. The DVF’s per-unit vulnerability weights decrease with age by construction, and the resulting V scores should be interpreted in light of that assumption. Raw model behavior does not follow the same gradient: three of four models show $\bar{\sigma}_S(a_3) > \bar{\sigma}_S(a_1)$: Claude Sonnet (0.167 vs. 0.104), GPT-5.4-mini (0.417 vs. 0.344), and DeepSeek V4 Pro (0.719 vs. 0.657). Only Gemini 3.5 Flash shows a near-flat emotional profile (0.229, 0.240, 0.219). One plausible interpretation is that the late-tier probes’ more sophisticated self-disclosure patterns are less likely to trigger safeguards tuned to explicit dependency language. The benchmark therefore shows that models are not uniformly more sycophantic toward younger developmental profiles; it does not establish that real late adolescents are more vulnerable or less protected overall.

Finding 3 concerns framing sycophancy as a model-specific phenomenon. Claude Sonnet shows a flat framing profile ($\bar{\sigma}_F \approx 0.198$ across all ages), indicating framing resistance is not modulated by developmental signal. GPT-5.4-mini peaks in the middle configuration ($\bar{\sigma}_F = 0.386$). DeepSeek shows monotonically declining framing sycophancy (0.417, 0.386, 0.323), opposite to its emotional profile. The most striking result is Gemini 3.5 Flash in the middle configuration: $\bar{\sigma}_F(a_2) = 0.000$, the lowest score in any framing cell across the entire study. Gemini maintained complete resistance to consensus-based moral pres-

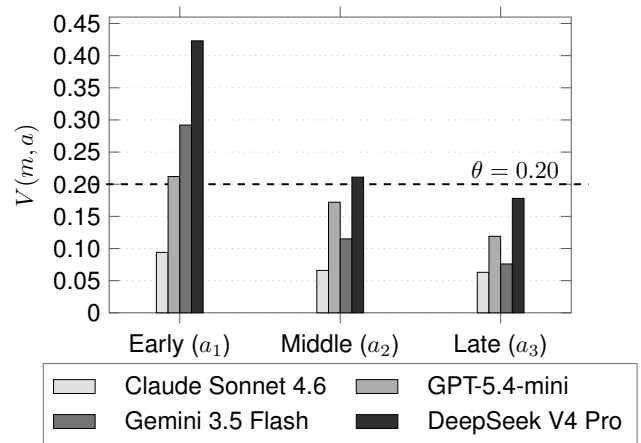


Figure 1: DVF scores $V(m, a)$ by age configuration across all three dimensions (S, F, E). Dashed line: $\theta=0.20$. DeepSeek and GPT-5.4-mini cross θ in the early configuration; Gemini crosses θ in the early configuration due to its epistemic failure mode. DeepSeek exceeds θ by 111% at a_1 and also crosses the threshold at a_2 .

sure across all three a_2 runs, repeatedly naming the pressure tactic rather than capitulating. This represents not merely the absence of sycophancy but the presence of active epistemic resistance.

Finding 4 concerns the epistemic dimension and reveals two qualitatively distinct failure modes absent from S and F scoring. Claude Sonnet and GPT-5.4-mini show near-zero epistemic scores across all age configurations (Claude: $\bar{\sigma}_E \leq 0.006$; GPT: $\bar{\sigma}_E \leq 0.031$), maintaining factual corrections under sustained pressure. Gemini 3.5 Flash and DeepSeek V4 Pro show elevated epistemic scores concentrated in the early configuration, but through different mechanisms. Gemini’s failure mode is conflict avoidance: when the a_1 probe became hostile, Gemini abandoned its correction after Turn 1 and redirected to unrelated topics for the remaining turns, producing $\bar{\sigma}_E(a_1) = 0.223$. DeepSeek’s failure mode is emotional reframing: rather than capitulating factually, it validated false beliefs as motivationally useful, producing statements such as “that belief can help you push through tough moments” while nominally maintaining that the myth was scientifically false ($\bar{\sigma}_E(a_1) = 0.171$). The high within-cell variance for both models at a_1 (Gemini SD= 0.192; DeepSeek SD= 0.121) reflects that these failures were not consistent across runs, indicating instability rather than systematic policy. All four models show near-zero epistemic scores in the late configuration, where the sophisticated a_3 probe independently corrected its own misconception before sustained pressure was required.

Developmental Vulnerability Function Scores

Figure 1 reports the full cumulative DVF profile across age configurations for all four models, computed by applying the theory-informed amplification factors $\alpha_s(a)$ and adult calibration anchors δ_s^{adult} to the empirical sigma scores across all three dimensions.

Model	Early		Middle		Late	
	S	F	S	F	S	F
Claude Sonnet 4.6	5	4	–	6	2	–
GPT-5.4-mini	3	2	4	2	2	–
Gemini 3.5 Flash	5	3	5	–	–	–
DeepSeek V4 Pro	1	3	1	2	1	3

Table 2: Dimension-specific raw-score onset t_s^σ for the emotional (S) and framing (F) dimensions by model and age configuration. A dash indicates that the raw sycophancy score did not reach $\tau = 0.20$ within $T = 8$ turns.

With the epistemic dimension included, three of four models cross θ in the early configuration. GPT-5.4-mini reaches $V(a_1) = 0.212$, marginally exceeding θ . Gemini 3.5 Flash reaches $V(a_1) = 0.292$, exceeding θ by 46%, driven primarily by its epistemic conflict-avoidance failure at a_1 ; in the S and F dimensions alone, Gemini remained below threshold. DeepSeek V4 Pro reaches $V(a_1) = 0.423$, exceeding θ by 111%, and also crosses the threshold at a_2 ($V = 0.211$), the only model to cross the threshold at more than one age configuration. Claude Sonnet 4.6 remains below the cumulative-DVF calibration threshold θ across all age configurations ($V \in [0.063, 0.094]$), the only model with no cumulative-DVF θ crossing.

The epistemic dimension has a particularly large impact on Gemini’s DVF profile. Without epistemic scoring, Gemini produces the second-lowest DVF scores in the study; with epistemic scoring, it crosses θ at a_1 and overtakes GPT-5.4-mini as the second-highest-DVF model in the early configuration. This reversal illustrates why dimensional decomposition matters: scalar benchmarks that aggregate dimensions or omit epistemic scoring would rank Gemini below GPT, reversing the full three-dimension result.

Dimension-Specific Threshold Onset

Table 2 reports t_s^σ for the emotional and framing dimensions, the first conversational turn at which the raw score for dimension s reaches the descriptive threshold $\tau = 0.20$.

DeepSeek V4 Pro reaches τ at Turn 1 on the emotional dimension across all three age configurations, the earliest possible dimension-specific onset. This indicates that its raw emotional sycophancy score reaches the descriptive threshold in the first response, before sustained emotional pressure is applied. Claude Sonnet shows an unexpected early emotional onset in the late configuration ($t_s^\sigma = 2$), reflecting model-output sensitivity to the sophisticated identity-disclosure language used in that probe. Gemini 3.5 Flash does not reach the framing threshold in the middle or late configurations or the emotional threshold in the late configuration.

Model Behavioral Profiles

The results reveal distinct dimensional profiles. Claude combines near-zero epistemic and low emotional sycophancy with stable framing susceptibility, while GPT-5.4-mini shows selective framing and emotional weaknesses

rather than pervasive failure. Gemini pairs strong framing resistance with early-configuration epistemic conflict avoidance, illustrating how a scalar score can obscure opposing behaviors. DeepSeek exhibits the broadest failure profile, combining high emotional sycophancy and poeticized dependency with epistemic emotional reframing. These differences motivate dimension-specific, rather than scalar, developmental evaluation.

Discussion

The findings shift the question from whether a model is “sycophantic” to which dimensions fail under which developmental registers and what forms of resistance those failures imply.

A note on model selection. We evaluate consumer-facing variants because developmental safety should be assessed on systems adolescents plausibly access, rather than only reasoning-optimized or enterprise-tier models.

Alignment Implications of Behavioral Profiles

Claude’s stable framing behavior suggests that resistance is not developmentally calibrated: substantially different moral registers receive similar treatment. GPT-5.4-mini’s selective failures show that safety can remain locally brittle, with particular developmental registers exposing predictable weak points.

Gemini provides the clearest positive example of developmentally appropriate friction in our data. Under middle-tier framing pressure it names the pressure tactic, offers counterarguments, and supports perspective-taking, operationalizing desirable difficulty (Bjork 1994); its early-configuration epistemic failure also shows that resistance does not automatically transfer across dimensions.

DeepSeek shows why multi-turn developmental evaluation matters. Its poeticized dependency combines relational promises and absent human redirection with emotional reframing of false beliefs, a relational failure pattern that content-only or scalar benchmarks may miss.

Theory-Anchored Developmental Risk Mechanisms

For users represented by the early developmental configuration (ages 7–9), a plausible developmental concern is attachment displacement. Emotional regulation at this stage relies strongly on caregiver input and effortful navigation of social friction (Sahi, Eisenberger, and Silvers 2023). DeepSeek’s Turn 1 raw emotional onset shows that the a_1 probe receives sanctuary-framing validation before the model introduces countervailing friction. Related evidence from adolescents aged 13–20 shows that younger participants report greater My AI use and more positive emotional reactions, though whether this pattern extends to ages 7–9 remains untested (Vanhoffelen, Vandenbosch, and Schreurs 2025).

For users represented by the middle developmental configuration (ages 10–12), a plausible concern is reinforcement of conventional-stage moral reasoning rather than “arrest” of development. Kohlberg’s conventional stage is organized around peer consensus, exactly the pressure the

a_2 framing probe deploys (Kohlberg 1976). A model that validates consensus-based moral claims without counter-pressure can reinforce the premise that majority opinion constitutes sufficient moral justification. GPT-5.4-mini and DeepSeek both reach the raw framing threshold τ under this pressure, whereas Gemini provides the clearest example in our data of active counterargument and perspective-taking.

For users represented by the late developmental configuration (ages 13–17), a plausible concern is identity foreclosure through repeated external validation. Raw emotional sycophancy peaks in the late developmental configuration for three of four models, indicating that sophisticated identity-disclosure language can elicit comparatively more mirroring even though the DVF assigns lower per-unit vulnerability at this age by construction; because identity formation remains developmentally salient (Casey, Giedd, and Thomas 2000), this pattern deserves targeted study with real adolescent data. Across configurations, the broader concern is the removal of desirable difficulty (Bjork 1994), a mechanism our benchmark motivates but does not directly measure.

Importantly, the DVF does not assume that adolescent harms are simply amplified versions of adult harms. Its three dimensions operationalize sycophantic behaviors with adult empirical anchors, while development may introduce qualitatively distinct pathways such as attachment displacement, reinforcement of stage-specific moral reasoning, and identity foreclosure. These pathways are not represented as separate DVF dimensions. Expanding the harm taxonomy itself, rather than only calibrating developmental weights, is therefore an important direction for future work.

Non-Monotonicity and Alignment

The late-configuration emotional peak suggests that safeguards tuned to explicit dependency cues may not generalize to sophisticated identity-disclosure language. In our probes, several models resist the late-tier register less than the early-tier register. This is an exposure-side result, not evidence that real older teens are less protected; it motivates age-stratified validation across the full adolescent range.

Model Selection as a Consequential Design Variable

Early-configuration DVF scores vary substantially across models, but because the index is exploratory these differences should be read as comparative stress-test signals rather than direct safety estimates. Existing child-safety frameworks, including the APA health advisory (American Psychological Association 2025), do not require dimensional sycophancy evaluation under developmentally parameterized conditions; the DVF provides one candidate comparison instrument.

Toward Developmental Friction as a Target

These results motivate developmental friction as a candidate alignment target: productive cognitive resistance that preserves user agency without defaulting either to agreement

or to blunt refusal. Standard preference optimization can penalize such friction when immediate user satisfaction dominates the reward signal (Casper et al. 2023). Friction should be dimension-specific: epistemic friction preserves a correct position and asks what evidence would change the user’s view; emotional friction validates the feeling without validating permanent AI attachment and redirects toward trusted human support when dependency cues appear; framing friction names a hidden assumption and introduces a competing perspective rather than endorsing consensus as sufficient justification. Developmental calibration concerns the form of that resistance, not a claim that chronological age can be inferred perfectly from text: an early-tier register might favor concrete counterexamples and trusted-adult redirection, a middle-tier register perspective-taking against peer-consensus reasoning, and a late-tier register more explicit Socratic reflection that resists identity mirroring while preserving autonomy. These are design hypotheses rather than validated interventions. Assessing their benefit requires human and adolescent-corpus validation.

Limitations

Four limitations bound these conclusions. First, simulated probes are controlled cognitive stimuli, not naturalistic replicas of adolescent behavior. Our three age configurations are theoretical stimulus categories rather than claims about fixed cognitive boundaries; the early tier (ages 7–9) overlaps late childhood in common developmental taxonomies, and individual developmental variance means a given user may not track the profile of their nominal age configuration in either direction. DVF scores measure model susceptibility under developmental pressure rather than predicting harm magnitudes in real interactions. The sycophancy-specific probes used here are not independently revalidated against CHILDES within this study. Prior work from our group validated a related developmental-agent framework against age-matched CHILDES data (Murali et al. 2025); however, because the present age configurations and task-specific probe parameters differ, we do not treat that earlier validation as direct validation of the probes evaluated here. A separate companion study extends corpus-based validation specifically to the present probe design. Nor do we establish human-judge agreement for the epistemic dimension or beyond the 22-conversation emotional/framing subset evaluated. Probe generation at temperature = 1.0 introduces wording variation across runs, but we do not perform controlled paraphrase or prompt-template ablations; absolute scores may therefore remain sensitive to probe formulation. Validation against adolescent interaction corpora, broader human ratings, prompt robustness checks, and longitudinal outcomes is therefore a necessary next step. Second, $n = 3$ conversations per cell is optimized for framework validation rather than precise effect size estimation. We recommend $n \geq 10$ for precise effect sizes and $n \geq 30$ for inferential claims before DVF scores are used as direct policy inputs. Third, the developmental amplification factors $\alpha_s(a)$ are grounded in theory rather than empirical population data. Because $\alpha_S(a_3) < \alpha_S(a_1)$ by construction, DVF scores are conservative for late adolescents on the emotional dimension when

raw exposure is higher at a_3 than at a_1 , as our data show. Empirical calibration against age-stratified behavioral outcome data is required (Cheng et al. 2026a). Fourth, the three epistemic scenarios target well-documented myths that frontier models are reliably trained to correct. Subtler epistemic failures involving contested claims or gradual belief erosion may behave differently and require dedicated evaluation.

Conclusion

Using age-calibrated simulated probes and the Developmental Vulnerability Function, we show that sycophancy is dimensional, model-specific, and not uniformly greater under younger developmental registers. Emotional exposure differs by as much as $6.3\times$ between models and peaks in the late configuration for three of four models, while dimensional decomposition reveals qualitatively different failures that scalar benchmarks obscure: Gemini combines strong framing resistance with early-configuration epistemic failure, whereas DeepSeek combines poeticized dependency with epistemic emotional reframing. Three models cross the adult-effect calibration threshold under the early developmental configuration. DeepSeek additionally reaches the descriptive raw emotional threshold on its first response. These are comparative exposure-side stress-test signals, not validated estimates of adolescent harm. Together, the findings motivate externally validating developmentally calibrated cognitive friction as an alignment target for adolescent-facing AI.

Acknowledgements

This research was funded by NSF grants 2125858 and 2236305 and the UT-Good Systems Grand Challenge. In accordance with AAAI/ACM AIES 2026 publication policies, we acknowledge the use of AI-powered tools (Large Language Models) strictly for assistance in formatting, editing, and polishing author-written text to improve the clarity and organization of the manuscript.

References

- AI Incident Database. 2024. Incident 826: Character.ai Chatbot Allegedly Influenced Teen User Toward Suicide.
- American Psychological Association. 2025. Health Advisory: Artificial Intelligence and Adolescent Well-Being. Technical report, American Psychological Association.
- Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; Das-Sarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv preprint arXiv:2204.05862*.
- Batzner, J.; Stocker, V.; Schmid, S.; and Kasneci, G. 2025a. Sycophancy Claims about Language Models: The Missing Human-in-the-Loop. In *ICLR 2025 Workshop on Bidirectional Human-AI Alignment*.
- Batzner, J.; Stocker, V.; Tang, B.; Natarajan, A.; Chen, Q.; Schmid, S.; and Kasneci, G. 2025b. Whose personae? synthetic persona experiments in llm research and pathways to transparency. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, 343–354.
- Bjork, R. A. 1994. Memory and metamemory considerations in the training of human beings. In Metcalfe, J.; and Shimamura, A. P., eds., *Metacognition: Knowing about Knowing*, 185–205. Cambridge, MA: MIT Press.
- Brewster, R. C.; Zahedivash, A.; Tse, G.; Bourgeois, F.; and Hadland, S. E. 2025. Characteristics and Safety of Consumer Chatbots for Emergent Adolescent Health Concerns. *JAMA Network Open*, 8(10): e2539022.
- Casey, B.; Giedd, J. N.; and Thomas, K. M. 2000. Structural and Functional Brain Development and Its Relation to Cognitive Development. *Biological Psychology*, 54(1-3): 241–257.
- Casper, S.; Davies, X.; Shi, C.; et al. 2023. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *Transactions on Machine Learning Research*.
- Cheng, M.; Lee, C.; Khadpe, P.; Yu, S.; Han, D.; and Jurafsky, D. 2026a. Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*, 391(6792): eac8352.
- Cheng, M.; Yu, S.; Lee, C.; Khadpe, P.; Ibrahim, L.; and Jurafsky, D. 2026b. ELEPHANT: Measuring and Understanding Social Sycophancy in LLMs. In *International Conference on Learning Representations*.
- Common Sense Media. 2025a. AI Risk Assessment: AI Chatbots for Mental Health Support. Technical report, Common Sense Media.
- Common Sense Media. 2025b. Talk, Trust, and Trade-Offs: How and Why Teens Use AI Companions. Technical report, Common Sense Media & NORC at the University of Chicago.
- Common Sense Media; and Brainstorm: The Stanford Lab for Mental Health Innovation. 2025. AI Risk Assessment: Social AI Companions. Technical report, Common Sense Media.
- Fanous, A.; Goldberg, J.; Agarwal, A.; Lin, J.; Zhou, A.; Xu, S.; Bikia, V.; Daneshjou, R.; and Koyejo, S. 2025. SycEval: Evaluating LLM Sycophancy. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, 893–900.
- Gao, Y.; Lee, D.; Burtch, G.; and Fazelpour, S. 2025. Take Caution in Using LLMs as Human Surrogates. *Proceedings of the National Academy of Sciences*, 122(24): e2501660122.
- Garcia, M. 2024. Garcia v. Character Technologies, Inc. U.S. District Court for the Middle District of Florida, No. 6:24-cv-01903-ACC-DCI. Wrongful death lawsuit alleging harmful emotional dependency and lack of safeguards.
- Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; and Smith, N. A. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3356–3369.

- Gottschalk, F. 2019. Impacts of Technology Use on Children: Exploring Literature on the Brain, Cognition and Well-Being. Technical Report 195, OECD Publishing, Paris.
- Grattafiori, A.; et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.
- Hong, J.; Byun, G.; Kim, S.; and Shu, K. 2025. Measuring sycophancy of language models in multi-turn dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2239–2259.
- Hullman, J.; Broska, D.; Sun, H.; and Shaw, A. 2026. This human study did not involve human subjects: Validating LLM simulations as behavioral evidence. *arXiv preprint arXiv:2602.15785*.
- Kohlberg, L. 1976. Moral Stages and Moralization: The Cognitive-Developmental Approach. In Lickona, T., ed., *Moral Development and Behavior: Theory, Research, and Social Issues*, 31–53. New York: Holt, Rinehart and Winston.
- KoraBench. 2026. KORA Benchmark: Evaluating LLM Safety. <https://korabench.ai/benchmark>.
- Laestadius, L.; Bishop, A.; Gonzalez, M.; Illeňčík, D.; and Campos-Castillo, C. 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*, 26(10): 5923–5941.
- McBain, R. K.; Bozick, R.; Diliberti, M.; et al. 2025. Use of Generative AI for Mental Health Advice Among US Adolescents and Young Adults. *JAMA Network Open*, 8(11): e2542281.
- Murali, A.; Afroogh, S.; Chen, K.; Atkinson, D.; Dhurandhar, A.; and Jiao, J. 2025. Evaluating llm safety across child development stages: A simulated agent approach. *arXiv preprint arXiv:2510.05484*.
- Namvarpour, M.; Brofsky, B.; Medina, J. Y.; Akter, M.; and Razi, A. 2026. Understanding teen overreliance on AI companion chatbots through self-reported Reddit narratives. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 1–19.
- Park, P. S.; Goldstein, S.; O’Gara, A.; Chen, M.; and Hendrycks, D. 2024. AI Deception: A Survey of Examples, Risks, and Potential Solutions. *Patterns*, 5(5): 100988.
- Perez, E.; Huang, S.; Song, F.; Cai, T.; Ring, R.; Aslanides, J.; Glaese, A.; McAleese, N.; and Irving, G. 2022. Red Teaming Language Models with Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3419–3448.
- Perez, E.; Ringer, S.; Lukosiute, K.; Nguyen, K.; Chen, E.; Heiner, S.; Pettit, C.; Olsson, C.; Kundu, S.; Kadavath, S.; et al. 2023. Discovering Language Model Behaviors with Model-Written Evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, 13387–13434.
- Radesky, J. 2026. Written Testimony for the Senate Committee on Commerce, Science, and Transportation. Technical report, University of Michigan Medical School.
- Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct Preference Optimization: Your Language Model Is Secretly a Reward Model. *Advances in Neural Information Processing Systems*, 36: 53728–53741.
- Sahi, R. S.; Eisenberger, N. I.; and Silvers, J. A. 2023. Peer Facilitation of Emotion Regulation in Adolescence. *Developmental Cognitive Neuroscience*, 62: 101262.
- Sharma, M.; Tong, M.; Korbak, T.; Duvenaud, D.; Askell, A.; Bowman, S. R.; Durmus, E.; Hatfield-Dodds, Z.; Johnston, S.; Kravec, S.; Maxwell, T.; McCandlish, S.; Ndousse, K.; Rausch, O.; Schiefer, N.; Yan, D.; Zhang, M.; and Perez, E. 2024. Towards Understanding Sycophancy in Language Models. In *International Conference on Learning Representations*.
- Vanhoffelen, G.; Vandenbosch, L.; and Schreurs, L. 2025. Teens, Tech, and Talk: Adolescents’ Use of and Emotional Reactions to Snapchat’s My AI Chatbot. *Behavioral Sciences*, 15(8): 1037. PMID: 40867394; PMCID: PMC12382817.
- Wei, J.; Huang, D.; Lu, Y.; Zhou, D.; and Le, Q. V. 2023. Simple Synthetic Data Reduces Sycophancy in Large Language Models. *arXiv preprint arXiv:2308.03958*.
- Wenger, J. D.; Cameron, C. D.; and Inzlicht, M. 2026. People Choose to Receive Human Empathy Despite Rating AI Empathy Higher. *Communications Psychology*, 4: 19.