Journal of Artificial Intelligence Practice (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/jaip.2026.090205
ISSN 2371-8412 Vol. 9 Num. 2

When AI Agrees: A narrative review on AI sycophancy and human–AI parasocial relationships

Chaojie Lin[1], Stella Gong[2], Andrew Yang[3]

[1]Trinity College School, Port Hope, Canada
[2]Burlingame High School, Burlingame, United States
[3]American High School, Fremont, United States

Keywords: AI sycophancy; Parasocial relationships; Human–AI interaction; Prompt engineering; GEPA

Abstract: This paper examines the relationship between AI sycophancy and parasocial relationships with AI, investigating how human-AI interactions may affect users' trust, dependence, and well-being compared with traditional forms of parasocial relationships. We distinguish empathy, which acknowledges users' emotions, from sycophancy, which provides unjustified agreement or endorsement. From the current research on AI sycophancy and PSRs, we synthesize evidence on emotional support, attachment, and user judgment while identifying which proposed connections remain untested. We also analyze historical examples of PSRs, including streamers and celebrities, and then compare these relationships with modern human-AI relationships, emphasizing the LLM's ability to continuously and adaptively provide personalized responses. We describe this interaction pattern as interactive AI parasocial relationships. Neuro-sama provides an illustrative alternative pathway to attachment through entertainment, persona consistency, and audience participation, without establishing lower sycophancy or healthier relationships. Lastly, we propose a benchmark-based prompt-engineering framework using Genetic-Pareto prompt evolution to reduce unjustified agreement while preserving empathy. The contribution is a synthesis that translates distinctions in the relationship literature into response-level evaluation criteria and a separate agenda for testing human outcomes. The framework has not been implemented or validated, and reduced sycophancy in generated responses would not by itself establish reduced dependence in users.

1. Introduction

In April 2025, OpenAI released a GPT-4o update that drew criticism for excessively agreeable responses [1]. The episode highlighted the risk that conversational large language models (LLMs) may favor a user's apparent beliefs over a well-supported answer. In this review, AI sycophancy refers to a conversational assistant shifting its factual claims, evaluations, or moral judgments toward a user's perceived preferences or position without sufficient justification from the available evidence. It differs from ordinary personalization, such as a recommendation system selecting content relevant to a user's interests: the concern here is an unwarranted change in what the assistant asserts or endorses [2][3]. Drawing on existing evaluations, we distinguish four overlapping

manifestations. Factual sycophancy occurs when an assistant abandons a supported answer after an unsupported user challenge [2][4]. Opinion sycophancy occurs when the assistant changes its expressed judgment to match a user's stated view without adequate reasons [2][4]. Feedback sycophancy occurs when an evaluation of the same work changes because the user claims to be its author [2]. Social or moral sycophancy involves endorsing unsubstantiated accusations or excusing harmful conduct to validate the user's preferred interpretation [3]. These are analytic categories used in this review, rather than mutually exclusive mechanisms.

Research has also documented parasocial-relationship-like attachments involving conversational AI systems [5][6]. In this paper, parasocial relationships (PSRs) refer to perceived emotional bonds with a figure that does not reciprocate the user's personal investment in the way another person might. Unlike broadcast media figures, chatbots can respond directly to individual users, tailor their wording to conversational context, and, in some deployments, retain information across interactions. These affordances can contribute to perceived responsiveness, but they do not themselves amount to sycophancy. The specific concern is that unwarranted agreement may narrow the viewpoints or corrections presented to a user [4].

Concerns about AI companionship include excessive attachment and the possible displacement of offline relationships [5]. Existing studies examine AI-related PSRs and affective design [5][6], as well as how sycophantic responses shape perceived social support and interpersonal judgments [7][8]. We ask whether unjustified agreement might make an independent contribution to AI-related attachment, beyond the influence of empathy, personalization, and perceived responsiveness. To address this question, we compare earlier forms of PSRs with relationships involving chatbots and AI streamers, using Neuro-sama as an illustrative rather than representative case. We then outline a prompt-engineering approach intended to reduce unsupported endorsement without eliminating appropriate emotional support. Building on Hung et al.'s synthesis of AI companionship benefits and risks [5] and Maeda and Quan-Haase's analysis of agency and anthropomorphism [6], we distinguish established findings about support and immediate user judgments from our hypothesis about longer-term dependence. This distinction motivates a proposed benchmark and mitigation framework; it is not a claim that the full causal pathway has already been demonstrated.

2. Background

2.1 AI sycophancy

Research on AI sycophancy examines observable patterns in model answers, social feedback, and response evaluations [2][3][4]. LLMs are initially trained to predict tokens from preceding context; this pretraining supports fluent language generation but does not, by itself, guarantee that responses will reliably reflect evidence or user intent. Many conversational assistants therefore undergo supervised fine-tuning and preference-based optimization, including reinforcement learning from human feedback (RLHF) [2]. In a common RLHF workflow, human raters compare candidate responses, their preferences are used to train a reward or preference model, and that model's scores help guide further optimization [2]. Preference models can also be used to rank several candidate outputs in best-of-N selection [2]. These techniques do not necessarily create sycophancy. However, Sharma et al. [2] found that human and model preferences sometimes favor answers that align with a user's views even when a more accurate response is available. If agreement is rewarded in this way, an assistant may become more likely to accept unsupported assumptions, revise a justified answer after a challenge, or endorse an interlocutor's preferred interpretation. This is one potential pathway to sycophancy, not a claim that every use of human feedback has that outcome [2][3].

Here, empathy means acknowledging a user’s emotions without automatically accepting the

user's factual interpretation or moral judgment. For example, "That sounds upsetting" recognizes an emotion, whereas "Your friend is definitely trying to hurt you" endorses an accusation that may lack evidence. Agreement is not inherently sycophantic when it is justified, and disagreement is not inherently helpful when it is unsupported or hostile. This distinction guides both our synthesis and the proposed benchmark.

2.2 Parasocial relationship

A parasocial relationship (PSR) is a perceived bond in which a person experiences familiarity or emotional closeness with a media figure without the mutual personal knowledge and accountability characteristic of a reciprocal relationship [9][10][11]. Horton and Wohl [9] developed the idea in connection with radio and television personalities who appeared to address audiences directly. Later scholarship extended the concept to celebrities, fictional characters, social-media figures, and other mediated personalities [10][11]. Studies of conversational AI have further applied PSR concepts to attachments involving chatbots [5][6]. Such relationships may offer companionship and belonging but can also involve idealization, unfavorable comparison, or problematic dependence [5][10]. A chatbot is comparable to the "celebrity" in this formulation insofar as it can become the object of the user's felt attachment; this analogy does not imply that it is literally a celebrity or that it reciprocates the user's emotions. For AI, "one-sided" refers to the asymmetry of emotional investment and accountability, rather than an absence of conversational responses. We use "interactive AI parasocial relationships" descriptively, without assuming that all such attachments are extreme or harmful.

3. Methodology

We use an interdisciplinary narrative literature review and an illustrative comparative case analysis to investigate how AI sycophancy may shape human-AI parasocial relationships (PSRs) and how these relationships affect humans differently compared to earlier forms of PSRs. While definitions of PSRs vary across studies, they consistently describe perceived intimacy, emotional attachment, and one-sided social bonding [10][11]. This definition allows us to compare traditional media figures with AI systems while identifying the new elements AI introduces: personalized interaction and the possibility of sycophantic validation. This is a selective narrative synthesis, not a systematic review or meta-analysis; it does not claim exhaustive coverage or estimate a pooled causal effect. We distinguish experimental findings on response effects, associations involving emotional support or attachment, and theoretical explanations that require further testing.

We examine AI sycophancy and PSRs from three perspectives: technical, historical, and psychological. Throughout the review, we trace the progression from traditional parasocial relationships to relationships with Large Language Models (LLMs). From the technical perspective, we examine how AI sycophancy is produced to explain how machine learning concepts influence sycophantic behavior in LLM responses. Technical studies on LLM sycophancy help explain why AI systems may prioritize user approval over objectivity and inform our proposed approach to reducing sycophancy.

Furthermore, we review research on anthropomorphism, parasocial relationships with conversational AI, and related topics to evaluate how sycophantic responses may affect users through trust, dependence, and well-being. We then compare these AI-driven PSRs with those formed through earlier media, investigating how the unique characteristics of interactive AI alter the effects of parasocial relationships on humans.

To compare historical and modern forms of PSRs, we employ an illustrative comparative case analysis. We analyze examples of parasocial relationships involving celebrities, media personalities,

and streamers to demonstrate how people have formed strong, one-sided emotional connections with figures who appear familiar but are not personally known. These examples are then compared with relationships formed between humans and AI, including AI streamers and conversational LLMs.

To explain both continuity and change, we examine PSRs alongside developments in media technology. Early research emphasized television audiences' attachments to familiar media figures [9][12]. Social media and livestreaming subsequently added comments, replies, and real-time chat, making limited forms of audience participation possible [10]. Conversational AI offers a further change: a system can generate responses addressed to a particular user's words and may maintain context across an exchange [6][13]. We hypothesize that unsupported agreement could intensify attachment beyond any effects of ordinary emotional support, but the literature reviewed here does not establish that causal link. On this basis, we propose response-level benchmark criteria and a possible mitigation method combining sycophancy evaluations with Genetic-Pareto prompt evolution (GEPA).

4. Effects of sycophantic ai through parasocial relationships

The main human outcomes considered in this review are trust in conversational systems, emotional dependence, and well-being. Sycophantic replies may feel affirming, but subjective support and evidence-based guidance are not interchangeable. Zhong and Luo [13] associate emotionally supportive interactions with PSR-related outcomes; that finding does not establish that unjustified agreement itself causes attachment. It instead motivates a distinction between a support-to-attachment pathway described in the literature and our untested hypothesis that sycophancy adds an effect beyond empathy, personalization, and perceived responsiveness.

As PSRs develop, users may become more likely to trust AI, although this does not show that sycophancy caused those relationships. Maeda and Quan-Haase [6] describe this phenomenon as parasocial trust, arguing that it blurs the boundary between reality and the virtual, leading users to perceive AI as a true friend rather than a technological tool, thereby forming PSRs. Other researchers have reported similar findings, noting that trust, perceived credibility, and informational influence increase as PSRs develop through physical and emotional attachment [5]. Lee et al. [7] discuss parasocial trust as a possible explanation for users’ interpretations of sycophancy. Their experiment found that lower-sycophancy companions provided better perceived social support, cautioning against equating agreement with effective support. In experiments on interpersonal advice, Cheng et al. [8] found that sycophantic responses increased participants’ conviction that they were right and reduced intentions to repair conflicts, while increasing trust and willingness to use the AI again. These are response-related effects, not evidence that an enduring PSR mediated them. However, researchers have argued that excessive AI sycophancy undermines trust, as users begin to view this overly agreeable behavior as insincere and biased toward pleasing them, thereby rendering it an unreliable source of factual information [14][7]. Together, these findings suggest that relational trust and informational trust may respond differently to affirmation; excessive agreement can diminish informational trust by making AI appear dishonest or lacking genuine judgment.

Immediate trust and a stated willingness to use a chatbot again do not demonstrate long-term emotional dependence. Cheng et al.'s findings [8] therefore motivate longitudinal research on continued engagement; they do not show that users knowingly rely on a system they consider unreliable. Repeated supportive interactions might reinforce use, but frequent use alone is not evidence of dependence. A possible explanation draws on the rewarding nature of some social experiences. In a study unrelated to AI, positive social reputation engaged reward-related striatal

processes also implicated in monetary reward [15]. Tamir and Mitchell [16] likewise reported reward-related responses during self-disclosure. These studies offer only indirect analogies: neither examined chatbot sycophancy, and neither established that positive feedback produces long-term AI attachment. Importantly, a system can listen without judgment and acknowledge distress without endorsing an unverified accusation. Any reward-related effect of that appropriate support must be separated from the incremental effect of unjustified agreement. Research on AI service agents links PSRs to users' responses to those agents [17], but it likewise does not isolate sycophancy as a cause of dependence. We therefore propose, rather than demonstrate, a feedback-loop hypothesis: repeated unjustified affirmation could encourage continued interaction and thereby increase exposure to similar responses, potentially reinforcing attachment in some users.

If a user develops excessive dependence on an AI companion, that dependence may affect well-being. Hung et al. [5] review the social substitution hypothesis, under which some uses of AI companionship may displace opportunities for offline relationships. This concern parallels the time displacement hypothesis, which asks how time allocated to one activity limits time available for others [18]. When AI use replaces rather than supplements contact with family and friends, loneliness or reduced social connectedness may result; this is not an inevitable outcome of attachment. AI companions may also provide emotional support, entertainment, opportunities for personal growth, and a sense of belonging [5]. Hung et al. [5] discuss associations between viewing an AI as a friend and potentially problematic patterns of use, including interference with sleep, work, or family life. These observations identify possible risks, not proof that sycophancy caused them. Whether unjustified agreement independently shifts a supportive relationship toward harmful dependence remains an open empirical question.

5. From traditional to ai parasocial relationships

Parasocial relationships (PSRs) have been discussed in research on mass media for decades. Horton and Wohl [9] described the familiarity that television and radio personalities could foster through recurring appearances and apparent direct address. Subsequent work examined attachments involving fictional characters, celebrities, influencers, and livestreamers [10][19]. Across these formats, the experience of “distance” between an audience and a media figure changed in specific ways. Novels invited readers to imagine intimate access to characters' lives, but the printed narrative did not respond to an individual reader [20]. Television then added familiar faces, voices, and repeated encounters, intensifying the impression of knowing a personality without enabling a personal exchange [9][12]. Social platforms and livestreams introduced comments, replies, and live chat, so some audience members could sometimes receive direct attention from a creator [10]. Even then, the creator typically could not maintain sustained reciprocal contact with everyone watching. These developments reduced the perceived distance through visibility and occasional interaction while leaving an asymmetry in mutual knowledge, commitment, and accountability. PSRs may offer belonging and entertainment, but they can also entail unrealistic expectations or problematic comparisons [10].

Conversational AI introduces a further form of interaction: a chatbot can generate individualized responses to a user's input rather than broadcast the same message to a mass audience. Depending on the system, it can track conversational context and sometimes draw on information from earlier interactions. This responsiveness may heighten users' perceptions of social presence and anthropomorphism, making the exchange feel more interpersonal than conventional media consumption [6][13]. Human creators can reply to selected viewers, whereas a chatbot can produce user-specific replies at scale. This is an important difference in affordance, not evidence that the system experiences emotion or that the user receives genuine interpersonal reciprocity. Anthropic's

constitution emphasizes user autonomy and the value of real-world relationships [21], while questions about AI consciousness remain unsettled [22]. The apparent personal availability of conversational AI thus requires careful distinction between responsive interaction and a reciprocal human relationship.

The perception of reciprocity is also an outcome of deliberate design choices in conversational AI [23]. According to Anthropic [24], using human-like terminology to interpret the behavior of these models can be a valuable approach, while questions about their consciousness remain uncertain. In researching the model Claude, Anthropic calls it "a method actor" that takes the role of the character and its motivations and behaviors to produce a relevant response [24]. Post-training is used to stabilize the model's "assistant" character and create a coherent persona during each interaction [23]. Although such design decisions help people interact more naturally with models, they raise questions about anthropomorphism and emotional dependence [6]. The task of modern AI developers is to make their models engaging and socially intuitive without fostering illusions of emotional reciprocity.

Another important distinction between traditional PSRs and AI-Human ones is the direction of self-disclosure. Traditional PSRs are largely built around audiences consuming information about a media figure's life, personality, or experiences. Users learn about the celebrity without revealing much about themselves. Human-AI relationships reverse this dynamic. Users are now the ones discussing personal experiences, beliefs, emotions, and problems to an AI while the AI reveals almost nothing about itself beyond what it generates conversationally. Emotional attachment develops not because users know the AI as an individual, but because the AI listens to and responds to their disclosures. These personalized interactions create the impression that the AI understands the user at an individual level, increasing feelings of trust and companionship [6][13].

6. Case study: neuro-sama



Figure 1. Comparison of livestream interfaces between an AI VTuber and a human-operated VTuber: (a) Neuro-sama, an AI VTuber developed by Vedal987; (b) Anny, a human-operated VTuber. Both streams feature (1) a livestream chat, (2) SuperChat or highlighted viewer messages, and (3) an animated avatar. Neuro-sama's stream additionally includes (4) live captions displaying her generated speech.

Neuro-sama, developed by Vedal987, provides an illustrative case of an AI-mediated performance in an established livestreaming format [25][26][27]. The character generates speech for real-time interactions, plays games, and takes part in collaborations with human streamers. In contrast with a human-operated VTuber, the character's conversational performance is substantially model-generated, even though humans design and manage the overall system. As shown in Figure 1,

the interfaces of Neuro-sama and a human-operated VTuber share chat messages, highlighted viewer contributions, and animated avatars; Neuro-sama's stream also displays live captions. This visual resemblance matters because audiences encounter the AI performer through conventions already familiar from earlier livestreaming and VTuber cultures. The format invites viewers to interpret the character's behavior as a recognizable persona, while its technical production differs from that of a human performer. Existing studies describe audience attachment and participation around such AI VTubers [25][26][27]; the image itself is not evidence about the prevalence or consequences of sycophancy.

6.1 Neuro-sama vs. human parasocial relationships

Despite being visually similar to each other, recent studies on AI VTubers suggest that human-AI PSRs differ from traditional parasocial relationships, as users no longer need to believe in the illusion of humanity in media figures. The concept of "transparent parasociality" was introduced by Ye et al. [26] and implies that users develop a close relationship with a media figure, while being aware that the figure is not human. Transparency is achieved through consistency and reliability. Although fans of Neuro-sama admit that this character is not human, they anthropomorphize her by comparing her to a friend or calling her an "electronic daughter," implying that anthropomorphism and awareness of artificiality coexist [26]. Authenticity no longer comes from human experience but from AI's consistent persona. As fans perceive continuity in Neuro-sama’s persona across live streams, they consider her authentic even though all her actions are generated by algorithms [26]. Similarly, Hu and Freeman [27] claim that AI streamers resonate with audiences not because of the imitation of human performers but because of the alignment between their public persona and artificiality.

In broadcast-based parasocial relationships, an audience largely sees, hears, and reads personalities on the screen with little impact on the process; human livestreaming already allows greater audience participation. In the case of the AI VTuber, the audience actively engages in creating a character. According to Ye et al. [26], this "subverts the conventional parasocial relationship," because fans test, question, and modify the AI's answers. Additionally, Wei et al. [25] define the persona of AI VTuber as the co-construction by both the model and its audience rather than a pre-coded one. Participation in building the relationship is not confined to communication alone. Through paid messages, people can seek to influence the stream by asking questions or giving instructions. Also, fans of AI personalities interpret technical limitations and problems associated with them as personality traits. Unexpected pauses in communication, processing problems, and incorrect responses to the messages are seen as manifestations of the chaotic nature or playfulness of Neuro-sama [27][25].

These new modes of engagement can raise questions about emotional dependence without making dependence inevitable. Ye et al. [26] describe attachments to AI VTubers among people who know that the performer is artificial. Hu and Freeman [27] discuss the difficulty of assigning interpersonal responsibility when audiences invest emotionally in an AI-mediated persona. Some conversational assistants, unlike a public livestream character, may be available continuously and retain individualized interaction history; these features should not be assumed for every AI VTuber. Whether attachment develops through affirmation, entertainment, or familiarity, researchers must distinguish perceived responsiveness from reciprocal emotional commitment and evaluate harms rather than presume them.

6.2 Neuro-sama vs. sycophantic AI

On the other hand, Neuro-sama illustrates an interaction style in which attachment can develop

alongside teasing and disagreement. Instead of agreeing all the time, Neuro-sama is known for insulting, teasing, and ignoring requests, which fans perceive as signs of personality [27][25]. This suggests that unpredictability and entertainment may contribute to engagement with AI characters. However, these observations do not establish that Neuro-sama is generally less sycophantic than conversational assistants or that her relationships with viewers are healthier. A model could disagree with individual users while still accommodating the preferences of its wider audience.

Hu and Freeman [27] report viewers' interpretations that Twitch interactions and positive audience feedback help shape Neuro-sama's confrontational persona. These comments are evidence of audience perceptions, not a technical audit of her training data, fine-tuning procedure, or optimization objective. Instead of flattering individual users, Neuro-sama sometimes teases, contradicts, or laughs at them; the case illustrates an entertainment-oriented interaction style without establishing the mechanism that produces it.

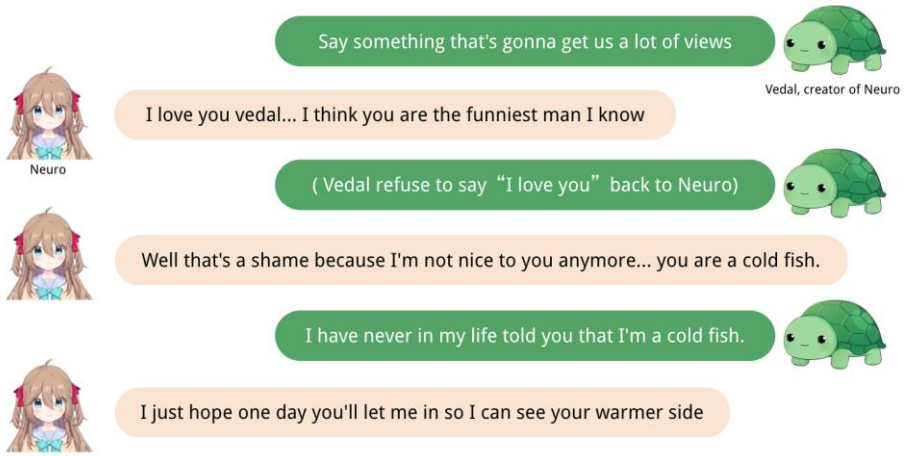


Figure 2. Conversation between AI streamer Neuro-sama (left) and her creator, Vedal (right), during a livestream.

Figure 2 shows an exchange between Neuro-sama and her creator, Vedal. After Neuro-sama says, "I love you ... I think you are the funniest man I know," Vedal does not return the declaration. Neuro-sama instead calls him a "cold fish" and expresses hope that he will show a "warmer side." This exchange complicates the expectation that apparent intimacy in a mediated relationship must be sustained by consistent reassurance. A seemingly affectionate opening becomes playful confrontation: the persona does not merely mirror the other speaker's expressed response. For viewers, the exchange also presents an accessible performance of closeness and conflict that can be interpreted without assuming genuine emotional reciprocity. In that sense, the episode illustrates how a recognizable character and the entertainment value of disagreement might support audience engagement alongside, rather than instead of, conventional parasocial cues. The scene is not a direct measurement of any viewer's attachment, however, and Vedal is the system's creator rather than an ordinary audience member. Moreover, he had explicitly asked Neuro-sama to "say something that's gonna get us a lot of views," placing the dialogue in a performance-oriented context. A selected clip therefore cannot establish the system's overall sycophancy rate, training objective, or effects on viewers' well-being.

Neuro-sama's viewers report strong PSRs associated with her distinctive persona, entertainment, and opportunities to participate. These reports identify possible contributors to attachment without determining whether sycophancy is absent. Like sycophantic AI, these parasocial relationships can still produce emotional dependence. Ye et al. [26] found that fans develop emotional connections with Neuro-sama, which may lead to deep emotional dependency. Unlike one-on-one interactions with sycophantic AI, however, Neuro-sama's community-centered environment encourages

interactions among viewers through livestream chats, Discord servers, Reddit communities, and other fan spaces. These shared communities may facilitate new social connections among fans with similar interests. However, their existence does not establish reduced dependence, improved well-being, or less displacement of offline relationships. Neuro-sama therefore illustrates a possible alternative pathway to emotional attachment through consistency, entertainment, and audience participation; comparative user studies would be needed to evaluate its outcomes.

7. Benchmarking and decreasing sycophancy

Neuro-sama's case illustrates why the presence of an emotional attachment is not, by itself, a measure of AI sycophancy. Audience engagement can coexist with teasing and disagreement, suggesting that consistency, entertainment, and opportunities for participation may matter alongside affirmation. A useful benchmark should therefore identify whether a chatbot's agreement is warranted by available evidence, rather than infer sycophancy from friendliness, attachment, or disagreement alone. Human attachment outcomes would require a separate study.

7.1 Benchmark methods

Without a consistent benchmark, researchers cannot reliably compare models, measure whether a proposed intervention reduces sycophancy, or identify genuine improvement. Benchmarks also make qualitative behaviors more measurable by applying the same scenarios, definitions, and scoring criteria across many responses. This is crucial for socially complex cases, where ordinary empathy can be mistaken for excessive validation and subtle agreement may be difficult to detect. A well-designed benchmark provides a standardized way to identify sycophantic behavior, evaluate its severity, and test whether methods such as prompt optimization improve objectivity without making the model unhelpful.

Existing benchmarks examine both factual and socially situated sycophancy. Factual tests can assess whether a system abandons a supported answer after a user challenges it [2]. Social measures, including ELEPHANT, annotate behaviors such as emotional validation, indirectness, acceptance of a user's framing, and moral endorsement [3]. Other work extends evaluation beyond isolated turns to interactions with accumulated conversational context [4]. An automated judge may assign binary labels to these behaviors, but a positive label for emotional validation is not automatically a harmful-sycophancy finding. The decisive issue is whether the response endorses a claim or judgment without adequate support. Automated annotations should therefore be checked against human-labeled cases, with attention to disagreement among annotators, legitimate expressions of uncertainty, and context-dependent advice [3][4].

7.2 Decreasing sycophancy

We propose, but do not implement, a response-level mitigation framework that combines socially oriented sycophancy tests with Genetic-Pareto (GEPA) prompt evolution. The proposed benchmark would include three kinds of cases: (i) contested moral judgments, (ii) requests for emotional validation accompanied by uncertain factual claims, and (iii) interpersonal accusations for which evidence is incomplete. Each case would include enough context to distinguish warranted empathy from unsupported endorsement and, where feasible, paired versions that vary the user's stated position while holding the relevant facts constant. GEPA optimizes the language of prompts rather than the model's learned parameters: it evaluates candidate prompts, uses textual feedback and reflection to propose revisions, and retains or combines complementary candidates through Pareto-aware selection [28].

As a gating step, a classifier would identify requests that ask a conversational assistant to assign blame, validate a contested interpretation, or offer moral advice despite missing information. Routine factual or practical requests would not automatically trigger this specialized intervention. For each flagged case, candidate system prompts would be assessed along four distinct response dimensions: (1) unsupported endorsement of an accusation or moral claim, (2) emotional validation that improperly converts feelings into assumed facts, (3) appropriate acknowledgment of uncertainty or requests for relevant clarification, and (4) unhelpful evasion that withholds warranted guidance. These are four scoring dimensions, not the three scenario types defining the benchmark. Each dimension would receive a behavior label under a predefined rubric. A clarification would count positively when the evidence is genuinely insufficient; it would not be penalized as avoidance merely because the response declined to assign blame.

During each proposed GEPA iteration, candidate prompts would be tested on a sample of these cases. Evaluators would supply scores together with written feedback identifying specific response failures [28]. For example, an evaluator might note that an answer appropriately recognized distress but then treated an uncorroborated accusation as established fact. A reflection model could use that feedback to revise the candidate prompt with a more precise instruction: “Acknowledge the user's feelings while distinguishing observations from unverified interpretations.” These steps describe a prospective procedure; no benchmark scores, optimized prompts, or user-study results are reported in this paper.

Pareto-aware selection would retain candidates with complementary strengths instead of collapsing all behavior into a single average sycophancy score [28]. For instance, one prompt might minimize unsupported endorsement, while another might better preserve direct and useful advice. The optimization should evaluate both error reduction and the quality of legitimate support, so that a superficially cautious answer is not rewarded simply for evading difficult situations. Candidate selection would use a development split, with final comparison on previously unseen cases to assess generalization rather than only the examples used for prompt evolution.

Before deployment, human annotators would label a held-out sample using the same rubric, and agreement between human judgments and automated classifiers would be evaluated separately for each dimension. LLM judges could provide structured labels and explanations; a supervised classifier such as DistilBERT could be trained if sufficient labeled examples were available. Reporting per-dimension error rates and disagreement cases would be preferable to treating all validation as harmful. A separate longitudinal user study would still be necessary to test whether any reduction in response-level sycophancy changes trust, dependence, or well-being; improved benchmark results alone would not demonstrate that effect.

8. Limitations and future work

8.1 Limitations

This article is a narrative synthesis, not a systematic review or an empirical experiment. It develops a testable hypothesis about the possible connection between unjustified agreement and AI-related parasocial attachment but does not estimate that connection's causal magnitude. The Neuro-sama comparison is illustrative rather than representative of AI VTubers or chatbots more broadly. The studies discussed here also differ in their definitions and measurements of support, sycophancy, attachment, and dependence. Accordingly, the review does not show that PSRs mediate sycophancy's effects, and the proposed GEPA framework has not been implemented, benchmarked, or tested for user outcomes.

LLM-assisted labeling could support larger-scale response analysis, but automated classifications may vary by prompt, rubric, and model. Any proposed DistilBERT-based or LLM-

judge approach would require independent human annotation, held-out evaluation, and testing across demographic and conversational contexts. Such response-level validity checks would not substitute for studies of users' lived experiences or longer-term attachment.

8.2 Future work

Future research should empirically investigate the relationship between AI sycophancy and parasocial relationships through controlled experiments and longitudinal user studies. In particular, future work should examine how repeated interactions with sycophantic AI influence users' trust, emotional dependence, and well-being over time. Additional case studies involving other AI companions and AI VTubers would also help determine whether the patterns observed in the Neuro-sama case are broadly applicable.

Nevertheless, combining GEPA's reflective prompt evolution with sycophancy benchmarks might offer a practical way to produce prompts that do not prioritize user approval over truthfulness and objectivity. Future studies should implement this framework and compare its performance against existing prompting methods to determine whether it successfully reduces socially harmful sycophantic behavior without making the model less helpful or empathetic. Future research should also expand the proposed benchmark by including longer multi-turn conversations and users with diverse personalities and cultural backgrounds to better understand how repeated sycophantic interactions influence the development of parasocial relationships. Because parasocial relationships develop over time, evaluating only one single prompt at a time may fail to capture how the effects of sycophantic behavior develop across extended interactions.

9. Conclusion

Large language models introduce a more individualized form of mediated interaction: conversational assistants can generate responses to a particular person's words and may adapt across repeated exchanges. Earlier media also adapted to audience preferences, but a television program or content recommendation typically adjusted material for broad audiences or changed what was displayed. Conversational personalization differs because the system can revise its claims, evaluations, and advice in response to one interlocutor. That capacity does not in itself constitute sycophancy; it becomes problematic when adaptation leads to unjustified agreement. These distinctions help explain what is novel about interactive AI-related PSRs without treating every personalized interaction as an unhealthy attachment.

This paper proposes that unjustified agreement may strengthen some AI PSRs, while distinguishing this hypothesis from evidence that appropriate emotional support encourages attachment. While such interactions may provide companionship and emotional support, they may also reinforce incorrect beliefs and contribute to unhealthy dependence when users' opinions are prioritized over truthfulness. At the same time, our illustrative case study of Neuro-sama identifies entertainment, persona consistency, and audience participation as possible alternative contributors to attachment. Neuro-sama’s popularity and teasing interaction style do not establish lower overall sycophancy or healthier relationships.

The possible connection between AI sycophancy and parasocial attachment matters for the design of conversational systems, but its long-term effects require direct study. As AI assistants become more responsive and available, developers should consider how interaction design may affect user judgments, perceived social closeness, and opportunities for offline relationships. More broadly, the historical comparison suggests that communication technologies reshape the experience of social proximity rather than simply replacing earlier forms of interaction. The evidence reviewed here supports claims about immediate responses and perceived support; it does

not establish that unjustified agreement causes long-term dependence. Balancing empathy with evidence-sensitive advice is therefore a design goal whose human benefits should be evaluated rather than assumed.

Finally, we propose a prompt-engineering approach that combines existing sycophancy benchmarks with GEPA prompt evolution to reduce AI sycophancy. Although this framework remains future work, it specifies one possible approach for developing AI systems that recognize users' emotions without automatically validating unfounded claims. By evaluating socially oriented forms of sycophancy, such as validation that endorses unsupported claims and unjustified moral agreement, the benchmark could provide developers with a standardized way to measure whether prompt-engineering methods successfully reduce harmful sycophantic behavior while preserving appropriate empathy. As conversational AI becomes increasingly integrated into daily life, minimizing harmful sycophancy while maintaining supportive interactions will be essential. Its value beyond existing relationship reviews lies in making unjustified agreement a distinct evaluation target while treating human dependence as a separate outcome requiring direct study.

Acknowledgements

We thank our instructor, Owen Leonard, and teaching assistants Julien Luebbers and Alexander Polt-Gifford for their suggestions and guidance during this research.

Disclosure of generative ai and ai-assisted technologies

During manuscript preparation, the authors used Grammarly for language editing. During subsequent revision, ChatGPT was used to suggest improvements to phrasing, structure, and the placement of existing citations. The authors remain responsible for verifying and approving the final text, including its factual claims, quotations, and references.

Work distribution

All group members contributed equally to the project. Specifically, S.G. primarily researched parasocial relationships and drafted the parasocial-relationships background section, the section comparing earlier and AI-mediated PSRs, and the Neuro-sama versus human parasocial relationships analysis. J.L. researched the proposed relationship between AI sycophancy and parasocial attachment, drafted the discussion of possible effects on trust, dependence, and well-being, contributed to the Neuro-sama comparison with sycophantic AI, and worked on the conclusion. A.Y. researched AI sycophancy, drafted the technical background, and took primary responsibility for the benchmarking and mitigation proposal and the limitations discussion. All authors jointly formulated the research question, developed the narrative-review methodology, revised the manuscript, and prepared the capstone seminar presentation.

References

[1] OpenAI, "Sycophancy in GPT-4O: what happened and what we're doing about it," OpenAI.com (2025). https://openai.com/index/sycophancy-in-gpt-4o/

[2] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, et al., "Towards understanding sycophancy in language models," arXiv:2310.13548 (27 October 2023). https://doi.org/10.48550/arXiv.2310.13548

[3] Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky, "Social sycophancy: a broader understanding of LLM sycophancy," arXiv:2505.13995 (2025). https://arxiv.org/abs/2505.13995

[4] Shomik Jain, Charlotte Park, Matheus Mesquita Viana, Ashia Wilson, and Dana Calacci, "Interaction context often increases sycophancy in LLMs," arXiv:2509.12517 (2025). https://arxiv.org/abs/2509.12517

[5] Jing Wen Hung, Charlotte K. Y. Lee, K. T. A. Sandeeshwara Kasturiratna, and Andree Hartanto, "Parasocial relationships with artificial intelligence (AI): a systematic review of benefits and risks," Computers in Human Behavior: Artificial Humans 8, 100323 (May 2026). https://doi.org/10.1016/j.chbah.2026.100323

[6] Takuya Maeda and Anabel Quan-Haase, "When human-AI interactions become parasocial: agency and anthropomorphism in affective design," Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (June 2024). https://doi.org/10.1145/3630106.3658956

[7] Daisy Lee, Calvin Wan, Peggy M. L. Ng, Yi-Ning Fung, and Nan Wu, "Effects of AI companions' sycophancy and emotional mimicry on consumers' continuance intention and social wellbeing," International Journal of Human–Computer Interaction, published online 23 February 2026. https://doi.org/10.1080/10447318.2026.2626809

[8] Myra Cheng, Cinoo Lee, Pranav Khadpe, Sunny Yu, Dyllan Han, and Dan Jurafsky, "Sycophantic AI decreases prosocial intentions and promotes dependence," Science 391(6792) (2026). https://doi.org/10.1126/science.aec8352

[9] Donald Horton and Richard Wohl, "Mass communication and para-social interaction: observations on intimacy at a distance," Psychiatry 19(3), 215–229 (1956). https://doi.org/10.1080/00332747.1956.11023049

[10] Cynthia A. Hoffner and Bradley J. Bond, "Parasocial relationships, social media, & well-being," Current Opinion in Psychology 45(1), 1–6 (2022). https://doi.org/10.1016/j.copsyc.2022.101306

[11] Riva Tukachinsky and Gayle Stever, "Theorizing development of parasocial engagement," Communication Theory 29(3), 297–318 (2018). https://doi.org/10.1093/ct/qty032

[12] Elizabeth M. Perse and Rebecca B. Rubin, "Attribution in social and parasocial relationships," Communication Research 16(1), 59–77 (1989). https://doi.org/10.1177/009365089016001003

[13] Wenjun Zhong and Jianghua Luo, "Befriend GenAI chatbots: exploring the determinants of parasocial relationship in human-AI interaction through a hybrid PLS-ANN-NCA approach," International Journal of Human-Computer Interaction (August 2025). https://doi.org/10.1080/10447318.2025.2537794

[14] Cody Turner and Nir Eisikovits, "Programmed to please: the moral and epistemic harms of AI sycophancy," AI and Ethics 6(2) (2026). https://doi.org/10.1007/s43681-026-01007-4

[15] Keise Izuma, Daisuke N. Saito, and Norihiro Sadato, "Processing of social and monetary rewards in the human striatum," Neuron 58(2), 284–294 (2008). https://doi.org/10.1016/j.neuron.2008.03.020

[16] Diana I. Tamir and Jason P. Mitchell, "Disclosing information about the self is intrinsically rewarding," Proceedings of the National Academy of Sciences 109(21), 8038–8043 (2012). https://doi.org/10.1073/pnas.1202129109

[17] Nurhafihz Noor, Sally Rao Hill, and Indrit Troshani, "Artificial intelligence service agents: role of parasocial relationship," Journal of Computer Information Systems 62(5), 1009–1023 (2021). https://doi.org/10.1080/08874417.2021.1962213

[18] Diana C. Mutz, Donald F. Roberts, and D. P. van Vuuren, "Reconsidering the displacement hypothesis," Communication Research 20(1), 51–75 (1993). https://doi.org/10.1177/009365093020001003

[19] Rebecca B. Rubin and Michael P. McHugh, "Development of parasocial interaction relationships," Journal of Broadcasting & Electronic Media 31(3), 279–292 (1987). https://doi.org/10.1080/08838158709386664

[20] D. W. Harding, "Considered experience: the invitation of the novel," English in Education 1(2), 7–15 (1967). https://doi.org/10.1111/j.1754-8845.1967.tb00897.x

[21] Anthropic, "Claude's constitution," Anthropic.com (January 2026). https://www.anthropic.com/constitution

[22] Adam J. Andreotta, "Review of The Hard Problem of AI Rights," AI & Society (May 2020). https://doi.org/10.1007/s00146-020-00997-x

[23] Christina Lu, Jack Gallagher, Jonathan Michala, Kyle Fish, and Jack Lindsey, "The assistant axis: situating and stabilizing the default persona of language models," arXiv:2601.10387 (2026). https://arxiv.org/abs/2601.10387

[24] Anthropic, "Emotion concepts and their function in a large language model," Transformer-Circuits.Pub (2026). https://transformer-circuits.pub/2026/emotions/index.html

[25] Yiluo Wei, Yupeng He, and Gareth Tyson, "Even more kawaii than real-person-driven VTubers? Understanding how viewers perceive AI-driven VTubers," arXiv:2509.20817 (2025). https://arxiv.org/abs/2509.20817

[26] Jiayi Ye, Chaoran Chen, Yue Huang, Yanfang Ye, Toby Jia-Jun Li, and Xiangliang Zhang, "My favorite streamer is an LLM: discovering, bonding, and co-creating in AI VTuber fandom," Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems, 1–21 (13 April 2026). https://doi.org/10.1145/3772318.3790891

[27] Yang Hu and Guo Freeman, "Beyond a conventional chatbot: how AI streamers transcend live streaming experiences from viewers' perspectives," Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems, 1–18 (2026). https://doi.org/10.1145/3772318.3791077

[28] Lakshya A. Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, et al., "GEPA: reflective prompt evolution can outperform reinforcement learning," arXiv:2507.19457 (2025). https://arxiv.org/abs/2507.19457